Be a part IEEE CAI—an exciting two-day event covering the applications of AI and key AI verticals that impact industrial technology and innovations.

At IEEE CAI you'll gain insights from over 30 leading artificial intelligence innovators on the following topics:

- AI in Healthcare/ Life Sciences
- AI in Transportation/ Aerospace
- AI in Energy
- Industrial AI
- AI for Earth Systems Decision Support
- Ethical & Societal Implications of AI

https://cai.ieee.org/2023
8
Intent-Aware Interactive Internet of Things for Enhanced Collaborative Ambient Intelligence

16
Democratizing Data-Driven Agriculture Using Affordable Hardware

36
Nvidia Hopper GPU and Grace CPU Highlights
Computer Vision

8 Intent-Aware Interactive Internet of Things for Enhanced Collaborative Ambient Intelligence
CHAORAN HUANG, LINA YAO, XIANZHI WANG, QUAN Z. SHENG, SCHAHRAM DUSTDAR, ZHONGJIE WANG, AND XIAOFEI XU

16 Democratizing Data-Driven Agriculture Using Affordable Hardware
RANVEER CHANDRA, MANOHAR SWAMINATHAN, TUSHER CHAKRABORTY, JIAN DING, ZERINA KAPETANOVIĆ, PEEYUSH KUMAR, AND DEEPAK VASISHT

Artificial Intelligence

26 The AI Chip Race
GUANSONG PANG

28 Taking a Measured Approach to Investing in Information Infrastructure for Attaining Leading-Edge Trustworthy Artificial Intelligence
JAMES BRET MICHAEL

Parallel Processing

32 PyOMP: Multithreaded Parallel Programming in Python
TIMOTHY G. MATTSON, TODD A. ANDERSON, AND GIORGIS GEORGAKOUDIS

36 Nvidia Hopper GPU and Grace CPU Highlights
ANNE C. ELSTER AND TOR A. HAUGDAHL

Blockchain

42 The Evolution of Nonfungible Tokens: Complexity and Novelty of NFT Use-Cases
ANDREW PARK, JAN KIETZMANN, LEYLAND PITT, AND AMIR DABIRIAN

48 Elliptic Curve Pairings
JOSHUA BRIAN FITZGERALD

Departments

4 Magazine Roundup
7 Editor’s Note: From Computer Vision to Robot Helpers
53 Conference Calendar

Subscribe to ComputingEdge for free at www.computer.org/computingedge.
The IEEE Computer Society’s lineup of 12 peer-reviewed technical magazines covers cutting-edge topics ranging from software design and computer graphics to Internet computing and security, from scientific applications and machine intelligence to visualization and microchip design. Here are highlights from recent issues.

**Computer**

**Redefining Circular Cities: Regulation, Governance, Infrastructure, and Technology**

This article from the December 2022 issue of *Computer* explores major global urban challenges driven by climate change, outlines the European Union policy framework, and discusses new human-centric urban trends and technologies shaping circular urban planning for city leaders, technology experts, and academics.

**Computing**

**The Influence and Contribution of Jack Dongarra to Numerical Linear Algebra**

This article from the July/August 2022 issue of *Computing in Science & Engineering* is dedicated to Jack Dongarra on the occasion of him receiving the 2021 ACM Turing Award. The article concentrates primarily on his contributions to numerical linear algebra, particularly on the development of algorithms and software to reliably and efficiently solve linear algebra problems. Quoting from one of Jack’s own recent talks, his tenets have been “accuracy, community, innovation, performance, portability, productivity, readability, and reliability.”

**Annals**

**Computing Counterinsurgency: The Hamlet Evaluation System (HES) and Databasing During the Vietnam War**

In 1967, the US military in South Vietnam launched the Hamlet Evaluation System (HES), a new type of data-processing system comprising a demographic database and monthly surveys for monitoring the progress of its counterinsurgency effort. The HES had a considerable impact on strategic and political decisions. A close examination of its conception and practices of data collection, processing, and analysis reveals that the system could not adequately capture the dynamics of war, especially refugee movements. Although several revisions steadily refined the system, policymakers and the military only embraced highly aggregated results. Read more in this article from the October–December 2022 issue of *IEEE Annals of the History of Computing*.

**IEEE Computer Graphics and Applications**

**SUBPLEX: A Visual Analytics Approach to Understand Local Model Explanations at the Subpopulation Level**

Understanding the interpretation of machine learning (ML) models has been of paramount importance when making decisions with societal impacts, such as transport control, financial activities, and medical diagnosis. While local explanation techniques are popular methods to interpret ML models on a single instance, they do not scale to the understanding of a model’s behavior on the whole dataset. In this article from the November/December 2022 issue of *IEEE Computer Graphics and Applications*, the authors outline the challenges and needs of visually analyzing local explanations and propose SUBPLEX, a visual analytics approach to help users...
understand local explanations with subpopulation visual analysis.

**Intelligent Systems**

**Knowledge-Based Entity Prediction for Improved Machine Perception in Autonomous Systems**

Knowledge-based entity prediction (KEP) is a novel task that aims to improve machine perception in autonomous systems. KEP leverages relational knowledge from heterogeneous sources in predicting potentially unrecognized entities. In this *IEEE Intelligent Systems* September/October 2022 article, the authors provide a formal definition of KEP as a knowledge completion task. Three potential solutions are then introduced, which employ several machine learning and data mining techniques. Finally, the applicability of KEP is demonstrated on two autonomous systems from different domains: autonomous driving and smart manufacturing.

**Internet Computing**

**Toward Decentralized Cloud Storage With IPFS: Opportunities, Challenges, and Future Considerations**

The Interplanetary File System (IPFS) is a novel decentralized storage architecture that provides decentralized cloud storage by building on founding principles of P2P networking and content addressing. IPFS is used by more than 230,000 peers per week and serves tens of millions of requests per day, which makes it an interesting large-scale operational network to study. While it is used as a building block in several projects and studies, its inner workings, properties, and implications have only been marginally explored in research. Thus, the authors of this article from *IEEE Internet Computing*’s November/December 2022 issue provide an overview of the IPFS design and its core features, along with the opportunities that it opens and the challenges that it faces because of its properties.

**MultiMedia**

**Transferring Deep Gaussian Denoiser for Compressed Sensing MRI Reconstruction**

Deep neural networks have achieved the most outstanding performance in compressed sensing magnetic resonance imaging (CS-MRI) reconstruction by learning the potential structures of images from training samples. However, the required data comprising hundreds of subjects are usually rare. In this article from *IEEE MultiMedia*’s October–December 2022 issue, the authors transfer the easy-to-get deep Gaussian denoisers trained with natural images for artifact reduction in the iterative recovery process without the use of full-sampled MRI data. They first train a set of deep Gaussian networks with natural images and then incorporate them into a play-and-plug
framework that is built by modifying the proximal gradient algorithm with the classical momentum strategy.

**Personalized Gestures Through Motion Transfer: Protecting Privacy in Pervasive Surveillance**

With the growing ubiquity of pervasive sensing and toward ambient intelligence, pervasive surveillance becomes a very real privacy threat, where private gesture interaction is likely to be observed and automatically interpreted by other (even benign) pervasive intelligence tools. The authors of this article from *IEEE Pervasive Computing*’s October–December 2022 issue propose motion transfer, the example-guided modification of motion to translate from default motion and gesture interaction alphabets to personal ones. Apart from privacy, incentives to use personalized gesture interaction alphabets include convenience and physical handicaps (i.e., inability to conduct certain movements). The authors demonstrate the concept using motion transfer in RGB-video data.

**Distributed Self-Sovereign-Based Access Control System**

Distributed ledger technology has paved the way for new innovative approaches in the field of security and privacy. Self-sovereign identity contributed to empowering individuals to have control over their digital identifiers. This *IEEE Security & Privacy* article from the November/December 2022 issue presents a self-sovereign-based access control system that utilizes both technologies.

**Infrastructure From Code: The Next Generation of Cloud Lifecycle Automation**

The authors of this article from the January/February 2023 issue of *IEEE Software* identify 14 fundamental cloud infrastructure procedures (CIPs) applicable to software development processes on the public cloud and their associated challenges. They then evaluate the capabilities of leading cloud automation technologies, such as infrastructure as code, and pinpoint their gaps in enabling the CIPs.

**VCFN: Virtual Cloth Fitting Try-On Network**

Virtual cloth fitting network has an increasing demand with a growing online shopping trend to map target clothes on reference subject. Previous research depicts limitations in the generation of promising deformed clothes on the wearer’s body while retaining features like logo, text, and wrinkles. The authors of this November/December 2022 *IT Professional* article propose a model that learns thin-plate spline transformations to warp images according to body shape, followed by a try-on module. The former model combines deformed cloth with a rendered image to generate a composition mask and outputs target body without blurry clothes while preserving critical requirements of the wearer. Experiments are performed on the Zalando dataset, and the model produces richer details and promised generalized results.
Editor's Note

From Computer Vision to Robot Helpers

When computer vision technology is implemented in robots, new applications in our homes and businesses become possible. These robots can analyze visual data in real time and take appropriate actions based on their conclusions. This ComputingEdge issue gives two examples of computer vision and robotics coming together in innovative ways.

The authors of “Intent-Aware Interactive Internet of Things for Enhanced Collaborative Ambient Intelligence,” from IEEE Internet Computing, propose a smart home system in which robots equipped with 3D depth cameras identify scenarios and assist humans with household tasks like making tea. IEEE Micro’s “Democratizing Data-Driven Agriculture Using Affordable Hardware” describes how aerial imagery from autopilot drones can be integrated with sensor data to help farmers predict crop outcomes and conserve resources.

Artificial intelligence (AI) is another field making an impact in everyday life, and the research is progressing rapidly. “The AI Chip Race,” from IEEE Intelligent Systems, details efforts to develop cheaper and faster AI accelerators. The author of IEEE Security & Privacy’s “Taking a Measured Approach to Investing in Information Infrastructure for Attaining Leading-Edge Trustworthy Artificial Intelligence” considers technical, policy, and legal concerns related to AI infrastructure.


Finally, this ComputingEdge issue covers blockchain trends. The authors of “The Evolution of Nonfungible Tokens: Complexity and Novelty of NFT Use-Cases,” from IT Professional, focus on the history of this blockchain-based technology. Computer’s “Elliptic Curve Pairings” describes a way to construct zero-knowledge proofs for privacy in the blockchain transaction ledger.
Intent-Aware Interactive Internet of Things for Enhanced Collaborative Ambient Intelligence

Chaoran Huang and Lina Yao, University of New South Wales, Sydney, NSW, 2052, Australia
Xianzhi Wang, University of Technology Sydney, Ultimo, NSW, 2007, Australia
Quan Z. Sheng, Macquarie University, Macquarie Park, NSW, 2109, Australia
Schahram Dustdar, TU Wien, 1040, Vienna, Austria
Zhongjie Wang and Xiaofei Xu, Harbin Institute of Technology, Harbin, 150001, China

The Internet of Things (IoT) enables the connection of a broad range of artifacts with advanced sensory technologies and produces massive amounts of data to support ambient intelligence. While the potential of IoT systems is widely recognized, little work has demonstrated a system with the ability to execute autonomously in the real world. Inspired by the success of robotics in specialized IoT environments, we propose an end-to-end solution for a generic, interactive ambient intelligence system where robots can assist humans in conducting activities in IoT-enabled smart homes. We evaluate the solution using implementations of public benchmarks on open-source platforms and use several activities to demonstrate the effectiveness of the proposed solution in real life.

The Internet of Things (IoT) promises to integrate digital and physical worlds by connecting artifacts and building networks of them. Traditionally, IoT devices work on their own, due to incompatible techniques or protocols. The Web of Things (WoT) provides IoT applications with standardized descriptions of actions, events, and properties of things to extend Web protocols, thereby supporting broader interactions between smart objects. With plenty of commercially available IoT products, recent research and industry have successfully instrumented connected things in our daily lives. However, these connected things still cannot work fully autonomously but require human intervention. The existing research primarily focuses on recommending the next actions to take, yet cannot control devices directly to execute those actions. Inspired by multiagent systems, Internet of Robotic Things (IoRT) enables robots to gain awareness of the environment and adapt their actions. However, the limited related research focuses on specialized scenarios, such as workplaces and smart cities; it still faces challenges for conducting physical interactions for daily tasks. There remains a significant gap to bring robotics into home applications to support streamlined physical human–machine interactions.

We envision a scenario (Figure 1) to demonstrate our goal toward ambient intelligence and to motivate this study. Bob is preparing a bowl of cereal as breakfast, and usually, he finishes the meal with a cup of tea. The system detects a series of events in the kitchen at the time, such as movements of a cabinet door, a bowl, the fridge door, cereal containers and such, thereby recognizing that Bob is preparing cereal as breakfast. Based on Bob’s daily routine, the system envisages a
the Web of Things (WoT)\(^\text{a,b}\) ally, IoT devices work on their own, due to incompatible descriptions. However, these commercially available IoT products, recent research extend Web protocols, thereby supporting broader extensions of actions, events, and properties of things to provide IoT applications with standardized descriptions. \(^\text{b}[\text{Online}].\) Available: https://www.w3.org/TR/wot-architecture/
\(^\text{a}[\text{Online}].\) Available: https://www.w3.org/TR/wot-thing/

In this work, we propose a novel framework named Intent-aware Interactive Internet of Things for implementing ambient intelligence to facilitate seamless collaborations between humans, smart objects, and robots based on a unified IoT platform.\(^\text{2}\) Our framework consists of four tasks: 1) fusing sensory data from robots and smart homes to infer human intent based on past and current human behaviors; 2) combining the inferred intent with observations of the robot (via a 3D camera) to recognize semantic context and prompt physical interactions; 3) defining objectives for the robot; and 4) commanding the robot to take actions toward these objectives. Our main contributions are summarized as follows:

1) We propose an end-to-end solution for interactive ambient intelligence, where robots (also called robotic assistants) assist humans in conducting physical activities in IoT-enabled smart homes.
2) We engineer generic Web-based descriptions to facilitate interactions of smart-home devices. They enable the system to not only decide but also execute actions in the physical world.
3) We train reinforced learning-based robotic assistants to exploit successful and failure experience. We also apply a policy continuation strategy and a Hindsight Experience Replay method to expedite the learning process.
4) We implement the system with the Deepbot framework in the OpenAI Gym\(^\text{c,19}\) environment to demonstrate some typical activities in real life.

