AFRL-AFOSR-VA-TR-2021-0042

DDDAS-as-a-Service: Dynamic Resource Management Algorithms and Systems Software for an Infosymbiotics Hosting Platform

Gokhale, Aniruddha
VANDERBILT UNIVERSITY
230 W 41ST STREET FL 7
NEW YORK, NY, 37203-2416
US

06/01/2021
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-06-2021	Final	01 Feb 2018 - 31 Jan 2021

4. TITLE AND SUBTITLE
DDDAS-as-a-Service: Dynamic Resource Management Algorithms and Systems Software for an Infosymbiotics Hosting Platform

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-18-1-0126

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Aniruddha Gokhale

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
VANDERBILT UNIVERSITY
230 W 41ST STREET FL 7
NEW YORK, NY 37203-2416
US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2021-0042

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This document reports progress on the research investigations of the AFOSR DDDAS project titled DDDAS-as-a- Service: Dynamic Resource Management and Systems Software for an Infosymbiotics Hosting Platform. It reports on research outcomes along multiple dimensions of research conducted in this project. This includes resource management solutions for applications that use the continuum of resources from the edge to the cloud, deployment optimizations, tooling and benchmarking efforts, and ongoing work on adversarial machine learning. Keywords: Dynamic resource management, model learning, simulation-based optimizations, cloud infrastructures for DDDAS applications.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			ERIK BLASCH
U	U	U	UU	9	**19b. TELEPHONE NUMBER** *(Include area code)* 426-7311

Standard Form 298 (Rev.8/98)
Prescribed by ANSI Std. Z39.18

DDDAS-as-a-Service: Dynamic Resource Management and Systems Software for an Infosymbiotics Hosting Platform

Final Report for AFOSR DDDAS #FA9550-18-1-0126

Faculty: Aniruddha Gokhale, Xenofon Koutsoukos and Eugene Vorobeychik‡
Students: Shashank Shekhar, Anirban Bhattacharjee, Yogesh Barve, Shweta Khare, Robert Canady
Department of Electrical Engineering and Computer Science
Vanderbilt University, Nashville, TN 37235, USA
Contact Email of PI: a.gokhale@vanderbilt.edu

This document reports progress on the research investigations of the AFOSR DDDAS project titled DDDAS-as-a-Service: Dynamic Resource Management and Systems Software for an Infosymbiotics Hosting Platform. It reports on research outcomes along multiple dimensions of research conducted in this project. This includes resource management solutions for applications that use the continuum of resources from the edge to the cloud, deployment optimizations, tooling and benchmarking efforts, and ongoing work on adversarial machine learning.

***Keywords*-Dynamic resource management, model learning, simulation-based optimizations, cloud infrastructures for DDDAS applications.**

I. Introduction

A. Motivation

The landscape of next-generation DoD missions as well as civilian missions, which tackle significant societal and environment problems (e.g., efficient and secure power grid or smart and connected cities), is rapidly changing. In this new context, applications must exploit the increasingly sensor-driven world in order to make informed decisions for their correct behavior. This sensor- and other edge devices-filled operating space has resulted in the Internet of Things (IoT) paradigm that is enabling a range of smart services in a variety of domains as shown in Figure 1.

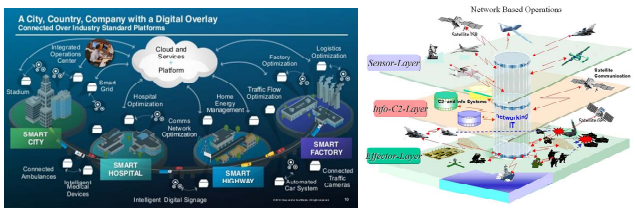

Fig. 1. Emerging Landscape driven by Internet-of-Things

‡Subawardee due to move to Washington University, St. Louis

A common requirement for these applications is autonomous resilience stemming from the substantial uncertainty in the operating environment and infeasibility of human intervention to address any system anomalies. Thus, applications are increasingly required to execute in a distributed set of resources ranging from the network/compute edge all the way to high performance compute clusters, such as HPx [1], and centralized cloud architectures. Moreover, the defining boundaries of the next-generation DDDAS applications can no longer be statically defined; rather application capabilities may themselves shrink or grow depending on mission needs and decisions based on instrumented data.

As an example, the author in [2] mentions that the United States Air Force was the second largest buyer of sensors between 2013–2015, which have been used primarily in munitions for guidance, and for platforms that provide intelligence, surveillance and reconnaissance (ISR) capabilities – all of which are heavily data-driven. A quote from another article [3] reflects the current DoD thought process: *"The Air Force wants to tap into the connected network of sensors and devices known as the Internet of Things, but doesn't want to constantly relay information back to a centralized data center to process it. Data culled from the Internet of Things could help improve communications and military logistics, as well as aid in monitoring service members' health, but the Air Force wants to process the data locally – closer to the point of collection."*

In these new operating scenarios, the following requirements hold: (a) applications expect the systems software to deliver the stringent response time requirements they impose on the execution of application logic despite fluctuations in availability of resources, such as CPU and network bandwidth, (b) application scale is expected to be large and distributed in nature, and (c) applications must make sense for their missions based on the very large volumes of information emitting from the numerous sensors, which in turn requires *Big Computing for Big Data processing.* These traits are representative of Infosymbiotics that is manifested in systems deployed across a spectrum of resources ranging from the centralized cloud all

the way to the sensors – a vision that has been laid out by Dr. Frederica Darema.

B. Research Focus and Contributions

The *dynamic data driven applications systems* (DDDAS) [4], [5] principles are a promising approach to address the need to manage and control the next generation of cyber-physical systems. Despite the substantial success of the DDDAS paradigm over the years, to date it has predominantly been applied in application domains where the sources of instrumentation are often defined statically with seldom any dynamic discovery of sources, and the tasks of model learning and simulation are conducted in resource-rich environments, such as high performance clusters and traditional cloud data centers. Even our earlier DDDAS project [6] was predicated on these assumptions where the models of individual DDDAS applications execute independently in a centralized cloud.

