ADDITIVE MANUFACTURING OF SiC FOR NUCLEAR APPLICATIONS

Adam Bratten, Department of Materials Science and Engineering
Maalavan Arivu, Department of Materials Science and Engineering

ABSTRACT
The purpose of this research is to produce monolithic silicon carbide (SiC) and SiC fiber-reinforced SiC matrix (SiC/SiC) composites using the Ceramic On-Demand Extrusion (CODE) process. Presently, only feedstock trials have been performed. The most promising feedstock produced to date incorporates liquid polycarbosilane (PCS), SiC nanopowder, and oleic acid. A commercial 3D printer is being ordered which will then be modified with air extrusion nozzles for the CODE process.

1. INTRODUCTION
Silicon carbide (SiC) is a non-oxide ceramic material with high specific strength and thermal stability. Further, it retains its mechanical properties well across a broad thermal range. Integrating SiC fibers possessing a very large length-to-diameter ratio into a matrix has been shown to drastically improve the mechanical properties. This improvement occurs even for introduction of SiC fibers into a monolithic SiC matrix, known as a SiC/SiC composite.

Because of their strength, oxidation resistance, and thermal stability, SiC/SiC composites have been proposed for use in light water nuclear reactor systems [1]. However, current SiC and SiC/SiC articles are produced using chemical vapor deposition (CVD) or chemical vapor infiltration (CVI). This process can be slow and inefficient, so more effective methods of producing SiC/SiC composites are needed for feasibility in this field. Additionally, CVD is not capable of producing dense parts with a complex architecture. This limitation restricts the parts which could potentially be replaced by SiC/SiC based on feasibility of production.

One common answer for the form factor restriction on materials/components is additive manufacturing (AM). Unfortunately, many popular methods for additively manufacturing SiC (such as binder inkjet printing and selective laser sintering [SLS]) are not efficient and tend to produce ceramic bodies with low relative density and/or microstructural inconsistencies that adversely affect the mechanical properties [2].

Ceramic On-Demand Extrusion (CODE) is an AM process akin to fused deposition modeling (FDM) or the most commercialized form of AM, 3D printing [3-7]. It involves the continuous extrusion of a ceramic suspension to produce a single cross-section of a computer assisted design (CAD) model. The cross-section is then briefly subjected to infrared irradiation to enhance the drying profile before the proceeding cross-section is extruded. This allows for deposition of a high-solids loading paste to yield a highly dense finished body. The CODE process was selected as the primary candidate for this project because of the high final density of resultant parts. Thus, the high relative density exhibited by articles produced with CODE is a desirable trait for structural materials.

2. PROCEDURE
The polymer ceramic precursor polycarbosilane (PCS) was selected as the matrix for the CODE feedstock paste. This material polymerizes (cures) around 200°C and pyrolyzes to form SiC near 1200°C. To increase the viscosity of the PCS and promote formation of dense SiC upon pyrolysis, silicon carbide nanopowder was blended with the PCS. Oleic acid or polyethylene glycol (PEG) were added the suspension of SiC nanopowder and any crystalline PCS that formed during curing.

2.1. Feedstock Processing
The feedstock components were blended using a magnetic stir rod. The blend was placed under a Dymax UV lamp for 20-30 minutes to cure the PCS. The stir rod distributed the granules of solid PCS that formed during curing to prevent premature agglomeration. The gel that formed following this process was then loaded into plastic syringes, degassed, and stopped to facilitate air extrusion. The ratios of binder, matrix, and powder were adjusted to reach optimal rheology for deposition.
3. RESULTS

Preliminary tests with UV curing of pure PCS without stirring showed formation of a crystalline layer atop the liquid (Fig. 1). This crystalline layer is thought to have inhibited the curing processes of the liquid below it by absorbing the UV light.

With the addition of a stir rod during UV curing, the PCS formed a homogenous gel (Fig. 2). This gel continued to until solid even after the UV lamp was turned off. This indicates that the polymerization process taking place is self-sustaining after some point, since there is no additional energy input progressing the reaction.

Since the magnetic stir rod was effective at dispersing the PCS crystallites, SiC nanopowder was introduced to the system with ~5 weight percent oleic acid or PEG to disperse the powder and facilitate homogenous mixing. The result was a gray gel with a nearly identical viscosity to the pure PCS gel (Fig. 3). Thus, the dispersants were effective at producing homogeneous PCS-based gels that may be appropriate for the CODE process.

Figure 1: Formation of a crystalline film atop liquid PCS that prevented the continued curing of the remaining bulk.

Figure 2: The addition of a stirring rod produced the desired result of distributing crystallites and enabling the entire batch to polymerize.

Figure 3: Introduction of SiC nanopowder and organic liquid dispersants produced a flowable gel with comparable rheological behavior to the pure PCS.

Feedstock composition optimization led to a mixture of 53 weight percent PCS, 43 weight percent SiC nanopowder, and 4 weight percent oleic acid.

4. FUTURE WORK

The future goals of this project are outlined in the original proposal. The work performed thus far has served to set up the remaining project goals. The overall scope of work will follow the flow chart in Fig. 4. UV exposure time during the curing process will also be varied to produce the most workable gel and ensure predictable feedstock behavior during extrusion. The gels will be sintered at high temperature and characterized to assess the densities achieved using the compositions appropriate for CODE. Acquisition and modification of a commercial 3D printer will enable implementation of the CODE process using the gels that have been already formulated.

4.1. AM of SiC and SiC/SiC composite by the CODE process

The SiC green body will be printed using the CODE process. After printing, the SiC green body will be sintered in vacuum. The steps and parameters for AM of SiC will guide AM of SiC/SiC composite. For AM of SiC/SiC composite, SiC powders will be mixed with SiC short fibers, PCS, water, acetone and binder. The contents of SiC
short fibers, PCS, water, acetone and binder will be controlled and adjusted. Then SiC/SiC green body will be printed using CODE. After printing, the SiC/SiC green body will be sintered in vacuum.

The additively manufactured and sintered SiC and SiC/SiC composite will be subjected to visual inspection and density measurement. If cracks are visually detectable, the AM process parameters are not acceptable, and they need to be adjusted and improved. Similarly, if the density of the product is low, the AM process parameters need to be improved. The CODE process will be adjusted and improved until visually there are no cracks and high density of the products is achieved.

The additively manufactured and sintered SiC and SiC/SiC composite will be subjected to visual inspection and density measurement. If cracks are visually detectable, the AM process parameters are not acceptable, and they need to be adjusted and improved. Similarly, if the density of the product is low, the AM process parameters need to be improved. The CODE process will be adjusted and improved until visually there are no cracks and high density of the products is achieved.

4.2. Mechanical testing of additively manufactured sic and sic/sic composite

After SiC and SiC/SiC composite without cracks and with high density are additively manufactured, they will be subjected to room-temperature and high-temperature mechanical testing. SiC and SiC/SiC composite would be used at high temperature in nuclear reactors, and therefore high-temperature mechanical properties are important. The high-temperature mechanical testing will be carried out up to 1100°C. Specifically, compression and three-point bending tests will be performed. From compression tests, stress-strain curves will be obtained, and strength and strain-to-failure values will be determined. From three-point bending tests, stress-strain curves will be obtained and fracture toughness will be determined.

4.3. Microstructural characterization of additively manufactured sic and sic/sic composite

As microstructure governs mechanical properties, the microstructures of the additively manufactured SiC and SiC/SiC composite will be characterized in detail using X-ray diffraction, scanning electron microscopy, and transmission electron microscopy. X-ray diffraction will be performed to study the phase.

4.4. Comparison between additively manufactured SiC and SiC/SiC composite and their commercial nuclear grade counterparts

The microstructures and mechanical properties of additively manufactured SiC and SiC/SiC composite will be compared to those of their commercial nuclear grade counterparts produced by chemical deposition/infiltration, which are available from the vendor or in the literature. The mechanical properties of the additively manufactured SiC and SiC/SiC composite need to be reasonably similar to those of their nuclear grade counterparts, in order to qualify the potential nuclear applications of the additively manufactured SiC and SiC/SiC composite in the future.

4.5. Anticipated results

This research will help to assess necessary modifications to the CODE process for production of non-oxide ceramics. It will also facilitate the evaluation of properties and performance of the additively manufactured SiC and SiC/SiC composites for potential nuclear applications.

5. CONCLUSIONS

Feedstock optimization testing on a PCS-based gel has been performed. The optimal composition tested contained 53 weight percent liquid PCS, 43 weight percent SiC nanopowder, and 4 weight percent oleic acid. The mixture was cured under direct UV light for 30 minutes and then allowed to cool in air. This process yielded a highly viscous gel with an apparently homogeneous dispersion of PCS granules and SiC powder. The gel became a solid after sufficient cooling and degassing, so slight modification to the UV curing procedure is needed. Sintering of the
feedstock gel will be performed to ensure that densities required to serve as a functional structural material. A commercial 3D printer is being ordered that will then be fitted with air extruders and modified to execute the CODE process, at which point further mechanical and microstructural analyses can be performed.

6. ACKNOWLEDGMENTS
The authors would like to acknowledge the support of the Intelligent Systems Center for making this project possible. They would also like to thank Drs. Haiming Wen and Ming Leu for their guidance.

7. REFERENCES


ABSTRACT
Additive manufacturing of ceramics via laser direct deposition is particularly challenging owing to high thermal gradients and thermally induced high cracking tendency. Therefore, it is necessary to have an improved understanding of the effects of processing conditions and material compositions on the quality of deposited ceramic parts. For this paper, thin wall structures of high purity ceramics were fabricated with commercially pure and zirconia doped alumina powders. The effects of zirconia dopants, varying from 0 wt.% to 10 wt.%, were studied. The microstructure and compositions of the manufactured specimens were analyzed using scanning electron microscopy (SEM), energy dispersive x-ray spectroscopy (EDS), and electron backscatter diffraction (EBSD) to characterize grain size, orientation, and distribution. Grain size distribution varied within the deposited ceramic parts due to the non-uniform temperature distribution during printing. The added Zirconia was found to mainly accumulate within grain boundaries. A finer microstructure was observed with Zirconia as doping materials in the printed samples.

1. INTRODUCTION
High-purity alumina ceramics are among one of the mostly widely used ceramic materials. Excellent electrical insulation, compressive strength, corrosion and wear resistance, low density and bio-inertness make these materials lucrative for multiple applications. Alumina ceramics, though exhibiting high hardness and elastic modulus, have relatively poor fracture toughness. Mechanical properties of ceramic composites are closely related to their microstructure and compositions. Thus, to fabricate high-quality ceramic composites, it becomes necessary to study and subsequently control these aspects. Toughening effect has been observed during previous studies in additively manufactured ceramics that were doped with zirconia [1]. This has been attributed to fine microstructure within the resulting enhanced microstructure of the ceramic composites. The anisotropy between alumina and zirconia maintain zirconia localized around alumina grain boundaries. This instigates crack bridging and deflection phenomenon which aids improvement in toughness [2], [3].

Laser direct deposition was used for fabrication of the samples. Relatively high efficiency, simplicity of the set-up coupled with ability to print large samples made this method a preferred approach for fabrication. Also known laser engineered net shaping, this additive manufacturing process involves ceramic powder being fed into melt pool. The powder thus melts and solidifies into a dense ceramic deposit onto the substrate. This process is continued through layer-by-layer deposition to fabricate ceramic parts of desired dimension and geometry.

Previous additive manufacturing of alumina ceramics mainly focused on either pure ceramics or ceramic composites with the specific eutectic ratio [4], [5]. The effects of smaller percentages of zirconia dopant onto the microstructure of additively manufactured high-purity ceramics has not been thoroughly investigated. In this study, size, orientation and distribution of grains within additively manufactured zirconia doped alumina has been examined. Specimen comprising of pure alumina, 5 wt.% and 10 wt.% zirconia doped alumina was studied. To retain favorable properties of high-purity alumina ceramics, maximum dopant volume of 10 wt. % was investigated.

2. FABRICATION
2.1. Powder Preparation
Commercially pure (99.7%) alumina powder (A3500UG kindly provided by Almatix), was used for fabrication. The D-50 particle size for this powder is quoted as 2.2 µm. Partially stabilized zirconia (3 mol-% yttria) from TOSOH, was used as dopant. Zirconia doped alumina ceramic powder mixture was prepared through ball milling process where the powders were agitated in distilled water with zirconia media overnight before being sieved and processed using rotary evaporator to remove media and solvent. The processed powders were then dried in an oven at 100 ºC to eradicate moisture and facilitate powder flow during the printing process.

2.2. Experiment Setup
The additive manufacturing set-up comprised of a 10.6 µm wavelength, 1.7 kW computer controlled CO2 laser (Convergent Energy Ultimate Model) with Gaussian beam intensity profile. The laser was complemented by a 3-axis computer numerically controlled (CNC) motorized table. The laser beam’s spot-size was maintained at 2.5 mm as a focal distance to 127mm ZnSe lens that was employed. This spot-size allows adequate resolution for deposition and at the same time facilitates clear observation during the parametric study. X2W powder feeder by Powder Motion Labs was used to feed the powder into the melt pool.
AD998 (99.8%) pure alumina plates by CoorsTek were used as substrate. They provided improved thermal compatibility between the deposited alumina tracks and substrate. A silicon nitride ceramic heating element by Induceramic was used to preheat the substrate to 450 °C. This helped in reducing the thermal gradient and therefore, thermal stresses near the substrate-deposition interface which facilitated bulk fabrication through additive manufacturing.

The entire set-up was encased within a collapsible glove box which was purged with pure Argon gas to maintain neutral / controlled environment around melt-pool and subsequently, the deposited part. A 3D rendered depiction of set-up is illustrated in Fig. 1.

A single-track parametric study was conducted to optimize the processing parameters where the powder flow rate, laser power, spot size and scanning speed was varied. For our abovementioned set-up, the optimum parameters are enlisted in following

<table>
<thead>
<tr>
<th>Powder flow rate</th>
<th>Laser power</th>
<th>Laser spot size</th>
<th>Laser scan speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.85 g/min</td>
<td>275 W</td>
<td>2.5 mm</td>
<td>700 mm/min</td>
</tr>
</tbody>
</table>

Table 1. Processing parameters for laser powder forming

2.3. Sample preparation

To facilitate measurements of single-track geometries as well as SEM, EDS and EBSD characterization of bulk specimen, polished cross-sections were necessary. For this, the samples were imbedded within VeriDur acrylic and then cut using diamond saw to expose the cross-section. The samples were then polished in accordance to the ASM standards [6].

3. RESULTS AND DISCUSSION

3.1. Microstructural analyses of thin walled structures

Fractured surfaces of thin walled structures fabricated using abovementioned additive manufacturing process were observed under SEM to characterize their microstructures. FEI Quanta 600F Environmental SEM was used for observation. Specimens were sputter coated with 25nm thick platinum coating to minimize charging and obtain better image resolution. Secondary electron (SE) images as well as backscattered secondary electron images (BSE) were studied during preliminary analyses.

![Image of laser fabrication set-up with an (inset) image of actual setup near the substrate.](image1)

![Fig. 1. Illustration of laser fabrication set-up with an (inset) image of actual setup near the substrate.](image2)

![Image of SE (left) and BSE (right) images of pure (A, B), 5 wt.% zirconia doped (C, D), and 10 wt.% zirconia doped (E, F) alumina ceramics, respectively.](image3)

Fig. 2. SE (left) and BSE (right) images of pure (A, B), 5 wt.% zirconia doped (C, D), and 10 wt.% zirconia doped (E, F) alumina ceramics, respectively.

Comparing and contrasting between two images made it easier to distinguish different phases inherently present in doped specimen. A side-by-side comparison between the SE and BSE images for all three types of additively manufactured specimens is illustrated in Fig. 2. SE (left) and BSE (right) images of pure (A, B), 5 wt.% zirconia doped (C, D), and 10 wt.% zirconia doped (E, F) alumina ceramics, respectively: Fig. 2.

It is apparent from preliminary observation that the grain sizes for pure alumina samples are noticeably larger than doped
specimens. Microscopic pores were detected for all three variants. Spherical nature of the pores suggests trapped air as plausible explanation. Increase in scan speed and/or laser power is generally recommended to address this issue [7].

Prominent second phase can be observed in BSE images for 5 wt.% and 10 wt.% doped samples which was absent in pure. This illustrated the doping effects of zirconia through the rapid solidification during the fabrication process. The distribution of doping elements was further examined by EDS as shown in section 3.1.2. Further studies will be performed to investigate the effects of doping elements in controlling microstructure during solidification processes.

3.1.1. SEM characterization for grain size measurement
As aforementioned, discrepancy between the grain sizes amongst the specimen was prominent. Upon further investigation, it was determined that the grain sizes varied significantly within the thin wall structure, depending on the location. Four specific locations were observed under the SEM, schematic for which, along with grain structure, is illustrated in Fig. 3 and Fig. 4, respectively.

It can be observed that the samples with 5 wt.% doping yielded smallest grain size followed by 10 wt.%. Pure alumina specimens yielded largest average grain structures. The variation across the specimen in the grain size was also significant in the pure alumina specimen, with less variation observed in 10 wt.% specimen; 5wt.% specimen having almost consistent microstructure, irrespective of the location within the sample.

Generally, location ‘C’ is where the largest grains were observed, regardless of the type of specimen. This may be attributed to the large local thermal gradient due to argon propelled powder stream. In pure alumina samples, noticeable voids were observed, localized around location ‘A’. This phenomenon was absent elsewhere across pure samples and completely non-existent in doped specimens. Similarly, intra-granular porosity that was observed in pure samples, was largely eliminated in doped specimen, due to inherently smaller grain size and subsequently, denser material.

![Fig. 3. Schematic diagram illustrating specific locations at which the microstructures were observed.](image)

![Fig. 4. Microstructure of (a) pure, (b) 5 wt.%, and (c) 10 wt.% zirconia doped alumina at different locations (A, B, C, D) depicted in Fig. 3.](image)
3.1.2. EDS characterization to observe composition

The composition of the phases observed in the SEM images were identified using EDS. The primary composition observed in pure alumina specimen was, expectedly, aluminum and oxygen; whereas doped specimen showed aluminum and zirconia along with oxygen as their primary composition. The BSE images with EDS mapping for primary composition is illustrated in Fig. 2.

Peculiarly, zirconia was detected along alumina grain boundaries. This is a typical behavior in ceramic composites comprising of materials with different binary equilibrium phases. During solidification of molten ceramic power, zirconia particles precipitate out and settle along the grain boundaries, forming interfaces with alumina [8]. Irrespective of the variation in grain sizes, this effect was consistently observed across the specimen, regardless of location being observed within the sample.

![Fig. 2. BSE image and EDS mapping showcasing primary composition of the ceramic composites (a) pure alumina, (b) 5 wt.% zirconia doped alumina, and (c) 10 wt.% zirconia doped alumina.](image)

Previous studies have illustrated possible interfacial bridging and subsequent toughening of ceramics based on the position of the dopants within their microstructure [3], [9], [10]. Further studies will be performed to evaluate the effects of doping elements on cracking formation within the deposited ceramic parts.

3.1.3. EBSD characterization to quantify grain size distribution and orientation

EBSD mapping was plotted using FEI Scios DualBeam FIB SEM. The data was mapped near the ‘D’ region as illustrated in Fig. 3. The procured images were first segmented into grains using FIJI image analyses package and then analyzed using Dream3D. Each grain was fitted with ellipse, with major and minor axes. The axes were defined using ‘build direction’ as reference. Inclination per grain was measured and plotted, with clockwise orientation being assigned a positive value and anti-

![Fig. 3. EBSD analyses and ensuing grain volume frequency vs. orientation plots for (a) pure, (b) 5 wt.%, and (c) 10 wt.% zirconia doped alumina.](image)
clockwise as negative. Frequency vs. orientation histograms were thus generated for each specimen, normalized by total cross-sectional area. The height of the bar represents fraction of area oriented in the corresponding inclination angle on orientation axis.

Clearly, from the orientation and size distribution, it may be concluded that doping alumina specimen with zirconia has a noticeable change in the size and the orientation of the grains within ceramic. 5 wt.% doping led to reduced grain size and a much diverse orientation compared to pure and 10 wt.% samples. There was no dramatic distinction to be drawn between 10 wt.% and pure as far as the orientation and distribution is concerned. Only distinction being pure alumina ceramics were likely to be oriented perpendicular to the build direction while 10 wt.% zirconia doped specimen had slight inclination towards positive orientation along with the perpendicular growth similar to pure specimens. The average grain size distribution and orientation observed corroborates with the grain size measurement tallied in 3.1.1 as pointed out that the average size at position ‘D’ was similar for pure and 10 wt.% zirconia doped specimens.

Growth of the crystals usually occur in the direction with larger thermal gradient. With laser heating up from the top and heating element elevating the temperature from the bottom, largest thermal gradient was likely experienced by the grains in direction perpendicular to the build direction.

4. CONCLUSIONS
An additive manufacturing technique for high-purity alumina ceramics using a blown-powder approach and CO$_2$ laser was established in this study. The changes in microstructure owing to grain refinement in high-purity alumina ceramic through small quantities (< 10 wt.%) of zirconia doping was studied. Noticeable degree of change in general grain size and orientation was observed between additively manufactured pure and doped specimens. This opens new avenues with multiple applications of ceramic composites, while retaining the primary characteristics of major constituent. Further research would allow quantification that could enable optimization of grain sizes and orientation to yield additively manufactured, fully dense ceramic composites with controllable mechanical properties.

5. ACKNOWLEDGMENTS
The author appreciatively acknowledges the support of the Intelligent Systems Center for research presented in this paper; also, extends special thanks to Dr. Jeremy Watts for his assistance with specimen preparation for analyses.

6. REFERENCES

© 2019 Aditya R. Thakur
Distributed State Estimation and Tracking Scheme using Active Passive Sensor Networks

Akhilesh Raj, S. Jagannathan and Tansel Yucelen

Abstract—This paper proposes a novel adaptive neural network (NN) based distributed state estimation scheme for a heterogeneous sensor network (HSN) to estimate the state vector of an unknown nonlinear process/target by using sensed output when the target input remains unknown. The active nodes in the HSN can sense the target output based on the detection range. By using a connected graph the active nodes will communicate their estimated state vector from their adaptive NN observer to other passive nodes in the neighborhood that cannot sense the target so as to estimate the target state vector. Next, a subset of nodes in the HSN, referred to as mobile nodes, track the moving target using the estimated state information and a state feedback controller. Considering the communication exchange being dictated by a connected graph, it is shown that the distributed state estimation error, the NN observer weight estimation error, and the tracking errors are uniformly ultimately bounded. Simulation results verify the theoretical claims.

I. INTRODUCTION

Distributed state estimation of nonlinear systems using a heterogeneous sensor network (HSN) can have a wide range of applications in the field of military and civilian environments such as target tracking, search and rescue operations. Since, all the practical systems are inherently nonlinear, a nonlinear estimation scheme is needed. The target state estimation with linear dynamics using a HSN is addressed in our previous effort [1]. In contrast, a novel distributed state estimation scheme by a HSN for a moving target with nonlinear dynamics is proposed here.

In a HSN, the active nodes can sense the target output based on its sensing range while the passive nodes rely on information exchange of its neighbors for the estimation of target state vector despite unreliable communication and faulty sensors. Therefore, in a HSN, a distributed sensing and estimation scheme is needed wherein the nodes, irrespective of their role and location, should estimate the target state vector and communicate the estimated target state information to the nodes in their neighborhood. A neural network (NN) observer based distributed state estimation scheme is required in order to estimate the nonlinear target dynamics and its state vector. By designing an observer at each node in a HSN and using information exchange, we will be demonstrate that both active and passive node can converge close to the target state vector.

Observers for non linear systems have been studied in the literature [2]–[5]. In [2], the design of an observer and output feedback controller was accomplished independent of each other by using a high gain observer. Typically nonlinear observers require system dynamics and high gain observers can use the separation principle while in general nonlinear observers have to overcome it. A more generalized observer design involving a neural network, is introduced in [6] when the system dynamics are not known. However, NN observers require the input to be considered known, which is a stringent assumption. Motivated by this fact, in this paper, we relax the assumption of known input to the nonlinear system and propose a novel adaptive NN observer to estimate the state vector of a moving target with unknown dynamics. In [3] the distributed estimation of states in a multi-agent system with unknown inputs is explained. The estimation algorithm rely on the measured output and the known dynamics to estimate the states of the system. In case of the target tracking applications using the HSN, the knowledge about the target dynamics is not practical.

First we will be present the distributed state estimation scheme by designing a NN observer at each node of a HSN. The NN approximates the nonlinear dynamics of the target and the observer estimates its state vector by using only the measured/sensed target output and NN approximation. We will use the observer design similar to the approach given in [6] which was utilized in our previous effort [1]. The NN observer is tuned by using the measured output in the case of an active node. A dedicated connected graph will provide the communication between the active and passive nodes. The passive nodes in turn estimate the target state vector by using the information exchange.

Subsequently, a subset of nodes are considered mobile and by using a new communication topology, standard feedback controller and information exchange among active and passive nodes, the mobile nodes are shown to track the moving target with a bounded error. By using Lyapunov analysis, the tracking error, estimation error and NN observer weight estimation errors are shown to be ultimately bounded.

The main contributions of this paper include the: 1) development of an adaptive NN observer for estimating the state vector of a target with uncertain dynamics; 2) design of NN observer based distributed adaptive estimation scheme with fixed node roles in a HSN for both active and passive
nodes; 3) extension of the estimated scheme for tracking a moving target with uncertain dynamics when a subset of nodes is mobile and 4) verification of the proposed approach using simulation.

Next after presenting the notations and brief background, the problem statement and NN observer design at a given node in a HSN is introduced. Subsequently, the distributed state estimation scheme is presented.

II. NOTATIONS USED AND BACKGROUND

The mathematical notations used in this paper are fairly standard. Specifically, $\mathbb{R}$ is the set of real numbers, $\mathbb{R}^n$ depicts the set of $n \times 1$ real column vectors, $\mathbb{R}^{n \times r}$ is the set of $n \times r$ real matrices, $\mathbb{R}^{n \times n}$ represents a symmetric positive definite $n \times n$ matrix, $\mathbf{0}_n$ depicts an $n$ dimensional column vector of 0, $\mathbf{1}_n$ depicts an $n$ dimensional column vector of 1, $(\cdot)^T$ depicts the transpose operator, $(\cdot)^{-1}$ represents the inverse, $(\cdot)^+$ is the generalized inverse, $\lambda_{\text{min}}(A)$ is the minimum eigen value of the matrix $A$, $\lambda_{\text{max}}(A)$ is the maximum eigen value of the matrix $A$, $\text{diag}(\mathbf{a})$ is the diagonal matrix with vector $\mathbf{a}$ as diagonal, $\otimes$ depicts the Kronecker product, $\| . \|$ depicts the Euclidian norm, $\| . \|_F$ represents the Frobenius norm, $tr$ is the trace operator, $I_n$ is an $n \times n$ Identity matrix, $( . )^{(i)}$ depicts the $i^{th}$ derivative of a vector or a scalar.

$G$ represents the generalized undirected network graph, $d_i$ is the degree of the $i^{th}$ node, $\mathcal{A}(G)$ is the adjacency matrix of $G$, $\mathcal{D}(G) = \text{diag}(d)$, $\mathcal{d} = [d_1, ..., d_N]^T$, $\mathcal{L}(G)$ is the Laplacian of the graph $G$ computes as $\mathcal{L}(G) = \mathcal{D}(G) - \mathcal{A}(G)$, and $i \sim j$ depicts the neighborhood information of the $i^{th}$ node.

Fig. 1 represents the schematic of a heterogeneous sensor network depicting active and passive nodes. Given a moving target, the active nodes sense and exchange information to the passive nodes through a predefined communication protocol. Throughout the paper an HSN is represented by using a connected graph. A graph is connected if given any two vertices $P_i$, $P_j$, there is a path from $P_i$ to $P_j$. In this paper the connected graph $G$ defines the neighborhood information of network involving the active and passive nodes and the information exchange.

The graph $G$ in the estimation scheme uses the Laplacian matrix, which is a matrix associated with the consensus of information between the neighboring nodes in the graph $G$. It is represented by $\mathcal{L}(G)$ and is computed by subtracting the adjacency matrix $\mathcal{A}(G)$ of $G$ from the diagonal matrix $\mathcal{D}(G)$ whose diagonal elements are degree associated with each node in $G$. In other words, $\mathcal{L}(G) = \mathcal{D}(G) - \mathcal{A}(G)$ where $\mathcal{A}(G) \in \mathbb{R}^{N \times N}$ is the adjacency matrix with the element $a_{i,j} = 1$ if connection exists from $j^{th}$ node to the $i^{th}$ node and $a_{i,j} = 0$ otherwise, $\mathcal{D}(G) = \text{diag}(\mathcal{d})$, and $\mathcal{d} = [d_1, ..., d_N]^T$. Degree of the $i^{th}$ node $d_i$ is the total number of connections associated with that node.

An easy way to compute $\mathcal{D}(G)$ is by summing each element in the $i^{th}$ row of $\mathcal{A}(G)$ and writing it in the place of $a_{i,i}$ and making all other elements in the row zero. Here it is important to note that in the Laplacian matrix the sum of elements in a row is zero indicating $\mathbf{1}_n$ is the eigenvector of $\lambda = 0$ and it is a positive semidefinite for undirected graph [7]. It is also required that there should be at least one spanning tree associated with the graph to ensure convergence of estimated target information. In addition, for a sensor network with active and passive nodes, at least one node should be active to satisfy the requirement of collective observability as defined later in the paper. Finally, this translates to the existence of a graph for communication among the nodes.

III. PROBLEM STATEMENT

In this section we present the problem statement and nonlinear NN observer design for the state estimation scheme associated with each node in the HSN. Since the target dynamics are considered nonlinear and uncertain, a NN based function approximator will be employed in the adaptive observer design. In the first subsection, we design the NN observer to estimate the state vector of the target. Subsequently in the next subsection the problem statement for the proposed distributed estimation is explained.

A. Target Description and Observer Design

In this subsection state estimation for an unknown process or target with nonlinear dynamics is discussed. A nonlinear target dynamics of the form

$$\dot{x}(t) = f(x(t)) + h(x(t))u(t),$$

is considered where $x(t) \in \mathbb{R}^n$ is the state vector, $f(x(t)) \in \mathbb{R}^n$, $h(x(t)) \in \mathbb{R}^{n \times m}$ represent smooth nonlinear functions and $u(t) \in \mathbb{R}^m$ is the control input with $x(t)$, $f(x(t))$, $h(x(t))$ as well as $u(t)$ are considered unavailable to the HSN. The measured output $y(t) \in \mathbb{R}^p$ at each active node is given by

$$y(t) = Cx(t)$$

where $C \in \mathbb{R}^{p \times n}$ denotes the output vector that is considered known. For convenience, the time variable $t$ will be dropped from now on. The objective is to design a NN based observer for the target given by (1) subject to the observation given by (2) when both the target dynamics as well as its input are unknown to the HSN. Next we make the following assumption on the control law for the target.

Assumption 1: There exists a state feedback control law for the target given by

$$u = h^{-1}[Kx + r],$$

where $K \in \mathbb{R}^{n \times n}$ is an unknown gain matrix and $r$ is the reference input.

By using Assumption 1 we represent the target/ process dynamics in (1) as

$$\dot{x} = f(x) - h(x)u$$

$$= f(x) - Kx - r$$

(4)
Adding and subtracting $Ax$ in equation (4), we get
\[ \dot{x} = Ax + F(x) - r, \]
where $F(x) = -(A - K)x + f(x) \in \mathbb{R}^n$ is the uncertain dynamics of the target and $A \in \mathbb{R}^{n \times n}$ is a design matrix.
Under the assumption that $(A, C)$ is observable, the goal of the observer is to estimate the state vector of the target by approximating the unknown nonlinear function $F$ through a NN approximator in (5) by using the measured output. Next consider the adaptive observer of the form
\[ \dot{x} = Ax + W^T \sigma(x) + L(y - \hat{y}) \]
where $L \in \mathbb{R}^{n \times p}$ is a design matrix such that $A_e = A - LC$ is Hurwitz and $\hat{y}$ is a function estimate of $F(x)$. By employing the NN universal approximation property in (5) we get
\[ \dot{x} = Ax + W^T \sigma(x) - r + \epsilon \]
where $W \in \mathbb{R}^{n \times l}$ denotes the unknown constant target NN weight matrix. The function $\sigma(x) \in \mathbb{R}^l$ represents the NN activation function vector which forms a basis, $\epsilon \in \mathbb{R}^n$ is the NN reconstruction error and $l$ depicts the number of neurons in the hidden layer. Next by incorporating the NN approximator, rewrite the NN observer as
\[ \dot{x} = \dot{\hat{x}} = Ax + \tilde{W}^T \sigma(\hat{x}) + L(y - \hat{y}) \]
where $\tilde{W} \in \mathbb{R}^{l \times n}$ is the estimated weight matrix, $\sigma(\hat{x}) \in \mathbb{R}^l$ is the activation function of the NN which forms a basis. Therefore the target state estimation becomes obtaining the uncertain NN weight matrix $\tilde{W}$ in (8).

Now we define the state estimation error $e = \hat{x} - x$ and the output estimation error as $\tilde{y} = \hat{y} - y$. The estimation error dynamics become
\[ \dot{e} = \dot{\hat{x}} - \dot{x} = A_e e + \tilde{W}^T \sigma(\hat{x}) + W^T \sigma(x) - r + \epsilon \]
where $A_e = A - LC$ and weight estimation error given by $\dot{W} = \dot{\tilde{W}} - W$. Here we choose the activation function $\sigma$ such that it is Lipschitz continuous. Therefore we can express $\hat{\sigma} = \hat{\sigma}(x) - \sigma(x)$, in such a way that it satisfies $\hat{\sigma} < \rho e$ where $\rho$ is the Lipschitz constant. The error dynamics can therefore be rewritten as
\[ \dot{e} = A_e e + \tilde{W}^T \sigma(\hat{x}) + \rho W^T e - r + \epsilon \]
Next we will design a NN weight update law such that the stability of estimation error dynamics given by (10) is ensured. The following assumption is needed before we state the theorem.

**Assumption 2:** The target weights, $W$, activation function, $\sigma$, and reconstruction error $\epsilon$ in (7) are assumed to be bounded above such that $||\sigma|| \leq \sigma_M$, $||W|| \leq W_M$ and $||\epsilon|| \leq \epsilon_M$. The target output and its previous values as well as the desired trajectory $r$ are also considered bounded with $||y|| \leq y_B$ and $||r|| \leq \bar{r}$. We also assume that the Lipschitz constant $\rho$ and $W_M$ are known.

In the following theorem, it will be shown that the state estimation and weight estimation errors are bounded.

**Theorem 1:** Consider the target dynamics given by (1) approximated as (7) along with the NN dynamics given by (8). Let assumptions 1 and 2 hold. Then the observer state estimation and the NN weight estimation errors are uniformly ultimately bounded (UUB) when the actual NN weight matrix in (10) is updated as
\[ \dot{\tilde{W}} = -\beta \psi(C^+ E_1)^T P - \kappa \tilde{W} \]
provided the condition $\lambda_{\text{min}}(Q_0) > 2 \rho \lambda_{\text{max}}(P W^T)$ is satisfied, where $\beta, \kappa > 0$ represent user design parameter or learning rate, $C^+ \in \mathbb{R}^{n \times p}$ is the generalized pseudo inverse of $C$, $E_1 = [e_1 e_1^{(1)} ... e_1^{(n-1)}]$, $\psi = [\sigma(\hat{x}) \sigma(\hat{x})^1 ... \sigma(\hat{x})^{n-1}] \in \mathbb{R}^{l \times v}$, $\sigma_i \in \mathbb{R}^l$ depicts the $i^{th}$ derivative of the vector $\sigma(\hat{x})$ and $P \in \mathbb{R}^{n \times n}$ is the solution of the Lyapunov equation given by $A_i^T P + PA_i = -Q_0$ where $Q_0 \in \mathbb{R}^{n \times n}$ is a positive definite design matrix.

The NN observer from this subsection is extended to the case of distributed state estimation of HSN where spatially located nodes will estimate the position of a moving target with nonlinear dynamics. The following subsection extends the design of the observer to sensor networks under the assumption of collective observability.

**B. Distributed State Estimation**

Consider a sensor network with $N$ nodes connected over a network capable of locally exchanging information, according to an undirected graph topology given by $G$. If each node is subject to observation, we can write
\[ y_i = C_i x, \]
where $y_i \in \mathbb{R}^p$ and $C_i \in \mathbb{R}^{p \times n}$ denote observed output associated with the $i^{th}$ node and the associated node modality. If non identical sensor modalities are considered then each $C_i$ will be different for each node in the heterogeneous network.
Here each node has to estimate the state vector using both the observed output (12) for the case of active nodes and local exchange of information using the consensus term and communication topology dictated by the connectivity graph in the case of passive nodes. This attribute is even more important in the case of time varying nodes with different node modalities where all the nodes should have an agreement about the estimate of the target position [8]. Therefore the proposed distributed state estimation scheme should be capable of accommodating node modalities–fixed as well as time varying roles of the nodes as described in the next section.

**IV. DISTRIBUTED STATE ESTIMATION USING NN OBSERVER**

In this section the observer designed in the previous section is used for the distributed state estimation in HSN by assigning the roles apriori–active and passive. An active node can measure the target output defined by $y_i$, given by (12). When a node is passive it will not measure the target output i.e $y_i = 0_p$, whereas it can rely on the neighborhood for information exchange of estimated target state vector. The estimation scheme associated with each node is proposed in

this section and is proven stable by using a novel NN weight update law based on [9].

For target dynamics described by (1), consider a sensor network with \( N \) nodes whose measured output is given by (12) for an active node and zero for a passive node. Now define the NN based adaptive observer at an \( i^{th} \) node for the distributed estimation of target state vector as

\[
\dot{x}_i = A_i \dot{x}_i + \tilde{W}_i^T \sigma(\dot{x}_i) + g_i \tilde{L}_i (\dot{y}_i - y_i) - \alpha \sum_{j=1}^{N} (\dot{x}_i - \dot{x}_j)
\]

\[
\hat{y}_i = C_i \dot{x}_i
\]

where \( \dot{x}_i \in \mathbb{R}^n \) is a local state estimate of \( x \) for the \( i^{th} \) node, \( A_i \) is any asymptotically stable matrix, \( L_i \in \mathbb{R}^{n \times p} \) design matrix associated with \( i^{th} \) node such that \( A_i - L_i C_i \) is Hurwitz, \( \alpha > 0 \), \( y_i = 1 \) for active nodes and \( y_i = 0 \) for passive nodes. The value of \( g_i \) limits the measured output availability to the active nodes while the passive nodes have to rely on the local information exchange provided by \( \sum_{j=1}^{N} (\dot{x}_i - \dot{x}_j) \). By augmenting the local information to the vector \( \dot{x} = [\dot{x}_1 \ldots \dot{x}_N]^T \in \mathbb{R}^{Nn} \), stacking the measured outputs associated with all nodes into \( \dot{y} = [\dot{y}_1^T \ldots \dot{y}_N^T]^T \in \mathbb{R}^{Pn} \) and stacking the measured outputs associated with all nodes into \( y = [y_1^T y_2^T \ldots y_N^T]^T \in \mathbb{R}^{Pn} \) equation (13) becomes

\[
\dot{x} = A \dot{x} + \tilde{W}^T \Sigma(\dot{x}) + GL(\dot{y} - y) - \alpha \mathcal{G} \dot{x}
\]

\[
\dot{y} = C \mathcal{X}
\]

where \( \mathcal{W} = diag([W_1 \ldots W_N]) \in \mathbb{R}^{n \times Nn} \), \( \Sigma = [\sigma(\dot{x}_1)^T \ldots \sigma(\dot{x}_N)^T]^T \in \mathbb{R}^{n \times N} \), \( L = diag([L_1 \ldots L_N]) \in \mathbb{R}^{n \times pN} \), \( G = diag([g_1 \ldots g_N]) \in \mathbb{R}^{n \times Nn} \), \( \mathcal{A} = diag([A_1 \ldots A_N]) \in \mathbb{R}^{n \times Nn} \), and \( \mathcal{G} = \mathcal{L} \otimes I_n \in \mathbb{R}^{n \times n \times Nn} \). \( N \) corresponds to the number of sensing nodes and \( \mathcal{L} \in \mathbb{R}^{n \times n} \) is the Laplacian associated with the node connectivity graph \( \mathcal{G} \).

In the communication topology considered throughout this paper the graph \( \mathcal{G} \) is undirected, which implies that the connected nodes are capable of communicating among each other in either directions. Therefore, the Laplacian matrix \( \mathcal{L} \) as well as the \( \mathcal{G} \) are symmetric positive semi definite [7].

Stacking the error variables, \( e = [e_1^T, e_2^T, \ldots, e_N^T]^T \in \mathbb{R}^{nN} \), and operators associated with all the nodes we get

\[
\dot{e} = \dot{\mathcal{A}} e + \tilde{W}^T \Sigma(\dot{x}) - \alpha Ge + \mathcal{W}^T \Sigma - r - \varepsilon
\]

where \( \tilde{W} = diag([\tilde{W}_1 \ldots \tilde{W}_N]) \in \mathbb{R}^{n \times nN} \), \( \varepsilon = [\varepsilon_1 \varepsilon_2 \ldots \varepsilon_N]^T \in \mathbb{R}^n \) is the stacked approximation error such that \( \varepsilon \leq \varepsilon_M \), \( r \in \mathbb{R}^{n \times N} \) is the reference trajectory vector with \( r \) stacked \( N \) times and \( \Sigma = \Sigma(\dot{x}) - \Sigma(x) \). The following lemma and assumption is required before we proceed.

**Lemma 1:** (Proposition 8.1.2, Bernstein(2009)). If \( P_1 \in \mathbb{R}_+^n > 0 \) and \( S_1 \in \mathbb{R}^n \geq 0 \), then \( P_1 + S_1 > 0 \).

**Assumption 3:** Using Lemma 1 there exists a connectivity graph with Laplacian \( \mathcal{L}(\mathcal{G}) \) such that \( \Omega = (pG + \mathcal{G}P) \) is a symmetric positive semidefinite matrix with \( P \) being a user defined positive definite matrix and \( \mathcal{G} \) being the Laplacian [7].

Next in the following theorem for the given graph topology considered in a HSN, the state estimation and NN weight estimation errors are demonstrated to be bounded.

**Theorem 2:** Consider the target dynamics given by (1) modified as (7) and the distributes state estimation algorithm given by (IV). Let the Assumptions 1-3 hold and the local information exchange over the HSN follows the graph topology \( \mathcal{G} \). Then the state estimation error \( \varepsilon \) and the NN weight estimation error \( \tilde{W} \) are UUB when the unknown weight matrix is updated as

\[
\dot{\tilde{W}} = -\beta(G\Psi(C^+e_1)^TP - \kappa \tilde{W})
\]

provided the conditions \( \lambda_{\min}(Q_0) > 2\lambda_{\max}(\Gamma) + 1 \) and \( 2\kappa > \lambda_{\max}((I - G)P) \) are satisfied, where \( \Gamma = PW^T \rho \) and \( \rho \) is the Lipschitz constant, \( C^+ \in \mathbb{R}^{nN \times pN} \) is the pseudo inverse of \( C, C = diag([C_1 \ldots C_N]) \in \mathbb{R}^{pN \times nnN} \) and \( e_1 = [\hat{y}(i) \ldots \hat{y}(v-1)]^T \in \mathbb{R}^{Pn \times vN} \), \( \Psi = \Sigma(\dot{x}) \Sigma(\dot{x})^T - \Sigma(x)^T \in \mathbb{R}^{nN \times nN} \), \( \hat{y}(i) \) is the \( i^{th} \) derivative of the variable \( \hat{y} \), \( v \) is the observability index, and \( \beta > 0 \) and \( \kappa > 0 \) are user defined gain constants.

**V. TARGET TRACKING**

In the previous section, it was also proven that the state estimation and the NN weight estimation errors are bounded with the bound being a function of user defined parameters. In this section we will demonstrate how the proposed adaptive distributed estimation scheme can be employed in the target tracking applications. By classifying a set of nodes in the HSN as mobile, and by using the linear node dynamics for these mobile nodes we will demonstrate that the tracking error as well as the state and the parameter estimation errors are bounded.

Here we consider a subset of nodes, that are classified mobile, consisting of \( M \) nodes with the associated state vector and the system matrices carrying the subscript \( q \), where \( q = 1, 2, \ldots, M \). Similarly the connected graph \( \mathcal{G}^\prime \) dictates the flow of information among the \( M \) mobile nodes. Here we define a dedicated network \( \mathcal{G}^\prime \), exclusively for the set of mobile nodes to ensure the presence of a spanning tree for the information exchange among them. This is important to account for the changes that can happen in the number of nodes involved in tracking the target. If all the nodes associated with the HSN are mobile, then \( \mathcal{G}^\prime = \mathcal{G} \).

In the case of tracking the estimated target output \( \hat{\gamma} \) is used as the desired trajectory for the passive nodes that are mobile while the measured output as the desired trajectory for the active nodes that are mobile. The constant feedback gain \( K_\gamma \in \mathbb{R}^{m \times n} \), associated with the tracking controller of the \( q^{th} \) mobile node, can be easily designed with the knowledge of the node system matrices using pole placement or any other linear techniques such that the closed loop dynamics of the controlled system are asymptotically stable.

Now consider the target dynamics given by (1) and the state estimation scheme given by (14) with the NN update law given by (16). The mobile nodes are allowed to track the target. The dynamics for the \( q^{th} \) active node is given by...
where \( z_q \in \mathbb{R}^n \) is the state vector (position and velocity) associated with the \( q^{th} \) mobile node, \( \xi_q \in \mathbb{R}^{n \times n} \) and \( \zeta_q \in \mathbb{R}^{n \times n} \) represent the known system matrices associated with the mobile node, \( \omega_q \in \mathbb{R}^m \) is the state feedback control input associated with the \( q^{th} \) mobile node that is given by

\[
\omega_q = \pi_q z_q + \hat{y}_q.
\]

In the above equation, \( \pi_q \) is the constant state feedback gain such that \( \xi_q - \zeta_q \pi_q \) is Hurwitz. In equation (18), if the \( q^{th} \) mobile node under consideration is active, the estimated trajectory, \( \hat{y}_q \), can be replaced with the measured target output \( y \). Therefore equation (18) can be expressed as

\[
\omega_q = \pi_q z_q + y \tag{19}
\]

Combining equations (18) and (19) we can write the state feedback control \( \omega_q \) by incorporating the fixed node roles as

\[
\omega_q = \pi_q z_q + g_q y + (1 - g_q) \hat{y}_q \tag{20}
\]

where \( g_q \) is the classifying variable associated with the \( q^{th} \) mobile node that differentiates the active and passive nodes; i.e., \( g_q = 1 \) for the active nodes and \( g_q = 0 \) for the passive nodes. Therefore by substituting (20) in (17) we get

\[
\dot{z}_q = (\xi_q - \zeta_q \pi_q) z_q + \zeta_q g_q y + \zeta_q (1 - g_q) \hat{y}_q = \Xi z_q + \zeta_q g_q y + \zeta_q (1 - g_q) \hat{y}_q \tag{21}
\]

Since a subset of overall nodes track the target, the stability analysis is given for these mobile nodes which can be either active or passive. Now we define the tracking error for the \( q^{th} \) mobile node as \( \tilde{x}_q = z_q - x \). Therefore computing the tracking error dynamics we get

\[
\dot{\tilde{x}}_q = \Xi_q \tilde{x}_q + \zeta_q g_q y + \zeta_q (1 - g_q) \hat{y}_q - Ax - W^T \sigma(x) - r - \epsilon
\]

Adding and subtracting the terms \( \Xi \tilde{x}_q \) and \( B_q y \) to (22) and rearranging terms we get

\[
\dot{\tilde{x}} = \Xi_q \tilde{x}_q + (\Xi_q - A + B_q C_q) x + (1 - g_q) B_q C_q e_q - W^T \sigma(x) - r - \epsilon \tag{23}
\]

where \( \dot{\tilde{x}} = \Xi_q \tilde{x}_q + T_q x + (1 - g_q) J_q e_q - W^T \sigma(x) - r - \epsilon \). Now, rearranging the error variables associated with all the mobile nodes we get

\[
\dot{\tilde{x}} = \Xi \tilde{x} + T x + G^T M e - W^T M \Sigma M - r - \epsilon \tag{24}
\]

where \( \tilde{x} = \begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & ... & \tilde{x}_M \end{bmatrix}^T \in \mathbb{R}^{nM} \) is the augmented target state vector with the states stacked \( M \) times \( \Xi = A_0 \otimes I_M \in \mathbb{R}^{nM \times nM} \). Now, using the following theorem, we demonstrate that by using the weight update law given by (16) the state estimation, the NN weight estimation error and the tracking errors are bounded.

**Theorem 3:** Consider the target dynamics given by (1) and modified as (7). Let the distributed state estimation scheme for the set of mobile nodes be given by

\[
\dot{x}_M = A_M x_M + W^T M \Sigma M (x_M) + G_M L_M (y_M - y_M) - \alpha G^T M x_M
\]

where \( \dot{x}_M, y_M \in \mathbb{R}^{nM} \), and \( A_M, G_M, G_M \in \mathbb{R}^{nM \times nM} \). Let the Assumptions 1 through 3 hold and the local information exchange over the HSN follows the graph topology \( G \). Then the state estimation error \( e \), the NN weight estimation error \( \hat{W} \) and the tracking error \( \tilde{x} \) are uniformly ultimately bounded (UBB) when the unknown NN weight matrix is updated as

\[
W_M = -\beta (G^T M \Psi M (C^+_M e_1 M) )^T P_1 - \kappa \hat{W}_M \tag{25}
\]

provided the conditions given by \( \lambda_{min}(G_1^T > 2 \lambda_{max}(\Gamma) + 1 + \lambda_{max}^2(P_2 G_2^T M, J), \lambda_{min}(G_1^T > 0) \) and \( 2 \kappa > \lambda_{max}^2 (G_1^T M) \) are satisfied, where \( C^+_M e_1 M \in \mathbb{R}^{nM \times nM} \) is the pseudoinverse of \( C \) and \( e_1 M = \begin{bmatrix} \hat{y}_M y^{(1)}_M y^{(2)}_M \ldots y^{(n-1)}_M \end{bmatrix}^T \in \mathbb{R}^{pM \times eM} \). \( \hat{y}_M^{(i)} \) is the \( i^{th} \) derivative of the variable \( \hat{y}_M \). This is the observability index, and \( \beta > 0 \) and \( \kappa > 0 \) are user defined gain constants.

Figure 2 shows the algorithm behind the distributed state estimation and tracking.

**VI. ILLUSTRATIVE NUMERICAL EXAMPLES**

In this section the performance of the proposed scheme is evaluated with examples. The evaluation will be based on the standard performance metrics such as trajectories of the target and the nodes with respect to time, the RMS error of each node vs time and the phase plane plots of the target.
1) Example 1: Consider the two link planar robotic manipulator as a target given in [9]. The state vector consists of the joint coordinates and the angular velocities in x and y directions. This example evaluates the performance of the distributed state estimation scheme in (14) when the target dynamics traces the trajectory given by 
\[ r = [3\sin(\frac{2\pi}{5}); 3\cos(\frac{2\pi}{5})]. \] 
The output was measured using 
\[ C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \]
The matrices 
\[ A = \text{diag}([-6 - 8 - 6 - 8]) \]
and 
\[ L = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix} \]
were chosen such that \( A - LC \) is Hurwitz. An HSN consisting of 12 nodes with 6 active and 6 passive nodes were considered for the simulation. The values of \( \beta, \kappa, \alpha \) and \( Q_0 \) were chosen 5, 0.01,20 and 10I respectively. A one layer NN with 14 neurons was used to estimate the nonlinearity and a sigmoidal activation function was used in the input layer to form a basis. The initial conditions were all chosen random. The simulation results in the Figure 3 show the actual and the estimated outputs and phase plane plots shows the actual and estimated trajectory tracked by the manipulator. The RMS error plotted with respect to time shows the speed of convergence.

\[ \text{Fig. 3. Distributed state estimation with time varying node roles: a) output trajectory b) Average of the RMS of the estimation errors associated with each node, c) phase plot.} \]

2) Example 2: In this example the proposed distributed state estimation scheme will be used for tracking. The target system and the HSN architecture are taken exactly same as that of the previous example. All the nodes of the HSN are considered mobile in this example. A fourth order point mass system is considered as the mobile node dynamics and the feedback controller was designed using pole placement method with two poles each at -1.2 and -0.8. Figure 4 shows the outputs of the target and the mobile nodes. The convergence of the estimation and tracking errors are also shown in the figure.

VII. CONCLUSION

In this paper the distributed adaptive state estimation and tracking performance of active and passive nodes in a HSN for tracking a target with uncertain nonlinear dynamics is introduced. First an NN observer with consensus term was designed. The unknown weight matrix of the target dynamics was estimated by using the measured output provided the target is in the detection range of an active node. The passive nodes estimate the target state using information exchange via a connected graph provided at least one active node is present in the neighborhood.

The Lyapunov analysis demonstrated the boundedness of the tracking, estimation and NN weight estimation errors. It was observed that the number of active nodes play an important role in the speed of convergence and to approach the actual values both for estimation and tracking.

REFERENCES

VANET AND CLOUD-BASED SECURE TRAFFIC MONITORING

Ayan Roy
Department of Computer Science
Missouri University of Science and Technology
Rolla, MO 65409, USA
Email: ar3g3@mst.edu

Sanjay Madria
Department of Computer Science
Missouri University of Science and Technology
Rolla, MO 65409, USA
Email: madrias@mst.edu

ABSTRACT
In Vehicular Ad-hoc Networks (VANETs), neighboring vehicles communicate with each other and with the roadside infrastructure about the traffic using wireless communication. Therefore, message authentication and trustworthiness of messages become crucial for monitoring traffic. In this paper, we propose an edge cloud-based decision-making model to provide accurate traffic information for possible events like congestion at an ROI. Security of event-related responses, and velocity and GPS coordinates reported by neighboring vehicles at the time of a request by ROI for every vehicle are used for accurately determining the traffic flow. We evaluate the performance of the proposed model using VENTOS and SUMO simulators, and a simulated cloud environment where the consensus-based and the driver-rating based models fail to provide the correct prediction when the number of malicious vehicles is in majority. We also compare our model against an existing peer-based authentication model, and show that our model offers low communication overhead and better resiliency against security attacks.

Index Terms—Secure, Cloud-assisted, Traffic, VANET

1 INTRODUCTION
Vehicular Ad-hoc Networks (VANETs) aim to improve driver experience by allowing communication among vehicles for better traffic management [2]. Vehicles equipped with on board units (OBUs) can communicate via wireless communication with other OBUs as well as with the roadside infrastructure. However, the presence of malicious vehicles at an ROI can disrupt the traffic monitoring, and can persuade a system to infer incorrect traffic status. Malicious vehicles have the capability to launch attacks such as Sybil, Man in the middle, GPS data spoofing, hidden vehicle, and message tampering which are considered as the top 10 threats in VANET or Intelligent Transportation System (ITS) [5].

Although some significant amount of security work has been done in VANET/ITS, still some major challenges like satisfying conflicting goals such as authentication and anonymity, remains there [6]. GPS based applications can be used to offer alternative routes based on traffic information obtained from the local authorities. However, the information obtained from GPS does not guarantee privacy, and is not real-time because it does not provide information based on events that have just taken place [7]. Moreover, GPS information is not tamper-proof, and can be manipulated by a malicious vehicle to disrupt the traffic in a region or to impede other vehicles from entering the region for personal intent. Existing models mostly considers adversarial parsimony; a heuristic assuming that an attack involving a few malicious nodes is more likely than an attack that requires a collusion between a large number of malicious nodes [1]. However, there is a possibility of more malicious responses by collusion or Sybil attack than non-malicious responses, especially where the ROI is smaller or the number of vehicles is less.

In this paper, we propose an edge cloud-based traffic monitoring model with authenticating responses efficiently obtained from the vehicles, and provide an accurate traffic related information at an ROI. Our proposed model is able to handle scenarios where the number of malicious vehicles at an ROI outnumbers the non-malicious vehicles. Furthermore, we consider that malicious vehicles have the computational capability to send manipulated GPS data when requested. Furthermore, we compare with existing popular centralized approaches such as the reputation-based system [18], and the consensus-based system to show the resiliency and efficiency of the proposed model. We also make a comparison with a distributed solution such as peer-based authentication model [17] in VANET with respect to parameters such as the number of communications, message transmission time, and energy required for communication based on simulation parameters.

2 RELATED WORKS
To design secure VANETs, researchers have proposed threshold based authentication systems where they define a threshold for every vehicle that decides whether data is acceptable or not. Younes et.al. [8] proposed a method in which every registered vehicle at the ROI advertises information about the road condition. Every neighbor accepts the information only if it is valid. The vehicle closest to the RSU encrypts the data, and sends it to the RSU which determines the status of the traffic. However, the proposed model fails to respond in situations if the vehicle closest to the RSU maliciously drops the information or it purposefully accepts incorrect information, and transmits that to the RSU.
Shao et.al [9] proposed a dynamic threshold authentication technique in which every vehicle has a threshold for accepting messages. However, if the threshold for the vehicle is too high or too low, the model fails to perform efficiently. Caballero-Gill et.al [7] proposed a technique where the vehicles communicate with each other, and decide on an event. The authors propose a concept of reactive group where the vehicles at that group decides and agree on an event. The model fails to respond in case where the number of malicious vehicles are more.

Taie et. al [10] proposed a threshold based authentication, and event detection scheme in VANET. The model detects an event that has occurred at an ROI such as traffic congestion or accident. Furthermore, RSU keeps public-private key pair of each vehicle at the ROI, thus, needs to store a large number of public-private key pairs. It periodically broadcasts the revocation list about the possible malicious vehicles. The monitoring events are broadcasted by the vehicle based on certain criteria. At this stage, a malicious user can purposefully drop a message. Using the models, if majority vehicles are malicious then it will accept the monitoring event, and record the incorrect data. Hence, in such a scenario, the system fails.

3 SYSTEM MODEL, ASSUMPTIONS AND REQUIREMENTS

3.1 System Model

Figure 1 represents the architecture of our proposed model where there is a centralized server (cloud), an edge server for different regions, RSUs deployed at an ROI, and vehicles with installed with OBUs.

- **Edge Server**: It is associated with each region of a city, and deployed to analyze the traffic scenarios in a region like detecting events or regular traffic monitoring. It interacts with the vehicles deployed within an ROI under its supervision via the RSUs and securely detect the traffic condition based on the vehicular responses, thus, identifying the vehicles as malicious or non-malicious.

- **Centralized Server**: It is responsible for analyzing the traffic scenario of a city, state or a country. We assume that traffic congestion of a region has a cascading effect on other regions surrounding it. In order to efficiently study the traffic scenarios of a city, each of the edge servers sends their responses to the centralized server. It receives the responses of different ROIs from different edge servers, and determines the different traffic scenarios in a city.

3.2 Assumptions

We assume that no packet is dropped by the non-malicious vehicles, and each challenge and response packet is sent and received by the RSU as well as by the vehicles. The centralized server, and the edge server are trusted [16], and a majority of vehicles at the ROI participates in the decision-making. The requesting vehicle is assumed to be non-malicious. All the vehicles considered have a uniform id length. Every vehicle information that does not reach the edge server is not taken into consideration for the decision making. However, the edge server must receive at least one conflicting neighbor if there exists any malicious vehicle at the ROI. We further assume that every vehicle is registered with the centralized trusted authority with a unique public-private key pair, generated using the ElGamal digital signature scheme [15] associated with every unique vehicle id.

3.3 System Requirements

- **Decision Similarity Graph**: A new data structure called the Decision Similarity Graph (DSG) is maintained by the edge server in order to determine the malicious vehicles at an ROI. DSG is an ordered pair of the form:

  \[ DSG = (V, E) \]

  where \( V=V_1, V_2, V_3, \ldots, V_n \) represents the n vehicles from which the responses are received and \( E=E_1, E_2, E_3, \ldots, E_n \) represents the relationship among the vehicles based on their distances.

  Figure 2 represents a DSG maintained by the edge cloud. Every \( E_i \) connecting to \( V_i \) indicates that the two vehicles are within the transmission range. DSG is used by the edge server after the challenge-response step (which we will explain later in this paper) to identify the malicious as well as the non-malicious vehicles.

- **Schmidt-Samoa Cryptosystem [12]**: Every message exchanged between the vehicles and the edge server is encrypted using this public key cryptosystem. The algorithm is used at the edge server to generate the public-private key for encrypting the messages from the vehicles.

- **ElGamal Digital Signature Scheme [15]**: It is used to generate a private-public key pair corresponding to the unique \( V_{id} \) of each vehicle. The private key is kept by the vehicle for digitally signing a message while the public key is stored with the centralized trusted authority for authenticating the vehicle transmitting a signed message.
4 PROPOSED MODEL

Here, we present the steps in our proposed model. The abbreviations used have been explained in Table 1.

4.1 Service Request

The proposed model can handle both normal traffic monitoring, and ad-hoc requests. On the event of an ad-hoc request, the requesting vehicle, $V_{req}$ willing to enter the ROI, submits $req_{pkt}$ to the centralized cloud server of the form <$V_{req}, GPS_{req}, vel_{req}, ROI$>.

4.2 Key Generation by the cloud

The request from the vehicle is relayed to the appropriate edge server from the centralized server. The edge server generates a public-private key pair using the Schmidt-Samoa cryptosystem. Two large prime numbers ‘p’ and ‘q’ are chosen at random by the edge. $G_{pub}$ is generated using the equation:

$$G_{pub} = p^2 * q$$

while $G_{pri}$ is generated using the equation:

$$G_{pri} = G_{pub}^{-1} \mod (lcm(p-1, q-1))$$

and stored at the edge. It generates info_pkt=$<req, G_{pub}>$, and sends it to the RSUs deployed at an ROI.

4.3 Request Dissemination and Vehicle Response Acquisition

An RSU after obtaining the info_pkt through the wireless communication connecting to the edge server, broadcasts the data packet within its transmission range. Every vehicle deployed within the range of transmission of the RSU receives info_pkt.

A vehicle after receiving the info_pkt encrypts its $V_{id}$ using $G_{pub}$ extracted from the info_pkt, and sends it to nearby vehicles through a broadcast. It generates data_packet in the following form:

$$data_{packet} = < V_{id}, ds(V_{id}), cong, vel_{id}, GPS_{id}, enc_{id}(v_{neigh}) >$$

Every vehicle appends a digitally signed copy of its $V_{id}$ with the private key registered during the vehicle registration phase using the ElGamal digital signature scheme. Since each vehicle at an ROI integrates not more than predefined $n_{num}$ number of $enc_{id}(V_{id})$. Thus, some of the packets can be purposefully dropped by a malicious vehicle.

The edge server extracts $V_{id}$ and $ds(V_{id})$ from each of the data_packet. It uses the public key corresponding to the unique $V_{id}$ to generate $check_{id}$ from $ds(V_{id})$ and matches it with the $V_{id}$. Thereafter, it extracts the $V_{id}, vel_{id}, cong, GPS_{id}$ and $enc_{id}(v_{neigh})$ for every vehicle whose response has been received, and constructs the DSG in which every $V_{id}$ is the vertex and the edges connecting the vertex representing the neighborhood between vehicles. At this point, the server intends to find any possible POC among the vehicles deployed at the ROI. The algorithm to detect POC among the vehicles involves steps shown in Algorithm 3. If no POC is detected after the initial comparison with the neighbors, it checks the event recorded by $V_{id}$ under the same $RSU_{id}$ even if they are not neighbors.

If a POC is detected then the flag is raised, and the initial inspection based on the information obtained from the responding vehicles is performed following algorithm 4. If no decision can be made after an initial inspection, the server generates a challenge_pkt consisting of $V_{id}$ of the conflicting vehicle, expected coordinates based on GPS$_{id}$ of the conflicting vehicles, and time$_{id}$.

