ERDC TR-21-19

US Army Corps
of Engineers®
Engineer Research and
Development Center

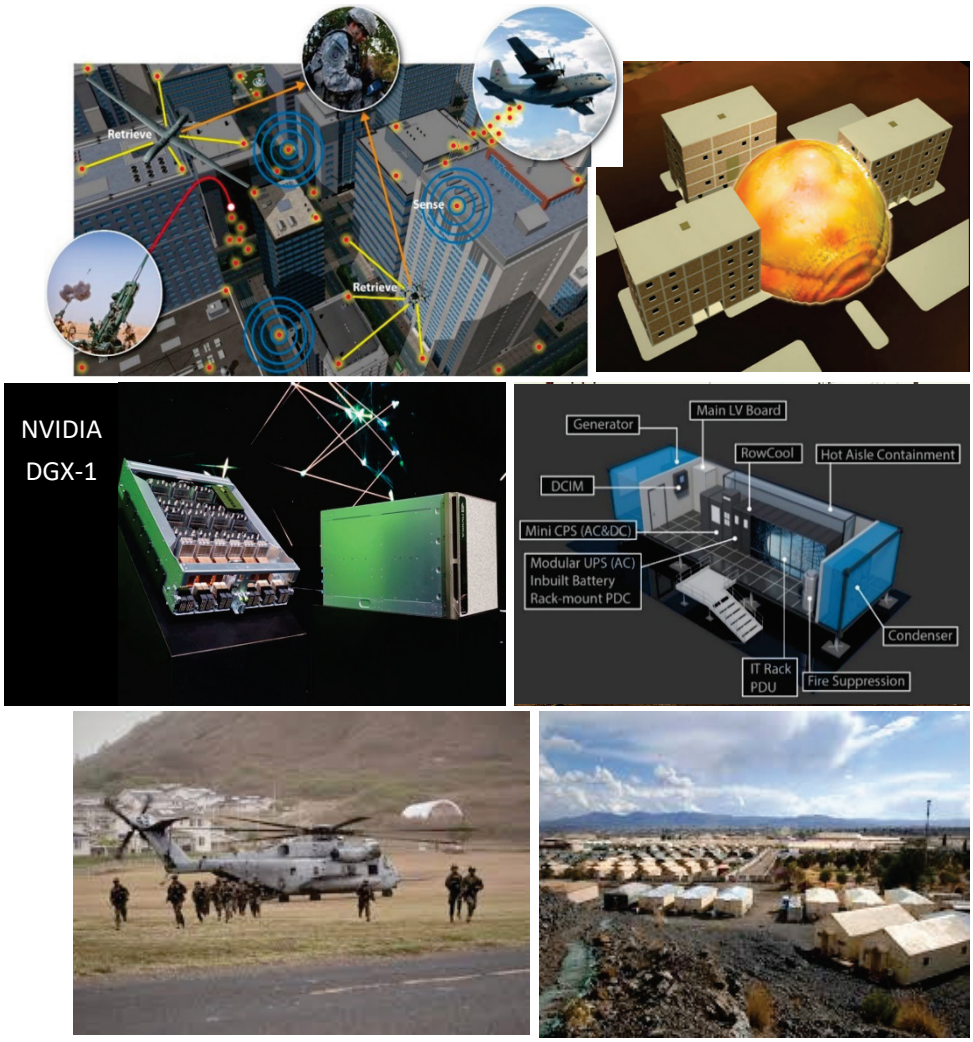
ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

Accelerating the Tactical Decision Process with High-Performance Computing (HPC) on the Edge

Motivation, Framework, and Use Cases

Alicia I. Ruvinsky, Timothy W. Garton, Daniel P. Chausse,
Rajeev K. Agrawal, Harland F. Yu, and Ernest L. Miller

September 2021



Engineer Research and
Development Center

The U.S. Army Engineer Research and Development Center (ERDC) solves the nation's toughest engineering and environmental challenges. ERDC develops innovative solutions in civil and military engineering, geospatial sciences, water resources, and environmental sciences for the Army, the Department of Defense, civilian agencies, and our nation's public good. Find out more at www.erdc.usace.army.mil.

To search for other technical reports published by ERDC, visit the ERDC online library at https://erdclibrary.on.worldcat.org/discovery.

ERDC TR-21-19
September 2021

Accelerating the Tactical Decision Process with High-Performance Computing (HPC) on the Edge

Motivation, Framework, and Use Cases

Alicia I. Ruvinsky, Timothy W. Garton, Daniel P. Chausse,
Rajeev K. Agrawal, and Ernest L. Miller

Information Technology Laboratory
U.S. Army Engineer Research and Development Center
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Harland F. Yu

Geospatial Research Laboratory
U.S. Army Engineer Research and Development Center
7701 Telegraph Road
Alexandria, VA 22315-3864

Final report

Prepared for U.S. Army Corps of Engineers
Washington, DC 20314-1000

Under ERDC FLEX-4, Funding account code U4371831

Abstract

Managing the ever-growing volume and velocity of data across the battlefield is a critical problem for warfighters. Solving this problem will require a fundamental change in how battlefield analyses are performed. A new approach to making decisions on the battlefield will eliminate data transport delays by moving the analytical capabilities closer to data sources. Decision cycles depend on the speed at which data can be captured and converted to actionable information for decision making. Real-time situational awareness is achieved by locating computational assets at the tactical edge.

Accelerating the tactical decision process leverages capabilities in three technology areas: (1) High-Performance Computing (HPC), (2) Machine Learning (ML), and (3) Internet of Things (IoT). Exploiting these areas can reduce network traffic and shorten the time required to transform data into actionable information. Faster decision cycles may revolutionize battlefield operations.

Presented is an overview of an artificial intelligence (AI) system design for near-real-time analytics in a tactical operational environment executing on co-located, mobile HPC hardware. The report contains the following sections, (1) an introduction describing motivation, background, and state of technology, (2) descriptions of tactical decision process leveraging HPC problem definition and use case, and (3) HPC tactical data analytics framework design enabling data to decisions.

Contents

Figures and Tables

Figures

Tables

Preface

This report is a deliverable product under the FLEX-4 (FIF) program, which is funded by the Programs Office (PO). The FIF is administered by the U.S. Army Engineer Research and Development Center (ERDC).

The work was performed by efforts of the Scientific Software Branch (SSB), Sensor Integration Branch (SIB), and Computational Analysis Branch (CAB) of the Computational Science and Engineering Division (CSED), ERDC, Information Technology Laboratory (ITL), Vicksburg, MS and the Information Generation and Management Branch (IGMB) of the Topography, Imagery, and Geospatial Research Division (TIGRD), ERDC, Geospatial Research Laboratory (GRL), Alexandria, VA.

At the time of publication, Mr. Timothy W. Dunaway was Chief, SSB; Mr. Quincy G. Alexander was Chief, SIB; Dr. Jeffery L. Hensley was Chief, CAB; and Ms. Katlyn N. Castillo was Chief, IGMB. Dr. Jerrell R. Ballard, Jr. was Chief, CSED; and Ms. Martha E. Kiene was Chief, TIGRD. The Deputy Director of ITL was Ms. Patti S. Duett and the Director was Dr. David A. Horner. The Deputy Director of GRL was Ms. Valerie L. Carney and the Director was Dr. Gary W. Blohm.

COL Teresa A. Schlosser was the Commander of ERDC, and Dr. David W. Pittman was the Director

1 Introduction

Background

The approach to artificial intelligence (AI) on the tactical edge is subdivided into three major research areas: (1) investigating hardware and architecture for a mobile and/or portable high-performance computing (HPC) environment that extends to the tactical edge; (2) investigating machine learning (ML) algorithms and internet of things (IoT) architectures and their consistency with portable HPC; and (3) operational workflow framework design for tactical decision processes with HPC on the edge.

- HPC Edge Architecture: This work includes investigating the state-of-the-art hardware, software, networking, and storage that are needed to provide and support data science capabilities within a mobile and/or portable HPC environment.
- ML Edge Analytics: Machine learning has already transformed data analytics with capabilities to find and analyze latent features of the large dataset. Can we revolutionize how data is analyzed on the battlefield, and then build capabilities that allow ERDC to solve this new class of problems?
- HPC on the Edge: This work aims to integrate the proposed innovation in hardware capability with the innovative analytic capability to most optimally and appropriately provide computationally intense decision support analytics on the tactical edge. The approach taken to address this task will be that of exposing an operational process while tackling an operational need. In other words, as the team directly builds tactical edge analytics for a real-world battlefield data processing need, the operational workflow undergone to implement the tactical edge analytic will begin to define a generalizable operational workflow for tackling tactical decision analytic problems with HPC on the edge.

Motivation

The rapid pace and diffuse nature of technological innovation (AI) across industries and nations is weakening the U.S.' monopoly on defense applications of technology. Advantages gained are temporary, as near-peer adversaries are able to rapidly leverage these increasingly ubiquitous capabilities. China laid out a development plan to become the world leader

in AI by the year 2030 and create a $150 B industry. Russian president Vladimir Putin stated, “Whoever becomes the leader in this sphere (AI) will become the ruler of the world” (Gigova 2017). The Department of Defense (DoD) must leverage AI and automation quickly and decisively to enable U.S. forces to operate more effectively and efficiently. Appropriately harnessed, through strategically-focused intent and effort, AI could become the greatest offset – alternatively, lacking vision and direction, it could become a strategic deficit.

AI has become a growing source of both solutions and problems in every type of enterprise. AI-based techs have become integrated into so many sectors, often causing huge changes, and it cannot be ignored from a military or global security standpoint.