**RELATED WORK**

Smart-home-related studies have attracted enormous attention, thanks to the availability of networked smart devices and improved computing capability. Early studies in the field concern Web management, localization, tracking, or activity recognition. CASAS\(^\text{3}\) is one of the earliest smart-home experimental platforms that tracks users with preinstalled devices in apartments. Ruan \textit{et al}.\(^\text{4}\) achieve device-free indoor human localization and tracking. Yao \textit{et al}.\(^\text{3,6}\) demonstrated a unified management system that integrates monitors physical devices and finds relevant things according to past human interactions. Shemshadi \textit{et al}.\(^\text{7}\) provide a framework for diversified relevance search in IoT, laying the foundation for IoT search engines. Their follow-up study\(^\text{2,21}\) reveals a real-time multilevel activity monitoring system for a personalized smart home, which

\(^\text{c}[\text{Online}].\) Available: https://gym.openai.com/
continuously tracks daily activities and conducts abnormal activity detection. These studies realize ambient intelligence in certain aspects or applications but are still far from being truly "smart" in the sense of directly assisting human activities. Currently, robots are usually dedicated to specialized tasks but they offer great flexibilities in terms of "taking actions" in a smart environment. The concept of Internet of Robotic Things (IoRT)\(^1\) aims to integrate robotics with IoT networks. Mahieu et al.\(^1\) investigated context-aware and personalized interactions on the IoRT. Vermesan et al.\(^8\) review IoRT and clarify some concepts with suggested architecture and applications. All these efforts form the base of our work.

**METHOD**

Our framework (Figure 2) explores ambient intelligence to enable autonomous robots to collaborate actively with humans in their daily activities in smart homes. The framework can infer human intent, compose actuation for interactions, and finally execute actions via robots.

**Problem Setup**

We first model human activities and develop a segmentation method for recognizing complex activities with awareness of concurring events and the subject intentions. Then, we finalize conscientious moves and

---

**Figure 2.** Workflow of the proposed system in our scenario (illustrated in Figure 1). First, the robot observes human activity and infers human’s current behavior (i.e., making cereal breakfast) and intent (i.e., having a cup of tea) based on IoT sensory data. Then, the system figures out the objects to interact and works out the procedure for making tea. Finally, the robotic arm is guided to execute the procedure and make the predicted intent a reality (by preparing a cup of tea).
devise instructions for the robot assistants. We use the moving robot with a 3D stereo camera as the visual guide for other cooperative robot assistants to collaborate with the human. The 3D camera enables the robot to scan the environment, monitor the subject’s activities, and infer the next move of the subject based on the usage history and basic rules preset or learned beforehand. After identifying interactable objects, the robot will receive step-by-step instructions by a smart and adaptive robotic control system.

Let $A \in \mathbb{R}^N$ be a set of activities, where $a_i \in A$ is the $i$th class of activity in the set. At time $t$, the system receives sensory input $x^A_t \in \mathbb{R}^d$, along with the predicted activity for this input:

$$x^A_t \rightarrow \tilde{y}^A_t = a_i.$$ 

Suppose $Q^A_t \in \mathbb{R}^d$ is the semantic description corresponding to the current input $x^A_t$. The system matches this description with the semantic keywords $Q^O \in \mathbb{R}^p$ of the objects detected by the robot $R_m$. With the contexts and historical usage information $(A_H)$, the system infers the future action of the human at time $t' + 1$, $y^A_{t+1} = \tilde{a}_j \in A$; accordingly, the action composer $\alpha(\cdot)$ produces executable procedure with a series of interactions $I \in R^M$ for the target $Q^A$ with objects to $Q^O$ assumed that in a short time duration $Q^O_{t'} + 1 = Q^O_{t'} = Q^O$, as illustrated below:

$$\{Q^A_t, Q^O, A_H\} \rightarrow \tilde{a}_j$$

Once receiving the interactions $I$, the robot sees each interaction $\iota_m$ as a goal and learns to reach these goals. We abstract all the connected and interactive things in a smart home environment as Web resources to support dynamic discovery, and perform hypermedia interactions as standardized Web interactions via normal Web protocols, such as HTTP. The device interactions are realized via Web services following WS-* standards, where JSON-based serialization help ease the implementation. Besides, we adapt a semantic description of Web-based Artifacts (based on Ricci et al.\textsuperscript{10}) as the first-class abstraction for a clear integration and deployment of multiagent system (MAS).

Intent-Awareness With Connected Things

We aim to predict the next human activity and generate the corresponding semantics. We adopt a workflow inspired by Triboan et al.\textsuperscript{11} a semantic theory-based approach for sensor event segmentation, to facilitate a semantic activity prediction.

Activity Modeling

The environmental context $(EC)$ consists of human subject $(H_n)$, location $(L_m)$, ambient characteristic $(AC_o)$, sensor characteristics $(S_p)$, and interactable objects $(Obj_q)$ of classes $(C)$:

$$EC = \{H_n, L_m, AC_o, S_p, Obj_q\}.$$ 

Sensor Environment $(SR)$ describes the semantic relationship $(R_c)$ between sensor events and objects where

$$SR = a_n (R_c, EC) \rightarrow R_c \rightarrow SE$$

and

$$SE = instance(R_c, S_p).$$

Since the actual activities performed by human may not match the system’s prior knowledge,\textsuperscript{11} we additionally consider human preferences $(Pref_f)$ below:

$$Pref_f = instance(R_c, a_n \cap Preference) \rightarrow R_c.$$ 

Semantic Decision

Based on the above relationships, we recognize activities using ontology-based semantic reasoning methods. Given a set of the streamed sensory events $E^n$ and possible active candidacies $A'$, we construct an activity thread $AT$, by conducting Terminology Box (T-Box) reasoning on regular, generic activities, and Assertion Box (A-Box) reasoning for user preferred activities:

$$A_T = \{tBox[A', E^n], aBox[Pref_f, Pref_f', E^n]\}.$$ 

We analyze the metadata of sensor event $e_m \in E^n$ for the corresponding entity $(ET_k)$ to deduce candidate relationships with activities; the concurrent activities can be inferred during this process using the above semantic reasoner.

Sensor Events Segmentation

We generalize the concept of sensory events to contain status and usage information gathered in the IoT networks. Inspired by Mallick et al.,\textsuperscript{12} we adopt a transaction-based segmentation as activities may be multiplexed. Given a set of sensory events $E = \{e_1, e_2, \ldots, e_i | e_i \in ES\}$, we segment it into multiple transactions, denoted by $t_{\iota_i} = \{e_i, e_{i+1}, \ldots, e_{\iota_j}\}$ in $Tr$. Suppose ProperCut is the transaction $t_{\iota}$ that matches exactly with activity $a_i$, OverCut includes the transactions that do not contain all transactions, and UnderCut 1\textsuperscript{1} over those transactions that involve multiple activities. Then, our goal becomes to minimize the
number of overcutting and undercutting transactions. To this end, we identify the minimal transactions of sensory events that contain complete activities for our downstream processing, the activity prediction, and use the MinMax algorithm to embed and cluster contextual information of sensors and IoT devices, where we use distances and temporal sequences for segmentation.

**Activity Prediction**
We infer a user’s possible next moves for our downstream processing, such as giving commands to robotic assistants. Then, we employ two stages of the method proposed by Altulyan et al.\textsuperscript{13} to suggest the next items to be used by the human:

1) **Complex activity recognition**: At the current stage, the system learns to recognize simple activities, such as the posture and movement of human. The next step is to recognize complex activities based on predefined ontology and rules.

2) **Recommendation**: At the current stage, the system prompts activities and items to be used by human or robot in the next step. We learn from past trajectories of activities with Q-learning.

**Semantic Searching for Interactive Internet of Things**
The system uses semantic search to identify the suitable objects and generate the procedure for a robot to act toward the inferred human intent. We use Mask R-CNN,\textsuperscript{14} a widely used pixel-level object segmentation model that outputs 1 or 0 to indicate whether a pixel belongs to an object, for object detection. We also combine objective detection with depth information to get 3D positions of objects of interest. Such position information is then used to adapt the learning goals and training/running processes of robots.

**Context-Aware Hypermedia Interactions**
We introduce an RL-based approach named Policy Continuation with Hindsight Inverse Dynamics (PCHID) to generate hardware commands for instructing robotic arms’ movements.

**Policy Continuation With Hindsight Inverse Dynamics**
PCHID\textsuperscript{15} supports self-supervised learning for goal-conditional tasks and can extrapolate the learned policy to complex and nonlinear Hindsight Inverse Dynamics (HID) applications. Specifically, PCHID introduces the goal into inverse dynamics as Hindsight Inverse Dynamics and outputs the minimal \(k\)-step actions to achieve the goal under the optimal policy. This enables us to complete a complex task in multiple steps, each achieving a different subgoal.

**Multistep Goals With Policy Continuation**
While PCHID\textsuperscript{15} makes the system aware of the hindsight goal, more sophisticated settings are introduced and it is considerable challenging. We extend the original Hindsight Experience Replay (HER) model with the ideal proposed by Sun et al.\textsuperscript{15} to let HID embrace \(k\)-step solubility for better optimization. As in the
Universal Value Function Approximators (UVFA), we have possible goals \( g \in G \) and a corresponding reward \( r_g : S \times A \rightarrow R \), where \( S, A \) are the state space and action space, respectively. At timestamp \( t \), we have \( r_t = r_g (s_t, g) \) and policy \( \pi : S \times G \rightarrow A \). The reward is extended from the (0, 1) binary problem set to the following:

\[
\begin{align*}
    & r (s_t, a_t, g) = \lambda |g - s^\text{object}_t|^p - |g - s^\text{object}_{t+1}|^p.
\end{align*}
\]

The hyperparameter \( p \in \{1, 2\} \), and \( \lambda \) may not be limited to 0 or 1 (to be discussed later in our evaluation and experiments).

Based on the UVFA model of HID, we apply the \( k \)-step solvable extension, thus decomposing the state-goal \( S \times G \) into \( S \times G = (S \times G)_0 \cup (S \times G)_1 \cup \ldots \)

**FIGURE 4.** Learning results for the task “Remove cup” on some selected intermediate episodes, where (a) task failed with a gripper of robotic arm hitting the desk; (b) the cup was gripped yet other items were knocked off; (c) the cup was gripped and removed successfully.

**FIGURE 5.** Testing success rates to training episodes for certain tasks. (a) Pick and lift. (b) Place cup. (c) Remove cup. (d) Press switch.
Internet of Things, People, and Processes

EVALUATION

We have implemented live demonstrations of some selected activities and evaluate the feasibility of our system in real-life applications based on simulation.

We test the task learning process of robotic assistants based on RLBench, an open source robotics-related manipulation benchmark that contains 100 daily activities. While RLBench uses the PyRep simulator, we implement our system with the Deepbot framework in the OpenAI Gym environment, which enables us to synchronize the simulation with Robot Operating System (ROS). Our setting involves a Lynxmotion AL5D robotic arm, a TurtleBot equipped with a 3D stereo camera, and some interactive items, such as kitchen utensils.

As aforementioned, the robotic assistant actuates based on the UAVA model of HID, an extension to DQN. We compare PCHID with DQN and its successor HER to demonstrate the effectiveness of our framework in RLBench tasks. Specifically, we set the reward to 0 when the final state is within the tolerable range of the subgoal for each step and otherwise. Figure 3 shows the overall success rates and rewards with respective to episodes of training. The PCHID method is effective and learns significantly faster than HER and the original DQN methods. Figure 4 shows some intermediate learning outcomes: a) and b) failed to complete the task of "Pick and lift" while c) succeeded at episode 20, 50, and 100. Figure 5 further demonstrates the effectiveness of PCHID in some selected tasks: a) Pick and lift; b) Place cups; c) Remove cups; d) Press switch. In our experiments, we set the steps k to 5—we observed a larger step tended to improve the results.

CONCLUSION

In this article, we proposed an interactive ambient intelligence system that enables robots to actively infer human intents and assist humans in accomplishing daily tasks. Our system relies on human activity prediction and reinforcement learning on IoT sensory data in a controlled smart-home environment. In the future, we will develop an interactive item-discovery method to dynamically expand the system’s knowledge base about the environment and prompt robotic actions.

ACKNOWLEDGMENTS

The work of L. Yao and Q. Z. Sheng was supported in part by Australian Research Council under Grant DP210101723 and Grant LP190100140. The work of Z. Wang and X. Xu was supported by the National Key Research and Development Program of China under Grant 2018YFB1402500 and the National Science Foundation of China under Grant 61832004 and Grant 61772155.

REFERENCES

a) and b) failed to complete the task of Figure 4 shows some intermediate learning outcomes: significantly faster than HER and the original DQN methods.

framework18 in the OpenAI Gym environment,19 which daily activities. While RLBench uses the PyRep simu-

technology to generate items, such as kitchen utensils.

Our setting involves a Lynx-based on the UVFA model of HID, an extension to DQN.20

We have implemented live demonstrations of some RLBench tasks. Speci
demonstrate the effectiveness of our framework in...
Democratizing Data-Driven Agriculture Using Affordable Hardware

Ranveer Chandra, Microsoft Corporation, Redmond, WA, 98052, USA
Manohar Swaminathan, and Tusher Chakraborty, Microsoft Corporation, Bangalore, 560097, India
Jian Ding, Yale University, New Haven, CT, 06511, USA
Zerina Kapetanovic, University of Washington, Seattle, WA, 98195, USA
Peeyush Kumar, Microsoft Corporation, Redmond, WA, 98052, USA
Deepak Vasisht, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

The world needs to sustainably grow more food to feed the growing population of the planet. Data-driven agriculture is a promising technique that can help farmers be more productive, reduce costs, and enable adoption of sustainable agricultural practices. However, the adoption of data-driven agriculture is limited by a lack of affordable technologies, for broadband, sensing, imaging, and insights. In this article, we present an overview of Project FarmBeats, a research project that started in 2014 to increase the adoption of data-driven agricultural practices. We provide an overview of the various components of the FarmBeats architecture, and details of the hardware innovations.

The world’s food production needs to increase by 50% by 2050 compared to 2010 levels to feed the growing population of the world.1 Meeting this demand is challenging, given the limited amount of arable land, reduced quality of soil health, and receding water levels. This problem is even more severe if we consider the challenge of nourishing the world, instead of just feeding the world, without harming the planet. One of the most promising approaches to address this challenge is data-driven agriculture. Using the latest advances in artificial intelligence (AI) and cloud computing, a farmer can be issued advisories, such as when to sow seed, and what pests are likely to occur in their farm. They can then use precision agriculture techniques to selectively apply farm inputs, such as water, nutrients, and pesticides. These advances will help the farmer become more profitable, reduce costs, as well as practice sustainable agriculture. This will also drive farm employment in the farm and the digital advisory ecosystem. According to the International Food Policy Research Institute, data-driven techniques can help us achieve this goal by increasing farm productivity by as much as 67% by 2050 and cutting down agricultural losses.2

Data-driven agriculture requires information about the farm. This is obtained from a variety of sources, including sensors, drones, tractors, weather stations, and satellite imagery. Field trials have shown that techniques that use sensor measurements to vary water application in the farm at a fine granularity (precision irrigation) can increase farm productivity by as much as 45% while reducing the water intake by 35%.3 Similar techniques to vary other farm inputs like seeds and soil nutrients have proven to be beneficial. More recently, the advent of aerial imagery systems, such as drones, has enabled farmers to get richer sensor data from the farms. Drones can help farmers map their fields, monitor crop canopy remotely, and check for anomalies. Over time, all of these data can indicate useful practices and make suggestions based on previous crop cycles, resulting in higher yields, lower inputs, and less environmental impact.