The steps to calculate the expected $RSU_{id}$ for the conflicting vehicles is shown in Algorithm 5.

$$challenge_{pkt} = < CV_1, expected_{rsu1}, time_1, CV_2, expected_{rsu2}, time_2 >$$

4.5 Challenge Packet Dissemination and Response Collection

When the RSU receives a challenge_pkt, it extracts $V_{id}$ for the assigned CV and time$_i$. It then generates a crypto
challenge', and broadcasts it within its transmission range after time. It further waits for a 'grace' amount of time to receive the 'crypto response'. The value of 'grace' is chosen to make reparation for any change in vel\textsubscript{CV} during the computation. The crypto\_challenge\_pkt is obtained using the XOR operation.

\[ \text{crypto}_\text{challenge}_\text{pkt} = CV_i \oplus \text{Decision}_\text{pkt} \]

where CV\textsubscript{i} refers to the conflicting vehicle assigned to the corresponding RSU. Every vehicle, within the transmission range of the RSU, receives crypto\_challenge\_pkt, and it generates the crypto\_response packet and broadcasts it.

\[ \text{crypto}_\text{response}_\text{pkt} = V_{id} \oplus \text{crypto}_\text{challenge}_\text{pkt} \]

Algorithm 2 POC detection

1: for every V\textsubscript{id} in data\_packet do
2: \hspace{1em} \text{neigh}[] = neighbors\{V\textsubscript{id}\}
3: \hspace{1em} for every neighbor in \text{neigh} do
4: \hspace{2em} if neighbor.cong != V\textsubscript{id}.cong then
5: \hspace{3em} POC\_detected = true
6: \hspace{2em} CV\textsubscript{i} = V\textsubscript{id}
7: \hspace{2em} CV\textsubscript{2} = neighbor
8: \hspace{2em} break all loops
9: \hspace{1em} end if
10: \hspace{1em} if POC\_detected = false then
11: \hspace{2em} Send "Everything is Ok"
12: \hspace{1em} end if
13: \hspace{1em} end for
14: \hspace{1em} end for

The crypto\_challenge\_pkt is obtained using leftshift and XOR operations.

\[ \text{Decision}_\text{pkt} = RSU_{id} \ll \text{leftnum} \]

where leftnum refers to the number of left shift operations performed. The RSU waits for 'grace' amount of time, extracts the information from the crypto\_response\_pkt, and matches it with the Decision\_pkt. Since leftnum is kept secret by RSU, and any keyword can be used by RSU apart from RSU\textsubscript{id}, it becomes difficult for CV\textsubscript{i} to brute-force the appropriate values of both. The RSU sends rsu\_response\_packet packet to the edge server.

\[ \text{rsu}_\text{response}_\text{packet} = \ll \text{rsu}_\text{id}, CV\textsubscript{i}, \text{response}_\text{obtained} > \]

where response\_obtained can be "received" or "not received".

4.6 Response Analysis and Decision Making

The edge server, on receiving the rsu\_response\_packet from the RSU, analyzes the response packet, and makes a decision about congestion, and flags vehicles as malicious depicted in Fig.4. Let us assume that the edge server concludes that the event at the ROI is event\textsubscript{i}. At this stage, the edge uses the DSG in order to find the neighbors of the CV having the same decision as event\textsubscript{i}, and considers them as non-malicious. On the other hand, every V\textsubscript{id} contradicting event\textsubscript{i} is flagged as malicious. This procedure is followed across the entire DSG until all the vehicles, i.e., the vertices in DSG are either flagged as malicious or non-malicious.

Algorithm 3 Initial Inspection based on vehicle information

1: \hspace{1em} for every vehicle v \in CV\textsubscript{i} do
2: \hspace{2em} if v.cong = "yes" then
3: \hspace{3em} if vel\textsubscript{v} \geq 50 then
4: \hspace{4em} DecisionMade = true
5: \hspace{4em} Decision = "No Congestion"
6: \hspace{3em} end if
7: \hspace{2em} else if v.cong = "no" then
8: \hspace{3em} if vel\textsubscript{v} < 50 then
9: \hspace{4em} for every neighbor \in neighbors[CV\textsubscript{i}] or neighbors[CV\textsubscript{2}] do
10: \hspace{5em} if neighbor.cong = "no" and vel\textsubscript{neighbor} \geq 50 then
11: \hspace{6em} CV\textsubscript{1} = "no" and vel\textsubscript{neighbor} \geq 50
12: \hspace{5em} Send \_challenge = true
13: \hspace{5em} end if
14: \hspace{5em} end if
15: \hspace{4em} end if
16: \hspace{3em} if Send\_challenge = false then
17: \hspace{4em} DecisionMade = true
18: \hspace{4em} Decision = "The road is Congested"
19: \hspace{3em} else
20: \hspace{4em} Generate\_Challenge\_Packet.
21: \hspace{4em} end if
22: \hspace{3em} end if
23: \hspace{1em} end if

4.7 Response from Edge to Centralized Server

The event, observed by the edge server deployed at a region is sent to the centralized server. The centralized server sends the response to the requesting vehicle in case of an ad-hoc request or is stored for traffic monitoring. The event recorded by every edge server is determined using the proposed model, and sent to the centralized server. The centralized server observes the traffic using the majority selection method, and decides the traffic scenario of multiple regions to be the one reported by majority of the edge servers. The purpose of the centralized server is to prevent any bottleneck that may result on account of enormous requests about different ROIs and allow graceful degradation.

5 EXPERIMENTAL RESULTS AND ANALYSIS

5.1 Experimental Configurations

To validate the effectiveness of the proposed model, several experiments were conducted using the VENTOS [13] simulator. The road network at ROIs and vehicles’ locations were simulated using SUMO [14]. The centralized server (cloud) was simulated in a separate workstation,
Windows 7 Enterprise 64 bit, Intel(R) Xeon(R) CPU E5-1620 v2 @ 3.70 GHz, and the edge server was simulated in a separate workstation, Windows 10, Intel Core i3-4005U CPU@ 1.7GHz(4 CPUs), and both servers were simulated under DevC++ environment.

Algorithm 4 Generating the challenge_packet

1: Initialize time $t$
2: for every $rsu \in RSUList$ do
3: if $CV_i.coords_x$ lies within $rsu.lowerCoord$ and $rsu.upperCoord$ then
4: $time_i = t$
5: $expected\_rsu_i = rsu$
6: else
7: $t++$
8: end if
9: end for

Fig. 3: Decision Tree for analysis

5.2 Experimental Results

We simulated congested as well as non-congested road scenarios, and compared them against the consensus-based model [1], the reputation-based model [18], as well as with the peer-based authentication model [17]. Table III gives the results of the experiments using these models.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Reputation-Based</th>
<th>Peer-Based</th>
<th>Consensus-Based</th>
<th>Proposed Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M = \text{5NM} = 5$</td>
<td>Works</td>
<td>Conditional</td>
<td>Works</td>
<td>Conditional</td>
</tr>
<tr>
<td>$M = \text{5NM} = 10$</td>
<td>Works</td>
<td>Conditional</td>
<td>Works</td>
<td>Does not Work</td>
</tr>
<tr>
<td>$M = \text{10NM} = 5$</td>
<td>Works</td>
<td>Conditional</td>
<td>Works</td>
<td>Conditional</td>
</tr>
<tr>
<td>$M = \text{10NM} = 10$</td>
<td>Does not work</td>
<td>Conditional</td>
<td>Works</td>
<td>Conditional</td>
</tr>
</tbody>
</table>

It can be seen from Table I that our proposed model outperforms both the consensus-based model and peer authentication model for most of the scenarios. The performance of the reputation-based model is highly dependent on the previous ratings of the vehicles. When the combined previous weight of the malicious vehicles is higher than the non-malicious vehicle, it fails to provide accurate results. This can happen when the malicious vehicles outnumber the non-malicious and hence, in such a scenario, the reputation-based model fails. Thus, the reputation-based model provides conditional responses dependent on the previous ratings of vehicles involved in the decision making process. Also, in case of a peer-based model, the efficacy of the model is highly dependent on the number of malicious vehicles that surrounds a non-malicious vehicle. This decreases as the non-malicious vehicles in the system increases.

As a result, the total transmission power required for the communication, and the total transmission time of the messages with respect to the simulation parameter is reduced in our proposed model. Figure 6 compares the proposed model with the peer authentication model based on the number of communications, transmission time and the transmission energy. Based on the results, we can conclude that our proposed model outperforms the peer authentication model based on resiliency against various attacks, communication, energy, and time.

6 SECURITY ANALYSIS

6.1 Defense against Confidentiality and Integrity attack

The proposed model provides resiliency against the integrity attack in the following ways:

6.1.1 Resiliency against data_packet attack

The data_packet is encrypted using $G_{pub}$ which is generated using the Schmidt-Samoa cryptosystem. In order to breach the integrity of the data, the attacker has to possess $G_{pri}$. Since it is kept secret by the edge server which is a trusted entity, and the attacker has to brute-force in order to retrieve data_packet of another $V_{id}$. As the difficulty is based on the integer factorization problem, it is not possible for the attacker to get the data in real-time.

6.1.2 Resiliency against crypto_challenge_packet attack

Assume that the size of the $RSU_{id}$ and the $V_{id}$ be uniform and call it, say len.

$enc\_space = a - z, 0 - 9!, @, #, $, % = 41$

where $enc\_space$ consists of the characters used to generate the Decision_pkt.

The number of possibilities for $Decision\_pkt$ is:

$\text{possibility} = 41 \times 41 \times 41 \times 41 \times 41 \times \cdots = 41^{\text{len}}$

For len = 5, the possibility is $115856201$. For a higher value of len, larger $enc\_space$, and different values of left_num which remains private to the assigned RSU, the value of possibility increases even more, thus, making it difficult for an attacker to perform a brute-force approach to obtain the Decision_pkt.

6.2 Defense against Sybil attack

To perform the Sybil attack, each malicious vehicle can impersonate to be some other vehicle, generate bogus information, and send the data to the cloud. Since these values are unique for every vehicle, we show the difficulty involved in guessing the above mentioned registered numbers for the vehicles.

Assume that the prime number for a vehicle has a sample space of 10000, i.e., $P(\text{Prime})=10000$

The probability of guessing reg_prime_number by an attacker is denoted by $P(\text{chosen})$ where
Fig. 4: Comparison of models based on (a) number of communications (b) total transmission energy (c) total transmission time

\[ P(\text{chosen}) = \frac{1}{10000} \]

The reg_sec_key corresponding to reg_pub_key is defined in the range \( 1 \leq \text{reg}_\text{sec}_\text{key} \leq (\text{reg}_\text{prime}\_\text{number} - 2) \)

The probability of choosing the correct reg_sec_key is denoted by \( P(\text{reg}_\text{sec}_\text{key}) \) where \( P(\text{reg}_\text{sec}_\text{key}) = \frac{1}{104,729} \) when \( \text{reg}_\text{prime}_\text{number} = 104,729(10000^{th} \text{ prime number}) \).

The Probability for the attacker for impersonating a vehicle is given by:

\[ P(\text{sybil}) = \frac{1}{10,000} \times \frac{1}{104,729} \]

From the above, we can infer that it is almost impossible for an impersonator to perform Sybil attack in real-time using the proposed model.

6.3 Defense against GPS data forging attack

A malicious vehicle cannot manipulate the location information to a large extent as it should be within the transmission range of the RSU known by the edge server. The edge server will generate the challenge packet based on the velocity, and the location information sent by the vehicles in conflict. The edge will make the challenge packet available for the vehicle based on the information it has provided. Thus, it is obvious that the vehicle sending the bogus information will not be able to respond. Also, in case of a congestion, if it reduces the speed to respond to the challenge packet, the other conflicting vehicle providing a response will be considered authentic, and the edge can deduce that the vehicle reporting congestion purposefully reduced its velocity to respond to the challenge packet. Hence, GPS data forging attack is rendered futile using the proposed model.

7 CONCLUSIONS

In this paper, we proposed a secure cloud-assisted traffic monitoring system for VANETs to provide the current traffic situation. The model is resilient to different types of attacks, and can respond in a situation even when the malicious vehicles at any region of interest outnumber the non-malicious vehicles. The use of a challenge-response strategy helps to prioritize even one vehicle’s report that is in conflict with the majority of vehicles regarding the traffic situation. A number of simulations have been performed to evaluate the model against various performance metrics and test cases, which show its superior performance and resiliency against three other well-known models.

REFERENCES

3D-PRINTING OF A MULTIFUNCTIONAL LITHIUM-ION ELECTRODE WITH FUSED DEPOSITION MODELING

Charng-Yuh Lee, Aditya R. Thakur, Dr. Xiangyang Dong

cl2w2@mst.edu, artt4b@mst.edu, dongxi@mst.edu

ABSTRACT
In this study, we studied the electrical properties of 3-D printed short carbon fiber/polylactic acid (PLA) cathode materials infused by lithium perchlorate. High-performance short carbon fiber/PLA electrode materials were fabricated. Different electrically conductive materials (short carbon fiber, and carbon black) were mixed with PLA, and their electrical conductivities were studied through four-point probe resistivity measurement. To demonstrate the applications of processed PLA materials in lithium-ion batteries, they were further infused by 1M lithium perchlorate solution. A lithium-ion battery structure was fabricated through assembling a coated 12K carbon fiber with the processed PLA cathode materials. It was shown to be capable of powering a white LED. The obtained PLA cathode materials were demonstrated to have potentials to achieve battery design with a complex structure through 3D printing.

1. INTRODUCTION
The desire to reduce waste and produce a complex geometrical structure, additive manufacturing (AM) has been an important topic in manufacturing engineering in the industry [1-3]. A component can be built up layer by layer precisely with wide various materials through AM based on a computerized 3D model. Thus, the needs of using cutting tools and coolants are decreased [1]. In the business environment, an accurate physical model has to be produced as quick as possible. To prototype a model rapidly through AM, fused deposition modeling (FDM) has been invented in 1991 [4].

Thermoplastic materials, such as polylactic acid (PLA) and acrylonitrile butadiene styrene (ABS), are melted and deposited layer-by-layer expeditiously through FDM. However, the lack of the mechanical property of the final products of FDM is caused by the narrow range of available materials and the limit of interfacial adhesion [5-6]. To develop this technique to satisfy the need of rapid prototyping, the methods to improve the mechanical property of the FDM processed products have been investigated [6].

The mechanical property of the matrix can be enhanced dramatically by adding fiber reinforcement. In order to increase the mechanical property, short fiber, such as glass fiber and carbon fiber, has been used as a reinforcing material. The tensile strength of the fiber-reinforced composite was improved with increasing fiber content [8-9]. Through SEM micrographs on the fractured section of the specimens after the tensile test, the interfacial adhesion between carbon fiber and the matrix can be investigated [10-11]. The flexural strength also could be increased by fiber addition[12]. For a short fiber-reinforced composite, the mechanical property was ruled by fiber load-transfer mechanism and the orientation of the short fiber [13-14]. The loading stress can be transferred to the added fiber from the matrix more by increasing the average length of the fiber or adding multiple reinforce fibers to improve the fiber/matrix interfacial mechanism [10, 15-16]. However, the mechanical property improvement is restricted to the limit of short fiber reinforcing mechanism [17-18]. Thus, the long fiber reinforcement is being considered a way to keep improving the mechanical property of the fiber-reinforced composites [19].

The tensile property of long fiber reinforced thermoplastic composites has been investigated, and the results demonstrated that the strength of the composites could be further improved [20]. The fatigue life of long fiber reinforced thermoplastic composites is higher than unreinforced thermoplastic composites because long fiber reinforced thermoplastic composites have higher thermal conductivity than unreinforced composites, which allows the hysteretic heat dissipates to the surroundings faster [21].

It also has been demonstrated that carbon fiber composite has special electronic performances [26-28]. Continuous carbon fiber added composite was found to be an outstanding electromagnetic interference shielding material [26]. Electrical conductivity of carbon fiber composites was raised with high fiber orientation [27]. Carbon fiber based zinc-carbon battery was produced, and the battery performance was not affected during the bending test [28]. Carbon fiber also can be an electrode in lithium-ion battery [29-30]. The lithium-ion storage mechanism was investigated through SEM and XRD [31]. The diffusion coefficient of lithium was predicted, and battery behavior was proved being ruled by lithiation rate [30-31].

FDM was utilized to develop printable lithium-ion batteries. However, the filaments, such as PLA, were not a conductor. To overcome this barrier, electric conductivity was first improved by adding carbon black, graphene, and short carbon fiber. [7, 32-33] The ion conductivity of PLA was increased by adding ethylene carbonate and lithium perchlorate. [34-35]. The PLA-anode and cathode were printed to produce a lithium-ion battery. After 100 cycles testing, 3D printed batteries were stable but the capacity was lower than the theoretical capacities [36].
In our study, carbon fiber/PLA lithium-ion battery was printed through a commercial 3-D printing machine based on FDM. We first improve the conductivity of PLA. The conductivity of the samples was measured through four-point probe measurement. A sandwich structure was designed to produce a carbon fiber/PLA lithium-ion battery. Processed PLA disks were printed through the commercial 3-D printing machine for the sandwich structure as top and bottom layers. The coated carbon fiber was the center layer for the battery. The sandwich structure was produced through compression molding. The performance of the carbon fiber/PLA lithium battery was tested after the infusion process through powering a white LED.

2. EXPERIMENTAL SETUP

2.1. Materials

The materials used in this study were PLA pellets (4340D, Filabot), super P (conductive carbon black, Alfa Aesar), continuous carbon fiber (T800, Toray), lithium iron phosphate (99.99%, Sigma-Aldrich), milled carbon fiber (MF80, Carbeiso), and lithium perchlorate (99.99%, Sigma-Aldrich). Dichloromethane (≧99.8%), Ethylene carbonate (98%) and ethyl methyl carbonate (99%) were purchased from Sigma-Aldrich.

2.2. Preparation of Carbon Fiber Anode

To prepare coated carbon fiber anodes, an electrolytic solution comprising of monoethoxy polyethylene glycol (350) monomethacrylate (SR550) monomer for coating, with 1 M of lithium perchlorate (LiClO₄) as supporting electrolyte for the electro coating process is prepared. This solution is mixed using magnetic stirrer until lithium salt particles are adequately dispersed in the monomer solution. 1:1.5 volume ratio of DMF is then added with continued stirring until the lithium salt particles are completely dissolved. The prepared solution is then agitated to eliminate air-bubbles.

The coating procedure uses this solution along with a customized Teflon fixture inside a high-purity argon purged glovebox. The process uses lithium foil as a reference electrode, aluminum foil as working electrode, a glass fiber mesh separator, conductive, long CFs (Toray 900) as counter electrodes and 2450 EC Keithley potentiostat equipped with three electrode assembly for polarization. A cyclic potential sweep using voltammogram is then applied to record passivation peak voltage which corresponds to the potential at which the conductive substrate (continuous carbon fibers) are coated by grafting polymers from the monomer solution. The grafting peak voltage is then applied for 400 seconds producing desired coating composition and thickness to the carbon fibers.

2.3. Preparation of 3D Printable PLA Electrode

In order to remove moisture, PLA pellets were dried at 85°C at least 4 hours in an oven. Dried PLA pellets were dissolved in dichloromethane (DCM) at 0.1g/mL under stirring until it was completely dissolved. Super P, short carbon fiber, and lithium iron phosphate were added to the solution. The solution was stirred on a magnetic mixer for at least 12 hours. Then, the solution was cast on a FET sheet in a fume hood at room temperature until the remaining solvent was evaporated. The cured composite was pelletized and dried in a drying oven at 85°C at least 4 hours. A single-screw extruder (EX2, Filabot) was used for the extrusion of processed PLA filaments. The extruder was preheated to 180°C at least 15 min. After preheating, the temperature was switched to 160°C for the extrusion. The filaments were stored in a sealed vacuum bag with bagged sodium silicate desiccant before using [7].

2.4. 3D Printing process

The 3-D printer (Snapmaker 3-in-1 3D Printer, snapmaker) was used in printing the deposited layer. The needed 3D models were built through Solidwork 2018, and the G-codes were produced by UltimakerCura 3.6.0. The temperature of the printing head was set at 210°C. Layer thickness (L) was selected as 0.2 or 0.5mm. The bed temperature was 50°C. Printing speed was fixed in 200 mm/min.

2.5. Electrical Conductivity Measurements

A four points collinear probe (S-302 Resistivity Stand, Lucas Signatone Corp.) and Graphical Potentiostat(2450-EC, Tektronix) were used for measuring the conductivity. Processed PLA with different compositions of conductive materials were tested for the electric conductivity at room temperature. Table 1. 40 wt.%SCF, 40 wt.%SCF+CB, 60 wt.%SCF, 60 wt.%SCF+CB and 80 wt.%SCF were tested for the conductivity before extrusion. 12 % CB, 15 % SCF and 20 % SCF were tested after printed. Only 12% CB was further printed as an electrode, and assembled with coated carbon fiber as a battery. Carbon fiber/PLA lithium-ion battery was infused in the solution of 1 M lithium perchlorate (LiClO₄) in 50/50 vol EMC/PC solvent for 24 hours[7].

<table>
<thead>
<tr>
<th>Specimen</th>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 % CB</td>
<td>12 vol% Carbon black + 88 vol% 4043D PLA</td>
</tr>
<tr>
<td>15 % SCF</td>
<td>15 vol% Carbon black + 85 vol% 4043D PLA</td>
</tr>
<tr>
<td>20 % SCF</td>
<td>20 vol% Carbon black + 80 vol% 4043D PLA</td>
</tr>
<tr>
<td>40 wt.%SCF</td>
<td>40 wt.% Short carbon fiber + 60 wt.% 4043D PLA</td>
</tr>
<tr>
<td>40 wt.%SCF+CB</td>
<td>40 wt.% Short carbon fiber + 55 wt.% 4043D PLA + 5wt.% Carbon black</td>
</tr>
<tr>
<td>60 wt.%SCF</td>
<td>60 wt.% Short carbon fiber + 40 wt.% 4043D PLA</td>
</tr>
<tr>
<td>60 wt.%SCF+CB</td>
<td>60 wt.% Short carbon fiber + 35 wt.% 4043D PLA + 5wt.% Carbon black</td>
</tr>
<tr>
<td>80 wt.%SCF</td>
<td>80 wt.% Short carbon fiber + 20 wt.% 4043D PLA</td>
</tr>
<tr>
<td>(93 vol% Short carbon fiber + 7 vol% 4043D PLA)</td>
<td></td>
</tr>
</tbody>
</table>
3. RESULTS AND DISCUSSION

3.1. Conductivity of Processed PLA

The electrical conductivity of three different constituents of printed processed PLA is shown in table 2. The electrical conductivity was measured in parallel to the printing direction to investigate the effects of mixing additional conductive material. The maximum conductivity obtained is $9.65251 \times 10^{-13}\text{S/cm}$ for a 20 vol% short carbon fiber addition. All specimens have bad electrical conductivities which mean volumes of the conductive materials are not sufficient to form a network structure [32]. Also, the nozzle was clogged during printing 20 % SCF because of high volumes of solid fillers (short carbon fiber) [7].

Table. 2 Conductivities of Longitudinal mode of the processed PLA

<table>
<thead>
<tr>
<th>Specimens</th>
<th>Conductivity (S/cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 % CB</td>
<td>$3.17007 \times 10^{-13}$</td>
</tr>
<tr>
<td>15 % SCF</td>
<td>$2.44081 \times 10^{-13}$</td>
</tr>
<tr>
<td>20 % SCF</td>
<td>$9.65251 \times 10^{-13}$</td>
</tr>
</tbody>
</table>

In order to improve the conductivity of PLA, 5 different combinations were investigated. Fig.1. Without carbon black, the electrical conductivity was increased slightly with the addition of short carbon fiber. The electrical conductivity was significantly increased with addition of carbon black. The highest conductivity (31.25 S/cm) was gained for addition of 60 wt.% short carbon fiber and 5 wt.% carbon black. The amount of conductive fillers is able to form a network structure and provides more conductive paths. [32,37]

3.2. Coated Carbon Fiber

Figure. 1 shows the SEM image of the coated carbon fiber. The coated carbon fiber has a coating layer which acts as a solid electrolyte and separator. The rough surface of the coating layer provides more surface area to participate in the reaction in the carbon fiber/PLA lithium-ion battery.

3.3. 3D Printed Carbon Fiber/Lithium-ion Battery

To make a carbon fiber/PLA lithium-ion battery, a sandwich structure was selected. Fig.3. A deposited PLA cathode layer was first printed, and the coated carbon fiber was placed on the top of the layer. Consecutively, the coated carbon fiber was covered by another deposited layer to form a sandwich structural. The sandwich structure was heated to 220°C and compressed to form a battery.

In order to demonstrate the capability of carbon fiber/PLA lithium-ion battery, we produced a sandwich structure battery, 12% CB longitudinal mode battery. Fig.4. The orientation of coated carbon fiber of 12% CB longitudinal mode battery is the same as the printing direction. The battery was infused for 24 hours before the testing. A white LED was powered by 12% CB longitudinal mode battery which demonstrated that carbon fiber/PLA lithium-ion battery has potentials to achieve complex structure design through 3D printing. Fig.5.
4. CONCLUSION
In this work, we reported on carbon fiber/PLA lithium-ion battery which produced by the 3D printing process and compression modeling, and the conductivity of processed PLA with different compositions of conductive fillers. The electrical conductivity was investigated by four-point probe measurement. 12% CB, 15% SCF and 20% SCF did not show the good conductivities because the volume of the conductive materials is not sufficient. To improve the electrical conductivity, carbon black and short carbon fiber were mixed with PLA simultaneously. The highest electrical conductivity obtained was 31.25 S/cm for 60 wt.% SCF+CB. To fabricate a composite lithium-ion battery, an assembled carbon fiber/PLA disk was first infused in the solution of 1 M lithium perchlorate (LiClO₄) in 50/50 vol EMC/PC solvent for 24 hours to increase the ionic conductivity. The obtained battery structure was found to be able to light a white LED.

5. ACKNOWLEDGMENTS
This research was sponsored by the Intelligent Systems Center.

6. REFERENCES
three dimensional continuous fibre reinforcement. Materials & Design, 90, 276-283.


Enhanced Mechanical Properties for 304L Stainless Steel Parts Fabricated by Laser-Foil-Printing Additive Manufacturing

Chia-Hung Hung
Department of Mechanical and Aerospace Engineer, Missouri University of Science and Technology
hungch@mst.edu

Abstract

This study shows the mechanical properties of 304L stainless steel parts fabricated by the laser-foil-printing (LFP) additive manufacturing can be enhanced as compared to parts fabricated by the powder-bed selective laser melting (SLM) technology. The tensile test results indicate that the LFP fabricated parts achieve ~15% and ~10% higher yield strength and ultimate tensile strength, respectively, compared to the SLM fabricated parts. This is mainly because the use of foil feedstock in LFP leads to a higher cooling rate during the solidification of molten metal than the use of powder bed in SLM, due to higher thermal conductivity in foils than powders. By using electron backscattered diffraction it is confirmed that the LFP parts have finer grain structures than the SLM parts. The LFP process also produces metal parts with an average oxygen content about 75% less than those by the SLM process, due to ~10 times of surface area in powders than foils.

1. Introduction

Additive manufacturing (AM) has been widely used in industry, particularly for fabricating complex-shaped three-dimensional (3D) parts that are otherwise too difficult to manufacture by conventional machining processes [1]. AM processes can be generally divided into seven categories according to the ASTM F42 Committee, including vat photopolymerization, material jetting, binder jetting, material extrusion, powder-bed fusion, sheet lamination, and directed energy deposition. Selective laser melting (SLM) is a powder-bed fusion process, while laser-foil-printing (LFP) can be regarded as a sheet lamination process.

SLM is a popular method for producing metallic parts and has commercially available production facilities, while LFP is a recently developed technology that has not been much explored [2]. SLM uses a laser beam to selectively melt metal particles in the powder-bed [3], while LFP employs a laser to weld foils together layer-by-layer. In SLM, each layer of the powder-bed thickness is limited to about 70 µm due to balling behavior and concerns of formation of pores [4]. The powder-bed includes a mixture of particles ranging from a few nanos to tens of microns. Normally the laser being used in SLM is limited to low powers (≤400 W), otherwise powders can be blown away by the recoil pressure for a high power laser. The powder size distribution can have significant effects on powder melting, powder-bed densification, pore formation, and consequently quality of manufactured parts. After laser welding, there is about 40% layer thickness reduction due to powder melting and densification. Hence, for each layer, the maximum thickness after densification is usually less than 50 µm. Due to the porous nature of the powder-bed with gaps between particles, thermal conductivity of the metal powder-bed is significantly lower than that of the bulk material of the same metal [5–7]. The combination of low laser power, small powder-bed thickness, and low thermal conductivity of the powder-bed leads to a slow production rate of the SLM process. Furthermore, it has been shown that balling and spattering behavior can occur due to the porous nature of the powder-bed, which could lead to the formation of pores/voids in the part [4,8,9]. Although the existence of pores in the part could be advantageous in certain biomedical applications, such as prosthetic devices and bone scaffolds [10], pores may behave as stress concentrators that will reduce the part’s strength, causing unexpected or premature failure of the manufactured components [11].

Laser-foil-printing is a laminated object manufacturing process recently developed at the Missouri University of Science and Technology. In this process, a 3D metal part is built layer-by-layer using a dual-laser system to weld a layer of metal foil on the substrate or previous layer and then cutting the cross-sectional contour for each layer. The dual-laser system consists of a continuous-wave laser for welding and a pulsed laser for cutting [12]. To facilitate automation, cutting the foil to the desired shape for each layer may be done in advance by laser or other methods before the foils are welded. The thermal conductivity and the corresponding cooling rate of the foil are high enough to create three-dimensional amorphous structures if desired [13]. In LFP, foil thickness ranges from tens of microns to a few hundred microns. By the use of metal foils, the possible formation of porosity can be significantly reduced as compared to powder-bed processes.

In this study, 304L stainless steel was selected as the material since it is widely used in industry and has high corrosion resistance and high weldability [14]. 304L stainless steel parts were built by a laboratory LFP system and by a commercial SLM system in an inert shielding gas environment. The tensile properties, micro-hardness, oxygen content, porosity, and microstructures of 304L parts fabricated by SLM and LFP were measured and compared. Electron backscattered diffraction was used to identify the phases present, in addition to grain
distribution, for the parts manufactured by each process to observe the differences in the microstructures.

2. Process overview and experimental setup

2.1 Selective laser melting (SLM)

304L SS powders produced by gas atomization (LPW Technology) were used as the feedstock. The particle size distribution was D10 = 19.2 µm, D50 = 27.5 µm, and D90 = 38.3 µm for the cumulative volume. The parts were fabricated on a commercial SLM system (Renishaw AM250) under an argon shielding atmosphere with an oxygen content of < 0.1%. The laser system consists of an infrared pulsed laser, a galvano-mirror scanner, and an F-Theta focal lens. The pulsed laser has a central wavelength of 1070 nm, beam quality M2 of 1.2, maximum pulse repetition rate of 100 kHz, pulse duration of 75 µs, and maximum average power output of 200 W. The laser beam was focused on the surface of the powder bed and the beam diameter was 68.4 µm.

In the processing of each layer, the powders were first spread on the substrate using a recoater blade, as illustrated in Fig. 1(a). The layer thickness (s1) was 50 µm. Then, a pulsed laser was applied to melt powders in selected regions with a striped scanning strategy (Fig. 1(b)). The laser power (P), radiation exposure time, pulse repetition rate, laser scan speed (v), and hatch space (h) were 200 W, 75 µs, 13.3 kHz, 800 mm/s, and 80 µm respectively. The laser energy input (E) was 50 J/mm³ based on Equation 1.

\[ E = \frac{P}{v \cdot h \cdot s_1} \]  

To build a 304L SS part using LFP, five steps are followed for each layer, as illustrated in Fig. 3. First, a layer of metal foil is placed on the substrate or the previously welded layer (see Fig. 3(a)). Next, spot welding is applied on the metal foil using the CW fiber laser (see Fig. 3(b)). The purpose of the spot welding is to fix the foil onto the previous layer to prevent the foil from possible thermal distortion/curving. The third step is pattern welding which uses a raster scan strategy, as shown in Fig. 3(c). The foils are welded under an argon shielding atmosphere with an oxygen content of ~ 1%. The fourth step is to cut the pattern’s contour using the UV pulsed laser (see Fig. 3(d)). Finally, the excess foil is removed (see Fig. 3(e)). Note the foil can be precut into the shape for each layer according to the CAD file of the part and then the foil parts are welded together layer by layer. In this study, for spot welding, the laser power was 400 W, the weld time was 0.5 ms, and the distance between spots was 1 mm. For pattern welding, the laser power was 400 W, the laser scan speed was 200 mm/s, the hatch space was 0.1 mm, and thus the laser energy input was 160 J/mm³ by applying Equation 1. For cutting the pattern’s contour, the laser power was 10 W and the cutting speed was 5 mm/s.

Note that the laser processing parameters used in the aforementioned SLM and LFP processes are different and they do not represent optimal conditions. However, the conditions used for the SLM and LFP processes were investigated to assure that the selected parameters can achieve good parts with porosity less than 1 % [15].
Fig. 3 Schematic illustrations of the five steps in LFP for the processing of each layer: (a) foil feeding; (b) spot welding; (c) pattern welding; (d) contour cutting; (e) excess foil removing.

2.3 Characterization

The parts fabricated by the SLM and LFP machines were cut off from the substrates for analysis. The oxygen content was measured by a carrier gas hot extraction machine (Leco TC500). The microstructure was studied by using an optical microscope (OM, Nikon Epiphot 200), X-ray diffraction (XRD, Philips X’pert MRD), and scanning electron microscope (SEM, Helios Nanolab 600) equipped with electron backscattered diffraction (EBSD). The specimens for OM, XRD, SEM, and EBSD analyses were polished using standard metallographic techniques with a final polishing using 0.04 µm silica. The specimens were then electro-etched using 1.5 volts with 70% nitric acid and 30% deionized water. The porosity was determined by calculating the area of pores on OM images of the cross-section with a total area of 7×5 mm². ImageJ was used to measure the area of pores. The surface area analyzer (NOVAe) was used to measure the surface area of the original powders using the Brunauer-Emmett-Teller theory. The EBSD patterns had the scanning area of 450×450 µm² and pixel resolution of 2 µm. The average grain size was calculated from the EBSD pattern by following the standard ASTM E2627-13.

The mechanical properties of the SLM and LFP parts were measured using tensile testing and micro-indentation. For the tensile test, the tensile strength along the layer building direction (i.e., the vertical direction, indicated by “V”) and the tensile strength along the laser scanning direction (i.e., the horizontal direction, indicated by “H”) were both tested, as shown in Fig. 4(a), since the horizontal and vertical directions are the two directions in which the part usually has either the highest or lowest strength [16,17]. Tensile test specimens were cut from the part using wire electrical discharge machining (EDM). The dimensions of the tensile specimen are shown in Fig. 4(b), with the thickness of 1 mm. The tensile test was conducted on an INSTRON testing machine with a clip-on extensometer at room temperature. The speed of the machine crosshead was maintained at 0.015 mm/mm/min. Seven specimens were tested for each direction, and the mean value with one standard deviation was reported. The fracture surfaces after tensile testing were analyzed using SEM (ASPEX-PICA 1020). The micro-hardness was measured using a Vickers micro-hardness tester (Struers, Duramin 5) with 981.2 mN load and 10 s load duration. The reported micro-hardness value was an average of seven measurements with one standard deviation.

3. Results and discussion

The oxygen content of the powder, the foil, the SLM part and the LFP part was first measured since it could affect the performance of mechanical properties of fabricated part [18]. The oxygen content of the powder is 242 ± 17 ppm. After the SLM processing, the part contains more oxygen content (345 ± 42 ppm). The oxygen content of the foil is 59 ± 1 ppm which is lower than the powder. After the LFP processing, the oxygen content of the part is increased to 90 ± 35 ppm. Even though the LFP part was built in an argon shielding atmosphere with much higher oxygen content than that the SLM part was built in, the LFP part has much lower oxygen content than the SLM part. This is because the surface area per unit volume of the powder (21.9 ± 0.9 µm⁻¹) is much higher (approximately ten times greater) than that of the foil (2.0 ± µm⁻¹), thus the powder has higher tendency to oxidation.

Fig. 5 OM images of the cross-sections: (a) SLM-H; (b) SLM-V; (c) LFP-H; and (d) LFP-V.
Figure 5 shows the typical optical microscope (OM) images of the horizontal and vertical cross-sections of the SLM part and the LFP part. Pores are observed on both the SLM part and the LFP part. The porosity of the SLM part in the horizontal and vertical direction was measured to be 0.04% and 0.1%, respectively. The LFP part has slightly higher porosity (0.5% in the horizontal direction and 0.28% in the vertical direction) than the SLM part. Although in general a high volume fraction of porosity could deteriorate the mechanical properties [19,20], the parts fabricated by both SLM and LFP processes were insignificantly affected due to their porosities less than 1%.

The EBSD patterns indicate that both the SLM and LFP parts consist of austenite phase. The EBSD was also used to measure the average grain size and grain distribution of the LFP and SLM parts in both horizontal and vertical directions, as shown in Fig. 6. The average grain sizes for SLM-H, SLM-V, LFP-H, and LFP-V are 8.9, 12, 7.9, and 9.1 μm, respectively, indicating that the LFP part has smaller grain size than the SLM part. Based on the SEM images and EBSD patterns in Fig. 6, the edges and the center of the laser scan tracks can also be identified and marked as yellow arrows and red dashed lines, respectively, to help understand the grain growth behavior during the melting pool solidification. For instance, by comparing SEM images and EBSD patterns from Fig. 6(e) to 6(h), the grains could be seen to have grown from the boundary of the melt pool to the center of melt pool where grains at the center are finer than grains at boundaries. The relationship between the grain size and the number of grains of EBSD patterns are summarized in Fig. 7. In Fig. 7(a)-(d), the number of grains of the LFP-H whose grain size is smaller than 4 μm is the highest among all. Fig. 7(e) presents the cumulative ratio of area of grain versus grain size in EBSD patterns.

The mechanical properties of the SLM and LFP parts were measured using tensile tests. Figure 8 shows representative tensile stress-strain curves for SLM and LFP parts in both vertical and horizontal directions. The average yield strength (YS) and ultimate tensile strength (UTS) with one standard deviation are given in Table 1. It can be seen that both the SLM part and the LFP part exhibit higher tensile strength along the horizontal direction than the vertical (building) direction, but lower elongation along the horizontal direction than the vertical direction, which is consistent with the results reported in [17]. The LFP part has higher tensile strength than the SLM part in both horizontal and vertical directions. In the horizontal direction, the LFP part is ~15% higher in YS and ~10% higher in UTS than the SLM part. In the vertical direction, the LFP part is ~14% higher in YS and ~11% higher in UTS. The micro-hardness of the LFP part is also higher by ~10% than that of the SLM part, as shown in Table 1, which is consistent with the higher tensile strength of the LFP part. However, the LFP part has lower elongation to failure (~13% less in horizontal direction and ~20% less in vertical direction) compared to the SLM part.

**Fig. 6** Corresponding SEM images and EBSD patterns of the grain distribution for (a, b) SLM-H; (c, d) SLM-V; (e, f) LFP-H; and (g, h) LFP-V. The red dash lines indicate the center of melt pool where has finer grains than grains at the boundaries.
Fig. 7 The number of grains vs. grain size of (a) SLM-H; (b) SLM-V; (c) LFP-H; and (d) LFP-V; (e) the cumulative area ratio vs. grain size in the EBSD patterns.

In tensile testing, the SLM part in the vertical direction exhibits the highest elongation but the lowest strength, while the LFP part in the horizontal direction has the highest strength but the lowest elongation. Thus, the fracture surfaces of the SLM part in the vertical direction and the LFP part in the horizontal direction from the tensile testing were examined and compared using SEM. The results are shown in Fig. 9. All fracture surfaces exhibit ductile fracture. The fracture surface of the SLM part in vertical direction contains dimples of ~25 µm in diameter and dense micro-voids, as shown in Figs. 9(a) and 9(b), respectively, while the LFP-H part shows a void of ~100 µm and sparse micro-voids, as shown in Fig. 9(c) and 9(d), respectively. Micro-voids are indicative of plastic deformation experienced by a ductile material in a typical fracture, and a high density of micro-voids usually indicates good ductility.

Fig. 9 SEM images showing the fracture surfaces of the tensile specimens: (a,b) SLM-V; (c,d) LFP-H.

The higher strength and lower ductility of the LFP part is mainly contributed by the smaller grain size of the LFP part compared to the SLM part. Di Schino [21] reported that when the grains of 304L become finer, the strength and hardness increased and accompanied by a reduction of ductility. This is because smaller grains will increase resistance for the motion of dislocations due to more grain boundaries impeding plastic deformation during the tensile test [22]. In addition, because the average grain size in the vertical direction in both SLM and LFP processes is larger than that in the horizontal direction, the strength in the horizontal direction is higher than that in the vertical direction. Fig. 7 shows that the LFP-H part has the largest amount of fine grains (less than 4 µm), thus the part fabricated by LFP-H has the highest YS and UTS.

4. Conclusions

The mechanical properties of 304L stainless steel parts produced by the LFP process using foils are measured and compared to the parts produced by the SLM process using powders. The comparison shows that the LFP part has higher tensile strength but lower ductility than the SLM part. The higher strength of the LFP part is attributed to the finer grains. The finer grains are due to faster heat dissipation leading to a higher cooling rate in the LFP process. In both LFP and SLM parts, the part strength in the horizontal (perpendicular to build) direction is higher than in vertical (build) direction. The larger surface area in powder vs. foil leads to higher oxygen content in the SLM part vs. the LFP part.

Table 1 Tensile and micro-hardness test results of the SLM and LFP parts.

<table>
<thead>
<tr>
<th>Process</th>
<th>Orientation</th>
<th>YS (MPa)</th>
<th>UTS (MPa)</th>
<th>Elongation (%)</th>
<th>Hardness (HV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLM H</td>
<td>510 ± 6</td>
<td>699 ± 8</td>
<td>72 ± 3</td>
<td>237 ± 10</td>
<td></td>
</tr>
<tr>
<td>SLM V</td>
<td>488 ± 4</td>
<td>645 ± 4</td>
<td>86 ± 3</td>
<td>229 ± 9</td>
<td></td>
</tr>
<tr>
<td>LFP H</td>
<td>585 ± 3</td>
<td>761 ± 5</td>
<td>64 ± 3</td>
<td>260 ± 9</td>
<td></td>
</tr>
<tr>
<td>LFP V</td>
<td>558 ± 13</td>
<td>715 ± 11</td>
<td>72 ± 5</td>
<td>254 ± 7</td>
<td></td>
</tr>
</tbody>
</table>
Acknowledgments

This work was supported by the Department of Energy [grant number DE-FE0012272] as well as by the Intelligent Systems Center and the Keith and Pat Bailey Professorship fund from the Missouri University of Science and Technology.

References

A SURVEY OF AUTOMATED FPGA-BASED ACCELERATOR GENERATION

Deacon Seals, Daniel Tauritz
Department of Computer Science
Missouri University of Science and Technology
Rolla, MO 65409, USA
djsmbc@mst.edu

Minsu Choi
Department of Electrical and Computer Engineering
Missouri University of Science and Technology
Rolla, MO 65409, USA
choim@mst.edu

ABSTRACT
Field programmable gate arrays (FPGAs) offer a potential solution to diminishing sequential computing performance, but often require low-level hardware design experience to utilize. This has resulted in poor adoption of FPGA-based program acceleration techniques by software developers. This paper surveys the state-of-the-art on automated FPGA acceleration research in high-level synthesis, acceleration-focused programming languages, dynamically partially reconfigurable computing, as well as application-specific research in financial modeling and convolutional neural networks.

1. INTRODUCTION
The demise of Moore’s Law means that we can no longer expect to experience significant improvements in general-purpose computer hardware performance as we have in the past [1]. The most common response to this problem has been to move towards parallel computing. By performing multiple computational tasks concurrently, software developers have been able to mitigate the lack of significant improvements to sequential computing performance. Parallel programming has been an established computing paradigm for several decades and the transition to the utilization of multi-core CPUs and vectorization like single instruction, multiple data (SIMD) operations has been gradual and began while Moore’s Law was alive and well [2]. More recent efforts have focused on GPU acceleration, which can utilize large-scale matrix operations to realize improved parallel computing performance compared to what can be achieved on multi-core CPUs alone [3]. However, not all tasks are suited for GPU acceleration, which usually requires the presence of matrix operations to exploit or highly regular sets of operations, and CPU-based parallelization can have significant overhead in fine-grained tasks that require frequent data passing and synchronization. One novel solution to these problems is the use of field programmable gate arrays (FPGAs) for program acceleration. Originally developed for the prototyping of digital hardware designs, FPGAs can be reconfigured to create custom application-specific hardware accelerators that overcome the limitations of other acceleration approaches.

While all acceleration and parallelization approaches come with a learning curve associated with new programming techniques, FPGA acceleration also requires an understanding of computer hardware design principles. This has resulted in a low adoption rate of FPGA acceleration by software developers, who are often not trained in hardware design nor familiar with the domain-specific hardware description languages (HDLs) used to program FPGAs. A substantial amount of research has been devoted to the automation of FPGA accelerator design, but this research has been largely disjointed and spread across multiple fields and domains. The purpose of this survey paper is to provide a unified overview of the research performed to automate FPGA-based accelerator generation and illuminate the state-of-the-art and most promising approaches across all fields.

2. MOTIVATION
Program acceleration offers a way to overcome the diminishing progress of sequential computing performance, but at the cost of increased programming complexity unless it can somehow be automated. Perhaps the most well-established method for overcoming diminishing sequential computing performance is CPU parallelization. Research into the automation of CPU parallelization is relatively mature, and mainstream tools like the GNU C++ compiler and Intel C++ compilers have automatic vectorization and/or automatic multi-threading [4-6], but this is still an active area of research. GPU acceleration is a more recent development approach, but results from manual GPU acceleration have solidified the method as a significant advancement for certain problem domains [7-9]. Automated GPU acceleration has yet to make its way to mainstream developer tools in the same way automated CPU parallelization has, but some automated GPU acceleration research has proven fruitful [10].

Despite the programming complexity of implementation, CPU parallelization and GPU acceleration have been widely accepted by performance-focused software developers. In contrast, FPGA acceleration has seen far less prominent adoption.
despite manual FPGA acceleration producing significant performance improvements [11-13]. Research into automated FPGA accelerator design is active, but this research is spread across multiple application domains and often varies in scope and purpose. The contribution of our research is to survey the state-of-the-art in automated FPGA acceleration as it exists today. This is a topic that has not been surveyed from the perspective of design automation and utilization, and this topic is in urgent need of a crosscutting multi-field survey due to the disjointedness of the existing research publications.

3. Methodology
In order to compare the existing work on automated FPGA acceleration, it is necessary to establish a set of characteristics on which the various works can be compared. For the purpose of this survey we will consider (1) the design space traversal method, (2) if an implementation requires knowledge of program parallelism, (3) if an implementation requires knowledge of pipelining or other hardware design concepts, (4) if an implementation requires a hand-made programmatic representation of the problem being accelerated, (5) if an implementation offers an API, and (6) if an implementation automatically optimizes what functionality is offloaded to the FPGA accelerator.

3.1. Design Space Traversal Method
In this context, a design space refers to the search space of all possible accelerator designs explored by a source. In other words: how does a particular automated FPGA acceleration approach decide the hardware accelerator design? This metric provides useful information to those seeking to perform research in this space as it clearly indicates what approaches have and have not been considered.

3.2. Knowledge of Program Parallelization
If an automated FPGA acceleration implementation requires an understanding of what tasks in a program can be parallelized, then we assert that the implementation in question inherently requires more background knowledge to utilize than an implementation that does not have this requirement. Requiring an understanding of things like data dependencies in a program may be within the understanding of most software developers, but we include this metric in order to consider the acceleration of novice-grade code as well as computer-generated code.

3.3. Knowledge of Hardware Design Concepts
Similar to the last metric, we also consider if an automated FPGA acceleration implementation requires an understanding of pipelining or any other hardware-design concepts as this indicates that a user inherently requires more background knowledge to utilize the implementation. Additionally, hardware design concepts are not normally within the scope of many software development tasks and the requirement of such knowledge is likely to result in poor adoption by software developers.

3.4. Requires Hand-Made Programmatic Representation
There are several automatic FPGA acceleration implementations that utilize novel design space traversal methods in order to design accelerators for a very specific task. These implementations are significant in that they provide a proof-of-concept for certain automatic FPGA acceleration approaches, but the necessity of a hand-made programmatic representation is considered a significant barrier to entry for most developers outside of the specific task(s) addressed in the source. These implementations are significant to the field of automatic FPGA acceleration, but this metric is included to clearly indicate that the implementation in question is not easily applied to arbitrary tasks or existing code.

3.5. Acceleration API
Some automatic FPGA acceleration implementations create a programmer-friendly API with predefined software components for software developers to utilize. This enables the manual acceleration of programs with a particular implementation and may enable the modification of code that has been accelerated. The utilization of an API does require a sufficient understanding of the acceleration mechanisms, but the bar is a lot lower.

3.6. Automated Offloading Optimization
An automated FPGA acceleration implementation that attempts to automate the process of deciding what tasks to offload from the CPU to the FPGA inherently requires less prerequisite knowledge to use than an implementation without this feature. The offloading scheme determined by an implementation may or may not be optimal or of acceptable performance, but the potential for accelerating novice or existing code is valuable regardless. While we cannot make decisive comments about the performance of all offloading schemes generated by all automated FPGA acceleration implementations, some search algorithms have been shown to produce human-competitive results in certain tasks [14].

4. SURVEY OF EXISTING PUBLISHED RESEARCH

4.1. High-Level Synthesis
High-level synthesis (HLS), in broad terms, is the process of generating hardware designs from high-abstraction-level algorithmic descriptions of the desired hardware behavior. While HLS does enable the transformation of high-level source code to hardware designs, it exists primarily as a tool for hardware developers to raise the level of abstraction and simplify development of complex designs. Additionally, HLS source code often relies on specialized libraries to utilize existing components that are often complex and extensively tested. In practice, high-level code used in HLS requires an extensive amount of modifying to achieve the desired hardware performance [15].

Some HLS tools, such as LegUp, aim to accept generic high-level language code [16]. LegUp is not explicitly designed to
accelerate a program that would otherwise run entirely on a CPU. Instead, LegUp runs an entire program on the FPGA through the use of a soft-core MIPS or ARM processor with custom accelerators for functions identified by the user in a separate file [16]. Like many HLS tools, the primary application of LegUp is in the development of embedded systems, but the FPGA designs generated by LegUp can be used as accelerators in CPU-FPGA systems and the exploration of the accelerator design search space is significant to this survey [17]. LegUp takes C source code as input and compiles it with an optimizing LLVM-based compiler to perform optimizations such as dead-code elimination, constant propagation, and loop unrolling before profiling the program entirely on the soft-core processor [16]. This profiling information is then provided to the user, who will then indicate in a separate input file what functions should be accelerated as well as hardware resource limits [17]. LegUp synthesizes hardware by implementing HLS algorithms as optimization passes performed on the LLVM intermediate representation (IR) code. Once the operations in the accelerated function have been optimized, hardware resources are allocated, operations are scheduled, and a finite state machine (FSM) of the accelerated function is generated and bound to the FPGA in a way that makes efficient use of resources like lookup tables (LUTs) and discrete signal processing units (DSPs) [16,17]. The specific details for each of the six comparative characteristics are as follows:

(1) LegUp traverses the design search space through the use of guided optimizations and FSM generation [17]. This approach limits the hardware designs to a set of architectures that may or may not be optimal for the functionality being accelerated. While LegUp does make use of optimizations like bitwidth minimization and loop pipelining that minimize data size and improve the performance of certain loop structures, respectively, the hardware architectures may not be optimal for accelerating certain code structures that could benefit from more unique parallelization approaches like those found in [11].

(2) The LegUp publications don’t mention any automated investigation of program parallelization outside of the parallelization of well-defined operations that are implemented in the FSM [16,17]. This means that the user would have to manually parallelize tasks more complex than the fundamental atomic operations implemented in the FSM.

(3) LegUp does automate the implementation of pipelining and the efficient use of LUTs and DSPs, but does require manual implementation of more course-grain parallelism as previously explained [16,17].

(4) LegUp does allow the user to express their functionality as C code, but does not require a full programmatic representation of the entire scenario in which the accelerator will be applied (e.g., a simulation of the problem the code functionality addresses).

(5) While LegUp does not create their own API to be used in the process of automatic acceleration, they do make use of existing frameworks like OpenMP and Pthreads that enable the user to manually and conveniently declare parallelism in their code. The utilization of an established framework is largely preferable to the creation of a new LegUp-specific API.

(6) LegUp does provide profiling information to inform the user about what functions may be worthwhile to accelerate, but this information does not definitively indicate what function or combination of functions could be offloaded and accelerated for optimal or near-optimal performance [16,17].

4.2. Programming Languages

One proposed solution to the problem of automatic program acceleration is to change the way in which programmatic functionality is expressed and create a domain-specific language. Spatial is a programming language and compiler that allows the user to define their program’s functionality in terms of accelerator functionality, explicit memory locality, and CPU functionality [18]. The specific details for each of the six comparative characteristics are as follows:

(1) Spatial traverses the design search space by applying an active learning-based multi-objective derivative-free optimizer to consider combinations of scheduling options, memory banking options, loop parallelization, and tile size while considering parameters set by the user [18]. A surrogate model to predict performance is “initially built using only a few hundred random design point samples and is iteratively refined in subsequent active learning steps” [18].

(2) Spatial requires the user to define their code using terms like “Foreach” that explicitly indicate to the compiler that the code inside is parallelizable and thus requires the user to understand parallelization concepts [18].

(3) If the user has an understanding of software-level parallelization, Spatial largely automates the process of making low-level hardware-level optimizations and thus does not explicitly require knowledge of hardware design concepts to produce acceptable-quality accelerator designs.

(4) For the purpose of indicating that Spatial is not able to accelerate existing code, as it is itself a self-contained programming language, we consider Spatial to require handmade programmatic representations. However, Spatial can be considered a high-level language and the necessity to hand-code situation simulations is thus not required to utilize Spatial.

(5) Spatial does not implement an API for existing languages since it is itself a self-contained programming language.

(6) While Spatial does enable automated search space traversal, the user is required to indicate what functionality should be implemented in hardware and thus what functionality to offload through the use of the “Accel” operator, which has both blocking and non-blocking versions [18].

4.3. Dynamic Partially Reconfigurable Computing

Dynamic partially reconfigurable computing is an active field of research that aims to utilize the online partially reconfigurable functionality that FPGAs have had for the last decade [19]. Utilizing online accelerator design space traversal can result in accelerators that adapt to changing computational scenarios and may outperform static accelerators as a result.
Publications like [19] have already surveyed this field in detail, but the scope of the research presented often focuses on manual partial reconfiguration (PR) implementations that traverse a limited search space. [19] explains this lack of automation by stating that “many published techniques for overcoming the limitations of vendor tools have slowly become obsolete as a result of the increasing heterogeneity of modern devices and less open access provided by (FPGA) vendors.” As a result, “most systems that use PR at present must be designed at a low level with detailed hardware design expertise required” [19]. There are some PR publications describing approaches that aim to overcome this drawback and are within the scope of this survey. One such approach highlighted by [19] is a high-level design framework called Partial Reconfiguration Amenability Test (PaRAT) [20].

PaRAT aims to automate the process of identifying the mutually exclusive and runtime-modifiable components of a program written in an HLS-specific variant of C [20]. The program is partitioned using the automated analysis of a directed acyclic graph that represents the program’s data and control dependencies [20]. The partitions represent hardware functionality that can be reconfigured at runtime to modify behavior, computation time, and power consumption. PaRAT then uses these partitions to enable the complete design space exploration of static hardware module and reconfigurable hardware module combinations [20]. This partitioning is normally a very complex and manual process that is difficult to achieve with existing HLS tools and the automation of this task enables automated design space traversal through the use of the PaRAT team’s previous research, FoRSE [20,21]. FoRSE is a formulation-level partial reconfiguration design space exploration tool that is capable of pruning the search space by three to four orders of magnitude to produce a pareto-optimal set of configurations that “tradeoff competing designer-designated implementation metrics” [21]. The specific details for each of the six comparative characteristics are as follows:

1. PaRAT itself does not necessarily perform significant design space traversal with regards to identifying an optimal hardware configuration, but it is designed to be compatible with FoRSE which generates a pareto-optimal set of hardware configurations for the user to choose from.

2. PaRAT isn’t strictly designed for program acceleration and instead enables the development of advanced accelerators by experienced users. As such, the user is required to have in-depth understanding of parallelism in order to efficiently integrate an accelerator developed with PaRAT/FoRSE.

3. PaRAT does abstract away the need for explicit PR-specific design concepts, but expects an HLS-focused representation of the behavior desired. As such, the user must have advanced hardware knowledge in order to effectively describe their functionality.

4. PaRAT does not require a handmade representation of the acceleration scenario, but this is largely due to the fact that the scope of PaRAT is hardware design and not explicitly accelerator optimization.

5. PaRAT does not have an API of any kind.

6. PaRAT is designed to generate advanced hardware designs, but does not provide explicit support for accelerator generation and thus does not contain functionality to determine the efficient offloading of tasks to an accelerator.

Another field related to dynamic partially reconfigurable computing is evolvable hardware, which focuses on the use of evolutionary algorithms for hardware design applications. Evolvable hardware applications can vary from a priori hardware design to online hardware configuration [22], but recent publications have the tendency to focus on the latter [23]. This approach does, however, have the potential to generate high-quality hardware accelerators that can take on software functionality and configure themselves [23]. This is an appealing prospect from the perspective of FPGA acceleration, but evolvable hardware research often focuses on a particular application domain such as image processing [23]. The research performed in [23] focuses on the evolution of systolic arrays — an array of reconfigurable processing elements (PEs) in a fixed topology — by an evolutionary algorithm running on a CPU onboard the FPGA. The specific details for each of the six comparative characteristics are as follows:

1. The work in [23] utilizes evolutionary algorithms to traverse the hardware design search space from a low level, but this is an objectively narrow search that only addresses the functionality specifically assigned to the hardware. However, once this behavior is accurately represented, the search space traversal method has the ability to completely automate the hardware design process [14,23].

2. One advancement made in [23] is the partitioning of a large systolic array into smaller arrays in order to enable automated exploration of parallelism during search space traversal. Despite this, the application of the dynamic accelerators generated using approaches like those in [23] would likely still require manual parallelization of the code not being accelerated.

3. Due to the fitness-metric-based search approach utilized by evolutionary algorithms, a knowledge of hardware concepts is not required when utilizing evolvable hardware accelerators like those discussed in [23].

4. While evolutionary algorithms are powerful for design space traversal [14], they require hand-made fitness metrics to assess the quality of a potential solution. This can be a more abstract task than direct hardware development and is presumably not an entirely foreign concept to software developers, but it showcases contrast to acceleration solutions that accept whole programs as inputs for acceleration.

5. An API is not proposed in [23].

6. While the accelerators that can be generated using evolvable hardware are potentially very powerful, the work in [23] does not consider the automatic acceleration of a whole program nor its optimal task offloading.

4.4.Application-Specific Automation Solutions

The issue of automated FPGA accelerator generation for generalized problems has been addressed by some projects like
LegUp and Spatial [16,18], but some significant publications have focused on the automation of accelerator generation for a specific application. We have already discussed the image processing accelerator developed in [23], but we will also discuss work in automated accelerator generation for financial modeling and neural networks.

As established in [11], option pricing is an active field of research where FPGA-based acceleration techniques are being applied. While the research performed in [11] details a particular accelerator design, the work in [24] focuses on automating reconfigurable accelerators for option pricing. This is accomplished by accepting option pricing simulation parameters and hardware constraints as input via an XML file. The specific details for each of the six comparative characteristics are as follows:

1. Design space traversal is accomplished through a basic algorithm that scales the design of the requested option model accelerator to the FPGA-based and user-defined hardware constraints [24].
2. The user is not required to have a knowledge of parallelism as the pre-defined models are automatically parallelized to the limits of the hardware constraints.
3. Hardware knowledge is not strictly required either, but a user can manually restrict the hardware resources used by the accelerator if they wish.
4. The framework outlined in [24] allows for non-programmatic representations of the option pricing parameters and thus does not require hand-made programmatic representations of the option pricing problem.
5. One of the goals and contributions made by [24] is the creation of pre-defined hardware accelerator libraries for option pricing. This is not a software-focused API, but the computation approach required by option pricing research doesn’t seem to require such an API.
6. There doesn’t seem to be a need for the automated offloading of tasks to an accelerator in this domain since the majority of the calculations are performed on the FPGA accelerator itself.

Convolutional neural networks (CNNs) also benefit from FPGA acceleration and this application-domain is an active field of research [25–27]. We will focus on [25] as it is a recent publication that represents the state-of-the-art in this application domain. fpgaConvNet is an automated CNN acceleration framework that aims to support a variety of regular and irregular CNN topologies expressed as a synchronous dataflow framework graph (SDGF) [25]. The specific details for each of the six comparative characteristics are as follows:

1. fpgaConvNet traverses the design search space by applying an analytical performance model that makes estimates of throughput and latency [25]. This performance estimation is then considered against the performance metric of interest to the user and used to guide the exploration of design points.
2. The scope of fpgaConvNet is not necessarily in accelerating an arbitrary task or in integrating a CNN into a separate program. As such, the user is able to generate a high-quality CNN without significant knowledge of parallelization.
3. fpgaConvNet automatically utilizes FPGA resources like LUTs and DSPs while it only requires the user to represent their CNN in terms of an SDFG and thus does not require hardware design knowledge to utilize its functionality [25].
4. The analytical performance model employed by fpgaConvNet precludes the necessity for a hand-crafted representation of the CNN application.
5. fpgaConvNet does not indicate that it has an API for integration of CNN functionality into generic programs.
6. Due to the application-specific nature of fpgaConvNet, automated accelerator offloading optimization is not in the scope of CNN acceleration.

The pertinent characteristics of the publications surveyed are represented in Table 1. This table provides a snapshot of the current approaches utilized in state-of-the-art FPGA acceleration research and is valuable for researchers interested in advancing this field through unexplored methods or those wishing to apply automated FPGA-based acceleration to their work.

<table>
<thead>
<tr>
<th>Research</th>
<th>Search Space Traversal Method</th>
<th>Requires Knowledge of Parallelization</th>
<th>Requires Knowledge of Hardware Design Concepts</th>
<th>Requires Hand-Made Programmatic Representation</th>
<th>Automation API</th>
<th>Automated Offloading Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>LegUp [16,17]</td>
<td>Fixed Optimization Passes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Utilizes existing APIs</td>
<td>Suggests offloading schemes to user</td>
</tr>
<tr>
<td>Spatial [18]</td>
<td>Active Learning</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PaRAT/FoRSE [20,21]</td>
<td>F gestures Pareto Front Generation</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Evolutionary Systolic Arrays [23]</td>
<td>Evolutionary Algorithm</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Option Pricing Accelerator Framework [24]</td>
<td>Basic Resource Constraining</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>fpgaConvNet [25]</td>
<td>Performance Model-Based Traversal</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1: Approach-Characteristics of Existing Automated FPGA Acceleration Research

5. CONCLUSIONS

This paper presents the state-of-the-art in automatic FPGA acceleration. This survey provides a first-of-its-kind comparison across multiple fields and places otherwise loosely related publications in the context of automated FPGA acceleration.

6. ACKNOWLEDGMENTS

This work has been supported by S&T’s Intelligent Systems Center.

7. REFERENCES


APPLICATION OF FORMAL METHODS IN COMPUTER SECURITY TO TELEMEDICINE

Fernaz Jasdanwala* Bruce M McMillin
{faj2k7, fl@mst.edu}
Department of Computer Science
Intelligent Systems Center
Missouri University of Science and Technology, Rolla, MO 65409-0350

ABSTRACT
Telemedicine is a technology-based alternative to traditional health care delivery. However, poor security mechanisms in telemedicine services can severely hamper the quality of care provided. Telemedicine security includes problems such as authorization, authentication, and availability that are common with other infrastructures such as banking. Apart from these there are many new challenges posed by telemedicine as well. Telemedicine requires information security and privacy as well as physical safety. For example, sensors that are located in a patients home or that interface with the patients body to detect safety issues or medical emergencies may inadvertently transmit sensitive information about household activities. Similarly, routine data transmissions from an app or medical device, such as an insulin pump, may be shared with third-party advertisers. Also, if RFID systems are used to monitor patients they may create new security and customer privacy issues. An attacker can hypothetically learn the patients informational preferences without their permission if they carry items with RFID tags, revealing privacy information. Patients should be able to trust the system which can be difficult without adequate security and privacy protections for underlying tele-health data and systems. This paper describes what security problems are reported most commonly by users of telemedicine and what security problems have not been sufficiently addressed in the literature.