Early AI questions for potential military applications (through the lens of a global security group):

1. What military applications of AI are likely in the near-term?
2. Of those, which are potentially consequential for the stability of strategic deterrence? Does AI change how we consider deterrence?
3. How does AI-assisted military systems affect regional stability?
4. What is the connection between regional stability and strategic deterrence?
5. What are the risks of unintended consequences and strategic surprises from AI?

The current defense community lacks common schema, terminology, and baselines for what constitutes AI. AI has numerous applications and approaches, but one of the biggest and most widely used is in analyzing “big data” in a way that provides value in one form or another. Narrow AI uses discrete problem-solving tools to perform specific narrow tasks. General AI encompasses technologies that are designed to mimic and recreate functions of the human brain. Narrow AI has had greater adoption and been shown to provide value in a variety of sectors. Crossover of AI into business applications has empowered predictive analytics in data-rich areas. The glut of sensors and cameras, coupled with existing data-centric resources, are driving AI opportunities to reveal hidden insights.

Objective

There is clearly a latent military potential to AI, but how long will it be until it actually shows up in practice? The DoD established a Joint AI Center in June 2018, leading to increased funding and research. Military applications with direct analogs to industry/academia AI applications (logistics, planning, analysis, transportation, etc.) have quickly adopted AI-supported data analytics throughout the defense and intelligence communities. Warfighting has separate and distinct applications that have had slower AI adoption. The primary categories of AI applications as it applies to warfighting are (1) those with effects primarily at the operational level of war and (2) those with effects primarily at the strategic level of the war. "AI applications at the operational level of war could have a very significant impact on the use of general-purpose military forces to achieve tactical objectives, and thus on the credibility of conventional deterrence. AI applications at the strategic level could have significant influence on political decisions about the scale and scope of war, escalation and de-escalation, and, by extension, strategic stability and deterrence" (Davis 2019).

A key focus of the DoD's strategy is identifying critical pathfinders to enable the Department to achieve its vision for AI:

- Continuing the efforts of the Algorithmic Warfare Cross Functional Team (AWCFT), known as Project Maven, to rapidly integrate and deploy commercial and government-developed AI capabilities in support of the counter-ISIS campaign.
- Establishing a DoD AI-focused center to innovate and deploy operational prototypes of AI systems across multiple areas of the defense enterprise and identify pathways to continuously apply AI technologies to a variety of use cases.
- Delivering world-class computational power at the tactical edge. The DoD will regain competitive advantage in embedded HPC to support "algorithmic warfare" by focusing on forward-deployable HPC in shipping containers to support in-theater tactical operations. Designs leverage world-class, power-efficient architectures to fuse tens of thousands of information sources.

The DoD's AI goals are:

- Establish cross-cutting foundations for AI. The Department will strengthen cross-cutting foundations and develop legal and policy frameworks to ensure it can successfully apply AI. It must emphasize interoperability in the systems it develops, while ensuring safety and security in their deployment. Additionally, it must pursue international, commercial, and academic partnerships to the greatest extent possible, while growing and nurturing its own capable AI workforce.
- Achieve military technology superiority. The Department will focus its investments and leverage commercial and academic investments in order to achieve military technology superiority over its adversaries in key areas, including core AI, machine learning, robotics, data analytics, advanced computing, and human-AI collaboration.
- Transform key DoD business functions. The Department will modernize and streamline its business operations by heavily leveraging commercial AI/ML products and investments.
- Build, field, and maintain AL/ML-based capabilities providing military superiority of the battlefield. To increase lethality, DoD will invest in applied capabilities development to link emerging technology with specific military capabilities and concepts of operation, rapidly transition capabilities, and focus on key application areas.

Three potential applications already identified (and being currently researched) for Operational goals:

- Omnipresent and omniscient autonomous vehicles
    - High priority for military applications of AI
    - Focus on unmanned systems at all levels (land, sea, air)
    - Conduct sophisticated battle tactics, adjust rapidly, report changes
- Big-Data-Driven M&S and Wargaming
    - AI has already seen some application to nuclear weapon systems
    - Increased interest in Wargaming approaches of AI, in order to explore and understand how dynamic conditions affect outcomes and decision making
- Focused Intelligence Collection and Analysis

  - Streams of intelligence data needs to be analyzed faster (without information overload)
  - Increased amount of data (coupled with increased variety of data from disparate sources) is requiring AI for collection and analysis problems

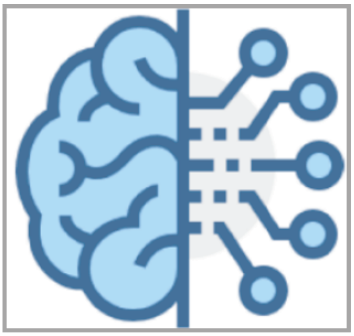
Four potential applications already identified (and being currently researched) for Strategic goals:

- System of Systems enabling Intelligence, Surveillance, and Reconnaissance (ISR)
  - Object identification is just the beginning; ISR is also needed for multi-domain situational awareness
  - Holistic awareness is the goal and is critical as the battlefield extends to all domains
  - Project Maven (Algorithmic Warfare Cross-Functional Team)
  - Military version of IoT can enable military advantages in traditional factors: speed and range
- Precision targeting of strategic assets
  - AI-empowered ISR makes it possible to locate, track, and target entities on the battlefield, thereby enabling the possibility of strikes against enemies
  - Makes the fundamental precepts of deterrence based on mutual vulnerability less certain until better understood
- Effective Missile Defense
  - Enable better target acquisition, tracking, and discrimination
  - Prevent ballistic-missile attack
- AI Guided Cyber
  - Discovery of network/data vulnerabilities by creating AI-guided probing, mapping, and hacking systems for defensive strategies
  - Offensive AI for location and collection, disruption, or disinformation
  - Defensive AI for detection of intrusions and search for debilitating anomalies

These all directly impact the speed of war. “The speed of war has changed, and the nature of these changes makes the global security environment even more unpredictable, dangerous, and unforgiving. Decision space has

collapsed, and so our processes must adapt to keep pace with the speed of war" (Garamone 2017). It has been said that “the military must make the most of its decision space, so military leaders can present options at the speed of war” (Garamone 2017). This can be done by “establishing a framework that enables senior leaders to make decisions in a timely manner” (Garamone 2017).

Figure 1. Humans versus AI.



Humans Excel at:
- Common sense
- Morals
- Imagination
- Compassion
- Abstractions
- Dilemmas
- Dreaming
- Generalization

AI Excels at:
- Harvesting Knowledge
- Pattern Identification
- Natural Language
- Machine Learning
- Eliminating Bias
- Endless Capacity

Approach

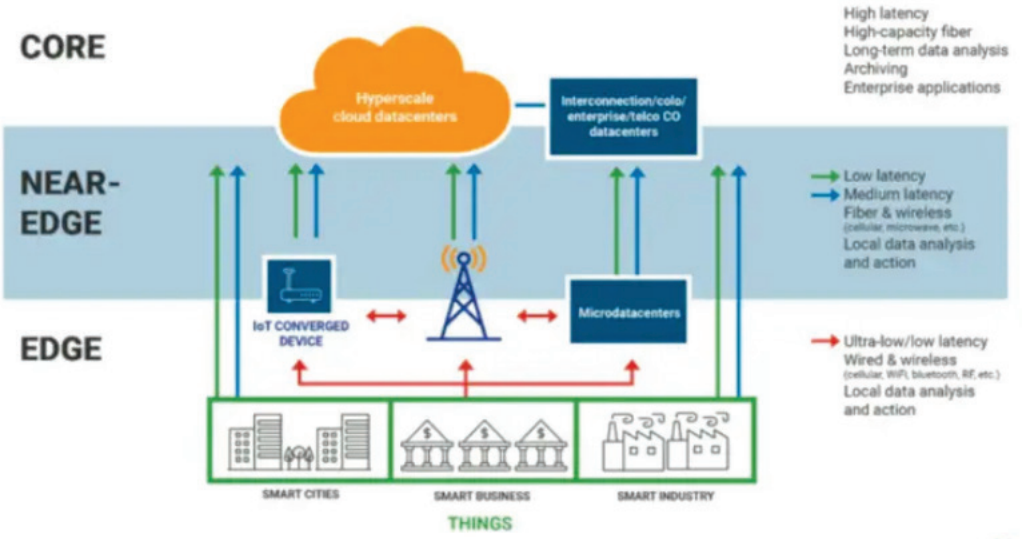
AI is a critical piece to the overarching goal of meeting the speed of war changes head on (Figure 1). AI provides commanders with situational understanding, as well as opens the door to efficient and effective processing, exploitation, and dissemination of information at speeds impossible for humans to match. AI allows systems to learn and adapt, accelerates operational tempo, makes soldiers smarter, and operates autonomously. Battlefield systems need to be resilient and capable of adapting to ever-changing situations. Growth of information has exceeded humans’ ability to rapidly analyze it and apply it to the decision-making process. Soldiers and machines work as an integrated cognitive system capable of greatly expanding the depth and breadth of data analysis. Machines focus on analyzing low-level details and allowing soldiers to focus their attention on higher-level strategy and planning.

AI has multiple potential effects on deterrence and stability. It also has a strong potential to erode stability by increasing perceived risk of surprise attack. Distorted data may lead AI systems to take unintended actions. AI does make mistakes, but these must not cause strategic instability or unnecessary escalation. AI's speed could accidentally accelerate something that could be de-escalated by other efforts. It could also misrepresent intentions when given different parameters pertaining to foreign and friendly platforms. AI is just one piece of the larger puzzle in keeping up with the speed of war.