While these techniques are promising, their adoption is limited to less than 20% of farmers, even in the United States, owing to the high cost of manual sensor data collection.4 The adoption is much more limited in the low- and medium-income countries, where most farmers are smallholder farmers.5

One of the primary reasons for limited adoption of digital agriculture technologies is the cost of these solutions compared to their value to the farmers. Good-quality agriculture sensors cost over a few hundred dollars. Drones with cameras are commercially available for...
The world needs to sustainably grow more food to feed the growing population of the world. Meeting this challenge is compounded by the paying capacity of the farmers, who are severely financially constrained. More than half of farmers in the United States need a second income to stay afloat. The situation is more dire for smallholder farmers. An average farmer in Sub-Saharan Africa makes less than $2 per day.

In this article, we present a holistic approach for building a more affordable data-driven agriculture system. First, to reduce the cost of sensors, we are exploring new modalities of sensing, such as RF sensing, and ways to reliably use low-cost sensors. Our work, called Strobe, leverages Wi-Fi on existing smartphones to sense soil moisture and soil electrical conductivity (EC). In another work, we propose the Fall-curve to detect sensor failures. Second, for aerial imaging of small farms, we use tethered helium balloons as an alternative to more expensive drones. Third, we reduce the number of sensors needed in a farm using multimodal AI. By combining aerial imagery and on-ground sensors, we need much fewer sensors to build heat-maps of farms. Fourth, we reduce the cost of connectivity by leveraging a communication technology that uses empty TV spectrum to carry wireless signals. Finally, instead of sending all the data to the cloud, we use edge computing—a computer in the farmer’s house or office that is capable of operating offline.

The above mentioned technologies were developed as part of the FarmBeats research project at Microsoft. The goal of FarmBeats is to democratize data-driven agriculture, such that any farmer can augment their knowledge of the farm, with data, and data-driven insights. As part of this initiative, we are developing new technologies, spanning AI, cloud, data, IoT, edge, robotics, networking, and hardware, to make data-driven agriculture more affordable. In this article, we will discuss the innovations in hardware, and the open questions.

DATA-DRIVEN AGRICULTURE: EXISTING APPROACHES AND CHALLENGES

Perhaps the biggest innovation in the last century, which led to the Green Revolution, is the use of genetic engineering. For example, genetically modified organism (GMO) soybean seeds are genetically engineered to handle pesticides better, and can also be planted much closer to each other. Although these innovations have helped, further increase in yield is needed to meet the growing demand.

Data-Driven Agriculture

Data can augment a farmer’s knowledge about their farm, and AI on that data can provide insights to the farmer. A promising data-driven technique is precision agriculture, which treats the farm as heterogeneous (land, livestock, fisheries, etc.), and uses variable treatment throughout the farm, such as variable seeding, fertilizer application, lime application, irrigation, feeding, among others. Precision agriculture is good for the overall farming ecosystem. It improves yield, reduces operating expenses, and is also good for the environment. For example, by only irrigating areas that need water, the farmer gets a healthy crop, and by not using water where it is not needed, the farmer saves money, while preventing surface runoff, and nutrient leaching. Similarly, by applying fertilizer only where needed, the farmer saves cost, and limits the damage caused by overfertilization. For livestock, adapting the feed for each animal using precision nutrition can help increase productivity for meat or dairy.

In addition to precision agriculture, data can also help a farmer with other digital advisories, such as planning what to grow, when to plant seeds, and when to harvest based on market prices and logistics overheads. Data-driven agriculture can also help a farmer create the right market linkages, provide access to financial tools and insurance, and help with agricultural research and development.

Data-driven tools can help farmers adapt to climate change. Farmers take a lot of actions based on past weather conditions. An unexpected change, either in temperature or precipitation, will significantly impact a farmer’s yield. Data-driven techniques can predict these changes and provide timely notifications to the farmer. In addition, agriculture can be a potential solution to climate change, by helping sequester carbon in soil. However, this requires the use of regenerative agricultural practices, such as no till, or reduced till, cover cropping, and nutrient management. The use of digital technologies can help a farmer adopt regenerative agricultural practices, without sacrificing profitability.

Existing Approaches

Agronomists have studied various aspects of precision agriculture, from defining more accurate management
zones, to improving prescription, to leveraging soil science, and plant physiology techniques. Remote sensing from satellite imagery and soil samples from the lab are the most commonly used data-driven techniques. Recent work has looked at technologies for gathering data from farms. Researchers have built specialized sensors, for measuring nutrients, water levels, and other such sensors. Practitioners have started using drones to get a spatial view of the farm, and mesh networks to gather data from sensors in the farm.

Challenges
We note that despite the promise of data-driven agriculture, it has not been widely adopted. For example, only 13% of small holder farmers in Sub-Saharan Africa have registered for digital services, and less than that are active.12

Even in the United States, precision agriculture is still in its infancy. The primary reason is cost and inaccuracy. Satellite data are sparse, with coarse spatial resolution, and lack temporal data below the clouds. On-farm data collection technologies are expensive to provide the return of investment (ROI) to the grower. Several commercial products build management zones, but they are often unable to capture the spatial and temporal variations in the field caused by climatic and soil variations. This variation is even more for soil organic carbon, which needs to be monitored for agricultural emissions and carbon sequestration.

FarmBeats: A PLATFORM FOR DIGITAL AGRICULTURE
FarmBeats was started as a research project in Microsoft Research in 2014, with a goal to enable data-driven agriculture. Since then, parts of the research have been shipped as a Microsoft product called Azure FarmBeats.13 Organizations such as Land O’Lakes14 and USDA ARS15 have since announced partnerships on Azure FarmBeats for their agricultural products.

The FarmBeats research system is innovating on the end-to-end system for digital agriculture, as shown in Figure 1. We work with partners to prototype agricultural services for farmers. Given any farm, which could be a polygon or a shape file, the system captures large amounts of data about the farm, from a variety of data sources, including sensors, drones, tractors, cameras, satellites, and weather stations. This includes both temporal, spatial, and historical data. The system then uses AI to combine these data in new ways to fill in gaps, and make predictions of what is likely to happen in the farm. This abstraction is available via APIs to partners, who then use their detailed agricultural knowledge to develop agricultural insights for growers.

The research system described in Figure 1 works as follows:

1) Sensors: The system recommends sensor locations based on knowledge of the farm. These are regions where the farmer should place the sensors.
Our research includes ways to reduce the cost of sensing using Wi-Fi signals on smartphones, and improve the fault tolerance of sensors.  

2) **Aerial imaging**: Drones are used to capture imagery from large parts of the farm, and to spray chemicals and water. We have researched ways to improve the battery life of drones, and intelligent path planning. We have also invented a low-cost way to image farms using smartphones with battery packs mounted on tethered helium balloons.

3) **Networks**: Since large areas in the farm do not have Internet access, and even if they do get a wireless signal reception, the connection is not affordable. We use a new technology that uses unused TV channels to send and receive data. Antennas on a farmer’s house or office send and receive signals over several miles, providing low cost, high-speed connectivity over long distances. In addition, we have come up with a new radio design for low-power, long-range, narrowband operation in the TV spectrum.

4) **Edge compute**: Not all the data generated in the farm need to be sent to the cloud. In some cases, it is extremely prohibitive. For example, a drone can generate several gigabytes of data in tens of minutes. Transmitting these data to the cloud over a few Mbps Internet connection will take extremely long. Instead, we perform large amounts of compute on a PC form factor device in the farmer’s house or office. We have also invented a new technique to send large amounts of data from the edge to the cloud by first identifying parts of the image that are more important, and then selectively sending the subframe fragments using progressive compression.

5) **Satellite imagery**: One of the key sources of data for a farm is from satellite imagery. FarmBeats ingests current and historical satellite data about a farm. One of the challenges in satellite imagery is clouds. Since over 70% of satellite data are covered by clouds, it is difficult to observe the farm through the clouds. We have invented a new technique, called SpaceEye, that combines imagery from optical satellites with RF signals from RADAR satellites to reconstruct satellite imagery below the clouds, with high accuracy.

6) **AI and computer vision**: We use AI on farm data to a) fill in gaps in collected data, and b) predict likely outcomes. Both of these are performed at the edge and in the cloud. We merge data across multiple data streams, such as sensors and aerial imagery, sensors and weather stations, drones and satellites, etc. For example, using multimodal techniques, we are able to combine local sensor data with weather station data to make very hyperlocal predictions of weather in the farm. We also use computer vision techniques to build 3-D orthomosaics, and create aerial timelapses.

We note that the architecture in Figure 1 is not designed to be growing facing. It is meant for other AgTech researchers to incorporate their agronomic expertise with the data collection and AI capabilities of FarmBeats. In the rest of this article, we discuss the hardware innovations in components 1–4 of Figure 1.

### RELIABLE, LOW-COST SENSING

Existing sensing technologies in a farm are expensive. They also require a lot of sensors in the farm, which further drives up costs. Our research has focused on ways to significantly bring down the cost of sensing by:

1) leveraging other forms of sensing, which are more commonly available, such as RF;
2) making low-cost sensors more reliable by detecting faults early;
3) using AI/ML techniques to enable low-cost sensors to function as more powerful weather stations;
4) reducing the number of sensors needed in a farm by combining sensors with multispectral imagery from drones and satellites.

We discuss the first two technologies in the following sections.

#### Sensing Using RF

Several technologies for measuring soil moisture and EC have been invented in the last few decades, including direct sensing techniques, which require soil to be extracted and dried out, as well as indirect sensing methods that measure surrogate properties of soil moisture and EC, such as capacitance, electrical, and nuclear response. However, their adoption is limited by the cost and accuracy. Even sub-1000 dollar sensors can fail to accurately measure soil EC and moisture. Our work, Strobe, leverages RF signals in existing Wi-Fi bands to bring down the cost to tens of dollars as well as achieving comparable performance to more expensive soil sensors.

RF-based soil sensing is enabled by the phenomenon that RF waves propagate slower and attenuate more in soil than in air because of soil’s larger permittivity and EC than air. Unlike prior RF-based solutions such as ground-penetrating radars and time-domain reflectometry that require a wide bandwidth from
hundreds of megahertz to a few gigahertz to achieve high accuracy, Strobe only exploits the 70 MHz of the available Wi-Fi spectrum in 2.4 GHz. Strobe overcomes the key challenge of limited bandwidth using a novel multiantenna technique. With a linear antenna array, Strobe measures the relative propagation time and amplitude of Wi-Fi signals received by different antennas, and then converts them to soil moisture and EC. Strobe can work with commodity multiantenna Wi-Fi cards, which brings down the cost to be tens of dollars. We expect the cost to be lower when the system is manufactured at a larger scale, e.g., tens of thousands of devices.

Reliable Sensors
Sensor hardware for agriculture, especially low-cost sensors, are prone to faults as they are exposed to harsh outdoor environments. Water, humidity, extreme temperatures, and insufficient power can all lead to faulty sensor hardware and, in turn, corrupt data. There are several real-world scenarios where existing data-centric and heuristic-based approaches have limitations in sensor fault detection. Our work, Fall-curve, focuses on identifying and isolating defective sensor hardware at the edge.

We exploit a unique hardware signature where the signature of a working sensor is distinct from a faulty one. To save power in an IoT device, a common practice is to power off the sensor when the data collection is not required. We observe that when a sensor is powered off, its output voltage signal gradually falls down to zero following a curve. The Fall-curve is exhibited due to the presence of active and parasitic capacitances in the sensor hardware. Consequently, the curve is unique to each sensor, i.e., its hardware components. Therefore, any hardware malfunction results in a different shape of the curve.

An advantage of using the Fall-curve is that the fault can be identified at the end IoT device without any contextual knowledge and historic sensor data. To do so, we use a lightweight nearest neighbor search algorithm running in the end IoT device. A dictionary of polynomial feature vectors extracted from the Fall-curves of nonfaulty sensors is prestored in the IoT device. The search is conducted on this dictionary against feature vectors extracted from the Fall-curve of the connected sensor.

Broadband in the Farm
In addition to other technologies, we use the TV white spaces (TVWS) to connect the farm. This technology refers to unused TV spectrum, in the very high frequency (VHF) and ultrahigh frequency (UHF) bands, which are legal to use in the United States, Canada, and several countries worldwide. Devices using this technology consult with a database to determine the available channels at a location, and operate in an available TV channel. Since this spectrum is in the lower frequencies, the signals can propagate much farther than signals in 2.4 GHz or 900 MHz of the spectrum, and also through dense leaves and crop canopies. For example, recent deployments in Africa have links operating at over 2 Mbps at 10 km when transmitting at 1 W. Furthermore, earlier this year, the Federal Communications Commission (FCC) approved the TVWS devices to operate at up to 10-W EIRP in rural areas, enabling extremely long-range networks.

Even though there are few free TV channels in metropolitan cities, the rural areas, where most farms are located, have more than 100-MHz TV spectrum. Given this large available bandwidth, we are able to support several devices, including high-bandwidth devices, such as cameras, operating at long distances using a single gateway device. Furthermore, since these devices operate in unlicensed spectrum, their cost is significantly less than comparable cellular networks. Farmers can set up the TVWS radio and antenna, as shown in Figure 2, to connect several miles around a farm.
Connecting Sensors

LoRa and SIGFOX IoT networks operate around the 2.4-GHz or 900-MHz band, which have limited propagation through crops and canopies. They can operate in 400-MHz “semilicensed band” in the United States, but that needs coordination with a local agency to get a very small sliver—12.5 KHz—of the spectrum, which can support very few sensors.

We propose the use of TVWS spectrum, which is abundantly available in rural areas, for IoT communication. It offers very long-range connectivity over tens of miles, even through crops and canopies. As part of the FarmBeats system, we designed and implemented a narrowband IoT radio that can operate in this spectrum (as seen in Figure 3). This opens the door for very large-scale network deployments where a single base station can support hundreds of devices across tens of miles.

Although our IoT radio offers longer range, it consists of low-cost off-the-shelf hardware components. It enables low-power LoRa communication over the TVWS spectrum by incorporating the SX1276 chip from Semtech as the LoRa (de)modulator. We modified the RF filters to operate outside the industry, scientific, and medical (ISM) band, including the VHF and UHF TV spectrum. The TV spectrum has strict regulatory restrictions on side-channel leakage and harmonics of the signal to protect the primary user from the harmful interference. It is also very wideband compared to the 900-MHz ISM band, and separate filters are required to operate in different bands within this spectrum. We designed a software configurable logarithmic periodic filter that enables narrowband communication in a continuous spectrum starting from 150 to 960 MHz including VHF, UHF, and ISM bands. The filter design limits the side-channel leakage and harmonics in other bands within the spectrum. We modified the base LoRa media access control (MAC) protocol, such that these transmissions do not interfere with existing TV reception. We provided our results to the FCC, including the design, and in November 2020, for the first time, the FCC approved regulations for use of IoT devices in the TVWS spectrum.