1 INTRODUCTION
While designing the security of a cyber-physical systems (CPS) several characteristics of such systems must be taken into account. Some of these characteristics are feedback between the cyber and physical environment, distributed management and control, uncertainty, real-time requirements, and geographic distribution [1]. Applications of CPS include high confidence medical devices and systems, assisted living, traffic control and safety, advanced automotive systems, process control, energy conservation, environmental control, avionics, instrumentation, critical infrastructure control (electric power, water resources, and communications systems for example), distributed robotics (telepresence, telemedicine), defense systems, manufacturing, and smart structures [2].

This paper considers the traditional methods of formal security with respect to a limited model of telemedicine. We consider two types of right leakage: Confidentiality and Integrity in the system and analyze the limited telemedicine scenario using formal methods. Most importantly we try to show whether these rights leakage can be prevented by certain formal methods or how powerful they are in recognizing the rights leakages and then explicitly redesigning the telemedicine framework based on the guidelines of the formal methods to overcome them.

2 APPLICATION OF FORMAL METHODS
We consider eight formal methods to analyze the telemedicine infrastructure in this section. We start each section with a brief overview of what the method encapsulates followed by the analysis of the system with respect to the rights leakage.

2.1 HRU
The Harrison-Ruzzo-Ullman (HRU) model is an access control model which provides an access control matrix (ACM) specifying who is allowed to do what. The capturing of this is difficult for cyber physical systems since interactions between types of access often overlaps. The HRU model also introduces the concept of the safety problem which states that given a system, an initial configuration Q, a right R, we say Q is safe for R if there is no sequence of requests that will write R in a cell of ACM that didn’t already contain it [3].

We consider four subjects: Health care providers HCP(P), HCP(Q) and Patients P, Q. Objects are patients records i.e RecordP and RecordQ. HCP(P) has read, write right over patient P’s record RecordP. For practicality, patients P and Q cannot write to its own healthcare records, but can only read them. New rights to the patient healthcare records can be added by one of the healthcare provider (primary), chosen by the patient with some consideration, i.e the patient has given explicit consent to its primary healthcare provider to add/modify rights over their records. We call them Super(P) and Super(Q) respectively. There is also a user Auditor which gets read-only rights over all the patient records. A user Unauthorized is also considered to
show rights leakage who has no access to the system initially. The HRU model for the above can be given as follows:

One of the rights leakage (Confidentiality) in this model occurs when the primary health care provider shares the records of its patients to others without the consent of the patient. The sharing can be reason of emergency (Patient unable to provide consent), personal gain for the primary HCP or an unauthorized personnel just stealing online records from the primary HCP since they are not being careful while securing patient records under them. For example, if health care provider HCP(P) wants to share patient Ps records to healthcare provider HCP(Q), he can easily do so by principle of attenuation of privileges. Moreover, since the patients records can be shared without the consent of the patients it constitutes a bad confidentiality leakage. An integrity violation of the records can also occur if a corrupt HCP modifies patient records for unnecessary treatments for monetary gains. However, this violation cannot be captured in ACM since the HCP already has write right over its patient’s record and they may chose to maliciously modify it. The HRU commands for all these leakages is as follows:

command share_right(Super(P),Unauthorized, RecordP)
    if read in a[Super(P), RecordP]
        then
            enter read into a[Unauthorized, RecordP];
        if write in a[Super(P), RecordP]
            then
                enter write into a[Unauthorized, RecordP];

command share_right_A(HCP(P),HCP(Q), RecordP)
    if read in a[HCP(P), RecordP]
        then
            enter read into a[HCP(Q), RecordP];
        if write in a[HCP(P), RecordP]
            then
                enter write into a[HCP(Q), RecordP];

command share_right_B(HCP(P),HCP(Q), RecordQ)
    if read in a[HCP(Q), RecordQ]
        then
            enter read into a[HCP(P), RecordQ];
        if write in a[HCP(Q), RecordQ]
            then
                enter write into a[HCP(P), RecordQ];

The modified HRU after the previous commands is given above.

2.2 T-G

Take-Grant model is a formal model for designing and analyzing the security of a system. It is represented as a directed graph whose vertices are the subjects and objects of the system and arcs are the access right set (read, write, take and grant) of subjects to objects [4]. The essence of this model lies in the fact that it helps us to formalize the concept of theft with its unique set of access rights, take and grant. Hence the safety of a system can be analyzed and formalized with the notation of the possibility of theft with the help of T-G protection model [4].

Consider the following Take-Grant protection system. Suppose patient P has two healthcare providers HCP(P) and HCP(P,Q). So, HCP(P) and HCP(P,Q) both have (read, write) rights over RecordP of patient P. This is similar to P consenting to getting healthcare from HCP(P) and HCP(P,Q). Also, patient Q’s healthcare provider is only HCP(P,Q). Hence HCP(P,Q) has (read, write) rights over RecordQ of patient Q. This is again similar to patient Q consenting to getting healthcare from HCP(P,Q). Again, we assume that the patient has already chosen a primary healthcare provider with some consideration who can modify the ACM for the patient record i.e give additional rights or revoke rights from other health care providers. We call them Super(P) and Super(Q) respectively. This is modelled in the diagram below.
However, the use of the Take-Grant protection model to model this kind of a scenario, WILL ALWAYS lead to an unauthorized sharing of rights. The conventional Take-Grant mechanism for exercising control over the movement of rights suffers from an unfortunate limitation: it cannot enforce strictly unidirectional channels for the flow of rights. That is, if rights can be moved directly or indirectly from the provider HCP(P) in the graph to another provider HCP(P,Q), then one cannot prevent rights from flowing in the opposite direction, from HCP(P,Q) to HCP(P).

It can be shown by construction that once a tg-path of arbitrary length has been formed between HCP(P) and HCP(P,Q) (after HCP(P) takes right from HCP(P,Q)), it becomes possible for HCP(P) to take any right that HCP(P,Q) has over RecordQ, without the involvement/authorization of patient Q. The scenario where HCP(P) takes rights from HCP(P,Q) can be anything ranging from HCP(P,Q) being a subsidiary healthcare provider of HCP(P) to HCP(P) taking rights from healthcare provider HCP(P,Q) without the knowledge of HCP(P,Q) (Theft). The scenario of emergencies where the patient is not able to give explicit consent will not constitute theft since the primary health care provider will choose to share the patient’s record with appropriate authorities rather than the overlooking authorities forcefully taking it from the primary health care provider. The commands for this operation can be given as:

\[
\text{HCP(P) takes((read,write) to RecordQ) from HCP(P,Q)}
\]

Another unauthorized leakage of rights in this model occurs when some unauthorized personal steals the patient’s record from the primary health care provider. The stealing of records from the primary HCP can be a result of inefficient security protocols being employed by HCP while securing the patient records under them. The commands for this operation can be given as:

\[
\text{Unauthorized takes((read,write) to RecordP) from Super(P)}
\]

The notion of theft in the T-G model helps us to analyze the confidentiality rights leakage and not the integrity rights leakage. Since unauthorized access can be modelled here there is confidentiality rights leakage in the system which can be identified but not prevented with this model.

### 2.3 BLP

The Bell Lapadula Model was proposed with a theory of multilevel security in mind so as to prevent unwanted information flow. This model primarily analyzes and prevents confidentiality leakages in a system [5]. It was initially intended to provide an access control framework in government and military applications. Each object in the system we are modelling is given a security level of the form (classification level, set of categories), maximum security level and current security level. The current security level can be changed dynamically. This model supports mandatory access control by determining the access rights from the security levels (high, low, etc) associated with subjects and objects and discretionary access control by checking access rights (read, write, append, etc) from an access control matrix [5]. The model then enforces two properties namely: simple security property (no read up) and *-property (no write down). This model does not always support the need to know principle except in rigid military situations [5].

For the telemedicine model, I am very skeptical that the lattice structure of BLP models are meaningful because information between health care providers and patients can flow in both directions. I have seen suggestions that doctors could have two different accounts, one to work on patient records, the other to write prescriptions and that they should log in and out between these two activities. This sounds very unrealistic to be represented in a rigid BLP structure which was designed for the military where there was no dynamic flexibility involved.

However, that being said we can construct a very limited and specific scenario of the healthcare system using the BLP model. For practicality, we assume that the patient cannot write to its own healthcare records but can only read them. Also, when a patient sees the healthcare provider for the first time, a record is opened with health care provider who writes patient’s data to it as when the patient comes in for any sort of treatment. The ACM for this is defined in Table 3.

The ACM states that each health care provider of a patient has a (read,write) right over their patient records. The Auditor
has read rights over all the patients in a particular group. This

group can be analogous to a hospital where all the patients

are treated and Auditor could be the administration of that

hospital. In another scenario, we can regard the group as all

the patients covered by a health insurance company where

the Auditor is the administrator of the insurance company.

Also, we assume that the patient has already chosen a primary

healthcare provider with some consideration who can modify

the ACM for the patient record i.e give additional rights or

revoke rights from other health care providers. We call them

Super(P) and Super(Q) respectively. The lattice for BLP model

can then be given as:

\[
\begin{array}{c}
\text{Auditor (RecordP, RecordQ)} \\
\text{HCP(RecordP, RecordQ)} \\
\text{Patient(P)(RecordP)} \\
\text{Patient(Q)(RecordQ)} \\
\text{Super (RecordP, RecordQ)} \\
\text{Unauthorized}
\end{array}
\]

The lattice essentially tells us that the patient can read only

his records from HCP and that Auditor has to be at a level

higher than HCP (read down, Simple Security Condition). The

HCP (including Super who is also a HCP) can read and

write to patient records whose access it has based on the ACM

(*-Property). Here the Auditors are at a high security level

while the HCPs and patients are at a low security level.

The integrity of data is maintained here as the patient cannot

write to his own record because the ACM does not allow

that. Also, the Auditor can only read patient records which

is analogous to how they actually work where they get full

access to temporary copy of the patients record. However, for

simplicity, I have considered they have only read rights over

all patient records in a practice.

A rights leakage can be given in the form of a corrupt HCP

(including Super who is also a HCP) which passes on the in-

formation to a third party for personal gain without the consent

of the patient (Confidentiality) or manipulate the data in some

way so as to give additional treatments to the patient which are

not necessary but which give a monetary advantage to the HCP

(Integrity). The rights leakage in the form of confidentiality can

be restricted to HCPs in the lattice and ACMs but it cannot be stopped all together. Thus Unauthorized user

cannot access patient records. Even though the Biba model

is an integrity model, it can neither identify nor stop when

a corrupt HCP modifies patient records for monetary/personal

gain.

2.5 Clark Wilson Integrity Model

The Clark-Wilson model like the Biba model is primarily

an integrity model. The model upholds integrity by making

sure that authorized users are not making unauthorized changes

[7]. It defines integrity by a set of constraints on the data

items and then uses Integrity Verification Procedures (IVPs)

to make sure that these constraints are conformed to it. It also

tries to keep the entire system in a valid state by using well-

formed transactions to move system from one valid state to

another [7]. The Clark Wilson Model captures all the possible

transactions in a system, with a focus on the integrity of the

data on which transactions are being performed.

So, for telemedicine the Constrained Data Items (CDIs) are

the medical records of the patients and also the associated

ACM, which specifies who has what kind of rights over a

medical record of a patient. Moreover, a patient has to give

explicit consent and be given a subsequent notification when

there is an update in its ACM.

The Transaction Procedures (TPs) are then obviously the

functions that update the medical records of a patient and
the corresponding ACMs. The Clark-Wilson Model provides ACMs in the form of triples which is a combination of Certification Rule 2, Enforcement Rule 1 and Enforcement Rule 2. The general triples for telemedicine captured so far can be given as follows:

1)(Patient, Read, \{Its own medical record\})
2)(Health Care Provider of the Patient,(Read, Write, Add other health care providers),\{Medical record of the patient\})
3)(Auditors, Read, \{Medical records of an entire practice\})

The first triple is fairly straightforward, saying every patient can read its own medical record. The second triple adds one additional functionality to healthcare providers, authorizing a primary healthcare provider (after being certified by the patient) who can add other healthcare providers for the patient. This can be understood easily if we consider the situation where a patient has to go to a pathologist at a different practice to get some blood work done or the case of referring a patient to a specialist from a general practitioner. The third triple allows only a temporary copy of all the records to be accessed by a auditor/insurer.

The Integrity verification procedures (IVFs) for this scenario can be: A doctor in a healthcare provider database must be genuine (verified with reasonable trust), The patient is duly consented and given subsequent notifications when there is a change in the ACMs.

The Certification Rule 4 (CR4) can be enforced by maintaining a log of all the times a patient record is opened or modified and another log when something is deleted from the patient record. This way, in case of a discrepancy the records can be fixed from the logged data.

One of the rights leakage in this model occurs in case of emergencies. In such a case, the patient cannot give explicit consent to who should access its record and the over looking healthcare provider needs to share the records with the appropriate authorities. This is not such a bad confidentiality rights leakage since as mentioned in triple 2, a health care provider is given the authority to add subjects and objects to the ACM with the consent of the patient, hence he is chosen with some trust in the first place.

The other scenario for rights leakage can be of a corrupt HCP modifying (Integrity) or sharing (Confidentiality) patient’s records. The Clark Wilson model can prevent the accidental modification/deletion of something in the patient record (by CR4) but the intentional and malicious intent of a corrupt HCP modifying patient records to give additional treatments for personal gain cannot be identified or stopped with this model.

2.6 Chinese Wall Model

The Chinese wall model creates a conceptual wall/barrier between two or more groups, usually as a means of restricting the flow of information. The subjects and objects which are in competition with each other are placed in one group known as Conflict of Interest (COI) class [8]. If a subject in a COI class tries to access an object within the same COI class then the access is denied [8].

The central idea in the Chinese wall model is that of conflict of interest, which is most often encountered in stock trading company or investment consultancy. However, in the case of telemedicine, conflict of interest is not a critical problem but patient confidentiality and authentication of medical records is. A health care provider has any patients without any issue of conflicts as well as there are many auditors/insurers who audit/insure more than one medical practice. However, there are certain HIPAA violations that can be addressed using the Chinese wall model. For example, if a patient is seeing a cardiologist and a psychiatrist then the cardiologist should not be able to read the patient’s psychiatric history.

For the telemedicine model, consider the conflict of interest scenario as shown in the figure below. We assume that HCP1(P)=HCP1(Q), HCP2(P)=HCP2(Q) and Super(P)=Super(Q). HCP1 can access object RecordP(A) and RecordQ(A), HCP2 can access object RecordP(B) and RecordQ(B), Super(P) can access object RecordP and RecordQ. Also, RecordP = RecordP(A) U RecordP(B) and RecordQ = RecordQ(A) U RecordQ(B). This is nothing but the CW-Simple Security Condition and CW-∗property.

![Fig. 5. Chinese wall model database](image)

The integrity rights leakage cannot be prevented using this model since a health care provider can always write to records of the patients under them to suit their needs. However, the confidentiality rights leakage can be prevented to a certain extent. This means that HCP2 cannot access RecordP(A) with HCP1 (according to CW-SSC), but this can be circumvented if HCP1 shares patient’s information with HCP2. Similarly, in case of emergencies the Super can grant access to patient records to appropriate authorities without the patient consent if that particular authority is not in the same COI class.

2.7 Information Flow Models (NI, NF, ND)

The concept of non-interference (NI) is to characterize the absence of information flow. It is a fundamental confidentiality concept which states that the system outputs the same trace of events with or without high level inputs and if it does not do so then it is not NI secure [9]. In other words, there is a confidentiality leakage.

Non-interference (NF) on the other hand is weaker counterpart of NI. NF security states that an operation like read or write performed by one subject should not be used to infer unauthorized knowledge by another subject in the domain [9].
In other words, removing a high level event leaves a valid system trace and not the same system trace.

The most important requirement of a secure system is that information should not be obtained by certain individuals on the system. In other words, if a user is not cleared to obtain certain information, they should not be able to deduce it indirectly by observing the system [9]. Thus, ND security states that a low level subject should not be able to deduce the the sequence of high level inputs that led to a visible change in the system which the low level subject can see [9]. More formally, by restricting the trace to only those events that are visible to the low level subject, if another trace can always be found with the same high level input events, the system is ND secure [9]. Since the subjects involved in the confidentiality and integrity rights leakage in this infrastructure are at the same security level in the BLP/Biba Model, the application of Information flow models like NI, NF and ND do not make sense.

2.8 MSDND

Multiple Security Domain Non-Deducibility (MSDND) helps to model systems like Cyber Physical Systems (CPSs) having traditional onion security (domains within domains) [10]. It was developed to address situations where ND applies as well as when ND should apply but there are problems because of the the lack of clear high-low domain separations. MSDND assumes an entity i as any part of the system capable of independent observation or action. The Event System (ES) creates multiple security domains, SDi, as viewed by each entity i in the model. These domains may, or may not, overlap [10].

A system is MSDND secure if there exists some world with a pair of states where one must be true and the other false (exclusive OR), but an entity i has no valuation function for those states. In security domain SDi, i simply cannot know which state is true and which is false [10].

Since the core concept of MSDND is that CPSs can not be accurately modelled by rigid high-low domains as actual CPSs have overlapping domains, one would think it can be used to model the telemedicine infrastructure which cannot be modelled on rigid hierarchical models like BLP/Biba. However, in MSDND, information is always leaked by observing the physical side. But the case with telemedicine and the rights leakage that I have chosen does not have an outsider observing the physical side rather it is the sharing / theft of information among health care providers which cannot happen by just observing the patient / HCP interaction. In fact, in real life when we see a HCP we just pass information to the HCP in a secure consulting room.

Moreover, if the patient records are compromised, there is a very little chance the patient knows about it. Whereas in the case studies(Automated vehicles, power/chemical plants) that I have reviewed so far suggests the controller/driver knows that something is not quite right by the combination of the inputs from the cyber system and physically observing the system.

Hence, implementing MSDND is not feasible for this infrastructure.

3 CONCLUSIONS

The different formal methods of security studied in this paper when applied to different CPSs like telemedicine may not always give an accurate security leak model. Some formal methods maybe more accurately used to capture the access rights, information flow or the complex design domains CPSs have. For example the automated vehicular systems can be modelled with information flow models like NI, ND or NF to determine the security design flaws since these CPSs have an information flow leak due to inherent observability but the same models do not make sense for CPSs like telemedicine.

According to my analysis for the telemedicine infrastructure, the security model that best captures the security policy and the model of the system is the Clark Wilson model. It is important to understand that not all models can be used to model security policy of all CPSs and thus understanding the nuances and the intent based on which all the security models were designed helps us in picking and choosing which model will best capture the security policy needed for a a particular CPS.

REFERENCES

APPLICATION OF SWITCHED ITERATIVE LEARNING CONTROL TO MULTI-AGENT SYSTEM COORDINATION

He Li
Missouri University of Science and Technology
Rolla, MO, USA
Email: hl6yc@mst.edu

Douglas A. Bristow
Missouri University of Science and Technology
Rolla, MO, USA
Email: dbristow@mst.edu

Robert G. Landers
Missouri University of Science and Technology
Rolla, MO, USA
Email: landersr@mst.edu

ABSTRACT
This paper proposes a switching algorithm under the Iterative Learning Control (ILC) framework for multivariable systems whose output channels cannot be accessed simultaneously. The developed scheme is named SILC for short and is applied to a multi-agent system coordination problem. Previous works have shown that zero tracking error cannot be accomplished for such systems with regular ILC, i.e., switching algorithm is not in effective. Zero tracking error, however, is guaranteed if the switching algorithm is in operation and the learning matrix is designed appropriately using a plant-inversion approach. The developed algorithm and the plant-inversion design are applied to a multi-agent coordination problem where each agent is required to track its own global trajectory while maintaining desired relative positions with respect to its neighbors. The global sensor can access only limited number of agents at any time, and thus, the switching algorithm is necessary in order to achieve perfect tracking for all the agents.

1 INTRODUCTION
Iterative Learning Control (ILC) is a feedforward control technique aiming at improving tracking performance of systems that execute the same task multiple times. Applications of ILC are found in industrial robots [1], wafer stages [2], etc. Given an initial input with the associated tracking error, ILC iteratively modifies the control input through an update algorithm to decrease the tracking error. Due to the repeated nature of this process and the learning-from-experience nature of ILC, the convergence of the control input and the tracking error is achieved.

Research on multi-agent system control have been flourishing in academia in the past few decades years, due to its promising applications on autonomous vehicles [3] and quadcopters [4] etc. A popular topic among these research is the study of finite-time perfect tracking through repeated training using ILC. An implicit assumption in fulfilling perfect tracking condition with ILC, i.e. zero tracking error, is that every agent is successfully measured. In a practical MAS, however, it is possible that information of some agents are not accessible while the other agents are being measured. For instance, in calibrating the coordination errors among multiple industrial manipulators with Laser Trackers (LTs) [5], less number of LTs is preferred due to its comparable cost as a robot. If one laser tracker is adopted, then only one robot can be measured at any time. A second example might be found in multi-UAV coordination in an indoor environment, where the global information of the agents are measured by a camera system or a laser system. For similar reason as in the first example, the metrology system such as an LT may not measure all the agents simultaneously. The loss of the measurements of some agents result in the singularity of the effective learning matrix $L$ if the system is controlled with ILC. Consequently, the zero error convergence condition $\rho(I - PL) < 1$ is violated. Motivated by this problem, a switching algorithm is developed in this work with which the agents are sequentially measured and zero tracking error is achieved.

The rest of the paper is organized as follows. Section 2 reviews the SILC algorithm developed in our previous works. In particular, matrix representation of multivariable ILC systems is introduced, which is widely adopted for time-domain convergence analysis; Non-zero error convergence problem for multivariable ILC systems with limited measurement access is discussed, and SILC algorithm is presented. In section 3, the SILC algorithm is applied to multi-agent system coordination. Relative trajectory tracking using relative-information based feedback controller and global trajectory tracking using SILC algorithm are discussed.

2 SILC OVERVIEW
2.1 Matrix representation of multivariable ILC systems
Consider a Linear Time-Invariant (LTI), Multi-Input Multi-Output (MIMO), discrete-time system

$$y[k] = P(z)u[k] + d[k], \quad (1)$$

where $k \in \mathbb{Z}$ is the time instant, $z$ is the forward time-shift operator such that $zz[k] = z[k + 1]$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$, respectively, refer to the control input, system output and exogenous disturbance signals where $m$ and $n$ denote...
the input and output dimensions, respectively. This work studies square multivariable systems, i.e., \( m = n \). The notation \( \mathbf{P}(z) \) represents the transfer function matrix which is a linear operator mapping the control inputs to the system outputs. The system \( \mathbf{P}(z) \) is assumed to be stable and of relative degree one throughout the paper, since it can be stabilized with a feedback controller if it is not stable, in which case \( \mathbf{P}(z) \) represents the closed-loop dynamics.

In ILC, the process described by (1) is repeated, and, thus, has finite time duration the length of which is denoted by \( N \). For the input signal \( u[k] \), \( 0 \leq k \leq N - 1 \), and for the output signal \( y[k] \), \( 1 \leq k \leq N \). Such a finite-dimensional system is typically written in a matrix form, known as lifted system, by stacking signals over the length \( N \), i.e.,

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_m
\end{bmatrix} = \begin{bmatrix}
P_{11} & P_{12} & \cdots & P_{1m} \\
P_{21} & P_{22} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
P_{m1} & \cdots & \cdots & P_{mm}
\end{bmatrix} \begin{bmatrix}
U_1 \\
U_2 \\
\vdots \\
U_m
\end{bmatrix} + \begin{bmatrix}
D_1 \\
D_2 \\
\vdots \\
D_m
\end{bmatrix}, \tag{2}
\]

where \( \mathbf{U} \in \mathbb{R}^{mN} \) and its element \( U_i \in \mathbb{R}^N \) represents the control signal in the \( i \)-th input channel stacked over \( [0, N - 1] \), i.e.,

\[
U_i = [u_i[0], \ldots, u_i[N - 1]]^T. \tag{3}
\]

The vectors \( \mathbf{Y} \) and \( \mathbf{D} \) are defined accordingly. The matrix \( \mathbf{P} \in \mathbb{R}^{mN \times mN} \) is interpreted as a \( m \times m \) partitioned block matrix where each block \( \mathbf{P}_{ni} \in \mathbb{R}^{N \times N} \) is a matrix representation of the transfer function \( P_{ni}(z) \) in \( \mathbf{P}(z) \) and is lower triangular Toeplitz, i.e.,

\[
\mathbf{P}_{ni} = \begin{bmatrix}
p_{ni}[1] & 0 & \cdots & 0 \\
p_{ni}[2] & p_{ni}[1] & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
p_{ni}[N] & \cdots & p_{ni}[N - 1] & p_{ni}[1]\end{bmatrix}, \tag{4}
\]

where \( p_{ni}[k], 1 \leq k \leq N, \) is the weighting sequence [6] of \( P_{ni}(z) \). For detailed explanation and construction of the matrix representation (4), the readers are referred to [7].

In ILC, the process (2) is executed repeatedly. Thus, the equation (2) is extended with another dimension which is indexed by iteration, i.e.,

\[
\mathbf{Y}(j) = \mathbf{P U}(j) + \mathbf{D}, \tag{5}
\]

where \( j \) denotes the iteration index. Note that \( \mathbf{D} \) is not indexed with \( j \) since the disturbance signal is assumed invariant in every iteration. Accordingly, the tracking error \( \mathbf{E} \) in the \( j \)-th iteration is

\[
\mathbf{E}(j) = -\mathbf{P U}(j) + \mathbf{\hat{E}}, \tag{6}
\]

where \( \mathbf{\hat{E}} = \mathbf{Y}_d - \mathbf{D} \), and \( \mathbf{Y}_d \) denotes the desired trajectory \( y_d[k] \) in lifted system over length \( [1, N] \). Similar as in (2), the error \( \mathbf{E} \) is expressed as

\[
\mathbf{E} = \left[ \mathbf{E}_1^T, \mathbf{E}_2^T, \ldots, \mathbf{E}_m^T \right]^T, \tag{7}
\]

where \( \mathbf{E}_i^T \) denotes the tracking error in the \( i \)-th output channel in lifted system. The purpose of ILC is to iteratively update the control input such that the output is as close as possible to the desired trajectory. A widely used ILC update equation [8], [9] is

\[
\mathbf{U}(j + 1) = \mathbf{U}(j) + \mathbf{LE}(j), \tag{8}
\]

where \( \mathbf{L} \in \mathbb{R}^{mN \times mN} \) is the learning matrix for the entire multivariable system and is also interpreted as a \( m \times m \) partitioned block matrix

\[
\mathbf{L} = \begin{bmatrix}
L_{11} & L_{12} & \cdots & L_{1m} \\
L_{21} & L_{22} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
L_{m1} & \cdots & \cdots & L_{mm}
\end{bmatrix}, \tag{9}
\]

in which each submatrix \( L_{ni} \) is responsible for learning the \( n \)-th output channel and updating the control in the \( i \)-th input channel. The resulting closed-loop iteration-domain error dynamics is obtained combining the equations (6) and (8),

\[
\mathbf{E}(j + 1) = (I - \mathbf{PL})\mathbf{E}(j). \tag{10}
\]

As an iterative algorithm, convergence of the ILC update equation (8) is important. Two convergence definitions are discussed in the literature, i.e., convergence and zero-error convergence [10].

**Definition 1.** (Convergence) A multivariable system (10) controlled with ILC is said to be convergent if there exists a constant vector \( \mathbf{U}_\infty \) such that

\[
\lim_{j \to \infty} ||\mathbf{U}(j) - \mathbf{U}_\infty|| = 0. \tag{11}
\]

**Definition 2.** (Zero-error Convergence) A multivariable system (10) controlled with ILC (8) is said to be zero-error convergent if it is convergent, and further,

\[
\lim_{j \to \infty} ||\mathbf{E}(j)|| = 0. \tag{12}
\]

It is straightforward to observe that convergence is less strict than zero-error convergence. Satisfaction of the latter one implies that of the former one, but not otherwise. Zero-error convergence is widely used in the literature. The necessary and sufficient condition for zero-error convergence is stated in Theorems 1. The necessary and sufficient condition for convergence and its proof will be published somewhere else.

**Theorem 1.** (Zero-error Convergence) A multivariable system (10) controlled with ILC (8) is zero-error convergent if and only if [10]

\[
\rho(I - \mathbf{PL}) < 1, \tag{13}
\]

where for an arbitrary square matrix \( \mathbf{A} \), \( \rho(\mathbf{A}) \) is the spectral radius of \( \mathbf{A} \).

**Proof.** Proof of this theorem directly follows asymptotic stability analysis for LTI systems and can be found in [10].

©He Li 2019
2.2 SILC algorithm for multivariable systems with limited measurement information

Multivariable systems may have limited access to its output channels, for instance, a multi-agent system which can acquire the global information of limited number of agents at any time. Let $S$ be a selection matrix and $S \in \mathbb{S}$ and $\mathbb{S} = \{S_1, S_2, \cdots, S_m\}$. Each $S_i \in \mathbb{R}^{mN \times mN}$ is interpreted as a block diagonal matrix which has $m$ block diagonal entries and each entry has dimension $N \times N$. The $i$th diagonal entry is an identity matrix whereas the other diagonal entries are zero, i.e.,

$$S_i = \text{blkdiag}(0, \cdots, 0, I, 0, \cdots, 0).$$

The selection matrix $S$ determines which output channel is currently available and thus used for learning by pre-multiplying the error vector by $S$, i.e., SE. Consequently, the ILC update equation for such systems and the resulting closed-loop error dynamics are

$$U(j + 1) = U(j) + \text{LSE}(j),$$

$$E(j + 1) = (I - \text{PLS})E(j).$$

Note that although the learning matrix $L$ is not changed, the effective learning matrix is essentially LS due to the incomplete measurements. It is straightforward to conclude from the equation (16) that such a system always violates the zero-error convergence condition (13) since the error dynamics matrix $I - \text{PLS}$ always has eigenvalues at one due to the singularity of the matrix $S$, which implies that

$$\rho(I - \text{PLS}) \geq 1.$$  

The inequality (17) indicates that zero tracking error is not possible for the system (15) and (16). Zero tracking error, however, is always desired in tracking problems. A switching algorithm is, therefore, proposed in [7], [11] to achieve this goal, with which the selection matrix $S$ is swept through its domain and at each $S$ the regular ILC is executed. The switching algorithm is shown in Algorithm 1. Theorem 2 is proposed in [7] which guarantees the tracking error converges to zero with the Algorithm 1.

**Theorem 2.** Let the learning matrix $L$ in (15) and (16) be

$$L = P^{-1} \Lambda,$$

where $\Lambda = \text{diag}(\Lambda_1, \cdots, \Lambda_m)$ is a block diagonal matrix with $\Lambda_i \in \mathbb{R}^{N \times N}$ for $\forall 1 \leq i \leq m$. Then, with the SILC algorithm in Algorithm 1, the tracking error in every channel of the multivariable system (2) converges to zero if $\rho(I - \Lambda) < 1$.

### Algorithm 1: SILC Algorithm

1. Initialization: $j = 0, U(j) = 0$
2. $Y(j) = PU(j) + D$
3. $E(j) = R - Y(j)$
4. for $i \leftarrow 1$ to $N_{sw}$ do
5. \hspace{1cm} $i = f(l)$ where $f(l) = \text{mod}(l, m)$
6. \hspace{1cm} $S = S_i$
7. \hspace{1cm} $j = 0$
8. \hspace{1cm} for $j \leftarrow 0$ to $N_{iter} - 1$ do
9. \hspace{2cm} $U(j + 1) = U(j) + \text{LSE}(j)$
10. \hspace{2cm} $E(j + 1) = (I - \text{PLS})E(j)$
11. \hspace{1cm} end
12. \hspace{1cm} $U(0) = U(N_{iter}), E(0) = E(N_{iter})$
13. end

denotes the set of directed edges, and $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{N_a \times N_a}$ is the weighting adjacency matrix whose diagonal elements are zero and off-diagonal elements are non-negative, i.e., $a_{ij} = 0$ and $a_{ij} \geq 0$ ($i \neq j$). A directed edge from the $i$th vertex to the $j$th vertex is denoted by an ordered pair $(i, j)$, representing that the $j$th agent could receive information from the $i$th agent.

The set $\mathcal{E}$ incorporates all effective edges. If a directed edge $(i, j)$ exists, then $a_{ij} = 1$, otherwise, $a_{ij} = 0$. For instance, the topology relationship of a three-agent system shown in Figure 1 is described by $G_1 = \{V_1, \mathcal{E}_1, A_1\}$, where $V_1 = \{1, 2, 3\}$, $\mathcal{E}_1 = \{(1, 2), (2, 3), (3, 1)\}$, and

$$A_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$  

![Fig. 1: Graphical representation of topology relationship of a three-agent system](image)

**3.2 Linear multi-agent systems and consensus problem**

For a multi-agent system whose topology is described by $\mathcal{G}$, suppose that all agents have the identical linear dynamics, i.e.,

$$x_i[k + 1] = Fx_i[k] + Gu_i[k],$$

$$y_i[k] = Hx_i[k],$$

where $x_i$, $u_i$, and $y_i$ respectively, represent the state, input, and output of the $i$th agent. The multi-agent system (20) is said to achieve consensus if the states of all agents asymptotically reach an agreement, i.e.,

$$\lim_{k \to \infty} \|x_j[k] - x_i[k]\| = 0, \forall i \neq j.$$  

\(\copyright\)He Li 2019
A widely used consensus protocol in the literature is $\mathbf{u}_i[k] = \mathbf{u}_{fbi}[k]$ where

$$\mathbf{u}_{fbi}[k] = K \sum_{j=1}^{N_a} a_{ij} w_{ij}(\mathbf{x}_j[k] - \mathbf{x}_i[k]),$$

(22)

where $K$ is the feedback gain and $W = \{w_{ij}\}$ is a weighting matrix. A necessary and sufficient condition [12] to achieve consensus (21) with the protocol (22) is that the digraph $\mathcal{G}$ incorporates a spanning tree.

3.3 Relative trajectory tracking using relative-information feedback control

While the consensus protocol (22) aims at achieving the agreement of the states among all agents, it is extended to a generalized protocol with which every two agents, provided that an edge exists, maintain a desired dynamic relative trajectory. For the agent $i$ and the agent $j$, suppose the edge $(i, j) \in \mathcal{E}$ and a relative trajectory $d_{ij}$ is desired between them. Then the extended control protocol is

$$\mathbf{u}_{fbi}[k] = K \sum_{j=1}^{N_a} a_{ij} w_{ij}(\mathbf{x}_j[k] - \mathbf{x}_i[k] - \mathbf{d}_{ij}[k]).$$

(23)

In the case $a_{ji} = 1$, then

$$\mathbf{d}_{ij}[k] = -\mathbf{d}_{ij}[k].$$

(24)

Note that if $\mathbf{d}_{ij}$ is a constant, then the relative state between the agent $i$ and $j$ asymptotically arrives at a constant with the control protocol (23). A special case is $\mathbf{d}_{ij} = 0 \ \forall (i, j) \in \mathcal{E}$, then the control protocol (23) is the same as the consensus protocol (22).

A unique feature of the control protocol (22) is that it only requires the relative information to be available, e.g. relative position and relative velocity etc, which facilitates the use of relative sensors where absolute sensor is not economic or has limited access. On the other hand, since relative information is used alone, it only guarantees the asymptotic relative trajectory tracking among agents, whereas there are infinite number of possible global trajectories for each agent which depend on their initial conditions.

3.4 Global trajectory tracking using SILC

Global trajectory tracking requires the global information of at least one agent to be available. The measurement system possibly can access the global coordinate of only one of the agents at any time. The control protocol (23) used in (20), however, has no control on the global trajectories. In this section, we separate the control signal $\mathbf{u}_i$ in (20) into two parts, a feedback control signal $\mathbf{u}_{fbi}$ and a feedforward control signal $\mathbf{u}_{ff}$, i.e.,

$$\mathbf{u}_i[k] = \mathbf{u}_{fbi}[k] + \mathbf{u}_{ff}[k].$$

(25)

The feedback control signal $\mathbf{u}_{fbi}$ is generated by the control protocol (23) and the feedforward control signal is generated the switching algorithm as in Algorithm 1. Substitute the equation (25) in to the equation (20) and concatenate the dynamics of all agents in a unified structure, we get

$$\ddot{\mathbf{x}}[k + 1] = \dot{\mathbf{F}}\ddot{\mathbf{x}}[k] + \dot{\mathbf{d}}_{rel}[k] + \mathbf{G}\ddot{\mathbf{u}}_{ff}[k],$$

(26)

where

$$\ddot{\mathbf{x}} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_{N_a}^T]^T, \quad \dot{\mathbf{d}}_{rel} = \sum_{j=1}^{N_a} d_{ij}, \quad \mathbf{G} = \text{blkdiag}(\mathbf{G}, \cdots, \mathbf{G}),$$

and

$$\ddot{\mathbf{F}} = \begin{bmatrix}
-K_{\sum_{j=1}^{N_a} a_{ij}} d_{ij} & 0 & \cdots & 0 \\
0 & -K_{\sum_{j=1}^{N_a} a_{ij}} d_{ij} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & -K_{\sum_{j=1}^{N_a} a_{ij}} d_{ij}
\end{bmatrix},$$

(28)

in which $deg_{in}(i) = N_a - \sum_{j=1}^{N_a} a_{ij}$, and $\mathbf{I}$ and $\mathbf{F}$ are identity matrices of the same dimension as $\mathbf{F}$. The function $\text{blkdiag}()$ constructs block diagonal matrices for $\mathbf{G}$ and $\mathbf{F}$, both of which have $N_a$ block diagonal entries.

In the concatenated structure (26), the first two items on the right side of the first equation is the closed-loop dynamics resulting from the feedback control (23), which is responsible for relative trajectory tracking. In the case that $\dot{\mathbf{d}}_{rel}[k] = 0$, the consensus problem is resolved. The third item is in charge of global trajectory tracking of a particular agent. The introduction of the switching algorithm to designing this item enables global trajectory tracking of all agents. To see how this works, the state space equation (25) is transformed into the transfer function representation, i.e.,

$$\ddot{\mathbf{y}}[k] = \mathbf{P}(z)\ddot{\mathbf{u}}_{ff}[k] + \mathbf{d}[k],$$

(29)

where

$$\mathbf{P}(z) = \mathbf{H}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G},$$

$$\mathbf{d}[k] = \mathbf{H}(z\mathbf{I} - \mathbf{F})^{-1}\dot{\mathbf{d}}_{rel}[k] + \mathbf{H}(z\mathbf{I} - \mathbf{F})^{-1}\ddot{\mathbf{x}}[0].$$

Since the equation (29) is in the same structure as the equation (1), the SILC algorithm and the associated theories and design methods apply.

Remark 1. The conversion from the state space representation (26) to the transfer function (29) assumes zero initial condition. The effect of nonzero initial condition on system response is considered as a disturbance and represented by $\mathbf{H}(z\mathbf{I} - \mathbf{F})^{-1}\ddot{\mathbf{x}}[0]$. In this work, it is assumed that the initial condition is invariant every time the system is reset, and, thus, resulting in repeated disturbance when the system is executed repeatedly.
3.5 Simulation study

Consider now an MAS system consisting of three agents which move in the horizontal plane. The dynamics of the \( i \)th agent is described by

\[
x_i[k+1] = x_i[k] + u_i[k], \quad i = \{1, 2, 3\},
\]

(30)

where the state vector \( x_i = [x_i, y_i]^T \) denotes the position of the agent in Cartesian coordinates. The topology relationship of this system is described by \( G_i \) as defined previously. The initial conditions of each agent respectively, are

\[
x_1[0] = [3.3, 3.4]^T, \quad x_2[0] = [5.5, 0.5]^T, \quad x_3[0] = [2.7, 0.7]^T.
\]

(31)

Each agent is expected to follow its previous agent and move in a loop along the triangle \( \text{ABC} \) as shown in Figure 2. The triangle is described by the vectors \( \text{AB} = [2, -2], \text{BC} = [-2, 0], \text{and CA} = [0, 2] \). The initial effective desired positions are described by \( \text{d}_{12}[1] = [-2, -2]^T, \text{d}_{23}[1] = [-2, 0]^T, \) and \( \text{d}_{31}[1] = [0, 2]^T \). The desired position starts at \( k = 1 \) due to the one-step delay in system dynamics. Each agent is expected to spend equal amount of time on all edges of the triangle, denoted by \( T_p \), and the overall time spent in a loop is \( T_f = 3T_p \). The complete desired relative trajectories are then automatically determined. Note that wherever the triangle is placed, the desired relative trajectories are not changed, and, thus, relative trajectory tracking applies. If the triangle is restricted globally, then the resulting desired trajectories are computed for each agent, and thus, global trajectory tracking applies. In the simulation below for global trajectory tracking, the triangle is placed at \( \text{A} = [1, 3], \text{B} = [3, 1], \) and \( \text{C} = [1, 1] \).

3.5.1 relative tracking using feedback control

Consider first the case where the MAS system (30) is controlled only by the relative-information based feedback controller (23). The feedback gain \( K = 0.08 \) is selected, and the weighting matrix \( W \) adopted is

\[
W = \begin{bmatrix}
1 & 0.4 & 0.6 \\
0.5 & 1 & 0.5 \\
0.8 & 0.2 & 1
\end{bmatrix}.
\]

(32)

The simulation results are shown in Figure 3, where the desired relative trajectories and the actual trajectories of each agent are depicted. Two observations are made from Figure 3. First, although large tracking errors are observed, each agent roughly tracks its preceding agent along the triangle. Second, the actual trajectories are far from the reference global trajectories. These observations demonstrate that the feedback controller (23) achieves relative-trajectory tracking, but fail to achieve global-trajectory tracking since no global measurements are used in control.

3.5.2 global tracking using SILC

The MAS system (30) is now required to achieve global trajectory tracking as shown in Figure 2. In order to accomplish this goal, the control \( u_i \) is modified to incorporate a feedforward control signal, i.e., (25) is used. A transfer function matrix \( P(z) \) is obtained by following the process in the section 3.4, and the resulting system has six inputs and six outputs. Note that each agent has two outputs, both of which are measured whenever an agent is accessed. Therefore, the resulting six-input six-output system is regarded as a virtual three-input three-output system by grouping together the two output channels of each agent. The plant-inversion design method with \( \Lambda = I \) is then adopted to obtain the learning matrix \( L \) for this three-input three-output system.

The trajectories of each agent, after 1, 2, 3, 13, 19 and 30 switches of learning respectively, are depicted in Figure 4. As shown in Figure 4, upon the completion of the first switch, where the global information of the agent 1 is acquired, the agent 1 succeeds in tracking its global reference while agent 2 and agent 3 are trying to follow agent 1 along the triangle. Although large tracking errors are observed, similar observations are obtained for the other switches. Further, Figure 4 also demonstrates that as the number of switches increases, for instance, upon the completion of the 30th switch, all agents tend to perfectly track their global reference trajectories. It is here emphasized that in realizing the global tracking in this simulation, the Algorithm 1 is not directly applied. Instead, the feedforward control is updated only for the agent whose global information is currently measured. The reason why this modification is adopted will be publish somewhere else.
Fig. 4: Actual trajectories of each agent for the 1st, 2nd, 3rd, 13th, 19th and the 30th switch when the switching algorithm is in operation for global tracking.

4 SUMMARY AND CONCLUSIONS
A switching algorithm is introduced in this paper for multivariable Iterative Learning Control (ILC) systems which can measure a limited number of output channels at any time. This algorithm is proposed in our previous works in which the convergence conditions are proved. This paper focuses on applying the switching algorithm to a multi-agent coordination problem, where three agents are required to track their own global trajectories while maintaining desired relative positions at the same time. A low-level feedback controller is adopted for relative-trajectory tracking based on the relative information, whereas the high-level switching algorithm is used to generate a feedforward control signal utilizing the global information. The simulation results demonstrate that the tracking errors of every agent are significantly reduced after 19 switches. The current work focuses on the convergence of the error. Future work would study the monotonic convergence conditions and the robustness of the designing methods.

5 ACKNOWLEDGMENTS
The authors gratefully acknowledge the financial support for this research provided by the Intelligent Systems Center at Missouri University of Science and Technology.

REFERENCES

©He Li 2019
MICROWAVE CURING FOR FIBER REINFORCED POLYMER COMPOSITES

Siva Sai Krishna Dasari  
Department of Mechanical and Aerospace Engineering  
Advisors: Dr. K. Chandrashekhara and Dr. Thomas Schuman

ABSTRACT
Microwave curing technologies has many advantages over the traditional thermal curing methods for the manufacturing of fiber reinforced polymers, especially the processing speed and energy efficiency. Energy can be instantaneously transferred through applied electromagnetic fields and heat is generated based on dipolar rotational interactions. Microwave curing processes have been used for glass fiber composites but there are significant challenges associated with microwave curing of carbon fiber composites. Efficient heating may be difficult due to high dielectric loss associated with carbon fibers. Laminate quality will be highly dependent on the uniformity of the electromagnetic field in the material. In this work, a multiphysics three-dimensional model was developed to study the composite curing behavior and temperature distribution of a carbon/epoxy prepreg system (IM7/Cycom 5320-1) in the presence of microwave radiation. Microwave heating depends on the thermal conductivity, convective heat transfer, surrounding temperature, intensity of the electromagnetic field and the geometry of the sample. The anisotropic dielectric properties of a composite are incorporated into the simulation model. This model can be used to optimize process parameters to cure thick and complex shaped composite parts. A cure cycle optimized to microwave energy is developed.

1. INTRODUCTION
Advance manufacturing techniques provide gateway to tailor the material properties in order to achieve high performance, environment friendly engineered structures. Advance manufacturing techniques reduces the fuel consumption and emissions and helping combat climate change [1]. Fiber reinforced polymers (FRP) are light-weight materials. With high strength-to-weight ratio, FRP can be tailored for modern high performance structural applications allowing for efficient engineering solutions to severe and varying operating conditions e.g. dynamic and impact events on composite aircrafts [2-3].

Autoclave is the dominant process used for manufacturing high performance composites, especially for the aerospace industry. However, autoclave process is time consuming, expensive and exhaustive. The temperature and pressure requirements add demand on tooling materials. Moreover, growing demand for polymer composites in critical structures gave rise to an urgent need for a reliable and rapid composite repair and fastener-less joining, e.g. adhesive bonding especially in the maintenance, repair and overhaul (MRO) sector in aerospace. In adhesively bonded structures, the quality of the bonded composites has a strong dependency with the variabilities caused by process parameters such as temperature, curing duration and rate [4].

The term curing in thermosetting polymers refers to the transition of liquid resin and hardener components to a solid material. Curing is initiated when the components are stoichiometric and physically mixed together. Polyester, phenolic and epoxy resins are among the mostly used thermoset polymers with epoxy popularly used in high performance composite structures.

The curing of thermosetting polymer systems can be expressed mathematically through cure kinetic equations. Several researchers have worked on the mathematical modeling and simulation of composite curing kinetics [6-8]. The cure kinetics equations can be used to evaluate temperature distribution during cure of thermoset resin systems. Initial research on cure simulation of composites used one or two-dimensional finite difference analysis and was applicable for simple geometries.

Microwave curing is an alternative advanced curing method which can result in reduced cure times and reduced energy consumptions. Microwave curing processes have been used for glass fiber composites but there are significant challenges associated with microwave curing of carbon fiber composites. Efficient heating may be difficult due to high dielectric loss associated with carbon fibers. Laminate quality is highly dependent on the uniformity of the electromagnetic field in the material. Another major challenge is related to arcing of carbon fiber bundles, which can result in very high localized temperatures and damage to surrounding materials [5]. Finite element model was used to study the microwave curing process in a cost-effective manner. The model can be used to optimize process parameters to cure thick and complex shaped composite parts.

2. MATERIALS
In this study Cycom 5320-1 unidirectional prepreg system (Cytec Industries) was used. Cycom 5320-1 with a fiber areal weight of 145 g/m2 is a toughened epoxy resin system. The unidirectional prepreg contains 33% resin by weight which is highly suitable for Out-of-Autoclave curing. The manufacturer recommended cure cycle for Cycom 5320-1 is shown in Figure-
1. The kinetics of common Out-of-Autoclave (OOA) systems, Cycom 5320 and MTM 45-1 were characterized by Kratz et al. [9].

$$\alpha = \frac{\Delta H_t}{\Delta H_0}$$  \hspace{1cm} (1)

where, $\Delta H_0$ is the total heat of reaction and $\Delta H_t$ is the heat of reaction at time, $t$. Using multiple isothermal DSC experiments, the cure kinetic equation can be derived. For the carbon/epoxy prepreg system used in the current study, the cure kinetic equation is defined in Equation 2(a) and Equation 2(b). This equation accounts for the interplay between kinetics-controlled and diffusion-controlled reaction mechanisms [10].

$$\frac{d\alpha}{dt} = \sum_{i=1,3} K_i^{m_i}(1-\alpha)^{n_i}$$  \hspace{1cm} (2a)

$$K_i = A_i e^{-E_{A_i}/RT}, \quad l = i, j$$  \hspace{1cm} (2b)

where $A_n$ is the Arrhenius constant, $E_{A_n}$ is the activation energy, $R$ is the universal gas constant, $m_i$ and $n_i$ are reaction order-based fitting constants, $D_f$ is the diffusion constant, $T$ is the temperature, $\alpha_{co}$ is the critical degree of cure at absolute zero, and $\alpha_{cT}$ accounts for the increase in critical degree of cure with temperature. The cure kinetics equation which accounts for the effects of prepreg out-time was developed for Cycom 5320–1 OOA prepreg system [10]. Values of the parameters are given in Table 2. The out time effects were not considered in this study.

3.2. Thermal Model

In the thermal portion of the model was deduced form the 1st law of thermodynamics. The thermal equilibrium equation (Equation 3) was used to solve for energy balance in the laminate portion. The heat generation is dependent on both position in the laminate and time. In the system there are two sources of thermal energy. The heat is generated with in the system by external microwave and also heat is generated by the exothermic chemical reaction.

$$\rho c_T \frac{\partial T}{\partial t} = \nabla \cdot (k \nabla T) + v_m \rho m H_u \frac{\partial \alpha}{\partial t} + 2 \pi f \varepsilon'' E_{rms}^2$$  \hspace{1cm} (3)

where $\rho_c$ is the density of composite, $c_T$ is the specific heat capacity of composite, $T$ is temperature, $t$ is time, $k_c$ is the thermal conductivity of composite, $v_m$ is the resin volume fraction, $\rho_m$ is the resin density, $H_u$ is ultimate heat of reaction of the system and $\alpha$ is the degree of cure, $f$ is frequency of microwaves, $\varepsilon''$ is dielectric loss factor, $E_{rms}$ is root mean square power of the microwaves.

3.3 Multiphysics Cure Modeling

A thermo-chemical model was built in Comsol Multiphysics software (Figure 2). The model is built to simulate the curing of composite parts. The partial differential equations module was used in the Comsol Multiphysics software to address the cure kinetics process in the laminate. The thermal model was incorporated using the heat transfer module. The cross section of the multiphysics model used, is shown in Figure 3. The model reflects the actual manufacturing layup used to cure composite components.
A heat flux boundary condition was applied on all outer surfaces according to Equation 4. This corresponds to the cure cycle for composite manufacturing.

\[ Q = h(T_{\text{ext}} - T_{\text{boundary}}) \]  \hspace{1cm} (4)

Where Q is the heat generated, h is the convection heat transfer coefficient, \( T_{\text{ext}} \) is the temperature of the autoclave and \( T_{\text{boundary}} \) is the temperature at the external surface of the layup. The convective heat transfer coefficient was set to 150 W/m²K.

4. RESULTS AND DISCUSSION

Degree of cure and the temperature variation during composite cure were studied. The heat is generated due to the microwaves and resin cure reactions. The degree of cure is maximum at the center of the laminate. Degree of cure through thickness is shown in figure 4. The degree of cure throughout the cure cycle is shown in figure 5.

The maximum temperature at the center of the laminate is 400°F observed at 230 minutes from the start of cure cycle. Figure 6 is the temperature distribution at part center and the boundary. The time require for achieving same degree of cure using microwaves is less than that of autoclave curing.

5. CONCLUSIONS

In this study a new multiphysics simulation model was designed to study the microwave curing process of carbon fiber reinforced thermoset composite structures. Anisotropic properties of the composite and asymmetric microwave distribution in the laminate was incorporated into the model. The cure cycle was modified over the traditional thermal cure cycle to accommodate the energy due to the microwaves and to incorporate a gradual ramp in the curing phase. The simulation predicts the degree of cure and temperature distribution in the laminate. Compared with the traditional thermal curing and microwave curing is faster, energy efficient and convenient for remote use.

6. ACKNOWLEDGMENTS

Support from Intelligent Systems Center is gratefully acknowledged. The author would like to thank Manoj Reddy Rangapuram for their contribution to this work.

7. REFERENCES


8. APPENDIX

Table 1 Dimensions of the three dimensional model

<table>
<thead>
<tr>
<th></th>
<th>Length (cm)</th>
<th>Width (cm)</th>
<th>Thickness (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laminate</td>
<td>30</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Mold</td>
<td>40</td>
<td>40</td>
<td>0.5</td>
</tr>
<tr>
<td>Caul Plate</td>
<td>30</td>
<td>30</td>
<td>0.5</td>
</tr>
<tr>
<td>Consumable</td>
<td>-</td>
<td>-</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 2 Thermal properties used in the model

<table>
<thead>
<tr>
<th></th>
<th>Density (kg/m³)</th>
<th>Specific heat (J/kgK)</th>
<th>Conductivity (W/mK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laminate (prepreg)</td>
<td>1591.6</td>
<td>1260</td>
<td>x 3.302</td>
</tr>
<tr>
<td>Mold</td>
<td>2700</td>
<td>8.1</td>
<td>y 3.302</td>
</tr>
<tr>
<td>Caul Plate</td>
<td>2700</td>
<td>8.1</td>
<td>z 0.506</td>
</tr>
<tr>
<td>Consumable</td>
<td>314.2</td>
<td>0.119</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 Cure Kinetic parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1 (s^{-1})$</td>
<td>1.48×10^7</td>
</tr>
<tr>
<td>$E_{A1}/R (K)$</td>
<td>1.02×10^4</td>
</tr>
<tr>
<td>$m_1$</td>
<td>0.17</td>
</tr>
<tr>
<td>$n_1$</td>
<td>9.3</td>
</tr>
<tr>
<td>$A_2 (s^{-1})$</td>
<td>8.3×10^4</td>
</tr>
<tr>
<td>$E_{A2}/R (K)$</td>
<td>8.54×10^3</td>
</tr>
<tr>
<td>$m_2$</td>
<td>0.7</td>
</tr>
<tr>
<td>$n_2$</td>
<td>0.87</td>
</tr>
<tr>
<td>$D_2$</td>
<td>97.4</td>
</tr>
<tr>
<td>$a_{c0,2}$</td>
<td>-1.6</td>
</tr>
<tr>
<td>$a_{cT,2} (K^{-1})$</td>
<td>5.7×10^{-3}</td>
</tr>
<tr>
<td>$A_3 (s^{-1})$</td>
<td>6.39×10^7</td>
</tr>
<tr>
<td>$E_{A3}/R (K)$</td>
<td>8.94×10^3</td>
</tr>
<tr>
<td>$m_3$</td>
<td>1.65</td>
</tr>
<tr>
<td>$n_3$</td>
<td>16.63</td>
</tr>
<tr>
<td>$A_4 (s^{-1})$</td>
<td>9.8×10^4</td>
</tr>
<tr>
<td>$E_{A4}/R (K)$</td>
<td>7.1×10^3</td>
</tr>
<tr>
<td>$m_4$</td>
<td>1.66</td>
</tr>
<tr>
<td>$n_4$</td>
<td>3.9</td>
</tr>
<tr>
<td>$D_4$</td>
<td>63.3</td>
</tr>
<tr>
<td>$a_{c0,4}$</td>
<td>-0.6</td>
</tr>
<tr>
<td>$a_{cT,4} (K^{-1})$</td>
<td>3.0×10^{-3}</td>
</tr>
</tbody>
</table>
Distributed Optimization Scheme for Deep Neural Network Weight Tuning

R. Krishnan, S. Jagannathan and V. A. Samaranayake

Abstract—Learning the weights of a deep neural network (DNN) involves optimizing a cost function with respect to all the weights of the DNN. However, this optimization problem is non-convex and presents with challenges such as vanishing gradients, local minima and inefficient learning. In this paper, a novel distributed learning methodology is introduced to address these challenges for the problem of classification. In the proposed approach, layer-wise cost functions are derived with the use of additional variables. Finally, these layer-wise cost functions are optimized to learn the weights. Overall, it is demonstrated in simulation with nine data-sets that the optimization problem is convex and does not suffer from the challenges mentioned before.

I. INTRODUCTION

Deep neural networks (DNN) can be very effective in big data environments when efficient learning schemes for training DNN weights can be designed. The popular approach used in the literature for training DNN weights is back-propagation [1] wherein a cost function is first defined, and then the DNN weights are tuned by obtaining the gradient of the cost function. The effectiveness of gradient-based techniques has been well established in the literature [2]. However, it has also been demonstrated in [2]–[5] that gradient-based approaches can suffer from critical drawbacks such as: (1.) vanishing gradients, where learning signals diminish with an increase in the number of layers in the DNN [3]; (2.) unsatisfactory solutions and local minima, due to improper initial conditions [2]; (3.) lack of a proper mechanism for parallelization across layers [4].

In the literature, the use of alternate activation functions such as relu, leaky relu, drelu, etc [3], [5]; and learning rate schedules [3] are common ways of addressing the vanishing gradient problem. However, it has been shown that these approaches [3], [6] does not mitigate the vanishing gradient problem completely and only provide a workaround for satisfactory performance [5]. In addition, non-convexity of the optimization problem for DNN renders their learning phenomenon very susceptible to initial conditions [7] due to which local minima and unsatisfactory solutions are observed. One solution to the problem is the introduction of sustainable strategies that are heuristic in nature [2], [8]. Alternatively, several methods [9], [10] have been proposed that convert a non-convex problem into a convex one by either using a smoothing parameter [9] or a predetermined transformation [10]. The drawbacks of these approaches [9], [10] include the need to know the exact transformation or the smoothing parameter.

These drawbacks have motivated further research in alternate non-gradient-based learning frameworks [11]–[14] that are either computationally expensive or suffer from inefficiencies while generalizing to different neural network architectures such as convolutional neural networks, stacked auto-encoder, etc. In light of these issues, a principled methodology is introduced in this paper to address challenges in terms of: (1) the non-convexity of cost function; and (2) lack of any structured approaches that parallelize over layers in deep NN.

To address the non-convexity problem, layer wise cost functions are derived at each DNN layer by introducing additional variables such that each of these layer-wise cost function is convex and can be easily optimized. The additional variables are introduced in such a way that the output at a given DNN layer depends on the additional variables instead of the previous layer. Subsequently, these layer-wise cost functions are independently optimized to learn the additional variables and the DNN weights at each layer.

As the optimization problem with additional variables is convex, the issues due to non-convexity of cost function are not observed in contrast with [7] and other SGD-based approaches [15]. Though, our approach appears similar to [10] in terms of defining a transformation of the cost function, it is still different due to extra variables instead of pre-determined transformation. Finally, our approach is similar in spirit to [13], [14] with key differences that include (1.) a variable for every data-point in the data-set that is not maintained; and (2.) our approach can be generalized to architectures such as CNN in order to work with images efficiently.

Since, each of these layer-wise problems are independent, the methodology provides opportunities for high scalability and parallelization. To enable such learning, within the proposed approach a novel implementation algorithm is proposed that enables parallelization across both the data-batches and the layers in the DNN. In contrast with [16]–[18], the cost function is split across DNN layers and data-points and parallelization is enabled. Furthermore, the cost of communicating weights between different segments of the DNN optimization problem is much lower compared to [16]–[18].

The benefits of this approach are demonstrated in simulation on a total of eight data-set where many of them belong to big data. In the next section, the preliminaries and subsequent notations are discussed. Next the proposed approach is introduced in section II, with section III discussing the results, followed by conclusions in Section IV.
II. BACKGROUND AND PRELIMINARIES

Let there be a manufacturing process and the data at an instance be denoted as $x$. Consider the problem of classification where the objective is to determine, whether $x$ belongs to any one of the $F$ categories or not. To predict these categories, the basic idea is to estimate a map $\phi(.)$, referred to as classifier, that transforms the data-point $x$ into a probability space where the magnitude of $\phi(x)$ indicates the membership of $x$ to different categories. Magnitude of these probabilities signify which category $x$ belongs to. The underlying problem in classification is to learn the map $\phi(.)$.

To learn $\phi(.)$, one usually assumes that a data-set $X$ with labels $Y$ is available and it represents all the necessary information. The overall learning then involves approximating $\phi(.)$ through a parametric map such as neural networks, spline functions, etc [19].

In this paper, DNN is chosen to be the parametric map of choice and each training instant is denoted as $k$ with $x_k$ denoting the data-sample that is randomly sampled from $X$. Next, we assume that $\phi(.)$ can be approximated using a DNN [20] with an ideal set of weights denoted by $\theta = [W^{(1)} \cdots W^{(d)}]^T$ and write

$$
\phi(x_k) = y(x_k; \theta) + \varepsilon
= f^{(d)}(W^{(d)} \cdots (f^{(1)}(W^{(1)}(x_k)))) + \varepsilon,
$$

where $y \in \mathbb{R}^{F \times 1}$ represents the ideal output/labels for $\phi(.)$ and the target weights includes the layer-wise bias. Let the variable $d$ denote the total number of layers in the DNN and the estimated map at any training instant $k$ be denoted as $\hat{y}(x_k)$. With estimated weights denoted as $\hat{\theta} = [\hat{W}^{(1)} \cdots \hat{W}^{(d)}]^T$, the estimated map may be written as

$$
\hat{y}_k(x_k) = \hat{y}_k(x_k; \hat{\theta})
= f^{(d)}(\hat{W}^{(d)} \cdots (f^{(1)}(\hat{W}^{(1)}(x_k))))).
$$

It can be observed that the estimated map is dependent on the training instances but the ideal map is independent of $k$. To estimate $\phi(.)$ using $\hat{y}_k$, we minimize the difference between the labels(ideal parametric map) and the DNN output(estimated parametric map) at any training instance $k$ such that the difference goes to zero as $k \to \infty$.

Formally, define a loss function $\ell(y, \hat{y}(k))$ such that it represents the difference between the labels and the NN output. Since, we cannot work with the entire data-set at one, we work with a sample of data and calculate the expected value of loss at any training instance. Let $E[.]$ denote the expectation operator and write the overall cost $\hat{J}$ as the expected value of the loss to get

$$
\hat{J} = E_{\forall x_k \in X, y \in Y}[\ell(y, \hat{y}(x_k))].
$$

Note that we replace the ideal map with labels from the data-set while calculating the expected values. For brevity of notation, we suppress the sub-scripts for the expected values from hereon. Formally, the learning objective is then obtained as the minimization of this cost function $\hat{J}$ to estimate the weights such that

$$
\theta^* = \arg \min_{\theta \in \Omega} \hat{J}(k),
$$

where $\Omega$ represents the parameter space. With this setup, the proposed approach is described next.

III. OVERALL METHODOLOGY

It was mentioned before in the introduction that there are several drawbacks with a traditional gradient based approach when the function to optimize has functional nesting(the form denoted in Eq. (1)). In DNN-terminology, this refers to the presence of multiple layers that leads to the non-convexity of the optimization [3].

To alleviate this issue, the principal idea is to disassociate the learning problem across the layers such that each layer can be optimized independently. We accomplish this by introducing additional variables at each layer of the DNN and optimizing these additional variables along with the weights.

To elaborate, let us start by introducing a variable $z^{(d)}$ to replace the top-most layer in the DNN such that the cost function can be rewritten as

$$
\hat{J}(\hat{y}; \hat{\theta}, \lambda) = J(\hat{y}(z^{(d)}); \hat{\theta}, \lambda)
= E[\ell(y, \hat{y}(z^{(d)}))],
$$

where $\hat{y}(z^{(d)}) = f^{(d)}(z^{(d)})$ is the output of the DNN that depends on $z^{(d)}$. It follows that the loss function depends on $z^{(d)}$. The advantage as observed is that the optimization of the cost function does not have any nesting therefore it is equivalent to a one layer NN optimization problem. One layer NN optimization problem is relatively simpler to the original problem and has been well studied in the literature [19]. However, as optimization procedure is now dependent on the additional variables instead of the topmost layer in the DNN, the additional variables must mimic the topmost layer in the DNN. This can be achieved by adding a constraint into the optimization problem such that

$$
z^{(d)}(x) \approx W^{(d)} f^{(d-1)}(x).
$$

It can be observed that these additional variables are local observers that estimate the information required for learning the weights at layer $d$.