Defining the edge

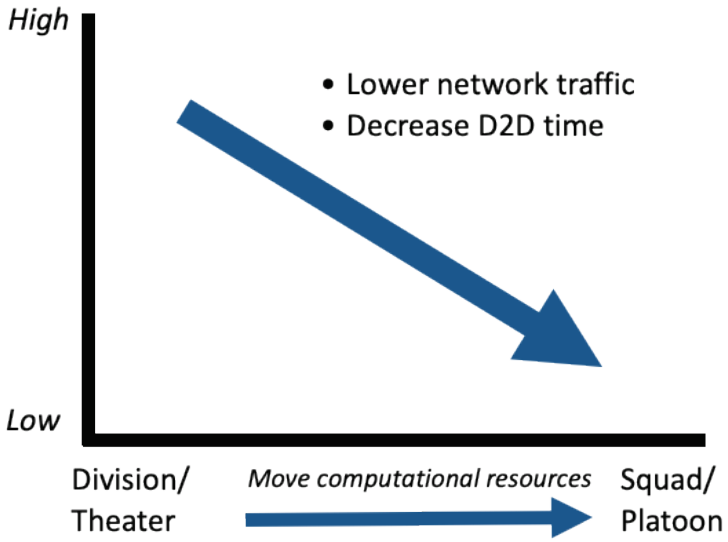
The definition of the edge is relative to the position within the enterprise. For example, mobile network operators (MNOs) consider the edge as the end of their radio access network (RAN) and a large opportunity for multi-access edge computing (MEC). Datacenter service providers may view edge as infrastructure deployed at key locations to minimize communication latency. HPC and edge server vendors view devices at remote sites as their edge. The data processing requirements can range from ultra-low latency and real-time latency at the edge, through medium latency and local data processing at the 'near edge,' to high latency and high-capacity storage and networking in centralized datacenters.[1]

Figure 2. Defining the edge.



[1] 451 Research Report: Data Analytics at the Edge, 2019.

Figure 3. AI on edge effect on network limitations.

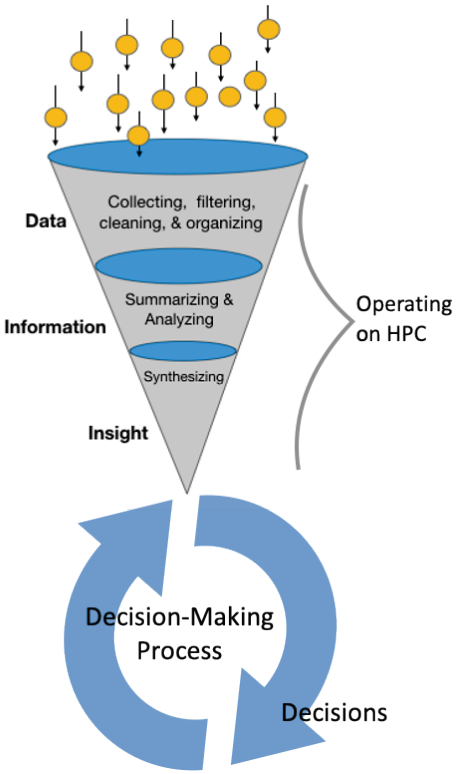


AI on-edge computing will decrease network traffic demands by taking the computation to the data rather than having to move the data to these powerful computer resources. This will decrease the amount of time it will take to move from data to decisions (Figure 2). Edge computing utilizes smaller and more powerful equipment when moving computational resources closer to the sources of data, thus reducing network traffic and bringing the computational resources closer to the front line (Figure 3). It brings together computations needed to analyze data to the sources of data; Data to Decision (D2D) time is reduced, insights and information extracted from data are highly compressed and easier to share, and more computing power is available in smaller footprint devices. The conventional method collects data at lower-echelons (tactical edge) and transfers data to high-echelons for processing and analysis. This method results in problems scaling as data sources continue to grow. It burdens the networks by transferring the unprocessed data. Data transfers become a major bottleneck, which impacts battlefield tempo.

Objective

An effective command-and-control (C2) system must provide the user with an operational picture, support the planning process, and enable the reception, processing, and transmission of information.

Figure 4. AI and HPC.



The objective of this effort is to investigate AI and HPC working together to advance key military capabilities, such as situational awareness (Figure 4). HPC plus AI on the Edge provides near real-time exploitation of local data, supports a high level of autonomy, computing assets are moved closer to data sources, increases the depth and breadth of the data analysis, reduces time to make decisions, and insight provides ground for decisions.

Organization of this report

This report is organized into six chapters:

- Chapter 1 represents an overview of the situation at hand and the purpose of the report.
- Chapter 2 provides the background and current available software and hardware infrastructure.
- Chapter 3 investigates the potential military use cases to be investigated.
- Chapter 4 provides a proposed approach to the problem.
- Chapter 5 summarizes the report.
- Chapter 6 covers future work plans and development.

2 Background

There are two aspects of this system that are essential in order to be able to enhance the decision-making process. The AI support will only be as good as the Edge devices and the IoT communications between them.

Edge computing

There are clear distinctions between edge computing and traditional cloud computing. The motivating idea to edge computing is to bring computing closer to the data source (Satyanarayanan 2017). As a new computing paradigm, it changes traditional computing archetypes by moving computing to the edge of the network.

Definitions of edge computing abound and can be found in Cao et al. (2020). Shi et al. (2016) and Shi and Dustdar (2016) defined the concept of edge computing with respect to cloud and IoT computing as, "Edge computing refers to the enabling technologies allowing computation to be performed at the edge of the network, on downstream data on behalf of cloud services and upstream data on behalf of IoT services. Here we define "edge" as any computing and network resources along the path between data sources and cloud data centers." In Shi et al. (2019) the benefits of the new paradigm are described. "The edge of the Internet is a unique place. Located usually just one hop away from associated end devices, it offers ideal placement for low-latency offload infrastructure to support emerging applications such as augmented reality, public safety, connected and autonomous driving, smart manufacturing, and healthcare" (Shi et al. 2019).

Satyanarayanan, from Carnegie Mellon University in the United States, presents edge computing very similarly to Shi et al. (2019) indicative of the consensus of defining this capability in the research community. "Edge computing is a new computing model that deploys computing and storage resources (such as cloudlets, micro data centers, or fog nodes, etc.) at the edge of the network closer to mobile devices or sensors" (Satyanarayanan 2017). Moving towards a slightly more technical characterization of edge computing capability, Liu et al. (2019) describes edge computing systems as "manag[ing] various resources along the path from the cloud center to end devices, shielding the complexity and diversity of hardware and helping developers quickly design and deploy novel applications."

The definition by Liu et al. (2019) introduces the idea of a path between data center and data producer. The tactical edge can be identified along that path as “the first tactical mile,” or the warfighters at the “tip of the spear” who are directly involved in mission execution (Staff 2005). The Committee on National Security Systems (CNSS) defined the tactical edge as "The platforms, sites, and personnel (U.S. military, allied, coalition partners, first responders) operating at lethal risk in a battle space or crisis environment characterized by 1) a dependence on information systems and connectivity for survival and mission success, 2) high threats to the operational readiness of both information systems and connectivity, and 3) users are fully engaged, highly stressed, and dependent on the availability, integrity, and transparency of their information systems” (Hansche 2019). As edge computing capabilities continue to evolve, the impact on tactical edge is undergoing major transformations (Shilawat 2018).

Figure 5. Nvidia DGX-1.

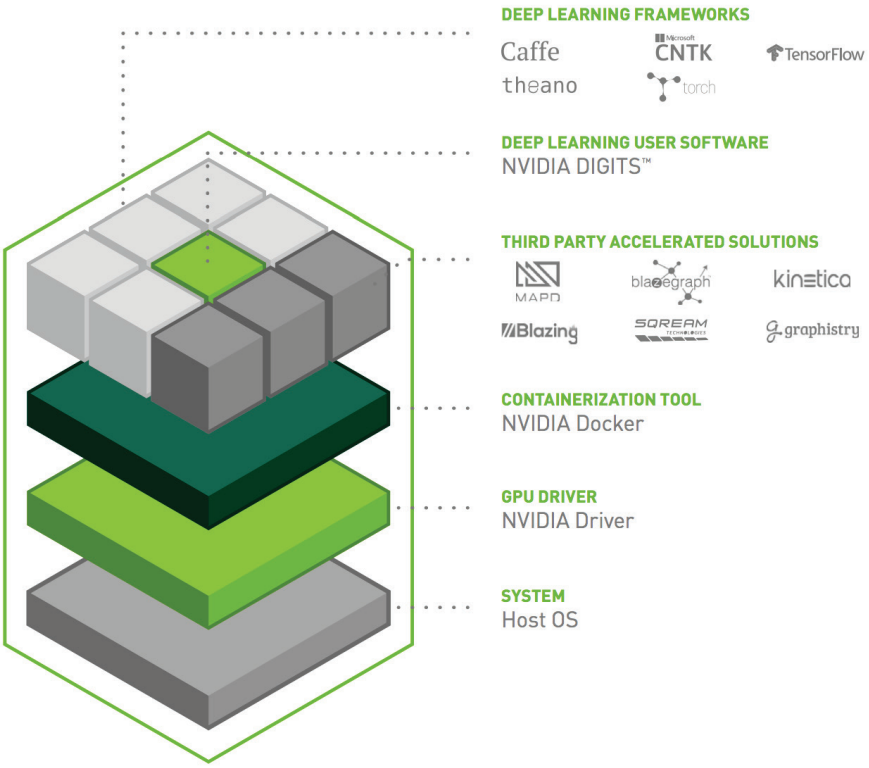
Figure 6. Militarized mobile HPC.