To address these current limitations, and expand the focus and applicability of the DDDAS paradigm to address the modern-day challenges, this project investigated the opportunistic use of the spectrum of resources ranging from the cloud to the edge to host applications and adaptively migrate its components based on a variety of factors including resource availability, QoS needs, workloads, contention from other applications and several other factors. It developed a number of tools and conducted benchmarking studies. Moreover, it also developed deployment optimizations for edge-based applications. Finally, it helped start a new research direction that combined resource management with adversarial machine learning. This report summarizes the accomplishments during the three year project. It also discusses opportunities for future research.

C. Project Team

The following team members participated on the project.

1) Aniruddha Gokhale, PI of the overall project at Vanderbilt University.
2) Xenofon Kousoukos, Co-PI at Vanderbilt University.
3) Yevgeniy Vorobeychik, Co-PI initially at Vanderbilt University and then as a subawardee at Washington University in St. Louis.
4) Graduate students: Shashank Shekhar, Anirban Bhattacharjee, Shweta Khare, Yogesh Barve and Robert Canady all at Vanderbilt University.

D. Report Organization

The rest of this report is organized as follows: Section II summarizes the cloud/edge resource management research including serverless serving and edge-based distributed model learning; Section III summarizes the tools and techniques developed during this research; Section IV summarizes ongoing work and future directions specifically in the context of adversarial perturbations to DDDAS models; Section V reports on DoD transition stories; Section VI reports on the PhD dissertations and publications stemming from this research; and finally Section VII provides concluding remarks.

II. Contributions to Cloud/Edge Dynamic Resource Management

This project made numerous contributions to enhancing the state-of-the-art in dynamic resource management for cloud/edge-based applications. This comprised migration of applications from the cloud to the fog, adaptive management of applications between the edge and fog considering user mobility, vertical elasticity, serverless serving, managing model re-learning at the edge and deployment optimizations at the edge. Each such effort is described briefly and the resulting publication(s) are mentioned.

*a) **Vertical Elasticity for Dynamic Resource Management**:* Elastic auto-scaling in cloud platforms has primarily used horizontal scaling by assigning application instances to distributed resources. Owing to rapid advances in hardware, cloud providers are now seeking vertical elasticity before attempting horizontal scaling to provide elastic auto-scaling for applications. Vertical elasticity solutions must, however, be cognizant of performance interference that stems from multi-tenant collocated applications since interference significantly impacts application quality-of-service (QoS) properties, such as latency. The problem becomes more pronounced for latency-sensitive applications that demand strict QoS properties. Further exacerbating the problem are variations in workloads, which make it hard to determine the right kinds of timely resource adaptations for latency-sensitive applications. To address these challenges and overcome limitations in existing offline approaches, we designed an online, data-driven approach which utilizes Gaussian Processes-based machine learning techniques to build runtime predictive models of the performance of the system under different levels of interference [7]. The predictive online models are then used in dynamically adapting to the workload variability by vertically auto-scaling co-located applications such that performance interference is minimized and QoS properties of latency-sensitive applications are met.

*b) **Performance Interference-aware Migration from Cloud to Fog**:* An increasing number of interactive applications and services, such as online gaming and cognitive assistance, are being hosted in the cloud because of the elastic properties and cost benefits of distributed data centers. Despite these benefits, the longer and often unpredictable end-to-end network latencies between the end user and the cloud can be detrimental to time-critical response to the applications. Although technology enablers, such as Cloudlets or Micro Data Centers (MDCs), are increasingly being leveraged by cloud infrastructure providers to address the network latency concerns, existing efforts in re-provisioning services from the cloud to the MDCs seldom focus on ensuring that the performance properties of the migrated services are met. This research demonstrates the application of Dynamic Data Driven Applications Systems (DDDAS) principles in the systems software layer to address these limitations by: (a) determining when to re-provision; (b) identifying the appropriate MDC and a suitable host within that MDC that meets the performance

considerations of the applications; and (c) ensuring that the cloud service provider continues to meet customer service-level objectives while keeping its operational and energy costs low [8], [9]. Empirical evaluations using a setup comprising a cloud data center and multiple MDCs composed of heterogeneous hardware validate the capabilities of the INDICES (Intelligent Deployment for ubiquitous Cloud and Edge Services) framework to process DDDAS methods. It should also be noted that the capabilitis created through INDICES are aimed to satisfy a broad set of applications requiring real-time data delivery and thus also satisfy the support requirements of environments enabling DDDAS-based applications.

*c) **High Availability Cloud/Edge Services to Support User Mobility**:* Fog/Edge computing is increasingly used to support a wide range of latency-sensitive Internet of Things (IoT) applications due to its elastic computing capabilities that are offered closer to the users. Despite this promise, IoT applications with user mobility face many challenges since offloading the application functionality from the edge to the fog may not always be feasible due to the intermittent connectivity to the fog, and could require application migration among fog nodes due to user mobility. Likewise, executing the applications exclusively on the edge may not be feasible due to resource constraints and battery drain. To address these challenges, we developed the URMILA resource management middleware that makes effective trade-offs between using fog and edge resources while ensuring that the latency requirements of the IoT applications are met [10], [11]. We evaluate URMILA in the context of a real-world use case on an emulated but realistic IoT testbed.