Next, lets introduce another variable $z^{(d-1)}$ that is constrained to mimic the penultimate layer in the DNN such that $z^{(d)}$ depends on $z^{(d-1)}$. It follows that the constraint on variable $z^{(d)}$ can be rewritten as

$$
z^{(d)} \approx W^{(d)} f^{(d-1)}(z^{(d-1)}).
$$

As mentioned earlier, this dependence introduces an additional constraint as

$$
z^{(d-1)} \approx W^{(d-1)} f^{(d-2)}(z^{(d-2)}).
$$

With this line of thought one may define a variable at each layer of the DNN constrained to mimic all the information...
prior to that particular layer in the DNN. Due to this, the overall optimization problem can be rewritten as
\[
\theta^* = \arg \min_{\theta \in \Omega} J(\hat{y}(z^{(d)}); \hat{\theta}, \hat{\lambda})
\]
subject to
\[
\begin{align*}
z^{(d)} &\approx \hat{W}^{(d)} f^{(d-1)}(z^{(d-1)}) \\
z^{(d-1)} &\approx \hat{W}^{(d-1)} f^{(d-2)}(z^{(d-2)}) \\
&\vdots \\
z^{(2)} &\approx \hat{W}^{(2)} f^{(1)}(z^{(1)}),
\end{align*}
\] (6)
where \(z^{(1)}(x) = \hat{W}^{(1)} x\). Observe that each constraint is dependent on the weight, the activation function at a layer and the additional variables corresponding to the layer below. As a result, the optimization problem from Eq. (4) can now be rewritten as shown in Eq. (6). The only difference in between these problems is the presence of additional variables and the constraints. Furthermore, observe that \(\approx\) in Eq. (6) signifies that the constraints are soft.

One might also observe that a straightforward unconstrained DNN optimization problem is being changed to introduce constraints. These constraints also introduce more variables and more parameters that must be optimized. However, the main advantage is the fact that an inherently non-convex optimization problem is converted into a series of sub-problems that can be solved independently. This introduces opportunities for parallelization and efficient scalability.

The constraints introduced into the new problem are already present in a standard DNN implicitly. However, in a traditional case, these constraints generally take the form of "input to each hidden layer in the DNN must be equal to the output of the previous layer" and we do not think of them as constraints. By introducing these additional variables, these constraints become explicit. The advantage as observed is that an optimization problem that depended on nested functional form of DNN does not depend on nested functions anymore. The lack of nested function also results in an optimization problem that is convex. However, this approach requires that the constraints are satisfied.

To enforce these constraints, we introduce a generic penalty function \(\psi(.)\) [21] such that
\[
J(\hat{y}, \hat{\theta}, \hat{Z}, \hat{\lambda}) = E[\ell(y, \hat{y}(z^{(d)})) + \sum_{i=1}^{d} \lambda_i P(W^{(i)}) + \sum_{i=2}^{d} \gamma_i \psi(z^{(i)}, W^{(i)} f^{(i-1)}(z^{(i-1)}), \) (7)
where \(Z = [z^{(1)} \ldots z^{(d)}]\). Observe that \(\gamma_i\) is the penalty coefficient for weight at the layer \(i\) and \(\psi(.)\) denotes a penalty function of choice. The choice of these penalty coefficients decide the exactness with which these constraints are enforced. In a traditional DNN, these constraints are enforced exactly, which is not the case in this formulation. It follows that at the cost of some imprecision in the constraints at each layer of the DNN, we achieve an optimization problem that is convex and is equivalent to the original optimization problem. In other words, each of these constraints are soft.

By introducing additional variables, the proposed approach lends itself to parallelization because all the layers in the DNN are not connected as is traditional with deep neural networks. Furthermore because the additional variables \(Z\) replicate the data at the hidden layers in the DNN. The convergence of the Zs before the weight updates is important and the learning rate for updating \(Z\) is chosen to be larger than the learning rule for the weights.

The advantages and disadvantages of the proposed work are summarized next

**Remark 1.** The following advantages are achieved due to the proposed methodology

1) In contrast with [15], the issue of vanishing gradients is not present because a cost function is obtained for each layer.

2) In contrast with other distributed learning mechanisms [18], the cost of communication between different nodes

---

![Fig. 1: Illustration for the proposed approach.](image-url)
**Algorithm 1** Algorithm For the Proposed Approach

1. **for** each epoch $k$
2. **Input**: $\mathcal{X}$
3. **for** each batch in $\mathcal{X}$
4. Compute Cost $H$
5. **for** $j = 1 \rightarrow T$
6. Expectation Step:
   \[
   \hat{Z}(k+1, j+1) = \hat{Z}(k) - \eta_1 \Delta \hat{Z}(k+1, j) H
   \]
7. **end for**
8. Learning Step:
   \[
   \hat{\theta}(k+1) = \hat{\theta}(k) - \eta_2 \Delta \hat{\theta}(k) H
   \]
9. \[
   \hat{\lambda}(k+1) = \hat{\lambda}(k) + \eta_3 \Delta \hat{\lambda}(k) H
   \]
10. **end for**

is reduced significantly less compared to [18]. Furthermore, in contrast with [16], [17], the proposed approach is designed to parallelize across different layers in the DNN as well as different data-samples.

3) The major disadvantage of the proposed approach is that these use of these additional variables introduce a computational load. However, with an appropriate parallelized algorithm, the impact of such a load can be minimized.

4) One big disadvantage is that due to use of these additional variables, we have introduced some extra tuning parameters that must be chosen for optimality.

Finally, the proposed approach is multiple data-sets in the next section.

**IV. RESULT AND DISCUSSION**

The details of all the data-sets analyzed in the section are illustrated in Table II.

The performance is first demonstrated on the Mnist data-set.

1) MNIST data-set: To start with, we compare the proposed learning approach with and without the Zs for a varying number of hidden layers in the DNN. For this experiment, tanh activation functions are used. The learning rates $\eta_1$ for weights are set at 0.0001 whereas batch size is chosen as 128. The number of hidden layer neurons and the learning rate $\eta_2$ is chosen as 150 and 0.1 respectively. Zs are updated at each step for 100 times. The lambdas are kept fixed for this experimentation and the results are recorded in Fig. 2a. With an increase in the number of hidden layers in the DNN, the gap in the accuracy’s achieved by the two approaches increase. Also, it is observed that the model with Zs performs significantly better than its counterpart.

Next, a seven hidden layers DNN model is trained for hundred iterations and the results are illustrated in Fig. 2a. The training accuracy for the case without the Zs does not improve beyond 15 epochs. On the other hand, the model with Z achieves an accuracy of 90% in less than 25 iterations. For the case without the Zs, this behavior is seen due to vanishing gradient [3] in deep neural networks. The problem does not occur in the model where Zs are used because the weights at each layer learn locally. To empirically demonstrate this, the progression of the cost function with respect to the number of training iterations is shown in Figure 2c. The case with the Zs converges faster relative to the case without the Zs.

Next, the behavior of the methodology with respect to different parameters is studied. These results are shown in Table I, where one parameter is changed while the others are kept fixed. Table Ia shows the results for different activation functions. Table Ib shows the comparison for different $\lambda$ values. Table Ic illustrates the impact of varying $\eta_1$ while Table Id illustrates the impact of $\eta_2$.

It is observed from Table Ic, that when the value of $\eta_1$ is chosen as 0.1, the learning phenomenon is unsatisfactory because the weights converge faster than the corresponding Z variables. This creates a peculiar problem, since Z variables denote the data at output of every layer, the weights learn to predict based on the values of Z. However, if Z variables do not reflect the data before the weights learn, the weights tend to learn from wrong representation of data thus leading to unacceptable performance.

From our analysis, it was observed that the best case results were obtained when the DNN is chosen with four hidden layers containing fifty hidden neurons each. The batch size is chosen as 128 with tanh activation function. The regularization term $\alpha$ was fixed at 0.00001, whereas the penalty $\gamma$ is chosen at 10. The learning rates $\eta_1$ for the weights and $\eta_2$ were fixed at 0.0001 and 0.1 respectively. For each epoch, Zs were updated 100 times. The best case test accuracy achieved for the Mnist data-set at this parameter setting was 94.32%. Therefore, these parameters are used from here on for analysis on the Mnist data-set. Next, the performance is analyzed on the problem of classification.

**A. Classification**

Next, we analyzed the performance of the proposed model on the data-sets mentioned in Table IV and these results are presented in Table IV. To demonstrate the ability of the methodology to work with different architectures, we test CNN, DAE, SAE and MLNN where the parameters are summarized in III.

The ability of the model to handle large complex problems by breaking it into simpler sub-problems and thus achieving faster learning is well illustrated by the testing accuracy results for the large data-sets such as Gas, Rolling and mnist, where the model has a very high testing accuracy (more than 90%). Since the data-sets Gas and Rolling are larger than the mnist data-set, hence for faster learning of the large data, the value of $\eta_2$ for these two data-sets was set higher than that for mnist data-set.

Next, Four learning paradigms, namely DFA (Direct Feedback Alignment) [12], EDL (Error Driven Learning) [28], FA (Feedback Alignment) [29] and EDL (Error-driven Learning) [28], are tested with a total of ten data-sets. For all the learning paradigms, a multilayer NN with 10 hidden layers is utilized. The results along with the model parameters are given in
TABLE I: (a-d) Accuracy's for different parameter changes

<table>
<thead>
<tr>
<th></th>
<th>$f^{(i)}$</th>
<th>$\lambda$</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>tanh</td>
<td>94.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>87.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sigmoid</td>
<td>74.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>94.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>tanh</td>
<td>94.19</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>93.83</td>
<td>22.13</td>
<td>12.61</td>
</tr>
<tr>
<td></td>
<td>sigmoid</td>
<td>94.19</td>
<td>27.12</td>
<td>12.47</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>94.33</td>
<td>12.47</td>
<td>12.61</td>
</tr>
<tr>
<td>(c)</td>
<td>tanh</td>
<td>94.19</td>
<td>0.001</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>93.83</td>
<td>22.13</td>
<td>12.61</td>
</tr>
<tr>
<td></td>
<td>sigmoid</td>
<td>94.19</td>
<td>27.12</td>
<td>12.47</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>94.33</td>
<td>12.47</td>
<td>12.61</td>
</tr>
<tr>
<td>(d)</td>
<td>tanh</td>
<td>94.19</td>
<td>0.001</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>93.83</td>
<td>22.13</td>
<td>12.61</td>
</tr>
<tr>
<td></td>
<td>sigmoid</td>
<td>94.19</td>
<td>27.12</td>
<td>12.47</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>94.33</td>
<td>12.47</td>
<td>12.61</td>
</tr>
</tbody>
</table>

TABLE II: % accuracies for the various data-sets with the proposed framework and four different learning algorithms with generalization error in the parenthesis.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>$p$</th>
<th>$n$</th>
<th>$J^p$</th>
<th>Model Parameters</th>
<th>DFA</th>
<th>FA</th>
<th>SGD</th>
<th>EDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rolling [22]</td>
<td>11</td>
<td>35000</td>
<td>4</td>
<td>$\alpha = 0.01, b = 100, d = 7, hln = 100, \lambda = 0.001, n_{neigh} = 2, \rho = 1$</td>
<td>99(0.0001)</td>
<td>99(0.0001)</td>
<td>99(0.6)</td>
<td>99(0.0)</td>
</tr>
<tr>
<td>Sensorless [23]</td>
<td>48</td>
<td>78000</td>
<td>11</td>
<td>$\alpha = 0.01, b = 100, d = 8, hln = 100, \lambda = 0.001, n_{neigh} = 2, \rho = 1$</td>
<td>94(0.008)</td>
<td>94(0.008)</td>
<td>95(0.008)</td>
<td>94(0.002)</td>
</tr>
<tr>
<td>MNIST [24]</td>
<td>784</td>
<td>72000</td>
<td>10</td>
<td>$\alpha = 0.01, b = 100, d = 8, hln = 100, \lambda = 0.001, n_{neigh} = 2, \rho = 1$</td>
<td>92(0.1)</td>
<td>94(0.2)</td>
<td>95(1)</td>
<td>97(0.6)</td>
</tr>
<tr>
<td>NotMNist [25]</td>
<td>784</td>
<td>81000</td>
<td>10</td>
<td>$\alpha = 0.01, b = 100, d = 9, hln = 100, \lambda = 0.001, n_{neigh} = 2, \rho = 1$</td>
<td>90(1)</td>
<td>82(1.8)</td>
<td>88(2)</td>
<td>92(0.8)</td>
</tr>
<tr>
<td>CIFAR10 [26]</td>
<td>3072</td>
<td>50000</td>
<td>10</td>
<td>$\alpha = 0.01, b = 100, d = 10, hln = 100$</td>
<td>81(7.21)</td>
<td>80(8.00)</td>
<td>82(12.11)</td>
<td>84(8.77)</td>
</tr>
<tr>
<td>Arcene [27]</td>
<td>10000</td>
<td>100</td>
<td>2</td>
<td>$\alpha = 0.01, b = 100, d = 9, hln = 100, \lambda = 0.001, n_{neigh} = 2, \rho = 1$</td>
<td>69(0)</td>
<td>61(0)</td>
<td>51(0)</td>
<td>78(0)</td>
</tr>
<tr>
<td>Dexter [27]</td>
<td>20000</td>
<td>300</td>
<td>2</td>
<td>$\alpha = 0.01, b = 100, d = 8, hln = 100, \lambda = 0.001, n_{neigh} = 2, \rho = 1$</td>
<td>71(0.7)</td>
<td>77(0.9)</td>
<td>81(1.1)</td>
<td>80(1)</td>
</tr>
<tr>
<td>Gisette [27]</td>
<td>5000</td>
<td>6000</td>
<td>2</td>
<td>$\alpha = 0.01, b = 100, d = 8, hln = 100, \lambda = 0.001, n_{neigh} = 2, \rho = 1$</td>
<td>92(0.25)</td>
<td>93(0.21)</td>
<td>91(0.3)</td>
<td>98(0.07)</td>
</tr>
</tbody>
</table>

(a) Accuracy's of both models for 10 different sized hidden layers with tanh activation
(b) Progression of accuracy's of both models for 7 hidden layers
(c) Progression of the Total Cost for the two models

Fig. 2

Table II. Consistent performance is observed across data-sets with the proposed methodology and the results on large dimensional data-sets is very good too as observed from the results.

V. CONCLUSIONS

In this paper, a mathematical construct was introduced to transform a non-convex optimization into a convex one. The proposed approach appears to converge faster compared to SGD and other error-driven learning regimes. Furthermore, a 15% improvement in accuracy is observed with the use of additional variables while optimizing deep architectures. It is demonstrated both in simulation and theoretically that the proposed learning approach with or without the additional variables is equivalent. Overall, the distributed learning procedure introduced here is demonstrated to be an efficient alternative to traditional learning procedures. The issue of incidental endogeneity is left for future effort.

ACKNOWLEDGMENT

This research was supported in part by an NSF I/UCRC award IIP 1134721 and the Intelligent Systems Center.
TABLE III: Parameters of the different models used in this paper.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>$n_{conv} = 2$, $n_{layers} = 2$, filter size = 5, pool shape = (3,3), $h_l=128$, $a = 0.01$, $b = 100$, $\lambda = 0.001$, $n_{epochs} = 2$, $\rho = 1$</td>
</tr>
<tr>
<td>MLNN</td>
<td>$d = 7$, $h_l=128$, $a = 0.01$, $b = 100$, $\lambda = 0.001$, $n_{epochs} = 2$, $\rho = 1$</td>
</tr>
<tr>
<td>SAE</td>
<td>$n_{init} = 3$, $h_l=128$, $a = 0.01$, $b = 100$, $\lambda = 0.001$, $n_{epochs} = 2$, $\rho = 1$</td>
</tr>
<tr>
<td>DAE</td>
<td>$n_{init} = 3$, $h_l=128$, $a = 0.01$, $b = 100$, $\lambda = 0.001$, $n_{epochs} = 2$, $\rho = 1$</td>
</tr>
</tbody>
</table>

TABLE IV: % accuracies for the various data-sets with the proposed framework and four different architectures with generalization error in the parenthesis.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>MLNN</th>
<th>CNN</th>
<th>SAE</th>
<th>DAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rolling</td>
<td>99(0.0001)</td>
<td>99(0.0001)</td>
<td>99(0.6)</td>
<td>99(0.0)</td>
</tr>
<tr>
<td>Sensorless</td>
<td>94(0.008)</td>
<td>94(0.008)</td>
<td>95(0.008)</td>
<td>94(0.002)</td>
</tr>
<tr>
<td>MNIST</td>
<td>94(0.2)</td>
<td>95(1)</td>
<td>97(0.6)</td>
<td>97(0.6)</td>
</tr>
<tr>
<td>NotMNist</td>
<td>88(2)</td>
<td>92(0.8)</td>
<td>97(0.6)</td>
<td>97(0.6)</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>81(7.21)</td>
<td>80(8.00)</td>
<td>82(12.11)</td>
<td>84(8.77)</td>
</tr>
<tr>
<td>Aracne</td>
<td>69(0)</td>
<td>61(0)</td>
<td>51(0)</td>
<td>78(0)</td>
</tr>
<tr>
<td>Dexter</td>
<td>71(0.7)</td>
<td>77(0.9)</td>
<td>81(1.1)</td>
<td>80(1)</td>
</tr>
<tr>
<td>Gisette</td>
<td>92(0.25)</td>
<td>93(0.21)</td>
<td>91(0.3)</td>
<td>98(0.07)</td>
</tr>
</tbody>
</table>

REFERENCES

Near-infrared broad band chiral plasmonic metasurface absorbers

Leixin Ouyang,¹ and Xiaodong Yang¹

¹Department of Mechanical and Aerospace Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA

Abstract: Chirality is common and essential in many fields, such as biology, chemistry and physics. Metasurfaces can provide extraordinary optical properties for many promising applications. Here, a new type of near-infrared broadband chiral plasmonic metasurface absorbers based on the combination of several double-rectangle patterns is designed and experimentally demonstrated to achieve large circular dichroism (CD) in a broad wavelength range. The maximum chiral absorption can reach to 0.7 and the average CD is around 0.5 within the operation wavelength range from 1.35 µm to 1.85 µm. The high CD can lead to a local temperature difference of the absorber. The high-contrast reflective infrared images are realized by changing the wavelength and polarization of incident light. The broadband chiral metasurface absorbers promise future applications in many areas such as optical filters, thermal energy harvesting, optical communication, and chiral imaging.

© 2018 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

1. Introduction

Chirality exists in many fields: like the two spiral kinds of DNA. Chirality has its special properties, and the study of chirality leads to many recovery of the nature and many applications [1, 2]. Metasurface has drawn much attention and provides optional breakout from nature materials [3, 4]. There are usually a pair of enantiomeric forms for optical chiral materials [5], which are two non-superimposable mirror images with different responses to the left-handed and right-handed circularly polarized (LCP and RCP) light. However, the chiral broad band absorbers with high efficiency are not attainable from natural materials. Thus different kinds of metamaterials and metasurfaces chiral optical designs are well developed: three-dimension (3D) structures like Kite-shaped opening/deposition [6], double-layer ellipses [7] and double-layer E structure [8]; different top-layer designs like Gamma-like array [9], gammadion pattern [10] single-L pattern [11] and two-L shape [12]; negative top-layer designs like C-shape array [13], bars array [14], squares array [15] and spots array [16]; Liquid-crystal-loaded chiral metasurfaces [17]; generative model computer-aid design [18] and deep-learning computer-aid optimization for chiral design [19].

Broadband absorption can be applied to many fields, especially for energy harvesting, bio & chemical measurement, optical filter and imaging. Here, we demonstrated a broad band chiral plasmonic metasurface absorbers working in the near-infrared wavelength region. The designed chiral metasurface absorber, consisting of a three-layer metal-dielectric-metal structure, can selectively absorb the LCP or RCP normal incident light. The results show 70% average chiral absorption and above 50% average CD in absorption at the resonance wavelength range from 1.35 µm to 1.85 µm. According to the previous research, the high CD can be reached and the effective wavelength can be tuned, thus the broadband metamaterial absorbers are developed using multiple chiral double-rectangle resonators of one unit cell in near-infrared regime. To elucidate the mechanism of optical chiral absorption among the effective wavelength range, electric field distributions and temperature distributions are mapped for the LCP and RCP incidence light. In order to further visualize the difference between the narrow-band and broadband absorbers and the potential applications of the chiral metasurface absorbers, the high-contrast near-infrared reflective chiral images of a Circle like
patterns depending on different incident polarizations are demonstrated. The results can be applied to many applications such as optical filters [9], thermal absorbers [8, 11], optical communication devices [7, 8, 10, 12], and chiral imaging [15, 20] and holograms [14, 21].

![Diagram of broad band chiral absorbers](image)

**Fig. 1.** (a), (b) Schematics of the broad band chiral metasurface absorbers in Form A and Form B. (c) Top view schematics in Form A with different geometric parameters: \( P_x = 1500 \text{ nm}; P_y = 400 \text{ nm}; L_{1-5} = 350, 410, 475, 540, 610 \text{ nm}; W_{1-5} = 135 \text{ nm}; \) \( OP_{1-5} = 10 \text{ nm} \). (d) SEM image of the broad band chiral absorber in Form A. Scale bar represents 500 nm.

### 2. Design and characterization of chiral metasurface absorbers

The chiral metasurface absorbers are composed of a top 55 nm-thick gold (Au) layer patterned with double-rectangle resonators, a 130 nm-thick silica (SiO\(_2\)) spacer layer and a 200 nm-thick gold ground plane on a glass substrate, with thicknesses denoted as \( t_s, t_d \) and \( t_m \), respectively. The schematics of the designed unit cells for two enantiomer forms are shown in Figs. 1(a) and 1(b). The narrow-band chiral absorbers (one double-rectangle resonator in unit cell) are also presented as reference and broadband chiral absorbers have six resonators in one unit cell. For the single double rectangle resonator, the length of the rectangle length \( L \) can determine the resonate wavelength; thus to achieve broadband absorption, the arrangement of several resonators on one unit cell is feasible. For each double-rectangle absorber, it consists of two connected gold rectangles with overlapped space \( OP \), rectangle width \( W \), and rectangle length \( L \), and the resonators are marked as 1-5. The rectangular unit cell has the periods of \( P_x \) and \( P_y \). The shifting of two combined simple rectangle resonators is helpful to form the mirror-break pattern for the chiral absorber; and according to Fabry-Pérot cavity [9] the three-layer structure also enhances the absorption of the incident lights. Due to the thick bottom gold layer the transmission is negligible \( (T = 0) \), thus the absorption can be calculated as \( A = \frac{I - R}{I} \) (where \( A, T \) and \( R \) denote the absorption, transmission and reflection, respectively). The two enantiomer forms (Form A and Form B, in Figs. 1(a) and 1(b)) create two different response for incident CPL, which is opposite to the other, so that the absorption performance of the chiral absorber in Form A is mainly analyzed. The Au-SiO\(_2\)-Au multilayer is deposited on a glass substrate, with the Au layer deposited by sputtering and the SiO\(_2\) layer deposited by electron-beam evaporation. Then the designed double-rectangle pattern array is milled into the top gold layer using a Focused ion beam (FIB) system (FEI Helios Nanolab 600), thus our chiral absorber is free from the dedicated milling nor the complicate 3D fabrication. Figs. 1(c) and 1(d) show the diagram of the broad band absorber and the scanning electron microscope (SEM) images of the top view of the fabricated chiral absorbers in Form A, where the geometric parameters \( P_x, P_y \) and \( L_{1-5} \) of the single resonators inside an unit cells are arranged in order to reach a broad band absorption, while \( W_1 = 135 \text{ nm} \) and \( OP_{1-5} = 50 \text{ nm} \) are set as constants.

The optical absorption performance are measured with a Fourier transform infrared spectrometer (FTIR, Nicolet 6700) connected to an infrared microscope. The CPL is gained by setting a linear polarizer and a quarter-wave plate before the light source of the FTIR.
Numerical simulations are also performed to investigate the optical responses of the chiral absorbers under both circular polarizations, where the permittivity of gold is taken from the experimental data with the imaginary part increased by three times and the refractive index of SiO$_2$ is set as a constant, with a value of 1.45. Figure 2(a) and Figure 2(b) plot the simulated and measured absorption spectra of the chiral metasurface absorber in Form A under LCP and RCP incidence. The CD in absorption defined as CD=$|A_{LCP} - A_{RCP}|$ is high, and as shown in Figure 2(a) the average CD is over 0.5 within the range from 1.35 $\mu$m through 1.85 $\mu$m. The slight difference between the simulated and measured absorption spectra could be caused by the defects and containment arising from the fabrication process as well as unexpected coupling effect between the neighboring resonators. As references, Figure 2(b) also plots the measured absorption spectra for single-resonator chiral absorbers marked as Ref.1 to Ref.5: the single-resonator cannot cover the wide absorption band, and however there is also tradeoff between the wider absorption band and a higher CD.

![Simulated and measured absorption spectra of the chiral metasurface absorber in Form A (Sample 3) under LCP and RCP incidence.](image1)

The electric field and temperature distributions of the broad band chiral absorber in Form A are gained, which is helpful to reveal the different responses to the incident circularly polarized lights. As shown in Figure 3(a), the combination of the five double-triangle absorbers broaden the absorption regime: the resonators in one unit cell share different length and works at different resonate wavelength, as well as it shows that the LCP and RCP incidences interact differently with the absorber. The simulated electric field distributions also shows that nearing resonators can cause the coupling effects and small peak shifts. Furthermore, the heat transfer equation $\nabla \cdot (-k \nabla T) = q$ is solved to gain the temperature distribution, where $T$ is the temperature, $k$ is the thermal conductivity and $q$ is the heat generation density in metal, $q(r)=(\omega/2)\text{Im}[\varepsilon(\omega)]|\varepsilon_0|E(r)^2$. The top surface of the air and bottom substrate are set as the fixed room temperature of 273 K and the side boundaries are periodic. The incident lights can be absorbed by resonators and heat is generated mostly inside the top-layer double-rectangle pattern. Due to the much lower heat conduction ability of the surrounding air and silica, the top gold patterns can reach a higher temperature then the temperature gradient will lead to a temperature equilibrium. Correspondingly, the top gold patterns have the highest temperature distributions and accordingly the highest temperature positions shift from the smaller resonators to the larger resonators as the increases of incident lights wavelength. Figure 3(b) shows the local temperature increase of the chiral absorber ($\Delta T = \left|T_{LCP} - T_{RCP}\right|$) are caused by the absorption of incident LCP or RCP light and the incident light power of 50 $\mu$W per unit cell: when the wavelength is short, the short resonators can reach a higher temperature than the other resonators; however, when the wavelength is large, the resonators with a long length have a high temperature.
Next, the chiral metasurface absorbers are fabricated to gain the high-contrast near-infrared reflective images. As shown in Figure 4(a), the circled logo is combined by six cells, Cell 1 through Cell 6 share different absorbers: Cell 1 and Cell 4 with opposite forms of broadband chiral absorbers that can work within 1.35–1.85 µm; Cell 2 and Cell 5 shares mirror narrow chiral absorbers, which resonate around 1.3 µm; Cell 3 and Cell 6 both work on wavelength at 1.55 µm but they are mirror images to each other. Then the logo is illuminated under two incident lights of different wavelength, 1.31 µm and 1.58 µm, along with a combination of linear polarizer and an achromatic quarter-wave plate; then the beam is focused normally onto the six-cell logo using a 20x objective lens. The reflected light from the sample is directed and captured by an infrared camera. Figure 4(a) is the SEM image of the six-cell logo with a zoom-in image, the diameter of the logo is around 45 µm. Figure 4(b) shows the reflected images of the six-cell logo under LCP and RCP lights with two different wavelength. Due to the different working wavelength and the two forms of absorbers, the logo shows several dark areas images under these incident lights: when the incident LCP (RCP) light has a wavelength of 1.31 µm, the Cell 4 & Cell 5 (Cell 1 & Cell 2) are darker than surrounding areas; but if it comes to 1.58 µm LCP (RCP) light, only Cell 4 & Cell 6 (Cell 1 & Cell 3) are dark. In other words, the dark area of the six-cell logo can be tuned by switching the incident wavelength as well as polarization.
3. Conclusion

In summary, we have designed and demonstrated a broadband chiral absorber working in the near infrared regime with a wide absorption band (1.35 μm ~ 1.85 μm) and large circular dichroism (CD>0.5). The broad band absorber is based on the double triangle structure which can be fabricated without complicate progress, and the combination of several resonators broaden the working range. Meanwhile, the near-infrared reflective chiral images are gained with distinct responses to different incident polarizations and wavelength, presenting the imaging and sensing application of the broad band chiral absorbers. These results confirm the advancing properties in optical filters, imaging and sensing, thermal harvest and optical communication.

Acknowledgment

The authors acknowledge the support from the Intelligent Systems Center and the facility support from the Materials Research Center at Missouri S&T.
References

DIRECT LASER DEPOSITION OF FE-CR, FE-NI AND FE-NI-CR ALLOYS AS A PATHWAY TO ADDITIVE MANUFACTURING OF HIGH-ENTROPY ALLOYS

Matthew Luebbe, Department of Materials Science and Engineering

Maalavan Arivu, Department of Materials Science and Engineering

ABSTRACT
High-entropy alloys are a novel class of alloys with promising properties for a number of applications. Some of those applications, especially nuclear energy, will require additive manufacturing of these alloys. The high degree of complexity of high-entropy alloys requires elemental powders deposited by powder-fed processes such as direct laser deposition (DLD), but such processes have been known for compositional variation. To study this, binary and ternary alloys of Fe-50%(Ni, Cr), Fe-30%(Ni, Cr), and Fe-Ni33%-Cr33% were fabricated via DLD and characterized with scanning electron microscopy (SEM), energy dispersive X-ray spectroscopy (EDS), and electron backscatter diffraction (EBSD). Initial results show large compositional variation from the intended compositions, most of which can be accounted for by differences in particle size, indicating this variable must be carefully controlled in future experiments.

1. INTRODUCTION
High-entropy alloys are a new class of alloys which consist of multiple (usually 5+) principal elements in near-equal proportions. The high configurational entropy of multiple different atoms in the lattice tends to stabilize solid-solution phases and promote single-phase microstructures [1]. In addition, this and other effects lead this class of alloys to have unique combinations of properties, including high strength-weight ratio, high temperature strength, high wear and oxidation resistance, and high thermal structural stability [1,2]. Their structure also suggests they should be resistant to irradiation, making them promising for nuclear energy applications. However, many nuclear applications require complex internal geometry, suggesting the use of additive manufacturing. Most studies done thus far on high-entropy alloys have involved cast specimens in the as-cast state [3].

Additive manufacturing (AM) is a new manufacturing technology that can produce 3D components with the aid of CAD models, allowing production of internal geometry. Powder bed processes like Selective Laser Melting (SLM) are more common, but powder-fed methods like DLD allow the use of elemental powders instead of prealloyed powders, making it more convenient for the production of high entropy alloys. Many studies have been done on additive manufacturing of high-entropy alloys [4,5], but most of those studies have been performed on Co-containing alloys, which is not ideal for nuclear applications. As such, several non-Co-containing HEAs such as FeNiMnCr, have been developed for this application [6].

To develop an alloy of this composition for additive manufacturing, first several alloys of simpler composition were chosen. For this paper, two binary alloys each of Fe-Ni and Fe-Cr were fabricated, as well as an equimolar ternary alloy of Fe-Ni-Cr. Fe-Ni-Cr especially have been extensively studied with respect to stainless steels, and some studies have even been done on additive manufacturing of this ternary alloy system [7-9]. To construct alloys of varying compositions such as those in Reference [7], it is necessary to use a powder-fed AM process such as DLD. However, DLD requires optimal powder flow to achieve the correct composition, which requires the powders to be spherical via gas-atomization and in the correct size ratio based on their density [10]. Smaller powders get accelerated faster out of the powder feeder than the larger ones.

2. EXPERIMENTAL
2.1. Materials
Commercially available gas-atomized pure powders of Fe, Ni, and Cr were obtained and analyzed for particle size distribution. The average size for the iron powders was around 50 μm, while the average size for the other two powders was around 150 μm. The powders were combined in the required ratios and mixed using a mechanical mixer before being added to the powder feeder.

2.2. Experiment
A hybrid cell combining a DLD system with a CNC system was used to deposit the powders. The process was performed under Ar gas flow to prevent oxidation of the powders during the process. The powders were deposited onto a 304 stainless steel substrate containing both Ni and Cr to limit elemental segregation and embrittlement during the DLD process. The deposit geometry was a simple thick line, and the samples were sectioned perpendicular to this line to observe the microstructure and composition close to the deposit-substrate interface as well as in the bulk.

3. RESULTS
The overall composition of the alloys was determined using an EDS point scan, and any compositional variations were observed using EDS elemental mapping. The grain boundary structure was
determined by EBSD, which also gave a map of different phases present in the structure.

3.1. Fe-Ni alloys
3.1.1. Bulk Composition and Microstructure

Table 1: EDS quantitative results for Fe-50%Ni (A) and Fe-30%Ni (B)

<table>
<thead>
<tr>
<th>Element</th>
<th>Wt% A</th>
<th>Wt% B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>0.18</td>
<td>0</td>
</tr>
<tr>
<td>Ti</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>Cr</td>
<td>11.50</td>
<td>4.39</td>
</tr>
<tr>
<td>Mn</td>
<td>0.71</td>
<td>0</td>
</tr>
<tr>
<td>Fe</td>
<td>66.25</td>
<td>78.99</td>
</tr>
<tr>
<td>Ni</td>
<td>21.23</td>
<td>16.52</td>
</tr>
</tbody>
</table>

As observed in Table 1, both Fe-Ni binary alloys are more iron-rich than expected, and there is also significant Cr contamination, either from the powders or from residue powder left over in the feeder reservoir. As indicated in Fig 1, the Cr contamination mostly segregated out of the matrix, changing the local crystal structure. The Fe-50%Ni alloy, whose phase map is not shown, is entirely fcc, even around Cr-rich regions. Figure 2 shows the grain structure of both binary alloys. Both alloys have elongated grain structure with numerous low angle grain boundaries shown by red lines. However, the scale of the grains in Fe-50%Ni is much larger.

Figure 1: EDS and EBSD maps of Fe-30%Ni away from the interface: (a) Fe, (b) Ni, (c) Cr, and (d) phase map. Blue is bcc phase and yellow is fcc phase.

Figure 2: EBSD of (a) Fe-50%Ni and (b) Fe-30%Ni with grain boundaries highlighted. Black lines correspond to misorientation angles greater than 15°, while red lines correspond to 2-15° misorientation.

3.1.2 Interface

As observed in Fig 3, the Fe-Ni alloys and the 304 substrate did not interpenetrate, so a clear transition can be seen. The stainless steel is fcc and has equiaxed grains, while the low-Cr Fe-30%Ni alloy is bcc with elongated grains due to the rapid cooling associated with DLD.

Figure 3: EDS and EBSD maps of Fe-30%Ni at the interface: (a) Ni, (b) Cr, (c) phase map, and (d) grain structure. Blue is bcc phase and red is fcc phase.
3.2. Fe-Cr alloys

3.2.1. Bulk Composition and Microstructure

Table 2: EDS quantitative results for Fe-50%Cr (A) and Fe-30%Cr (B)

<table>
<thead>
<tr>
<th>Element</th>
<th>Wt% A</th>
<th>Wt% B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0.62</td>
</tr>
<tr>
<td>O</td>
<td>0</td>
<td>0.28</td>
</tr>
<tr>
<td>Ti</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>Cr</td>
<td>32.72</td>
<td>10.55</td>
</tr>
<tr>
<td>Fe</td>
<td>65.17</td>
<td>87.35</td>
</tr>
<tr>
<td>Ni</td>
<td>0.95</td>
<td>1.02</td>
</tr>
</tbody>
</table>

As observed in Table 2, both Fe-Cr alloys are more iron-rich than expected, with Fe-30%Cr noticeably deficient in Cr. In addition, there are noticeable O and C impurities that manifest themselves as oxides and carbides in the Fe-30%Cr alloy. Figure 4 shows the Cr and O rich regions that formed in both Fe-Cr alloys, as well as the phase map showing Cr$_2$FeC$_6$ forming in Fe-30%Cr.

Figure 4: EDS and EBSD maps of Fe-30%Ni away from the interface: (a) Fe, (b) Ni, (c) Cr, and (d) phase map. Red is bcc, blue is fcc, and green is Cr$_2$FeC$_6$. Note the CrO enrichment appear to correspond to fcc regions on the phase map.

3.2.2. Interface

Figure 5 shows that these alloys show interpenetration and segregation effects near the interface between deposit and substrate. Figure 5a and c shows the Ni-rich substrate has grown into the deposit, while Fig 5b and d shows that Cr has segregated into grain-like shapes, although Fig 6 does not indicate grain boundaries there. Figure 6 shows the vast difference between the equiaxed grains of the stainless steel substrate and the grain structure of the Fe-Cr alloys.

Figure 5: EDS maps of Fe-Cr alloys near the interface: (a) Fe-50%Cr Ni map, (b) Fe-30%Cr Ni map, (c) Fe-50%Cr Cr map, and (d) Fe-30%Cr Cr map.

Figure 6: EBSD of (a) Fe-50%Ni and (b) Fe-30%Ni with grain boundaries highlighted. Black lines correspond to misorientation angles greater than 15°, while red lines correspond to 2-15° misorientation.
3.3. Fe-Ni-Cr ternary

Table 3: EDS quantitative results for Fe-33%Ni-33%Cr

<table>
<thead>
<tr>
<th>Element</th>
<th>Wt%</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.70</td>
</tr>
<tr>
<td>O</td>
<td>0.37</td>
</tr>
<tr>
<td>Cr</td>
<td>23.37</td>
</tr>
<tr>
<td>Mn</td>
<td>0.55</td>
</tr>
<tr>
<td>Fe</td>
<td>56.13</td>
</tr>
<tr>
<td>Ni</td>
<td>18.88</td>
</tr>
<tr>
<td>Total</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The data in Table 3 indicates that the equiatomic ternary alloy Fe-Ni-Cr is also iron-rich and Ni- and Cr-poor. Figure 7 shows that Cr and O enriched regions occur as in the other alloys, but the entire alloy is still fcc according to the EBSD results. The grains are larger and more uniform than the grains of other alloys observed in this paper. This may be due to early appearances of the high-entropy effect which favors single-phase solid solution formation.

Figure 7: EDS and EBSD maps of Fe-33%Ni-33%Cr away from the interface: (a) Ni, (b) Cr, (c) O, and (d) grain structure.

4. CONCLUSIONS

Overall, all of the alloys studied were found to be more iron-rich than the original powder mixtures were. Since smaller powders are accelerated faster out of the powder feeder, this behavior is observed [10]. Thus, future work will require taking particle sizes into account. In addition, although Cr has been known to segregate in particles in other alloys at these concentrations, the fast cooling rates near the surface appeared to enhance these tendencies in the Fe-Cr binary alloys tested.

5. ACKNOWLEDGMENTS

The authors would like to thank Dr. Frank Liou and his group in the Mechanical and Aerospace Engineering Department of Missouri S&T for providing the DLD hybrid cell used for deposition of the samples. They are also grateful for funding support from the innovation grant from Strategy, Institutional Research and Assessment at Missouri S&T. They would also like to thank Wenyuan Cui, Ph.D. student in the department of mechanical engineering for helping with the deposition. Finally, they would like to thank the Intelligent Systems Center for their support.

REFERENCES

Diagnosing Dementia Using Smart Chair, Wearable Device, and Testing Apps

Marium Yousuf, Sajal Das, Devin Burns
Missouri University of Science & Technology
Department of Computer Science
*Department of Psychology

Early and reliable diagnosis of dementia can help patients and their caretakers to take timely actions, in terms of treatments and effective future planning. Sedentary positions can reveal a lot about patients’ bodily movements, and so we use a smart chair installed with force sensitive resistors that record pressure points of a patient while taking a screening test. We expect that using combination of a smart chair, a wrist device, and screening test on an iPad, it is possible to diagnose most kinds of dementia earlier than the current methods with reliability.

I. Introduction

Dementia is an umbrella term for the diseases that result in cognitive decline, symptoms of which can include memory loss or ability to perform usual activities [1]. This paper discusses dementia and its symptoms, how it affects an individual diagnosed with it, ways to treat or avoid it, and better, possibly quicker, and more effective approaches to diagnose than the current methods. Our focus is the combination of using a smart chair, a wearable device, and a Saint Louis University Mental Status (SLUMS) examination score, which yields signs of dementia.

Dementia has multiple forms, Alzheimer’s disease being the most common [1], and its prevalence includes multiple stages. These stages begin with an asymptomatic phase when there are no apparent signs of dementia present, then they transition into a phase where the symptoms are not severe enough to be considered dementia, however, there are some cognitive changes such as forgetfulness, lack of focus, and behavioral inconsistency leading towards the diagnosis of Mild Cognitive Impairment (MCI) [2]. Final stage is when the symptoms of dementia are severely apparent and there is a drastic decline in the life quality of a patient.

Our objective is to develop and use an operational smart chair and a wrist device that records patients movements and pressure points during the clinical session when they take the screening test.

II. Related Works

Multiple studies are available on smart chairs that record bodily movements through sensors, on tests that are used as screening test to diagnose dementia, and on wearable devices that measure an individual’s cardiovascular conditions. However, the amalgamation of the three to diagnose dementia and its forms, to the best of our knowledge, has not yet been thoroughly explored. We are currently looking at the ground truth data (see the Data Analysis section below) available for patients that either have dementia or MCI and analyzing their measurements recorded using the smart chair and wearable device, and complementing their sensed data with the scores from the SLUMS test. First, we are going to explain the available work on why early diagnoses of dementia is helpful and desired. Second, we are going to compare the test available that help diagnose dementia and justify our choice of using SLUMS. Lastly, we are going to look at research done on using chairs to improve quality of life and describe how our initiative of using a smart chair for dementia patients can help in conducting reliable diagnoses.

A. Desirability of Early Diagnoses

Symptoms are not present when dementia is developing, due to which the diagnoses can be easily missed [2]. On early diagnoses, the patients, as well as the care takers of the patients, can gain sufficient amount of time to prepare for the necessities required for the treatment. During this asymptomatic stage, there may be precursor biological and neurological changes occurring, which might lead to early symptomatic stage [2], where the patient starts to suffer with symptoms like minor memory loss or mild cognitive impairment (MCI).

So, what does it mean to be diagnosed early? According to the Cambridge Dictionary, “early,” means “near the beginning of (a period of time), or before the usual, expected, or planned time.” While “near the beginning of time,”
suggests the diagnoses when there is no evidence of benefit or harm \cite{3}, “before the usual, expected, or expected,” suggests premature revelation, which indicates possibility of harm. Therefore, “early” is good enough if it allows early treatment, as well as the ability to plan future effectively such that access care and support that can be made available or possible by advanced decision making \cite{3}.

Our goal is to develop screening tools and diagnostic approaches that have high sensitivity and specificity: where sensitivity measures the correct identification of patients with dementia, while specificity measures the correct identification of people without dementia \cite{4}. In other words, our goals include conducting a session with a smart chair, wearable device, and the SLUMS score such that the percentage of people correctly identified and correctly reject is high. Our major goal is to be able to identify symptoms of dementia at comparatively earlier stages such that the severe consequences can be avoided or timely treated. However, currently, there are no treatments available for dementia that slow down or prevent its prevalence \cite{3}.

B. Why SLUMS?

There are several screening tests that are used to diagnose dementia, including Mini-Mental State Exam or Folstein Test \cite{5} and Montreal Cognitive Assessment \cite{6}. Tariq et al. \cite{7} conducted a comparison between SLUMS and Mini-Mental State Examination for detecting dementia and MCI. In their study, they revealed that although both SLUMS and MMSE are able to identify dementia with high sensitivity and specificity, MMSE doesn’t perform as well as SLUMS for MCI \cite{7}.

SLUMS examination is a screening method for kinds of dementia \cite{8}, which consists of 11 items that measure a patients’ memory, attention, orientation, and executive functioning such as planning, organizing, sequencing, and abstracting \cite{7,8}. Following from the comparison between MMSE and SLUMS from above, it has also been found out that SLUMS is a more difficult than MMSE. Total of 30 points is possible in both tests, however, SLUMS score is observed to be approximately five points lower than that of MMSE, which is why likely to be more sensitive to MCI \cite{8}.

C. What does a chair tell us?

Simple pressure sensors installed in a chair can yield results about users’ activity by recording information on different postures, hand movements, and head movements \cite{9}. The body, which is the main physical entity between the user and the chair, produces numerous amount of information about the physiological parameters, body motions, as well as cognitive processes \cite{10}. Arnrich et al. \cite{9} describes work-related stress as a potential cause of hindrance faced in body functions. Back, neck, shoulder, arm-wrist-hand problems are a few examples that can result from abnormal stress levels. Our smart chair has pressure sensors on the backrest. The force sensitive resistors on the backrest sense body movements of the patient and wirelessly collects and stores this data in the receiver for stable and fast outcomes.

Sedentary positions can reveal a great amount of information about our stress levels. Either at work or home the we sit or relax can if our stress level is normal or not. Breathing, sitting position, body movements, can reveal about the pressure points. Lewy body dementia (LBD), a kind of dementia, resembles Parkinson’s Disease and causes involuntary bodily movements in patients \cite{11}. Recording such movements and observing the delays in response to a SLUMS score can direct clinicians’ attention towards dementia, especially when there’s no prior diagnosis available.

III. Data Analysis

The variety of data recorded includes SLUMS score for each of the questions, pressure sensor data for all the 9 sensors, the data from the wearable device for 296 patients. Out of these, there is a ground-truth data available for 21 patients. Ground-truth data contains clinical diagnosis identifying patients with dementia, MCI, or normal (non-dementia and non-MCI) conditions.

Before conducting any analysis, a lot of data cleaning and organizing was necessary.

A. Cleaning

Given the volume of the data, the preliminary preparation to structure the data required:

1) manually assigning headers, which were initially included as repeated values in columns corresponding to their data value

2) importing new files, using Python, into Jupyter Notebook for analysis to make sense of the data

3) plotting to compare and find meaningful implications
B. Organization

The 21 patients for which we have the ground-truth data available, we separated the pressure sensor data, wearable device data, and the SLUMS score for each patient. Reading folders in Python for each patient, we stripped the columns with recorded time to convert it to date_time() to be able to conduct time-wise analysis. Some data files were separated using commas, while some were separated with whitespaces, making it difficult to read them altogether.

C. Analysis

Figure 1 shows the plot of the pressure sensors for a session of a single patient. As it can be observed, there are total 8 sensors, with a few of them producing bad values such that it is hard to see any meaningful activity. Deleting these sensors, in figure 2, the pattern can be observed recorded from the pressure sensors installed on the chair. Notice that these pressure sensors are installed on different positions, so the variations depict the pressure points of different movements. The slot between 15:46 and 15:50, there is some missing data (the straight lines), which tells patient might have possibly leaned away from the backrest, disallowing any values to get recorded for a brief period.

In figure 3, using the starting time, we manually added the time, converting it to seconds, to visualize how much time the patient took to answer each of the 12 questions. The first blue vertical line is the starting time, possibly of when the patient was done being set up with the devices. Black vertical lines show the starting time for each question, while red lines show the ending time. It can be observed from the figure that some questions, like questions 5, 8, and 12, took longer than other questions for this patient.

Note these figures are for a single patient only, who does not have dementia.

IV. Expected Results and Discussion

Plotting these for all the 21 patients, we expect to see a pattern such that patients with dementia share some common traits observed in the analysis. Similarly, the patients with MCI, and those without any signs of MCI, are expected
Fig. 3  Vertical lines on Fig. 2 representing starting and ending times for SLUMS questions to show some similar behavior as other patients with the similar diagnosis. On success, we would like to look at the Receiver Operator Curves for sensitivity vs. specificity and compare them to the ground truth data. The analysis and results of which will enable us to look at other patients, for which we don’t have the ground-truth data available, and yield results for those to identify if they have any signs of dementia.

Moreover, given the current structure of the data, we also plan to find ways to automate plotting and converting of start/end time of questions from SLUMS, to find any patterns and difference in the data according to each response between patients who have dementia and those who do not. We would also like to find out the which sensor records which plot to better understand the different curves in a single figure.

In conclusion, we have the ample amount of data to conduct analysis and investigate if we can identify signs of dementia in patients who have taken the screening tests using our smart chair and wearable device. So far, we have the preliminary analysis for the 21 patients for which we have the true diagnosis available. Plan is to conduct careful analysis to see if the pressure data shows any difference between the patients with dementia and the patients with not.

References
Spatial Attention Mechanism for Weakly Supervised Fire and Traffic Accident Scene Classification

Md Moniruzzaman  
Department of Computer Science  
Missouri University of Science and Technology  
Rolla, MO, USA  
mmqgr@mst.edu

Zhaozheng Yin  
Department of Computer Science  
Missouri University of Science and Technology  
Rolla, MO, USA  
yinz@mst.edu

Ruwen Qin  
Department of Engineering Management and System Engineering  
Missouri University of Science and Technology  
Rolla, MO, USA  
qunr@mst.edu

Abstract—During the past ten years, on average there were near 16.5 thousands of hazardous materials (hazmat) transport incidents per year resulting in $82 millions of damages. Prompt, accurate, objective assessment on hazmat incidents is important for the first-responders to take appropriate actions timely, which will reduce the damage of hazmat incidents and protect the safety of people and the environment. Therefore, one of the most important steps is to automatically detect transport incidents, such as fire and traffic accidents. In this paper, we introduce a simple and yet effective framework that integrates the convolutional feature maps of deep Convolutional Neural Network with a spatial attention mechanism for fire and traffic accident scene classification. Our spatial attention model learns to highlight the most discriminative convolutional features, which is related to the regions of interest in the input image. We train our network in a weakly supervised way. In other words, without the requirement of precise bounding box annotating the exact location of fire or traffic accidents in the image, our network can be learned from the only image-level label. In addition to the image-based traffic scene classification, the model is also applied on a set of collected videos for real-world applications. The proposed model, a simple end-to-end architecture, achieves promising performance on fire scene classification from images, and traffic accident scene classification from both images and videos.

I. INTRODUCTION

A substantial amount of hazardous materials (hazmat), such as flammable liquids and poisonous gases, need to be transported to locations of consumption or disposals. During the past ten years, on average there were near 16.5 thousands of hazmat transportation incidents per year resulting in $82 millions of damages [1]. When a transportation incident occurs (e.g., fire, traffic car accident), prompt and effective emergency response is critical to minimize the impact of the incident. For example, fire caused by hazmat accidents contains hazardous materials and has a dangerous influence on the environment, human health, and other valuable properties. Image-based fire detection (e.g., traffic surveillance cameras) are effective in large open areas. However, there are challenges to design an automatic image classification algorithm to tell if an image contains a fire or not. Fig.1 shows some samples of the fire image dataset collected by us, from which we can see that some non-fire images have the appearance similar to the fire images. In this paper, we utilize the deep Convolutional Neural Network to classify whether an image contains the fire or not.

In addition to the fire detection in images, we also explore the general traffic accident scene classification in images and videos, as traffic accidents can cause serious injuries, which also require rapid assistance to reduce the additional rescue minute. The traffic accident images and videos were acquired by both traffic surveillance cameras and cameras on vehicles. Some samples of the traffic accident video dataset collected by us are shown in Fig.2.

Deep Convolutional Neural Network: With the recent availability of powerful GPUs, effective optimization algorithms, and a large amount of human-annotated image data [13], Convolutional Neural Networks (CNN) [14]–[19] have achieved significant progress for the task of image classification. CNNs have the ability to learn meaningful feature representations from the large quantities of data for a wide range of tasks. In addition to image classification, CNNs pre-trained on ImageNet [13] contribute greatly in, object detection [23], [24], video classification [26], semantic segmentation [25], and many other tasks.

Spatial Attention: Despite the recent advance, image classification using deep CNN still has challenging research questions to address. Most of the state-of-the-art methods [14]–[19] employ CNN over the entire image region to compute the feature maps by convolution followed by standard pooling (average or max) operation or fully-connected layer for the classification, without highlighting the features extracted from the most relevant spatial regions. But, usually an object in an image does not occupy the entire spatial domain. Some of the pixels in the entire spatial domain are less or not relevant to the target class. Therefore, the motivated research question is: from the convolutional feature maps which features should get more importance to highlight the most discriminative regions?
of the input image? To address this challenge, we leverage the spatial attention mechanism on top of the convolutional feature maps to emphasize the most significant features. In other words, the spatial attention mechanism learns to focus on the most relevant parts of the input image.

**Weakly-Supervised Learning:** Most of the modern deep learning algorithms [20]–[22] are fully supervised, which rely on human-labeled annotations, such as the precise bounding box and the segmentation mask for training. But, in practice, collecting such accurate annotations are expensive and time-consuming. Building a training dataset with only image-level annotation is much easier than the bounding box or segmentation mask annotations. Therefore, the motivated research question arises: given a weakly labeled image dataset (i.e., each image in the training set has a label but which portion of the image contains the target class is unknown), how can we effectively train a deep learning algorithm? To address this challenge, in this paper, we use weakly-supervised learning that reduce the amount of human level intervention by using the image or video dataset that are partially labeled (e.g., “fire”, “accident”, etc.). In other words, without ever providing the network with information about the location of the target class, we train our network by utilizing only image-level labels.

**Our Work:** In this paper, first, we utilize deep Convolutional Neural Networks with a spatial attention mechanism for fire classification. We introduce an end-to-end spatial attention model for weakly supervised fire classification from images using pre-trained CNN networks. Our method starts by adopting existing VGG [15] networks (e.g., VGG-16, VGG-19) pre-trained on ImageNet [13] data with only image-level supervision (no bounding box or segmentation mask annotating the precise region of interest in the image) for feature extraction. The extracted features from CNNs are passed through our spatial attention model to get the attentionally-pooled feature representation, which is then processed by a classification layer for the final image-level classification. In addition to the classification, our approach can also locate the fire regions in the image by our spatial attention model.

Second, we generalize our learned spatial attention model from the fire classification dataset to traffic accident classification. For this purpose, we simply transfer the learned attention weights of our spatial attention model to traffic accident image and video datasets.

**II. RELATED WORKS**

**Fire Detection and Classification:** There are several works for vision-based fire detection and classification [2]–[4]. Healey et al. [2] used a purely color-based model for automatic fire detection. Spectral, spatial and temporal models of fire regions were developed for vision-based fire detection [3]. Temporal wavelet features in addition to ordinary motion and color cues were utilized to detect fire and flame [4]. There are a limited number of machine learning based fire classification and detection approaches [5], [6]. Ko et al. [5] introduced a vision sensor-based fire detection method, which used wavelet coefficients as the input to the support vector machines (SVM) classifier with a radial basis kernel (RBF) for the fire-pixel verification. Zhang et al. [6] used temporal shape features as an input to the artificial neural networks for real-time forest fire detection. Our approach differs from them as we use deep convolutional features with an attention model for fire classification and localization.

**Traffic Accident Classification:** There are not many literature studies on traffic accident data for classification based on deep learning methods. Ess et al. [7] presented a segmentation based method to recognize traffic scenes in front of moving vehicles with respect to the road topology and the existence of commonly encountered objects. Geiger et al. [8] proposed a probabilistic generative model for multi-object traffic scene understanding. Gupta et al. [9] presented a computer vision based algorithm for detecting and classifying vehicles in monocular image sequences of the traffic scene. Lan et al. [10] used the histogram of oriented gradient (HOG) and support vector machine (SVM) for real-time automatic obstacle detection in urban traffic. Shiau et al. [11] developed a forecasting model based on data mining technology for road traffic accident classification. Agarwal et al. [12] presented a
Fig. 2. Frames of sample videos collected in our traffic accident video dataset.

Fig. 3. The architecture of our approach. (a) Feature extraction, (b) Spatial attention network, and (c) Classification module.

A hybrid model based on logistic regression with a wavelet-based feature extraction process for traffic incident detection.

III. APPROACH

The workflow of our network is illustrated in Fig.3. First, we obtain a pre-trained CNN network and extract the last convolutional feature maps by passing the image through pre-trained CNN network. Second, we apply the spatial attention model on top of the last convolutional feature maps to get the attentionally-pooled feature vector. Third, we pass the attentionally-pooled feature vector through a classification module to get the final classification scores on the fire and the general traffic accident classification tasks.

A. Feature Extraction

An important component in our approach is feature extraction. In our approach, we choose two pre-trained networks, namely VGG-16 and VGG-19. These two pre-trained networks are almost similar, except more convolution and max-pooling layers in VGG-19 network. We use the pre-trained 2D CNN models (VGG-16, VGG-19) trained on ImageNet dataset [13] with only image-level supervision to extract the last convolutional feature maps. The feature maps after the last convolutional layer preserve the spatial information of the input image, denoted as \( X \in \mathbb{R}^{k_1 \times k_2 \times f} \), where \((k_1 \times k_2)\) denotes the spatial dimension of the feature maps and \(f\) is the number of feature map, as shown in Fig.3(a). These feature maps are utilized to describe the visual content of the input image and passed to the next layers for recognition.

B. Spatial Attention Mechanism

The feature maps \( X \in \mathbb{R}^{k_1 \times k_2 \times f} \) are converted to 2D matrix \( Y \in \mathbb{R}^{k \times f} \), where \( k = (k_1 \times k_2) \), as shown in Fig.3(b). Each row of matrix \( Y \) maps to different overlapping regions in the input space. Our spatial attention model learns to focus its attention on these \( k \) regions. Formally, our spatial attention model learns an attention weight vector \( a \in \mathbb{R}^{f \times 1} \) and computes attention score vector \( y \), which indicates the feature importance from \( k \) regions:

\[
\mathbf{y} = \mathbf{Y} \mathbf{a}, \quad \text{where} \quad \mathbf{y} \in \mathbb{R}^{k \times 1}
\]  

The attention score vector \( \mathbf{y} \) is passed through a softmax layer to get the normalized attention scores-

\[
y_{\text{softmax}}^{(i)} = \frac{\exp(y^{(i)})}{\sum_{j=1}^{k} \exp(y^{(j)})} \quad \text{where} \quad i = 1, \ldots, k
\]

where, \( y^{(i)} \) denotes the \( i \)th dimension of \( \mathbf{y} \) and \( y_{\text{softmax}} = [0, 1]^k \) denotes the normalized attention scores. After that, the attentionally-pooled feature vector \( \mathbf{v} \), which is the feature representation for the classification module, is computed by

\[
\mathbf{v} = (\mathbf{Y}^T \mathbf{y}_{\text{softmax}}) \in \mathbb{R}^{f \times 1}
\]
C. Classification Module

Now, we aim to classify the image into the predefined class categories based on the attentionally-pooled feature vector \( v \), as shown in Fig.3(c). We learn linear mapping \( W \in \mathbb{R}^{C \times f} \) (\( C \) is the number of classes) and compute the \( C \)-dimensional score vector \( s \) from attentionally-pooled feature vector:

\[
s = Wv, \quad \text{where} \quad s \in \mathbb{R}^{C \times 1}
\]  
(4)

Finally, the score vector \( s \) is passed through the softmax layer to get the normalized classification scores:

\[
s_{\text{softmax}}^{(i)} = \frac{\exp(s^{(i)})}{\sum_{j=1}^{C} \exp(s^{(j)})} \quad \text{where} \quad i = 1, ..., C
\]  
(5)

where, \( s^{(i)} \) denotes the \( i \)th dimension of \( s \) and \( s_{\text{softmax}} = [0, 1]^C \) denotes the normalized classification scores.

D. Attention Model for Video Recognition

Classifying videos instead of images adds a temporal dimension in addition to the visual appearance in individual frames. Therefore, in addition to the spatial attention model, we use Long-term Short Memory (LSTM) [28] for video recognition, which is able to address variant-length input and capture the long term temporal dynamics.

Since a video contains a sequence of frames, we extract the last convolutional feature maps obtained by pushing the video frames through pre-trained network. Formally, given a video with the duration of \( T \) frames, at each time step \( t \), we extract the last convolutional feature maps \( X_t \in \mathbb{R}^{h_1 \times k_2 \times f} \), which are passed through the spatial attention mechanism to get the attentionally pooled feature vector \( v_t \in \mathbb{R}^{f \times 1} \). The outputs of the spatial attention mechanism \( v_t \) are passed through the recurrent sequence learning module (Long Short-Term Memory (LSTM)). The weight parameters of LSTM maps the input \( v_t \) and previous time step hidden state output to an output feature vector \( o_t \in \mathbb{R}^{f \times 1} \), which is the feature representation for the classification module. The outputs of LSTM at each time step, are then fed into the classification module, which produces classification scores \( s_{\text{softmax}}^{(i)} \) for each frame. Finally, the classification scores of each frame of a video are averaged to get the final video-level label prediction. The overall process for video recognition is shown in Fig.4. It should be noted that we pass the outputs of LSTM to the classification module for video recognition, instead of directly passing the outputs of the spatial attention model to the classification module for image recognition.

IV. EXPERIMENTAL RESULTS

A. Implementation Details

We use Keras (Tensorflow backend) python API to implement our network architecture. The input image is resized to \( 224 \times 224 \times 3 \) pixels, which is passed through the pre-trained 2D CNN model (VGG-16, VGG-19). We extract the last convolutional layer of the VGG-16 or VGG-19 network, which produces \( 14 \times 14 \times 512 \) feature maps. The feature maps are fed into our spatial attention module, which is a trainable layer. Our spatial attention model learns attention weights \( (a \in \mathbb{R}^{f \times 1}) \) to get attentionally-pooled feature vector \( (v \in \mathbb{R}^{f \times 1}) \), which is the feature representation for the classification module. The classification module learns a linear mapping \( W \in \mathbb{R}^{C \times f} \) to transform the feature representation \( v \) into a \( C \)-dimensional feature vector.
In this paper, we introduce a new weakly-supervised framework for fire classification from images, and accident scene classification from both images and videos. We use pre-trained deep CNN features and employ a spatial attention mechanism to address the challenge of highlighting the most discriminative features for fire classification. To see the effectiveness of our approach, we also transfer the learned weights of our spatial attention model to a generalized traffic accident dataset for classification. We performed extensive experimental evaluation and showed that our model performs better than other existing support vector machine (SVM) and artificial neural network (ANN) based methods. To test the robustness of our approach, we use VGG-16 and VGG-19 pre-trained networks for our experiments. As shown in Table I and Table II, we get consistent performance for both VGG-16 and VGG-19 networks. As VGG-19 network has a few additional convolutional and max-pooling layers compared to VGG-16 network, we get slightly better performance on average for VGG-19 network.

**Image-Based Accident Scene Classification:** We transfer the spatial attention model that we learned from the fire dataset to the Traffic Accident Image dataset. Table III shows the comparison results of the transfer learning approach with the baseline approach for the extracted features from pre-trained VGG-19 network (here, we only use pretrained VGG-19 network, as we get better performance for VGG-19 features compared to VGG-16 features on the fire dataset) on the Traffic Accident Image dataset. For the baseline approach, we configure the network without attention pipeline. For this purpose, the extracted feature maps \( X \in \mathbb{R}^{k_1 \times k_2 \times f} \) are average pooled to get \( f \)-dimensional feature vector and passed through the classification module for the classification. As shown in Table III, our approach (achieves 93.72% average accuracy) outperforms the baseline approach (achieves 88.95% average accuracy) on Traffic Accident Image dataset, which means our spatial attention model learns generalized features that can be effectively used in traffic accident and hazardous materials incident classification.

**Video-Based Accident Scene Classification:** We evaluate the performance of our video recognition framework (spatial attention model + LSTM) on Traffic Accident Video dataset for real-time traffic accident scene classification. As shown in Table IV, we performed the ablation studies on our framework by comparing three configurations on Traffic Accident Video dataset. Over all the three configurations, the combination of spatial attention model and LSTM achieves the best performance.