A major component for edge computing available to the DoD is the Nvidia DGX-1 AI supercomputer (Figure 5). This supercomputer has been militarized and mobilized for edge computing through DoD High Performance Computing Modernization Program (HPCMP) efforts (Figure 6).

Table 1. Processing ability comparison.

	Dual Xeon	DGX-1
FLOPS (CPU + GPU)	3 TF	170 TF
AGGREGATE NODE BW	476 GB/s	768 GB/s
ALEXNET TRAIN TIME	150 Hours	2 Hours
TRAIN IN 2 HOURS	>250 Nodes	1 Node

Figure 7. Nvidia DGX-1 software stack.



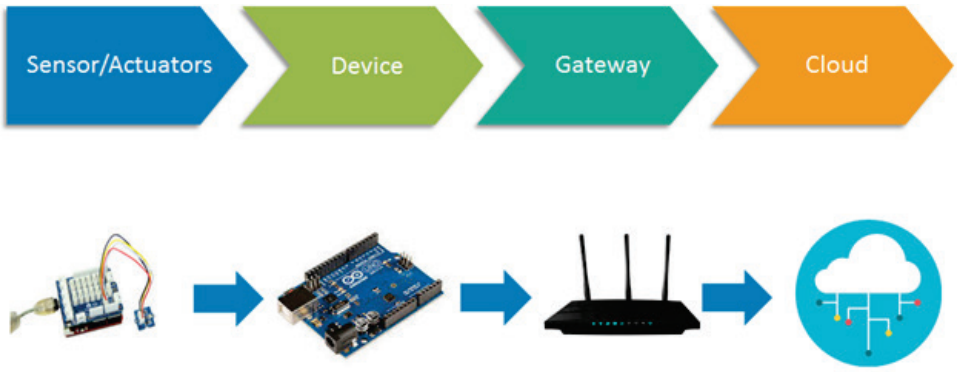
The DGX-1 is considered a datacenter in a box (Table 1). It has been engineered specifically for AI software, contains the densest computing node ever produced, and is equivalent to 250 servers. DGX-1 uses an approach termed GPGPU (General-Purpose computing on Graphics Processing Units). CUDA platform is a software layer that provides the AI direct access to the GPU's virtual instruction set and parallel computational elements, for the execution of compute kernels. DGX-1 utilizes Full Deep Learning OpenSource Software Stack, Docker-based software containers, NVIDIA-maintained Container Registry, Most popular deep learning frameworks available and optimized, Ubuntu-based for ease of administration, and Full SDK support for custom application development (Figure 7).

Internet of Things

The rapid swell in the number of connected devices has facilitated increased interest in the IoT as a means of gaining value for a number of proposed use cases in industry, academia, and government. On the battlefield, the IoT can be used to contribute capability that individual devices cannot provide on their own. Four major types of IoT have been identified, each with specific requirements and different kinds of values provided to the stakeholders. Three of these types; Massive, Broadband, and Critical IoT, have direct correlations when considered for direct military application and therefore provide value to the warfighter and the overall mission when implemented in the battlespace. The fourth kind, Industrial IoT, can also provide value to the DoD as a means of improving production and repair schedules, improving availability, and determining degradation, but, as of yet, has little direct effect on the use cases of the warfighter and command in the battlespace.

Before discussing the three kinds of enabling IoT types directly, it is important to note the overall architecture of the IoT in terms and roles that are easily understood. Figure 8 provides a generic overview of IoT architectures. It is important to note that every IoT implementation includes edge devices that act as the primary data sources operating within some defined environment, internet gateways that act as a means to collect the data from the variety of edge devices, edge IT that provides the ability to pre-process data for data centers or do quick-turn analytics on streaming data, and data centers/clouds that act as a place to store, analyze, and aggregate data. All stages are underpinned by basic needs that face any modern network, including security protocols and protections, definitions of the ecosystem that manages how devices and stages interact, and services that provide value from the data captured. Any IoT implementation must include all of these things at a minimum. In the next few sections, the focus will be placed on how different IoT focuses necessitate different technologies and associated infrastructures.

Figure 8. IoT Architecture.



Massive IoT

The addition of super-computing edge resources, such as mobile deployable HPCs, enables Massive IoT. The collection and collation of the available data provides massive datasets that can be examined by powerful computing resources in order to understand the operating environment as a whole and make decisions accordingly. Enabling large-scale data analytics that provide insights is the primary goal of Massive IoT approaches. Properly trained AI algorithms could provide additional value in a Massive IoT use case. One example would be the identification of latent features of the data when considered across multiple time stamps or locations that provide an insight as to enemy movements or tactics.

Massive IoT leverages the immense amount of collective data being created by a variety of low-cost edge devices in order to provide holistic insights into the operating environment. While each edge device only captures a subset of the data, the smaller volumes are gathered at the enterprise level where they can be examined using more powerful computing resources. The flexibility of large-scale computing resources would enable a variety of AI and ML use cases, as well as providing generalized capability for other computationally taxing efforts.

Broadband IoT

Stronger network connections with low latency and high throughput, such as 5G networks, enable Broadband IoT. The large data volumes traversing the network allow the enterprise level to understand the operating space in near real-time and make decisions with up-to-date information. AI algorithms can again provide value within a Broadband IoT use case,

potentially by coordinating a swarm of autonomous drones to react across a network connected battlespace.

Broadband IoT utilizes the faster connection speeds of an advanced network in order to enable connections from the enterprise to the edge devices in a way that allows them to react quickly and efficiently. The ability to leverage the computing power at the enterprise level enables more complicated reactions than in critical IoT, and it aids in the orchestration of multiple edge devices to meet a common goal. The availability of the feedback from the enterprise provides flexibility in what the edge devices can accomplish and how it can be accomplished.

Critical IoT

The connection between edge devices, such as a variety of sensors and devices observing the same events in real-time, enables Critical IoT. Critical IoT allows Artificial Intelligence (AI) applications to leverage the immediacy and availability of incoming data to react quickly to changes in the operating space. A prime example in the military space is the identification and geolocation of potential threats. The speed and connectivity of the small local network connecting several small-edge devices on soldiers could provide multiple data points for an AI application to consider, thereby improving accuracy and speed in providing the information to the users and making their reaction more effective.

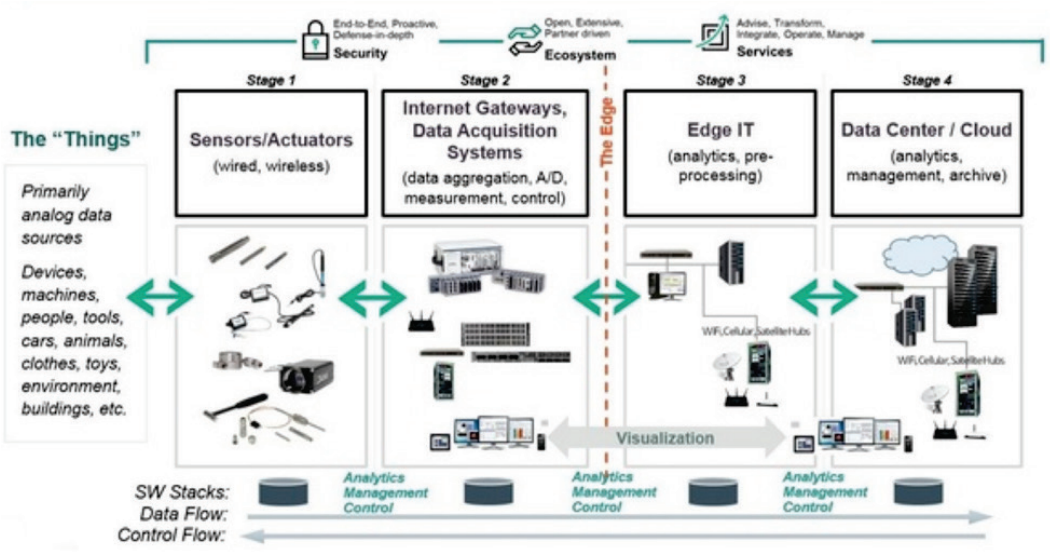
An entity interacting with its environment will often need to make decisions about how to respond to incoming signals, but it is in the high-leverage situations where Critical IoT becomes a necessity. A self-driving car utilizes a number of sensors in order to detect and identify objects in its path as it maneuvers on a street, and it must make the decision to stop in real-time. The car does not have time to connect to external systems as in Broadband IoT or analyze the breadth and width of data in Massive IoT to decide what action to take. The system must respond immediately and thus it must understand the desired outcomes, the incoming signals it might encounter, and how to interpret the signals from potentially disparate sources to achieve its goal.

3 Tactical Decision Process Leveraging HPC

It is important to note that while many of the use cases require trained AI and ML algorithms to be properly implemented, it is not a requirement of IoT to include AI and ML in order to provide value.

However, AI at the Edge does require an existing IoT architecture (Figure 9). All AI/ML approaches require data, both for training and to interpret when deployed. Thus, every AI on the Edge use case would require a data source, which in these scenarios can only be reasonably provided by edge devices. The connection between the edge devices and the AI/ML algorithms is at the discretion of the practitioner, but some form of the basic IoT architecture is required.

Figure 9. The 4 Stage IoT solutions architecture.