*d) **Performance-aware and Cost-effective Serverless Serving**:* Pre-trained deep learning models are increasingly being used to offer a variety of compute-intensive predictive analytics services such as fitness tracking, speech and image recognition. The stateless and highly parallelizable nature of deep learning models makes them well-suited for serverless computing paradigm. However, making effective resource management decisions for these services is a hard problem due to the dynamic workloads and diverse set of available resource configurations that have different deployment and management costs. To address these challenges, we developed a distributed and scalable deep-learning prediction serving system called Barista [12], which makes the following contributions. First, we show a fast and effective methodology for forecasting workloads by identifying various trends. Second, we formulate an optimization problem to minimize the total cost incurred while ensuring bounded prediction latency with reasonable accuracy. Third, we propose an efficient heuristic to identify suitable compute resource configurations. Fourth, we propose an intelligent agent to allocate and manage the compute resources by horizontal and vertical scaling to maintain the required prediction latency. Finally, using representative real-world workloads for an urban transportation service, we demonstrate and validate the capabilities of Barista.

*e) **Supporting Online Distributed Machine Learning at the Edge**:* Deep Learning (DL) model-based AI services are increasingly offered in a variety of predictive analytics services such as computer vision, natural language processing, speech recognition. However, the quality of the DL models can degrade over time due to changes in the input data distribution, thereby requiring periodic model updates. Although cloud data-centers can meet the computational requirements of the resource-intensive and time-consuming model update task, transferring data from the edge devices to the cloud incurs a significant cost in terms of network bandwidth and are prone to data privacy issues. With the advent of GPU-enabled edge devices, the DL model update can be performed at the edge in a distributed manner using multiple connected edge devices. However, efficiently utilizing the edge resources for the model update is a hard problem due to the heterogeneity among the edge devices and the resource interference caused by the co-location of the DL model update task with latency-critical tasks running in the background. To overcome these challenges, we developed Deep-Edge, a load- and interference-aware, fault-tolerant resource management framework for performing model update at the edge that uses distributed training [13]. DeepEdge makes the following contributions. First, it provides a unified framework for monitoring, profiling, and deploying the DL model update tasks on heterogeneous edge devices. Second, it presents a scheduler that reduces the total re-training time by appropriately selecting the edge devices and distributing data among them such that no latency-critical applications experience deadline violations. Experiments conducted using a real-world DL model update case-study based on the Caltech dataset and an edge AI cluster testbed validate the efficacy of the framework.

III. Tools and Techniques for Designing and Deploying Data-driven Applications

In this section we describe the tools and techniques we developed to make it easy for developers to develop and deploy a DDDAS-based system. Additionally, we discuss efforts at benchmarking on emerging hardware accelerator devices. Finally, we also describe load balancing and deployment optimization algorithms we defined to enable edge-based deployment of data-driven application components.

*a) **Deployment Orchestration Tooling**:* Users of cloud platforms often must expend significant manual efforts in the deployment and orchestration of their services on cloud platforms due primarily to having to deal with the high variabilities in the configuration options for virtualized environment setup and meeting the software dependencies for each service. Despite the emergence of many DevOps cloud automation and orchestration tools, users must still rely on specifying low-level scripting details for service deployment and management. Using these tools required domain expertise along with a steep learning curve. To address these challenges in a tool-and-technology agnostic manner, which helps promote interoperability and portability of services hosted across cloud platforms, we developed preliminary ideas on a GUI based cloud automation and orchestration framework called

CloudCAMP [14]. CloudCAMP uses model-driven engineering techniques to provide users with intuitive and higher-level modeling abstractions that preclude the need to specify all the low-level details. CloudCAMP's generative capabilities leverage a built-in knowledge-base to automate the synthesis of Infrastructure-as-Code (IAC) solution that subsequently can be used to deploy and orchestrate services in the cloud.

b) ***Tools for Streamlining Machine Learning****:* With the proliferation of machine learning (ML) libraries and frameworks, and the programming languages that they use, along with operations of data loading, transformation, preparation and mining, ML model development is becoming a daunting task. Furthermore, with a plethora of cloud-based ML model development platforms, heterogeneity in hardware, increased focus on exploiting edge computing resources for low-latency prediction serving and often a lack of a complete understanding of resources required to execute ML workflows efficiently, ML model deployment demands expertise for managing the lifecycle of ML workflows efficiently and with minimal cost. To address these challenges, we developed an end-to-end data analytics and serverless platform called *Stratum* [15], [16]. Stratum can deploy, schedule and dynamically manage data ingestion tools, live streaming apps, batch analytics tools, ML-as-a-service (for inference jobs), and visualization tools across the cloud-fog-edge spectrum.

c) ***Automating Benchmarking, Instrumentation and Model Building****:* Services hosted in multi-tenant cloud platforms often encounter performance interference due to contention for non-partitionable resources, which in turn causes unpredictable behavior and degradation in application performance. To grapple with these problems and to define effective resource management solutions for their services, providers often must expend significant efforts and incur prohibitive costs in developing performance models of their services under a variety of interference scenarios on different hardware. This is a hard problem due to the wide range of possible co-located services and their workloads, and the growing heterogeneity in the runtime platforms including the use of fog and edge-based resources, not to mention the accidental complexities in performing application profiling under a variety of scenarios. To address these challenges, we developed FECBench (Fog/Edge/Cloud Benchmarking), an open source framework comprising a set of 106 applications covering a wide range of application classes to guide providers in building performance interference prediction models for their services without incurring undue costs and efforts [17]. FECBench makes the following contributions. First, we developed a technique to build resource stressors that can stress multiple system resources all at once in a controlled manner, which helps to gain insights into the impact of interference on an application's performance. Second, to overcome the need for exhaustive application profiling, FECBench intelligently uses the design of experiments (DoE) approach to enable users to build surrogate performance models of their services. Third, FECBench maintains an extensible knowledge base of application combinations that create resource stresses across the multi-dimensional resource design space. Empirical results using real-world scenarios to validate the efficacy of FECBench show that the predicted application performance has a median error of only 7.6% across all test cases, with 5.4% in the best case and 13.5% in the worst case.