**V. Conclusions**

In this paper, we introduce a new weakly-supervised framework for fire classification from images, and accident scene classification from both images and videos. We use pre-trained deep CNN features and employ a spatial attention mechanism to address the challenge of highlighting the most discriminative features for fire classification. To see the effectiveness of our approach, we also transfer the learned weights of our spatial attention model to a generalized traffic accident dataset for classification. We performed extensive experimental evaluation and showed that our model performs better than other existing support vector machine (SVM) and artificial neural network (ANN) based methods.
TABLE IV
ABSTRACTION STUDY OF DIFFERENT ARCHITECTURES ON TRAFFIC ACCIDENT VIDEO DATASET FOR VGG-19 FEATURES.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Baseline approach</th>
<th>Spatial attention model (Ours)</th>
<th>Spatial attention model + LSTM (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial-1</td>
<td>78.84</td>
<td>81.39</td>
<td>82.69</td>
</tr>
<tr>
<td>Trial-2</td>
<td>79.16</td>
<td>84.67</td>
<td>86.53</td>
</tr>
<tr>
<td>Trial-3</td>
<td>77.44</td>
<td>80.10</td>
<td>82.69</td>
</tr>
<tr>
<td>Trial-4</td>
<td>77.81</td>
<td>81.18</td>
<td>84.61</td>
</tr>
<tr>
<td>Average</td>
<td>78.31</td>
<td>81.84</td>
<td>84.13</td>
</tr>
</tbody>
</table>

Fig. 5. Visualization of our spatial attention map. Our spatial attention model learns to locate the fire and accident regions in the image.

better than the baseline approach, which did not use any attention pipeline. The proposed framework is also efficient and easy to implement.

REFERENCES

Polarization dependent dual band metasurface thermal emitter has been developed. This type of technology is a key component for low cost, compact mid-infrared gas sensing system. In this work, we propose a metasurface emitter design that can sense the vibration of C-O bond and C-H bond respectively in CO$_2$ and CH$_4$ gas. The metasurface is based on an Au/Al$_2$O$_3$/Au structure. The emitter exhibits an emissivity of 0.90 and 0.84 at 3.95 μm and 3.32 μm respectively. The metasurface is combined with an electrical source to produce the infrared waves by applying a voltage. Later on, a gas sensing system is proposed where the metasurface thermal emitter can be incorporated into a CO$_2$-CH$_4$ sensing system. It can detect both gases at the same time. It will pave the way for a new level of highly integrated, low-cost optical gas sensors.

Keywords: metamaterials, metasurface, mid-infrared sensing, thermal emission, optical sensing.
microelectromechanical system (MEMS) heaters with metamaterial perfect emitter structures. Exhibiting a resonance quality factor of 15.7 at the center wavelength of 3.96 μm and an emissivity of 0.99, the demonstrated emitter is a spectrally narrow and efficient light source. Owing to its spectrally tailored, nondispersive emission, additional filter elements in a free-space optical gas sensing setup become obsolete. Their light source was tunable in emission wavelength through the mid-infrared wavelength band. Researches are going on to find an energy efficient on-chip thermal light source, which is thermally stable up to typical application temperatures of 800 K and can sense two or more gases particularly at the same time.

In this work, we propose a metasurface thermal emitter designs that can sense the vibration of C-O bond (90% emissivity) and C-H bond (84% emissivity) respectively in CO₂ and CH₄ gas. The metasurface is integrated with a voltage source to produce the infrared rays by applying voltage. The emitter exhibits an emissivity of 0.98 and 0.96 at 3.95 μm and 3.32 μm respectively. Finally, a setup schematic including the emitter as a light source for NDIR gas measurements has been proposed.

Metasurface Thermal Emitter

The structure of the metasurface thermal emitter as a mid-IR light source is illustrated in Figure 1. The metasurface is connected to a function generator via BNC cable. The function generator will be used as the voltage source. As it can generate the rectangular waves, it reduces the necessity of the chopper. Two BNC cables

![Concept of Metasurfaces Thermal Emitter (MTE) Light Source](image)
will be connected at the positive and the negative potential terminal of the metasurface thermal emitter (MTE) from the function generator. The function generator produces rectangular waves so that it will be easier for the lock-in amplifier to differentiate the signals for the pyroelectric detectors.

**Unit Cell Definition and Emission Spectra**

The profiles of the MTE unit cells are shown in Figure 2. In Figure 2, the unit cell comprises a cross-shaped hollow design (Au), which is separated from a metal backplane (Au) by means of a dielectric spacer layer (Al₂O₃). The macroscopic response of the metamaterial to electromagnetic waves can be homogenized by applying effective medium theory, and the MTE will be modeled as a quasi-homogeneous slab with effective material parameters εeff and μeff. The top resonator results in a resonant effective permittivity εeff. Hence, the unit cell geometry defines the effective material parameters of the metamaterial. These will be used to tailor its emission properties.

![FIG. 2. MPE Unit Cell for CO₂ Sensing](image)

Figure 2 shows the unit cell for CH₄-CO₂ sensing. It is composed of a 65 nm thick gold (Au) layer patterned with a cross-shape hollow design, a 250 nm thick Alumina (Al₂O₃) spacer and a 200 nm thick gold ground plane. The period of the structure is 1500 nm and 1520 nm along x-axis and y-axis respectively. The other parameters are shown in the figure 2. The clearance distance between the boundary and Lₓ and Lᵧ are 50 nm and 220 nm respectively. The emission spectra is shown in figure 3 for CH₄ and CO₂ sensing. There is an emissivity of 0.90 at 3.95 µm and 0.8 at 3.32 µm where the C-O and C-H bonds vibrate.

The emission spectra is shown in Figure 4(b) for both CO₂ and CH₄ sensing. There is an emissivity of 0.90 at 3.95 μm and 0.8 at 3.32 μm where the C-O and C-H bonds vibrate.

All the simulations were performed in CST Microwave Studio Suite in the RF module platform to determine the emissivity at 3.95 μm and 3.32 μm.

**Voltage and Temperature Distribution**

Figure 4 shows the voltage distribution of a microstructure. In this structure, our metasurface thermal emitter is patterned with a trench. It will show the voltage distribution along the whole structure. It will also provide information how much voltage is needed to get the emission spectra from the metasurface. The approximate voltage is needed to heat up is more than 1 V.

![FIG. 4. COMSOL Result for Voltage Distribution](image)

Figure 5 shows the temperature distribution of the microstructure. From this figure, we can determine the maximum surface temperature that can be achieved by applying the voltage. Also, it will help us to determine the
FIG. 5. COMSOL Result for Temperature Distribution

positions on the metasurface from where the heat is mostly generated. The heat is mostly generated just after the edges of the negative cross structure. So, the more the metal at those positions, the more heat will be generated.

All the simulations were performed in COMSOL Multiphysics in the Joule Heating platform to determine the amount of voltage to generate maximum possible temperature. The maximum temperature should be lower than the melting point of gold and alumina. In our simulation, we tried to maintain the maximum possible temperature around 700 K.

**Gas Sensing Setup Schematic**

Finally, a setup schematic which can be applicable in real life gas sensing has been proposed. There are two key features of the metamaterial light source, on-chip integration and high spectral efficiency, are exploited for demonstration of NDIR gas sensing at the example of CO₂-CH₄. A schematic representation of the experimental setup is provided in Figure 6.

The setup consists of a thermal light source (our MTE or a blackbody emitter as a conventional light source) and a dual-channel pyroelectric detector.

Both the emitter and the detector are connected to a gas mixing system, allowing for variable CO₂, CH₄ and humidity levels. On the downstream side of the measurement container, reference sensors for total mass flow (MF), temperature (T), and relative humidity (RH) will be connected. The source is driven by a periodic, rectangular voltage signal which makes it enable for lock-in amplification of the preamplified pyroelectric signal. The temperature of the emitter surfaces on the membranes at each voltage will be determined by comparing their emission spectra with those of equivalent emitters mounted on a temperature-controlled substrate. At the maximum voltage of 1V, the temperature reached 700 K. This is how we will determine the amount of voltage and frequency of the power source. In order to get the same emission spectra from a blackbody we will need to raise the voltage which will cost more energy. So we can see that, the metasurface thermal emitter will require less energy than the conventional blackbody emitter. This property makes it more energy efficient than the conventional blackbody emitter. The maximum emittance is 0.90 and 0.84 for CO₂ and CH₄ respectively which are nearly that of a perfect blackbody. A Fourier transform infrared (FTIR) spectrometer will be used for the measurement. The wavelength of 3.95 μm is absorbed by CO₂ molecules due to the asymmetric stretching vibration. Similarly, 3.32 μm is absorbed by CH₄ molecules due to the same reason. So by measuring emissivity at these two particular wavelength we can ensure the existence of these two gases.

**Conclusion**

In summary, we propose a metasurface thermal emitter (MTE) structure as an efficient, mid- infrared light source for molecule sensing. The MTE shows an emissivity of 0.90 at the wavelength of 3.95 μm and 0.84 at the wavelength of 3.32 μm. It will render the light source attractive for highly integrated, free-space optical gas sensing applications. Employing the MTE as a light source in a gas sensing system of real life, we can measure the concentration of CH₄ and CO₂ at the same time. The proposed light source may open the way for a new generation of highly integrated, low-cost NDIR gas sensors. Furthermore, by changing the parameters of the design we will be able to shift the emissivity at different wavelength to sense different gas species like CO, N₂O etc.
References


[31] Lochbaum, Alexander; Fedoryshyn, Yuriy; Dorodnnyy, Alexander; Koch, Ueli; “On-chip narrowband thermal emitter for mid-IR optical gas sensing,” ACS Photonics, 2017.
Development of a temperature sensing system based on a longwave infrared (LWIR) thermal camera for smart aerospace additive manufacturing

Muhammad Roman$^1$, Jie Huang$^1$, Zhaozheng Yin$^2$, Edward C. Kinzel$^3$, and Ming C. Leu$^3$

$^1$Department of Electrical and Computer Engineering, Missouri S&T
(mrhmcmst.edu, jieh@mst.edu)
$^2$Department of Computer Science, Missouri S&T
$^3$Department of Mechanical and Aerospace Engineering, Missouri S&T

Abstract
The temperature history across the powder bed during the selective laser melting (SLM) process is very useful in determining the quality of the build. Dimensional inaccuracies can be related with the temperature distribution on the powder bed during different stages of additive manufacturing. The aim of this project is to develop an optical temperature measurement system using a high resolution longwave infrared (LWIR) thermal camera. This LWIR camera, installed on the Renishaw AM250 machine at Missouri S&T, will measure 2-D temperature maps of the build at different steps of the SLM process. A 3-D temperature map will be created out of these 2-D maps and the data analysis will allow to relate temperature distribution with the manufacturing flaws.

**Keywords**: selective laser melting, thermal imagining, longwave infrared, machine learning

1. Introduction

Selective laser melting is a 3-D printing technique which employs a high power density laser to melt and fuse metallic powder to create three dimensional parts [1]. In order to achieve high quality parts, uniform thickness of layer and proper temperature distribution across the bed is required. Any non-uniformity in the layer and/or heat buildup in some portion of the bed results in dimensional inaccuracies in the build [2]. In this work, a thermal modeling technique is developed to quantify the properties of the powder bed. Temperature of the bed at different steps of the selective laser melting is measured and a 3D temperature map of the build is created. This 3D map is correlated with the part performance.

Currently, a laser melting machine, Renishaw AM250, is used in MST for experiments involving additive manufacturing. In this system, a shortwave infrared (SWIR) camera is used to measure high temperature surrounding the melt pool. This SWIR gets images of the melt pool through a
quartz window. The SWIR camera will be relocated in the system and a low resolution (320x240) (for now) LWIR thermal camera (FLIR-T240) will be installed as shown in Fig.1. The camera will get images from outside the chamber, so it will require a window to see the SLM process. As the quartz window is not transparent in longwave infrared range, a ZnSe window will replace quartz window for the LWIR camera. Thermal images will be collected through FLIR research IR, the data acquisition software of the camera.

![Figure 1: Integration of thermal imaging camera into SLM system](image)

These thermal images will be collected at different stages of the SLM process. These stages include: temperature distribution around the laser spot during laser heating, temperature distribution of the build when the next powder layer deposition is about to begin, temperature distribution of the bed immediately after the deposition. The 2D thermal images of all the stages will be fed to a data processing algorithm in MATLAB, and a 3D thermal map will be constructed. Analysis of 3D thermal map will allow to quantify the properties of SLM process. Correlating thermal map with the non-uniformity will help in determining the cause which could be any of the following: change in the layer height due to wiper misalignment or damage, non-uniform temperature of the bed due to heater failure or heat buildup, and cracks due to residual stress. There are also other factors which might influence the quality of the build [3][4]. Thermal modeling will not only help in determining the cause but may also avoid it by raising an early alarm in real time monitoring.

2. Experiments

Some preliminary experiments were performed to test the idea of thermal imaging system in additive manufacturing to ensure the quality of manufactured parts. Using a handheld LWIR thermal camera (FLIR T-420), images of the powder bed were captured at different stages of the SLM as shown in Fig.2. Due to the non-transparent nature of glass at longwave infrared, images were captured through the open door of AM250 machine. The FLIR Research IR was used to acquire data from the camera. Both the images, IR and visible, were captured using the same FLIR T-420 camera. The purpose of this experiment was to see the temperature distribution at various stages of the SLM process and relate temperature features with the properties of powder
bed. Moreover, to see the effect of machine’s malfunctioning due to a damaged part, powder wiper was damaged before coating one of the layers, and temperature distribution was monitored as a result of wiper damage.

![Figure 2: Thermal images of powder bed (top row), visible images (bottom row)](image)

Fig.2. shows temperature features at different stages of the build during selective laser melting. These features convey meaningful information about the SLM process. Fig.2a is the 2D thermal map of the powder bed after the build was completed and chamber was cooled down. It is evident from the image that loose powder was not as effectively heated as the solid parts. In this experiment, wiper was intentionally damaged before coating the layer. The effect of wiper damage can be seen in the image. The Fig.2b was captured immediately after coating a new layer of the powder. A cold powder bed can be seen over a completed first part. The 3rd image was taken after 10 s delay. This image can be used to quantify the diffusion rate of heat, and the information about the powder composition and layer thickness can be extracted from the diffusion rate.

This analysis shows the potential of LWIR thermal imaging in additive manufacturing.

### 3. Conclusion

Thermal imaging of the build during the selective laser melting is a useful technique to characterize the performance of the build. A longwave IR camera was used to capture thermal images at different stages of the build. Data from the LWIR camera was acquired using FLIR Research IR, a data acquisition software, and was stored and processed in a PC. Analysis of thermal images revealed meaningful information about the powder bed and SLM machine.

**Future Work**
The quartz window will be replaced with the ZnSe window for the longwave IR thermal camera. A low resolution thermal camera (FLIR T-240) will be integrated into the SLM machine, Renishaw AM250, to capture images during additive manufacturing. The FLIR T-420 will be calibrated through the ZnSe window using black bodies created by SLM machine. Images from the data acquisition software, FLIR Research IR, will be fed to a machine learning algorithm for the analysis.

Later, a high resolution (640x480 with 30 fps) LWIR thermal camera will replace the existing low resolution (320x240) camera.

References


ABSTRACT

Model transformation tools assist system designers by reducing the labor-intensive task of creating and updating models of various aspects of systems, ensuring that modeling assumptions remain consistent across every model of a system, and identifying constraints on system design imposed by these modeling assumptions. We have proposed a model transformation approach based on abstract interpretation, a static program analysis technique. In this paper, we present the foundations of our proposed approach to model transformation. We define model transformation in terms of abstract interpretation and prove the soundness of our approach. This work provides a methodology for relating models of different aspects of a system and for applying modeling techniques from one system domain, such as smart power grids, to other domains, such as water distribution networks.

Index Terms—Modeling, Model transformation, Formal methods, Abstract interpretation

1 INTRODUCTION

The multitude of functional and non-functional requirements for critical infrastructure cyber–physical systems (CPSs) present many challenges to system designers. A smart grid must be able to supply all its customers; it must be fault-tolerant in the face of component failure; it must be secure against physical and cyber attacks; and it must achieve all these goals with efficient infrastructure. To meet all these requirements, designers must integrate physical components, cyber control software and hardware, and processes for human operators into a complete system. This is a truly daunting task, but one that can be facilitated by model-based design and evaluation.

A vast body of literature has been published on various modeling formalisms that capture system performance, dependability, safety, and security. No single modeling formalism can encompass all aspects of system performance and dependability, necessitating the labor-intensive and error-prone process of creating multiple system models and propagating changes across all of these models. Furthermore, designers must be careful that these models remain consistent with each other, i.e., that the assumptions made about the system by one model are not contradicted by those of any other model. For instance, a dependability model for a smart grid where two power lines are assumed to be connected in parallel is not compatible with a power flow analysis where the lines are placed in series.

One way to alleviate these challenges is through model transformation, which enables automated or semi-automated transformations between modeling formalisms. These transformations can ensure that modeling assumptions are consistent across every model of a system by verifying that any model can be transformed into any other. This approach can also identify constraints on system design imposed by these assumptions. Such a model transformation approach should meet two design constraints. First, it should be applicable to a broad range of systems and a variety of modeling formalisms in order to be useful to designers of complex systems. Second, it should be sound—it should be possible to prove that the result of a transformation is correct and consistent with the initial models.

In our earlier work [1], we proposed a model transformation approach based on abstract interpretation, a static program analysis technique. Models are seen as abstractions of the semantics of a system—its structure and behavior. Through this lens, provably correct model transformation becomes the problem of defining sound mappings from system semantics to model semantics and vice versa. By composing these mappings, we can develop sound transformations between modeling formalisms.

This work proposes a formalization of system and model semantics, leading to a formalization of sound model transformation. Our research contribution formalizes the research approach we outlined in [1] and incorporates several improvements from feedback we have received since publication of that work.

The structure of this paper is as follows: in Section 2, we briefly summarize related model transformation and formalization techniques. Section 3 presents our formalization of system and model semantics and describes how we use this formalization to create a method for model transformation. Section 4 summarizes our work and discusses future directions for our research.
2 RELATED WORK

In the literature, model transformation refers to two different but related concerns. One concern is integrating models of different parts of the system into a complete system model; this is more specifically called heterogeneous model composition. The other concern is transforming one type of model for a system to a different model of the same system or a related system.

Model transformation research specific to CPSs primarily focuses on building hierarchical models [2]–[5]. Hierarchical models allow different model types to be combined together to model complex systems. Each component can be modeled in a convenient formalism; the hierarchical model is then simulated by simulating each sub-model in tandem.

The Ptolemy modeling software [6] performs hierarchical modeling and model composition [7], [8]. As such, Ptolemy makes it easy to build and link small models. Hierarchical models can consist of heterogeneous sub–models, allowing different parts of the system to be expressed using different types of models [8]–[11]. Model composition is achieved in part by defining ontologies of system properties, e.g., units of model inputs and outputs. Based on these ontologies, Ptolemy can perform conversions of values transmitted between sub–models and check for incompatibilities which indicate modeling errors. Ptolemy also enables heterogeneous model evaluation: it provides choices for both the modeling language and the solution or simulation technique used to evaluate the model [12]. However, Ptolemy does not offer methods for transforming one system–level model to another. In addition, it is focused on models of system function and lacks facilities for modeling non-functional attributes.

OsMoSys [13] and SIMTHESys [14] are modeling systems motivated by model–driven engineering. Their approach to model transformation is based on techniques from software engineering. Graph–based models, such as Petri Nets and Fault Trees, are described using an object–oriented notation. Every model has associated interfaces which allow models of different types to be composed and evaluated. OsMoSys features compositional models and interfaces with external tools to evaluate them [15]. SIMTHESys provides a language in which users can describe new modeling formalisms for use with OsMoSys. Both are capable of modeling both functional and non-functional aspects of a system [16]. However, neither are focused on the problem of model transformation.

Möbius [17] is another modeling tool that supports hierarchical modeling. It supports several modeling formalisms, including block diagrams and Petri nets, and additional formalisms including stochastic timed systems can be included via external modeling tools [18], [19]. While this feature offers considerable flexibility in modeling, Möbius is constructed around a modeling workflow that builds and evaluates hierarchical models and has little support for model transformation. Its model composition method is based on object–oriented design principles and is applicable to many state–based model formalisms.

AToM³ [20] is capable of both model transformation and model composition. It uses metamodels to describe specific modeling languages, then defines transformations between metamodels to transform models [21]. Models are graph-based and transformations take the form of graph rewriting rules [22]. However, there is no hierarchy of models, so introducing a new model requires writing transformation rules from the new model to each model that AToM³ implements.

CHESS [23] provides a modeling language for describing systems and includes several model transformation methods specific to creating dependability models. CHESS is based on the Unified Modeling Language (UML); transformations are based on graph rewriting rules. CONCERTO [24] extends CHESS by introducing modeling techniques for non-functional system attributes such as dependability [25]. However, CONCERTO is focused on multi–formalism modeling [29]. It takes an algebraic approach to relating models: each formalism is described as a coalgebra—a mathematical system useful for describing arbitrary transitions among arbitrary states [30], [31]. The coalgebras corresponding to each formalism are placed in a lattice, which provides a structure for determining how to transform one model into another. Model transformations can be used to relate different models of the same system; for example, it is possible to combine a functional system model with a model of that system’s power consumption. However, Rosetta lacks many features required for CPS modeling, especially support for hybrid discrete–continuous formalisms.

Each of these model transformation tools offers a partial solution to the model transformation problem; however, none of them present a solution that is generally applicable. Some frameworks place constraints on the behavior of transformation functions (e.g., class inheritance transformation). Others apply only to specific formalisms. Furthermore, only Rosetta offers an approach that can be proven to be correct. The work presented in this paper aims to address these shortcomings by providing a model transformation approach that relates a wide variety of modeling formalisms in a provably sound fashion, and yields results that are sufficiently specific to be meaningful.

3 ABSTRACT INTERPRETATION OF MODELS

The foundation of our approach is abstract interpretation [32], [33], a formalism for developing sound semantic abstractions. In this work, system semantics are represented in terms of properties that hold for the system. Such properties might include information about components, their reliabilities, and how they are interconnected. Models are abstractions of system semantics—they concern certain properties of the system, but not others. Thus, generating a model from a system’s properties, then deriving properties of that system from the generated model, may result in some of the initial properties not being present in the derived properties. This is a necessary effect of
abstraction—we cannot derive properties from a model if the model does not capture those properties. To define mappings from system properties to models and vice versa, both domains need to allow for this potential loss of precision.

3.1 Properties

We first define how system semantics are represented. Lattices (see [34]) offer a useful formalism for describing the nature of approximation. We define a complete \( \mathcal{P}(\text{Properties}) \) lattice ordered by specificity: for \( p_1, p_2 \in \mathcal{P}(\text{Properties}), p_1 \sqsubseteq p_2 \) means that the constraints in \( p_1 \) and \( p_2 \) are not contradictory and that \( p_1 \) places the same or more constraints on a system than \( p_2 \) does. For example, \( p_2 \) could constrain the reliability of a component to fall in the range \((0, 1]\), whereas \( p_1 \) could require that component to have a reliability of 0.95.

The meet (denoted as \( \cap \)) of two elements of \( \mathcal{P}(\text{Properties}) \) places the constraints of both elements on a system; the join (denoted as \( \sqcup \)) implies satisfaction of the constraints of either element. Suppose \( p_1 \) requires a component’s reliability to fall in \([0.8, 1.0]\) and \( p_2 \) constrains it within \([0.75, 0.9]\). Then \( p_1 \cap p_2 \) will require it to be in \([0.8, 0.9]\) and \( p_1 \sqcup p_2 \) within \([0.75, 1.0]\). \( \cap \) and \( \sqcup \) extend this concept to subsets of \( \mathcal{P}(\text{Properties}) \).

For certain \( p_1, p_2 \in \mathcal{P}(\text{Properties}) \) are contradictory, \( p_1 \cap p_2 \) will result in a constraint that is impossible to satisfy. If \( p_1 \) requires a component to have a reliability in \([0.5, 0.7]\) and \( p_2 \) requires it in \([0.9, 1]\), then it is impossible for any component to meet both constraints. In this paper, we require that every element of \( \mathcal{P}(\text{Properties}) \) to be satisfiable except for \( \bot \), the “impossible” constraint. Therefore, for this example, \( p_1 \cap p_2 = \bot \). Note that \( \forall p \in \mathcal{P}(\text{Properties}). \bot \sqsubseteq p \).

To summarize, each element of the \( \mathcal{P}(\text{Properties}) \) lattice describes one or more systems. In the general case, \( p \) describes a set of systems, all of which meet the constraints in \( p \). If every constraint in \( p \in \mathcal{P}(\text{Properties}) \) has exactly one possible choice, \( p \) will describe a single system.

3.2 Models

We now consider how modeling formalisms can be represented in this lattice framework. As a given element of the \( \mathcal{P}(\text{Properties}) \) lattice may not define a single system, we must account for the possibility that the lattice may not specify the system well enough for a single model to be abstracted from it. If \( p \in \mathcal{P}(\text{Properties}) \) does not constrain the reliability of a component to a single value, a single reliability model cannot be abstracted from \( p \). Instead, we abstract a set of models, one for each possible assignment of the component’s reliability, subject to the constraints of \( p \).

Therefore, in the same way that the \( \mathcal{P}(\text{Properties}) \) lattice is defined, we also define the domain of each modeling formalism to account for the nature of potentially imprecise system specifications. To ensure that this approach is broadly applicable, we define this domain using structure external to the modeling formalism itself. Thus we do not have to require, say, that a reliability model formalism be able to express the concept of a component having a range of possible reliabilities.

We use a powerset lattice to provide this extra structure. For a given model formalism, the set \( \text{Model} \) contains all possible models expressible in that formalism. The powerset lattice \( \mathcal{P}(\text{Model}) \) then forms a lattice ordered by specificity: for \( M_1, M_2 \subseteq \text{Model}, M_1 \subseteq M_2 \) indicates that \( M_1 \) contains fewer possible models describing a system, and thus places more constraints on the system, than \( M_2 \) does. Likewise, \( M_1 \cap M_2 \) produces a set of models that fit the constraints associated with \( M_1 \) and with \( M_2 \); \( M_1 \cup M_2 \) produces a set of models where constraints from either may hold.

Singleton sets (i.e., sets of the form \( \{m\}, m \in \text{Model} \) correspond to fully-specified models, and \( \emptyset = \bot \) corresponds to an “impossible” system—one with contradictory modeling requirements.

To make the notation clearer and more consistent, we will define \( \text{Model} = \mathcal{P}(\text{Model}) \), as the powerset lattice of the original set of models, \( \text{Model} \). For the powerset lattice \( \mathcal{P}(\text{Model}) \), we will use the rounded operators (\( \sqsubseteq, \sqcup, \sqcap \)) to prevent confusion with the square operators of the lattice \( \mathcal{P}(\text{Properties}) \), and of lattices in the abstract.

3.3 Correctness

In this work, we represent the set of systems by \( \mathcal{S} \). We think of these systems abstractly; thus, we do not concern ourselves with the representation of \( \mathcal{S} \) or its elements. When we speak of a system \( s \in \mathcal{S} \), we understand \( s \) to be the system to be modeled.

Any system \( s \in \mathcal{S} \) is described by a number of elements of \( \mathcal{P}(\text{Properties}) \). To formalize this notion, we use a correctness relation to relate a system to properties (and later, models) that describe it. We suppose a relation \( R_p : \mathcal{S} \to \mathcal{P}(\text{Properties}) \) where \( s R_p p \) if and only if \( p \) describes the system \( s \). We must assume the existence of \( R_p \), since the properties of the system being designed are determined by the designer. However, abstract interpretation allows us to induce correctness relationships between systems and models based on \( R_p \)—in other words, abstract interpretation enables sound transformations between system properties and system models.

**Definition 3.1:** A correctness relation \( R_L : \mathcal{S} \to \mathcal{L} \) relates systems to elements of a lattice \( \mathcal{L} \). Two attributes hold for \( R_L \):

(i) If \( s R_L l_1 \) and \( l_1 \sqsubseteq l_2 \), then \( s R_L l_2 \).

(ii) If \( \forall l \in \mathcal{L'} \sqsubseteq l, s R_L l \), then \( s R_L \sqcap \mathcal{L'} \).

In terms of \( \mathcal{P}(\text{Properties}) \) and its correctness relation \( R_p \), Property (i) states that we can relax correct constraints without violating their correctness. The reverse does not hold, otherwise, the inconsistent constraint \( \bot \) would describe every system. The formalization of relaxation of constraints as described by Property (i) allows us to generalize constraints and therefore plays a crucial role in modeling abstraction.

Property (ii) requires that for any set of constraints \( \mathcal{L'} \) there exist a “best” constraint that correctly describes any system described by every constraint in \( \mathcal{L'} \). We can apply this property to the constraints derived from several models to narrow down our description of a given system’s properties. In this sense, it allows us to derive a specific result from a number of more
general results. Note that the converse of (ii) follows from (i), so (ii) could also be written as a biconditional.

3.4 Abstraction and Concretization

Given a correctness relation $R_F$ for Properties, we desire to define a mapping between Properties and a modeling formalism $\text{Model}$ that induces a correctness relation $R_M : S \rightarrow \text{Model}$. Furthermore, this mapping must allow for the modeling domain to abstract system constraints. For instance, a topology model should be able to discard constraints on component reliability.

The formalism of choice for this task is a Galois connection:

**Definition 3.2:** A Galois Connection $(P, \alpha, \gamma, M)$ between two complete lattices $P$ and $M$ consists of a pair of monotone functions $\alpha : P \rightarrow M$ and $\gamma : M \rightarrow P$ for which the following relationships hold:

\[
\begin{align*}
(\gamma \circ \alpha)(p) & \sqsubseteq p \quad (1) \\
(\alpha \circ \gamma)(m) & \sqsubseteq m \quad (2)
\end{align*}
\]

We refer to $P$ as the concrete domain, $M$ as the abstract domain, $\alpha$ as the abstraction operator, and $\gamma$ as the concretization operator.

In terms of models and properties, $\alpha$ abstracts a model, $m$, from a set of constraints on a system, $p$, and $\gamma$, derives, or concretizes, system constraints from a model of that system. Relationship (1) states that abstracting the model $m$ from constraints $p$, then concretizing constraints from that model, results in constraints that are at most more general than those of $p$. In other words, abstraction may relax constraints irrelevant to the model formalism, but it cannot produce a model that implies constraints that contradict $p$. Relationship (2) requires that Properties be able to completely capture the constraints imposed by each model formalism, meaning that if constraints are concretized from a model, $m$, of a system, any other model abstracted from these constraints will be as least as specific as the original model, $m$. Concretization may introduce additional constraints, but in practice, the $\sqsubseteq$ of (2) will often be strict equality in practice.

Next, we show that each Galois connection induces a correctness relation $R_M$ on the abstract domain.

**Theorem 3.1:** Given a Galois connection $(P, \alpha, \gamma, M)$ and a correctness relation $R_F : S \rightarrow P$, the relation $R_M : S \rightarrow M$ defined by $s R_M m \iff s R_F \gamma(m)$ is a correctness relation.

**Proof:** We must show that (i) and (ii) from Definition 3.1 hold for $R_M$. Take $s \in S$ and $m_1, m_2 \in M$.

\[
\begin{align*}
s R_M m_1 & \wedge m_1 \sqsubseteq m_2 \\
\iff & s R_F \gamma(m_1) \wedge m_1 \sqsubseteq m_2 & (\text{Defn. of } R_M) \\
\iff & s R_F \gamma(m_1) \wedge m_1 \sqsubseteq \gamma(m_2) & (\gamma \text{ monotone}) \\
\implies & s R_F \gamma(m_2) & (\text{Prop. (i) for } R_F) \\
\iff & s R_M m_2 & (\text{Defn. of } R_M)
\end{align*}
\]

The proof of (ii) uses the fact that $\gamma$ is completely multiplicative, that is, $\square \{ \gamma(m) \mid m \in M \} = \gamma(\square M)$. Take $s \in S$ and $M' \subseteq M$.

\[
\begin{align*}
\forall m \in M' s R_M m & \iff \forall m \in M', s R_F \gamma(m) & (\text{Defn. of } R_M) \\
\implies & s R_F \gamma(\square M') & (\text{Prop. (ii) for } R_F) \\
\iff & s R_M \gamma(\square M') & (\text{Multiplicativity of } \gamma) \\
\iff & s R_M m & (\text{Defn. of } R_M)
\end{align*}
\]

Put in terms of models and system properties, if we define a Galois connection between Properties and the lattice for a given modeling formalism $\text{Model}$, then every correct collection of system constraints abstracts to a correct model and every correct model concretizes to a correct collection of system constraints. Therefore, we have developed a provably sound definition of the nature of model abstraction.

3.5 Model Transformation

Given this formalization of system and model semantics, we can now formalize the problem of model transformation. Suppose we have a properties domain and two modeling formalisms with associated Galois connections to the properties domain $(\text{Properties}, \alpha_{\text{Pl}}, \gamma_{\text{Pl}}, \text{Model}_1)$ and $(\text{Properties}, \alpha_{\text{Pl2}}, \gamma_{\text{Pl2}}, \text{Model}_2)$. Furthermore, we have a correctness relation $R_F$ which induces correctness relations $R_{\text{Pl1}}$ and $R_{\text{Pl2}}$.

**Definition 3.3:** A model transformation from $\text{Model}_1$ to $\text{Model}_2$ is a semantically sound mapping $\tau^{\text{Pl2}}_{\text{Pl1}} : \text{Model}_1 \rightarrow \text{Model}_2$. That is, if $m_1 \in \text{Model}_1$ is correct, then $\tau^{\text{Pl2}}_{\text{Pl1}}(m_1)$ is also correct.

We can define $\tau^{\text{Pl2}}_{\text{Pl1}}$ by first concretizing constraints from $m_1 \in \text{Model}_1$, then abstracting an element of $\text{Model}_2$ from it.

**Theorem 3.2:** The mapping $\tau^{\text{Pl2}}_{\text{Pl1}}(m_1) = (\alpha_{\text{Pl2}} \circ \gamma_{\text{Pl1}})(m_1)$ is sound.

**Proof:** Take $s \in S$ and $m_1 \in \text{Model}_1$.

\[
\begin{align*}
s R_{\text{Pl1}} m_1 & \iff s R_F \gamma_{\text{Pl1}}(m_1) & (\text{Defn. of } R_{\text{Pl1}}) \\
\iff & s R_F (\gamma_{\text{Pl2}} \circ \alpha_{\text{Pl2}} \circ \gamma_{\text{Pl1}})(m_1) & \text{Eqn. (1), Prop. (i)} \\
\iff & s R_{\text{Pl2}} (\alpha_{\text{Pl2}} \circ \gamma_{\text{Pl1}})(m_1) & \text{Defn. of } R_{\text{Pl2}}
\end{align*}
\]

To sum up the transformation process: begin with a model $m_1 \in \text{Model}_1$. Concretize properties of the system from $\{m_1\}$, then apply $\tau^{\text{Pl2}}_{\text{Pl1}}$ to produce a set of models $M_2' \subseteq \text{Model}_2$. Finally, select a model from $M_2'$ by introducing information about the system not present in $m_1$.

Figure 1 illustrates the domains, mappings, and relationships present in this formalization of model transformation.
3.6 Selection and Specificity

Recall that the elements of $\mathfrak{Model}$ are sets of models. To concretize properties of a single model $m \in \mathfrak{Model}$, we first map it to $\{m\} \in \mathfrak{Model}$, then apply $\gamma$. Conversely, for a set of models $M$ produced from an abstraction operation, each model in that set equally captures the system constraints from which $M$ was abstracted. If $M = \emptyset$, then the chosen modeling formalism cannot reason about the given system constraints. If $M = \{m\}$, then the abstraction process has produced a single model describing the system. Otherwise, the system constraints lack some information about the system that is relevant to this modeling formalism. In this case, the user must introduce new information about the system by selecting one model from this set. For example, one may have to provide information about component reliability when selecting a reliability model.

We represent this selection process as a function $\sigma : \mathfrak{Model} \rightarrow \mathfrak{Model}$; the definition of $\sigma$ depends entirely upon the exact system being modeled. While the known system constraints may not be precise enough to indicate exactly which model in the set is correct, they still indicate that the correct model is in the given set of models. Therefore we can constrain $\sigma$ to not produce a model which we know is incorrect even when we do not have enough information about the system to produce a single model.

Definition 3.4: The function $\sigma : \mathfrak{Model} \rightarrow \mathfrak{Model}$ is a selection operator if the following conditions hold:

\begin{enumerate}
  \item $\sigma(m) \in m$
  \item If $s R \alpha m$, then $s R \gamma \sigma(m)$
\end{enumerate}

Given a selection operator, we can incorporate the newly introduced information back into the system properties domain, allowing future transformations to include these constraints and therefore produce more specific results. Take $p \in \mathfrak{Properties}$ such that $s R \alpha p$. Derive the exact model of formalism $\mathfrak{Model}$ by letting $m := (\sigma \circ \alpha)(p)$. By definition of $\sigma$ we know $s R \gamma \sigma(m)$, so $s R p \gamma \{m\}$. Finally, we can construct a more specific element $p' \in \mathfrak{Properties}$ by $p' := p \cap \gamma \{m\}$. The correctness of $p'$ follows from property (ii) for $R \alpha$, and by definition of $\cap$, $p' \subseteq p$.

Figure 2 depicts the relationship between these given functions and domains.

4 CONCLUSION AND FUTURE WORK

In this paper, we have demonstrated a formalization of model and system semantics. Models abstract system semantics; therefore, we can derive, or concretize, constraints on a system from models of it. Conversely, given constraints on a system, we can abstract a set of models that are consistent with those constraints.

To formalize the soundness of this approach, we apply abstract interpretation, which defines a correctness relation between systems and constraints. If our abstraction and concretization mappings between a given modeling formalism and system constraints form a Galois connection between the two domains, we can show that these mappings and the correctness relation for system constraints induce a correctness relation between systems and the models of the modeling formalism.

Through this lens, the process of model transformation becomes the process of concretizing system properties from one model, then abstracting a second model from these properties. We show that this process is sound; that is, if the initial model is correct, then the final model will also be correct.

Future work will take several directions. We are currently working on relating models of different aspects of a system—in this case, reliability and topology. This work will demonstrate both how topology affects system reliability by introducing dependencies between components and how reliability, via the same dependencies and the constraints on system functionality, constrains the choice of topologies for which that definition of reliability holds.

We plan to further extend the work of this paper to other model types and other choices of system properties. Expanding the possible transformations will allow us to relate modeling techniques from various system domains; for example, we may apply a water distribution network analysis technique to a power grid, or incorporate both cyber and physical models into a cyber–physical model.

Another avenue of research is to expand the formalization of models and systems to other metamodeling tasks. A salient challenge in the design of complex systems is that of heterogeneous model composition: combining component models that use various modeling formalisms into a single model of a complete system. The abstraction and concretization functions defined in this work provide a basis for developing these connections. It may even be possible to perform this...
composition at a higher level, enabling the creation of hybrid modeling formalisms and associated solution and evaluation techniques.

Finally, the task of developing this approach into a tool for system designers will certainly present its own challenge. Such a tool must be interactive and scalable to complex, real-world systems, all without requiring the user to have a deep understanding of the underlying theory.

REFERENCES


©2019 by the authors
COMPOSITE SANDWICH STRUCTURES WITH ADDITIVELY MANUFACTURED METALLIC CORES

Okanmisope Fashanu
oaff4x@mst.edu
Department of Mechanical and Aerospace Engineering
Advisors: Dr. K. Chandrashekhara and Dr. Joseph Newkirk

ABSTRACT
Sandwich composite structures offer numerous advantages over conventional monolithic materials. Some of such advantages include lightweight, excellent damping properties, high bending, torsional stiffness, and superior thermal insulation. A recognized failure mode in sandwich composite structures is the delamination of the facesheets from the core. In the present work, the mechanical performance of a sandwich structure with an irregular core is evaluated using finite element analysis. A nonlinear model for 304L stainless steel is developed. The honeycomb core was designed in-house using Nx and Comsol commercial software while the honeycomb homogenization and sandwich performance was evaluated using Abaqus commercial software.

1. INTRODUCTION
Sandwich structures continue to gain usage in industries that require a combination of good mechanical performance and low weight [1]. These structures consist of two thin and stiff facesheets separated by a lightweight core. This configuration allows for high stiffness and moment of inertia when compared to a monolithic plate; facesheets absorb most of the in-plane loading (bending and stretching), while the core absorbs most of the shear load [2]. In the aerospace industry where specific stiffness of a component is of great importance, honeycomb-core sandwich structures are of great interest [3]. Honeycomb-core sandwich structures are manufactured by bonding the facesheet to the honeycomb core using an adhesive film. This type of structure is known to exhibit low in-plane strength and is vulnerable delamination [4-7]. The performance of a honeycomb-core sandwich structure is dependent on the load direction, core topology and the quality of adhesion between the core and the facesheets.

Many researchers have studied the performance of traditionally made honeycomb-core sandwich structures through experiments and numerical modelling [8-12]. For example, Ijaz et al [8] investigated the use of strain energy homogenization to estimate the bending stiffness of a sandwich structure. They observed that by homogenizing a representative volume element and using the resulting estimated elastic properties as input during their simulation they obtained accurate bending response as well as saved computational time. Okada et al [9] investigated peel fracture in honeycomb-core sandwich structures with emphasis on the role of fillets in fracture energy and energy absorption. It was observed during their study that the by selecting the right fillet size the critical strain energy release rate (Gc) of the sandwich can exceed that of a monolithic laminate made using the same material as the facesheet of the honeycomb-core sandwich. They concluded that the size of resin fillets (menisci) played an important role in the energy absorption of the sandwich structures upon fracture. Grove et al [10] investigated the effects of processing parameters on the peel strength of co-cured honeycomb sandwich panels. It was observed that curing temperature, heating rate, vacuum pressure and vacuum time played an important role in obtaining a strong skin to core bond. El-sayed et al [11] investigated the use of computational models in the prediction of interfacial delamination in sandwich structures. They observed that the cohesive layer concept was adequate in predicting sandwich delamination.

During the bonding process, the presence of flaws at the interface act as delamination initiation sites when the exterior surface is subjected to impact force. Also, at low relative densities (thin cell walls) shear stress induced during loading can peel the face-sheets off the honeycomb surface. Interfacial delamination, delamination between the face-sheets from the honeycomb, can lead to catastrophic failure of the sandwich structure. Therefore, it is important to improve the strength of the bond between the core and the facesheets. In this study, a honeycomb structure (to be manufactured using additive manufacturing) is designed and evaluated. This new (irregular) honeycomb was designed to have an increased wall thickness close to the bonding area whilst retaining similar relative density with the control (regular honeycomb). Finite element analysis was used to investigate the bond strength of this new honeycomb. The honeycomb was designed in-house using Nx commercial software while honeycomb homogenization and sandwich performance was evaluated about using Abaqus.

2. METHOD
2.1. Specimen description
The specimen used in the analysis is shown in Figure 1. It consists of a 304L stainless steel honeycomb core sandwiched between two thin carbon fiber reinforced polymer (CFRP)
facesheets. One sandwich structures is evaluated in this study (irregular honeycombs with a larger contact area) Figure 2. The irregular honeycomb-core was designed using Nx commercial software and Comsol multiphysics. Nx was used to draw the CAD file while Comsol multiphysics was used to optimize the design. The sandwich structure was modelled with the layup arrangement [45/0/-45/90/core/90/-45/0/45].

2.2. Finite element analysis

In order to evaluate the mechanical performance of the sandwich structure, simulation mimicking the modified cracked sandwich beam (CSB) test (Figure 3) was performed using the commercial finite element analysis software Abaqus [13]. The purpose of this analysis was to evaluate During the CSB simulation, interlaminar damage alone was considered and intralaminar damage was ignored since the aim of the simulation was to model the interfacial fracture toughness. The material properties of the CFRP and adhesive used for the FEA was obtained from literature while the core material model for 304L stainless steel properties was obtained from experimental testing of a solid 304L sample.

3. THEORY

The configuration of the sandwich structure gives the sandwich its exceptional properties. The transfer of load from the facesheet to the core can be classified as a mechanistic process that involves interfacial bond and friction. Interlaminar damage (Delamination), the separation between adjacent bonded layers, is a crack growth problem and can be investigated using fracture mechanics. Currently, cohesive zone models are the most widely used models for predicting delamination in sandwich structures. This is because this model describe delamination over a range of application in which just the strength or energy criterions would be inadequate. The modelling and simulation in this study is based on the interlaminar damage on the cohesive zone. Damage initiation within the zone was modelled using the quadratic stress criterion as can be expressed as:

\[
\left( \frac{\sigma_n}{\sigma_1} \right)^2 + \left( \frac{\sigma_{s1}}{\sigma_2} \right)^2 + \left( \frac{\sigma_{s2}}{\sigma_3} \right)^2 = 1
\]  

(1)

Where \( \sigma_n \) is the stress in pure normal mode, \( \sigma_{s1} \) is the stress in the first shear direction, \( \sigma_{s2} \) is the stress in the second shear direction, and \( \sigma_1, \sigma_2, \sigma_3 \) are the peak strength values in the same directions. Considering compressive normal stresses do not open cracks then:

\[
\langle \sigma_n \rangle = \sigma_n \text{ for } \sigma_n > 0 \text{ and } \langle \sigma_n \rangle = 0 \text{ for } \sigma_n \leq 0
\]  

(2)

After the conditions for damage initiation at the interface are satisfied, further loading causes stiffness degradation. This stiffness degradation is controlled by damage variables that assume values between 0 (undamaged state) and 1 (fully damage). As damage progresses, a mixed mode critical energy rate \( G_c \) is determined using linear softening in an energy based Benzeggagh-Kenane law [14] given by:

\[
G_c = G_{I,C} + \left( G_{II,C} - G_{I,C} \right) \left[ \frac{G_{II}}{G_I} \right]^{\eta}
\]  

(3)

Where \( G_{m,c} \) is the total critical strain energy release rate associated with delamination in mode m, and \( \eta \) is the semi-empirical criterion exponent applied to delamination initiation and growth.
4. RESULTS

4.1. Core design
The honeycomb-core wall was designed to have a higher thickness close to the surface as shown in Figure 4a. The extended surface lead to the creation of a stress concentration site (Figure 4b). Filets have been used to reduce stress concentrations in I-beams thus fillets were added to the structure A parametric study on the effects of the fillet size on the maximum von misses stress was carried out using Comsol and the results are shown in Figure 5. From the figure, it can be seen that there is no clear relationship between the fillet size and the maximum von Mises stress. However, when the fillet size was equal to the length of the extension ‘b’ the maximum von Mises stress was minimum. For the remainder of this study the honeycomb core used had a fillet radius of 0.1 mm. It is important to note that the maximum stress was recorded at the same location in all cases (Figure 4c).

![Figure 4: Honeycomb-core wall](image)

![Figure 5: Relationship between the fillet size and maximum stress](image)

4.2. Development of core material model
The core material model was developed by the compression of three selectively laser melted solid 304L cylinders. The tested samples were of ratio 1:1 (diameter 6.35mm, height 6.35 mm) since the aim of testing was to develop the materials nonlinear model. The crosshead speed of the frame was fixed to obtain a strain rate of \(5 \times 10^{-3} \text{s}^{-1}\) in the sample. Changes in the force and displacement were tracked during tests and used to calculate the stress-strain curve. From the tests, the average yield strength (0.2% offset) was 479±7 MPa. The Young’s modulus was not evaluated because of the size of the sample and therefore was taken from literature [15] to be 168 GPa.

The use of empirical models to depict the material flow behavior of a material strongly depends on the number of parameters in the particular model. The model used in the description of the flow behavior of a particular material represents a typical shape and any attempt to fit the model beyond this distinctive shape will only give an average value of fit parameters (will not truly represent the material’s property). In this study, three empirical models: Hollomon [16], Ludwik [17] and Ludwigson [18], were investigated for the flow behavior of SLM 304L stainless steel. Figure 6 shows an example using the three empirical models. It can be seen that the Ludwigson model fits the experiment data most when compared to the other two relationships which is in agreement with Choudhary et al [19].

![Figure 6: Plot of true stress vs true strain for experimental and different empirical equations](image)

The Hollomon and Ludwigson curves converge at about 20% plastic strain, which was expected since the Ludwingon model is an extension of the Hollomon model. Since the Ludwigson predicted the flow behavior over a wider range of plastic strains (0.03 – 0.5 plastic strain) the Ludwigson (Equation 4) was used in the in this study. The constants shown in Table 1 are the Ludwigson constants where the \(K_1\) and \(n_1\) values can also be used for the Holomons relation (Equation 5) since they are the same constants.

\[
\sigma = K_1 \varepsilon^n + \exp(K_2 + n_2 \varepsilon) \tag{4}
\]
\[
\sigma = K_1 \varepsilon^{n_1} \tag{5}
\]

Where \(K_1\) and \(n_1\) are the strength coefficient and strain hardening exponent respectively.

| \(K_1 [10^9]\) | 1350 |
| \(n_1\) | 0.31 |
| \(K_2\) | 19.3 |
| \(n_2\) | -20 |

Table 1: Variables for the Ludwigson formulation
5. CONCLUSIONS
In the study, the mechanical performance of an irregular honeycomb-core sandwich structure was investigated. A finite element model based on the cohesive zone model was developed to predict the fracture toughness of interface between the core and the facesheet. 304L stainless steel material nonlinear model was also developed. In the future, the cohesive zone model will be refined and experimental validation will be conducted.

8. ACKNOWLEDGMENTS
Support from the Advanced Manufacturing Signature (AMS) seed fund is gratefully acknowledged. The author will also like to thank Manoj Rangapuram and Siva Dasari for their contribution to this work.

9. REFERENCES
COMPOSITE SANDWICH STRUCTURES WITH ADDITIVELY MANUFACTURED METALLIC CORES

Okanmisope Fashanu  
oaff4x@mst.edu  
Department of Mechanical and Aerospace Engineering  
Advisors: Dr. K. Chandrashekhara and Dr. Joseph Newkirk

ABSTRACT

Sandwich composite structures offer numerous advantages over conventional monolithic materials. Some of such advantages include lightweight, excellent damping properties, high bending, torsional stiffness, and superior thermal insulation. A recognized failure mode in sandwich composite structures is the delamination of the facesheets from the core. In the present work, the mechanical performance of a sandwich structure with an irregular core is evaluated using finite element analysis. A nonlinear model for 304L stainless steel is developed. The honeycomb core was designed in-house using Nx and Comsol commercial software while the honeycomb homogenization and sandwich performance was evaluated using Abaqus commercial software.

1. INTRODUCTION

Sandwich structures continue to gain usage in industries that require a combination of good mechanical performance and low weight [1]. These structures consist of two thin and stiff facesheets separated by a lightweight core. This configuration allows for high stiffness and moment of inertia when compared to a monolithic plate; facesheets absorb most of the in-plane loading (bending and stretching), while the core absorbs most of the shear load [2]. In the aerospace industry where specific stiffness of a component is of great importance, honeycomb-core sandwich structures are of great interest [3]. Honeycomb-core sandwich structures are manufactured by bonding the facesheet to the honeycomb core using an adhesive film. This type of structure is known to exhibit low in-plane strength and is vulnerable delamination [4-7]. The performance of a honeycomb-core sandwich structure is dependent on the load direction, core topology and the quality of adhesion between the core and the facesheets.

Many researchers have studied the performance of traditionally made honeycomb-core sandwich structures though experiments and numerical modelling [8-12]. For example, Ijaz et al [8] investigated the use of strain energy homogenization to estimate the bending stiffness of a sandwich structure. They observed that by homogenizing a representative volume element and using the resulting estimated elastic properties as input during their simulation they obtained accurate bending response as well as saved computational time. Okada et al [9] investigated peel fracture in honeycomb-core sandwich structures with emphasis on the role of fillets in fracture energy and energy absorption. It was observed during their study that the by selecting the right fillet size the critical strain energy release rate ($G_c$) of the sandwich can exceed that of a monolithic laminate made using the same material as the facesheet of the honeycomb-core sandwich. They concluded that the size of resin fillets (menisci) played an important role in the energy absorption of the sandwich structures upon fracture. Grove et al [10] investigated the effects of processing parameters on the peel strength of co-cured honeycomb sandwich panels. It was observed that curing temperature, heating rate, vacuum pressure and vacuum time played an important role in obtaining a strong skin to core bond. El-sayed et al [11] investigated the use of computational models in the prediction of interfacial delamination in sandwich structures. They observed that the cohesive layer concept was adequate in predicting sandwich delamination.

During the bonding process, the presence of flaws at the interface act as delamination initiation sites when the exterior surface is subjected to impact force. Also, at low relative densities (thin cell walls) shear stress induced during loading can peel the face-sheets off the honeycomb surface. Interfacial delamination, delamination between the face-sheets from the honeycomb, can lead to catastrophic failure of the sandwich structure. Therefore, it is important to improve the strength of the bond between the core and the facesheets. In this study, a honeycomb structure (to be manufactured using additive manufacturing) is designed and evaluated. This new (irregular) honeycomb was designed to have an increased wall thickness close to the bonding area whilst retaining similar relative density with the control (regular honeycomb). Finite element analysis was used to investigate the bond strength of this new honeycomb. The honeycomb was designed in-house using Nx commercial software while honeycomb homogenization and sandwich performance was evaluated about using Abaqus.

2. METHOD

2.1. Specimen description

The specimen used in the analysis is shown in Figure 1. It consists of a 304L stainless steel honeycomb core sandwiched between two thin carbon fiber reinforced polymer (CFRP)
One sandwich structures is evaluated in this study (irregular honeycombs with a larger contact area) Figure 2. The irregular honeycomb-core was designed using Nx commercial software and Comsol multiphysics. Nx was used to draw the CAD file while Comsol multiphysics was used to optimize the design. The sandwich structure was modelled with the layup arrangement [45/0/-45/90/core/90/-45/0/45].

2.2. Finite element analysis

In order to evaluate the mechanical performance of the sandwich structure, simulation mimicking the modified cracked sandwich beam (CSB) test (Figure 3) was performed using the commercial finite element analysis software Abaqus [13]. The purpose of this analysis was to evaluate During the CSB simulation, interlaminar damage alone was considered and intralaminar damage was ignored since the aim of the simulation was to model the interfacial fracture toughness. The material properties of the CFRP and adhesive used for the FEA was obtained from literature while the core material model for 304L stainless steel properties was obtained from experimental testing of a solid 304L sample.

3. THEORY

The configuration of the sandwich structure gives the sandwich its exceptional properties. The transfer of load from the facesheet to the core can be classified as a mechanistic process that involves interfacial bond and friction. Interlaminar damage (Delamination), the separation between adjacent bonded layers, is a crack growth problem and can be investigated using fracture mechanics. Currently, cohesive zone models are the most widely used models for predicting delamination in sandwich structures. This is because this model describe delamination over a range of application in which just the strength or energy criterions would be inadequate. The modelling and simulation in this study is based on the interlaminar damage on the cohesive zone. Damage initiation within the zone was modelled using the quadratic stress criterion as can be expressed as:

\[
\left(\frac{\langle \sigma_n \rangle}{\sigma_1}\right)^2 + \left(\frac{\sigma_{s1}}{\sigma_2}\right)^2 + \left(\frac{\sigma_{s2}}{\sigma_3}\right)^2 = 1
\]

Where \( \sigma_n \) is the stress in pure normal mode, \( \sigma_{s1} \) is the stress in the first shear direction, \( \sigma_{s2} \) is the stress in the second shear direction, and \( \sigma_1, \sigma_2, \sigma_3 \) are the peak strength values in the same directions. Considering compressive normal stresses do not open cracks then:

\[
\langle \sigma_n \rangle = \sigma_n \text{ for } \sigma_n > 0 \text{ and } \langle \sigma_n \rangle = 0 \text{ for } \sigma_n \leq 0
\]

After the conditions for damage initiation at the interface are satisfied, further loading causes stiffness degradation. This stiffness degradation is controlled by damage variables that assume values between 0 (undamaged state) and 1 (fully damage). As damage progresses, a mixed mode critical energy rate \( G_c \) is determined using linear softening in an energy based Benzegag-Kenane law [14] given by:

\[
G_c = G_{I,c} + \left(G_{II,c} - G_{I,c}\right)\left[\frac{G_{II}}{G_I - G_{II}}\right]^{\eta}
\]

Where \( G_{m,c} \) is the total critical strain energy release rate associated with delamination in mode m, and \( \eta \) is the semi-empirical criterion exponent applied to delamination initiation and growth.
4. RESULTS

4.1. Core design

The honeycomb-core wall was designed to have a higher thickness close to the surface as shown in Figure 4a. The extended surface lead to the creation of a stress concentration site (Figure 4b). Filets have been used to reduce stress concentrations in I-beams thus fillets were added to the structure. A parametric study on the effects of the fillet size on the maximum von Mises stress was carried out using Comsol and the results are shown in Figure 5. From the figure, it can be seen that there is no clear relationship between the fillet size and the maximum von Mises stress. However, when the fillet size was equal to the length of the extension ‘b’ the maximum von Mises stress was minimum. For the remainder of this study the honeycomb core used had a fillet radius of 0.1 mm. It is important to note that the maximum stress was recorded at the same location in all cases (Figure 4c).

![Figure 4: Honeycomb-core wall](image)

![Figure 5: Relationship between the fillet size and maximum stress](image)

4.2. Development of core material model

The core material model was developed by the compression of three selectively laser melted solid 304L cylinders. The tested samples were of ratio 1:1 (diameter 6.35mm, height 6.35 mm) since the aim of testing was to develop the materials nonlinear model. The crosshead speed of the frame was fixed to obtain a strain rate of $5 \times 10^{-3} \text{ s}^{-1}$ in the sample. Changes in the force and displacement were tracked during tests and used to calculate the stress-strain curve. From the tests, the average yield strength (0.2% offset) was 479±7 MPa. The Young’s modulus was not evaluated because of the size of the sample and therefore was taken from literature [15] to be 168 GPa.

The use of empirical models to depict the material flow behavior of a material strongly depends on the number of parameters in the particular model. The model used in the description of the flow behavior of a particular material represents a typical shape and any attempt to fit the model beyond this distinctive shape will only give an average value of fit parameters (will not truly represent the material’s property). In this study, three empirical models: Hollomon [16], Ludwik [17] and Ludwigson [18], were investigated for the flow behavior of SLM 304L stainless steel. Figure 6 shows an example using the three empirical models. It can be seen that the Ludwigson model fits the experiment data most when compared to the other two relationships which is in agreement with Choudhary et al [19].

![Figure 6: Plot of true stress vs true strain for experimental and different empirical equations](image)

The Hollomon and Ludwigson curves converge at about 20% plastic strain, which was expected since the Ludwigson model is an extension of the Hollomon model. Since the Ludwigson predicted the flow behavior over a wider range of plastic strains (0.03 – 0.5 plastic strain) the Ludwigson (Equation 4) was used in the in this study. The constants shown in Table 1 are the Ludwigson constants where the $K_1$ and $n_1$ values can also be used for the Hollomons relation (Equation 5) since they are the same constants.

$$\sigma = K_1\varepsilon^n_1 + \exp(K_2 + n_2\varepsilon)$$

(4)

$$\sigma = K_1\varepsilon^n_1$$

(5)

Where $K_1$ and $n_1$ are the strength coefficient and strain hardening exponent respectively.

<table>
<thead>
<tr>
<th>$K_1$ [$10^3$]</th>
<th>1350</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>0.31</td>
</tr>
<tr>
<td>$K_2$</td>
<td>19.3</td>
</tr>
<tr>
<td>$n_2$</td>
<td>-20</td>
</tr>
</tbody>
</table>

Table 1: Variables for the Ludwigson formulation
5. CONCLUSIONS
In the study, the mechanical performance of an irregular honeycomb-core sandwich structure was investigated. A finite element model based on the cohesive zone model was developed to predict the fracture toughness of interface between the core and the facesheet. 304L stainless steel material nonlinear model was also developed. In the future, the cohesive zone model will be refined and experimental validation will be conducted.

8. ACKNOWLEDGMENTS
Support from the Advanced Manufacturing Signature (AMS) seed fund is gratefully acknowledged. The author will also like to thank Manoj Rangapuram and Siva Dasari for their contribution to this work.

9. REFERENCES
ON MAXIMIZING TASK THROUGHPUT IN IOT-ENABLED 5G NETWORKS UNDER LATENCY AND BANDWIDTH

Ajay Pratap, Ragini Gupta, Venkata Sriram Siddhardh Nadendla and Sajal K. Das
Department of Computer Science, Missouri University of Science and Technology, Rolla, MO 65409 USA
E-mail: {ajaypratapf, rg5rv, nadendla, sdas}@mst.edu

ABSTRACT
Fog computing in 5G networks has played a significant role in increasing the number of users in a given network. However, Internet-of-Things (IoT) has driven system designers towards designing heterogeneous networks to support diverse demands (tasks with different priority values) with different latency and data rate constraints. In this paper, our goal is to maximize the total number of tasks served by a heterogeneous network, labeled task throughput, in the presence of data rate, latency constraints, and priorities regarding computational needs. Since our original problem is NP-hard, we propose an approximate efficient solution based on graph-coloring techniques. Through simulation analysis, laboratory testbed experiments and performance evaluation comparison with other resource allocation algorithms, we demonstrate the effectiveness of the proposed algorithm in a multi-tier IoT enabled 5G network.

Index Terms—Resource Allocation; PRB; 5G; IoT; fog; Graph.

1 INTRODUCTION
Despite the availability of high-performance computing abilities at the base station (BS), the main challenge in designing an IoT network is to facilitate user mobility which results in limited resources in terms of battery power, bandwidth and computational capacity [1]. Therefore, fog computing has been proposed in 5G networks to reduce the computational load at the BS and support the ever-increasing number of mobile IoT devices [2]. The IoT devices generate tasks via sensing information from a physical phenomenon, and send pre-processed data to a gateway node called fog access point (FAP). Upon receiving the data from IoT devices, FAP executes the task and sends the response back to respective IoT device.

For example, a smart-health IoT network designed to serve stroke patients in a rehabilitation center. While it is necessary to continuously monitor various signals such as blood pressure, heart rate and blood sugar levels in multiple patients, there are other tasks such as fall detection (typically detected using accelerometers, gyroscopes and surveillance cameras) that play a crucial role in the avoidance of accidents during the rehabilitation period. Therefore, tasks such as fall detection take precedence over processing blood sugar readings. In other words, IoT devices typically generate heterogeneous demands (multi-priority tasks) that require diverse resource requirements (e.g. bandwidth, computational power) in the presence of non-identical latency constraints. In such a scenario, BS should prioritize tasks that need to be served and allocate necessary resources accordingly to different FAPs, via integrating heterogeneous constraints and dynamic network environments [3].

Related Work
A few attempts have been made to address similar problems in recent literature. For example, Samie et al. have proposed a novel resource management scheme for IoT devices in [1], where they have reasoned out the need for a discrete number of resources at different stages of operation in the context of a smart health application. On the other hand, Zhang et al. have addressed the resource allocation problem using a three-tier solution, which is based on Stackelberg games and matching theory. This approach is not applicable when the IoT devices will have heterogeneous non-uniform latency constraint. In [4], a joint radio and computational resource allocation in the IoT Fog computing model has been studied. In [5] authors have proposed a joint energy and latency optimization framework for IoT enabled fog access radio. The authors proposed a knapsack based approach to solve the optimization problem. Latency and non-sharability of limited resources are the main drawbacks of this approach. The authors have assumed the predefined static capacity of each FAP and as soon as the total demand of resources goes higher than the capacity, IoT requests start falling down. Dynamic adaptation and prioritization of IoT devices over limited available resources is another shortcoming of the existing works.

Our Contributions
In this paper, our goal is to maximize the total number of heterogeneous tasks (a.k.a. task throughput) served by the 5G network in the presence of data rate/latency constraints, along with device preferences regarding computational needs. Given that this problem is NP-Hard, we propose a graph-coloring based algorithm with pseudo-polynomial time complexity, to find an efficient solution. The novelty in our solution approach lies in our system framework where FAP nodes relay the task requests submitted by the IoT devices to the BS, so that the BS can centrally allocate optimal resources. Based on the priority of the tasks and network connectivity, the BS first identifies all the high priority tasks and allocates necessary resources to appropriate FAPs. If there are any residual resources that remain unassigned (or can be reused whenever FAPs are non-interfering with each other), the BS allocates them to serve the low priority tasks. Our solution approach outperforms state-of-the-art algorithm in the literature because it relies on the notion of
of reuse of resources whenever FAPs are non-interfering with each other.