Use case 1: Cybersecurity

The fundamental feature of edge computing is providing data analytics capability locally rather than sending all the traffic to the cloud and relay the results back to the edge. This reduces the latency due to sending the data to a remote, centralized location and receiving the analytics results back. Edge-based cybersecurity analytics tools will be able to deliver better performance and efficiency and lower cost. However, Cybersecurity edge data analytics also require the systems that are compliant to the government regulations and be aware of security breaches that can occur at the edge. The computing power and storage at the edge cannot match that of traditional data centers,

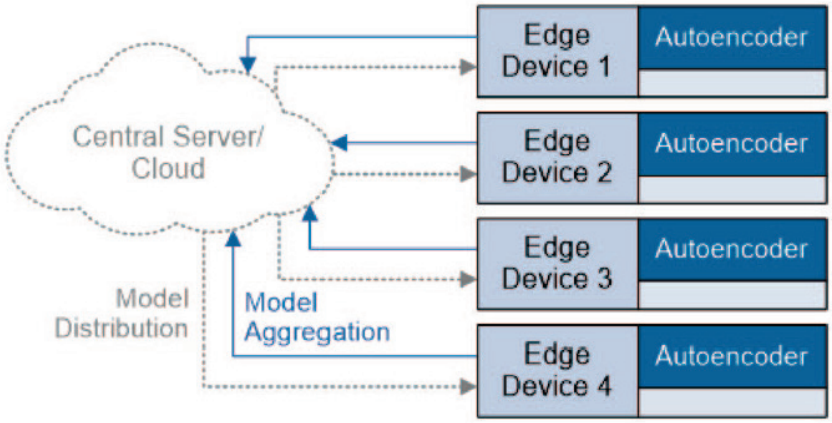
therefore, the analytics applications that can be deployed at the edge will need to be carefully selected. The best approach is prioritizing the applications at the edge versus in the remote cloud.

The emerging trend of "edge computing" also brings cybersecurity challenges and opportunities to the researchers. In Pan and Yang (2018), the following five challenges were identified:

1. Massive numbers of vulnerable IoT devices.
2. NFV-SDN Integrated Edge Cloud Platform.
3. Data Privacy and Security.
4. Offloading and Interaction between Edge and IoT Devices.
5. Trust and Trustworthiness.

Any edge deployment must consider the above challenges as well as the ones specific to their own environment.

Figure 10. Architecture of anomaly detection at the edge.



Technica uses autoencoders on each edge device to identify anomalies and learn from new observations and update the model. The updated models from each edge device are aggregated, sent to the centralized server, and distributed back to the edge devices (Schneible and Lu 2017). Figure 10 represents the high-level architecture of this model.

One potential use-case was created while researching cybersecurity edge analytics solutions to the warfighters. In this case, the cybersecurity data is collected from a large number of sensors, saved on DVDs, and sent to a contracting company for analytics. The entire process is not only very slow, but also has the potential of data being lost/stolen in transit. One

cybersecurity data analytics project that has the potential to be deployed on the edge and can benefit from HPC capability is Deep Learning-based Cybersecurity Log Analytics. The DoD currently maintains a network known as the Defense Research Engineering Network (DREN), which provides various DoD sites across the nation connectivity to HPC resource centers. To ensure the security of the DREN system, a customized HPCMP Intrusion Detection System (IDS) was developed. Over time, HPCMP IDS has accumulated massive quantities of valuable cybersecurity data, which necessitates a form of automation in the process of reviewing this data. The current research is primarily focused on utilizing the Zeek intrusion detection system data, but a variety of other datasets are collected within the HPCMP IDS.

The dataset is composed primarily of Zeek-IDS log files, separated by log type. Included among the Zeek data types are http, dns, and conn log files. The alert files are stored in JSON format to be examined by human analysts, who then label them as either "normal" or "bad," where normal indicates benign network activity, and bad indicates potentially malicious activity. These JSON files are downloaded and converted into a .csv format for deep learning model development. The majority of work done so far has focused on the http dataset, as it contains both a substantial number of alerts, as well as a relatively even split of normal and malicious alerts. The data undergoes several preprocessing steps before training, including feature selection, data encoding, normalization, and deletion of redundant records. A simple neural network-based model was developed on Onyx using a single GPU node.

Table 2. Model results.

Data Type	Size of data before encoding	Size of data after encoding	Accuracy
http	(201977,15)	(201977, 16854)	92.4%
dns	(261642, 22)	(261642, 6127)	96.9%
conn	(200000, 14)	(200000, 7682)	84.4%

Input to the neural network requires the data to be in a numerical format. Raw strings are not valid input and must be transformed or "encoded" into some numerical form. The selected features include data that is categorical and highly diverse, which makes traditional data encoding methods ineffective. Features that include information such as IP addresses may contain thousands of unique values, even in a relatively small set of

samples. Therefore, multiple methods were examined to encode this data in such a way that could be used as a feature for neural networks. One of these ways was the technique known as one-hot encoding. In one-hot encoding, each unique feature is separated into an individual feature column, which contains either a 1 or a 0. A 1 represents the presence of a feature in the sample, and a 0 otherwise. This is a fast and easy way of encoding data, but it increases the number of feature columns proportionally to the number of unique features present in a column, which results in large increases in memory usage and runtime. For these experiments, 80%/20% ratio was used for training and testing. The results of this model are shown in Table 2.

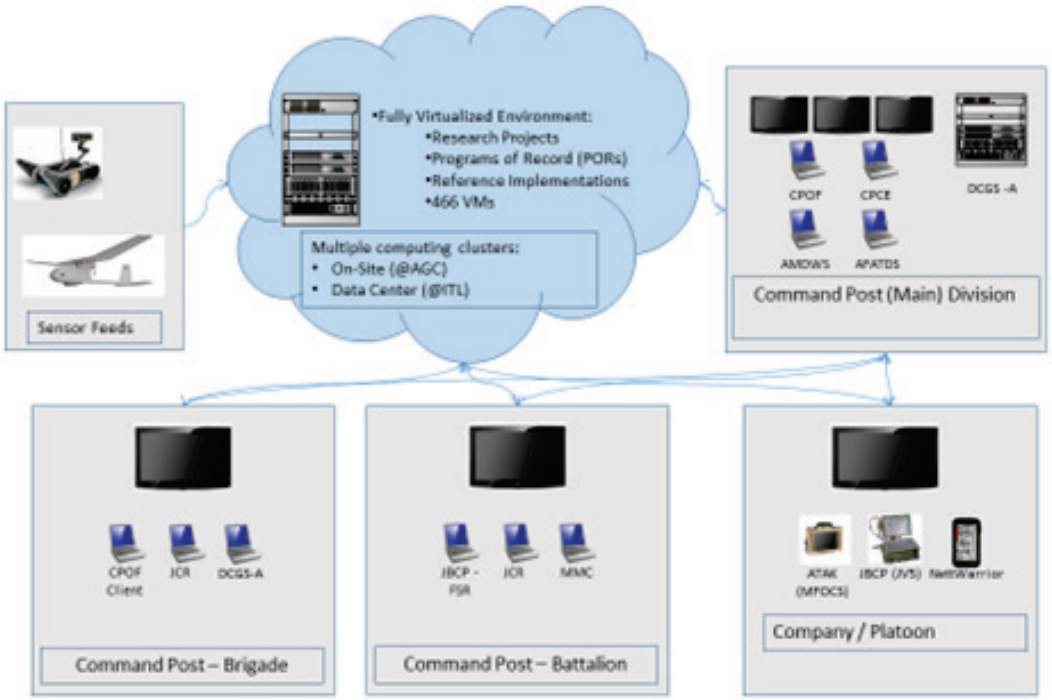
Use case 2: Near real-time operational plan management

AGE Node

The Army Geospatial Enterprise (AGE) Node consists of physical and virtual resources configured to simulate deployed Army computing environments at multiple echelons. Each computing environment contains software and systems used by Soldiers in the field. The AGE Node infrastructure can also replicate channel bandwidth limitations that friendly forces may face when operating communications links downrange. The Geospatial Research Lab (GRL) hosts the AGE Node's physical presence in a room stocked with multiple terminals and displays for exercises, demonstrations, and presentations.

The ERDC Information Technology Lab (ITL) provisions virtual computing resources used to run associated software of the deployed and research prototype variety. Figure 11 shows an example of the AGE Node physical layout.

Figure 11. AGE Node physical layout.



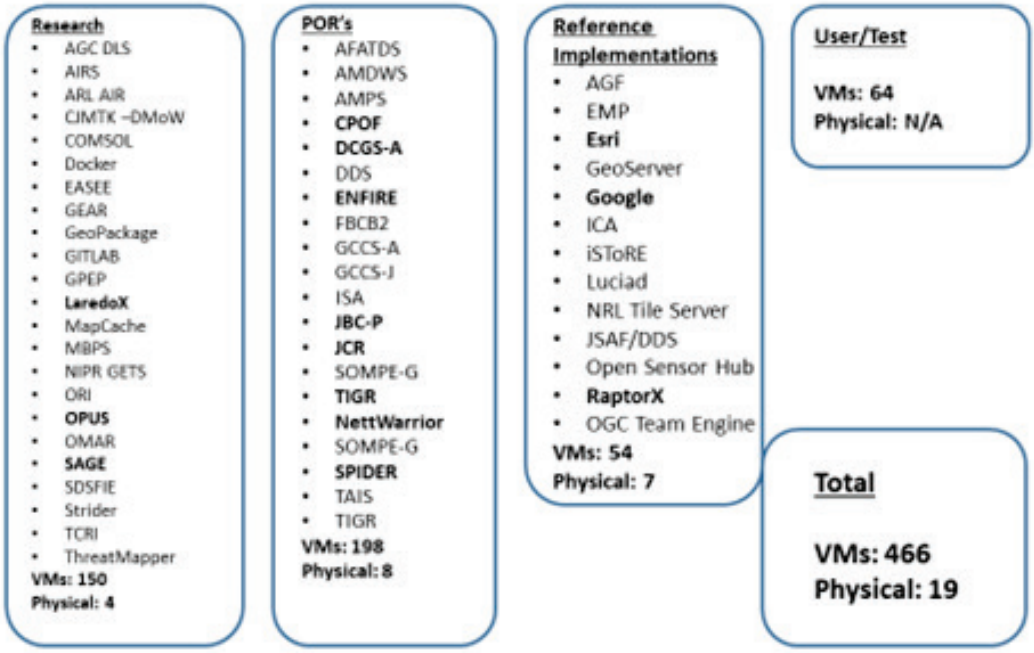
Within the context of this project, the AGE Node provides a safe environment to perform several key tasks:

- Understanding how current operational systems interoperate and are used by Soldiers.
- Experimenting with different machine learning / artificial intelligence (ML/AI) approaches within a realistic development environment.
- Demonstrating the utility or effectiveness of ML/AI algorithms by conducting limited user tests / exercises.