LOW-COST SMALL-SCALE AERIAL IMAGING

In spite of recent advances in unmanned aerial vehicle (UAV) technology, a few factors limit their adoption for smallholder farmers. First, UAVs consume a large amount of power to stay afloat, resulting in very short battery life (few tens of minutes for most commercial UAVs). Second, there are several regulatory restrictions associated with UAV usage. Finally, UAVs require high capital investment. Commercial quadrotors that can last for 30 min and are reliable outdoors cost over $1,000. This is further compounded by the fact that the UAV batteries have finite charge cycles and need to be replaced frequently if the UAV is used often.

In a previous work,10 we present a long-term low-cost aerial imagery platform called TYE (for Tethered eYE). TYE is a tethered aerial camera that floats in the air, at a few hundreds of feet, due to the lift provided by a balloon, as shown in Figure 4. We utilize a lighter-than-air gas (such as helium) filled reusable tethered balloon system to carry a payload (a camera and some additional hardware). Unlike UAVs, TYE is low cost, does not suffer from regulatory restrictions, and can last for several days. TYE operates in two different modes: 1) Static-TYE, where the balloon is tethered to a stationary point, for long-term unmanned aerial imagery applications, such as surveillance and crowd monitoring; and 2) Mobile-TYE, where the tether is movable, for spatial mapping applications that require TYE to map a large area, such as crop canopy estimation in farms. Raw visual imagery from TYE is extremely hard for humans to view and understand.
The balloons are highly susceptible to arbitrary lateral motion and camera rotations induced by wind. This constantly changes the field of view of the TYE camera across subsequent frames. To overcome this problem, we designed a custom mount to reduce camera mobility, and leverage the gyroscope to eliminate frames with extreme motion. We then utilize techniques from computer vision to correct for the camera motion in software. As a result, TYE produces consistent views for the user as if the camera was stable in the air, in spite of arbitrary physical camera motions.

While long-term imagery applications of TYE do not require any human involvement, the spatial mapping application for aerial imagery of a farm requires TYE to be maneuverable like drones. Thus, Mobile-TYE requires a person to move the balloon around while holding the tether. In this mode, wind-induced motion of the balloon causes the balloon-path to be uncorrelated with the human path. This leads to novel path planning challenges, since most area-coveraging algorithms are designed for UAVs that can follow tight global positioning system (GPS)-controlled paths by exerting turbine forces to counter the force of wind. In order to overcome this challenge, we propose a novel path-planning algorithm that ensures area coverage, with minimum human motion, in spite of balloon motion caused by wind. Furthermore, we implemented this algorithm as a mobile application for the user to keep adapting their path in response to the balloon motion.

We have implemented TYE as a software–hardware system, using helium balloons. We support two camera designs: a GoPro and a smartphone. Furthermore, we evaluated the feasibility of TYE in two different applications: 1) aerial survey of crops in agriculture (both static-TYE and mobile-TYE) and 2) flood monitoring to identify the flow of water in a flood (static-TYE).

AI FOR AFFORDABILITY

In addition to reducing the cost of hardware, AI can help make digital agriculture affordable by 1) reducing the number of sensors needed, and making sensors and drones more functional using multimodal AI, by combining with other sources of data, such as weather station data, or satellite imagery; and 2) improving the ROI in on-farm hardware by using AI to advise farmers for optimal strategies on seeding, spraying, harvesting, and trading decisions to make farming more sustainable and profitable (see Kumar et al.’s work21).

Multimodal AI on Farms

Inference based on multimodal data provide a more holistic perspective of the farm. Our work develops some key technologies to combine multiple modalities: 1) combining multiple temporal data streams (such as on-farm sensors with weather stations) representing various geospatial scales; and 2) combining spatial data stream (such as drone or satellite imagery) with temporal data stream (e.g., from sensors) sampled at various temporal resolutions.

Consider a specific scenario—It is springtime in Eastern Washington, USA, and the temperature is slightly above freezing. A farmer is preparing to fertilize their fields of wheat and lentils as winter runoff and frost are nearly finished. The plants are susceptible to fertilizer at freezing temperatures, so the farmer checks forecasts from the local weather station, which is about 50 miles away. The three-day outlook shows temperatures above freezing. The farmer rents equipment and starts fertilizing the farm. But at night, the temperature in parts of the fields drops below freezing and kills around 20% of the crops. This is unfortunately a common situation on farms, since climatic parameters can vary over short distances and even between sections of the farm.

To address this problem and others, we developed DeepMC, a multiscale encoder–decoder framework to combine weather station forecasts (which is collected and predicted at a coarser geo-spatial scale) with sensor data (which is collected at a highly localized geo-spatial scale) to predict microclimates on the farm. This framework is called DeepMC (see Kumar et al.’s work18). DeepMC predicts various microclimate parameters with over 90% accuracy at IoT sensor locations deployed in farms around the world.

Some of our other work on multimodal AI focuses on spatiotemporal data sets. We combine multispectral spatial data streams collected through either satellites or drones with IoT sensor-based temporal data sets to generate high-resolution heatmaps of soil and crop parameters on the farm. Large number of sensors on the farm create operational challenges for operating machinery on farms, in addition to the added cost of procuring and maintaining sensors on the farm. We keep the deployed sensors at the minimum by computing the optimal sensors needed for the desired resolution of the predicted outcomes. In addition, the technique is also used to generate heatmaps for various soil and crop parameters such as soil moisture, soil temperature, and soil nutrient content on the farm. Our technique uses a fusion mechanism to combine spatial data sets with temporal signals sampled at various temporal resolutions.4

Optimal Advisories

Farmers synthesize various parameters to make operational decisions on the farm. Typically, these decisions can be categorized into three groups:
1) Strategic: These are decisions based on long-term impact and are not dependent on day-to-day variabilities, such as resource allocation—how many workers to hire? What crops to grow in which section for this year?

2) Tactical: Scheduling decisions that are based on day-to-day operations—such as when and where to spray?

3) Real-time: Decision that are dependent on real-time feedback: such as operational automation—flight planning for drones, tractors, etc.

Most of these decisions are made under various uncertainties—nature, human, or machinery related. We developed a sequential decision making framework to advise operations on the farm by taking in insights collected by the FarmBeats system and compute optimal actions for strategic, tactical, and real-time decision-making. This system solves for optimal strategies by combining farm-specific guidelines with natural, market, and policy-based rules with signal dynamics. These AI technologies help democratize agriculture by bringing actionable insurgents and advisories in the hands of individual users in an affordable and quality manner.21

SUMMARY AND FUTURE WORK

Innovations in hardware can help democratize digital agriculture. Low-cost sensors can be made more fault tolerant using Fall-curves. Alternative sensing methodologies, such as RF sensing using Wi-Fi, can enable any farmer with a smartphone to get data about their farm. Low-cost imaging will help growers to get aerial data at low cost, and new low-cost networking techniques, such as the use of unused TV channels, can further bring down the cost of existing solutions by not relying on satellite connectivity. AI techniques can also reduce the need for expensive sensors, and also reduce the number of sensors needed in the farm.

However, we note that we have only scratched the surface in making sensing of the farm more affordable. A lot more innovation is needed, in technology, business, and policy, to further reduce costs, and increase adoption of digital agriculture techniques.

On the technology side, we need to develop more inexpensive forms of sensing. Recent work on using microelectromechanical systems (MEMS) sensors for sensing soil, and audio sensors for measuring rain, are very promising. In addition to sensors, we need lower cost and energy efficient drones that can reduce the overhead in procuring aerial imagery. Cloud and AI that can reach to all farmers worldwide, for example, using an affordable edge. Furthermore, research is needed on user interfaces, to make digital agriculture more usable by farmers who operate devices with soiled hands, and are often not the most technology savvy.

Business and policy innovation are also needed to further increase the adoption of digital agriculture solutions. New business models, such as sensing as a service, in which people on the field can carry a sensor to different parts of the farm instead of using multiple sensors, or new business models where farmers get paid to use digital techniques, can help drive adoption. Similarly, policy innovation is needed as well. Similar to subsidies for irrigation and farm equipment, there needs to be subsidies for digital technologies, to drive the adoption of digital agriculture technologies.

REFERENCES


RANVEER CHANDRA is the CTO of Agri-Food and the Managing Director of Research for Industry at Microsoft, Redmond, WA, USA. His research focuses on networking and systems. He is an IEEE Fellow, and has received multiple best paper awards, and recognitions, such as the MIT Technology Review Top Innovator Under 35 (TR-35), and the America’s 50 most Disruptive Innovators by the Newsweek Magazine. Contact him at Ranveer@microsoft.com.

MANOHAR SWAMINATHAN is a researcher with the Technologies and Empowerment Group, Microsoft Research India, Bengaluru, India. His persistent aspiration is to engineer a better quality of life for marginalized communities using his expertise in a range of fields including education, mobile computing, IoT, HCI, AR and VR. His research focuses on introducing computational thinking for children in schools for the blind in India using the methodology of Ludic Design for Accessibility. Contact him at swmanoh@microsoft.com.

TUSHER CHAKRABORTY is a Research Software Engineer at Microsoft Research India, Bengaluru, India. His work focuses on the Internet of Things (IoT) and wireless sensor networks (WSNs). He builds end-to-end real-world networking systems that light up new business opportunities for the industries with novel research directions. Contact him at tusherc@microsoft.com.

JIAN DING is a Ph.D. student with the Computer Science Department, Yale University, New Haven, CT, USA. Her research interest is wireless systems, with a focus on the physical layer of 5G massive MIMO system and RF sensing. She was the recipient of the Best Paper Honorable Mention Award at Mobicon 2019. Contact her at jian.ding@yale.edu.

ZERINA KAPETANOVIC is a Ph.D. candidate with the Electrical and Computer Engineering Department, University of Washington, Seattle, WA, USA. Her research focuses on low-power wireless communication, battery-free sensing, and the Internet of Things. She was the recipient of the Microsoft Research Dissertation Grant. Contact her at zerinak@uw.edu.

PEEYUSH KUMAR is a Senior Research Scientist at Microsoft Research, Redmond, WA, USA. His research is broadly in the field of data-driven decision making under uncertainty, reinforcement learning, game theory, and broader AI, particularly, in industry applications to sustainability, supply chain and logistics, agriculture, energy, and healthcare decision making. Contact him at pekumar@microsoft.com.

DEEPAK VASISHT is an Assistant Professor at the University of Illinois Urbana-Champaign, Urbana, IL, USA. His research focuses on mobile computing and wireless networking. He was the recipient of two best paper awards, the ACM SIGCOMM Doctoral Dissertation Award, and the Microsoft Ph.D. Fellowship. Contact him at deepakv@illinois.edu.
PURPOSE: Engaging professionals from all areas of computing, the IEEE Computer Society sets the standard for education and engagement that fuels global technological advancement. Through conferences, publications, and programs, IEEE CS empowers, guides, and shapes the future of its members, and the greater industry, enabling new opportunities to better serve our world.

OMBUDSMAN: Direct unresolved complaints to ombudsman@computer.org.

CHAPTERS: Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

AVAILABLE INFORMATION: To check membership status, report an address change, or obtain more information on any of the following, email Customer Service at help@computer.org or call +1 714 821 8380 (international) or our toll-free number, +1 800 272 6657 (US):
  - Membership applications
  - Publications catalog
  - Draft standards and order forms
  - Technical committee list
  - Technical committee application
  - Chapter start-up procedures
  - Student scholarship information
  - Volunteer leaders/staff directory
  - IEEE senior member grade application (requires 10 years practice and significant performance in five of those 10)

PUBLICATIONS AND ACTIVITIES

Computer: The flagship publication of the IEEE Computer Society, Computer, publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.

Periodicals: The Society publishes 12 magazines, 19 journals.

Conference Proceedings & Books: Conference Publishing Services publishes more than 275 titles every year.

Standards Working Groups: More than 150 groups produce IEEE standards used throughout the world.

Technical Communities: TCs provide professional interaction in more than 30 technical areas and directly influence computer engineering conferences and publications.

Conferences/Education: The society holds more than 215 conferences each year and sponsors many educational activities, including computing science accreditation.

Certifications: The society offers three software developer credentials.

COMPUTER SOCIETY OFFICES

Washington, D.C.: Los Alamitos:
2001 L St., Ste. 700, 10662 Los Vaqueros Cir.,
Washington, D.C. 20036-4928; Los Alamitos, CA 90720;
Phone: +1 202 371 0101; Phone: +1 714 821 8380;
Fax: +1 202 728 9614; Email: help@computer.org
Email: help@computer.org

MEMBERSHIP & PUBLICATION ORDERS
Phone: +1 800 272 6657; Fax: +1 714 821 4641;
Email: help@computer.org

EXECUTIVE COMMITTEE

President: Nita Patel
President-Elect: Jyotika Athavale
Past President: William D. Gropp
First VP: Hironori Washizaki
Second VP: Grace A. Lewis
Secretary: Carolyn McGregor
Treasurer: Michela Taufner
VP, Membership & Geographic Activities: Fernando Bouche
VP, Professional & Educational Activities: Deborah Silver
Interim VP, Publications: Greg Byrd
VP, Standards Activities: Annette Reilly
VP, Technical & Conference Activities: Grace A. Lewis
2023–2024 IEEE Division VIII Director: Leila De Floriani
2022–2023 IEEE Division V Director: Cecilia Metra
2023 IEEE Division V Director-Elect: Christina M. Schober

BOARD OF GOVERNORS

Term Expiring 2023:
Jyotika Athavale, Terry Benzel, Takako Hashimoto, Irene Pazos
Viana, Annette Reilly, Deborah Silver

Term Expiring 2024:
Saurabh Bagchi, Charles (Chuck) Hansen, Carlos E. Jimenez-Gomez, Daniel S. Katz, Shixia Liu, Cyril Onwubikpo

Term Expiring 2025:
Ilkay Altintas, Nils Aschenbruck, Mike Hinchey, Joaquim Jorge, Rick Kazman, Carolyn McGregor

EXECUTIVE STAFF

Executive Director: Melissa Russell
Director, Governance & Associate Executive Director: Anne Marie Kelly
Director, Conference Operations: Silvia Ceballos
Director, Information Technology & Services: Sumit Kacker
Director, Marketing & Sales: Michelle Tubb
Director, Membership Development: Eric Berkowitz
Director, Periodicals & Special Projects: Robin Baldwin

IEEE BOARD OF DIRECTORS

President & CEO: Saifur Rahman
President-Elect: Thomas M. Coughlin
Director & Secretary: Forrest (Don) Wright
Director & Treasurer: Mary Ellen Randall
Past President: K. J. Ray Liu
Director & VP, Educational Activities: Rabab Ward
Director & VP, Publication Services & Products: Sergio Benedetto
Director & VP, Member & Geographic Activities: Jill Gostin
Director & President, Standards Association: Yu Yuan
Director & VP, Technical Activities: John Verboncoeur
Director & President, IEEE-USA: Eduardo Palacio

EXECUTIVE STAFF

Executive Director: Melissa Russell
Director, Governance & Associate Executive Director: Anne Marie Kelly
Director, Conference Operations: Silvia Ceballos
Director, Information Technology & Services: Sumit Kacker
Director, Marketing & Sales: Michelle Tubb
Director, Membership Development: Eric Berkowitz
Director, Periodicals & Special Projects: Robin Baldwin

IEEE BOARD OF DIRECTORS

President & CEO: Saifur Rahman
President-Elect: Thomas M. Coughlin
Director & Secretary: Forrest (Don) Wright
Director & Treasurer: Mary Ellen Randall
Past President: K. J. Ray Liu
Director & VP, Educational Activities: Rabab Ward
Director & VP, Publication Services & Products: Sergio Benedetto
Director & VP, Member & Geographic Activities: Jill Gostin
Director & President, Standards Association: Yu Yuan
Director & VP, Technical Activities: John Verboncoeur
Director & President, IEEE-USA: Eduardo Palacio

EXECUTIVE STAFF

Executive Director: Melissa Russell
Director, Governance & Associate Executive Director: Anne Marie Kelly
Director, Conference Operations: Silvia Ceballos
Director, Information Technology & Services: Sumit Kacker
Director, Marketing & Sales: Michelle Tubb
Director, Membership Development: Eric Berkowitz
Director, Periodicals & Special Projects: Robin Baldwin

IEEE BOARD OF DIRECTORS

President & CEO: Saifur Rahman
President-Elect: Thomas M. Coughlin
Director & Secretary: Forrest (Don) Wright
Director & Treasurer: Mary Ellen Randall
Past President: K. J. Ray Liu
Director & VP, Educational Activities: Rabab Ward
Director & VP, Publication Services & Products: Sergio Benedetto
Director & VP, Member & Geographic Activities: Jill Gostin
Director & President, Standards Association: Yu Yuan
Director & VP, Technical Activities: John Verboncoeur
Director & President, IEEE-USA: Eduardo Palacio
The AI Chip Race
Guansong Pang, School of Computing and Information Systems, Singapore Management University, Singapore, 178902

The strong demand of computing power for artificial intelligence (AI) and machine learning is accelerating the race to develop cheaper and faster AI chips. The AI chip market was valued $10.6 billion in 2021 and the total revenue is expected to reach $79.8 billion by 2027.1 To be part of the market, tech giants from different countries have been successively joining the race, while AI chip startups attracting billions of dollars are taking off like a rocket.