d) ***Insights on Emerging Hardware Accelerators****:* Hardware accelerator devices have emerged as an alternative to traditional CPUs since they not only help perform computations faster but also consume much less energy than a traditional CPU thereby helping to lower both capex (i.e., procurement costs) and opex (i.e., energy usage). However, since different accelerator technologies can illustrate different traits for different application types that run at the edge, there is a critical need for effective mechanisms that can help developers select the right technology (or a mix of) to use in their context, which is currently lacking. To address this critical need, we proposed a recommender system to help users rapidly and cost-effectively select the right hardware accelerator technology for a given compute-intensive task [18]. Our framework comprises the following workflow. First, we collect realistic execution traces of computations on real, single hardware accelerator devices. Second, we utilize these traces to deduce the achievable latencies and amortized costs for device deployments across the cloud-edge spectrum, which in turn provides guidance in selecting the right hardware.

e) ***Load Balancing Edge-based Streaming Applications****:* Advances in Internet of Things (IoT) give rise to a variety of latency-sensitive, closed-loop applications that reside at the edge. These applications often involve a large number of sensors that generate volumes of data, which must be processed and disseminated in real-time to potentially a large number of entities for actuation, thereby forming a closed-loop, publish-process-subscribe system. To meet the response time requirements of such applications, we developed techniques to realize a scalable, fog/edge-based broker architecture that balances data publication and processing loads for topic-based, publish-process-subscribe systems operating at the edge, and assures the Quality-of-Service (QoS), specified as the 90th percentile latency, on a per-topic basis [19]. The key contributions include: (a) a sensitivity analysis to understand the impact of features such as publishing rate, number of subscribers, per-sample processing interval and background load on a topic's performance; (b) a latency prediction model for a set of co-located topics, which is then used for the latency-aware placement of topics on brokers; and (c) an optimization problem formulation for k-topic co-location to minimize the number of brokers while meeting each topic's QoS requirement. Here, k denotes the maximum number of topics that can be placed on a broker. We show that the problem is NP-hard for $k \geq 3$ and present three load balancing heuristics. Empirical results validate the latency prediction model and the performance of the proposed heuristics.

f) ***Optimal Placement of Graph-structured Applications at the Edge****:* Many IoT applications found in cyber-physical systems, such as smart grids, must take control actions in response to critical events, such as supply-demand mismatch,

which requires low-latency processing of streaming data for rapid event detection and anomaly remediation. These streaming applications generally take the form of directed acyclic graphs (DAGs), where vertices represent operators and edges represent the flow of data between these operators. Edge computing has recently attracted significant attention as a means to readily meet the requirements of latency-critical IoT applications due to its ability to provide low-latency processing near the source of data. To accrue the benefits of edge computing, the constituent operators of these applications must be placed in a manner that intelligently trades-off inter-operator communication costs with the cost of interference incurred due to co-location of operators on the same resource-constrained edge devices. To address these challenges and to substantially simplify the placement problem for DAGs of arbitrary sizes and topologies, we developed an algorithm [20] that first transforms any arbitrary stream processing DAG into an approximate set of linear chains. Subsequently, a data-driven latency prediction model for co-located linear chains is used to inform the placement of operators such that the makespan, defined as the maximum latency of all paths in the DAG, is minimized. We empirically evaluate our algorithm using a variety of DAG placement scenarios on a Beagle Bone cluster, which is representative of an edge computing environment.

IV. Ongoing Work and Future Directions

This section reports on ongoing work and opportunities for future work that builds on our recent and ongoing accomplishments. A significant part of our ongoing work and plans for future work center around adversarial machine learning, detection and defense mechanisms, and its impact on application QoS. We surmise this will add another dimension of complexity to dynamic resource management.

a) ***Studying the Impact of Adversarial Perturbations on Data-driven Prognostics****:* Deep learning has shown impressive performance across a variety of domains, including data-driven prognostics. However, research has shown that deep neural networks are susceptible to adversarial perturbations, which are small but specially designed modifications to normal data inputs that can adversely affect the quality of the machine learning predictor. We studied the impact of such adversarial perturbations in data-driven prognostics where sensor readings are utilized for system health status prediction including status classification and remaining useful life regression [21]. We found that we could introduce obvious errors in prognostics by adding imperceptible noise to a normal input and that the hybrid model with randomization and structural contexts is more robust to adversarial perturbations than the conventional deep neural network. Our work demonstrated the limitations of current deep learning techniques in pure data-driven prognostics, which indicates a potential technical path forward. To the best of our knowledge, this work is the first to investigate the implications of using randomization and semantic structural contexts against current adversarial attacks for deep learning-based prognostics.

b) ***Dynamic Data Repair Approach to Defend Against Adversarial Perturbations****:* For power distribution networks with connected smart meters, current advances in machine learning enable the service provider to utilize data flows from smart meters for load forecasting using deep neural networks. However, recent research shows that current machine learning algorithms for power systems can be vulnerable to adversarial attacks, which are small designed perturbations crafted on normal inputs that can greatly affect the overall performance of the predictor. Even with only a partial compromise of the network, an attacker could intercept and adversarially modify data from some smart meters in a limited range to make the load predictor deviate from normal prediction results. In this work [22], we leveraged the dynamic data-driven applications systems (DDDAS) paradigm and proposed a novel data repair framework to defend against these kinds of adversarial attacks. This framework complements the predictor with a self-representative auto-encoder and works in an iterative manner. The auto-encoder is used to detect and reconstruct the likely adversarial part in the input data. Different reconstruction results come up given different sensitivity levels in detection. As new data flows in each iterative time step, the service provider continuously checks the error of the previous prediction step and dynamically trades off between different detection sensitivity levels to seek an overall stable data reconstruction. Case studies on power network load forecast regression demonstrated the vulnerability of current machine learning algorithms and correspondingly the effectiveness of our defense framework.