For the purposes of this report, efforts will be focused at the tactical level – those systems that are deployed at brigade echelons and below. A list of available software systems can be seen in Figure 12. Of particular note is the Map Based Mission Planning system (MBMP). This project has since transitioned to the Joint Planning System (JPS) and will serve as the U.S. Army's primary interface for mission planning once implemented. The work will focus on the use of JPS within a tactical formation, homing in on the enemy courses of action (COA) contained within the Operations Plan (OPLAN) or Order (OPORD). These COAs can be used as a sort of 'ground truth' to evaluate enemy behaviors attempting to identify which COAs they are taking. The main challenge here is translating the COA text into a set of unique, potential observables in order to confirm the selection of a

particular COA. This could involve the use of natural language processing / understanding to perform such a task.

Figure 12. AGE Node systems.



Since actual operational data will be hard to obtain, additional software may be deployed to the AGE Node such as One Semi Automated Forces (OneSAF) or Joint Semi Automated Forces (JSAF) in order to generate synthetic data. There is a current effort underway to develop an API for JPS to interoperate with a SAF in order to simulate 'war game' friendly courses of action. Our work can take this a step further by linking the Integrated Sensor Architecture (ISA) to JSAF and simulate a deployed sensor network that is tracking enemy and friendly movement across the area of operations. This synthetic data can then be ingested, analyzed with our algorithms, and movements classified in terms of the COA taken. Details on ISA are in the following section.

The Integrated Sensor Architecture (ISA) is a software-based framework for sensor integration and interoperability. It uses a publish-subscribe mechanism to achieve dynamic discovery of assets and capabilities as well as automated delivery of relevant messages to consumers. ISA relies on availability of certain hardware components when deployed, namely:

- Networking – connections to wired or wireless networks must be provided.

- Computing – processing capacity, system memory, and local storage must be enough to run ISA [1]in addition to other tasks for which the platform was provisioned.
- Authoritative time source – ISA software components rely on time synchronization in order to provide reliable delivery of messages in degraded communications environments.

Key to the operation of ISA is a distributed network of controllers that provide three services that the ISA developers have designated as critical to proper functionality:

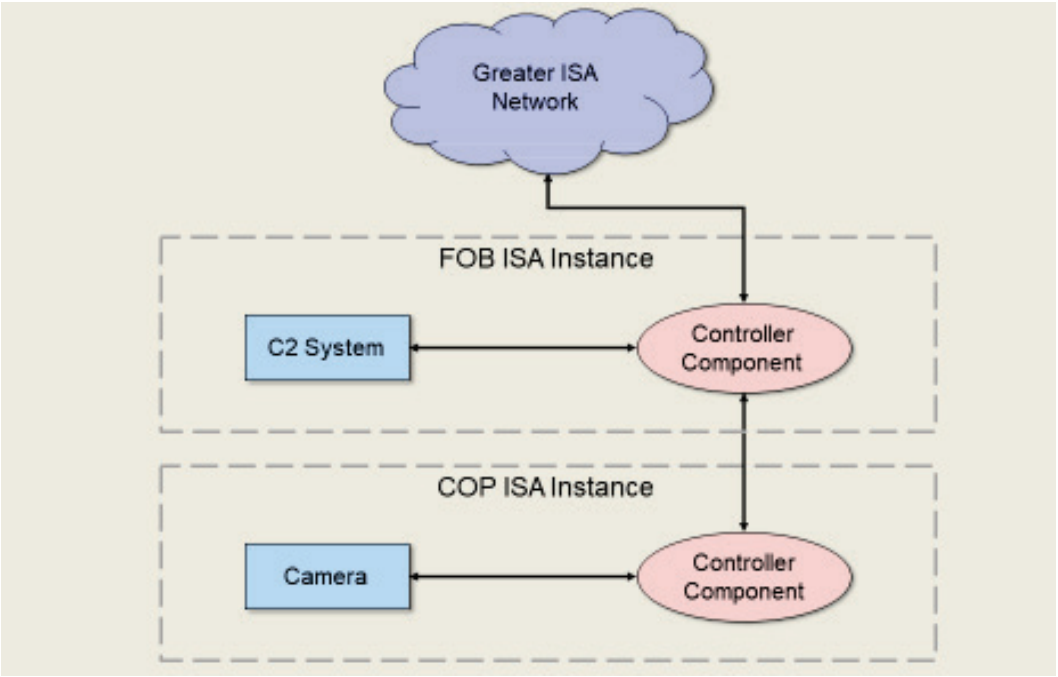
- Communication service – handles registration, authentication, availability monitoring and request handling for ISA components.
- Authorization service – mediates access between users and components based on designated policy.
- Live data service – provides information on status of ISA components as well as access to observable data that the component has recorded.

Additionally, having multiple controller instances deployed provides a degree of fault tolerance given the occasional ad hoc nature of tactical networks.

Within the context of this project, Figure 13 serves as an example of how ISA can be employed to experiment with ML/AI algorithms for mission command. Here, a controller instance is attached to the C2 system that is planned on instrumenting, namely JPS. Since there is no client component associated with JPS as of this writing, one would need to be created in order to interface with an ISA controller. The JPS ISA controller can then query another controller for simulated sensor capabilities attached to a simulation system such as OneSAF or JSAF.

[1] https://confluence.di2e.net/display/ISA/Downloads

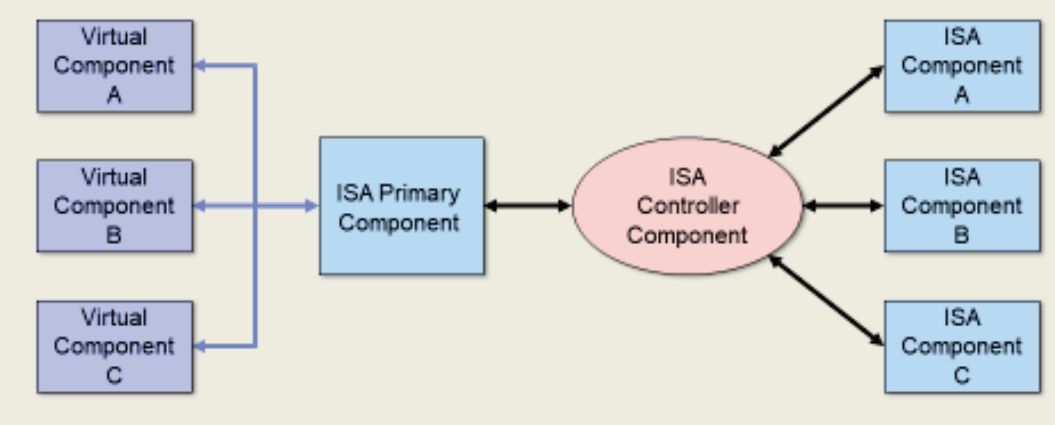
Figure 13. ISA composition.



Currently, it is unknown whether either simulation system can replicate a deployed sensor network. If so, it would need to be determined if the simulation already has an ISA client component embedded within the system. If the system does not feature ISA-compliant output, additional software would need to be developed that translates the simulated observables from the SAF data model to the ISA data model as well as implement a client component to connect to a controller. If, however, the simulation cannot replicate a sensor network at all, it may be a better option to create the observable data directly, outside of the simulation. A client component would still need to be developed to interface with a controller but would have more fine-grained control over the observables sent to JPS.

This paradigm is expanded to the Internet of Battlefield Things (IoBT) where synthetic data are being created from devices other than Army sensors, such as handheld devices or surveillance cameras. All of these would need client components, but probably would not differ much in terms of actual implementation. Figure 14 is an example of how the resulting architecture, consisting of a simulated Army sensor network meshed with IoBT, might be structured.

Figure 14. ISA virtual components.