AI chips are hardware accelerators specifically designed to accelerate AI and machine learning-based applications. They generally include graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and certain types of application-specific integrated circuits (ASICs) specialized for AI calculations.2 Deep neural networks (DNNs) are the cutting-edge, computationally intensive AI systems that these accelerators are tailored to. As a popular machine learning approach, DNNs consist of two key stages—training and inference—DNN models are fed with large-scale data to extract useful patterns during training, and they are then used to make predictions for unseen data during inference.

General-purpose chips, such as central processing units (CPUs), have strong sequential operation capability, but they cannot provide sufficient performance for techniques like DNNs that require intensive parallel computation and high-bandwidth memory. Specialized AI chips can be up to thousands of times faster than CPUs for training and inference of DNNs.1,2

Nvidia recently released H100 GPUs, offering about an order-of-magnitude leap compared to its precedent A100 GPUs.3 GPUs have been a dominant hardware tool to accelerate AI systems, especially excelling at training computationally costly DNN models. It enjoys the strong support of the parallel computing platform compute unified device architecture, and it is a type of widely commercialized AI chips.

Different from Nvidia, tech giants have been heavily investing on ASICs, as ASICs are relatively easier to design and have lower barriers to entry.4 Google’s tensor processing units (TPU) is one of such types. Google released its latest TPU v4 in May 2021, doubling the performance over the prior TPU v3 chips.5 Amazon recently released the second generation of its own AI chip trainium, claiming big cost-effective performance over Nvidia’s A100 chips.6 These specialized chips are used in their own AI research and applications, or their computing cloud to offer cloud service to external customers.

Intel and AMD-Xilinx instead focus on the FPGA market. FPGAs are more customizable than ASICs, but using FPGAs a requires lengthy, and complicated development process, which might be skirted by the latest FPGAs of AMD-Xilinx, VCK5000.7 Intel recently also introduced Agilex M-Series FPGAs, claiming the world’s fastest FPGAs with in-package high bandwidth memory.8 Compared to GPUs, ASICs and FPGAs typically have better inference efficiency.

Tech giants from countries other than the United States have also joined the race. Chinese tech giant Baidu introduced the second generation of Kunlun AI chips last year. Huawei unveiled two AI chips—Ascend 910 and Ascend 310—before United States sanctions. LG also vouched to develop their own AI chips.8 The


AI chip race among countries is accelerating among startups. The front runners of Chinese AI chip startups, such as Cambricon and Illuvatar CoreX, have released their own AI chips; other startups, such as Biren and MetaX, have attracted huge investment from Chinese venture capital to challenge the dominance of United States in the AI chip market. There are also a rising number of United States AI chip startups, some of which have introduced new AI chips, such as SambaNova Systems and Cerebras Systems. Active investment has also been funneled into many other young AI chip makers, such as British startup Graphcore, Japanese startup PEZY Computing, and Canadian startup Tenstorrent, which have recently claimed new AI chips with a large improvement in efficiency for deep learning.

Given the tremendous market value and the strategic significance to each country, the AI chip race among tech giants, startups, and countries is expected to further accelerate in the future. The market share among GPUs, FPGAs, and ASICs is likely to change, as well.

**REFERENCES**


Taking a Measured Approach to Investing in Information Infrastructure for Attaining Leading-Edge Trustworthy Artificial Intelligence

James Bret Michael, Associate Editor in Chief

As we ramp up our investment in our artificial intelligence (AI) infrastructure, concerns about the reliability, resiliency, and security of that infrastructure need to be addressed. Competition, whether political, economic, or otherwise, along with factors such as the rapid evolution of information technology, contributes to an ever-present tension when choosing between taking an expedient or a measured approach to investing in information infrastructure. We see this pressure playing out for AI. Our ability to conduct responsible AI (RAI) research and development goes beyond addressing data and algorithm requirements. It also requires the availability and proper use of state-of-the-art information infrastructure.

There are world leaders and industry executives who claim with certitude that market and technological dominance in harnessing AI are key to obtaining and maintaining strategic advantage over other nations and industry competitors. There are also policymakers who advocate for swift government-directed and funded expansion of information infrastructure, drawing analogies to major infrastructure projects, such as the U.S. government’s investment in the Dwight D. Eisenhower National System of Interstate and Defense Highways and the public-private partnership that resulted in the establishment of the Internet.

But what exactly is AI infrastructure, what makes such infrastructure leading-edge, and why take a calculated approach to investing in it? Today’s AI infrastructure consists of a combination of general-purpose and AI-specific components, all of which in some way support end-to-end AI workflows. Some examples of general-purpose infrastructure for AI are cloud storage and 5G mobile networks. Examples of AI-specific infrastructure components include such things as integrated circuits whose architecture has been optimized for performing deep learning (e.g., for a convolutional neural network) and tools for artifact management and metadata storage for developing machine learning models.

Of course, an AI infrastructure that is considered today to be leading edge will not be so for long, given technological advances that enable the adoption of more efficient and effective end-to-end AI workflows. For example, in the context of machine learning, Neil Conway, of Hewlett Packard Enterprise, argues that many of today’s cutting-edge AI infrastructure components are designed to support “a single person developing a single model with a cluster of machines.” This is unfortunate, as those infrastructure components, along with shadow and siloed AI infrastructure, are not aligned with end-to-end machine learning workflows that involve collaborative and distributed development of machine learning models. Conway also states that “many tools are either too generic or narrow technical point solutions,” which means they too are not congruent with end-to-end AI workflows. Conway and other AI practitioners argue that this results in them spending inordinate amounts of time and effort on infrastructure management and writing boilerplate software to support their end-to-end AI workflow processes.
There are many ways in which the available infrastructure can distract the attention of AI practitioners from their focus on using machine learning to solve customers’ problems. For instance, if the infrastructure does not provide an adequate framework for fault tolerance capabilities specific to machine learning, then AI practitioners need to write software to perform the necessary saving and/or loading of checkpoints. In addition, according to one of my colleagues, when working with “a custom model and especially if transitioning between different operating systems or precisions, then the only way to reload the model checkpoint is to redefine the model on the specific operating system (e.g., this is typical of loading a model on a Raspberry Pi ARM core that was trained on an Intel- or AMD-based system) and load weights of the model rather than loading the overall model. If you develop custom layers in the model, then you need to provide definitions of those on the target platform.”

Conway notes that AI practitioners can make errors in implementing the checkpointing functionality, but even if they get it right, the engineer will still need to update his or her model’s code as changes are made to the underlying AI infrastructure. For the distributed training of models, checkpointing is not necessary, and a framework (which is part of the infrastructure) handles the distribution.

We can expect that tomorrow’s cutting-edge AI infrastructure components will become more specialized for AI, supporting scalable end-to-end AI workflows while reducing the amount of manual work that AI practitioners need to do to get the infrastructure to support tasks such as tuning hyperparameters. Among other things, we will see the continued growth of infrastructure support for edge AI, due to the need to move the development and operation of AI capabilities to be close to the sources of data, whereas today’s commercial AI infrastructure relies heavily on cloud services to support end-to-end AI workflows.

So why take a measured approach to improving AI infrastructure? What may first come to mind is that enhancing AI infrastructure requires the use of scarce funding (e.g., obtaining economies of scale) and challenging-to-find human capital. However, another, just as compelling, reason is that AI practitioners need to consider how those improvements will affect the dependability of the resulting infrastructure, end-to-end AI workflow, and AI applications that ride on top of the infrastructure. The U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway lists lines of effort (LOE) specific to AI infrastructure, such as LOE 2.1.1, which calls for the acquisition and sustainment of a test and evaluation, verification, and validation framework tailored to support the AI product lifecycle and for use across the entire enterprise, and LOE 3.2.4, the “Fund the Development and Piloting of New Resources and Tools That Augment the RAI [Responsible AI] Toolkit.”

Consider edge AI, which I wrote about in this column in 2021. 5G New Radio (NR) and commercial satellite communication services can make it possible to implement end-to-end AI workflows at the network edge. However, security vulnerabilities have been identified in 5G NR, such as those related to the Global Navigation Satellite System. In addition, a recent report released by a group of researchers advocates for the military forces of the People’s Republic of China to acquire and employ capabilities to degrade, damage, and destroy SpaceX’s Starlink satellite Internet service (https://www.starlink.com/). So far, Starlink’s infrastructure (i.e., satellites, ground stations, and user terminals) has demonstrated its resilience against electronic warfare and cyber-based attacks conducted by the Russian Federation. The service has also been able to withstand the loss of satellites during a geomagnetic storm, in part due to the sheer number of Starlink satellite constellations in orbit. We play a continuous cat-and-mouse game to make these and other general-purpose infrastructure components, such as containerization, virtualization, and cluster management, sufficiently secure and resilient for use in mission- and safety-critical edge AI applications. Safety and security, among other aspects of dependability, are explicitly called out in Google’s Responsible AI Practices, as is guidance about applying AI infrastructure, such as “Rule #5: Test the infrastructure independently from the machine learning.”

However, how do we attract, retain, educate, and train the people we need to responsibly use, operate, and manage advanced AI infrastructure?

Now, consider the addition of a specialized AI commercial-off-the-shelf tool to an AI infrastructure, obviating the need for AI practitioners to write code
to perform AI-specific version control for machine learning models. This tool that supports reproducibility in a collaborative end-to-end AI workflow can be a double-edged sword. Yes, the AI practitioner is freed up from writing code and managing the AI infrastructure to achieve this capability, and he or she will therefore have less opportunity to introduce coding errors that are exploitable security vulnerabilities. In addition, there may be fewer instances in which changes to the tool would require the user to modify his or her AI application. Nevertheless, making the inner workings of the AI-specific version-control functionality hidden from the user results in the AI practitioner having to trust to some degree the reliability of the tool, but this is unavoidable and part of supply chain risk management. As we add, modify, and remove this and other tools from AI toolchains, what are the emergent behaviors of the integrated infrastructure, and what impact do those behaviors have on our ability to trust in the claims about the dependability of the AI infrastructure? How do we construct AI infrastructure so that as it evolves, legacy production AI systems continue to function correctly? How will design choices regarding an AI infrastructure impact our ability to meet goals associated with explainable and accountable AI?

These and other RAI technical concerns, along with policy and legal concerns, such as governance and even international humanitarian law (e.g., in the case of attacks on Starlink as a military objective), need to be considered in a calculated manner, in concert with balancing short- and long-term goals for improving private and shared AI infrastructure. If you are working at the intersection of applying leading-edge AI infrastructure and ensuring that AI infrastructure is dependable, trustworthy, and responsible, then please consider submitting an article to IEEE Security & Privacy to share your experiences and best practices for the benefit of your fellow readers.

**REFERENCES**


**DISCLAIMER**

The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

**JAMES BRET MICHAEL**, Associate Editor in Chief
Don’t miss IEEE Quantum Week 2023—the IEEE International Conference on Quantum Computing and Engineering (QCE) bridging the gap between the science of quantum computing and the development of the industry surrounding it. In-person registration space is limited. Register today!

qce.quantum.ieee.org
PyOMP: Multithreaded Parallel Programming in Python

Timothy G. Mattson and Todd A. Anderson, Intel Corporation, Hillsboro, OR, 97124, USA
Giorgis Georgakoudis, Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA

Python is a widely used language in scientific computing. When the goal is high performance, however, Python lags far behind low-level languages such as C and Fortran. To support applications that stress performance, Python needs to access the full capabilities of modern CPUs. That means support for parallel multithreading. In this article, we describe PyOMP, a system that enables OpenMP in Python. Programmers write code in Python with OpenMP, Numba generates code that compiles to LLVM, and the resulting programs run with performance that approaches that from code written with C and OpenMP. In this article, we provide an update on the PyOMP project and explain how to install it and use it to write parallel multithreaded code in Python.

Python is number one! According to the PYPL and TIOBE indices, more people are programming with Python than any other programming language. With its vast ecosystem of modules to draw on, Python programming is highly productive. Unfortunately, Python programs are notoriously slow and extract only a small fraction of the performance available from a system.

Programmers in high-performance computing (HPC) are often told to prototype their ideas in Python, but when an efficient, production worthy code is needed, use a “real programming language” such as C or C++. We believe, however, that the time has come to embrace Python, not only for prototyping new algorithms or as a flexible workflow manager, but as a first-class language for implementing HPC applications.

Achieving high performance with Python can be challenging. We highlight two issues in this article. First, Python by default is interpreted. This prevents many of the compiler optimizations HPC programmers take for granted. We can work around this problem by writing kernels in C that are exposed through a Python wrapper. This is a great solution for common algorithms used over and over again. Programmers, however, often need new or customized algorithms. These would need to be written from scratch, ideally in Python. We address this problem with the Numba\textsuperscript{3} just-in-time (JIT) compiler. Numba maps Python code onto LLVM and can achieve performance on par with a traditional compiled language such as C. The JIT process adds overhead, but the result is cached so that subsequent calls do not incur the JIT overhead.