2 SYSTEM MODEL

Consider a network shown in Fig. 1, where \( M \) heterogeneous IoT devices request bandwidth and computational resources to BS regarding their respective tasks. Let \( \mathbb{I} = \{I_1, \ldots , I_M\} \) denote the different IoT devices in the network. Assume that the \( i^{th} \) device \( I_i \) generates a task \( s_i \in \Psi \), where \( \Psi \) represents the set of tasks that the network can execute. Assume that there are \( K \) fog access points (FAPs) in the network labeled as \( F = \{F_1, \ldots , F_K\} \), which are equipped with CPU cycle rates \( \mathbb{C} = \{c_1, \ldots , c_K\} \). In other words, \( F_k \) can process a task at the rate of \( 1/c_k \) computations per unit cycle. In such a case, the goal of the BS is to maximize the total number of tasks served, via finding appropriate pairs of IoT devices and FAPs in order to reduce the overall latency in the system, while simultaneously increasing the overall productivity in terms of resource utilization.

Let \( \Phi_k \) denote the set of all IoT devices that are within the physical proximity of \( F_k \). Note that \( \Phi_k \) can also be interpreted as the coverage area of \( F_k \). Therefore, it is natural for BS to assign \( F_k \) to all the IoT devices within \( \Phi_k \) in order to minimize latency. However, it is also possible that the coverage areas of two nearby FAPs can overlap, which leads to interference (consequently, a reduction in the achievable data rate) in the communication between the IoT devices within the overlap region and the corresponding FAPs [6]. Furthermore, once the BS matches an IoT device to a FAP, the IoT device shares the task details to the FAP using one or more Physical Resource Blocks (PRBs), which is the smallest unit of communication resource assigned by the BS [7], [8]. We have considered Orthogonal Frequency Division Multiple Access (OFDMA) model where, a PRB comprises of 180 KHz bandwidth (\( \Delta f \)) and 0.5 ms time frame [9].

Interference can occur whenever the following possibilities happen: (i) the same PRB is assigned to two IoT devices \( I_i \) and \( I_j \) which are within the coverage area of \( F_k \), or (ii) the same PRB is assigned to two IoT devices, wherein one of them (say \( I_i \)) is in coverage areas of both \( F_k \) and \( F_{k'} \) and the other (say \( I_j \)) is in the non-overlapping regions of either \( F_k \) or \( F_{k'} \). In other words, we have the following two conditions:

\[
\begin{align*}
(C_1) & \text{ Given } n^{th} \text{ PRB, we have } \sum_{i=1}^{M} y_{k,i} = 1, \text{ for all } F_k \in F. \\
(C_2) & \text{ Given } n^{th} \text{ PRB, we have } y_{k,i} + y_{k',j} = 1, \text{ for all } I_i \in \Phi_k \cap \Phi_{k'} \text{ and } I_j \in \Phi_k \setminus \Phi_{k'}, \text{ for any } F_k \text{ and } F_{k'} \in F.
\end{align*}
\]

2.1 Execution Latency

Let \( x_{k,i} \) denote the BS’s match between the IoT device \( I_i \) with the FAP \( F_k \), i.e.,

\[
x_{k,i}(y_{k,i}) = \begin{cases} 
1, & \text{if } \sum_{n \in \mathbb{N}} y_{k,i} = 0 \\
-\infty, & \text{otherwise.}
\end{cases}
\]

Given that the network can execute any given task in \( \Psi \), let us assume that any FAP in the network can execute a specific task \( s \in \Psi \) in a total of \( \lambda_s \) CPU cycles. Therefore, total time taken to execute a task \( s_i \) generated by \( I_i \) at \( F_k \) is given by,

\[
\beta_{k,i}(y_{k,i}) = |x_{k,i}(y_{k,i})| \cdot \lambda_s \cdot \frac{1}{c_k}
\]

3 PROBLEM FORMULATION

Our goal is to maximize the task throughput, i.e. the total number of tasks executed in the network, which is given by,

\[
\eta(Y) = \sum_{k \in \mathbb{I}} \sum_{i \in \mathbb{I}} x_{k,i}(y_{k,i}).
\]

However, we also want to ensure that the high-priority tasks are given the largest number of PRBs, which is denoted as \( d \). The BS can assign the remaining resources to low priority tasks, where each task can at most have \( d \) PRBs. This can be formulated as the following problem statement:

\[
\begin{align*}
\max_{\mathbb{Y}} & \quad \eta(Y) \\
\text{subject to} & \quad 1. \alpha(y_{k,i}) + \beta(y_{k,i}) + \gamma(y_{k,i}) \leq \tau_i, \text{ for all } i \in \mathbb{I} \\
& \quad 2. \sum_{i \in \mathbb{I}} x_{k,i}(y_{k,i}) \leq 1, \\
& \quad 3. y_{k,i} + y_{k',j} \leq 1, \text{ for all } j \in Int_{i}, k' \in k \cup Int_k \\
& \quad 4. y_{k,i} \in \{0,1\}, x_{c,i} \in \{0,1\} \\
& \quad 5. \sum_{n \in \mathbb{N}} w_i y_{k,i} = d_i, \text{ for all } i \in \mathbb{I}, \\
& \quad 6. \sum_{n \in \mathbb{N}} y_{k,i} \leq d_i, \text{ for all } i \in \mathbb{I},
\end{align*}
\]

\( \text{(P1)} \)

where Constraint 5 enforces the BS to allocate exactly \( d \) number of PRBs for each high-priority task.

4 PROPOSED SOLUTION

Using this graphical representation of interference within the network, we will first design an algorithmic solution to our resource allocation problem based on graph-coloring approach.

4.1 Interference Graph Construction

As discussed in above Section 3, IoT devices belonging to interfering FAP should not be assigned with the same PRBs.
In order to avoid the interference constraint we construct interference graph as follows. We assumed that BS will construct a graph showing the interference scenarios among different FAPs. A FAP considers the other FAP as its neighbour if strength of control packet transmitted from other FAP is more than a predefined threshold. We assumed that each FAP can identify its neighboring FAPs and they report to BS, and BS maintains a interference graph accordingly as shown in Fig. 2. Let interference graph be \( G(V, E) \), where, vertex \( V \) represents set of FAPs and edge \( e_{k,k'} \in E \) represents interference relation between any two FAPs \( F_k \) and \( F_{k'} \) such as defined below:

\[
e_{k,k'} = \begin{cases} 1, & \text{if } F_k \text{ is a neighbor FAP of } F_{k'} \\ 0, & \text{otherwise.} \end{cases}
\]

Let interference graph be \( G(V, E) \), where, vertex \( V \) represents set of FAPs and edge \( e_{k,k'} \in E \) represents interference relation between any two FAPs \( F_k \) and \( F_{k'} \) such as defined below:

\[
e_{k,k'} = \begin{cases} 1, & \text{if } F_k \text{ is a neighbor FAP of } F_{k'} \\ 0, & \text{otherwise.} \end{cases}
\]

4.2 Proposed Algorithm

We design the graph coloring based resource allocation algorithm in order to maximize the number of high priority IoT devices and at the same time provide the required number of PRBs to low priority IoT devices as given in formulated Problem P1. Let’s say the total number of required PRBs of high priority IoT devices as minimum requirement of a FAP whereas, the total number of required PRB’s demand including high priority and low priority IoT devices as maximum quota of FAP in-order to achieve the required transmission rate between IoT and respective FAP. On the other hand the execution latency depends on the CPU cycle rate of FAP. For the sake of simplicity, we assume each co-CPU task will be allocated an equal share of the total CPU rate of respective FAP. To fulfill the minimum requirement and maximum quota of FAPs, we consider that BS will first assign the minimum required PRBs of FAPs and then try to fulfill the maximum quota of respective FAPs. Let, \( D_k^{\min} = \sum_{i \in [k \cup w_i]} d_i \) and \( D_k^{\max} = \sum_{l \in [k \cup w_i]} \sum_{n \in \mathbb{N}} y_{k,i} \) be the minimum requirement and maximum quota of PRBs of a particular FAP \( F_k \), respectively. Consequently, we can write \( D_k^{\min} = \sum_{i \in [k \cup w_i]} d_i \) and \( D_k^{\max} = \sum_{l \in [k \cup w_i]} \sum_{n \in \mathbb{N}} y_{k,i} \) as the minimum requirement and maximum quota of a FAP \( k \). Thus, each FAP divided into two dummy FAPs dubbed as steady FAP and elaborate FAP. The steady FAP keeps track of minimum requirement whereas elaborated FAP undertakes the remaining quota of device. Let \( k^m \) and \( k^g \) represent the steady and elaborated FAP of original FAP \( F_k \), respectively. The maximum quota of steady and elaborated FAP can be written as

\[
D_k^{\min} = D_k^{\min} \text{ and } D_k^{\max} = D_k^{\max} - D_k^{\min}, \text{ respectively.}
\]

If we reserve the \( k^{\min} \) PRBs for steady FAPs means, minimum requirement of original FAP is fulfilled and the remaining resources \( R = N - \sum k^{\min} \) will be assigned to the elaborated FAPs. Moreover, to find out how many number of PRBs are required for steady FAPs is not so trivial. For example, let there be 10 FAPs with minimum PRB requirement of 2 and none of them interfere with each other. Thus, instead of reserving 20 resources, we can just keep 2 resources and by reusing it, the FAPs can fulfill their minimum requirement. In order to sort out this problem, let \( R \) resources are assigned to elaborated FAPs and \( N - R \) resources are fixed for steady FAPs. Thus, we first assign the resources to steady FAPs and then remaining \( R \) resources assigned to the elaborated FAPs.

4.2.1 Resource Allocation Algorithm

Given the input of transformed FAPs and resources, the Algorithm 1 shows the steps of resource allocation procedure. We have introduced different rules for steady and elaborated FAPs in-order to fulfill the minimum and maximum required PRBs. Let \( A_n \) be the set of FAP that the BS has not tried for PRB \( n \) allocation. In each round BS selects the set of non-interfering FAPs for PRB \( n \) (Line 5). In the other-word, BS finds the maximum weighted independent set among FAPs based on priority and \( Z_n^{\max} \) is the best set of FAPs that BS can send message for PRB \( n \) in the current round. If for all the PRBs \( n \in \mathbb{N} \) the set \( Z_n^{\max} \) become empty then Algorithm 1 terminates (lines 8-9). The BS sends message \( msg < n \) to all the \( k \in Z_n^{\max} \) for PRB \( n \) and update candidate list \( A_n \). Upon receiving message \( msg \) \( n \) the FAP \( k \) adds PRB \( n \) into its waiting list (lines 15-16).

We have considered different rules for steady and elaborated FAPs to select the valid PRBs. A steady FAP \( k \) for which waiting list \( W_k \) is not empty, selects its required \( D_k^{\min} \) resources, and reset its waiting list (lines 18-20). For the elaborated FAPs, we follow the following method. Let, \( count \) be the counter to estimate the number of resources allocated to all elaborated FAPs and \( l \) is the index of elaborated FAP initialize to 1 (line 22). We first put all the allocated PRBs of elaborated FAPs into waiting list and then we sequentially check the elaborated FAP into a specific order for PRB allocation. Let index \( l \) be the index of FAP that is being considered.

4.3 Computational Complexity Analysis

As every time whenever the BS sends message to a FAP, it removes that FAP from its candidate list (as shown in line 4 of Algorithm 1). At the end, for each resource, the set of interference free FAP assigned with the same resource. As we have \( N \) PRBs, each time BS selects FAP by looking into the maximum weighted independent graph. Thus, the computational time complexity to allocate the valid resources to each steady FAP be \( O(KN\rho) \) where, \( \rho \) is time complexity for finding MWIS in a graph. When, the Algorithm 1 assigns the resource to a elaborated FAPs (lines 26-36), all the elaborated FAPs needs to be traverse and that may take \( O(K) \) computational time.

\[\text{at} \text{ } D \text{ } \text{ } D \text{ } \text{ } D \text{ } \text{ } D \text{ } \text{ } 1\text{ } \text{ } 1\text{ } \text{ } 1\text{ } \text{ } 1\]
Algorithm 1 Resource Allocation Algorithm

Input: Maximum quota of FAPs $\bar{D}^\text{max}_k, \forall k \in \mathcal{F}$, transformed interference graph $\bar{G}$.
Output: Set of PRBs allocated nodes i.e., $\bar{m}(k), \forall k \in \mathcal{F}$.

1: $\forall k \in \mathcal{F}, \bar{m}(k) = \phi$, waiting list $W_k = \phi$.
2: $\forall n \in \mathcal{N}, \bar{m}(n) = \phi$, candidate list $A_n = \mathcal{F}$.
3: while $\exists A_n \neq \phi$ do
4: for all resource $n$ with $|A_n| > 0$ do
5: $Z_n := \text{FAPs those satisfy } k \in A_n, \forall k' \in \mu(n), e_{k,k'} = 0$.
6: Find the maximum weighted independent set on $Z_n$ as $Z_n^\text{max}$.
7: end for
8: if $\forall n, Z_n^\text{max} = \phi$ then
9: Return $\bar{m}$
10: else
11: for all FAP $k \in Z_n^\text{max}$ do
12: BS sends $msg < k, n >$
13: $A_n = A_n \cup \{k\}$
14: end for
15: Upon receive $msg < k, n >$ on FAP $k$
16: FAP $k$ updates its waiting list $W_k = W_k \cup \{n\}$.
17: end if
18: for all FAP $k$ for which $W_k \neq \phi$ do
19: if $k \in \mathcal{F}$ then
20: FAP $k$ accepts $D_k^\text{max}$ resources from $W_k \cup \bar{m}(k)$ and reset $W_k = \phi$.
21: else
22: $\text{count} = 0$, Index $l = 1$
23: for all $k \in \mathcal{F}$ do
24: $W_k = W_k \cup \bar{m}(k), \mu(k) = \phi$
25: end for
26: while $\text{cnt} < R$ and $\exists k', W_k' \neq \phi$, $|\bar{m}(k')| < \bar{D}_k^\text{max}$ do
27: if $|\mu(k')| < \bar{D}_k$ and $W_k' \neq \phi$, then
28: FAP $k'$ selects PRB $n$ from set $W_k'$, such that
29: $\bar{m}(k') = \bar{m}(k') \cup \{n\}$
30: $W_k' = W_k' - \{n\}$
31: $\text{count} = \text{count} + 1$
32: end if
33: $l = l + 1$
34: end while
35: $W_k = \phi$
36: end for
37: end while

complexity. Thus, the proposed resource allocation method takes total $O(NK^2\rho)$ computational time complexity.

5 PERFORMANCE STUDY

In this section, we evaluate our proposed method based on the following environments. We have assumed that the set of FAPs and IoT devices are deployed randomly in the network area of 500 m x 500 m underlying a cellular BS. The bandwidth is considered as 20 MHz accordingly, the maximum number of available PRBs found as 100 [10]. We have assumed that FAP with a radius of 20 m, randomly deployed in the network. Transmit power of IoT devices is set to 25 dBm. A distance between IoT and FAP is considered between 10 - 15 m. Like [11], [8] we also considered that a PRB can carry 504 bits of data at 64 QAP modulation scheme. The network model is considered as the same [4], [12]. The noise power -114 dBm. The required latency, data size, and corresponding CPU cycles are determined by specific device types. The latency(execution latency) requirement of each IoT device is randomly distributed within [0.1 - 5] minutes. We have considered CPU frequency of FAP processor as 1.4 GHz and computational complexity of the task as 10 computation cycles/bit [5]. For the sake of simplicity we have assumed the response time i.e., downlink latency as a random variable $\gamma_{k,i}(y_{k,i}) = \delta t$, $\delta t \in [0, 1]$ for any IoT device [4].

5.1 Utility as a Surrogate to Task Throughput

In our results, we assume that a task generated by an IoT device will need one PRB to get it executed. Therefore, the total number of tasks served by our proposed algorithm (a.k.a. task throughput) is equal to the total number of PRBs assigned by the BS. In other words, the fraction of assigned PRBs to the total PRBs demanded by IoT device is a monotonically increasing function of the total number of tasks served by the proposed algorithm. As a result, we evaluate the performance of the proposed algorithm by determining utility based resource allocation in three case scenarios of an IoT framework. In the field of network communication, the utility of resources determines the ratio between the amount of the PRBs that are allocated to IoT devices to number of required PRBs of that device i.e., $\text{Utility}_i = \frac{1}{d_i} \sum_{n \in \mathcal{N}} y^n_{k,i},$ for any $F_k \in \mathcal{F}$. In the below, we define the utility of higher priority and lower priority IoT devices based on formulated Problem (P1).

$$\text{Utility}_i = \begin{cases} 1, & \text{if } w_i = 1 \\ \leq 1, & \text{otherwise.} \end{cases}$$ (5)

A) Impact of IoT devices on utility: For an experimental analysis, a set of FAPs ranging from 50 to 250 was chosen. For each set of FAPs, the proposed algorithm is estimated to obtain utility by setting the maximum demand of PRBs i.e., $D$ from 10 to 20. The utility is calculated as an average

Fig. 3. Comparison between number of IoT devices and utility value.
value. Comparison of utility with respect to FAP is shown in Fig. 3. From the graph, we can see that, for a specific demand of resources when there is an increase in total PRB demand, there is a gradual decrease in utility with a sudden fluctuation as number of FAPs approach to 250. This is because as the number of IoT devices increase, there is a throttle for resources in the network leading to a performance decline of the network and hence lower utility value. Additionally, as the demand for resources increase for the same number of FAPs, utility of network resources decrease due to scarcity of resources to meet a higher demand.

B) Impact of interference link density on utility: Fig. 4 demonstrates the graph for utility v/s interference of link density. Link density is defined as a ratio between actual edges in interference graph to maximum possible edges. This is performed to study impact of network interference on utility of resources among network devices. Utility is different for different link density values. From the result, we can observe that as link density increases, resource utility starts decreasing gradually. This is because for a given set of FAPs when there is a higher link density there is stronger interference in wireless link that lead to lower resources re-usability. This causes decline in network resource utility of devices.

C) Impact of number of allocated resources on utility: In Fig. 5, we compare the utilities along with an increment of the number of allocated PRBs. For a given set of IoT devices, as the number of allocated resources in the network increase, there is a sharp increase in the utility as now more resources can be assigned to each IoT device. If the number of allocated resources are small, the devices will not be serviced as per their demand due to interference constraint and hence, smaller utility. However, if the total resources in the network are sufficient, the effective resources allocated to each device is higher, which results in higher utility for the network. Additionally, it is evident from the graph that as the number of IoT devices increase the utility will decrease for the given amount of resources. This is because for a specific amount of allocated resources, as the number of devices in the network increase, more subgroups of non-interfering devices will be formed that can lead to a division of resources into more number of portions. Due to this, some devices may fall short of meeting their minimum resource requirement which causes a drop in the utility of the network resources.

5.2 Latency Evaluation

In this sub-section, we present a prototype of the proposed architecture to compare the performance of the proposed scheme. Further, we compare the total latency of the developed prototype with the obtained numerical results. We assumed that two PRBs are dedicated to each device in order to transmit the data to respective FAP.

5.2.1 Prototype Specification

To create a prototype of the proposed model, we use a laptop, an android mobile phone, and one raspberry pi as IoT devices sending data over Wi-Fi to another raspberry pi which is working as a FAP in our model. The camera enabled laptop takes photos in every interval of 1 ms and forwards the images to the central FAP. Another raspberry pi working as IoT device continuously sends stored images to the raspberry pi working as FAP. Additionally, the android mobile phone sends some continuous text messages through a client application to the FAP raspberry pi over the Wi-Fi network. On the FAP side, upon receiving the data from all three IoT devices, the raspberry pi FAP performs local data processing and once the execution is completed it sends response messages to each respective IoT device. We estimate the proposed scheme on different size of data varying from 4 Mb to 209.92 Mb by

Fig. 6. Prototype setup with mobile, laptop, raspberry pi and cloud with their communication medium.
assuming the task deadline as 1 minute. With the increase of data size, the total time consumed to execute the task also increases. As from Fig. 7 we can observe that the total consumed latency in numerical result and prototype model is almost the same. However, with the increase in data size, the total latency consumed by the prototype model is increasing. The main reason for this difference is that, laptop and smartphone used as IoT devices are not dedicated devices and these devices are running several other applications in parallel while in use as an IoT device in the prototype model.

5.3 Joint PRB and Latency Evaluation

To the best of our knowledge, most of the existing work for resource allocation in the IoT-enabled network did not consider the priority of tasks execution keeping limited PRB and latency constraint into consideration. So, in order to compare our proposed scheme with the existing work [5], we draw a random topology of 20-100 number of IoT devices getting services from 5 FAPs. We assumed that each IoT device will send 1 Mb of data to get executed at FAP. We can conclude from Fig. 8, with the increase in the number of PRBs in the network total latency, is minimized. The reason from this observation is that when the more number of PRBs are allocated for the same task, the total achievable data rate goes high and propagation latency becomes low, respectively. Our proposed scheme gives lower latency compared to the existing work. The reason is that unlike the existing work our proposed scheme always try to maximize the re-usability of PRBs while avoiding interference among FAPs and this phenomenon improves the achievable data rate between FAP and IoT device consequently, our proposed scheme results lower latency in the networks.

Fig. 8. Joint PRB and latency comparison with existing work.

6 CONCLUSION

In this work, we proposed a novel framework for 5G networks where the BS identifies appropriate pairs of IoT devices and FAPs and allocate necessary resources to maximize the total number of tasks served. We proposed a graph-coloring based algorithm to solve it in a computationally tractable manner. Through numerical result and prototype model, we have demonstrated the effect of different parameters over the utility and latency of IoT devices in different environments. In the future, the proposed scheme can be applied to specific applications via considering specific applications by modifying the set of constraints accordingly.

7 FUTURE WORK

In the current implementation, the lab testbed experiments do not incorporate the interference between FAPs which will be addressed as an extended part of this work. Additionally, we will scale the testbed with addition of new nodes and sensors to compute the task throughput under the latency and available bandwidth constraint in accordance with the proposed algorithm. The experiment results will be evaluated against the simulation output and literature work.

REFERENCES

A REVIEW OF EXTRUSION FREEFORM FABRICATION METHODS FOR CERAMIC ADDITIVE MANUFACTURING

Sachin Choudhary, Department of Mechanical and Aerospace Engineer, Missouri University of Science and Technology
Sc35w@mst.edu

Wenbin Li, Department of Mechanical and Aerospace Engineer, Missouri University of Science and Technology
wl9ff@mst.edu

Ming C. Leu, Department of Mechanical and Aerospace Engineer, Missouri University of Science and Technology

ABSTRACT
The Extrusion Freeform additive manufacturing method (EFF) is a very promising AM method for ceramics, which has shown its capability in manufacturing ceramic parts of high density, multiple materials and functionally graded materials with very low wastage of material. This paper studies in detail the technology capabilities and limitations of extrusion-based ceramic additive manufacturing methods, the current progress and challenges and the recent trend of material extrusion-based ceramic additive manufacturing.

1. INTRODUCTION
Several AM techniques for fabricating 3D ceramic components are binder jetting[1], material extrusion [2]–[5], vat photopolymerization [6]–[9], powder bed fusion [10], [11], directed energy deposition [12]–[15], etc. These have found applications in various areas as: (i) Catalyst support materials (e.g., thermal and electrochemical water splitting). [16] (ii) filter-based materials for environmental applications (e.g., heterogeneous catalyst support, photocatalytic device and heavy metal ion separation). [17] (iii) Biocompatible scaffold materials for biomedical applications. [18]–[21] (iv) Electro ceramic materials for energy-storage, sensing, and energy-harvesting applications. [22–25] (v) Light weight structures have a potential advantage in defense in heat-exchanger applications. [26] These techniques did not produce the parts with good mechanical properties due to crack sensitive nature of ceramics which heavily depends upon defects, and they would be expected to exhibit poor mechanical properties even at 80% relative density. Extrusion-based methods are among the most popular approaches for freeform fabrication of ceramic components due to the simplicity and low cost of their fabrication system, high density of their fabricated parts, their capability of producing parts with multiple materials [27] including functionally graded materials [28], and the low amount of material wasted during processing. Here the Technology capabilities and limitations of the EFF method are discussed.

2. Technology capabilities and limitations
Material extrusion-based ceramic additive manufacturing is a subset of AM techniques that rely on the extrusion of viscous ceramic or ceramic precursor suspensions of suitable rheological behavior from a small nozzle, which is computer-controlled for layer-by-layer deposition to form 3D structures. A variety of synonyms exist that may lead to confusion. For instance, Extrusion Freeform Fabrication, Robocasting (RC), Direct Ink Writing (DIW), etc. In this chapter, the term Extrusion Freeform Fabrication (EFF) will be used to represent all varieties of material extrusion-based ceramic additive manufacturing techniques. A simple schematic diagram of an EFF process is given in Figure 1. As EFF is not a one-step AM process, the resultant green body needs to undergo a debinding process to burn off the organic additives and a subsequent sintering process to densify the structures.

Figure 1. Schematic diagram of material extrusion-based ceramic additive manufacturing process

2.1. Types of feedstock
The feedstock used in material EFF are usually in two types of form, i.e., paste-form and filament-form. The paste-form
feedstock is extruded by mechanisms that are capable of dispensing liquid or viscous fluid, such as the positive displacement extruder. The filament-form feedstock, however, is fed by a set of squeezing gear or pulley into a heated chamber (is usually referred to as liquefier), where the solid-state filament undergoes phase transformation to become viscous fluid, and is then extruded through the nozzle for deposition. This squeezing-gear-based filament extruder is widely used in material extrusion-based additive manufacturing of polymers, such as the fused deposition modeling (FDM) process. The two different ways of extrusion-deposition are illustrated in Figure 2.

![Figure 2](image1.png)  
**Figure 2.** (a) Paste feedstock extruded by a plunger-syringe mechanism.

![Figure 3](image2.png)  
**Figure 3.** (b) Filament feedstock extruded by a squeezing gear mechanism.

The paste-form feedstock is used in EFF are closely related to the colloidal suspensions that are used in conventional slurry-based ceramic shaping processes, such as tape casting, slip casting. The filament-form feedstock, however, is similar to an injection molding feedstock, where ceramic powders, polymer binders were mixed and shaped into filaments of 1–3 mm diameters to fit the squeezing gear-based filament extruder available on the market.

### 2.2. Methods of Shape Retention

After the feedstock is deposited to a target layer, a consolidation process follows so that the deposited material can maintain is shape from being deformed by other loads such as gravity, acceleration, nozzle dragging, etc. An important distinction between all these EFF technologies is the method of shape retention. There are phase changing techniques such as the melting-solidification of filament-form feedstock prepared from ceramic powder and thermoplastic polymers, and the crystallization (frozen) process of an aqueous paste-form feedstock. The other method without a phase transition is the evaporation of the solvent to induce stiffening of the suspension to maintain the shape. The latter technique may also utilize reversible gel feedstock whose rheological properties often solely serve as the shape retaining property. Another hybrid between phase transformation and sol-gel technique involves a heated thermo-reversible hydrogel, which is deposited into a cold bath to induce a sol–gel transition. Furthermore, shape retention can be achieved through the usage of UV-curable inks.

### 2.3 Overview of existing EFF processes

The feedstock type and shape retention method being used are the main distinction that defines different EFF processes, as shown in Table 1. The Extrusion Freeforming (EF) [31] was the first technique to utilize extrusion of ceramic pastes (organic-based) to produce three-dimensional components, and is classified as category-1 in Table 1. Pastes of alumina in liquid acrylic monomers were heated to melt state and deposited onto a heated plate to form a layer. A phase transformation process (melting-solidification) was involved to retain the shape of the deposited material. The process was further improved and more complex geometries with other materials such as silicon nitride were fabricated. Similar processes including Multiphase Jet Solidification [32], and Thermoplastic 3D Printing (T3DP) [33] were developed by Landers et al. and Scheithauer et al., respectively, with some deviations regarding the process details, a schematic of T3DP process is shown in Figure 3, which is representative for all similar process in this category. A special case in this category is the Freeze-form Extrusion Fabrication (FEF) process. In the FEF process [34], a high solids loading (> 50 vol %) aqueous paste containing 1-4 vol% organic additives is extruded in a freezing environment to solidify the paste after its deposition. The FEF setup is shown in Figure 4. Freeze drying is then used to remove the water content before sintering. This process is also capable of producing complex and functionally graded parts made of different materials such as alumina, zirconium diboride, boron carbide, zirconium carbide, and bio-active glasses [35].

![Figure 4](image3.png)  
**Figure 4.** Freeze forming process.

| Table 1. Classification of EFF processes according to feedstock form and shape retention method. |
|---------------------------------------------|-------------------------------------------------|
| Feedstock Type                           | Shape Retention Method                          |
| Paste                                    | Evaporation                                     |
| Filament                                 | Phase Transformation, Melting-Solidification   |
| Extrusion                                | Sol-Gel                                         |
| Injection                                | UV-Curable Inks                                 |

2
EFF processes using filament feedstock and rely on phase transformation to retain part shape fall into category-2 in Table 1. Danforth [36] introduced the concept of Fused Deposition of Ceramics (FDC). A commercial Fused Deposition Modeling (FDM) system from Stratasys Inc. (Eden Prairie, MN, USA) were used to extrude ceramic-loaded thermoplastic filaments. The filament was liquefied, extruded, and re-solidified immediately due to cooling because of the fast heat transfer from the printed suspension to the surrounding atmosphere or to the surrounding atmosphere. Since then, they have significantly improved their process and have been able to produce high-quality parts made of different materials for various applications, especially sensors and actuator [37]. A synonym for FDC and FDM is Fused Filament Fabrication (FFF). Hence, some similar process were referred to as FFF instead of FDC. Several ceramic samples fabricated by FDC (FFF) process are demonstrated in Figure 5.

**Table 1**: EFF processes using filament feedstock and rely on phase transformation to retain part shape fall into category-2 in Table 1. The feedstock types are Paste, Filament, Paste, Paste, and Paste. The EFF processes are EF, MJS, T3DP, FEF, FDC, FFF, RC, UVRD, and EPP. The Gelation processes are 3D-Plotting and 3DGP. The Light curing processes are Paste and Paste. The Liquid evaporation processes are Paste, Paste, Paste, Paste, and Paste. The Phase transformation processes are Paste, Paste, Paste, Paste, and Paste. The Feedstock type is Paste, Paste, Paste, Paste, and Paste.

**Figure 4**: A schematic of the T3DP process, and a photo of heating syringe extruder used in the T3DP process.

**Figure 5**: Apparatus of the FEF process in a thermo-insulated chamber equipped with a triple-extrusion device for extruding and blending multiple paste feedstocks.

**Figure 6**: Mullite specimens fabricated by FDC.

RC [38] is a renowned material extrusion-based AM process of ceramics. In this process, typically an aqueous suspension from ceramic materials (e.g. alumina, silica, lead zirconate titanate, hydroxyapatite, silicon carbide, and silicon nitride) is prepared and extruded on to a plate to dry and maintain its shape. RC can produce grid or thin-wall structures for various applications, especially bio-fabrication [39]. Originally, a heated plate was used as the building substrate the RC processes to expedite the paste drying. However, along with more and more innovations carried out to improve the RC process, including modifying feedstock properties, controlling liquid phase evaporation, many variants of RC has been introduced by researchers in recent years. One famous variant is named as Direct Ink Writing or Direct Writing (DIW or DW), which was intensively studied, especially for fabricating sparse ceramic structures such as scaffold, thin walls, etc. Generally, DIW can be treated as a new name of RC, where a heated substrate was rarely used for drying, instead the deposited ceramic materials were dried at ambient temperature. A schematic of RC or DIW/DW is shown in Figure 6.
6. Another important variant is called Ceramic On-Demand Extrusion (CODE) process. In this process, aqueous ceramic pastes and deposited on a substrate located in a tank designed to hold a fluid medium. Once the deposition of each layer is completed, a liquid feeding subsystem pumps oil into the tank surrounding the layer to preclude undesirable water evaporation from the sides of the deposited layers. The level of the liquid is controlled so that it is maintained at a level that is about one layer below the top surface of the part being fabricated. Infrared radiation is then used to uniformly dry the deposited layer for shape retention. This layered uniform radiation drying approach minimizes the water content gradient in the fabricated part and thus enables the CODE process to produce crack-free thick ceramic parts with complex geometries, followed by post-processing. A schematic of the CODE process, as well as a photo of the CODE process setup are shown in Figure 7.

Figure 7. A schematic diagram of a typical RC experimental setup, and a photo showing the printing process of a ceramic scaffold.

A few studies have been carried out using UV radiation to cure the deposited ceramic feedstocks, which belong to category-4 in Table 1. Hazan et al. prepared monomer-based UV curable colloidal alumina and hydroxyapatite pastes for additively fabricate 3d structures through robotic deposition (UVRD). Depending on rheological properties, alumina paste and hydroxyapatite with 45 vol% and 32 vol% solids loading were selected as favorable feedstocks for EFF. The deposited layer exhibited good flow behavior, and sound shape retention and was able to withstand one upcoming layer. Hence, the deposition was interrupted after a deposition of 2 layers and a UV lamp was positioned above the freshly deposited structure for 2 min for curing. A variety of architectures such as high aspect ratio 2-layer periodic lattices, multilayers, single and multilayer spiral are printed successfully, as shown in Figure 8. Faes et al. studied extrusion-based additive manufacturing of ZrO₂ parts using photo initiated polymerization (EPP), where a UV curable resin was mixed with ZrO₂ powders and extruded through a syringe extruder to fabricate 3D parts. Different pastes containing 22.5–55 vol. % zirconia particles were prepared, and a dispersion containing 30 vol.% zirconia showed the most desirable properties such as a suitable homogeneity and shape stability.

Figure 8. A schematic diagram of the CODE process, and a photo of the CODE apparatus.

Ren et al. developed an EFF process using gelation for shape retention, which falls into category-5. The gelation occurred at room temperature by blending gelation initiator and catalyst to a slurry containing ceramic particles, monomer, cross-linking agent, dispersant and solvent. Slurry with 50 vol. % of ZrO₂ particles were mixed with catalyst and initiator and deposited through a screw extruder. The deposited feedstock then underwent a gelation process to settle down and hence retain its shape to form a 3D geometry. This process was named as 3D gel-printing (3DGP). The schematic of the 3DGP process, and sample specimens fabricated by 3DGP are shown in Figure 9, and Figure 10, respectively.
for which several AM systems have been available on the market. Material extrusion-based AM for ceramics (EFF) has been intensively investigated, improved by many researchers, however, has not been commercialized for fabricating advanced ceramics. Very few studies were found on the remaining three techniques, i.e., powder bed fusion, direct energy deposition, and sheet lamination.

Among the four main ceramic AM techniques, i.e., binder jetting, material jetting, vat photopolymerization and material extrusion (EFF), the EFF process holds significant advantages of system simplicity and low cost. The binder jetting process, although holds excellent capability of fabricating geometrically complex ceramic components, advantage in scalability, is often subjected to low density (60% - 90%) of the final products mainly due to the low compaction of the ceramic powder. The vat photopolymerization has exhibited superior part density (above 99%) and geometrical complexity, however, the difficulty of burning out of binder in the post-processing is a hassle which increases the production time and limits the maximum wall thickness (usually < 1cm) in the part. Moreover, the range of ceramic materials that can be used in vat photopolymerization would be narrowed down since the light scattering and absorbing properties of ceramic particles need to be considered. The material jetting technique has the advantages on both dimensional resolution and capability of fabricating ceramic components with multiple materials. Feedstocks with low viscosity are usually required in this process to facilitate the jetting of ‘inks’, hence low solids loading is often preferred, which could potentially lead to a higher shrinkage after sintering. Similar to the binder jetting and vat photopolymerization process, the material jetting process would also suffer from the difficulty of burning out organic binders. The material extrusion (EFF) process, generally has lower capability of fabricating geometrically complex features, as well as lower dimensional resolution when compared to the other three processes. The contradictory between dimensional resolution (nozzle diameter) and productivity also leads to the difficulty of scaling up the EFF fabrication. However, a special case where the EFF is advantageous is the fabrication of scaffolds with purposefully introduced porosity, which are usually used for chemical reaction and bio implant due to its high surface area to volume ration. In this case, EFF can easily 60-100 μm pore size resolution, which may not be achievable with other ceramic AM techniques such as the vat photopolymerization, as it is impossible to remove uncured resin or powder from small mesh-like cavities. The remaining resin will eventually complicated the burning out process and may lead to sample cracking. Besides, EFF holds the advantage of fabricating ceramic components with multiple materials such as functionally graded material (FGM). Moreover, EFF has a wide range of starting material (feedstock). Both organic-based and aqueous feedstock can be used in EFF, where the latter feedstock could address the binder burnout issue in the post-processing of the other three processes, enabling the fabrication of ceramic components with thick (>1cm) walls.

Without the need of considering light scattering and absorbing properties of ceramic components, the EFF process is advantageous for high accuracy, high dimensional resolution, andducibility of fabricating geometrically complex features. It is also advantageous for fabricating scaffolds with purposefully introduced porosity, which is often used for chemical reactions and bio implants. Moreover, EFF has a wide range of starting materials (feedstocks), including organic-based and aqueous feedstocks. The latter feedstocks can address the binder burnout issue in the post-processing of other processes, enabling the fabrication of ceramic components with thick (>1cm) walls. Without the need of considering light scattering and absorbing properties of ceramic particles, the EFF process is advantageous for high accuracy and high dimensional resolution.
properties, EFF also has more flexibility on ceramic particle size and color than the vat photopolymerization process. Among the EFF processes, each category (see Table 1) also hold distinct pros and cons. For FDC (FFF), the filament feedstock preparation is also burdensome and requires several steps. The organic binder removal procedures for categories 1, 2, 4, 5 is difficult and time-consuming, and sometimes causes severe warpage or other defects. It might require multiple cycles with different atmospheres. The feedstocks used in category-3 usually have minimal amount (< 5wt% ) of remaining organics before post-processing, the absence of high concentration of organic binder would result in the lack of strength of the as-deposited part and the dried part, however, it can avoid the issue of binder burnout in the other categories, which is the major advantage. All in all, the EFF process is advantageous in fabricating scaffolds, as well as ceramic components with multiple materials, such as Functionally Graded Materials (FGM). Especially, by using feedstocks with minimal organic binders, such as in the CODE process, EFF will have unique strength for fabricating highly dense, thick ceramic components.

7. CONCLUSIONS

The various EFF (Extrusion Freeform Fabrication) methods were discussed, pros and cons for each processes were studied. It is observed that a lot of EFF methods are not suitable as a manufacturing method owning to poor part quality. Material extrusion-based ceramic additive manufacturing technique was discussed in detail as the most suitable method in ceramic additive manufacturing. Owing to the various pros and cons associated with different EFF processes each has its own specific application area. The Ceramic On Demand Extrusion (CODE) process was described. Among all the EFF processes CODE process has displayed high capability of fabricating ceramic components with multiple materials and highly dense objects of good mechanical properties and greater thicknesses.

8. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of this research by the Intelligent Systems Center.

9. REFERENCES


Abstract—This paper describes the modeling and control of Extreme Fast Charging (XFC) station that can enable dc fast charging of multiple electric vehicles (EVs). The present objective is to develop a model suitable for future grid studies during charging transients. In the station, power is drawn directly from the 12.47 kV medium-voltage grid and the rating of the station is as high as 1 MW. The XFC stations are expected to charge the electric vehicle within 5 min to 10 min and this eventually leads to host of new challenges for its acceptance in distribution systems.

The basic architecture consists of 2 Level Voltage Source Converter (2L-VSC) which forms an interface between ac grid and dc bus, a Dual Active Bridge Converter (DAB) which enables bidirectional power flow during charging and discharging of battery and dc-dc converter at the final stage of feeding power to the vehicle.

The proposed strategy not only enables charging or discharging of EV but also improves the performance of the grid in terms of efficiency, reliability and stability. Simulations were done using MATLAB and PLECS and the results prove the feasibility of proposed grid integration model.

I. INTRODUCTION

The advent of electric vehicles (EVs) dates back to late 19th century and today due to the advancements in smart grid technologies, EVs were found to be the potential substitutes for the combustion engine vehicles. Decreasing energy consumption in the transportation sector is extremely important for energy independence and thus one of the strong motivations behind EV fast charging technologies [1].

The grid integration with an EV poses lot of challenges. Hence, this paper presents a smart control system that deals with the proper power exchange between the ac grid and dc bus and thus ensuring a quality power. A Dual Second Order Generalized Integrator (DSOGI) was chosen for Quadrature Signal Generation (QSG). Inverter control was achieved using the cascaded PI control. A Phase Locked Loop (PLL) system was implemented to track the grid frequency and set the reference values from the deviations of grid frequency and voltages. The conventional inverter control may not address the grid imbalances during faulty conditions, where negative component plays a key role. Thus, we go for positive and negative sequence extraction in this topology. Reference frame transformations were implemented on voltage and current to set the reference values for power, voltage and current.

The outline of paper is as follows: section II briefly describes the design of the charging station, section III illustrates the controls used, section IV depicts the simulation results to prove the efficacy of the proposed controller and section V presents the conclusions.

II. FAST CHARGING STATION DESIGN

The configuration of proposed XFC Station was shown in Figure 1. The charging station is comprised of 3 main stages, 1. 2L- Voltage Source Converter, 2. Dual Active Bridge Converter and 3. DC DC Buck Boost Converter. The converter is connected directly to a 12.47 kV MV grid, through an LCL Filter to filter the harmonics.

A. LCL Filter Design

The design of LCL filter is very critical in order to filter the harmonics produced within the grid interfaced system. We have many methodologies to design an LCL filter and we discuss one such method that’s suitable for large scale grid connected systems as shown in the reference [2]. The base impedance ($Z_b$) and base capacitance ($C_b$) are calculated as,

$$Z_b = \frac{E_n^2}{P_n} \quad (1)$$

$$C_b = \frac{1}{\omega_b Z_b} \quad (2)$$

Filter capacitance is designed in such a way that, power factor variation is within 5%,

$$C_f = 0.05 C_b \quad (3)$$

Inductor value at the grid side ($L_1$) is designed by assuming a current ripple of 10% of rated value. The equations of maximum current and inductor ($L_2$) are given below,

$$I_{max} = \frac{P_n \sqrt{2}}{3 V_{ph}} \quad (4)$$

$$\Delta I_{L_{max}} = 0.1 I_{max} \quad (5)$$

$$L_2 = \frac{V_{dc}}{6 f_{sw} \Delta I_{L_{max}}} \quad (6)$$

$$L_1 = r L_2 \quad (7)$$

where, $V_{dc}$ is the dc link voltage, $f_{sw}$ is the switching frequency, $P_n$ is the rated active power, $V_{ph}$ is the phase voltage on ac side, the constant $r$ is the ratio between inductance at the grid side and the converter side.
B. Voltage Source Converter

A simple 2L VSC is connected to the power grid through the LCL filter as shown in Figure 2. The switching pulses are fed to the IGBT or MOSFET switches from Sinusoidal pulse width modulation block.

![Fig. 2. 2-Level Voltage Source Inverter](image)

C. Dual Active Bridge Converter

DAB is a bidirectional, Controllable, dc dc Converter that has high power capabilities [3]. It has 8 semiconductor switches, a high frequency transformer and dc Link capacitors. The DAB forms an important stage in the configuration and due to its symmetry, it enables the bidirectional power flow from the Grid to Vehicle and Vehicle to Grid. It also provides galvanic isolation between ac grid and dc Bus.

The topology is as shown in Figure 3, where $V_{in}$ and $V_{out}$ are DC Link voltages on either sides of the converter, $L_k$ is the magnetizing Inductance.

![Fig. 3. Dual Active Bridge Converter](image)

1) Inductor Sizing: The inductor sizes are dependent on frequency. As shown in the equation below, the magnitude of energy transfer inductance is directly related to $V_{dc}$: dc link voltage, $f_{sw}$: switching frequency, $d$: duty ratio, $P_{max}$: maximum power output and $n$: no of turns of the transformer. The expression of magnetizing or energy transfer inductance is given below,

$$L_k = \frac{(1 - |d|)dV_{in}V_{out}}{2f_{sw}nP_{max}}$$  \hspace{1cm} (8)

D. DC-DC Buck Boost Converter

DC-DC Buck boost converter operates in either charging mode or discharging mode at an instant of time depending on the grid conditions. It consists of 2 IGBT Switches that are operated by complimentary control signals to ensure bidirectional power flow. A PI controller is used to ensure the value of current fed to the energy storage to be 870 A.

III. CONTROL SYSTEMS USED

Effective control methodologies were used at each stage of the charging station to obtain optimum results and maintain power quality. Converter control ensures a smooth power exchange between ac grid and dc bus by maintaining the active and reactive Power close to the nominal values. DAB Control helps to obtain ripple free dc Voltage by changing the phase shift corresponding to the perturbations, using Phase Shift Modulation (PSM) technique [4].

A. Converter Control

The voltage and current control loops were implemented in $dq$ reference frame by employing Clarke and Park Transformation as in [5]. Dual current Control scheme is implemented, which controls both positive sequence and negative sequence currents simultaneously [6]. The transfer function of DSOGI-QSG scheme is shown as

$$D(s) = \frac{v'}{v}(s) = \frac{k\omega s}{s^2 + k\omega s + \omega^2}$$ \hspace{1cm} (9)

$$D(s) = \frac{qv'}{v}(s) = \frac{k\omega^2}{s^2 + k\omega s + \omega^2}$$ \hspace{1cm} (10)

Where, $\omega$ is the resonant frequency, $k$ is the damping factor.

These quadrature signals will be converted into positive and negative sequence voltages using the following simple arithmetic operations,

$$V_{\alpha}^+ = \frac{V_{\alpha} + qV_{\beta}}{2}$$ \hspace{1cm} (11)
$$V_{\beta}^+ = \frac{V_{\beta} - qV_{\alpha}}{2}$$ \hspace{1cm} (12)
$$V_{\alpha}^- = \frac{V_{\alpha} - qV_{\beta}}{2}$$ \hspace{1cm} (13)
$$V_{\beta}^- = \frac{V_{\beta} + qV_{\alpha}}{2}$$ \hspace{1cm} (14)

The quadrature signals obtained from the above transfer function are fed to Synchronous Reference Frame Phase Locked Looped(SRF- PLL) system, which allows for a proper adaptation of resonant frequency to the actual network conditions under grid frequency variations [7]. The structure of DSOGI based PLL is shown in the Figure 4. We have PI Con-
reference frame, rotating in clockwise direction. The proposed methodology is as shown in Figure 5 and Figure 6. Also, $dq$ decoupling terms, $\omega L$ and feed forward voltage signals are added to obtain more precise reference values of voltage in $dq$ frame. Reverse transformation of $dq$ to $abc$ axis is applied and fed to sinusoidal pulse width modulation block to obtain the gating signals.

![Fig. 5. PI Controller in positive reference frame](image)

![Fig. 6. PI Controller in negative synchronous reference frame](image)

B. Dual Active Bridge Control

An accurate modeling of DAB is necessary to maintain a stable dc Output of 1150 V, 870 A, thus making it a 1 MW Charging port. Figure 7 shows the schematic of PI Controller used in order to maintain a stable dc Output voltage. The values of proportionality constants are obtained from the transfer function derived based on Generalized Average Modeling(GAM) technique to represent their high frequency ac Conversion stage as shown in [8]. The transfer function is derived from the following equations,

$$\frac{d}{dt}x = Ax + B\Delta d$$  
$$y = Cx$$  

where,

$$x = \begin{bmatrix} \Delta v_o0 & \Delta i_{1R} & \Delta i_{1I} \end{bmatrix}$$  
$$y = \begin{bmatrix} \Delta v_o0 \end{bmatrix}$$  

$$A = \begin{bmatrix} -\frac{1}{R_o} & -\frac{4\sin(D\pi)}{\pi L_o} & -\frac{4\cos(D\pi)}{\pi L} \\ \frac{2\sin(D\pi)}{\pi L_o} & \frac{\pi C_o}{\pi L} & 0 \\ \frac{2\cos(D\pi)}{\pi L} & -\omega_s & \frac{-R_s}{\pi L} \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{\pi C_o}{2V_o0} \frac{2V_o0\cos(D\pi)}{2V_o0\sin(D\pi)} \end{bmatrix}$$  

$$C^T = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

![Fig. 7. PI Controller for Vdc](image)

The values of various parameters used in DAB converter are presented in Table I.

### IV. SIMULATION RESULTS

To validate the proposed model, simulation of each stage was individually created within MATLAB and PLECS environment. The actual circuit involving all the parameters and thus high frequency power electronic switching is modeled within PLECS and the entire control circuit is created within MATLAB software.

The 2L-VSC is fed with a 12.47 kV three phase ac voltage. The dual loop current control is implemented to ensure a dc voltage of 22 kV and dc current of 45.45 A, thus providing a 1 MW power to the DAB. Three phase ac voltage at the grid, dc voltage and current fed from VSC to DAB are obtained in the following results,

![Fig. 8. $V_{abc}$ from the grid](image)

The DAB is separately modeled without integrating with the 2L-VSC assuming the ideal input conditions. Hence, a constant current source of 45.45 A is connected across an input capacitor charged to 22 kV and the voltage loop is modeled with the help of PSM techniques to obtain desired output voltage of 1150 V and current of 870 A. The simulation results
obtained were very close to the reference values as shown in Figure 11 and Figure 12. The rated power of 1 MW is obtained as shown in Figure 13.

V. CONCLUSIONS

This paper presents a model suitable for future grid studies during charging transients. The proposed power controller based on DSOGI-PLL takes into account the changes in negative sequence component of voltages and currents under unbalanced grid conditions. The Phase locked loop is integrated with the control circuit to track frequency variations from time to time. The proposed model not only accounts for the challenges imposed while enabling fast charging, but also improves the reliability of the main power grid by providing power from the energy storage back to the grid under faulty conditions.

The efficacy of system design and the controllers used can be seen in the section IV. Future work regarding the project would be testing the model under faulty conditions, accounting for switching transients, complete the circuit and its model from grid side up to the energy storage. This paper provides a deep insight for any researcher working in the field of fast EV charging.

VI. ACKNOWLEDGEMENTS

This project was supported by Intelligent System Center (ISC) at Missouri University of Science and Technology. I would like to express my deep gratitude to Dr. Jonathan W Kimball, my advisor for giving me an opportunity to work on this project. His valuable and constructive insights were of a great help for the development made in the project. His patience and willingness to help throughout the course of research has been very much appreciated.

I am thankful to Dr. Mehdi Ferdowsi and Dr. Rui bo for serving as committee members, sparing their time in reviewing my work and providing valuable suggestions. I would like to thank my lab mate Kartikeya Jayadurga Prasad for his timely help at various stages of the project.

Lastly, I would like to acknowledge the constant support and love from my parents. My progress wouldn’t have been so successful without them.
REFERENCES


Effect of Virtual Inertia on DC-Link Capacitor Condition Monitoring in a PV-Statcom

Sara Yazdani and Mehdi Ferdowsi

Abstract—The reliability of power electronic converters is dependent on the efficient performance of DC-Link capacitors. Accordingly, the application of techniques to monitor their operation is essential since these elements are quite susceptible to failures as they are exposed to most stringent conditions. On the other hand, the concept of virtual inertia is widely applied to transform the energy stored in DC-Link capacitors to compensate the kinetic energy formerly produced by synchronous machines. In this paper, a new technique based on imposing a virtual inertia control on the outer loop of the dc-bus controller along with a Model Reference Adaptive System has been implemented. This new technique provides an online estimation of DC-Link capacitor value which can serve as a candidate to enhance the reliability of DC-Link capacitor operation on the grounds of effective energy transformation. The performance of the proposed control scheme is verified in FPGA hardware cosimulation through several case studies.

Index Terms—Capacitors, Model Reference Adaptive System (MRAS), Condition Monitoring, Static synchronous compensator (STATCOM), Hardware-in-the-Loop (HiL)

I. INTRODUCTION

By the penetration of more PV-based power converters in modern power systems, various technical provisions and grid codes have been established to maximize their power production capacity. Recently, the distinctive concept of utilizing PV solar farms as STATCOM during nighttime has been proposed to support grid functions as well as generating energy during daytime and has attracted a considerable attention. [1] and [2]. PV-STATCOM is basically a Voltage Source Converter (VSC) and is more utilized to increase the connectivity of neighboring wind farms and enhancing the power transmission [3] and [4].

Along with all the benefits, PV solar panels are considered as asynchronous sources and they can affect the electromechanical modes of the remaining synchronous machines in a large interconnected power system. As PV arrays do not possess any mechanical rotations, they lack the inertial response to frequency disturbances which may result in displacing synchronous machines and thereby affecting the modes of operation. Therefore, PV panels are required to provide grid support services such as inertial response in a similar way to conventional synchronous generators. Several virtual inertia based control is proposed for PV arrays to enhance the inertia of hybrid PV-battery DC-Link [5].

DC-Link capacitors are the significant components used to shape the DC-Link in power electronic converters. As a passive device, they are widely applied to minimize the voltage fluctuations in the dc-link and balance the instantaneous power difference caused by load switching or provide sufficient energy during the hold-up time. In the solar market, capacitors play a critical role to convert the DC power produced by solar cells into AC power. Inverters typically make extensive use of large-sized capacitors that store electricity. However, along with semiconductor switches, capacitors are the most reliability vulnerable power electronic components [6]. Failure rate of capacitors in field operation of power electronic systems is fairly high [7]. The compulsion in cost reduction of capacitor design and their exposure to coarse working environment adds up to the challenges of designing capacitors with built-in reliability facilities. Moreover, effective dc-link design solutions and condition monitoring techniques on the present capacitors is indispensable to prevent maintenance and secure the reliability.

A proper control design on the stored energy in the DC-Link capacitor can help to make a virtual inertia response by varying the DC-Link voltage during frequency disturbances [8]. In this paper, we have applied a similar methodology in which the stored energy in the DC-Link capacitor is utilized by varying the voltage during disturbances. A current control methodology augmented by power regulator has been developed which can considerably mitigate the short-term oscillations without incorporating a massive storage device.

In addition, as the capacitors failures are mainly caused by gradual deterioration which leads the capacitor to wear-out, an estimation method based on model reference adaptive system (MRAS) has been proposed to monitor the capacitance value during the operation. In this method, the voltage of the DC-Link is measured through the sensors while the reference voltage of the Vdc is calculated based on the energy expression of the DC-Link capacitor for a particular power rating. The virtual inertia (J) of the system is user defined and by considering the normal operation of the controllers the DC-Link capacitance is estimated. The difference between the output of the estimator and the nominal value will give out the instantaneous value of the dc-bus capacitor without interrupting the control action. This method is not dependent on the type of the capacitor and it does not require the extraction of the low frequency and switching frequency components of the capacitor[9]. The advantage of this method is that it directly calculates the capacitance and its accuracy is not dependent on the voltage and current components as they are not used in the calculation process.

The rest of the paper is organized as follows: In section II, the operation principle of STATCOM and synthetic inertial control of the DC-Link is discussed. Section III addresses the MRAS structure and the estimation process of the DC-Link capacitance and the stability analysis. Cosimulation results and discussions are presented in Section IV and the conclusion is included in Section V.
II. VIRTUAL INERTIA BASED CONTROL OF PV-STATCOM

Basically, as STATCOM is treated as a voltage source converter, its potential to play a role in frequency regulation has been neglected. However, the electric energy stored in the STATCOM DC-Link capacitor has a large electric energy storage that can help to significantly improve the inertial response. An appropriate control design can extract the energy of the DC-Link and compensate for the kinetic energy required to enhance the frequency response of the system. STATCOM is normally connected to weak grid bus at PCC and the whole system is composed of three main parts: an inverter, a coupling inductor and a controller as Fig. 1. By adjusting the amplitude and the phase angle of the converter output voltage by dividing it into capacitive and inductive mode, the power exchange between the inverter and the power system is obtained. Adjusting the magnitude of the output voltage can control the reactive power and adjusting the phase angle controls the real power transferred Fig. 2.

As PV-STATCOM has no real inertia, the synthetic inertia, known as "Virtual Inertia", is embedded into the control structure to emulate the inertial response of a synchronous machine. The DC-Link in STATCOM is responsible to make the inertial response of the system. Therefore, by controlling the DC-Link voltage in a way that the capacitor energy emulates the kinetic energy of a synchronous machine, the frequency response enhances.

DC-Link voltage variations in PV-STATCOM are related to the irradiation changes in daytime. However, the minimum voltage rating cannot be less than 0.9 p.u. and the maximum voltage cannot exceed the voltage ratings of the power semiconductor devices and DC-Link capacitors. By considering that each phase has its own separate capacitor, the power generation is determined by the following equation:

\[ P_m - P_e = \frac{d}{dt} \left( \frac{1}{2} CV_{dc}^2 \right) \]  
\[ (1) \]

where, \( P_m \) is the generated power and \( P_e \) is the demanded power. The right side of the equation is the energy stored in the DC-Link capacitor. By considering the fact that the generated power is equal to the derivative of the kinetic energy stored in an equivalent synchronous machine, the above equation turns out to be:

\[ P_m - P_e = \frac{d}{dt} \left( \frac{1}{2} JS\omega^2 \right) \]  
\[ (2) \]

By equating (1) and (2), the reference \( V_{dc} \) voltage will be calculated as:

\[ V_{dc} = \omega \sqrt{J/C} \]  
\[ (3) \]

As grid code has set a stringent frequency drop rate 0.1-1 (Hz), the rate of variations in \( V_{dc} \) is determined by the rate of
change in the inertia of the system. As it can be figured out from Figs. 3 and Fig. 4, the variations are in a quite a small range and therefore, the estimation of the capacitance can be derived from the equations discussed above.

The overall structure of the controller has been depicted in Fig. 5. The control structure has been implemented in $dq$ reference frame. The first loop is a control loop over the reactive power which ends up in producing the reference voltage in $q$ axis. The second loop is a control loop over the DC-Link voltage which is augmented by a loop that controls the active power and it calculates the reference voltage in $d$ axis. In the next section, the MRAS structure and the estimation of the DC-Link capacitance and the engagement of the virtual inertia in the control of $V_{dc}$ is discussed.

III. MRAS STRUCTURE AND $V_{dc}$ CAPACITANCE ESTIMATION AND STABILITY ANALYSIS

In this paper, a direct MRAS methodology has been applied to estimate the capacitance of the DC-Link voltage. By utilizing this method, the estimated capacitance gives the capability to monitor the DC-Link capacitor performance and prevent unprecedented failures of operation. The philosophy behind model reference adaptive control is to create a closed loop controller with parameters that can be adjusted based on the error between the output of the system and the desired response of the reference model asymptotically with time for any bounded reference input signal. In this methodology, certain information about the plant (here the $V_{dc}$) is used to directly find the convergent ways to adapt the controller parameters. In this way, the tracking errors in the presence of uncertainties and variations in plant is reduced.

In this study, by considering $V_{dc}$ as the reference model, the output of which is processed through a compensator and it produces $\sqrt{\frac{J}{C}}$. When this value is multiplied by $\omega$ will give out the $V_{dc}$ reference. This loop will be stabilized if and only if the estimated capacitor value and the real value of capacitance are the same.

In order to prove the convergence of this estimation, in this paper, by proving that the overall system is passive due to the non-autonomous nature of the overall estimation scheme, it is shown that the estimator converges to the real value of DC-Link capacitor. The prove of passivity comes as follows.

A simplified diagram of MRAS is presented in Fig. 7. Considering that structure, the overall closed loop structure is given by:

$$\frac{Y}{V_{dc}} = \frac{K_i}{K_i + s}$$

(4)

Considering the above, the state space model of the system is given by:

$$\dot{x} = -K_i x + K_i V_{dc}$$

(5)

By choosing the Lyapunov function as $V = \frac{1}{2} x^2$, the derivative will be:

$$\dot{V} = x [-K_i x + K_i V_{dc}]$$

(6)

From (6), it can be very well proven that $\dot{V} \leq uY$, where $u = V_{dc}$ and $Y = x$. This shows that the system is passive and therefore it is globally asymptotically stable which proves the convergence of the estimator. More detailed explanations can be found in [10].

This passivity based stability analysis proves that this control strategy is sufficient to make a control package which not only provides the system with appropriate frequency response, but also it provides a mechanism to monitor the $V_{dc}$ capacitance which adds up to the reliability of the system. As long as this control method is not sensitive to the range of the parameters of the system, like the inductive characteristics of capacitors in different frequencies, this algorithm is unlikely to fail in abnormal measurement noises. By utilizing this control strategy, PV-STATCOM becomes can provide possible competitive inertial energy and along with the suitable frequency response, it is also supported by a reliability provision to
monitor the DC-Link capacitor behavior. In the next section hardware cosimulation results further prove the accuracy of this control methodology.

IV. HARDWARE COSIMULATION RESULTS AND DISCUSSIONS

Real time implementation of the proposed control scheme using Xilinx System Generator for DSP as a design tool to define, test and implement high-performance DSP algorithms plays an important role in rapid prototype hardware development of any control system. This discretized version of the controller is realized in Xilinx System Generator environment and later on it is implemented on FPGA, ZedBoard Zynq-7000 Development Board. Using this virtual environment to test the control performance before applying it to the target plant can replace risky or expensive physical testbenches. Also, the inherent parallel processing capability of the FPGA helps in accelerating the performance of the controller in discrete domain. In this work, the joint test action group, (JTAG) cosimulation interface has been used as it requires minimum resources on FPGA to achieve an optimum hardware cosimulation interface. The FPGA clock frequency is chosen to be around 120 MHz at which all modules are found to be compatible with each other. As the whole control structure was implemented in System Generator, the switching circuit was realized in PLECS. Logical blocks have been used to design the control and the sinusoidal waveforms were generated using Cordic algorithm. Floating point representation has been utilized for successive computations. This method does not have the cumbersome drawbacks of fixed point calculations. A suitable pipelining method overcomes the unwanted delay effects on the results. The discretized version of the controller has been realized by using the Tustin or bilinear approximation which yields the best frequency-domain match between the continuous-time and discretized systems.

The results in Fig. 8 demonstrate the control action on the DC-Link voltage and it shows perfect tracking between the reference and the actual value. In Fig. 9, the estimation of the capacitor value is shown. It has converged to the actual value in quite a short time. It shows that the estimation technique has an accurate performance. The results for reference and actual values of current in $d$- and $q$-axes confirm the control methodology in tracking the reference values. This also goes similar for the reactive power. In this control scheme the grid voltage and current have very low THD and show sinusoidal stability Fig. 13. For further verifications, the results for the voltage and current have been taken from a DAC and have been in an oscilloscope.

V. CONCLUSION

A virtual inertia based control methodology has been augmented with an adaptive estimation methodology (MRAS) to provide a control package to give inertial response to the frequency oscillations and also monitor the behavior of the
DC-Link capacitor in PV STATCOM. The capacitor condition monitoring method has been verified through passivity based stability analysis and it has been proved that the control structure can boost the reliability of the system regardless of the impacts of environmental disturbances like temperature and it give the capacitance without relying on the inductive characteristics of the capacitor in some switching frequency ranges or the current that passes through the capacitor. The future can focus on applying this method on more delicate topologies under different environmental conditions.

**REFERENCES**


Microscale Dynamics and Mechanisms of Flash Sintering Process revealed by in-situ x-ray imaging technique

S. Mohammad H. Hojjatzadeh\textsuperscript{a}, Dr Lianyi Chen\textsuperscript{a,}\textsuperscript{b}, Dr. Fatih Dogan\textsuperscript{b}

\textsuperscript{a}Department of Mechanical & Aerospace Engineering
\textsuperscript{b}Department of Materials Science and Engineering,

Emil addressees: sh7z7@mst.edu (S. Mohammad H. Hojjatzadeh), chenliany@mst.edu (Lianyi Chen)
doganf@mst.edu (Fatih Dogan),

Abstract

Sintering is a crucial processing step for obtaining the necessary mechanical stability and rigidity of ceramic bulk materials. Flash sintering is a recent technique to quickly densify ceramics in the presence of an electric field by which full densification can occur at a very short time. However, the underlying mechanism of flash sintering of ceramics is still puzzling. Here, we report the in situ x-ray imaging of flash sintering of 3YSZ (yttria-stabilised zirconia). We have analyzed the microscale dynamics of ceramic particle sintering behavior during flash sintering, and tested the mechanisms proposed previously about flash sintering.

Introduction

Flash sintering is a process in which an electric field is applied to a sample while it is heated in a conventional furnace and the densification often takes place at a much lower temperature and at a densification rate orders of magnitude higher than conventional sintering methods\cite{1}. This process was first reported by Raj et al. in 2010 \cite{2}, and named “flash sintering”. Initially demonstrated for 8-mol\% yttria-stabilised cubic zirconia (8YSZ), flash sintering have been successfully applied to other materials such as zinc oxide (ZnO\textsubscript{2}), alumina (Al\textsubscript{2}O\textsubscript{3}), cobalt manganese oxide (Co\textsubscript{2}MnO\textsubscript{4}), titanium oxide (TiO\textsubscript{2}), and strontium titanate (SrTiO\textsubscript{3})\cite{3}.