Map-based planning system

The Command Post Computing Environment (CPCE), under the direction of Product Manager Tactical Mission Command (TMC), provides a software infrastructure framework (common interface, data and services) upon which current Warfighter capabilities can be converged and future capabilities can be built.[1] The portfolio contains:

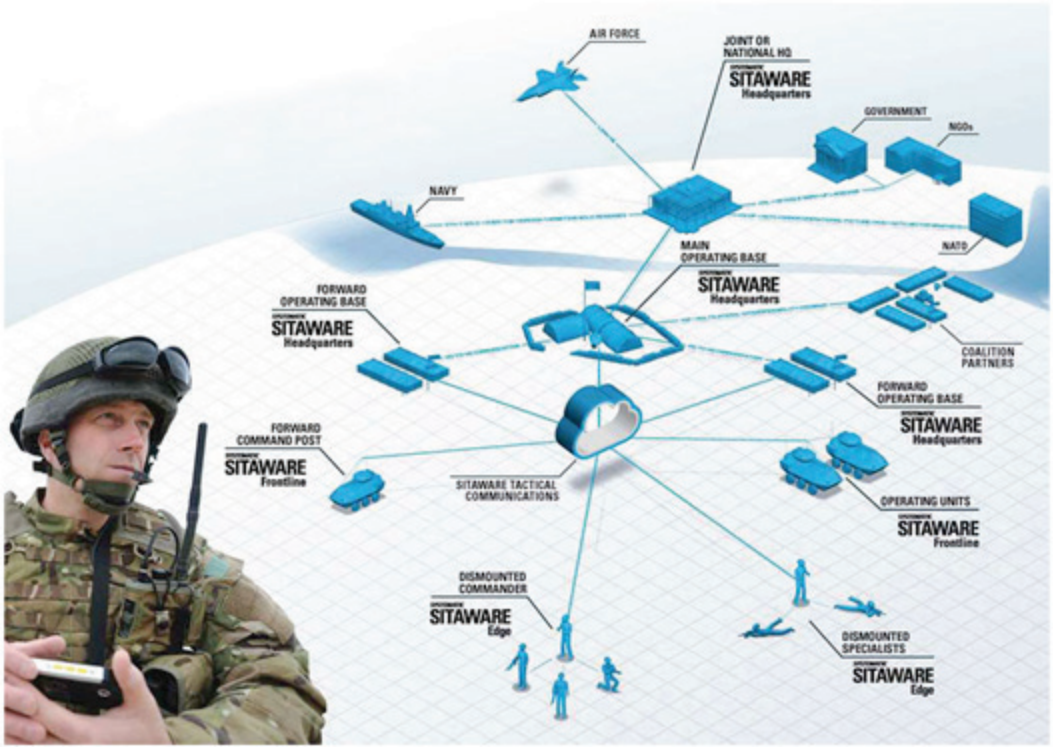
- Command Post Computing Environment
- Command Post Integrated Infrastructure
- Fire Support Command and Control
- Joint Battle
- Command-Platform
- Mounted Computing Environment
- Mission Command Cyber
- Strategic Mission Command

SitaWare by Systematic, Inc. is the base software of the CPCE on which MBMP is developed. Systematic, Inc. is a leading provider of simple and reliable C4I integration software solutions for the Department of Defense.

Their products and services (Figure 15) provide warfighters enhanced situational awareness at all levels of command, world-class interoperability with international allies and coalition forces, and greater security abroad.

[1] https://peoc3t.army.mil/mc/cpce.php

Figure 15. The SitaWare World of Interoperability.



SitaWare Edge is the new simple, lightweight, easy to use Android software designed for the dismounted domain.[1] SitaWare Edge is part of the SitaWare Suite, which is compliant with the latest standards, making it fully interoperable with other systems, coalition partners, and NATO member countries. When deployed in an operational theater, SitaWare Edge can be used with SitaWare Frontline and SitaWare Headquarters to provide a unified C2 system from the tactical edge to the highest levels of command.

Within the context of this project, MBMP serves as the planning review mechanism on which models and simulations can be evaluated for rapid decision support impacting current and future plans. OneSAF or JSAF can simulate the status of the plan based on data feeds coming from Edge devices in the field to evaluate and recommend plan updates as needed.

OneSAF modeling and simulation environment

OneSAF is the Army's next generation entity-level simulation that provides a composable, distributed, and scalable simulation of real-world

[1] https://www.systematicinc.com/products/n/SitaWare/edge/

battlefield situations using validated physical models and doctrinally correct behavior models. It can support analysis, acquisition, planning, testing, training, and experimentation. OneSAF allows users to compose a wide range of complete simulation systems from a set of component-based tools, develop new or extend existing tools, as well as compose new single or multi-resolution entities, units, and associated behaviors from existing physical and behavioral software components.

OneSAF also accurately and effectively represents activities within the Army warfighting functions to include:

- Intelligence
- Movement and maneuver
- Fire support
- Protection
- Sustainment
- Command and control

OneSAF hardware/system performance requirements are impacted numerous distributed simulation-related factors. For this paper, two dimensions of scalability were explored: fidelity and entity count. Increases in either dimension require additional hardware and communications resources to properly support the forces representation in the OneSAF model.

OneSAF supports variable levels of fidelity, which makes it possible to tailor the simulation in order to maximize satisfaction of diverse use cases. Here, fidelity refers to the faithfulness of the model to the real-world object being modeled. Often fidelity is equated to model detail and computational requirements, as they are typically, but not necessarily, related. For most entities, units, behaviors, and even terrain, OneSAF supports three levels of fidelity: low, medium, and high.

For example, low fidelity dismounted infantries (DIs) move, sense, and shoot, but they only move in a simple pattern along routes designated by the simulation operator and will not avoid obstacles. Medium-fidelity OneSAF entities will also move, sense, and shoot, but they have more sophisticated mobility, sensor, and weapon models than their low-fidelity counterparts. So, for example, medium-resolution DIs will exhibit more realistic movement behaviors along routes and will avoid obstacles. High-

fidelity entities have all capabilities and attributes of medium-fidelity entities; however, they are enhanced in some way. They might have a high-fidelity missile fly-out model, rather than Army Materiel Systems Analysis Activity (AMSAA) provided probability of hit/probability of kill (Ph/Pk) values. They might have sensor models that have greater detail, or they might have a higher-fidelity mobility model. OneSAF's composable architecture provides for complete fluidity among the various physical and behavioral models that can be instanced as a part of an entity. For example, the medium-fidelity mobility model can be freely combined with the low-fidelity communication model and a high-fidelity acquisition model to create an entirely new, customized entity. There are no rigid rules that determine the fidelity of any given entity; it is up to the creator to assess and properly define them. For this reason, it is impossible to easily estimate global performance characteristics of entities based on their stated fidelity alone.

Entity count is tied more specifically to the size of the force being represented. However, entities may be used to represent more than just dismounted infantry or vehicles. When used to also represent communication devices (such as radios or cell phones) and sensors, the entity count dimension can quickly grow with increased fidelity in battlefield environment representation.

Different applications require different fidelity / entity count requirements. Staff training, COA development and high-level COA analysis can be supported using low-fidelity, high-entity-count simulations. Medium-fidelity simulations are more suitable to support battlefield functional area (BFA)-specific training, detailed COA analysis, mission rehearsal, mission execution monitoring, and some types of experimentation. They can also support concept definition and tradeoff analyses. High-fidelity, low-entity-count simulations, on the other hand, work well to support detailed analyses appropriate for research, system tradeoff analyses, etc.

OneSAF provides the option to support a mixture of entity fidelities in the same exercise. The user can select medium- or high-fidelity entities in areas where detailed modeling of battlefield conditions is needed but use low-fidelity entities elsewhere as "wrap-around" to fill in the rest of the battlespace. This allows the computational resources to be focused on modeling items of interest in the simulation scenario. This composable

approach allows users to tailor the simulation to meet their specific domain requirements.

HPC offers the OneSAF M&S community a chance to increase the simulation representation along both dimensions of scalability, where workstation or network resources currently limit the simulation ability to scale and force the simulation architect to limit one dimension or the other.

As the use of OneSAF on HPC grows, changes are being rolled into the baseline, allowing OneSAF's computational capabilities to grow as computational hardware advances toward more parallel, cluster-type systems. It also provides DoD engineers and researchers a valuable tool for scaling the simulation environment to meet a more demanding, higher fidelity environment.

Integrating MBMP with OneSAF

Cole Engineering demonstrated OneSAF integration with SitaWare.[1] This proof-of-concept demonstrated two capabilities. First, OneSAF could run on a laptop for tactical simulations. Second, the functionality was embedded into Sitware, the base software for the Command Post Computing Environment (CPCE).

Although the simulation still requires longer processing times, incorporating advances in high performance and graphics computing could reduce simulation run times. Incorporating simulations is important to help leaders understand where gaps or risks exist in executing a given plan. Cole Engineering also demonstrated how a third-party developer could apply the SitaWare Software Development Kit to create additional capabilities for future CPCE versions.

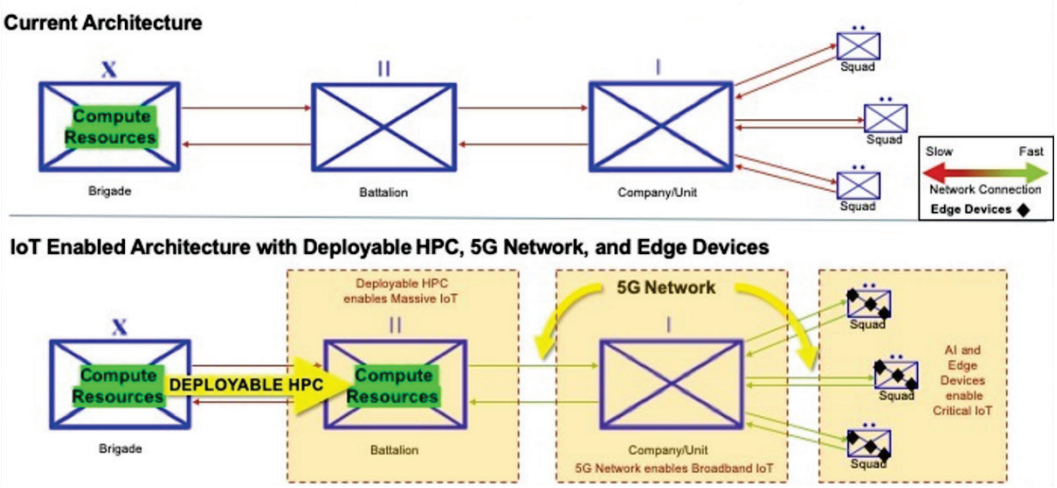
[1] https://www.army.mil/article/200727/mission_command_battlelab_hosts_technology_demo

4 HPC Tactical Data Analytics Framework Design

Data analytics framework for tactical data on portable HPC

Proposed upgrades to the technologies that are deployed at various levels of military units and how they enable Artificial Intelligence, Edge Computing, and the Internet of Things (Figure 16). Each of these IoT approaches have a variety of potential AI implementations that provide value to the warfighter that is not currently available. The current infrastructure does not have compute resources at the proper level to provide Massive IoT applications, nor does it have a network efficient enough for most Broadband IoT applications. Moving large-scale compute resources closer to the edge would quicken Massive IoT decision-making and enhance the impact of decisions made on the existing data.