The second problem arises from how multithreading is handled in Python. In order to avoid the issues that make concurrency challenging (e.g., data races and thread safety), Python uses a Global Interpreter Lock (GIL) to assure that only one thread at a time makes forward progress. This makes Python programming safer, but it prevents the performance benefits that comes from running multiple threads in parallel on a multicore CPU.

The solution is to enable parallel multithreading in Python. We recently reported on our work\textsuperscript{1} to add explicit parallel multithreading to Python. This system maps the most popular interface for multithreading in HPC (OpenMP\textsuperscript{2}) onto Python. We used this with Numba and a version of LLVM that includes interfaces to OpenMP. This allowed us to run Python code at speeds on par with analogous programs written with C and OpenMP. We called this system PyOMP.

In this article, we describe PyOMP and how to install it on your own Linux system. We present results showing similar behavior between C and PyOMP code for the simple cases we have tested to date. We then close with

1521-9615 © 2021 IEEE
Digital Object Identifier 10.1109/MCSE.2021.3128806
Date of current version 16 December 2021.
PyOMP API

For over 25 years, OpenMP has been the primary way to write multithreaded code in HPC. Almost everyone working in HPC is at least familiar with OpenMP. In this section, we highlight the key features of PyOMP by considering a program ([see Figure 1(a)] that carries out a numerical integration, the result of which should approximate \( \pi \).

OpenMP is tied to the compiler. It is an explicit API where you tell the compiler what to do in order to generate a multithreaded program. In this case, the compiler is Numba. It transforms the Python code into LLVM and bypasses the GIL. In Figure 1, we import Numba and the specific features from PyOMP that we need in lines 1–3. The JIT compiler is applied to the function with the `@njit` decorator on line 4.

```python
1 from numba import njit
2 from numba import openmp
3 numba_context as openmp
4
5 @njit
6 def piFunc(NumSteps):
7     step = 1.0/NumSteps
8     sum = 0.0
9     with openmp("parallel private(x) shared(numThreads)"):
10         with openmp("single"):
11             numThreads = omp_get_num_threads()
12             with openmp("for reduction(+:sum)"):
13                 for i in range(NumSteps):
14                     x = (i+0.5)*step
15                     sum += 4.0/(1.0 + x*x)
16     pi = step*sum
17     return pi
18
19 pi = piFunc(100000000)
```

(a)

```bash
OMP_NUM_THREADS=8 python piLoop.py
conda install numba -c drtoodl3 -c conda-forge --override-channels
```

(b)

**Figure 1.** The \( \pi \) program. (a) The code, (b) how to run the program while requesting eight threads, and how to install PyOMP.

some thoughts on future work. Our goal with this article is to grow a community of programmers around PyOMP. We hope you will use it and join with us to help PyOMP evolve into a mainstream environment for parallel computing in Python.

In C, C++, and Fortran, the constructs from OpenMP are expressed through compiler directives such as (for C and C++):

```c
#pragma omp parallel for
```

Python lacks an analog to the pragma of C and C++. We could add directives as comments (as was done for OpenMP in Fortran) but that would not match the programming style of Python programmers (something they refer to as being "pythonic"). Working with experienced Python programmers, we found a pythonic way to insert OpenMP compiler directives into Python. We do this with the Python `with` statement. This statement defines the context under which the associated block of statements execute.

We see this in our \( \pi \) program on line 9, where we stipulate that the block of code attached to the `with` statement is to be executed in parallel by a team of threads. Each thread has its own copy of the variable \( x \) and they share a single copy of `numThreads`. Later in the program on line 13 we indicate that the iterations of the loop on lines 14–16 are to be divided among the
Table 1. Summary of the elements of openmp included in PyOMP.

<table>
<thead>
<tr>
<th>OpenMP construct, function, or clause</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>with openmp(&quot;parallel&quot;):</td>
<td>Create a team of threads that execute a parallel region</td>
</tr>
<tr>
<td>with openmp(&quot;for&quot;):</td>
<td>Used inside a parallel region to split up loop iterations among the threads.</td>
</tr>
<tr>
<td>with openmp(&quot;parallel for&quot;):</td>
<td>A combined construct equivalent to a parallel construct followed by a for.</td>
</tr>
<tr>
<td>with openmp(&quot;single&quot;):</td>
<td>One thread does the work while the others wait for it to finish</td>
</tr>
<tr>
<td>with openmp(&quot;task&quot;):</td>
<td>Create an explicit task for deferred execution of work within the construct.</td>
</tr>
<tr>
<td>with openmp(&quot;taskwait&quot;):</td>
<td>Wait for all tasks in the current task to complete.</td>
</tr>
<tr>
<td>with openmp(&quot;barrier&quot;):</td>
<td>All threads arrive at a barrier before any proceed.</td>
</tr>
<tr>
<td>with openmp(&quot;critical&quot;):</td>
<td>Only one thread at a time can execute the code inside a critical region</td>
</tr>
<tr>
<td>reduction(op: list)</td>
<td>A clause used with for to combine values with op across all the threads</td>
</tr>
<tr>
<td>schedule(static: chunk)</td>
<td>A clause used with for to control how loop iterations are scheduled onto threads</td>
</tr>
<tr>
<td>private(list)</td>
<td>A clause on parallel, for, or task. Create a private copy of the variables in the list.</td>
</tr>
<tr>
<td>firstprivate(list)</td>
<td>A clause on parallel, for, or task. Private but the variables equal original value.</td>
</tr>
<tr>
<td>shared(list)</td>
<td>A clause on parallel, for, or task. Make the variables in the list shared</td>
</tr>
<tr>
<td>default()</td>
<td>A clause that forces explicit definition of variables as private or shared.</td>
</tr>
<tr>
<td>omp_get_num_threads()</td>
<td>Return the number of threads in a team</td>
</tr>
<tr>
<td>omp_get_thread_num()</td>
<td>Return an ID ranging from 0 to the number of threads minus one</td>
</tr>
<tr>
<td>omp_set_num_threads(int)</td>
<td>Set the number of threads to request for subsequent parallel regions</td>
</tr>
<tr>
<td>omp_get_wtime()</td>
<td>Call before and after a block of code to measure the elapsed time.</td>
</tr>
<tr>
<td>OMP_NUM_THREADS = N</td>
<td>Set the default number of threads until reset by omp_set_num_threads()</td>
</tr>
</tbody>
</table>

Note: This includes constructs (using the Python with statement), clauses that modify constructs, functions from the OpenMP runtime library, and a single environment variable.

To the experienced OpenMP programmer, the meaning of PyOMP code should be straightforward since we use the same syntax/semantics as in the OpenMP standard. We include the most commonly used elements of OpenMP in PyOMP. This set is the well-known OpenMP Common Core shown in Table 1. We don’t have room for detailed explanations, but in the SciPy paper introducing PyOMP, a we use versions of the π program in Figure 1 to explain the OpenMP Common Core and the three fundamental design patterns of OpenMP applications.

1) **SPMD:** Each thread runs the same code. The thread ID and the number of threads are used to split up loops and manage data. This is the pattern most familiar to MPI programmers.

2) **Loop-level parallelism:** This is the pattern shown in Figure 1. Compute-intensive loops are found and modified so there are no loop carried dependencies (other than reductions) so the program is correct regardless of the order loop-iterations are executed. The loop iterations are then shared across the threads.

3) **Divide and conquer with tasks:** The function being parallelized is converted into a recursive function, which for each call splits the loop into two parts calling the function on each part. This continues until the size of the block of remaining iterations is small enough to compute directly.

**RUNNING PYOMP PROGRAMS**

To use PyOMP, you need to install our version of Numba that understands OpenMP. Internally, PyOMP’s Numba uses a fork of LLVM tracking the publicly available, community version of LLVM. We packaged these components together so you can easily install PyOMP as a conda package [see Figure 1(b)]. Currently, this conda package only supports Linux.

When we run a PyOMP program, we usually set an OpenMP environment variable that requests a given number of threads. We show this for eight threads in Figure 1(b). Requesting a specific number of threads is optional. A system supporting OpenMP has a default

---

*aThe SciPy paper* that defines PyOMP is available at http://conference.scipy.org/proceedings/scipy2021/tim_mattson.html

*bhttps://github.com/IntellLabs/numba.git*

*https://github.com/ggeorgakoudis/llvm-project/tree/pyomp*
The SciPy paper introducing PyOMP, we use versions well-known OpenMP Common Core shown in Table 1. The openMP elements of OpenMP are included in PyOMP. This set is the reduction in the same syntax/semantics as in the OpenMP program in Figure 1 to explain the OpenMP patterns. The code for the loop-level parallelism pattern is shown in Figure 1(a). For the other patterns and the analogous code written in C, the code is described in our SciPy paper. The results (in seconds) are provided in Table 2. We did not include the time for the JIT compilation since this compilation is only needed the first time a function is encountered and can be cached on disk between runs.

Notice that for these simple programs, C-OpenMP and PyOMP performance are roughly the same. Given that Python is typically orders of magnitude slower than C, this is an excellent result.

### CONCLUSION

PyOMP brings parallel multithreading with OpenMP into Python. This is a transparent, Open Source project with the system available as a conda package (at least for Linux).

There is much work to do. We showed simple benchmarks with the $\pi$ program where we matched the performance from analogous C programs. Those $\pi$ programs, however, do not benefit from vectorization and other optimization passes a C compiler provides. When we move to matrix multiplication where vectorization and memory management are key to performance, PyOMP runs at about one third the performance of the C code. This is still a dramatic improvement over native Python, but it does indicate that we have much to do before we reach the full potential of this technology.

Also, we currently only support the OpenMP Common Core. We have a number of additional constructs from OpenMP that we hope to add to PyOMP. For example, we are eager to move the directives used for programming GPUs into PyOMP. We do not see any technical hurdles preventing us from growing the scope of PyOMP to cover the full range of OpenMP capabilities and hope to do so soon.
At GTC 2022, Nvidia announced a new product family that aims to cover from small enterprise workloads through exascale high performance computing (HPC) and trillion-parameter AI models. This column highlights the most interesting features of their new Hopper graphical processing unit (GPU) and Grace central processing unit (CPU) computer chips and the Hopper product family. We also discuss some of the history behind Nvidia technologies and their most useful features for computational scientists, such as the Hopper DPX dynamic programming (DP) instruction set, increased number of SMs, and FP 8 tensor core availability. Also included are descriptions of the new Hopper Clustered SMs architecture and updated NVSwitch technologies that integrate their new ARM-based Grace CPU.

Systems with GPUs have impacted scientific computing since CUDA was first released in June 2007. The first CUDA tutorial was held at SC07 the same year. After just three years, Nvidia GPUs were featured as accelerators in several of the Top-10 of the top500.org list. Nvidia was, at that time, still viewed primarily as a company focused on the gaming market. Today, Nvidia is marketing itself as an AI technology company that continues to build products of interest to computational scientists.

In this article, we will highlight the most interesting features of the new Hopper and Grace computer chips and the Hopper product family (see Figure 1), just announced by Nvidia at their GTC 2022 event (Nvidia Hopper Architecture in-Depth—Nvidia Technical Blog) and detailed in the Hopper Whitepaper (Nvidia Hopper GPU Architecture). We will include some of the history behind the Nvidia technologies and their most useful features for computational scientists.

**CONNECTING THE CPU AND GPU—INTEL AND PCIE**

In early 2009, Intel sued Nvidia over their 2004 chipset licensing agreement that let Nvidia make core-logic (chipsets) for Intel in exchange for Intel licensing Nvidia’s 3-D, GPU, and other patents. Nvidia made the chips for their nForce series chipset and the two-chip ION Platform. ION’s combination of Nvidia and Intel chipsets provided a $10 billion advantage over Intel-only chipsets and was popular in Apple laptop computers, etc. Intel claimed their agreement did not cover their newer Nehalem architecture, which featured an integrated memory controller. Nvidia chose to countersue. The $1.5 billion settlement in January 2011 barred Nvidia from making CPUs with Intel’s x86 technology. Nvidia GPU chips have since been relegated to connecting to Intel CPUs via the relatively slow PCIe buses. This is also why, at present, IBM’s Power CPUs are connected to Nvidia’s GPUs over NVLINK, while no such connections exist for Intel CPUs. AMD have similarly connected their CPUs and GPUs via their HyperTransport technology.

**NVIDIA INTERCONNECTS: MELLANOX, NVLINK, AND NVSWITCH**

By announcing the new ARM-based Grace CPU at GTC2022 (available in 2023), Nvidia will again provide fast links between the CPU and GPU as they will be connected via their upgraded NVLink technology. Nvidia’s fourth-generation NVLink technology provides $1.5 \times$ higher bandwidth compared to the previous generation, and improved scalability for multi-GPU system configurations. A single Nvidia H100 Tensor Core GPU supports up to 18 NVLink connections for a total bandwidth of 900 gigabytes per second (GB/s)—over $7 \times$ the bandwidth of PCIe Gen 5.

Nvidia also announced in March 2019 that they had reached an agreement with Mellanox Technologies to buy the company for $6.9 billion (Nvidia press
The Hopper GPU, introduced as Nvidia H100 Tensor Core GPU, is implemented using the Taiwan Semiconductor (TMSC) 4N process customized for Nvidia with 80 billion transistors. The H100 architecture, that will also be part of the H100 SuperPOD (Figure 2), includes several noteworthy architectural advances.

The custom Nvidia H100 SXM5 module houses the H100 GPU and HBM3 RAM chips, and also provides connection to other systems via their fourth-generation NVLink and PCIe Gen 5 ports (see Figure 3). Note that these modules do not include display connectors, Nvidia RT Cores for ray tracing acceleration, or an NVENC encoder since they, like the A100, are data center modules.

The H100 GPU consists of up to 144 SM per full GPU, which have many performance and efficiency improvements over earlier versions.

An overview comparing the A100 and H100 architectures is shown in Table 1.

HOPPER FEATURES USEFUL FOR SCIENTIFIC COMPUTING

Key new features useful for Scientific computing include: (Nvidia Hopper GPU Architecture)

- 2× faster clock-for-clock performance per SM contributes significantly to 3× faster FP32 and FP64 instructions.
- Fourth-generation tensor cores, which are up to 6× faster chip-to-chip compared to A100, announced to deliver 2× the MMA (Matrix Multiply-Accumulate) computational rates of the A100 SM on equivalent data types, and 4× using the new FP8 data type, compared to old FP16 (Figure 4).
- New DPX instructions that should accelerate DP algorithms by up to 7× over the A100 GPU. A more detailed description of DPX will follow in the next section.
- New thread block cluster feature (Figure 5) allowing programmatic control of locality at a granularity larger than a single thread block on a single SM. Note that this adds another synchronization layer. This will also be discussed in the next section.
- New asynchronous execution features, including a new tensor memory accelerator (TMA). TMA is designed to transfer large data blocks efficiently between global and shared memory. TMA also supports asynchronous copies between thread blocks in a cluster. There is also a new

EMBEDDED, AI, AND CRYPTO

The week before the announcement of the Intel-Nvidia settlement, Nvidia shares rose almost 30% on the announcement of their embedded ARM-based Tegra 2 chips. Nvidia has since made leaps in the embedded market and targeted their more powerful GPUs for AI and machine learning (ML), to the point where they today market themselves heavily as an AI tech company that also produces consumer-grade GPUs for gaming and workstation workloads, as well as other embedded technologies.