Flash sintering results in significantly less grain growth compared to conventional sintering methods which can offer nanostructured ceramic material with higher strength and toughness \cite{4}. Flash sintering, therefore, has a great potential for the manufacturing of high performance ceramics and novel materials, at a high production rate and significantly reduced energy consumption.

However, the fundamental mechanism of the flash sintering process is still not fully understood, which limits its further development. Different mechanisms have been proposed to explain the flash sintering event: (1) increase of diffusivity at particle boundaries due to rapid temperature rise at particle-particle junctions caused by joule heating\cite{5}, (2) liquid film capillary mechanism (rapid densification is explained by particle rearrangement led by spreading of the liquid due to local attractive capillary forces)\cite{6}, (3) nucleation and avalanche of
Frenkel pairs (vacancy-interstitial pairs)\[7\], (4) reduction of flow stress required for plastic deformation to nearly zero under applied electrical field \[8\], (5) reduced rate of grain growth under the applied electrical field, (6) increased diffusion kinetics, due to interaction between the external field and the space charge field\[9\]. It is still not clear as to which mechanism (or a combination of mechanisms) are responsible for observing flash sintering event.

So far, the reported data that are collected during the flash sintering process, are macroscopic data that include dimensions of samples (which were used to calculate shrinkage, shrinkage rate, relative density, densification rate), current and voltage (which were used to calculate conductivity, power, power density), electroluminescence, photoluminescence, temperature of the furnace and sample, and x-ray scattering pattern. These macroscopic data only show average information about the samples during flash sintering, which could not fully prove or disprove the proposed mechanisms for flash sintering. The microscale data on how individual particles sinter together, and how pores disappear during flash sintering are in high demand. Unfortunately, the non-transparency of the samples to visible light and the very short duration of the flash event pose a huge challenge to the characterization of this process in microscale.

The objectives of this proposed research is to characterize the microscale dynamics of particle sintering behavior during flash sintering and test the mechanisms proposed previously about flash sintering, especially the liquid film capillary mechanism.

**High speed X-ray imaging technique**

High-speed (up to 1 MHz) high-energy (which can penetrate more than 5 mm thick alumina) synchrotron radiation x-ray imaging at the Advanced Photon Source in Argonne National Lab was used to characterize the microscale dynamics of the flash event during flash sintering. The schematic of the x-ray experiment setup is shown in Fig. 1. The details of high speed x-ray imaging technique has been explained elsewhere \[10,11\]. The ceramic powder used to study flash sintering was 3YSZ (yttria-stabilized zirconia) with an average particle diameter \((d)\) of 120 \(\mu\)m. Experiments were performed under laser irradiation with and without applying the electric filed under argon gas atmosphere. During the experiments the laser diameter \((D4\sigma)\) was larger than 400 \(\mu\)m. The frame rate of 50 kHz was used to capture the details of particles sintering.

**Fig. 1.** Schematic of high speed synchrotron x-ray imaging technique used to study the microscale dynamics of flash sintering.
Sintering without applying electric field
We first performed a series of laser sintering experiments without applying electrical field to identify the role of electric discharge in flash sintering process.
When the exposure time and laser power ($P$) were 6 ms and 20 W, respectively, we observed only a negligible movement of the particles within powder layer and no particle sintering was identified. Fig. 2 shows a sequence of x-ray images when exposure time was increased to 10 ms (under laser power of 20 W). After 1.5 ms, one melted particle is observed to fly away into the gas atmosphere due to laser localize heating and creation of vapor jet between powder particles [11]. While droplet is flying, the internal gas pore (marked by red dotted circle at $t_0 + 1.5$ ms) expands due to continuous laser heating and bursts outward, ejecting several smaller droplets and creates a cavity inside the droplet ($t_0 + 2$ ms). After 2.5 ms, the droplet comes back into the powder layer again. Because the droplet view is blocked by frontier particles in the powder bed, we performed image processing using ImageJ[12] to identify its further interaction with nearby particles in the powder bed (Fig. 3).
As Fig. 3 shows, the droplet (outlined with black dotted circle) starts moving within the powder bed until starts merging with another melted particle after 2 ms (outlined with blue dotted circle). Sintering of two spherical droplets completed after $\sim$3.5 ms and consequently a bigger droplet with the diameter of $\sim$160 μs forms. This experiment reveals that laser heating without applying an electric filed was enough to melt the 3YSZ particle to near its melting point (2973 K) and to form a bigger droplet after $\sim$8 ms.

Sintering assisted with electric field (flash sintering)
Figure 4 shows the experiments under laser power of 20 W and applied electric field of
25 V. Under localized laser heating, the particle (outlined with red at \( t_0 \)) starts to melt and a liquid droplet will form after 0.3 ms. The droplet then starts to spread over nearby particles as a result of capillary mechanism and wettability, after 1 ms [13]. After \(~1.2\) ms, a droplet completely wets two nearby particles, melt them and a bigger droplet forms. Simultaneously, pores formed into the droplet can coalesce, expend or escape from the liquid droplet. Interestingly, when electric field exists, the primary melted droplet do not maintain the shape of a sphere and easily flow over nearby particles and wet them (compare Fig. 4 with Figs. 2 and 3). This shows the formation of a temperature surplus in the powder particle and dramatic decrease in viscosity of the droplet by applying the electric field and has the implication that the rapid heating in flash sintering is due to temperatures exceeding the conventional sintering temperature by several hundreds of degrees [9].

In flash sintering, only grain boundaries might be locally heated to melting temperatures when the highest electric fields are applied [14]. However, as laser heating is very localized, it is extremely challenging to avoid partial or complete melting of the particles, regardless of using a very large laser beam size (> 400 μm). To avoid and mitigate the melting of the particles, we tried to decrease the laser power to a minimum of 15 W under the electric field of 25 V (Fig. 5). Even though it takes longer for particle to form a liquid droplet which shows mitigated heating (3 ms compared to 0.3 ms) fully melting of the particles cannot be halted and still particles spreads over the nearby particles. As Fig. 5 shows, after 6 ms the droplet is in complete contact with the other two nearby particles as a result of wettability, and finally a bigger droplet combining four melted particles forms after 8 ms.

![Fig. 4. X-ray images showing the particles consolidation and droplet formation induced by capillary mechanism (\( P=20 \) w, \( E=25 \) V, \( X_p=10 \) ms).](image)

![Fig. 5. X-ray images showing the sequence of particles consolidation and formation of droplet induced by capillary forces (\( P=15 \) w, \( E=25 \) V, \( X_p=10 \) ms).](image)
The sliding of a solid particle over the liquid droplet surface due to surface tension is an important step in formation of the droplets and densification procedure in these experiments. This phenomena has been shown in Figure 5 at the time frame of $t_0 + 6$ ms, where a particle (marked by green dotted circle) is pulled over the surface of the droplet in the direction of the blue arrow. The particle sliding has also been observed when electric field was not applied. As shown in Fig. 6, a solid particles is dragged over the liquid surface of another particle caused by capillary forces.

![Image]

**Fig 6.** X-ray images showing the particle sliding over the liquid bridge formed on the surface of another particle ($P=15$ w, $X_p=10$ ms).

The proposed mechanism for ceramic particles coalescence

Figure 7 shows the overall mechanism of particles coalescence in this research. Particle coalescence under laser heating and electric fields starts by heating and melting of particle(s) (stages I and II). The melted particle then forms a liquid droplet with low viscosity and spreads over nearby particles induced by capillary action (stages III and IV). If laser heating and electric field are stopped, the capillary process also stops and the formed droplet rapidly becomes solid.

![Diagram]

**Fig 7.** Mechanism of particles coalescence in this research.

**Conclusions**

High speed x-ray imaging technique was utilized to study the flash sintering of ceramics powder; the following results were achieved:

- Laser irradiation of particles without applying electric field caused the formation of spherical droplets. The driving force of formation of a bigger droplet is the tendency to minimize the surface area and is controlled by capillary forces and wettability.
- Applying electric field accompanied with laser irradiation caused faster heating of the particles to the temperature point higher than melting temperate and formed a droplet with a very low viscosity that can easily spread over nearby particles. The main deriving
force for particle coalesce is identified to be surface tension and wettability.
- Because laser irradiation is localized, using the minimum laser power and large beam size could not mitigate fully melting of the particles.
- Further investigations with tuning laser powers to avoid fully melting of the particles are needed to simulate the real flash sintering conditions.

ACKNOWLEDGMENTS
The authors would like to acknowledge the support of the Intelligent Systems Center (ISC) at the Missouri University of Science and Technology (Missouri S&T).

Reference


ABSTRACT
In a battlefield or a disaster affected area where multiple groups operate, secure messages need to be exchanged by store and forward method via mobile nodes that form a delay tolerant network (DTN). In that environment, sending secure messages using public-private key cryptography would work only if every mobile node had all other nodes’ public keys. However, a trusted key infrastructure authority may not be always reachable or may be compromised to provide and validate a node’s public key as some nodes may also fall in another jurisdiction. In this situation, each node should take care of its own key generation, sharing and validation. Moreover, not any other single node other than the destination node should realize even a part of a secure message. Therefore, the goal is to pass messages securely to ensure this is readable only by the destination node while intermediate nodes help to carry the message forward towards the destination in a secure fashion.

1 INTRODUCTION
In application environments like international military coalitions or multi-party relief work in a disaster zone, passing secure messages using DTNs is challenging because existing public-private key cryptographic approaches may not be always accessible across different groups due to the unavailability of Public Key Infrastructure (PKI). In addition, connectivity may be intermittent so finding the reliable route is also difficult. Thus, instead of sending the complete message in a single packet, fragmenting the messages and sending them via multiple nodes can help achieve better security and reliability when multiple groups are involved. Therefore, encrypting messages before fragmentation and then sending both the data fragments and the key fragments (needed for decryption) provide much higher security. Keys are also fragmented as sending the key in a single packet can hamper security if it is forwarded to some corrupted nodes who may try to tamper or drop it. Hence, in this paper, we develop a scheme to provide improved security by generating multiple key-shares and data fragments and disseminating them via some intermediate nodes. In this fragmentation process, we also create a few redundant blocks to guarantee higher data arrival rate at the destination when message drop rate is higher like in the DTN environment.

Our performance evaluation when compared to the most closely related scheme like Multiparty Encryption [4] shows the improvement on minimizing the number of compromised messages as well as reduced bandwidth consumption in the network. We summarize our contributions in this paper as follows:

- We design a secure data dissemination algorithm based on fragmentation of data as well as keys which are then forwarded through trusted intermediate nodes for secure delivery.
- We show the improved performance of our algorithm through extensive simulations using a real dataset when compared with [4] based on message compromising rate and bandwidth usage and thus, our scheme provides a better security and reliability at the cost of a small increase in latency.
- We also provide integrity check for both key and data fragments, and through simulations, we show that our approach works better in forwarding uncorrupted messages in a highly intermittent DTN network compared to [4].

2 BACKGROUND AND RELEVANCE TO PREVIOUS WORK
For secure communication, to get rid of the necessity of public key, Identity Based Cryptography (IBC) has been proposed. The applicability of using IBC in DTN is discussed in [3]. However, they could not skip the importance of reachability to Private Key Generator (PKG) in their scheme and concluded that IBC is less preferable in energy restricted DTN nodes. [13] introduced distributed ID generation to avoid PKG but each node has to register its ID before being implemented in DTN. In summary, IBC always suffers from the compromised key of a node because it is directly related to the identity of master public key of PKG.

Another way to send private message is by fragmentation so that no node can get the entire message. Cabinas et al. [4] provided chaining algorithm for secure passing of fragmented message to address the issue of unavailability of Public Key Infrastructure (PKI). This algorithm uses the available public keys that a node collects via direct contact with others. Since MPE is close to our scheme, we use it for the performance
comparisons. To make this scheme more secure, a message can be encrypted with a key and then both the message and key are fragmented and sent via different nodes. Using Shamir’s secret sharing [11], it is possible to create several shares of the secret key from which a specific length subset is necessary to rebuild the key. The work in [9] and [8] have considered some recursive approaches to generate secret shares. Sending respective key of a message this way can provide security, but it is vulnerable if those key-shares pass through the same node. To handle this, encrypted data can also be fragmented to reduce this vulnerability because it is highly unlikely that enough key shares and data fragments both will pass through any random single node. Again, a normal fragmentation of data can incur data loss when one or more fragments are dropped due to bandwidth or resource constraints, link failure, faulty path selection or due to TTL limitation. Besides in DTN, seeking acknowledgement of packets is not feasible. Hence, creating multi-copy of a fragment or using erasure coding [5] like Cauchy Reed Solomon (CRS) [10] can be helpful to introduce some redundancy.

3 PROPOSED SCHEME

As we provide security by fragmenting both the keys and the data, we also generate few redundant fragments to help in ensuring higher data delivery in DTN environment where message drop rate could be higher. The overview of our scheme is as follows:

- Our algorithm first generates a random symmetric key for a data file and encrypts it with the key.
- Next it generates \( N \) key shares, fragments the encrypted data file and encodes the fragments with CRS matrix to generate \( N \) data fragments.

**Fig. 1: Key \( K \) is generated for Data \( D \).**

\( K \) with other hidden secrets are used to generate 8 key-shares, \( \{ K_1, K_2, \ldots, K_8 \} \). \( D \) after being encrypted by \( K \) is fragmented and encoded to 6 blocks, \( \{ D_1, D_2, \ldots, D_6 \} \). Finally, all the fragments are disseminated to different random relay nodes. [The number of key-shares and data blocks are independent and can be altered.]

- The key-shares and encoded file blocks are then forwarded to different relay nodes.
- When the DN gets \( N_{\text{key}} \) key-shares and \( N_{\text{data}} \) data fragments \(( K < N) \), it can regenerate the key and encrypted data. It then uses this key to retrieve actual data.

Figure 1 shows an example of the workflow for sending a message in our scheme.

3.1 Handling Key

We generate key-shares using RSS algorithm over Shamir’s scheme as it can hide additional \( (K_{\text{key}} - 2) \) secrets (such as configuration parameters). The \( K_{\text{key}} \) and \( N_{\text{key}} \) are network wide constants. Key-shares are plain packets and any node getting \( K_{\text{key}} \) number of shares are able to generate the key for a message. For this reason, more care should be taken while sending key-shares compared to sending data fragments. We limit the reachability of enough key-shares to any specific node by adopting the MPE fragmentation. The difference is that we **don’t send whole encrypted data along with every key-share** which would take significant bandwidth when data size is large. Again, an intermediate receiver needs to decrypt an encrypted key-share with its private key before encrypting with other IN or DN’s public key, which usually induces heavy computation costs. To reduce the energy consumption for an IN’s service, we use proxy re-encryption [12] that eliminates the necessity of decryption in each step except for the DN. Source creates several proxy re-encryption keys by using its own private key and other chosen INs’ public key. After receiving a key-share encrypted with a proxy re-encryption key built from its public key, an IN also creates a proxy re-encryption key using its own private key and other chosen IN or DN’s public key. This proxy re-encryption key is used to change the domain of the encrypted key-share from one IN to another IN or DN. Finally, only the DN needs to decrypt to realize the key-share. Fig 2 is describing an example to distribute the key-shares. Note that this recursive process of distributing key-shares may incur significant delay for few of them.

![Fig. 2: Key-Share Handling and Distribution via selected Intermediate nodes using Proxy Re-encryption](image-url)
3.2 Handling Data

Data is encrypted with a random symmetric key and then fragmented. The fragmented blocks are then encoded with CRS algorithm. These data blocks are dispersed by choosing different random relay nodes from the source. Note that it may be possible that in some relay node, \( K_{\text{data}} \) data blocks may have arrived and this node can apply Reed Solomon decoding to get the data. However, it can’t understand data as it is encrypted. Thus, it also need all \( K_{\text{key}} \) key shares which is very unlikely to happen for a random intermediate node for the same message.

4 CORRUPTED FRAGMENT

If any of the \( K_{\text{key}} \) shares or \( K_{\text{data}} \) fragments is corrupted and the destination node fails to detect, then, after decoding and decrypting, the destination node will get a faulty message. Though a packet could also get corrupted by a transmission error, more often in DTN, a packet gets corrupted by some malicious nodes or adversaries. In that case, if the destination node could know that some of the packets received are corrupted, waiting for additional correct fragmented packets would provide successful retrieval of the message. We here describe how to check for the integrity and correctness of received packets.

4.1 Integrity of Key Shares

We check the integrity of the key-shares using Merkle Hash Tree (MHT) [7]. As the number of total key-shares is network wide fixed, the structure of the MHT is also fixed and known to all. The idea is to send each key-share with necessary hash value so that the destination node can generate the root of the tree from each of them. In standard integrity check scheme using MHT, the calculated root is compared with the previously sent root which is not possible in our scheme. Therefore, the destination node will wait until it gets \( K_{\text{key}} \) number of shares with the same calculated root for each. We now describe the scheme by an example illustrated in figure 3.

Let the source node generates 8 key-shares, \( K_1, K_2, \ldots, K_8 \) by RSS algorithm. It now calculates hashes of each key-share as \( H_1, H_2, \ldots, H_8 \). Every other node of that tree is generated by the hash value of the concatenation of its children. Now source will disperse the shares with necessary hash value appended with them so that the destination node can generate the root. For example, for key-share \( K_1 \), \( [H_2, H_{3,4}, \text{and } H_{5,8}] \) are appended. Again, for key-share \( K_6 \), \( [H_5, H_{7,8}, \text{and } H_{3,4}] \) are appended. When the destination node gets \( K_{\text{key}} \) number of shares with the same calculated root value for each, it starts reconstructing the secrets.

4.1.1 Integrity of Data Fragments

The approach to check the integrity of data fragments is signing them in the source node and verifying the signature in the destination node. However, it requires source’s public key in destination. We can send this public key as another secret of the key shares but size of the public key should be small enough to send it as a secret. For this purpose we are using elliptical key cryptography (EC) in our scheme. The source node signs each data block and sends that signature by appending with the data block. Upon receiving \( K_{\text{key}} \) correct key shares, destination node can extract the public key of a source node. It will then discard every data block it received for which verification returns false with that key. If the verified correct data block number is less than \( K_{\text{data}} \), it will wait for new data block and verify each block until the total number of non-corrupted block is \( K_{\text{data}} \).

5 PERFORMANCE EVALUATION

We performed the simulation of our scheme using haggle-one-infocom2006-complete dataset [2] on ONE simulator [6] for 222 hours operating on 98 nodes and 152 jpg files of varying sizes. We used Spray and Wait routing protocol in binary mode with 6 copies of every packet to disperse them throughout the network. We also implemented MPE fragmentation algorithm on same configuration to compare with our scheme. We vary the percentage of malicious nodes and public key possession of nodes to show the performance variations. The parameters used in the simulations is listed in table I.

<table>
<thead>
<tr>
<th>( K )</th>
<th>32 byte random symmetric key</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>512KB, 1MB, 2MB, 4MB jpg image file</td>
</tr>
<tr>
<td>( K_{\text{key}}, N_{\text{key}} )</td>
<td>4, 8</td>
</tr>
<tr>
<td>( K_{\text{data}}, N_{\text{data}} )</td>
<td>8, 12 (for static scheme)</td>
</tr>
<tr>
<td>Data block size</td>
<td>512KB</td>
</tr>
<tr>
<td>Public Key</td>
<td>160bit elliptical key</td>
</tr>
</tbody>
</table>

TABLE I: Parameters used in the ONE simulator

We show two scenarios based on the nature of malicious nodes. First, to compare with the MPE scheme, we assume that the malicious nodes are semi-honest. They work together for an adversary to learn the entire message but do not corrupt any packet. Second, to emphasize the importance of integrity check in our scheme, we assume that the malicious nodes corrupt every packet relayed through it.

5.1 Comparison with MPE

Figure 4 shows the percentage of the compromised messages with varying percentage of malicious nodes. The number of compromised messages in our scheme will always be less than MPE because when a message is corrupted in our scheme, it means at least \( K_{\text{key}} \) number of key-shares and \( K_{\text{data}} \) number of data fragments are relayed through the malicious nodes and hence that message is already compromised in MPE scheme, but the opposite is not true. Hence, we have almost 30% less compromised messages.

Figure 5 shows the comparison of bandwidth consumed in the DTN network. When the message size is bigger, more bandwidth is needed to send data from source to destination.
Fig. 4: Percentage of compromised messages with varying percentage of semi-honest malicious nodes. (Here, the file size is 2MB and public key possession is 50%)

Fig. 5: Bandwidth (KB) for varying message size.

Here, we can see that MPE scheme consumes three times more bandwidth than our scheme for small files, and five times more for larger files. Bandwidth consumption in static scheme is little higher than dynamic scheme specially for shorter messages because the value \(N_{\text{data}} - K_{\text{data}}\) in former one is greater.

Figure 6 shows the delay distribution of receiving a message in the destination. This distribution has a long tail because after most of the messages arrive, less number of fragments are scattered throughout the network causing the remaining messages being delayed more. However, as the number of packets created is greater in our scheme, the average delay is around 0.5-1.2% longer than MPE. Again, as the number of packets required to generate a message in the Static scheme is always a few greater than Dynamic scheme comparatively for short messages, Static scheme incurs more delay. This delay can be reduced by applying more copies of packets which would increase bandwidth consumption or by decreasing \(K_{\text{data}}\) with a trade-off against security.

Figure 7 shows the computation energy needed for different tasks in our scheme such as key-share generation and reconstruction using RSS, data fragments encoding and decoding by CRS, cryptographic operations and integrity check feature. We measured the execution time for our application and then calculated the energy usage according to the power analysis of ARM Cortex-A53 [1]. It shows the CPU power consumption is 269mW when the CPU speed is 1,300MHz. Note that in MPE scheme, cryptographic computations with additional decryptions and key-shares related calculation are only performed as data is not fragmented. The energy consumed in our scheme is almost same as MPE (without the integrity check implementation). Again, with the integrity check included, which is an additional feature of our scheme to guarantee untempered data delivery, energy consumption is around three times compared to MPE. We can apply probabilistic integrity check to reduce that huge energy consumption. For example, 33% random packets will be checked to reduce to 1/3 energy.

5.2 Corrupted Node Effect

From figure 8, we can see that the delivery ratio rises if the number of fragments per message \(N_{\text{data}}/K_{\text{data}}\) increases. But for larger messages when \(N_{\text{data}}/K_{\text{data}} = 3\), the delivery ratio decreases. It is because this case causes too much redundancy. Therefore, necessary fragments for unreached messages often are dropped while collecting unnecessary redundant packet. Again, this figure shows, for a small message delivery ratio is little higher than the large message. It is because for a small message, if a very small number of fragments is dropped, the whole message becomes invalid, but this situation allows some more fragments of other small messages to be carried to the destination successfully. And, this small number of replicated fragments increase the possibility of successfully
delivering small messages. But, for a large message, even if a few fragments are missing, there may be enough redundant fragments available. Therefore, for a large message, there may be unnecessary extra fragments whereas for some other messages, missing a few fragments may prevent the message reconstruction at the destination. Note that message delivery is bounded by data delivery only as because data for a message can be reconstructed only when the respective key is available. There might be many other fragments for other messages in the destinations but cannot be reconstructed as data and shown as delivered because they lack respective keys.

Figure 9 shows the benefit of checking the integrity of packets. As the number of corrupted nodes increases, more key-shares or data-fragments are modified or corrupted. A single corrupted packet of a message contributes to generating a completely different faulty message. That is why without the integrity check feature, the corrupted message rate is high. From the figure, we can see that the rate of correct messages received is deteriorating from 52% to around 30% with the increase of corrupted nodes from 0 to 25% when integrity scheme is not present. However, the rate of deterioration is far less (from 52% to 37%) when the corrupted packets are omitted due to the integrity check feature. This deterioration still occurs when corrupted nodes increase because they corrupt more than \( N_{\text{key}} - K_{\text{key}} \) or \( N_{\text{data}} - K_{\text{data}} \) packets so that the it is unable to reconstruct the message. Thus, our approach guarantees to construct error-free messages by successfully skipping the messages that are impossible to reconstruct because of fewer correct fragments than necessary.

6 CONCLUSION

In this paper, we propose a scheme to securely send fragmented messages in a DTN environment through intermediaries where the PKI infrastructure or public key generator and distributor is unreachable either due to the non-availability or a compromised situation. For this purpose, we use RSS (Recursive Secret Sharing) and CRS (Cauchy Reed Solomon) algorithms to fragment both key as well as data and generate redundant packets which can provide better security and data retrieval in a highly intermittent environment where packet drop rate is much higher. We show the significance of our work when compared to the closely related MPE scheme designed for the same purpose with respect to the compromised message rate, bandwidth consumption and the effect on corrupted packets using real data set. In particular, it shows that our scheme works better in successfully delivering more non-compromised messages with reduced bandwidth consumption, in an environment where public key availability is very low.

REFERENCES

Catching the Thief Eventually

By Simon Thougaard

The following is a research proposal, that I have been working on in the last few weeks. The proposed work is new, but it is the culmination of months of study in the field of cyber-physical security. Since I have yet to conduct any experiments or simulations no such details are included. However I have a clear plan for how to proceed. This paper outlines the motivation, problem definition, review of related work and my proposed solution.

Introduction

When criminals obtain the details of victims credit card, they have to choose between two strategies to obtain money. They may either quickly withdraw/spend a lot of money, before a user discovers the crime. Or they may instead slowly withdraw/spend small amounts in order to avoid raising a red flag, or the attention of the victim. The second strategy can be extremely difficult to detect, because the illicit spending may look like ordinary spending. In this case, the criminal may choose to steal only a tiny amount every month, if that ensures that the theft is never noticed.

It has been shown that in Cyber Physical systems, such as the electric grid, attackers can effectively avoid being detected by carefully disguising the attack as noise. Such an attack is called a False Data Injection Attack (FDIA). Many attempts to detect FDIA have been proposed, but none that an attacker cannot side-step by simply slowing down the attack. I have been working on a defence scheme that:

- Guarantees detection eventually
- Cannot be avoided by slowing down the attack
- Works even if only a single node in a system remains uncompromised
- Only requires software updates to implement. No new hardware/sensors are needed.
- Works for any system with noisy measurements

The proposed defence scheme exhibits a few drawbacks:

- It is slow to detect attacks
- It only works if an attacker is doing an economic attack/theft

It is important to note that an attacker may attempt to cause damage to the system, which this defence scheme is not designed to handle. However any defence mechanism designed to quickly detect such an attack, can be easily augmented with this proposed work to also detect long term attacks such as power theft.

Since the defence scheme only relies on existing data gathered, it is essentially a very cheap way to completely prevent theft. By analysing the solution, it can be shown that the rate of false positives can be easily adjusted to essentially no detriment.
Related work

Renewable energy production in the US is reported to have doubled in the period from 2008 to 2018, mainly driven by solar and wind power. Renewable energy production is at an all time high, and expected to increase further.

The Mexican Federal Electricity Commission recently reported that 7% of all energy produced in Mexico in 2018 was stolen, at an estimated cost of US $1.57 billion.

In 2018, the energy consumption of Bitcoin miners was equal to that of Hungary. Even small changes in the price of energy have consequences for the profitability of such enterprises. It stands to reason that with the rise of complexity in the energy sector, and increased demand for cheap electricity in volatile industries, power theft will become an increasingly important issue.

Reliability of power systems have been extensively studied. In the event of faults in the electric grid, control center staff may need to make adjustments to the power generation or load distribution, in order to provide power and avoid further grid failures. Since power meters are noisy, and the power load constantly changing, state estimators are used to quickly estimate the actual state of the system. These state estimators are designed to handle noise in the system and provide reliable estimates which control center staff can use to take action.

If an attacker wishes to disrupt the power grid, and has access to one or more nodes in the grid, they can inject wrong values into the state estimators. Monticelli and Garcia (1983), proposed a method for “bad data detection” which relied on the assumption that bad data can be detected by a chi-squared test. A number of papers were published proposing similar bad data detection schemes.

In Liu et al (2011), these approaches were shown to be inadequate. Liu et al described a type of attack, False Data Injection Attack (FDIA), which could achieve arbitrary results without being detected, assuming that an attacker has access to the entire model of the system. FDIA works by carefully designing the “attack vector”, the false measurements injected into the state estimator, such that no alarm is raised. These attacks have been shown to be easy to design, and many potential solutions has since been proposed to detect them.

Some solutions focus on making a small set of nodes secure, and use measurements from those nodes to detect attacks in the rest. Other work focuses on describing the signature of an FDIA, and raising an alarm when detected. What they all have in common, is that they intend to detect an attack as quickly as possible, and that they depend on the attack vector to exceed some threshold. It should be obvious that if an attacker intends to never be detected, and they know what that threshold is, they can always avoid detection by simply slowing down their attack. To my knowledge, no work has been done to address this problem.
To understand how my proposed solution differs from existing work it is important to distinguish between two fundamentally different types of attack:

- A fast attack, where the goal is to sabotage operation. This attack is intended to be discovered when successfully carried out.
- A slow attack, where the goal is to gain something from the system. This attack is intended to never be discovered.

While these types of attack has been distinguished before (Liu et al 2011), attempts to prevent them tend to emphasize the fast attack. When doing so, they ignore the fact that a slow attack can always circumvent their detection. Similar to the credit card theft example, a patient attacker can always design their attack to stay below the threshold of detection, given that they know the threshold.

**Description of Work**

Suppose we look at a variable $x$ in a system, and $z_1$ and $z_2$, two independent measurements of $x$. Since the measurements are noisy, we can assume that $x=z_i+e_i$, where $e_i$ is some error in the measurement $z_i$. Measurements are generally assumed to have a gaussian distribution. A state estimator will attempt to estimate the real $x$, based on $z_1$ and $z_2$. Let that estimate be a simple average, $\bar{x} = \frac{z_1+z_2}{2}$. The residual $r_i$ is defined as the difference between the raw measurement and the estimate, $r_i = z_i - \bar{x}$. If $|r_i|$ is greater than a certain threshold, $\tau$, and alarm is raised and the bad data is considered detected. This model is a simplification of the models usually described in the literature. The main simplification being that $x$ is usually a vector, and not a single value, and the state estimation is often more complex than a simple average. However, it serves to explain the main thrust of my research.

Essentially any bad data detection scheme relies on the residual, to determine if a given raw measurement is outside expected range. The idea behind a FDIA, is to skew a measurement in one direction or another to effect change in the estimate. Suppose the meter providing $z_1$ is under the control of an attacker. If the attacker wants $\bar{x}$ to artificially increase, they can increase the reported $z_1$ by a certain amount, $a$. The state estimate will be $\bar{x} = \frac{z_1+z_2+a}{2}$. If $a$ is too large, the residual will exceed the threshold and trigger the alarm, so the attacker simply has to calculate the maximum possible attack.

$$r_1 = (z_1 + a) - \bar{x} = z_1 + a - \frac{(z_1+a)+z_2}{2}$$
$$\frac{1}{2}a = r_1 - z_1 + \frac{z_1+z_2}{2}$$
$$a = 2r_1 - z_1 + z_2$$

The maximum $a$ will be achieved when $r_1=\tau$, so the maximum value of $a$ can be easily calculated for each time-step. This is a simplified version of an FDIA.
Solutions have been proposed that more accurately estimate the state, or use an adaptive $\tau$ in order to catch the attack. But these solutions, if known to the attacker, can always be easily circumvented by simply lowering $a$.

My proposed solution is a simple extension to the state estimator, which will keep track of the residual over time. Suppose in the example, that for each time-step, $r_1$ is summed with the previous $r_1$ sum. Even if the threshold is not exceeded in any single time step, the residual sum will constantly increase at the rate of $a$. If there is no attack, the residual sum is not expected to increase or decrease over time. If a threshold, $\tau_{\text{sum}}$ is put on the residual sum, in addition to each individual residual, we have effectively capped the amount of $a$ the attacker can inject into the system over time. The key insight of this approach is that the attacker no longer benefits from slowing down. They will always be detected eventually, when the residual sum reaches $\tau_{\text{sum}}$.

The only way for the attacker to avoid detection is to inject a negative $a$ into the state estimator, to decrease the residual sum. However in the case of theft, this would be equivalent to giving back resources, in order to avoid being detected. There is no way to over or under report a measurement without eventually being detected.

I intend to conduct experiments and simulations to verify the efficacy of the proposed solution. While I am confident in the theoretical basis for the solution, a practical application to the electric grid may present challenges. I am currently working on this application.

**The Problem of False Positives:**
The statistical expectation of the residual sum will remain at 0, but the probability of it diverging from zero will increase over time. This is without even considering a technical bias, such as spillage or energy dissipation, which is expected to be present in any cyber physical system. However it is important to realize that in any attempt to detect FDIA, false positives are inevitable. Other works tend to present a tradeoff between false positives, and detection rate. With my proposed solution, the tradeoff is essentially between how much theft/spillage is tolerated before taking action, and the rate of false positives. However in the case of a thief that seeks to remain undiscovered indefinitely, the threat of getting caught will always outweigh the potential to steal a finite amount of resources. So the threshold for taking action can be set arbitrarily high, without attracting more theft. In the case of a bias in the measurements due to spillage, it can even be argued that my proposed solution is an effective tool at detecting and locating the spillage.

**The Problem of Accounting for Spillage:**
If alarms are consistently raised due to spillage, it could make it harder to detect real theft. With experimental data, it should be possible to quantify how much spillage already exists in a system. This quantity could then be subtracted from the residual sum. However it is important that this quantity is set exactly at, or slightly below the actual spillage rate. If the subtracted amount exceeds the actual spillage, theft is once again possible without being detected.
Conclusion

The solution is a useful and efficient tool to prevent any type of theft in any type of cyber physical system. By recognizing an explicit motive, a successful automated defence mechanism can be formulated. In a large system, where noisy and spillage is tolerated, the potential for nodes in the system to under- or over-report their usage is huge.

I have worked on a simple defence scheme to directly target slow resource theft without requiring any expensive changes to existing system architecture.

The scheme can hopefully be demonstrated to detect theft in a power system.

References

News articles used in motivation:


Literature:


A NOVEL CELL DESIGN OF ALKALINE-BASED ZINC-IODIDE FLOW BATTERY
FOR ENHANCING ENERGY AND POWER PERFORMANCE

Hiep Pham
MAE Department
Missouri S&T
Rolla, Missouri, USA
htp2n5@mst.edu

Tazdik Patwary Plateau
MAE Department
Missouri S&T
Rolla, Missouri, USA
tpd35@mst.edu

Jeffrey Smith
MAE Department
Missouri S&T
Rolla, Missouri, USA
jwsv34@mst.edu

Robert Landers
MAE Department
Missouri S&T
Rolla, Missouri, USA
landersr@mst.edu

Jonghyun Park
MAE Department
Missouri S&T
Rolla, Missouri, USA
parkjonghy@mst.edu

ABSTRACT
The performance of a novel cell design for Redox Flow Batteries (RFBs) utilizing a hybrid alkaline-based zinc-iodide flow system was identified. The cell design has great enhancements over traditional cell designs by enclosing components within the cell, which allowed for the removal of several components, reducing the size and complexity of the RFB, as well as eliminating issues associated with traditional cell designs such as leakage. The performance of the cell was characterized through polarization and discharge tests with varying currents and current densities, and the effects of these parameters were analyzed.

1. INTRODUCTION
Due to technological advances in the world, the consumption of energy has been increased in a great extent. However the primary sources for energy are focusing on the oil, gas and petroleum which pose issues such as the soaring prices of these mainstream energy resources and environmental pollution problems caused by the exhaust from solid, liquid, and gas combustion during energy production. Therefore, one of the most fundamental solutions is to replace fossil fuels with renewable energy sources. However, some alternative power sources face issues with the power generated. Renewable sources such as solar power and wind power cannot provide sufficient energy at all times [1, 2]. The specific timeframe hinders the more widespread adoption of photovoltaic energy generation. In the absence of solar radiation, additional energy comes generally from conventional energy generation. It also requires transitions from solar energy to conventional energy, and vice versa. Due to upgrading the solar panels, photovoltaic penetration increases, the transition rate between solar energy generation and conventional generation increases (particularly during mid-morning and late afternoon). On the other hand, availability and uncertainty of renewable resources arise the need of storing energy for a balanced power supply. So, power generation needs to be established ahead of time because of their dynamic constraints, e.g., startup/shutdown times and costs, ramp rates, etc. [3, 4].

One possible way to overcome these problems is to use energy storage devices. That is, the generated electricity at low-demand periods can be stored for use at high-demand periods. So, energy storage is needed to store the renewable energies for two distinct reasons: (1) to store electrical energy during the low-demand of electricity and to deliver this stored electrical energy during the high-level demand of electricity which can be maintained by shifting time on a daily basis; and (2) to mitigate the dynamic impact of the fluctuation from renewable energy output, stabilizing the grid.

There are many of energy storage system available, among them are Redox Flow Batteries (RFBs), which are reliable and alluring energy storage system especially for use in grid applications [2, 5]. A RFB is an energy storage device which performs electrochemical conversion and redox reaction between the elements in aqueous phase. The fuels, which are aqueous electrolyte solutions, are stored in the external tanks and transferred into the RFB according to the demand of flow. There are several favorable reasons for using RFB in grids, such as, excellent scalability and flexibility, independent size of power and energy, high cycling efficiency, long term sustainability, and quick responsiveness. In addition, these properties are suitable for supporting the grid application in generating electricity from renewable energy sources where RFBs can mitigate the demand of variation of operation times and discharge times.

However, traditional RFBs designs have limitations in terms of the cell structure. First, mechanically weak and brittle graphite materials are used in traditional designs, which makes it difficult to fabricate complex shapes and different flow field configurations. Additionally, the porous nature of the graphite exacerbates electrolyte penetration which decreases the
performance of the RFB. Secondly, traditional RFB designs utilize a stacked structure that includes multiple gaskets in order to prevent leakage, which complicates the assembly and increases the size and weight of the cells. Thus, it imperative to develop new cell designs which can overcome these limitations to improve the performance of RFBs.

Additionally, traditional RFBs have utilized all vanadium (vanadium oxide flow battery), iron, or chromium (Fe/Cr) based flow battery which show limited amount of energy density (<25 Wh l\(^{-1}\)) [7-9], where this low energy density excludes the flow batteries from the energy storage market. Therefore in order to compete in the market, the enhancement of the energy density of the flow battery is needed. In order to get high energy density hybrid flow batteries (such as Zn/Br flow batteries, Li-S and Li-iodine flow batteries) are used where one half-cell features a redox reaction with redox species (metal) that is not fully soluble. The energy densities of Zn/Br flow batteries and Li-S system show ~65 Wh l\(^{-1}\) and ~108 Wh l\(^{-1}\) [7, 10]. But the high corrosiveness and health hazardous of Br/Br\(_2\) redox couple limit the usage of Zn/Br flow batteries. Flammable components, poor scalability, low conductivity of non-aqueous electrolyte as well as expensiveness of the Li-S and Li-iodide flow batteries make the challenges for the flow batteries. Thus, there is need to find new redox species with high solubility and environmental friendliness to get higher energy density aqueous flow battery. Considering these factors, Zinc-Iodide flow battery is a good candidate as it possesses high energy and power density as well as high solubility [11-13].

In this study, a novel design is reported which can address the limitations of conventional flow battery system. In this work, a thermoplastic material, PVC (polyvinyl chloride), is used to substitute the graphite plate. PVC has many benefits like lightweight, better mechanical strength, impermeability, resistance to weathering, chemical rotting and corrosion [14, 15]. In addition, it can be cut, shaped and joined easily to make various designs. For these reasons, PVC is suitable for making the flow field channels for electrolyte distribution in the flow battery system. Our previous report using this design with vanadium-based electrolyte system showed significant improvement compared to traditional designs. In this report, the vanadium-based electrolyte will be replaced with the higher energy and power density alkaline-zinc-based electrolyte system (Fig. 1) and the performance of the new cell based on hybrid alkaline-zinc electrolyte is compared with previous report of traditional flow batteries based on vanadium electrolytes.

![Figure 1. Schematic diagram of Zinc – Iodide Flow Battery](image1)

2. EXPERIMENTAL

A lab-scale RFB with a single-cell configuration was designed, in which the thermoplastic material PVC (McMaster-Carr), 12.7 mm thick and 120 mm side long, was used to replace the graphite plate found in traditional designs. As shown in Fig.2b, the design was based on an embedded pocket-like structure, in which all components are placed inside the pocket. Consequently, several components including gaskets, insulators, and endplates are not necessary and can be removed to simplify the design.

![Figure 2. Geometrical details; (a) novel design cell with an 80 × 80 mm\(^2\) electrode area, (b) actual cell assembly](image2)
The flow field plates of each design included flow channels with a $4 \times 1.59$ mm$^2$ cross-sectional area and a width of 7.5 mm among the channels. The electrodes used were zinc foil (0.1 mm thick) as the anode and graphite foil as the cathode, both of which are cut to the size of the cell electrode area and with the flow channel design cut out. A graphite rod material (McMaster-Carr) that was 1/4 in. thick and 2.5 in. long (side) was fixed through each PVC compartment in the new design to serve as a current collector.

The commercially available cation exchange membrane (Nafion 117) was used as the separator to be placed between the two electrode materials. First, the membrane was treated with 3% H$_2$O$_2$ at 100°C for 1 h, and then soaked in deionized (DI) water at 100°C for 1 h. Then, the Nafion 117 membrane was treated with 0.5 M H$_2$SO$_4$ at 100°C for 1 h, and again washed in DI water at 100°C for 1 h. Finally, the membrane was treated in 1 M KOH at 100°C for 1 h to convert the membrane into a K$^+$ conductive membrane.

The catholyte and anolyte was prepared in accordance to reference [12]. Briefly, 6 M concentration catholyte was prepared by dissolving appropriate amounts of potassium iodide (KI) with iodine (I$_2$) in DI water. The anolyte was prepared with 6 M concentration of potassium hydroxide (KOH) in DI water. The volume of the solutions was controlled at 15 mL. All experimental measurements, including discharge performance and polarization curves, were conducted using a 4 Channel Battery Analyzer, BST8-20A-CST (MTI Corporation, CA 94804, USA) at various currents and at room temperature. No leaking was observed for the novel cell design. The results from these experiments are discussed in the next section.

3. RESULTS AND DISCUSSION

To identify the polarization characteristics of the cell, a polarization test was conducted at a flow rate of 700 mL/h, which plots the cell potential against current (or current density) under a certain flow rate. The measurements of the discharge polarization curves were made starting with a fully charged cell, at an open circuit voltage of 1.81 V, for a period of 3 hrs to a lower voltage limit of 0.3 V. The discharge performance results, in relation to the voltage and energy output over time, are shown in Fig. 4.

The results of the discharge polarization tests showed a maximum power output of 180 mW at 0.5 A. With application of initial low currents, the voltage profile decreases and the power appears to peak at 142 mW at 0.2 A. The power then begins to drop at 0.3 A to 111 mW. However, as the current density increases above 0.3 A, the voltage remains constant at approximately 0.3 V and, interestingly, the power begins to increase. Further investigation is needed to understand this result and to obtain the actual maximum power output of the hybrid alkaline-zinc cell.

The discharge performance of the cell was then identified at a flow rate of 700 mL/h and at a current density of 5 A cm$^{-2}$ relative to the area of the zinc and graphite foil electrodes. The measurements of the discharge performance were made starting with a fully charged cell, at an open circuit voltage of 1.81 V, for a period of 3 hrs to a lower voltage limit of 0.3 V. The discharge performance results are shown in Fig. 4.

From the discharge performance results, an energy output of 0.954 Wh was obtained with a corresponding capacity of approximately 0.7 Ah. Compared to literature values which utilize the hybrid alkaline-zinc flow system and to the previous
report utilizing the novel cell structure with vanadium flow, it is apparent that this result is not sufficient. Thus, further work is needed to improve the performance of this cell design with the alkaline-zinc flow system.

4. CONCLUSIONS
From the results of the novel RFB design with the hybrid alkaline-zinc flow, a maximum power of 180 mW at 0.5 A and energy of 0.954 Wh at 5 A cm\(^{-2}\) was obtained. From the polarization test, it was observed that the voltage remained constant with application of a current above 0.3 A and, subsequently, the power continued to increase. As a result, further investigation is needed to understand this observation and in order to further improve the performance of the cell.

5. ACKNOWLEDGMENTS
The authors gratefully acknowledge the financial support from the Intelligent Systems Center at Missouri University of Science and Technology.

6. REFERENCES
ABSTRACT
Machinability of difficult-to-cut metals like Inconel-718 used in aerospace and nuclear industries has gained a lot of interest over the years due to the need to reduce machining cost and improve machinability. This paper presents the results on experimental investigation of the effects of face-milling methods and cutting speed on high speed machining of Inconel-718 under conventional emulsion (7% concentration of semi-synthetic oil) cooling strategy, using shell mill with coated carbide inserts. Cutting force components and burr formation were measured by up and down milling. The lowest resultant cutting force and cutting life (lasts for only two passes), followed by down milling at 40 m/min (5 passes). The highest cutting force components and resultant cutting force but further increase from 40 m/min drastically increased the cutting force for up- and down milling. Thus cutting speed of 40 m/min is the optimum cutting speed for milling Inconel-718 that generates the lowest resultant cutting force and longer machining passes (longer tool life). Down milling generates lower cutting force components and resultant cutting force, and longer tool life (longer machining passes) than up-milling. The combination of down milling and cutting speed of 40 m/min generates the lowest cutting force and longer tool life (longer machining passes (14)), followed by down-milling at 30 m/min. Up—milling at all three-cutting speed of 30, 40, 50 m/min is not good for machining Inconel-718 as it generates the highest cutting force components and resultant cutting force and shortest tool life (shortest machining passes). Up milling at 50 m/min generates the highest cutting force and the shortest tool life (lasts for only two passes), followed by up-milling at 30 m/min (4 passes), and up-milling at 40 m/min (5 passes). The cutting forces and vibration were more stable and controlled under down milling than up-milling in face milling using coated carbide inserts on shell mill. In face milling, down-milling using optimum machining parameters: cutting speed of 40 m/min axial depth of cut of 2 mm, radial depth of cut of 6.25 mm, and chip load of 0.1 mm under 7% concentration of semi-synthetic emulsion cooling strategy is recommended when face milling Inconel-718 with coated carbide inserts.

Keywords Emulsion cooling strategy, Face milling methods, Inconel-718, High-speed face milling, Coated carbide inserts and Machinability

1. INTRODUCTION
Inconel-718 super alloy is known for its high impact, tensile and rupture strength. It is used extensively for both high-heat and cryogenic applications in aerospace and nuclear industries for components such as turbine blades, low and high-pressure disc compressors used in high temperature, high load and corrosion resistance compartments of jet engines, nuclear fuel parts due to its excellent properties such as high strength-to-weight ratio, ability to retain its properties at high temperatures, high corrosion and creep resistance [1]. However, Inconel 718 is characterized as a very “difficult-to-cut material”, because it poses severe problems during machining [2]. Enormous amount of heat is generated in the cutting zone due to high friction between the tool-chip interface and tool-workpiece interface and is not removed as quickly as it is generated due to its low thermal conductivity and specific heat capacity [3]. Therefore, the heat is mostly concentrated on the tool cutting edges, causing severe tool wear. In addition, due to high temperature and pressure at the cutting zone, some particles from the workpiece are welded on the cutting edges of the tool, which further exacerbates tool wear [4]. Its good mechanical strength and ability to work harden at elevated temperatures leads to high cutting forces and its high chemical affinity with many cutting tool materials leads to welding of the tool to form built-up edge (BUE), rapid tool wear and bad surface roughness [5]. Shell mills are applied in intermittent cutting operations and widely used in milling operations due to its versatility and its applications may range from aerospace and automotive to small tool and die shop, as it can be used for milling many features like conventional face, partial face, end, profile, pocket and surface contour milling. Some problems encountered while machining using intermittent cutting operations are high cutting forces, high cutting temperature, rapid tool wear, tool breakage, and chatter due to impacts, resistance and friction between tool and workpiece. These problems are further exacerbated while machining difficult-to-cut metals such as Inconel-718 at high machining speeds and other machining parameters [6]. Due to the problems mentioned above, care must be taken to select appropriate machining parameters and milling method when machining difficult-to-cut metals.
The demand to improve the productivity of machining difficult-to-cut materials in the aerospace and automotive industries by increasing material removal rate and cutting speeds require manufacturers to understand also the effect of milling methods and how it affects the cutting force components, workpiece surface roughness, tool wear and tool life. Hadi et al. [7] investigated the wear behavior of up- and down-milling on tool wear of PVD-coated carbide inserts and chip formation on ball nose end milling of Inconel-718 under MQL and depth of cuts of 0.5, 0.75 and 1.0 mm at 100, 120 and 140 m/min cutting speeds. It was observed that tool wear is less rapid in down-milling compared to up-milling but achieved only one machining pass above 0.75 mm depth of cut with maximum flank wear above 0.5 mm. Rodrigues et al. [8] investigated the tool life and wear mechanism during face milling of alloy 625 using a tool life criterion of 0.3 mm maximum flank wear (ISO 8688-2). The results show similar wear mechanism with other nickel-based alloys and lower wear rate. Aykut et al. [9] investigated tool wear, cutting forces and chip morphology in face-milling of cobalt based super alloy using PVD coated and uncoated hard metal inserts under dry cutting conditions. Investigated parameters were 0.25, 0.5 and 0.75 mm depth of cut and 30, 35 and 40 m/min cutting speed. It was observed that cutting forces increases with depth of cut and feed rate. Application of coolants and lubricants are important when machining using carbide tools, as they reduce cutting temperature, friction and tool wear, cutting force/power consumed, improve surface roughness and machinability. Due to high temperature obtained in machining Inconel 718, application of coolant is necessary to maintain the properties of the cutting tool and improve machinability. In order to maintain the temperature in machining coolants are supplied to cool and lubricate the cutting zone. Application of coolants reduce the friction and heat generated within the workpiece-tool and tool-chip interfaces and removal of the chips formed away from the workpiece and tool surfaces. Ezugwu et al. [10] investigated the application of high-pressure coolant supply at 20 l/min for all experiments. The coolant was supplied at high pressure via nozzle at 20 l/min flow rate. Emulsion coolant was characterized to obtain pH value, Viscosity, and Thermal conductivity of the fluid at room temperature as summarized in Table 1.

The use of 1- insert for comparative analysis of machinability of Inconel 718 under emulsion cooling strategy was adopted though not often practiced in the industry can be used for comparison of machinability of Inconel 718 under different cooling strategy and cutting parameters [11]. All experiments were conducted using 7% volume concentration of emulsion. Two cutting edge carbide inserts grade coated with advanced cobalt binder and AlTiN/TiN coating were used in all experiments. Each pass of the experiment consist of an axial depth of 2 mm, radial depth of 6.25 mm and length of cut for each pass of 88.98 mm. Factors considered in this research are milling method (Up- and Down-face milling) and high cutting-speeds (30, 40 and 50 m/min) in machining of Inconel-718 under high-speed face milling. Each milling method was conducted on one half end of experimental block.

2.1 Coolant and workpiece preparations

Conventional flood cooling used in all experiment was prepared using semi-synthetic fluid from VAL cool recommended for cutting Inconel-718. Coolant tank was drained and cleaned properly before mixing semi-synthetic fluid and distilled water of ratio of 7:93 respectively to form emulsion fluid used for all experiments. Emulsion of 7% by volume concentration was maintained through all experiments by checking concentration level to ascertain emulsion uniform concentration periodically for all experiments. The coolant was supplied at high pressure via nozzle at 20 l/min flow rate. Emulsion coolant was characterized to obtain pH value, Viscosity, and Thermal conductivity of the fluid at room temperature as summarized in Table 1.

Workpiece of dimension 150 x 75 x 37.5 mm Inconel-718 AMS 5596 plate and saw cut was purchased from Altemp alloys, and the material composition is summarized in Table 2. The workpiece was carefully drilled, counter-bored and tapped to obtain two holes for firmly clamping unto the dynamometer.

Table 1. Thermo-physical properties of emulsion coolant at 25 °C

<table>
<thead>
<tr>
<th>Semi synthetic Vol. Conc. (%)</th>
<th>Viscosity (cP)</th>
<th>Thermal conductivity (W/mk)</th>
<th>pH value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.0</td>
<td>1.08</td>
<td>0.698</td>
<td>9.2</td>
</tr>
</tbody>
</table>

Table 2 Chemical composition of Inconel-718

<table>
<thead>
<tr>
<th>Element</th>
<th>Ni</th>
<th>Cr</th>
<th>Fe</th>
<th>Nb</th>
<th>Mo</th>
<th>Ti</th>
<th>Al</th>
<th>Co</th>
<th>Cu</th>
<th>Mn</th>
<th>Si</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53.67</td>
<td>18.70</td>
<td>17.72</td>
<td>4.83</td>
<td>2.85</td>
<td>0.82</td>
<td>0.24</td>
<td>0.12</td>
<td>0.10</td>
<td>0.08</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.2 Measurement set-up

For each of the experiments, CNC codes for machining each pass under Down- and Up-milling and cutting speeds of 30, 40 and 50 m/min was written and transferred into Cincinnati VMC. The Inconel-718 workpiece with two holes was firmly clamped into a dynamometer with a bolt tightened using a torque wrench to evenly distribute load and reduce noise in the signal. The dynamometer was secured firmly into the VMC fixture. For each pass, the maximum 3-cutting force components in feed ($F_x$), perpendicular to feed ($F_y$), and axial ($F_z$) were processed from the acquired force signal and used to calculate the resultant cutting force component ($F_R$). The schematic of up- and down-face milling methods are shown in figure 1.

![Figure 1: Schematic of milling operation (a. Down-milling b Up-milling)](image)

![Figure 2: Experimental set-up with emulsion coolant and data acquisition system](image)
3 RESULTS AND DISCUSSIONS

3.1 Sample plots of acquired cutting force components for up- and down-face milling versus time

The sample plots of processed cutting force components signal (FX – perpendicular to feed force, FY – feed force, and FZ – axial force) against time acquired is represented in figures 3 and 4 for the first pass for up- and down-milling respectively at cutting speed of 40 m/min under emulsion cooling strategy using coated inserts on shell mill. The sample signals are further processed to determine the maximum cutting force for all passes in all experiments. As shown, in figure 3, the feed cutting force component gives the highest force magnitude and the lowest is the axial cutting force component. This is due to the direction of table feed in relation to the tool rotation in up-milling in which most resistant to chip formation is acted in the feed direction. In figure 4, down-milling cutting force perpendicular to feed component signal gives the highest force magnitude and the least is the axial. This can also be explained by the direction of table feed to tool rotation. In down milling, the tool impacts with the workpiece in the same direction as the feed motion giving a lower resistant compared to the resistance of the tool driving into the workpiece perpendicular to feed motion.

Figure 3 Cutting force signal versus time for 1st pass, cutting speed of 40 m/min and Up-milling. a FY – Feed force, b FX – Perpendicular to feed force, c FZ – Axial force
Figure 4 Cutting force signal versus time for 1st pass, cutting speed of 40 m/min and Down-milling. \textbf{a} \(F_y\) – Feed force, \textbf{b} \(F_x\) – Perpendicular to feed force, \textbf{c} \(F_z\) – Axial force

3.2 Maximum cutting force components, resultant cutting force for up- and down-face milling versus number of machine pass at varying cutting speeds

Further processed cutting force component signal (\(F_x\) – perpendicular to feed force, \(F_y\) – feed force and \(F_z\) – axial force) to determine the maximum cutting force for each pass is plotted against each machine pass for up- and down-milling and cutting speeds of 30, 40 and 50 m/min at constant chip load of 0.1 mm/tooth, radial depth of cut of 6.25 mm and axial depth of cut of 2.0 mm. The resultant force component is further calculated from the processed maximum 3-component force signals for each pass in all experiments. From the plots in figures 3 and 4 for up- and down-milling respectively it shows that the calculated resultant force data follow the same pattern as the highest cutting force component in all experiments. The maximum cutting force component in all experiments increases with increase of machining pass. The progression of cutting force component is shown to be relative to the progression of machined pass. At the selected constant machining parameters of chip load 0.1 mm/tooth, radial depth of cut of 6.25 and axial depth of cut of 2.0 mm, increasing the cutting speeds decrease the resistant to force motion in chip formation and removal as shown in figures 5 and 6. The increase in cutting speeds also increases the progression of maximum cutting force components and resultant force as the machining pass progresses. Up-milling using a coated carbide insert on shell mill shows that increasing the cutting speed to 50 m/min under the selected parameters can significantly increase the cutting force components and resultant force and reduce the number of machining passes achieved and increase machine and tool vibration. As shown in figure 8, the resistance to chip formation and significant maximum cutting force progression leads to significant burr formation. Down-milling at different cutting speed as shown in figure 7, shows a lower progression in cutting force components compared to up-milling shown in figure 8 in all cutting speeds. Similar progression of maximum cutting force component and resultant force was observed for the down-milling at cutting speeds of 30, 40 and 50 m/min.

Figure 5 Maximum cutting force components, resultant force versus number of machine pass for up-milling at varying cutting speeds \textbf{a} 30 m/min, \textbf{b} 40 m/min, \textbf{c} 50 m/min
3.3 Resultant cutting force study for up- and down-face milling versus number of machine pass at varying cutting speeds

Comparative study of the resultant cutting force of up- and down-milling at cutting speeds of 30, 40 and 50 m/min at constant chip load of 0.1 mm/tooth, radial depth of cut of 6.25 mm and axial depth of cut of 2 mm is shown in figure 10. The plot shows that the increase of resultant cutting force with machining pass in up-milling is more drastic for all investigated cutting speed compared to down-milling. More significant tool and machine vibration was observed in up-milling after a shorter number of machining passes compared to down-milling. The resultant cutting forces at cutting speed of 50 m/min in up- and down-milling was initially lower than the resultant cutting force at 30 m/min but they increased more significantly with increase in number of machining passes, and the experiment for 50 m/min cutting speed was terminated before that for 30 m/min cutting speed due to excessive tool and machine vibration. It is evident from the experiments that within the machining parameters investigated that down-milling at cutting speed of 40 m/min should be employed when face milling Inconel-718 under emulsion and using coated carbide insert to ensure longer number of machining passes before enormous tool and machine vibrations are experienced.

Figure 6 Maximum cutting force components, resultant force versus number of machine pass for down-milling at varying cutting speeds a 30 m/min, b 40 m/min, c 50 m/min

Figure 7 Resultant cutting force versus number of machine pass for up- and down-milling at 30, 40, and 50 m/min cutting speeds.
4. CONCLUSIONS

This face milling experiment studied the machinability of Inconel-718 at constant chip load, axial and radial depth of cut at varying cutting speed and milling methods under emulsion cooling strategy using coated carbide insert. From the results obtained for cutting force components, burr formation and tool wear (tool life) based on tool life criterion of 3.5 mm, within the investigated parameters the following conclusions can be deduced:

1. Down-milling generates lower cutting force components and resultant cutting force, and longer tool life (longer machining passes) than up-milling.
2. The combination of down milling and cutting speed of 40 m/min generates the lowest cutting force and longer tool life (longer machining passes (14)), followed by down-milling at 30 m/min.
3. Up—milling at all three-cutting speed of 30, 40, 50 m/min is not suitable for machining Inconel-718 as it generates the highest cutting force components and resultant cutting force, and shortest tool life (shortest machining passes).
4. Up milling at 50 m/min generates the highest cutting force and the shortest tool life (lasts for only two passes), followed by up-milling at 30 m/min (4 passes), and up-milling at 40 m/min (5 passes).
5. Cutting speed significantly affect cutting force components and resultant cutting forces. Increasing the cutting speed from 30m/min to 40 m/min drastically reduces the resultant cutting force, but further increase from 40 m/min to 50 m/min drastically increases the cutting force for up- and down milling.
6. Thus cutting speed of 40 m/min is the optimum cutting speed for face milling of Inconel-718 that generates the lowest resultant cutting force and longer machining passes (longer tool life).
7. Machinability of Inconel-718 when face milling under emulsion cooling strategy and high cutting speed can be improved by using down-milling operation.
8. Feed force is the largest cutting force component followed by perpendicular to feed and least is axial force in Up-milling and perpendicular to feed force is the largest cutting force component followed by feed force and least is the axial force in Down-milling.
9. The cutting force components when machining under cutting speeds of 40 m/min is increases gradually when compared to cutting speed of 50 m/min.
10. Appropriate combination of milling method and cutting speed can improve machinability of Inconel-718.
11. Cutting speed of 40 m/min and down-milling method is the most efficient way to improve machinability of Inconel 718 within the range of the parameters investigated in face milling with coated carbide insert under emulsion cooling strategy.

5. ACKNOWLEDGMENTS

The financial support from the Intelligent System Center (ISC) of the Missouri University of Science and Technology is greatly acknowledged. The financial assistance provided in the form of Graduate Teaching Assistantship by the Department of Mechanical and Aerospace Engineering at Missouri University of Science and Technology is also greatly acknowledged.

6. REFERENCES


Fabrication of AlCoCrFeNi High-Entropy Alloy Coating on an AISI 304 Substrate via a CoFe2Ni Intermediate Layer

Wenyuan Cui, Department of Mechanical and Aerospace Engineering, wc6z68@mst.edu
Frank Liou, Department of Mechanical and Aerospace Engineering, liou@mst.edu

ABSTRACT

Through laser metal deposition, attempts were made to coat AlCoCrFeNi, a high-entropy alloy (HEA), on an AISI 304 stainless steel substrate to integrate their properties. However, the direct coating of the AlCoCrFeNi HEA on the AISI 304 substrate was found to be unviable due to cracks at the interface between these two materials. The difference in compositional change was suspected of being the source of the cracks. Therefore, a new transition route was performed by coating an intermediate layer of CoFe2Ni on the AISI 304 substrate. Investigations into the microstructure, phase composition, elemental composition and Vickers hardness were carried out in this study. Consistent metallurgical bonding was observed along both the interfaces. It was found that the AlCoCrFeNi alloy solidified into a dendritic microstructure. The X-ray diffraction pattern revealed a transition of the crystal structure of the AISI 304 substrate to the AlCoCrFeNi HEA. An intermediate step in hardness was observed between the AISI 304 substrate and the AlCoCrFeNi HEA. The AlCoCrFeNi alloy fabricated was found to have an average hardness of 418 HV, while the CoFe2Ni intermediate layer had an average hardness of 275 HV.

1. INTRODUCTION

As a novel metallic alloy system, high-entropy alloys (HEAs) have received considerable attention in the past decade. The name HEA indicates that the mixing of the principal elements in the alloy leads to a substantial change in entropy. This change in entropy promotes the formation of a simple solid solution instead of complex compounds. One of the extensively studied HEAs is equiatomic AlCoCrFeNi, which shows high hardness, good wear behavior but low tensile ductility [1]-[4]. As-cast AlCoCrFeNi alloy showed a tensile elongation of 1.0%, while post-heat treatment, the elongation was increased to 11.7% [2]. Wang et al. studied the compressive properties of AlCrFeCoNi HEA prepared by vacuum arc melting. They found that this alloy showed large strain hardening and compressive strength up to 2004 MPa with a 32.7% compressive plasticity [4]. Munitz et al. reported the impact of heat treatment of AlCoCrFeNi HEA, in which the BCC (Body-centered cubic) matrix transformation occurred between 650 and 975 °C. This transformation led to a substantial increase in microhardness [3]. Most of these studies are based on material fabricated through processes such as casting and arc melting. Unlike these early studies, laser metal deposition (LMD) was implemented in this study.

LMD is capable of fabricating freeform three-dimensional metallic components [5] and has been used to fabricate several HEAs [6]-[10]. He et al. used laser cladding to produce FeCoCrNiAlTi3 (x = 0, 0.25, 0.5 and 1 atom %, respectively) coating on Q253 steel through the use of elemental powders. Addition of titanium was observed to improve the hardness and wear resistance of the HEA [8]. Similarly, FeCoCrAlCu HEA coating by laser cladding demonstrated good wear resistance under a dry sliding condition [11].

In this paper, the feasibility of coating an AlCoCrFeNi HEA on an AISI 304 stainless steel substrate was investigated. Sole LMD fabrication of AlCoCrFeNi HEA components is very costly due to the need for high-purity (i.e., 99.9%) raw powders of elements such as Co, Cr and Ni. AISI 304 stainless steel, on the other hand, is a low-cost structural material. However, AISI 304 is a soft material with low wear resistance. It is widely used in industrial facilities, transportation equipment and architectural applications. Therefore, by coating AlCoCrFeNi HEA on AISI 304, it can enhance the hardness of AISI 304 structures. This combination of materials could facilitate fabrication of components for applications that require both hardness and wear resistance.