c) ***Deploying Adversarially Robust Computer Vision Deep Learning Models at the Computing Edge****:* Edge-based, deep learning computer vision applications, such as those used in surveillance or traffic management, are becoming increasingly important. However, realizing these applications incurs a number of challenges. First, the constraints on edge resources precludes the use of large-sized, deep learning computer vision models. Second, the heterogeneity in edge resource types causes different execution speeds and energy consumption during model inference. Third, deep learning models are known to be vulnerable to adversarial perturbations, which can make them ineffective or lead to incorrect inferences. Although the first two challenges have received some attention in recent years by researchers, defending against adversarial attacks at the edge remains mostly an unresolved problem. This paper addresses this unresolved challenge to realize robust and edge-based deep learning computer vision applications. In preliminary work, we have utilized state-of-the-art (SOTA) object detection attacks from the TOG (adversarial objectness gradient attacks) attack suite on YOLOv3 and YOLOv3-tiny to test the robustness of these object detection models. We also explore a variation of adversarial training that augments the training data with multiple types of attacks. Our solution is then evaluated using the PASCAL VOC dataset where we are able to improve the robustness of YOLOv3-tiny models by 1-2% mean average precision (mAP), which is more significant than it sounds due to the way mAP is calculated and factoring

in how well they performed on clean data. Full YOLOv3 realized an improvement of up to 17% mAP on attacked data.

*d) **Adversarial Robustness of Deep Sensor Fusion Models**:* Deep sensor fusion for perceptual data processing has become increasingly important in high-stakes applications, such as autonomous driving. However, while there is considerable literature investigating vulnerabilities of single-sensor modalities, such as camera or LiDAR, vulnerabilities of deep sensor fusion architectures have received less attention. In preliminary work, we systematically study the robustness of deep camera-LiDAR fusion architectures for 2D object detection in autonomous driving, and make several experimental observations. First, we find that the fusion model is usually both more accurate, and more robust against single-source attacks than single-sensor deep neural networks. Furthermore, we show that without adversarial training, early fusion is more robust than late fusion, whereas the two perform similarly after adversarial training. However, we note that single-channel adversarial training of deep fusion is often detrimental even to robustness. Moreover, we observe crosschannel externalities, where single-channel adversarial training reduces robustness to attacks on the other channel. Additionally, we observe that the choice of adversarial model in adversarial training is critical: using attacks restricted to cars' bounding boxes is more effective in adversarial training and exhibits less significant cross-channel externalities. Finally, we find that joint-channel adversarial training helps mitigate many of the issues above, but does not significantly boost adversarial robustness.

V. DoD Transition and Outreach Efforts

This section reports on collaborative activities with DoD researchers, transition targets, and student internships at DoD.

1) **Methods to AFRL/DoD and Industry:** Contributions made to the AFRL StreamlinedML project were transitioned via Lockheed Martin. Our team was a subcontract to Lockheed Martin. That project has since ended for Vanderbilt University. The algorithms and tools from the different synergistic research projects are available in open source at https://github.com/doc-vu.
2) **Students to employment (trained under grant):** Dr. Shashank Shekhar graduated around the time the new grant started. He was funded primarily by our previous DDDAS grant. Three students defended their PhD dissertation in January 2020. Dr. Yogesh Barve is now a Research Scientist at Vanderbilt University; Dr. Anirban Bhattacharjee is now a postdoc at the US National Institute of Science and Technology (NIST); and Dr. Shweta Khare is an engineer at Amazon. Graduate student and US citizen Mr. Robert Canady did his summer internship at Wright State working closely with AFRL researchers in Dayton, OH, and will continue his internship in Summer 2021.
3) **DoD Internships:** US citizen graduate student Mr. Robert Canady was selected for a summer 2020 internship at Wright State University to work with Wright Patterson AFRL researchers. He will continue his internship in Summer 2021. Mr. Canady's research focus is on dynamic resource management for adversarially robust DDDAS systems.
4) **Additional Collaborations:** Our team collaborates with other researchers in our institute working on multiple different DARPA projects. PI Xenofon Koutsoukos is leading the DARPA Assured Autonomy project and also leads the NSA Lablet project. We are exploring the possibility of new research on secure software-defined networking systems. Although our team was unsuccessful in securing a short-term project from AFRL, Rome to evaluate different datastores, we have preliminary work conducted in this area and we plan to approach the AFRL researchers with our findings.
   We are also interacting with researchers on multiple ongoing NSF projects in our institute. PIs Xenofon Koutsoukos and Eugene Vorobeychik (from Washington Univ St.Louis) are involved in Cyber Physical Systems. PIs Gokhale and Koutsoukos are part of a NSF AI Convergence Accelerator Phase 1 project focusing on streamlining AI/ML for medical image processing where we are applying our DDDAS and StreamlinedML ideas to solve domain-specific problems. Finally, PI Gokhale is part of a project funded by Cisco for data-driven anomaly detection in their router devices.

VI. Dissertations and Publications

In this section we list the dissertations and publications resulting from this project.

A. Dissertations

The following PhD dissertations of doctoral students supported in part by the DDDAS grant resulted from the research conducted for this project.

1) Yogesh Barve, (PhD Computer Science, Vanderbilt University, Jan 2020), *Principles and Techniques for Performance Management and Validation of CLoud Hosted Distributed Applications,*
   [Current Activities:] Research Scientist, Institute for Software Integrated Systems, Vanderbilt University,.
2) Anirban Bhattacharjee, (PhD Computer Science, Vanderbilt University, Jan 2020), *Automating Deployment and Management of Data Analytics Services Across Distributed Systems,*
   [Current Activities:] Postdoc sponsored by US National Institute for Standards and Technology (NIST), University of Maryland, College Park, MD, USA.
3) Shweta Khare, (PhD Computer Science, Vanderbilt University, Jan 2020), *Resource Management Algorithms for Edge-Based, Latency-Aware Data Distribution and Processing,*
   [Current Activities:] Employed at Amazon in Bay Area, CA, USA.
4) Shashank Shekhar, (PhD Computer Science, Vanderbilt University, May 2018), *Algorithms and Techniques for*

Dynamic Resource Management across Cloud-Edge Resource Spectrum

[Current Activities:] Employed at Siemens, Princeton, NJ, USA.