Nvidia has profited greatly from their high-end GPUs being used for crypto mining, so much so that it has been hard the last couple of years for many gamers and scientists to obtained their newest most powerful consumer GPUs.

HOPPER GPU H100 OVERVIEW

The Hopper GPU product family with racked pods as background (extract from Nvidia source).

release on acquiring Mellanox for $6.9 Billion⁴). The Israeli-American company is known for its InfiniBand technology used between by high-end servers. This has led Nvidia to develop their interconnected technology further.

Their new NVLink Switch System can support clusters of up to 256 connected H100s and promises to deliver 9× higher bandwidth than InfiniBand HDR on Ampere. In addition, Nvidia is taking advantage of its Mellanox purchase with NVLink now supporting innetwork computing called SHARP, previously only available on InfiniBand. Nvidia states that it will deliver one exaFLOPS of FP8 compute performance while delivering 57.6 terabytes/s (TB/s) of All2All bandwidth.

⁴https://blogs.nvidia.com/blog/2022/03/22/nvidia-hopper-accelerates-dynamic-programming-using-dpx-instructions/
asynchronous transaction barrier for doing atomic data movement and synchronization.

- **HBM3** memory subsystem providing nearly a $2x$ bandwidth increase over the previous generation. The H100 SXM5 GPU is the world’s first GPU with HBM3 memory delivering a class-leading 3 TB/s of memory bandwidth.
- **50-MB L2 cache** (versus A100s 40 MB L2) reducing trips to HBM3.
- Second-generation Multiinstance GPU (MIG) technology provides approximately $3x$ more compute capacity and nearly $2x$ more memory bandwidth per GPU instance compared to A100.

**Tensor Memory Accelerator**

The newly added TMA enables asynchronous transfers of multidimensional blocks of data. An elected thread within a thread group takes on responsibility for interacting with the TMA by passing along a Copy Descriptor detailing the information the TMA needs to correctly transfer a multidimensional block of data, or tensor. The remaining threads are free to perform other instructions while the TMA operation is underway.

**Fourth-Generation Tensor Cores**

Fourth-generation tensor cores further improve upon the efficiency of the previous generation. Nvidia has now added support for an 8-bit floating-point data-type: FP8. They support two flavors of FP8, namely E4M3 and E5M2, enabling the choice between dynamic range or precision. The number following the E and the number following the M represent the number of exponent- and mantissa bits, respectively. Generic computations that natively match FP8 ranges are few and far between. In the cases where FP8 is sufficient, one can expect great performance improvements over, e.g., FP16. Nvidia expects its new DGX...
SuperPOD to be able to deliver 1 exaFLOPS of sparse FP8 compute.

**TABLE 1.** A100 versus H100—comparing main features (source: Nvidia).

<table>
<thead>
<tr>
<th>GPU features</th>
<th>A100</th>
<th>H100</th>
<th>H100</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU architecture</td>
<td>Ampere</td>
<td>Hopper</td>
<td>Hopper PCIe</td>
</tr>
<tr>
<td>GPU board form</td>
<td>SXM4</td>
<td>SXM5</td>
<td>PCIe Gen</td>
</tr>
<tr>
<td>factor</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>SMs</td>
<td>108</td>
<td>132</td>
<td>114</td>
</tr>
<tr>
<td>TPCs</td>
<td>54</td>
<td>62</td>
<td>57</td>
</tr>
<tr>
<td>FP32 cores/SM</td>
<td>64</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>FP32 cores/GPU</td>
<td>6912</td>
<td>16,896</td>
<td>14,592</td>
</tr>
<tr>
<td>FP64 cores/SM</td>
<td>32</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>(excl. tensor)</td>
<td>3456</td>
<td>8448</td>
<td>7296</td>
</tr>
<tr>
<td>INT32 cores/GPU (64 per SM)</td>
<td>6912</td>
<td>8448</td>
<td>7296</td>
</tr>
<tr>
<td>Tensor cores / GPU (4 per SM)</td>
<td>432</td>
<td>528</td>
<td>456</td>
</tr>
<tr>
<td>Peak FP64 TFLOPS (nontensor)</td>
<td>9.7</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>Texture units</td>
<td>432</td>
<td>528</td>
<td>456</td>
</tr>
<tr>
<td>Memory interface, all 5120 bit</td>
<td>HBM2</td>
<td>HBM3</td>
<td>HBM2e</td>
</tr>
<tr>
<td>Memory size</td>
<td>40 GB</td>
<td>80 GB</td>
<td>80 GB</td>
</tr>
<tr>
<td>Memory bandwidth (GB/s)</td>
<td>1555</td>
<td>3000</td>
<td>2000</td>
</tr>
<tr>
<td>L2 cache size</td>
<td>40 MB</td>
<td>50 MB</td>
<td>50 MB</td>
</tr>
<tr>
<td>Shared memory size/SM (config)</td>
<td>≤ 164 KB</td>
<td>≤ 228 KB</td>
<td>≤ 228 KB</td>
</tr>
<tr>
<td>Register file size/SM</td>
<td>256 KB</td>
<td>256 KB</td>
<td>256 KB</td>
</tr>
<tr>
<td>Total drawn power</td>
<td>400W</td>
<td>700W</td>
<td>350W</td>
</tr>
<tr>
<td>Transistors</td>
<td>54.2 billion</td>
<td>80 billion</td>
<td>80 billion</td>
</tr>
<tr>
<td>TSMC manufacturing process</td>
<td>7nm N7</td>
<td>4N</td>
<td>4N</td>
</tr>
</tbody>
</table>

**FIGURE 4.** Hopper floating point layouts (source: Nvidia).

**FIGURE 5.** H100 features clusters of thread blocks (source: Nvidia).

**NOVEL ARCHITECTURES**

**FIGURE 3.** A thread within a thread group takes on responsibility

**FIGURE 2.** Second-generation Multiinstance GPU (MIG) with 50-MB L2 cache and HBM3.

**Fourth-Generation Tensor Cores**

Fourth-generation tensor cores further improve upon the efficiency of the previous generation. Nvidia has undertaken a number of initiatives in the areas of precision and alignment. They support two flavors of FP8, namely FP8 compute and FP8 floating-point data. They include healthcare (e.g., genomics), robotics (e.g., path finding), quantum computing, and data science.

A simple example comes from the Fibonacci numbers. The nth Fibonacci number is known to be the sum of the two previous Fibonacci numbers. Finding the nth Fibonacci number is thus solved by recursively solving subproblems. Furthermore, subproblems of Fibonacci overlap. Other DP algorithms include Dijkstra’s shortest path, Floyd–Warshall all-pairs shortest path, and Smith–Waterman for sequence alignment.

DP problems benefit from the tabulation (building a solution bottom-up) and memoization (top-down) strategies. Both strategies store results of subproblems such that recomputation is avoided. The new DPX instruction set aims to speed up DP with specialized instructions that presumably exploit the characteristics of the DP problems.

Thread Block Clusters

The CUDA programming model has long encompassed threads, thread blocks, and grids. The Hopper architecture adds another level to the hierarchy: Thread Block Clusters. The new level in the hierarchy exists between Grids and Thread Blocks. Its functionality enables increased programmatic control of data locality.

Thread block clusters further the capabilities of, among others, the CUDA Cooperative Groups API. Thread Blocks participating in Thread Block Clusters are guaranteed to be scheduled concurrently, allowing for finer-grained parallelism across thread blocks.

*https://blogs.nvidia.com/blog/2022/03/22/nvidia-hopper-accelerates-dynamic-programming-using-dpx-instructions/
running on SMs. Going even further, Nvidia introduces a specialized SM-to-SM interconnect network, allowing SMs to exchange shared memory directly instead of through global memory.

Distributed Shared Memory (DSMEM)

DSMEM is Nvidia’s name for the new capabilities for sharing data between SM shared memories. The CUDA environment ensures a contiguous, virtual memory address space for the participating shared memories. Sharing data between SMs is thus done by simple pointer references. As mentioned earlier, the data exchanges between SMs are sped up by roughly 7× (Nvidia Hopper Architecture In-Depth—Nvidia Technical Blog).

Configurable Shared Memory

Like the A100, which has Compute Capability (CC) 8.0, the H100s CC 9.0 has the same parameters (e.g., warp size of 32, max warps/SM of 64, etc.) as the V100 (CC 7.0), except for the shared memory sizes per SM, which are configurable up to 96, 164, and 228 KB, respectively, for the V100, A100, and H100. Keeping data local to the SMs is thus made even easier.

Application Speedups Over A100

Nvidia highlighted four HPC applications when presenting potential speedups for their new Hopper architecture at GTC 2022: Climate modeling, genomics, lattice quantum chromodynamics, and 8000-point 3-D FFT. The FFT, in particular, was shown to perform about 30× faster over A100 on a multi-GPU H100 system where the GPUs were connected via an NVLink network. It will be interesting to see how our own applications perform on such systems.

Hopper Power Efficiency

The SXM5 form factor H100 is stated to have a thermal design power (TDP, the maximum heat dissipation a hardware component is designed to endure) of 700 W. This generated its fair share of discourse on social media with proponents and opponents of the seemingly high TDP. Opponents are discussing, e.g., direct operating costs from the power draw, cooling costs due to increased cooling needs, among other concerns. The TDP is relatively high, yet Nvidia states that the Hopper generation of GPUs is their most energy-efficient yet. How does one defend a TDP of 700 W for a GPU when previous generations have had TDPS around 3–400 W?

The SXM5 variant uses HBM3 memory modules. Utilizing the full bandwidth of HBM3 likely requires an increase in memory clock over the PCIe Gen 5 variant; the PCIe Gen 5 variant uses the lower bandwidth HBM2e modules. The SXM5 variant is also, in likelihood, going to have a faster boost clock than the PCIe Gen 5 variant. Increased memory and boost clock frequencies are, however, unlikely to be the factors pushing the TDP toward 700 W, due to power consumption scaling linearly with frequency under the same voltage conditions.

At the time of writing, it is challenging for us to estimate the mean power draw of the SXM5 variant in a general AI or HPC workload. Benchmarking the performance-to-power ratio of the H100 when power capped is another benchmark that could be interesting to investigate.

Note also that the Hopper architecture introduces additional capabilities for asynchrony. One of the significant benefits of asynchrony in this case is the potential for attaining high degrees of utilization. The increased asynchrony and potential for concurrency bodes well for latency hiding. The Hopper architecture might look bad on paper with a TDP of 700 W, but might look good when taking the performance-per-watt into account.

New AI and Security Features

AI techniques, especially those related to ML and deep learning, are increasingly useful for scientific computing as well. The H100 new Transformer Engine can accelerate Transformer model training and inference by dynamically choosing between FP8 and 16-bit calculations, which may deliver up to 9× faster AI training and up to 30× faster AI inference speedups on large language models compared to A100.

H100 also provides some features that enable safer multiuser environments, especially important in virtualized environments. The new Confidential Computing capability with MIG-level trusted execution environments (TEEs) supports up to seven individual GPU instances, each with dedicated NVDEC and NVJPEG units. Each instance now includes its own set of performance monitors that work with Nvidia developer tools. The H100 extends the TEE with CPUs at full PCIe line rate.

ARM and Grace

Announced for 2023, Nvidia’s ARM-based Grace CPU seems also noteworthy. USA’s Los Alamos National Laboratory and the Swiss National Computing Centre have already announced plans for Grace-based supercomputers (Nvidia’s Supercomputing CPU Puts Intel Under Pressure). ARM, originally a U.K. IP-only company, is now owned by the Japanese SoftBank Group Corp. Nvidia also tried buying ARM from them in a

---

1https://spectrum.ieee.org/nvidia-supercomputing-cpu-puts-intel-under-pressure
Dr. Grace Murray Hopper

Finally, it is nice to see computer chips named after a woman. Computer pioneer and Rear Admiral of the U.S. Navy Dr. Grace Brewster Murray Hopper (1906–1992) (Captain Grace M. Hopper Navy file, July 1981⁴) developed, among several other things, the first compiler called A-0, and also coined the computer term “bug” and “de-bug” after discovering an actual bug in a computer. She got her Masters and Ph.D. degrees from Yale University where she was the first woman to earn a Ph.D. in Mathematics. The Cray XE6 “Hopper” supercomputer at National Energy Research Scientific Computing Center is named after her (Five Fast Facts About Technologist Grace Hopper) as is Hopper Hall, the new center for Cyber Security Studies at the U.S. Naval Academy (USNA Hopper Hall Announcement).⁵

REFERENCES


ANNE C. ELSTER is a professor in HPC at the Norwegian University of Science and Technology, Trondheim, Norway, and also an associate editor of the Computing in Science & Engineering Novel Architectures Department. Contact her at elster@ntnu.no.

TOR A. HAUGDAHL is currently working toward his master’s degree in computer science specializing in autotuning for HPC at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. Contact him at thaugdahl@gmail.com.

Fall 2020 agreement, but the deal was terminated in February 2022 due to “regulatory challenges” in both U.S. and China.

Intel still dominates the top500.org list of the world’s largest supercomputers, but in 2021 Japan’s ARM-based Fugaku supercomputer (> seven million cores, running at 442 petaFLOPS) took the top spot.

The Grace CPU Superchip (Nvidia Grace CPU intro⁶) features two Grace cores connected via the NVLink-C2C technology thus providing up to 144 Arm v9 CPU cores. It claims to be the world’s first CPU using LPDDR5x memory with ECC and 1-TB/s total bandwidth. Its 900-GB/s coherent interface is 7× faster than PCIe Gen 5.

Nvidia’s Grace Hopper Superchip combines the Grace CPU and Hopper GPU architectures using Nvidia’s NVLink-C2C technology to deliver a coherent CPU+GPU memory model.

The system targets both HPC and AI applications and can provide a 30× higher aggregate system memory bandwidth to GPU compared to the DGX A100. Both the Grace and Grace Hopper superchips will run Nvidia’s software stacks, including Nvidia HPC, Nvidia AI, and Nvidia Omniverse.