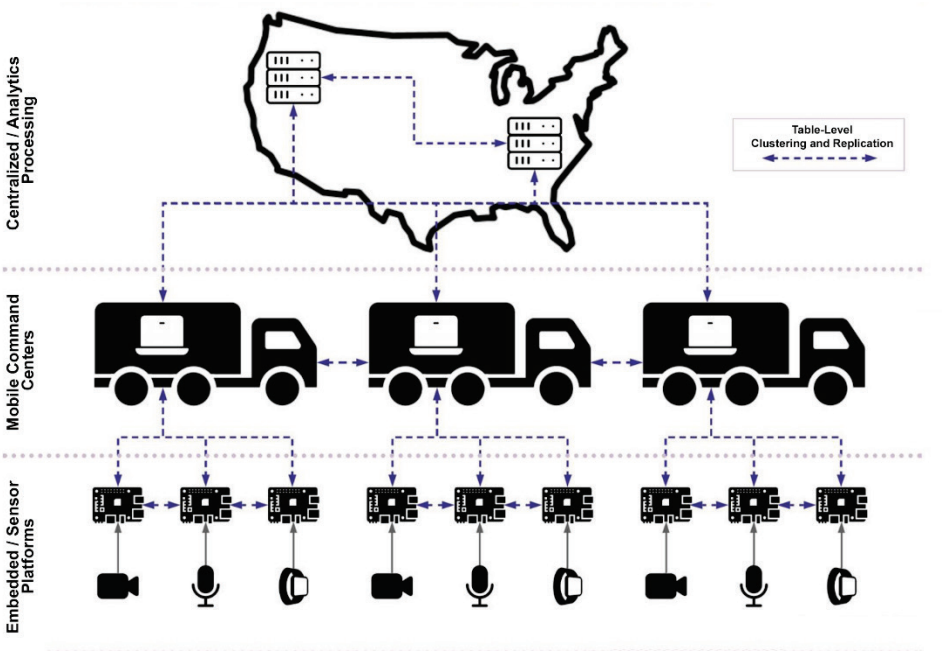
Figure 16. Technological upgrades that enable AI on Edge and IoT.



Proposed Army approach to IoT and AI/ML

The proposed architectural shift is shown in Figure 17. Enhancements and investments at all three phases (network, deployed edge devices, and large edge computing machines) of the proposed architecture provide value both for standard operations and in enabling data-driven responses by decision makers.

Figure 17. Proposed Architectural Shift.



5 Summary

In summary, IoT solutions have been investigated based on the understanding that collecting, analyzing, and rapidly converting large data sets into actionable intelligence is key for the DoD. One of the important means to technological superiority on the battlefield is the exploitation of information. Recent advances in AI and supercomputing provide the DoD the opportunity to push AI-based capabilities to the tactical edges of the network. AI (machine learning and deep learning technologies) provides an opportunity to gain new insights from vast quantities of data. Supercomputers are needed on the edge to host the AI. Future experiments will be needed to better understand the level of supercomputing resources necessary. Current identified commercial technology will continue to drive advanced capabilities in AI with future developments.

The investigation phase of this project has been completed and will continue with the acquisition and implementation of hardware and architecture for environments that extend to the tactical edge. Use cases have been developed and will define proof-of-concept creations to demonstrate the described software, hardware, and communication infrastructure.

6 Future Work

The Accelerating the Tactical Decision Process with HPC on the Edge project is a three-year project from FY19 through FY21. In FY19, the Commercial Off the Shelf (COTS) hardware and software solutions to IoT problems were identified. Industry standard infrastructures were also identified that will not suffice for military uses and therefore more research is required in this area to identify satisfactory solutions.

For the next stages, the team will begin to analyze ML and IoT tools to optimize computation and execution time on tactical edge computing. The end goal is to define an operational workflow design, testing and evaluation of the tactical decision process. A proof-of-concept demo will be created and then expanded upon to show a full-scale demonstration of capabilities on an HPC environment.

References

Davis, Z. S. 2019. Artificial Intelligence on the Battlefield: An Initial Survey of Potential Implications for Deterrence, Stability, and Strategic Surprise. https://cgsr.llnl.gov/content/assets/docs/CGSR-AI_BattlefieldWEB.pdf

Garamone, J. 2017. "*Dunford: Speed of Military Decision-Making Must Exceed Speed of War > U.S. DEPARTMENT OF DEFENSE > Defense Department News*." https://www.defense.gov/Explore/News/Article/Article/1066045/dunford-speed-of-military-decision-making-must-exceed-speed-of-war/.

Gigova, R. 2017. "*Who Putin Thinks Will Rule the World - CNN*." https://www.cnn.com/2017/09/01/world/putin-artificial-intelligence-will-rule-world/index.html.

Hansche, S. 2019. "Committee on National Security Systems." In *Official (ISC)2® Guide to the CISSP®-ISSEP® CBK®*, 597–633. http://www.cnss.gov.

Liu, F., Tang, Y. Li, Z. Cai, X. Zhang, and T. Zhou. 2019. "A Survey on Edge Computing Systems and Tools." *IEEE* 107(8): 1537-1562, doi: 10.1109/JPROC.2019.2920341.

Lorenzen, C., R. Agrawal and J. King. 2018. "Determining Viability of Deep Learning on Cybersecurity Log Analytics." *IEEE International Conference on Big Data* (Big Data) 4806-4811, doi: 10.1109/BigData.2018.8622165.

Shi, W., G. Pallis, and Z. Xu. 2019. "Edge Computing." *Proceedings of the IEEE* 107(8): 1474–81.

Shilawat, S. 2018. "Council Post: Step Aside, Cloud -- Here Comes Edge Computing." *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2018/10/30/step-aside-cloud-here-comes-edge-computing/#467837b333c1.

Staff, Joint. 2005. "Net-Centric Operational Environment Joint Staff." October (October), https://dodcio.defense.gov/Portals/0/Documents/netcentric_jic.pdf

.

Acronyms

Term	Meaning
ERDC	U.S. Army Engineer Research and Development Center
ITL	Information Technology Laboratory
GRL	Geospatial Research Laboratory
USACE	U.S. Army Corps of Engineers
IoT	Internet of Things
AI	Artificial Intelligence
HPC	High Performance Computing
ML	Machine Learning
HPCMOD	High Performance Computing Modernization
DoD	Department of Defense
MBMP	Map Based Planning System
OneSAF	One Semi-Automated Forces
IoBT	Internet of Battlefield Things

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
September 2021	Final	

4. TITLE AND SUBTITLE

Accelerating the Tactical Decision Process with High-Performance Computing (HPC) on the Edge: Motivation, Framework, and Use Cases

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Alicia I. Ruvinsky, Timothy W. Garton, Daniel P. Chausse, Rajeev K. Agrawal, Harland F. Yu, and Ernest L. Miller

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Information Technology Laboratory
U.S. Army Engineer Research and Development Center
3909 Halls Ferry Road, Vicksburg, MS 39180-6199;
Geospatial Research Laboratory
U.S. Army Engineer Research and Development Center
7701 Telegraph Road, Alexandria, VA 22315-3864

8. PERFORMING ORGANIZATION REPORT NUMBER

ERDC TR-21-19

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Corps of Engineers
Washington, DC 20314-1000

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

ERDC FLEX-4, Funding account code U4371831

14. ABSTRACT

Managing the ever-growing volume and velocity of data across the battlefield is a critical problem for warfighters. Solving this problem will require a fundamental change in how battlefield analyses are performed. A new approach to making decisions on the battlefield will eliminate data transport delays by moving the analytical capabilities closer to data sources. Decision cycles depend on the speed at which data can be captured and converted to actionable information for decision making. Real-time situational awareness is achieved by locating computational assets at the tactical edge.

Accelerating the tactical decision process leverages capabilities in three technology areas: (1) High-Performance Computing (HPC), (2) Machine Learning (ML), and (3) Internet of Things (IoT). Exploiting these areas can reduce network traffic and shorten the time required to transform data into actionable information. Faster decision cycles may revolutionize battlefield operations.

Presented is an overview of an artificial intelligence (AI) system design for near-real-time analytics in a tactical operational environment executing on co-located, mobile HPC hardware. The report contains the following sections, (1) an introduction describing motivation, background, and state of technology, (2) descriptions of tactical decision process leveraging HPC problem definition and use case, and (3) HPC tactical data analytics framework design enabling data to decisions.

15. SUBJECT TERMS

Battles – Decision making
Network-centric operations (Military science)
Electronic data processing
High performance computing
Machine learning
Internet of things
Artificial intelligence

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	SAR	43	