A US citizen student, Mr. Robert Canady was also supported by the grant and is currently focusing on adversarial machine learning and its impact on resource management as his PhD topic.

B. Publications

The following major publications listed in reverse chronological order have resulted from research conducted for this project.

1) Shashank Shekhar, Ajay Dev Chhokra, Anirban Bhattacharjee, Yogesh Barve, Shweta Khare, Guillaume Pallez, Hongyang Sun and Aniruddha S. Gokhale, INDICES: Applying DDDAS Principles for Performance Interference-aware Cloud-to-Fog Application Migration, in the *Handbook of Dynamic Data Driven Applications Systems,* Frederica Darema and Erik Blasch, Volume 2, Springer Nature publisher, 2021, To Appear.
2) Yi Li, Shashank Shekhar, Yevgeniy Vorobeychik, Xenofon D. Koutsoukos, Aniruddha S. Gokhale, "Simulation-Based Optimization as a Service for Dynamic Data-Driven Applications Systems," in the *Handbook of Dynamic Data Driven Applications Systems,* Frederica Darema, Erik Blasch, Alex Aved and Sai Ravela, 2nd Edition, Springer Nature publisher, 2021, To Appear.
3) Anirban Bhattacharjee, Yogesh Barve, Shweta Khare, Shunxing Bao, Zhuangwei Kang, Aniruddha Gokhale, and Thomas Damiano, Stratum: Towards Rapid Development and Deployment of Machine Learning Pipelines across Cloud-Edge, To Appear *Deep Learning for Internet of Things,* Al-Sakib Khan Pathan and Uttam Ghosh eds, CRC Press, 2021.
4) Xingyu Zhou, Robert Canady, Yi Li, Xenofon Koutsoukos, and Aniruddha Gokhale, Overcoming Stealthy Adversarial Attacks onPower Grid Load Predictions Through DynamicData Repair, Third International Conference on Dynamic Data-driven Applications Systems (DDDAS), Boston, MA, USA, Oct 2–4, 2020, pp. 102–109.
5) Xingyu Zhou, Robert Canady, and Aniruddha Gokhale, Overcoming Adversarial Perturbations in Data-driven Prognostics Through Semantic Structural Context-driven Deep Learning, Annual Conference of the Prognostics and Health Management Society (PHM), Virtual, Sept 2020, pp. 11.
6) Xingyu Zhou, Robert Canady, Shunxing Bao, and Aniruddha Gokhale, Cost-effective Hardware Accelerator Recommendation for Edge Computing, *3rd USENIX Workshop on Hot Topics in Edge Computing (HotEdge '20),* Virtual, June 25–26, 2020, pp. 7
7) Yogesh Barve, Himanshu Neema, Zhuangwei Kang, Hongyang Sun, Aniruddha Gokhale, and Thomas Roth, A Model-driven Middleware Integration Approach for Performance-Sensitive Distributed Simulations, The 23rd IEEE International Symposium on Real-time Distributed Computing (ISORC), 2021, Nashville, TN, USA, May 19–21, 2020, pp. 184–191.
8) Anirban Bhattacharjee, Ajay Dev Chhokra, Hongyang Sun, Shashank Shekhar, Shreyas Ramakrishna, Aniruddha Gokhale, Abhishek Dubey and Gabor Karsai, DeepEdge: An Efficient Framework for Deep Learning Model Update on Heterogeneous Edge, *4th IEEE International Conference on Fog and Edge Computing (ICFEC),* Melbourne, Australia, May 11–14, 2020, pp. 75–84.
9) Shashank Shekhar, Ajay D. Chhokra, Hongyang Sun, Aniruddha Gokhale, Abhishek Dubey, Xenofon Koutsoukos and Gabor Karsai, URMILA: Dynamically Trading-off Fog and Edge Resources for Performance and Mobility-Aware IoT Services, *Elsevier Journal of Systems Architecture (JSA),* vol. , no. , 2020, pp. 20, DOI:10.1016/j.sysarc.2020.101710.
10) Anirban Bhattacharjee, Yogesh Barve, Shweta Khare, Shunxing Bao, Zhuangwei Kang, Aniruddha Gokhale, and Thomas Damiano, STRATUM: A BigData-as-a-Service for Lifecycle Management of IoT Analytics Applications, *IEEE International Conference on Big Data (BigData' 19),* Los Angeles, CA, USA, December 9–12, 2019, pp. 1607–1612.
11) Shweta Khare, Hongyang Sun, Julien Gascon-Samson, Kaiwen Zhang, Yogesh Barve, Aniruddha Gokhale, and Xenofon Koutsoukos, Linearize, Predict and Place: Minimizing the Makespan for Edge-based Stream Processing of Directed Acyclic Graphs, *4th ACM/IEEE Symposium on Edge Computing,* Washington, DC, USA, Nov 7–9, 2019, pp. 1–14.
12) Anirban Bhattacharjee, Ajay Dev Chhokra, Zhuangwei Kang, Hongyang Sun, and Aniruddha Gokhale, BARISTA: Efficient and Scalable Deep Learning Prediction Serving using Serverless Computing, *IEEE International Conference on Cloud Engineering (IC2E),* Prague, Czech Republic, June 24–27, 2019, pp. 23–33.
13) Yogesh Barve, Shashank Shekhar, Shweta Khare, Anirban Bhattacharjee, Zhuangwei Kang, Hongyang Sun, and Aniruddha Gokhale, FECBench: A Lightweight Interference-aware Approach for Application Performance Modeling, *IEEE International Conference on Cloud Engineering (IC2E),* Prague, Czech Republic, June 24–27, 2019, pp. 211–221. **BEST PAPER AWARD.**
14) Shashank Shekhar, Ajay Dev Chhokra, Hongyang Sun, Aniruddha Gokhale, Abhishek Dubey, and Xenofon Koutsoukos, URMILA: Dynamically Trading-off Fog and Edge Resources for Performance and Mobility-Aware IoT Services, *IEEE Symposium on Real-time Computing (ISORC 2019),* Valencia, Spain, May 7–9, 2019, pp. 118–125.
15) Anirban Bhattacharjee, Yogesh Barve, Shweta Khare, Shunxing Bao, Aniruddha Gokhale, and Thomas Damiano, Stratum: A Serverless Framework for the Lifecycle

Management of Machine Learning-based Data Analytics Tasks, *2019 Usenix Conference on Operational Machine Learning (OpML),* Santa Clara, CA, USA, May 20, 2019, pp. 3.