However, the direct coating of AlCoCrFeNi HEA on AISI 304 is difficult due to the change in chemistry, thermal expansion and residual stress of the dissimilar materials. For example, the measured coefficient of thermal expansion (CTE, 10⁻⁶/K) for AlCoCrFeNi HEA was 9.03 (293–303 K), 12.47 (303–378 K) and 13.54 (423–773 K) [12]. However, the CTE values of AISI 304 were 14.7 (293 K), 16.3 (400 K) and 19.5 (700 K) [13]. An extensive network of cracks occurred when a TiVCrAlSi HEA was cladded on a Ti-6Al-4V substrate. This was attributed to the difference between the thermal expansion coefficients and residual stresses associated with the high cooling rate in laser cladding [14].

Therefore, to facilitate the dissimilar material bond, an intermediate layer was necessary and could accommodate the residual stresses and variation in chemistry change[15]. Intermediate layers of FeCr/V were used between AISI 316 stainless steel and Ti-6Al-4V to facilitate a similar material bond [5]. Currently, there are few studies available identifying
the viable intermediate layer between AlCoCrFeNi HEA and AISI 304. In this study, an attempt was made to coat the equiatomic AlCoCrFeNi HEA on the AISI 304 substrate using LMD. The objective was to obtain a strong bond between the two materials. We first demonstrated the issues with direct-coating the HEA onto the substrate. Then we proposed a candidate intermediate material and proved its viability.

2. MATERIALS AND METHODS

Elemental powders of gas-atomized aluminum (Al), chromium (Cr), cobalt (Co), nickel (Ni) and iron (Fe) from Atlantic Equipment Engineers Inc. were used as precursor materials. These powders, weighed in required ratios, were mixed using a Tursula mixer (Glen Mills Inc., Clifton, NJ, USA) for 1 h to obtain homogeneous blends. Commercially procured AISI 304 bar stock (dimensions: 2.75 inch × 2 inch × 0.25 inch) was used as the substrate material for the deposition. The particle size distribution of the elemental powders stated by the producer is as tabulated in Table 1. Elemental compositions (atom %) of the as-blended CoFe2Ni intermediate layer and AlCoCrFeNi alloy are given in Table 2.

Table 1. Particle size distribution of the precursor elemental powders.

<table>
<thead>
<tr>
<th>Materials</th>
<th>US Standard Mesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al</td>
<td>-100</td>
</tr>
<tr>
<td>Co</td>
<td>-100+325</td>
</tr>
<tr>
<td>Cr</td>
<td>-100</td>
</tr>
<tr>
<td>Fe</td>
<td>-100</td>
</tr>
<tr>
<td>Ni</td>
<td>-100+325</td>
</tr>
</tbody>
</table>

Table 2. Nominal compositions (atom %) of CoFe2Ni and AlCoCrFeNi alloy powder blends.

<table>
<thead>
<tr>
<th>Alloy</th>
<th>Al</th>
<th>Co</th>
<th>Cr</th>
<th>Fe</th>
<th>Ni</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoFe2Ni</td>
<td>0</td>
<td>25</td>
<td>50</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>AlCoCrFeNi</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

The laser deposition process was performed in an LMD system whose schematic representation is as seen in Figure 1. The heat source was a 1 kW continuous-wave YAG fiber laser (IPG Photonics, Oxford, MA, USA) at a 9.8 N load and a 10 s load duration. The powders were fed using a vibration X2 powder feed system procured from Powder Motion Labs. The powder was introduced into the melt pool through an alumina tube. A computer numerical control (CNC) table was used to facilitate the movement during the deposition. Argon gas was used to ensure an inert atmosphere and act as a carrier gas to deliver the powder mixture to the melt pool.

The AISI 304 substrates were cleaned with acetone to remove the impurities such as dirt and oil from the surface. A preheating scan was conducted by running the laser across the substrate surface. To ensure a successful start, the power of the initial five layers of the deposition was carried out at 750 W and 8.5% (3.36 g/min) powder feed rate. The remainder of the deposit was run at a power level of 550 W and 8.5% (3.36 g/min) powder feed rate. The thickness of each layer is 1 mm.

After laser deposition, vertical transverse sections of the specimens were cut using a wire electric discharge machine (Hansveld Industries Inc., Rantoul, IL, USA) and mounted in Bakelite for polishing and etching. The metallographic specimens were first ground using 240, 400, 600 and 800 grit silicon carbide papers and then polished using 15 μm, 9 μm and 3 μm diamond suspensions. The final step of polishing involved 0.05 μm colloidal silica suspension. To reveal the microstructure, the electrolytic etching was carried out in the nitric acid solution (70 mL nitric acid, 30 mL distilled water) at 5 V for 5 seconds. Scanning electron microscopy (SEM), energy dispersive X-ray spectroscopy (EDS) and electron backscatter diffraction (EBSD) were performed on Helios Nanolab 600 SEM (Thermo Fisher Scientific, Waltham, MA, USA). The SEM image was acquired by an Everhart-Thornley detector. The EDS element was analyzed by the factory standardizations provided in the Aztec software. The EBSD step size was selected to be 2.5 μm. EBSD data acquisition and analysis were conducted using Aztec and Channel 5 software, respectively. Grain size was measured by the line intercept method, and the misorientation angle was 10°. Optical microscopy images were collected using a Hirox optical microscope. X-ray diffraction patterns were collected using Philips X’pert MRD using Cu anode. The Vickers hardness was measured using a Struers Duramin hardness tester (Struers Inc., Cleveland, OH, USA) at a 9.8 N load and a 10 s load duration. The reported hardness results were the average of three indentations.

3. RESULTS AND DISCUSSION

3.1. Direct coating of AlCoCrFeNi HEA on AISI 304 substrate

The direct LMD of the AlCoCrFeNi HEA on the AISI 304 substrate will be discussed first. Figure 2a shows a portion of the vertical transverse section of the HEA deposit near the AISI 304 substrate. An area close to the crack zone, as marked in the dashed-line box, is shown in Figure 2b with high magnification. A network of cracks, mostly transverse and horizontal in orientation, were found to be prevalent. Cracking occurred at the bottom of this HEA deposit. This could be attributed to the mismatch between the thermal expansion coefficients. The CTE of this HEA was reported to be 9.03 (10⁻⁶/K, 293–303 K) while the value of AISI 304 was 14.7 (10⁻⁶/K, 293 K) [12], [13].
The elemental composition distribution along the interface between the HEA deposit and the AISI 304 substrate is shown in Figure 3. At the bottom of the melted metal, the composition mixing was significant during the laser deposition process (see Figure 3). The bottom of the deposit had high susceptibility of cracking in the transverse cross-section, as seen in Figure 2.

The variation in Vickers hardness across the HEA–AISI 304 direct coating is presented in Figure 4. The average Vickers hardness of the HEA deposits was 412 HV, while that of the substrate was 161 HV. Since the coefficients of thermal expansion are mismatched between HEA and the substrate, residual stresses were developed during the laser deposition process. The AISI 304 substrate had a high elongation rate from 28% to 50% in the temperature range of 300–500 °C [16]. However, the tensile elongations of the AlCoCrFeNi HEA were 1% (as-cast condition) and 11.7% (after heat treatment) [2]. A difference in ductility exists between the substrate and the HEA. Having an intermediate material to bridge these differences was deemed necessary.

3.2. A new transition route

A blend of Fe, Co and Ni powders was selected as the candidate intermediate material. Since they are among the constituents of the AlCoCrFeNi HEA, no special procurement was needed. A Fe–Co–Ni ternary phase diagram at 1073 K was compiled from experimental data [17]. Fe, Ni and Co have excellent mutual solubility, and no brittle intermetallic phases are expected. From the phase diagram, an atomic composition ratio of Fe, Ni and Co of 50%, 25% and 25%, respectively, was chosen. The selected ratio is expected to bridge the material composition gap between the AlCoCrFeNi HEA and AISI 304. This new transition route, AISI 304 substrate → CoFe₂Ni intermediate layer → AlCoCrFeNi HEA, was then carried out and characterized.

3.3. AlCoCrFeNi HEA-AISI 304 with an intermediate layer

3.3.1. Microstructure

The CoFe₂Ni intermediate layer was coated on the AISI 304 substrate using premixed elemental powder. Then, the AlCoCrFeNi HEA was coated on the intermediate layer by LMD. The intermediate layer composition was theorized to avoid the formation of intermetallic compounds and bridge the large gap in strength differences. Figures 5a and 5b show the optical images of etched surfaces of transverse sections of these deposits. Unlike the HEA–AISI 304 direct coating, no apparent cracks were observed, which indicated an improvement in bonding. However, issues of microporosity persisted. A dendrite microstructure was observed along the interface between the intermediate layer and the HEA.

A high-magnification secondary electron image of the AlCoCrFeNi HEA deposit is shown in Figure 6, where a two-phase dendritic microstructure was observed. The area fraction of the dendritic microstructure was ~52%, while the interdendritic area fraction was ~48%. The interdendritic region is named A, and the dendritic region is named B. The mean elemental compositions of A and B (average from three arbitrary points) were analyzed by EDS, and the results are listed in Table 3. It is shown that the atomic percentages of Al
and Ni were ~29% in A and ~41% in B. The percentages of Fe and Cr were ~54 atom % in A and 43 atom % in B. These results indicate that Fe and Cr were rich in A, while Al and Ni were rich in B. The composition of Co did not show evident differences between A and B. The mixing enthalpies between Fe–Cr, Fe–Ni, Fe–Al, Cr–Ni, Cr–Co, Cr–Al, Ni–Co, Ni–Al and Co–Al were −1, −2, −1, −7, −4, −10, 0, −22 and −19 kJ/mol, respectively [4]. The mixing enthalpy of Al and Ni was higher than other pairs, which indicated that Al and Ni tended to form atomic pairs and segregate. Similar results have been reported for the AlCoCrFeNi HEA, with this microstructure being attributed to the spinodal decomposition [2]–[4].

Table 3. Elemental compositions analyzed by energy dispersive X-ray spectroscopy (EDS) of the AlCoCrFeNi HEA shown in Figure 6.

<table>
<thead>
<tr>
<th>Elements (atom %)</th>
<th>Al</th>
<th>Co</th>
<th>Cr</th>
<th>Fe</th>
<th>Ni</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>16.2</td>
<td>16.8</td>
<td>23.4</td>
<td>30.2</td>
<td>13.4</td>
</tr>
<tr>
<td>B</td>
<td>23.5</td>
<td>15.7</td>
<td>19.4</td>
<td>24.2</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Figure 6. Secondary electron image of the AlCoCrFeNi HEA microstructure at a magnification of 10000.

XRD was used to identify the crystal structures of the intermediate layers and the HEA. A transition of the crystal structure was observed from the AISI 304 substrate to the AlCoCrFeNi alloy. The XRD patterns of the AISI 304 substrate, the CoFe₂Ni intermediate layer and the AlCoCrFeNi alloy are shown in Figure 7. The present phases and the corresponding crystallographic information are summarized in Table 4. The peak patterns of FCC were observed in the CoFe₂Ni intermediate layer, while BCC peak patterns were detected in the AlCoCrFeNi alloy. Löbel et al. found BCC and B2 (ordered BCC) phases in AlCoCrFeNiTiₓ (x = 0) when fabricated via arc melting [1]. Due to the same basic lattice structure and lattice parameters, the B2 ordered structure is very hard to detect from XRD, as the peak patterns of B2 and BCC are the same [1]. However, the evidence of the existence of the B2 phase was found from the EDS analysis above.

Table 4. Summary of phases detected by XRD analysis for AISI 304, CoFe₂Ni and the AlCoCrFeNi HEA.

<table>
<thead>
<tr>
<th>Alloy</th>
<th>Lattice</th>
<th>Space Group</th>
<th>Lattice Parameter (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AISI 304</td>
<td>FCC</td>
<td>Fm-3m (225)</td>
<td>3.5911</td>
</tr>
<tr>
<td>CoFe₂Ni</td>
<td>BCC</td>
<td>Im-3m (229)</td>
<td>2.87</td>
</tr>
<tr>
<td>AlCoCrFeNi HEA</td>
<td>BCC</td>
<td>Im-3m (229)</td>
<td>2.876</td>
</tr>
</tbody>
</table>

Figure 7. XRD pattern of the AISI 304 substrate, the CoFe₂Ni intermediate layer and the AlCoCrFeNi HEA.

The evolution in chemistry from the intermediate layer to the substrate was characterized by an EDS line scan first. The quantitative results are shown in Figure 8a. The EDS measured results of the AISI 304 substrate (Cr: ~18–19 atom %, Fe: ~70–72 atom %, Ni: ~9–10 atom % in Figure 8a) did not vary from the nominal AISI 304 elemental compositions. Mn (~1–2 atom %) was detected in the AISI 304 substrate by EDS but is not shown in Figure 8. The percentages of Co (~17–22 atom %) and Ni (~21–23 atom %) reduced, while the Fe (~54–56 atom %) content increased from the intermediate layer to the AISI 304 substrate. A small amount of Cr (~3–5 atom %) was present in the intermediate layer, because the substrate was mixed with the intermediate layer. The composition distribution from the HEA to the intermediate layer is shown in Figure 8b. The constituents of the AlCoCrFeNi HEA were detected by EDS (Al: ~16–17 atom %, Co: 19–20 atom %, Cr: ~17 atom %, Fe: ~25 atom %, Ni: ~20–21 atom %). The difference between the as-blended (20 atom %) and as-deposited aluminum (~16–17 atom %) percentages is suspected to be a consequence of inconsistency in capture efficiencies of the constituent powders, and evaporation due to differences in melting point. Al and Cr were present in the intermediate layer as seen in Figure 8b, and their total content was ~4–5 atom %.
3.3.2. EBSD

Figure 9a shows the inverse pole figure (IPF) map obtained from the bottom of the HEA section of the specimen. The measured area was approximately 3.4 mm x 1.2 mm of the cross-section parallel to the build direction (BD), which spanned from the left to the right of the specimen. The difference in color indicates the different crystallographic orientations. From Figure 9a, the overall constitution can be classified into two zones—the edge zone (1 and 3) and the middle zone (2). In areas 1 and 3, the grains were observed to be elongated along the build direction (see 1 and 3 in Figure 9a). The distributions of the intercept lengths (using 100 horizontal lines) in different areas are depicted in Figure 9b. The median linear intercept for areas 1 and 3 was 72.5 µm, while it was 127.5 µm for area 2. From the linear intercept distribution of area 2, 25% of the intercept values were greater than 300 µm, whereas only 14% of the intercept values were above 300 µm for areas 1 and 3. This grain morphology is likely to be a consequence of deposition toolpath and variation in cooling rate at edges and in the middle. Figures 9c and 9d show the {100}, {110} and {111} pole figures of different areas, which give the distribution of the pole density along the build direction. The pole figure of the areas 1 and 3 (Figure 9c) suggests that the orientations of the grains were close to the <100> direction. However, the grains were random in orientation and did not appear with obvious texture in area 2 (Figure 9d). Further study is necessary to investigate the impact of this toolpath on the grain morphology.

3.3.3. Vickers hardness test

Figure 10 gives the Vickers hardness distribution of the AlCoCrFeNi HEA deposited on the AISI 304 substrate with the CoFe₂Ni intermediate layer. The Vickers hardness of the CoFe₂Ni intermediate layer was around 275 HV, which could be attributed to the solid solution strengthening. The average Vickers hardness of the HEA deposit was in the range of 418 HV, because of the second-phase strengthening [2].
AlCoCrFeNi HEA had a BCC structure. The transition from FCC to BCC structure is also expected to enhance the hardness. The high hardness is expected to correlate with good performance in strength and wear resistance [1].

4. CONCLUSIONS

An AlCoCrFeNi HEA was coated on an AISI 304 substrate by laser metal deposition (LMD) technology. The coating on the substrate without and with the intermediate layer was characterized and discussed. The main conclusions are as follows:

Cracking was found to be prominent when the AlCoCrFeNi HEA was directly coated on the AISI 304 substrate due to the compositional change between HEA and the substrate.

Using an intermediate layer of CoFe2Ni improved the bond. The incorporation of the intermediate layer successfully eliminated crack formation in the deposit. XRD patterns revealed a transition of crystal structure from FCC in the AISI 304 substrate to BCC in the AlCoCrFeNi alloy. The evidence of a B2 phase in the AlCoCrFeNi HEA was also found in the EDS analysis results. The AlCoCrFeNi alloy fabricated by LMD was found to have an average hardness of 418 HV, while the CoFe2Ni intermediate layer had an average hardness of 275 HV.

5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from NSF (National Science Foundation) grants CMMI-1547042 and CMMI-1625736. The support from the Intelligent Systems Center (ISC) and Materials Research Center (MRC) is also appreciated.

6. REFERENCES


MULTI-PHYSICAL MODELING OF SELECTIVE LASER MELTING

Yaqi Zhu
Department of Mechanical and Aerospace Engineering
Missouri University of Science and Technology
Email: yzbxd@mst.edu

ABSTRACT
Numerical modeling based on physics profiles of Selective Laser Melting (SLM) process could help determine the set of optimal parameters to avoid defects and deformations in the final part. However, most multiphysical models of SLM process do not include the prediction of microstructure evolution for the final part. In this study, we develop a complete modeling framework composed of a thermal model of the powder pool and a phase-field model for microstructure evolution. The influence of different heat source scanning velocity on microstructure is investigated as an example using this model. This approach enable us to predict the morphology of final product under various parameters, which provide an important reference for making the product-oriented processing strategy.

1. INTRODUCTION
Selective Laser Melting (SLM) is one of the most promising techniques in additive manufacturing process. It allows parts fabricated from Computer aided design (CAD) models directly, enabling the produce of complex geometries and reducing the enormous manufacturing cost. SLM system is composed of a powder delivery system, a laser scanner system and a fabrication powder bed (Fig. 1). In SLM process, metal powders are delivered and spread in a thin even layer over the build surface by a piston and roller. Selected regions of the metal powder layer are then melt by a focused laser scanning, fuse and coalesce into a slice of the manufactured part. When one layer completed, the fabrication piston is lowered and a new layer of powders is deposited to repeat the process. With the layer-by-layer fabrication, a finished part could be obtained after removing all the loose powder.

To obtain a high-quality part in SLM processing, numbers of processing parameters need to be controlled. For example, for the laser heat source, laser power, laser speed, spot size, scanning velocity and strategy need to be considered. For the powder bed, the particle morphology, particle size distribution, surface roughness, and the basic property of material such as bulk density, thermal conductivity and heat capacity will decide the heat transfer, flow and phase transition of metal powders. For ambient environment, property of shield gas, ambient pressure and temperature would have an impact on the interaction of powder and circumstance. [2] To avoid the defects and deformations in the final part, a set of optimal parameters need to be determined for each new geometry and material which needs large amount of experiments and tests if solely relying on the experimental trial-and-error method.

Numerical modeling based on physics profiles of SLM process offers a way to predict the product quality and validate experiment results, reducing and saving the time and cost of manufacture process design. In early works of SLM modeling, the powder bed were regarded as homogeneous to simplify the mathematics. For example, Zäh and Lutzmann proposed a finite element model in terms of the temperature distribution during laser scanning. A general heat conduction and heat source model were included. The powders were considered as a continuum with effective properties in this model. [3] To introduce the non-homogeneity of powder bed, Xiao and Zhang divided the mush zone during melting process into two sub-region and obtained the temperature distribution and locations of melting interfaces by an integral approximation method. With this model, the shrinkage effect of powder bed could be investigated in single- and multi-line sintering of multiple layers. [4,5] But the description of its effect on melt pool geometry is limited in these models.

Körner and Attar et al. came up with a 2D lattice Boltzmann model instead of previous fluid dynamics models where individual powder particles are considered. The model contains beam absorption, powder heat convection, phase transformation and fluid motion, and capillarity and wetting between liquid and solid. Lattice Boltzmann method was then adopted to simulate the temporal evolution of one-particle distribution functions and predict the local powder melting and re-solidification processes on the powder level. [6-8] This model was proved to be much closer to the experimental results and revealed the complicated

Figure 1. Selective Laser Melting system [1]
mechanism in SLM processing. Then Gurtler et al. developed a transient three-dimensional numerical model to simulate the shape of melting pool and the hydrodynamics. The volume of fluid method was used to describe the free surface of molten steel. Khairallah et al. also proposed a 3D mesoscopic model, in which they have coupled a measured powder size distribution from experiment to improve the model’s resolution. [9] Further physics such as recoil pressure, Marangoni convection and radiation was added into this model to make it closer to realistic process. [10]

During SLM processing, a morphology evolution would occur from liquid to nucleation and fully crystallized in solidification of melt powders. The final formed microstructure mainly depend on the temperature profile of the melt pool and will have a key influence on the mechanical, thermal, transport and other properties of the product. However, in the researches mentioned above, no work include this process. Recently, Liu et al. studied the solidification in SLM processing by a 2D integrated phase field and thermal lattice Boltzmann model. They have got a dendrite growth of a single grain from a liquid under the release of latent heat of phase transformation and a fluid flow effect and revealed the evolution of a local microstructure of Al-Cu alloy in SLM processing. [11] But the growth of a single grain could not represent the morphology of the whole product. Therefore a model which could simulate complete and representative microstructure evolution is still in lack.

In this study, we developed a full complete model for SLM process, which will include (a) a microscale heat transfer in fluid model to account for the absorption of laser radiations by the metal powder, conductive and convective heat transfer process, and phase transition and wetting processes; (b) a mesoscale grain growth model to simulate the evolution of the grain structures during solidification. The finite volume approach coupled with a phase-field method will be adopted to predict the microstructure evolution.

2. METHODOLOGY

2.1. Heating and Melting Process

The heat transfer model geometry shown in Fig. 2 consists of an $6 \times 12$ array of metal powder of one size distribution ($R = 20 \ \mu m$) contained in an inert gas-filled layer. The moving heat source (laser beam) input is assumed to have a Gaussian distribution with a focus diameter of 130 $\mu m$ and it is modeled as a volume heat source. The moving track of the heat source is plotted in Fig. 2, which is multitrack with a spacing of 95 $\mu m$.

Depending on the absorbed laser intensity, the solid metal powder will be transformed into liquid and vapor phases. Thus, the calculation of thermal fields during the SLM process involves the solution of a coupled system of thermo-fluid dynamical equations. The heat conduction equation

$$
\rho C_p \frac{\partial T}{\partial t} + \rho \nabla \cdot (k \nabla T) + Q_{\text{Laser}} - Q_{\text{rad}} \tag{1}
$$

where $\rho$ is the density, $k$ is the thermal conductivity, $Q_{\text{Laser}}$ is the energy of laser heat, $Q_{\text{rad}}$ is the thermal radiation loss according to Stefan-Boltzmann law. Because a phase change occurs during melting process, latent heat of the powder material are accounted for by the effective properties defined in terms of the phase fraction variable.

The effective heat capacity can be expressed as below.

$$
C_p = \frac{1}{\rho} \left( \theta \rho_{ph1} C_{ph1} + (1-\theta) \rho_{ph2} C_{ph2} \right) \tag{2}
$$

The effective density can be calculated as

$$
\rho = \theta \rho_{ph1} + (1-\theta) \rho_{ph2} \tag{3}
$$

And effective conductivity is

$$
k = \theta k_{ph1} + (1-\theta) k_{ph2} \tag{4}
$$

In which $\alpha_m$ is the mass fraction, where

$$
\alpha_m = \frac{1}{2 (1-\theta) \rho_{ph1} - \theta \rho_{ph2}}, \quad X_{ph1} \quad \text{and} \quad X_{ph2} \quad \text{are property of solid and liquid phase, respectively.}
$$

2.2. Grain Growth Model

A phase-field method is applied to build the grain growth model, in which each grain in the microstructure is represented by an order parameter $\eta$. The temporal evolution of order parameter follows the Allen-Cahn kinetic equation:

$$
\frac{d \eta_i}{dt} = -L_i \left( \frac{\partial f(\eta_i, \eta_{i+1}, \ldots, \eta_p)}{\partial \eta_i} - \kappa \nabla^2 \eta_i \right) \quad i = 1, 2, \ldots, p. \tag{5}
$$

In which $f(\eta_1, \eta_2, \ldots, \eta_p)$ is the function of total free energy of the system and has a form of double-well function.

Figure 2. SLM model geometry
\[
\partial f (\eta_1, \eta_2, ..., \eta_p) = \sum_{i=1}^{p} \left( -\frac{1}{2} \eta_i^2 + \frac{1}{4} \eta_i^4 \right) + \sum_{i=1}^{p} \eta_i^2 \eta_j^2 
\]

So
\[
\frac{d\eta_i}{dt} = -L \left( \eta_i - \eta_i + 2 \sum_{j=1}^{p} \eta_j^2 \right) - \kappa \nabla^2 \eta_i, \quad i = 1, 2, ..., p.
\]

\( \kappa \) is the gradient energy coefficient, \( L \) is the kinetic coefficient, which is related to the mobility of grain boundary \( M \) and grain boundary width \( l_{gb} \), \( L = \frac{4M}{3l_{gb}} \). \( M \) is temperature dependent and is the main factor of the influence of temperature on morphology:

\[
M = \begin{cases} 
10^{-6} e^{-3.3 \times 10^5 Y^{-1}}, & T < T_M \\
1.96 \times 10^{-8} m^4 J^{-1} s^{-1}, & T >= T_M 
\end{cases}
\]

In our simulation, the coupled heat transfer model is based on heat transfer module of the finite element simulation software package COMSOL Multiphysics. Then, the temperature history from COMSOL simulation is approximated as a function of time and imported into phase-field model. The phase-field model is conducted in MATLAB software.

3. RESULTS

Figure 3 shows the simulation results for SLM process with two different scanning velocity. The position of moving heat source at these two cases is same. It can be seen from Figure 3(a) and (e) that with a slower scanning velocity, the temperature distribution is more homogenous, in which the difference between maximum and minimum of temperature is smaller than the value of faster scanning velocity. Meanwhile, the maximum of temperature also decreases. This can be explained as slower scanning velocity allows more sufficient transfer of heat between particles and particles, and between particles and environment. Besides, due to the more transferred heat, the phase fraction of vapor and liquid is also larger than that in faster case (Fig. 3(b) (c) (f) and (g)), which indicates more heat is consumed as latent heat for phase transition for slower scanning velocity.

To investigate the temperature variation of particles at different positions for various scanning velocities, three particles located at the beginning, middle and final stage on the scanning track of moving heat source are selected (as shown in Fig. 4(a)). It can be observed in Fig. 4 (b) that for particle p1, there are two peaks of temperature. The first one is caused by the direct heat absorbed from heat source, and the second peak is caused by the heat transfer from neighboring particles when heat source move to the end of second track. It is same reason for the

![Figure 3](image-url)

Figure 3. Temperature distribution and phase fraction of powder pool for scanning velocity of 0.1m/s at 6.6 ms (a, b, c, d) and 0.14 m/s at 4.8 ms (e, f, g, h).

![Figure 4](image-url)

Figure 4. (a) Schematic diagram of positions of three particles p1, p2 and p3, temperature profile of p1, p2 and p3 versus processing time under different scanning velocity (b) 0.1 m/s (c) 0.14 m/s.
temperature variation of particle p2 and p3, where three peaks of p2 correspond to one direct heating of laser source at second track and two heat transferred from left/right when heat source passes by at first/third track. The temperature of all particles is gradually decreasing as the source beam moving away and the slope of the plots tend to be identical in the end. Compared Figure 4 (b) and (c), the temperature peaks for 0.1m/s is higher and wider because of longer time of heat transferred. And the cooling rate for plots of scanning velocity of 0.1 m/s are also smaller than 0.14 m/s. A linear approximation is applied for the cooling temperature profile of particle p1 for being implemented in the phase-field model:

\[
T = -58.669 \times t + 1810 \quad \text{for } 0.14 \text{ m/s}
\]
\[
T = -56.2194 \times t + 1810 \quad \text{for } 0.1 \text{ m/s}
\]  

Figure 5. shows the microstructure of the steel sample after SLM processing. The model is dimensionless with the grid space \(dx = dy = 2\) and mesh size \(N_x = N_y = 200\). Different colors in the image represents grains with different orientations. The cooling rate of two samples are given in Eq. (11). It can be observed that grain size is evidently larger for a slower cooling rate case. The nucleation of melted sample will take place when cooling temperature is below melting point of steel (1810K). And grains keep growing when temperature continue decreasing. However, the microstructure would reach an equilibrium state after certain \(T\) because the driving force of the system is not enough to overcome the energy barrier for further evolution. A smaller cooling rate allows longer time for grain growth. Therefore, a larger grain size can be obtained as a result.

4. CONCLUSION

This study provide a couple-thermal model and phase-field model for SLM processing. The couple-thermal model could describe the heat transfer and phase transformation with different processing parameter and provide temperature information of particles at different locations and different time. By using the temperature history information from couple-thermal model, the microstructure evolution of product could be simulated in phase-field model. Therefore, the final microstructure information could be obtained. We conduct an investigation of influence of scanning velocity on the microstructure as an example. Based on the simulation results from coupled-thermal model, a slower scanning velocity results in a more homogeneous temperature distribution and large fraction of phase transformation. It also cause a broadening of temperature peaks for single particles and a smaller cooling rate. The cooling rate is demonstrated that has a large influence on the microstructure, where smaller cooling rate corresponding to a larger grain size in the final product. So the microstructure could be predicted for different processing parameters, which offers an important reference for the product-oriented experiment design. So far, we only consider the heat transfer and phase transformation in our thermal model. The flow behavior of melting powder is not included. And the model’s parameter have not be validated with practical experimental data.

We are going to further complete the model from these two points as the next step.

5. ACKNOWLEDGMENTS

This work is funded by the grant from Intelligent Systems Center (ISC) and PI’s startup.

6. REFERENCES


A REVIEW OF METALLIC FUNCTIONALLY GRADED MATERIALS FABRICATED BY DIRECTED ENERGY DEPOSITION

Yitao Chen  
Department of Mechanical and Aerospace Engineering  
Missouri University of Science and Technology  
Rolla, MO, USA

Lei Yan  
Department of Mechanical and Aerospace Engineering  
Missouri University of Science and Technology  
Rolla, MO, USA

Frank Liou  
Department of Mechanical and Aerospace Engineering  
Missouri University of Science and Technology  
Rolla, MO, USA

ABSTRACT
Functionally graded materials are composite materials with a gradual variation in composition from one component to the other. Nowadays, additive manufacturing of metallic functionally graded materials has become an important topic. The present paper reviews the state-of-art of metallic functionally graded materials fabricated by directed energy deposition, an additive manufacturing approach to build 3D metal parts. The research progress in metal-metal functionally graded materials, metal-ceramic functionally graded materials, and functionally graded coating will be elaborated. And the technical challenges in this field will also be discussed.

1. INTRODUCTION
A functionally graded material (FGM) is a two-component composite material with a compositional gradient from one component to the other [1]. With the rapid development of industry and human life, materials with single component become limited in use when the working conditions are complex and severe which require different materials properties in different locations of an individual part. While a two-component composite material can exhibit materials properties from both components, nowadays they have received quite a lot of concerns in various engineering fields. However, the manufacturing processes of conventional composite materials often merge two components directly. When working in the extreme environment, if two components which are directly joined are very different in materials properties, the part is susceptible to failure at the interface due to delamination [2]. To avoid this problem, a compositional gradient was introduced to replace the sharp interface [3]. As shown in Fig. 1(b), instead of the direct joining in Fig. 1(a), a compositional gradient between material A and material B will result in a smooth transition in material properties (F(x), G(x)) which can reduce the chance of failure resulted from the sharp interface. Metallic FGMs are FGMs made by metallic materials. Due to the important role of metals and their alloys in our everyday life, numerous research works have been done on metallic FGMs. The earliest concept of FGM was proposed in the 1980s in Japan, and the purpose was to reduce the thermal stress in metal-ceramic composites [3]. Later, FGMs, especially metallic FGMs have been widely used in many other areas such as aerospace, medicine, etc. [4].

![FGM example](image)

Fig. 1. (a) Directly join material A and material B with a sharp interface (b) FGM which a compositional gradient from material A to material B. Where F(x) and G(x) stand for different material properties.

Conventional methods for manufacturing FGMs include vapor deposition, powder metallurgy, centrifugal casting, etc. However, these traditional methods all have their own constraints. Vapor deposition only works well for thin films. FGMs fabricated by powder metallurgy and centrifugal casting rely on the shape of the mold and the volume is also limited by the mold [4]. These days, additive manufacturing (AM) is attracting a huge amount of interest in fabricating 3D parts. According to the definition of ASTM, AM is “the process of joining materials to make objects from 3D model data, usually
layer upon layer, as opposed to subtractive manufacturing methodologies [5]. The computer model of a 3D object can be constructed and sliced into thin layers by 3D computer model. Then the AM system will form each layer under the guidance of the toolpath and join all the layers together to make a 3D part. Compared with traditional subtractive manufacturing, AM can directly produce complex 3D parts without much tooling and assembly, and it is more material-saving [6]. Thus, AM becomes an important part of the manufacturing industry. Directed energy deposition (DED) is a major AM technique for metallic materials. As seen in Fig. 2, during the DED process, focused energy, usually a laser beam is applied to create a melt pool on a metal substrate. Metallic materials in powder or wire form are fed into the melt pool and solidify into a 2D solid layer. The tool path is guided by computer numerical control (CNC) to fill every layer, and the successive layers will be built until a 3D part is achieved [7]. Because of the feature of in situ material feeding, DED can flexibly change the material composition from layer to layer to achieve a compositional gradient along the building direction. Thus, DED is a suitable approach to make FGMs with a compositional gradient [8]. This technique is especially useful in industrial areas where multiple types of outstanding mechanical properties are required within one metal.

![Fig. 2. Schematic diagram of the mechanism of laser-based directed energy deposition.](image)

Nowadays, there are a decent amount of review articles in FGMs [2, 4, 8-12]. However, there is still more work to be done for obtaining a comprehensive understanding of the state-of-art of metallic FGMs fabricated through DED. This paper will summarize the current research progress in different types of metallic FGMs via DED and discuss the major technical challenges remained in this area in order to provide guidelines for future study.

2. RESEARCH IN METALLIC FGMs BY DED

For metallic FGMs fabricated by the DED processing method, they can be classified into metal-metal FGMs, metal-ceramic FGMs, and functionally graded coating. The details of the research progress on each of them will be elaborated in Sections 2.1, 2.2, and 2.3, respectively.

### 2.1. Metal-Metal FGMs

With the rapid development of human society, there is a need to study the joining of different metals and alloys in order to produce parts which can work in a harsh environment which requires comprehensive material properties and make the best use of properties from both components. Due to the difference in thermal expansion coefficient of two dissimilar metals or alloys, direct joining can cause serious cracking at the interface. FGMs have been designed and fabricated to resolve this issue.

![Fig. 3. Metal-metal FGM from 100% metal/alloy A to 100% metal/alloy B.](image)

By utilizing the benefit of composition control in DED, so far, metals and their alloys which are widely used as engineering materials, such as iron, nickel, and titanium, have been studied in metal-metal FGMs and the compositional grading starts from 100% metal/alloy A to 100% metal/alloy B (see Fig. 3). Stainless steel and nickel-based superalloy can be merged to combine the strength at elevated temperatures in nickel-based superalloy and the low cost and in stainless steel. Carroll et al. fabricated 304L/Inconel 625 FGM without compositional boundaries using DED [13]. Shah et al. explored the manufacturability of 316L/Inconel 718 FGM and the effects of laser power and powder flow rate on tensile properties [14]. Titanium alloys are strong, light and biocompatible, which are then considered as one of the best engineering materials in aerospace and medical engineering. Combining titanium alloys with stainless steel to take advantages of both alloys has also become popular in recent years. 304L stainless steel has been attempted to merge with Ti-6Al-4V with graded intermediate vanadium layers [15, 16]. In [17], 316L stainless steel has been joined with Ti-6Al-4V by inserting vanadium and chromium layers in between. Titanium alloys are also combined with aluminum alloys as they are able to form γ-TiAl intermetallics, which is a promising light weight structural material. Ti-Al has been synthesized by either grading pure titanium to pure aluminum [18] or alloying 100% Ti-6Al-4V and 100% AlSi10Mg [19]. In addition, pure titanium and titanium aluminate Ti-48Al-2Cr-2Nb have been integrated with a compositional gradient by DED in [20]. Improving the biocompatibility of Ti-6Al-4V has also been studied. Schneider-Maunoury et al. fabricated Ti-6Al-4V/Mo FGM, as adding non-toxic molybdenum element will help improve the biocompatibility and molybdenum can also tailor the overall Young’s modulus [21].

The tool & die industry needs composite materials with high thermal conductivity and excellent mechanical properties in
different regions. Combining the high conductivity of copper-based alloys and the outstanding mechanical properties of iron and nickel-based alloys using compositional gradients has therefore received great concerns [22, 23]. Graded H13 tool steel and high purity copper was successfully synthesized by Articek et al. [22]. Karnati et al. blended different composition of elemental copper and nickel powders to realize graded Cu/Ni structure [23]. These developments of metal-metal FGMs are summarized in Table 1.

**Table 1. Summary of metal-metal FGMs fabricated by DED.**

<table>
<thead>
<tr>
<th>Metal/Alloy A</th>
<th>Metal/Alloy B</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS304L</td>
<td>Inconel 625</td>
<td>[13]</td>
</tr>
<tr>
<td>SS316L</td>
<td>Inconel 718</td>
<td>[14]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>SS304L</td>
<td>[15, 16]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>SS316</td>
<td>[17]</td>
</tr>
<tr>
<td>Ti</td>
<td>Al</td>
<td>[18]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>AlSi10Mg</td>
<td>[19]</td>
</tr>
<tr>
<td>Ti</td>
<td>Ti-48Al-2Cr-2Nb</td>
<td>[20]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>Mo</td>
<td>[21]</td>
</tr>
<tr>
<td>Cu</td>
<td>H13</td>
<td>[22]</td>
</tr>
<tr>
<td>Cu</td>
<td>Ni</td>
<td>[23]</td>
</tr>
</tbody>
</table>

### 2.2. Metal-ceramic FGMs

Metal matrix composites (MMCs) belong to a type of composite material which mainly consists of two components, where a bulk metal matrix is reinforced by dispersing another material which acts as the reinforcement into the metal matrix [24]. Usually, the reinforcing materials are hard ceramic materials such as metal carbides which can significantly improve the strength and hardness of the metal matrix. In conventional MMCs, reinforcing materials are evenly dispersed into the matrix, which strengthens the matrix uniformly at every location. However, for many cases, it is not a satisfactory strategy to strengthen the entire bulk metal part with the same degree. For example, a cutting tool has different functions at different locations. Some specific locations need a high hardness while some other locations need to maintain a good toughness. Hence, a variation in the concentration of the reinforcing material with respect to the location will be more preferred (see Fig. 4). Between different locations, a compositional gradient can be introduced and DED is able to adjust the concentration of matrix and reinforcing materials. Therefore, the DED process also plays an important role in converting conventional MMCs into more flexible compositionally graded MMCs.

![Fig. 4. Changing conventional MMC into graded MMC.](image)

Titanium-based alloys are most widely researched in functionally graded MMCs by the DED process, and the most commonly used reinforcing component is titanium carbide (TiC). In 2003, Ti/TiC FGM was fabricated by Liu and DuPont [25], which was one of the earliest generation of FGMs fabricated by DED. In recent years, more research efforts have been focused on manufacturing Ti-based FGMs reinforced by TiC to improve hardness and wear resistance of titanium alloys. Laser-aided wire deposition also belongs to the category of DED. And since wire feeding process has a larger capture rate, attempts have also been made to use metal wire as the matrix while ceramic reinforcing materials are supplied as powders. Ti-6Al-4V/TiC functionally graded MMC has been fabricated by simultaneously feeding Ti-6Al-4V wire and TiC powder by Wang et al. [30], which provides a higher manufacturing efficiency, a lower cost, and a larger density. Tungsten carbide (WC) powder was also adopted to harden titanium alloys in [31]. Table 2 shows typical research in titanium-based FGM strengthened by TiC or WC with the corresponding properties studied.

**Table 2. Titanium-based metal-ceramic FGMs fabricated by DED.**

<table>
<thead>
<tr>
<th>Metal Matrix</th>
<th>Ceramic Reinforcement</th>
<th>Properties Studied</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ti</td>
<td>TiC</td>
<td>Microhardness</td>
<td>[25]</td>
</tr>
<tr>
<td>Ti</td>
<td>TiC</td>
<td>Tensile properties, hardness, wear resistance</td>
<td>[26]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>TiC</td>
<td>Microhardness, wear resistance</td>
<td>[27]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>TiC</td>
<td>Tensile properties, microhardness</td>
<td>[28]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>TiC</td>
<td>Microhardness</td>
<td>[29]</td>
</tr>
<tr>
<td>Ti-6Al-4V wire</td>
<td>TiC</td>
<td>Tensile properties, wear resistance</td>
<td>[30]</td>
</tr>
<tr>
<td>Ti-6Al-4V wire</td>
<td>WC</td>
<td>Microhardness</td>
<td>[31]</td>
</tr>
</tbody>
</table>

### 2.3. Functionally graded coating

Coating is an important method to tailor the surface properties of a bulk metal substrate. DED has been found to be able to coat different types of metals and ceramics on metal substrates to significantly improve the surface hardness, wear resistance, and thermal barrier properties. Also, repairing and remanufacturing the worn surface of a component can be done by DED with the aid of reverse engineering. It has been found that it may induce large stress if the coating material is directly deposited on the substrate because of the mismatch in thermal expansion coefficients. Therefore, similar to joining dissimilar metals, a compositional gradient is also necessary between the top coating surface and the substrate to ensure a smooth change in order to eliminate, or at least alleviate the stress issue (see Fig. 5). Materials for functionally graded coating might be either ceramics or hardfacing alloys. Although functionally graded coating can also be classified into metal-metal or metal-ceramic FGMs based on the types of coating materials, due to its special coating function, here it is discussed separately in an additional subsection.
Laser-based coating has been applied on the surface treatment of steels to enhance the service life of engineering components. Today, different types of steels strengthened by DED-processed graded coating have been reported. Compositionally graded alumina/Ni-Cr coating was deposited on 316 stainless steel with a lower porosity and a better bonding strength. The hardness of coating was in the range of 1800-2000 HV [32]. Compositionally graded Fe/SiC layer from 85/15 to 95/5 on the surface of mild steel substrates was developed. Hardness, wear resistance and corrosive resistance were significantly improved at the deposited surface [33]. Pure vanadium carbide (VC) was deposited on 304 stainless steel with 20% VC in the middle as an intermediate buffer layer. The wear rate was decreased by more than 80%. The abrasive waterjet test was also performed, and the result demonstrated that the 100% VC coating made the entire structure much more resistant to wear resistance and corrosive resistance were significantly improved at the deposited surface [33]. Functionally graded NiCr-WC coating was deposited on AISI 1025 carbon steel by Amado et al. [35]. In this study, the NiCr-WC was a mixture of Delero 30 NiCrBSi alloy and WC. The outer layer was 60% WC while compositions of three types of intermediate layers were set to 15%, 30%, and 45%, respectively. The 15% interlayer resulted in no crack. In [36], 316L was coated by graded nickel-based hardfacing alloy Ni-Cr-B-Si with no presence of crack. The functionally graded coating on titanium and titanium alloys was also studied. According to the application of titanium alloy in the biomedical field, improvement can be made in both tribological and biomedical properties if there is a graded hard biocompatible coating on titanium alloys [37, 38], which will extend the application of titanium alloys as implants.

Table 3. Functionally graded coatings on metal substrates via DED.

<table>
<thead>
<tr>
<th>Substrate material</th>
<th>Graded coating materials</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>316 stainless steel</td>
<td>AlOx/Ni-20wt%Cr</td>
<td>[32]</td>
</tr>
<tr>
<td>Mild steel</td>
<td>Fe/SiC</td>
<td>[33]</td>
</tr>
<tr>
<td>304 stainless steel</td>
<td>VC/304 stainless steel</td>
<td>[34]</td>
</tr>
<tr>
<td>AISI 1025 Carbon Steel</td>
<td>Ni-Cr-B-Si alloy/VC</td>
<td>[35]</td>
</tr>
<tr>
<td>316L stainless steel</td>
<td>Ni-Cr-B-Si alloy/316L stainless steel</td>
<td>[36]</td>
</tr>
<tr>
<td>Ti-6Al-4V</td>
<td>Co-Cr-Mo alloy/Ti-6Al-4V</td>
<td>[37]</td>
</tr>
<tr>
<td>Ti</td>
<td>TiO2/Ti</td>
<td>[38]</td>
</tr>
</tbody>
</table>

3. DISCUSSIONS

3.1 Material compatibility

DED has made it possible to fabricate various metallic FGMs by in situ varying material composition layer by layer. However, for incompatible systems, cracks are still frequently observed when merging dissimilar metals due to brittle secondary phases formed by two metal components. The ideal case is that no intermetallic phase can be created, such as in Cu/Ni FGM, Cu and Ni are completely soluble [23]. However, in most other commonly used bimetallic systems such as Ti-Fe and Fe-Ni, intermetallic phases are very common. Therefore, the in-depth study on metallurgical properties of bimetallic systems should be done to assist the material selection.

For incompatible systems, joining stainless steel and titanium alloy is attractive, but it is also difficult due to multiple types of potential intermetallic phases. Adding additional elements between stainless steel and titanium alloy has been tried in multiple research works. Researchers in [16] used DED to fabricate the FGM from 100% Ti-6Al-4V to 100% 304L by introducing graded vanadium with 25% increment. Cracks were still observed which were caused by the intermetallic FeTi phase. In [15], another FGM was fabricated from 100% 304L to 100% Ti-6Al-4V. However, the printing stopped in the middle due to the formation of σ-FeV phase. Therefore, both FeTi compound and σ-FeV phase should be avoided and alternative gradient designs are needed. In [17], chromium was also added in the transition part by taking advantage of the compatibility between chromium and vanadium. Both FeTi phase and σ-FeV were avoided. Although chromium and vanadium were not in the form of compositional gradient, the work expanded the scope in finding suitable intermediate materials to join dissimilar metals. For incompatible systems like the Ti-Fe system, it needs more efforts to explore more suitable materials for intermediate layers.

3.2 Controllable process parameters

Due to the difference in material properties of different components, a flexible and controllable set of parameters is of great importance in building FGMs with desired compositions.

As a composite material which requires non constant and precise concentration of components, the control of powder mixing is an essential part. Powders of both components can be premixed according to the designed composition and fed by a single nozzle. However, due to the difference in material density and powder size, premixed powders under carrier gas will undergo powder segregation and the resultant composition will deviate from expectation. In [39], the premixed powder flow under argon gas was studied and relationships in density and powder size between different types of powders were discovered to address the separation issue. With a more sophisticated feeding system, powders made from different materials can be delivered individually by multiple nozzles [19, 27]. Placing different powders in different hoppers can make better control of composition and avoid segregation [27].

If the mixing and feeding of different powders are in good control, there should be a reasonable set of processing parameters to ensure the correct composition within the fabricated parts. Thus, thermal properties of different metal and ceramic powders such as melting temperature and the laser
absorption rate should be considered, especially for materials with a high melting point such as metal carbides, or low laser absorptivity such as copper [23]. A more adjustable laser power is needed to follow the graded composition to control the melting of different components. For metal-metal FGMs, in [21], a constant laser power was applied in the entire gradient zone. Due to the great difference in the melting point between Ti and Mo, plenty of Mo particles were still not fully melted in the gradient zone. A continuously optimized laser power was also suggested in [15] when researchers found vanadium inclusions in the high-vanadium region at the gradient between Ti-6Al-4V and 304L stainless steel.

For metal-ceramic FGMs, unmelted particles can also be responsible for crack initiation. Although higher ceramic concentration can improve the hardness, a limit should be controlled to reduce the number of unmelted particles. And the powder segregation between ceramics and metal matrix may also be the reason for crack formation [27]. For metal-ceramic FGMs using wire as the matrix, a large laser power is needed to fully melt the metal wire, which is different from powder [30]. The interaction between wire and powder within the melt pool under a varying laser power should be further investigated.

4. CHALLENGES AND CONCLUSION
DED has a huge potential in the manufacturing of metallic FGMs. Currently, the mechanisms of both DED and FGMs still need a complete understanding [40]. Although a few works have been done in the feasibility of joining dissimilar alloys, more studies are required to reliably predict the resultant phase to avoid unwanted phases and control the formation of wanted phases which can be used to tailor material properties. Currently, the rule of material choice in metallic FGMs has not been systematically studied. There are research needs for a comprehensive database for DED-processed metallic FGMs [41]. Optimizing multiple parameters for depositing different types of powders and wires is critical to reaching the best-joining effect, which is also a challenging topic.

Today, most reported works of DED-processed metallic FGMs are still feasibility studies with very simple specimens. Moving specimens to real functional parts need the comprehensive development of system control, and the collaboration with CAD model representation. Thus, a global method for path planning should be studied with the aid of modeling and numerical simulation to realize functional parts rather than simple specimens. This requires more advanced process control in the future. Until now, the knowledge about DED and FGMs is still limited. Model representation have not been largely applied to real additively manufactured metallic FGMs by DED.

Therefore, the development of metallic FGMs via DED calls for a joint effort among various areas, including the development of a complete metal AM systems for design and manufacturing, the in-depth theoretical knowledge in materials science especially in phase transformation, and computer modeling.

5. ACKNOWLEDGMENTS
This project was supported by National Science Foundation Grants #CMMI-1547042 and CMMI-1625736, and the Intelligent Systems Center (ISC) at Missouri S&T. Their financial support is greatly appreciated.

6. REFERENCES
[13] Carroll, Beth E., et al. "Functionally graded material of 304L stainless steel and inconel 625 fabricated by directed energy deposition: Characterization and


Online social media has become a popular way to get up to date information around the world and medical misinformation is broadcasted easily among social media communities. People who rely on these platforms to get news are easy to be misled. It is necessary to detect and control medical misinformation spread online. One of the particular challenges for medical misinformation detection on social media is to catch the characteristics of misinformation among mass daily generated messages. Most existing methods cannot solve this problem effectively since they merely rely on visual information or text information separately. In order to solve this problem, we proposed an end-to-end deep learning network which leverages multiple-modalities information including images, comments by the users and also hashtags they use. The proposed model consists of three main channels: the hashtag channel will produce hashtag essential information; the comment channel will generate comment features; the image branch will extract image features. The outputs of these three individual channels will be used to make final predictions. To evaluate the proposed models performance, a real-world social media dataset is conducted on Instagram. The experimental results show that our proposed multiple-modalities models can outperform the single-modality models and achieve more than 90% testing accuracy.

1 INTRODUCTION
Childhood vaccine hesitancy is a complex public health issue which affects the United States overall children health situation [1]. According to the 2017 National Immunization Survey-Child, the number of children who do not receive vaccine doses is increasing by around 1% gradually since 2011 [2]. The increasing child undervaccination proportion is linked to parental delay or refusal of vaccines [3]. Parents who have negative attitudes of vaccines are often affected by the beliefs such as distrust of pharmaceutical industries, doubts of vaccine safety, and vaccine potential side effects [4], [5]. More importantly, people who doubt about vaccine safety and refuse vaccines often seek health information online[6]. Instagram, which is the largest image-sharing online mobile community, is becoming popular among US people. We have found more than 8,400 hashtags (discussion topics) related to anti-vaccine (“antivax”) messages in six months. People who rely on mobile social media are susceptible to visual antivaccination messages. To control the spread of antivax messages through social media community, it is urgent and necessary to develop a detection system. As well known, deep learning technologies have been achieved huge successes in computer vision and natural languages processing areas. The deep learning neural networks can be trained to detect anti-vaccination messages in social media community automatically.

Fig. 1: Examples of Antivax Posts in Instagram
2 RELATED WORKS
To investigate antivax information spread among the Internet, some previous papers have been worked on related topics in social media [7], [8]. However, most of them are based on survey methods. However, deep learning technologies can help researchers model such medical misinformation on social media. To completely exploit useful information on social media, researchers have to employ multiple source information such as image and textual contents in social media messages.

To exploit image information, some powerful deep learning models have been developed such as VGG-16/19 [9], ResNet [10], DenseNet [11], Inception network [12]. They have achieved significant success in the computer vision community.

To exploit textual information, words, sentences, and even paragraphs need to be converted to numerical vectors and then can be processed by deep learning models [13], [14], [15], [16], [17]. Some famous embedding algorithms can help us convert words into vectors such as Word2Vec [18], GloVe [19], and BERT [20].

Multiple-modality information can be used to do event detection in social media. In [21], the authors use both images and textual information to detect fake news on Twitter and Weibo. Similarly, in this paper, we also consider both images and textual information. Besides, we will also consider hashtags information in the antivax message with attention mechanism [22], [23]. To our best knowledge, this paper is the first one about antivaccination information detection using deep learning technologies.

3 METHODOLOGY
In this section, we first introduce the three components of the proposed antivax messages detection model: the hashtag channel, the comments channel, and also the image channel, then describe how to integrate these three components to learn the combination feature representations.

3.1 Model Overview
The goal of our model is to propose a reliable deep learning model to represent and detect antivax posts in the social media community. As shown in the above Fig. 2, to achieve this, the proposed deep learning model contains two major components: the multiple-modalities feature extractor and the final classifier. First of all, general posts on Instagram includes textual contents and images. In textual contents, users will write down their comments to present their attitudes and also include hashtags. Hashtags can be considered as keywords of the entire comments. Most of the posts will contain hashtags to enlarge their post effectiveness, especially for those antivaxers. These hashtags are used to help other users to search for and also easy to have concentration meaning of these posts. Hashtags will be an important factor to make final predictions. Therefore, the multiple-modalities feature of hashtags, comments and images will be extracted and concatenated together to form the final combination feature representations. Moreover, a new attention mechanism computation method is proposed in this paper. It will help the model learn the representation of textual information effectively. A binary classifier will be built on top of the multiple-modalities feature extractor to make predictions.

3.2 Multi-modal Feature Extractor
3.2.1 Hashtag Channel
The hashtags sequence consists of several pieces of words to reflect the main idea of the comments. These hashtags begin with a hash sign—“#”. For example, #children, #health and so on. These hashtags normally have individual discrete meanings. If this post is related to antivaccination, it will contain hashtags of rejecting vaccination information such as #vaccines, #vaccinefree, #vaccineharm, #vaccinekills, #MMR, #autism, #vaccinetruth, #antivaccine, #vaccineinjury, #vaccineskill. Antivaccine posts will always attach these hashtags, but other posts will occasionally have some of these hashtags even in the pro-vaccine posts. Therefore, hashtags information is important but we cannot only use hashtags to detect antivax messages. To generate hashtags representation, we will first transfer these hashtags into word embedding vectors. The embedding word vectors are represented with pre-trained Global vectors (GloVe). We first choose top-N popular hashtags from Instagram and then represented them into N word embedding vectors. For the n-th word in the hashtags group, the corresponding k-dimensional word embedding vector is denoted as

\[ h_n \in \mathbb{R}^k \]

where, \( o_n \) is the one-hot encoding representation of the n-th antivax word. \( E_h \) is the pre-trained embedding matrix. \( h_n \) is the corresponding word embedding representation of the n-th antivax word. We will use the mean value of these N hashtags to create a vector representation of antivaccine information.

\[ \bar{h}_{anti} = \frac{1}{N} \sum_{n=1}^{N} h_n \]

For any given Instagram post, we will use GloVe embedding matrix to represent j-th hashtag words with embedding vectors- \( h_j \). Then, we also use the combination value of all the hashtags vectors to represent the whole meaning of hashtags in the post. However, we want to use the attention mechanism to see which part these hashtags should contribute more to reflect antivaccine information. This attention mechanism is highly related to antivaccine information.

\[ \alpha_j = \frac{\exp(\bar{h}_{anti}^T h_j)}{\sum_{m=1}^{J} \exp(\bar{h}_{anti}^T h_m)} \]

\[ h = \sum_{j=1}^{J} \alpha_j h_j \]

where, \( J \) stands for \( J \) hashtags in the post and \( h \) is the final vector representation of hashtags.
3.2.2 Comment Channel

The comments contain sequence words and also including punctuations and emojis. These symbols do not affect the sentence semantic representation. They will be ignored in this project. The sequence words will be first changed into vector representations with pre-trained GloVe embedding. After that, we will consider the sentence-level attention mechanism which is inspired by [13]. However, we do not randomly initialize a context and learn the representation during training. We will compute the similarity with combination antivaccine vectors $\tilde{h}_{\text{anti}}$. Second, we employ a bidirectional LSTM layer to catch the context information of each target words among comments and compute the attention weights with antivaccine hashtag vectors and finally create the embedding representation of comments representation.

$$w_t = E_c o_t, t \in [1, T]$$
$$\tilde{b}_t = \tilde{b}_{\text{LSTM}}(w_t), t \in [1, T]$$
$$\bar{b}_t = \bar{b}_{\text{LSTM}}(w_t), t \in [1, T]$$

where, $o_t$ is the one-hot encoding of the t-th word in comments. $w_t$ is the embedding representation of the t-th word in comments. $T$ denotes the amount of words in each comment. We obtain the context representation of the words using concatenating the forward and backward encoding vectors together. It is written as

$$b_t = [\tilde{b}_t; \bar{b}_t], t \in [1, T]$$

After that, we introduce the attention mechanism to extract such words in the sentence that are important to the contribution of antivaccine and aggregate the representation of those informative words to form the comments information.

$$\beta_t = \frac{\exp(\tilde{h}_{\text{anti}}^T b_t)}{\sum_{s=1}^{T} \exp(\tilde{h}_{\text{anti}}^T b_s)}$$
$$c = \sum_{t=1}^{T} \beta_t b_t$$

where, $c$ is the final vector representation of comments.

3.2.3 Image Channel

The attached images of the Instagram posts are fed into the image channel of the proposed model. In order to efficiently represent visual features, we employ the pre-trained VGG-19 model and add one more fully-connected layer on top of the VGG-19 model to generate fixed size visual features.

$$m = f(W_m \text{NET}_{\text{VGG19}}(I))$$

where, $m$ is the output concentrate features; $f$ is the activation function of the fully-connected layer; $W_m$ is the weight matrix need to be learned during training; $\text{NET}_{\text{VGG19}}$ is the high-level outputs of VGG-19 model; $I$ is the matrix of the attached image.

The hashtags representation features $h$, comments features $c$ and visual image features $m$ will be concatenated to form the final multiple-modalities feature representation denoted as $p$

$$p = [h; c; m]$$

After we get the final multiple-modalities feature representation $p$, it will be fed into fully-connected layers and then make final predictions. The final layer is activated with a sigmoid function.

4 EXPERIMENT AND RESULTS

In this section, we first introduce a real-world social media dataset used in these experiments, then present the proposed models detection accuracies, and finally analyze the performance of different models.

4.1 Datasets

To fairly evaluate our proposed model, we conduct experiments on a real-world social media dataset which is collected from the Instagram website. The data samples usually contain both pictures and text content. The text content contains comments posted by the account users and also hashtags to emphasize posts contents. The data samples are collected between January 2015 to March 2019 by Instagram API toolkit [24]. We collect 1250 antivaccine related posts and 1250 non-antivaccine posts. The non-antivaccine posts include pro-vaccine posts and also general daily life posts. Most of the posts contain images, comments, and hashtags. However, some of the posts do not always contain comments and/or hashtags.
The rise of mandatory and forced vaccination

Legislation should send a chill up every single
American spine. The government can force you to accept any kind of medicine and inject any drug your will is totalitarianism at its worst! This egregious form of state power is metastasizing.

Dr. Ron Paul

Fig. 3: Examples of basic case and hard case in our dataset

The posts contain pictures, comments and hashtags are treated as basic cases; those posts do not have comments and/or hashtags are considered as hard cases. Fig. 3 displays some examples of basic cases and hard cases. To fairly evaluate the proposed model performance, the data set includes both basic and hard data samples. The details of this Instagram dataset is shown in Table I.

<table>
<thead>
<tr>
<th></th>
<th>Antivaccine</th>
<th>Non-antivirus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic case</td>
<td>747</td>
<td>1050</td>
</tr>
<tr>
<td>Hard case</td>
<td>503</td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td>1250</td>
<td>1250</td>
</tr>
</tbody>
</table>

TABLE I: The details of Instagram datasets which includes both antivaccine posts and non-antivaccine posts.

We split the whole dataset into the training and testing sets in 8:2 ration. In other words, 2000 samples are used as the training set and the rest 500 samples are used as our testing samples. During training, we use data augmentation for images with rotation, translation, and different scales to enlarge the training set to 4000 samples.

4.2 Various Model Comparison

In the proposed model, we leverage multiple modalities information including image and textual information to detect antivax messages. To get better embedding representations of hashtags and comments, we employ the attention mechanism on textual branches. In this section, we compare the performance of three different categories models: single-modality model, multiple-modalities model, and multiple-modalities model with attention mechanism.

Single-modality model: we use pre-trained VGG-19 to extract high-level features, and then they will be fed into fully connected layers to make predictions. During training, we freeze all layers weights except the final convolutional layer.

Multiple-modalities model: In this model, we consider images, comments, and hashtags to generate combination features to represent social media messages. The model has three channels: a hashtag channel, a comment channel and an image channel. The hashtag channel uses a pre-trained embedding model to represent hashtags information; the comment channel uses the same pre-trained embedding model as the hashtag channel to represent words in comments. On top of the embedding vectors, we add a bidirectional LSTM layer to catch context information of comments. Then, outputs of the three channels will be concatenated together and fed into fully connected layers to make the final prediction.

Multiple-modalities model with attention mechanism: the model architecture is as same as the above one only except the outputs of the hashtag channel, and the comment channel will be modified with adding attention weights to let the model learn which parts of the textual information should be paid much attention during training.

4.3 Experiment Results Analysis

Table II shows the best experimental results of both the single-modality model and multiple-modality models. We can

©Ned Flanders 2014
find that the proposed multiple-modality model outperforms the single-modality model. The test accuracy rate can achieve a little bit higher if adding an attention mechanism in the model architecture.

<table>
<thead>
<tr>
<th></th>
<th>Single</th>
<th>Multiple</th>
<th>Multiple-Att</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0.83</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>Testing</td>
<td>0.80</td>
<td>0.90</td>
<td>0.92</td>
</tr>
</tbody>
</table>

TABLE II: The experimental results of the single-modality model, the proposed multiple-modality model, and the proposed multiple-modality model with attention mechanism

When training these models, to fairly compare models performance, we use the same experiment settings. (1) For both single-modality model and the image channel of the proposed multiple-modality model, we fine-tune the pre-trained VGG-19 model with weights trained on our dataset. A fully-connected layer with 128 hidden units is added on top of the pre-trained model; (2) For the hashtag channel, we represent the words with pre-trained GloVe word embedding model which trained on Twitter dataset; (3) For the comment channel, we first represent the words with pre-trained GloVe word embedding model which trained on Twitter dataset. To get context information of single words, we involve a bidirectional LSTM layer with 64 hidden units; (4) For the attention mechanism, we choose top-10 popular antivax related hashtags from Instagram and compute the combination vector representation $F_{att}$. It will help the proposed model to represent hashtags better when involving attention mechanism; (5) After concatenating the final features, they will be fed into a fully-connected layer with 128 hidden units. The final output will be got after a sigmoid function; (6) We choose Adam optimization algorithm with learning rate from $2 \times 10^{-5}$ to $1 \times 10^{-4}$. The loss function is based on binary cross-entropy. The models are trained to achieve the lowest loss value after around 50 epochs with batch size 32. Our proposed model can achieve above 90% testing accuracy which is better than the image based deep learning model. It proves that leveraging multiple-modalities information is useful for antivax message detection. Our dataset consists of some hard cases which do not have hashtags and/or comments. These data samples make the proposed model performs similarly to the image based model. It can also explain why the proposed model does not outperform much compared with the image-based model even adding an attention mechanism.

5 CONCLUSIONS

In this paper, we study medical misinformation detection on social media. To detect antivax messages on social media community, we first construct a real-word antivaccination messages dataset. Then, we propose a deep learning model involving multiple modality inputs including hashtags, comments, and images to classify whether a given Instagram post is antivax related or not. Also, we proposed a new way to compute the attention weights to help the model learn textual information representation. Based on our experiment results, it shows that the proposed multiple-modality model with an attention mechanism is effective and outperforms the single-modality model.

6 ACKNOWLEDGMENTS

The authors sincerely acknowledge the support of the Intelligent System Center, Missouri University of Science and Technology for this research project.

REFERENCES