16) Yogesh Barve, Shashank Shekhar, Shweta Khare, Anirban Bhattacharjee, and Aniruddha Gokhale, UPSARA: A Model-driven Approach for Performance Analysis of Cloud-hosted Applications, 11th IEEE/ACM International Conference on Utility and Cloud Computing (UCC), Zurich, Switzerland, Dec 17–20, 2018, pp. 1–10.
17) Shweta Khare, Hongyang Sun, Kaiwen Zhang, Julien Gascon-Samson, Aniruddha Gokhale, Xenofon Koutsoukos, and Hamzah Abdel Aziz, Scalable Edge Computing Architectures for Low Latency Data Dissemination in Topic-based Publish/Subscribe, *The Third ACM/IEEE Symposium on Edge Computing (SEC),* Bellevue, WA, USA, Oct 25–27, 2018, pp. 214–227.
18) Shashank Shekhar, Hamzah Abdel-Aziz, Aniruddha Gokhale, and Xenofon Koutsoukos, Performance-Aware Vertical Elasticity for Latency-Sensitive Applications, *IEEE International Conference on Cloud Computing (CLOUD),* San Francisco, CA, USA, July 2–7, 2018, pp. 82–89.
19) Anirban Bhattacharjee, Yogesh Barve, Aniruddha Gokhale, and Takayuki Kuroda, CloudCAMP: A Platform for Automating Deployment and Management of Cloud Services, *IEEE International Conference on Services Computing (SCC), Work-in-Progress Track,* San Francisco, CA, USA, July 2–7, 2018, pp. 237–240.

VII. Conclusions

This document summarizes the major accomplishments made by our team on our AFOSR DDDAS #FA9550-18-1-0126 grant titled *DDDAS-as-a-Service: Dynamic Resource Management and Systems Software for an Infosymbiotics Hosting Platform.* The key research, educational and transition outcomes from this project are the following:

- Developed data-driven approaches to dynamic resource management across the continuum from the edge to the cloud.
- Developed a data-driven approach to support both model training and model inference in a cluster of edge devices.
- Developed a data-driven approach that minimizes the makespan of information processing flows and packs them in minimum number of edge devices, which helps to utilize edge devices optimally for real-time mission-critical applications.
- Defined a reusable framework that enables the scientific community to conduct benchmarking of systems based on Design of Experiments, collecting data from various experimental results, training performance interference models of multi-tenant systems from the instrumented data, and using the models in an inferencing stage for dynamic resource management thereby completing the entire DDDAS feedback loop.
- Starting new research directions in applying DDDAS principles to handle adversarial attacks on data-driven models.
- Conducted preliminary work on utilizing hardware accelerator devices for DDDAS applications and strategies to handle adversarial attacks on models.
- Defined strategies for robust sensor fusion for visual perception. This was work conducted by our subawardee from Washington University in St. Louis.
- Graduated four PhD students and one US citizen graduate student is currently making progress on his doctoral work.
- Collaborated with AFRL researchers and DoD researchers from Lockheed Martin for the AFRL StreamlinedML program.
- Student summer internship at Wright State/Wright Patterson AFRL in summer 2020 and 2021.
- Continuing to explore opportunities to transition and build on the existing work for additional DoD opportunities, such as NSA, DARPA and AFRL.

Research artifacts from this and synergistic projects are available at https://github.com/doc-vu.

Acknowledgments

A major part of this work was supported by AFOSR DDDAS Grant FA9550-18-1-0126. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AFOSR. We are thankful to the Program Manager of DDDAS Program, Dr. Erik Blasch, the former AFOSR Director Dr. Frederica Darema, other DDDAS PIs, and anonymous reviewers of our publications for their helpful feedback that helped improve the quality of our research.

References

[1] T. Sterling, D. Kogler, M. Anderson, and M. Brodowicz, "SLOWER: A performance model for Exascale computing," *Supercomputing Frontiers and Innovations*, vol. 1, pp. 42–57, Sep 2014.
[2] P. Goldstein, "DOD, DHS and NASA Are Driving Adoption of Internet of Things Sensors," http://www.fedtechmagazine.com/article/2016/06/dod-dhs-and-nasa-are-driving-adoption-internet-things-sensors, Jun. 2016.
[3] M. Ravindranath, "Air Force Wants to Analyze Internet of Things Data in Real-time," http://www.nextgov.com/emerging-tech/2016/03/air-force-wants-analyze-internet-things-data-real-time/126750/, Mar. 2016.
[4] F. Darema, "Dynamic Data Driven Applications Systems: A New Paradigm for Application Simulations and Measurements," *Computational Science-ICCS 2004*, pp. 662–669, 2004.
[5] E. Blasch, S. Ravela, and A. Aved, *Handbook of dynamic data driven applications systems*. Springer, 2018.
[6] A. Gokhale, X. Koutsoukos, and D. Schmidt, "Stochastic Hybrid Systems Modeling and Middleware-enabled DDDAS for Next-generation US Air Force Systems," aFOSR DDDAS-funded project (#FA9550-13-1-0227, Sept 1, 2013–Sept 30, 2016).
[7] S. Shekhar, H. A. Aziz, A. Bhattacharjee, A. Gokhale, and X. Koutsoukos, "Performance Interference-Aware Vertical Elasticity for Cloud-Hosted Latency-Sensitive Applications," in *IEEE International Conference on Cloud Computing (CLOUD)*, San Francisco, CA, USA, Jul. 2018, pp. 82–89.
[8] S. Shekhar, A. Chhokra, A. Bhattacharjee, G. Aupy, and A. Gokhale, "INDICES: Exploiting Edge Resources for Performance-Aware Cloud-Hosted Services," in *IEEE 1st International Conference on Fog and Edge Computing (ICFEC)*, Madrid, Spain, May 2017, pp. 75–80.

[9] S. Shekhar, A. D. Chhokra, A. Bhattacharjee, Y. Barve, S. Khare, G. Pallez, H. Sun, A. Gokhale, and G. Karsai, *DDDAS Handbook Volume 2*. Springer, 2020, ch. INDICES: Applying DDDAS Principles for Performance Interference-aware Cloud-to-Fog Application Migration.

[10] S. Shekhar, A. D. Chhokra, H. Sun, A. Gokhale, A. Dubey, and X. Koutsoukos, "URMILA: Dynamically Trading-off Fog and Edge Resources for Performance and Mobility-Aware IoT Services," in *To Appear in the IEEE Symposium on Real-time Computing (ISORC 2019)*, Valencia, Spain, May 2019, p. 8.

[11] S. Shekhar, A. D. Chhokra, H. Sun, A. Gokhale, A. Dubey, X. Koutsoukos, and G. Karsai, "URMILA: Dynamically Trading-off Fog and Edge Resources for Performance and Mobility-Aware IoT Services," *Elsevier Journal of Systems Architecture (JSA)*, Jan. 2020.

[12] A. Bhattacharjee, A. D. Chhokra, Z. Kang, H. Sun, and A. Gokhale, "BARISTA: Efficient and Scalable Deep Learning Prediction Serving using Serverless Computing," in *To Appear in the IEEE International Conference on Cloud Engineering (IC2E)*, Prague, Czech Republic, Jun. 2019, p. 10.

[13] A. Bhattacharjee, A. D. Chhokra, H. Sun, S. Shekhar, S. Ramakrishna, A. Gokhale, A. Dubey, and G. Karsai, "Deep-Edge: An Efficient Framework for Deep Learning Model Update on Heterogeneous Edge," in *4th IEEE International Conference on Fog and Edge Computing (ICFEC)*. Melbourne, Australia: IEEE, May 2020, pp. 75–84.

[14] A. Bhattacharjee, Y. Barve, A. Gokhale, and T. Kuroda, "A Model-driven Approach to Automate Utility of Cloud Services Deployment and Management," in *International Workshop on Clouds and (eScience) Applications Management (CloudAM), Collocated with the 11th IEEE/ACM International Conference on Utility and Cloud Computing (UCC)*, Zurich, Switzerland, Dec. 2018, pp. 109–114.

[15] A. Bhattacharjee, Y. Barve, S. Khare, S. Bao, A. Gokhale, and T. Damiano, "Stratum: A Serverless Framework for Lifecycle Management of Machine Learning based Data Analytics Tasks," in *USENIX Conference on Operational Machine Learning (OpML '19)*, Santa Clara, CA, USA, May 2019, p. 2.

[16] A. Bhattacharjee, Y. Barve, S. Khare, S. Bao, Z. Kang, A. Gokhale, and T. Damiano, "STRATUM: A BigData-as-a-Service for Lifecycle Management of IoT Analytics Applications," in *IEEE International Conference on Big Data (BigData' 19)*, Los Angeles, CA, USA, Dec. 2019, pp. 1607–1612.

[17] Y. Barve, S. Shekhar, S. Khare, A. Bhattacharjee, Z. Kang, H. Sun, and A. Gokhale, "FECBench: A Lightweight Interference-aware Approach for Application Performance Modeling," in *To Appear in the IEEE International Conference on Cloud Engineering (IC2E)*, Prague, Czech Republic, Jun. 2019, p. 10.

[18] X. Zhou, R. Canady, S. Bao, and A. Gokhale, "Cost-effective Hardware Accelerator Recommendation for Edge Computing," in *3rd USENIX Workshop on Hot Topics in Edge Computing (HotEdge 20)*. Virtual Event: USENIX Association, Jun. 2020, p. 8.

[19] S. Khare, H. Sun, K. Zhang, J. Gascom-Samson, A. Gokhale, and X. Koutsoukos, "Scalable Edge Computing Architectures for Low Latency Data Dissemination in Topic-based Publish/Subscribe," in *3rd ACM/IEEE Symposium on Edge Computing (SEC)*, Bellevue, WA, USA, Oct. 2018, pp. 214–227.

[20] S. Khare, H. Sun, J. Gascon-Samson, K. Zhang, Y. Barve, A. Gokhale, and X. Koutsoukos, "Linearize, Predict and Place: Minimizing the Makespan for Edge-based Stream Processing of Directed Acyclic Graphs," in *4th ACM/IEEE Symposium on Edge Computing*, Washington, DC, USA, Nov. 2019, pp. 1–14.

[21] X. Zhou, R. Canady, Y. Li, and A. Gokhale, "Overcoming Adversarial Perturbations in Data-driven Prognostics Through Semantic Structural Context-driven Deep Learning," in *Annual Conference of the Prognostics and Health Management Society (PHM)*. Nashville, TN, USA: Springer, Sep. 2020, p. 11.

[22] X. Zhou, R. Canady, Y. Li, X. Koutsoukos, and A. Gokhale, "Overcoming Stealthy Adversarial Attacks on Power Grid Load Predictions Through Dynamic Data Repair," in *Third International Conference on InfoSymbiotics/DDDAS 2020*. Boston, MA: Springer, Oct. 2020, pp. 102–109.