FOOTNOTES

⁴https://www.history.navy.mil/content/dam/nhhc/research/histories/bios/HopperGrace/Hopper.pdf
⁵https://www.energy.gov/articles/five-fast-facts-about-technologist-grace-hopper
⁶https://www.usna.edu/NewsCenter/2020/10/NAVAL_ACADEMY_OPENS_NEW_CENTER_FOR_CYBER_SECURITY_STUDIES_HOPPER_HALL.php

ABOUT THE AUTHORS

We have not had the chance to test any H100 systems ourselves, but our recent BAT - Benchmark suite for Autotuners⁷ was benchmarked on both our 2-node IBM AC922 Power systems with NVLinked V100 GPUs and CPUs and our NVIDIA DGX2 system, with 16 tightly coupled NVSwitched GPUs, so we are looking forward to running it on a H100 system.
The Evolution of Nonfungible Tokens: Complexity and Novelty of NFT Use-Cases

Andrew Park and Jan Kietzmann, Peter B. Gustavson School of Business, University of Victoria, Victoria, BC, V8P 5C2, Canada
Leyland Pitt, Beedie School of Business, Simon Fraser University, Vancouver, BC, V5A 1S6, Canada
Amir Dabirian, College of Business and Economics, California State University, Fullerton, CA, 92831, USA

Nonfungible tokens (NFTs) have recently drawn considerable attention, highlighted by a digital art piece that sold for $69M USD in early 2021. Though they have only just started receiving coverage by traditional media outlets and interest from casual observers, the foundations of NFT technology date back to advances in computer science in the late 1970s. In this article, we examine the emergence of NFTs, from their technical origins, the introduction of blockchain technologies and the first token-based collectibles that led to modern day NFT products. We categorize the current use cases for NFTs, introduce their potential future applications, and highlight the challenges managers face in incorporating them into their existing workflows. By presenting our NFT adoption framework, we offer managers strategies for evaluating the risks and benefits of NFTs.

In the digital space, it is very hard to tell apart an original file from the countless copies of it that can be made with very little effort. Every time we attach a document we crafted to an email, for instance, we send a copy of our work. While not a new concept, protecting original digital work is only now becoming a practical reality through the adoption of nonfungible tokens (NFTs). In fact, NFTs entered the public conscience in early 2021, when artist Mike Winkelmann, also known as Beeple, sold a digital painting for $69M USD (Crow and Ostroff, more than twice the amount that Picasso’s Les femmes D’Alger fetched in 1997. Since then, NFT transaction activity has boomed, with over a third of the $600M USD lifetime trading volume of NFTs having occurred in the first and second quarters of 2021. What the casual NFT observer may not realize is that the technological underpinnings of NFTs date back to 2008, with Satoshi Nakamoto’s seminal technical whitepaper on blockchains. While Nakamoto proposed a new cashless system, based on a fungible digital token they called Bitcoin, the underlying immutable and decentralized computer network first conceptualized in the late 1970s has led to a proliferation of other use cases, including other cryptocurrencies and now, NFTs. NFTs themselves have been in existence since 2017, when a little-known Vancouver-based company called Axiom Zen created nonfungible digital pets known as Cryptokitties. The pets became so popular among their niche buyers that they almost ground one of the most popular blockchain networks to a halt. However, it was not until Beeple’s sale that NFTs began receiving broader consumer and media attention.

In this article, we provide a description of the historical rise of blockchain and NFTs and the current use cases of NFTs. We also discuss some of the proposed applications of NFTs and their future potential in transforming certain industries. In these proposed applications, we present some of the challenges facing their adoption and offer a management framework to determine how and when to prioritize them. We conclude with a note of caution around the current hype surrounding NFTs and how managers can derisk the technical and market uncertainties surrounding their implementation.
History of Blockchain Technology and the Rise of NFTs

Much like the disintermediation that takes place when cryptocurrencies can be exchanged without the involvement of traditional banks, the key benefit of NFTs is that they can be freely traded between buyers and sellers without processing the transaction through a central authority. While traditional asset transfers (e.g., art or real estate) rely on these intermediaries to mitigate the risk of fraud, NFTs can operate without such intermediaries because of their underlying blockchain technology.

Blockchains are a relatively new phenomenon. However, the core components that make up the technology have deep roots in computer science spanning back over 40 years. Chaum\(^1\) described a proposed computer network where a group of untrusting parties can use distributed ledgers (or “vaults,” as originally described) to keep active records of information, and as long as a majority of the members in the network agree on the information in the shared ledger. A malicious member who wishes to change the ledger will get outvoted and dismissed. This proposed system is combined with the concept of cryptographic keys, which ensures that one party cannot make transactions on behalf of another party without possession of those keys.

A little over ten years later, Haber and Stornetta\(^2\) introduced the concept of immutability to the work that had been done on distributed and trustless systems up to that point. By utilizing hashing, which compresses large volumes of text (such as transaction data) into unique and predictable strings of fixed length, and linking those hashes, it becomes inordinately difficult for a nefarious party to modify the content. Konst\(^3\) expanded on this idea, suggesting that hashing sets of text and chaining them would provide authentic, ordered, and complete entries for record keeping in numerous industries. By hashing and chaining sensitive text, auditable log files could be produced where sensitive information is not needlessly disclosed, but any member of a computer network could verify that the log was genuine and had not been tampered with. This concept of linking hashes is the foundation of what is known today as “blocks” and “chains” in the context of blockchain technology.

The seminal paper that gave rise to modern day blockchains was written by an anonymous cryptographer named Satoshi Nakamoto, who in 2008, proposed a peer-to-peer cashless digital currency, Bitcoin, using the aforementioned advances in computer science.\(^4\) The significant contribution made by Nakamoto, which enabled the advent of cryptocurrencies, is the incorporation of rate limiting in the network, so that dishonest sellers are unable to double-spend their existing assets. This so-called rate limiting is achieved by another computer science concept, proof-of-work, which was derived from previous advances known as Hashcash.\(^5\) Thus, it is evident that the innovation of blockchain is not the creation of a novel technology, rather it is the amalgamation of previously disparate computer science concepts.

Since the introduction of Bitcoin, blockchains have evolved both in technical complexity and in utility, most notable with the incorporation of smart contracts in the Ethereum blockchain.\(^6\) Blocks in the Bitcoin blockchain are comprised of static transaction data that contains information on the sender, the receiver, the amount of Bitcoin being transferred, and a referent to the previous block. Smart contracts in the Ethereum blockchain are small programs embedded within blocks that can be executed based on the specific conditions laid out in the program. This flexibility has enabled the creation of thousands of alternative cryptocurrencies (also known as altcoins) and the emergence of NFTs.

Emergence of NFTs

NFTs are related to traditional cryptocurrencies in that they are tokenized representations of an item of value. However, each individual NFT has its own unique characteristics and/or identifiers, allowing for differentiation that is not seen in traditional cryptocurrencies such as Bitcoin and Ether. Thus, NFTs mostly take the form of unique collectibles in a variety of settings such as digital art, collectibles, gaming, and metaverses rather than as mediums of exchange in and of themselves. Therefore, two NFTs within the same ecosystem (such as two different art pieces being sold by an NFT auction house) may sell at significantly different price points, where this phenomenon is not seen in traditional cryptocurrencies: one Bitcoin is worth one Bitcoin. Smart contracts allow for the attachment of unique identifiers to each NFT and given that Ethereum is the most popular smart contract-based blockchain,\(^7\) most NFTs are currently traded on the Ethereum blockchain.

While there were flashes of projects that bear some semblance to modern NFTs in the mid 2010s (e.g., the Counterparty platform allowed for the trading of Pepe memes in 2015), the first notable introduction of NFTs was a project called Cryptokitties by Vancouver-based software company, Axiom Zen. Cryptokitties allowed for the creation and trading of digital cats, one of which fetched $172,000 USD.\(^8\) Even though Cryptokitties became immensely popular within the cryptocurrency community and somewhat within the digital gaming ecosystem, NFTs still did not enter the mainstream until early 2021. A timeline of the key events that led to the current state of NFTs is presented in Figure 1. In the next section, we discuss the current use cases of NFTs.
Current NFT Use Cases

Due to the novelty of the NFT ecosystem, the breadth of use cases is rather limited, centered around applications that digitally represent items of scarcity. NFT products that have notable levels of adoption can be organized into three main categories: art and collectibles, games and metaverses, and utilities and DeFi. Even within these categories, adoption is heavily skewed towards a handful of NFT products, with most exhibiting inconsistent and low levels of transaction activity. Given this, NFTs are still firmly placed within the “Early Adopters” phase of the Diffusion of Innovations process (see Figure 2), drawing mainly the attention of technology enthusiasts. In contrast, cryptocurrencies, in particular Bitcoin, have likely moved into the “Early Majority” category as evidenced by its inclusion in investment portfolios of numerous well-known individual and institutional investors, and large-cap companies such as Stanley Druckenmiller, Ark Investment Management, and Tesla. This section describes each of the three NFT categories in further detail.

Art and Collectibles

NFTs have provided a novel solution to a persistent issue within the digital art ecosystem: how does an artist maintain scarcity of a digital art piece when it can be so easily copied and distributed? In the material world, verifiers and authenticators can examine art pieces to determine if they are reproductions. While this process is not immune to forgery, it provides enough confidence to buyers that they are willing to transact on expensive art pieces.

In the NFT ecosystem, if a participant claims to own a digital art piece, any third party can examine the history of transactions on the blockchain to see if in fact the participant’s cryptographic key is associated with the piece. If the rightful owner decides to sell it again, a new transaction record will be imprinted on the blockchain so any third party can examine the chain of custody of that NFT. Therefore, although the NFT can be copied, as long as there is consensus within the blockchain network of the identifying hash of that NFT, anyone can verify its ownership. CryptoPunks, created by Larva Labs, is one of the most popular NFT art products; it is a series of 10,000 digital art pieces that are traded within the Ethereum blockchain.

Other NFT collectibles are often similar to art, but they do not necessarily need to be creative works. They can be anything that represents a scarce digital asset. Perhaps the best-known example of a collectible is the first tweet Jack Dorsey, Co-Founder and CEO of Twitter ever sent. Dorsey tokenized his tweet and sold it on a commonly traded NFT marketplace for nearly $3M USD. Corporations are now recognizing NFT collectibles as additional sources of revenue. The Toronto Raptors of the NBA recently released six NFTs that can be used to claim “exclusive perks, from first-look access, signed game-worn memorabilia and curated VIP experiences.”

Gaming and Metaverses

Blockchain-based game developers are attempting to import the concept of in-game collectibles and microtransactions through NFTs. A central authority, such as a traditional game developer that issues and keeps track of ownership of in-game assets is vulnerable to hacking and asset loss, especially in the event that the developer decides to stop supporting the game. NFT games allow for the persistence of these assets due to their permanent records on the blockchain. Even if the NFT game developer were to drop support for the game, the records of ownership of the assets within the game would be preserved, allowing for another developer to create a new game, to which the original participants would be able to seamlessly port their assets and privileges. Axie Infinity is an example of an NFT game that is based on tradable in-game items.

Metaverses are similar to NFT games but the impetus for participation is not necessarily competitive. Much like popular nonblockchain-based virtual worlds such as Sim City, Fortnite, and Second Life, NFT metaverses like Decentraland allow for peer-to-peer interactions where participants can interact, build communities, and trade digital assets. A virtual plot of land in Decentraland recently sold for $900,000 USD.

Utilities and DeFi

The third category of NFTs are comprised of products and services outside of gaming and collectibles that
attempt to replace traditional professional services. For example, the Ethereum Name Services is aiming to replace the way domain names are traditionally sold and maintained (through individual vendors and a central internet authority) by decentralizing the record of address ownership across the entire blockchain network. Other types of utility NFTs include access to real-world privileges; for example, a social media influencer may sell NFTs for a “shout-out” on one of their social media or YouTube posts. The aforementioned Toronto Raptors NFT is actually a hybrid of a collectible and a utility NFT.

DeFi, or Decentralized Finance, is a general term to describe any project that aims to reconfigure traditional finance workflows, whether they be related to banking, credit cards, mortgages, or money transfers. For example, using smart contracts on the blockchain, a lender and borrower can automate interest payments and a cryptocurrency can automatically leave an escrow account at predetermined intervals, eliminating the need for manual money transfers and arbitration in case of a dispute. While DeFi is still in its nascent projects such as JustLiquidity and BakerySwap are gaining more attention. These two projects allow NFT holders to earn interest when they keep their tokens in the community asset pool, which provides a secondary market for trading.

Future NFT Use Cases
Admittedly, the number of existing use cases for NFTs is still rather limited. However, it is also very promising. This section reviews the ambitious use cases for NFTs that have not yet reached an appreciable number of users. Following each example, we identify some challenges facing the NFT ecosystem that must be overcome before these envisaged use cases reach critical mass. Figure 3 depicts an adaptation of Lakhani and Lansiti’s\(^9\) framework of adoption of foundation technologies and maps the future NFT use cases described in this section. On the vertical axis, we categorize use cases by their novelty, to describe how much effort will be required to educate users about the problems NFT solves. The horizontal axis refers to how complex required coordination within the ecosystem will be to produce value with the NFT.

Expanded DeFi
While there are some nascent products in this category, DeFi is still very much in its infancy and its purported ability to disrupt existing financial institutions is far from being realized. Speculators have posited that NFTs could be created for individual mortgages, and then these NFTs could be traded on secondary markets. In theory, this new workflow would eliminate the need for centralized parties to issue and securitize mortgages. However, it would require massive coordination between securities and exchange commissions, other federal and local regulatory bodies, financial institutions, customers, and mortgage issuers.

---

**FIGURE 2.** NFTs and cryptocurrencies in the diffusion of innovations. Adapted from Singlemann.\(^{18}\)

**FIGURE 3.** NFT adoption framework. Adapted from Lakhani and Lansiti.\(^9\)
We draw from Lakhani and Iansiti’s work detailing the challenges in implementing foundational technologies. Ambitious DeFi applications are both highly novel and require significant coordination across a broad spectrum of parties, thus their mass adoption faces significant headwinds, as do similar transformative technologies.

Dematerializing Real World Assets
There is significant hype surrounding the efficiency gains that could be realized by removing central authorities. Title companies that keep track of valuable physical goods such as automobiles and land are one such example of such disintermediation. In theory, any physical asset such as a land parcel could be represented by a hash on the blockchain, allowing for a transparent, public record of ownership of the asset. This would allow sellers and buyers to transact directly without waiting for the results of a query from a land title office. Consequently, transaction speeds should increase, and transaction costs should decrease. In fact, this idea of tracking land ownership is already being piloted in Honduras, due to chronic problems caused by traditional manual land registry practices, including duplicate titles, unauthorized changes, and bribery. Should this pilot prove successful, it is likely that records management authorities in other jurisdictions will consider using an NFT-based system, especially if the incumbent method of tracking ownership is flawed. Since governments can relatively easily migrate to an NFT-based registry, with little need for approval from other parties, the degree of novelty is high, and the complexity of coordination is low, placing this use case in the “localization” quadrant in Figure 3.

There remains a significant challenge with dematerializing physical assets, particularly if the item is small and portable. A land parcel is physically verifiable, as it is tied to a specific physical address. Smaller, mobile items such as jewelry or even automobiles are much more technically challenging. A piece of jewelry can, in good faith, be represented as an NFT by hashing an image or the dimensions of the piece. However, a nefarious party could take a similar image or piece of jewelry, create a new NFT, and claim it is the original. It is unclear what technological advances are required to create a tighter link between a mobile physical good and a digital representation of it. Thus, at this stage, any attempt to replace paper certificates of authenticity with NFT-based certifications will only yield marginal efficiency gains, if any at all. However, certificate authorities, buyers, sellers, and vendors would all need to be on board with such a dramatic infrastructure change. Therefore, this particular use case would be an example of “substitution” in Figure 3.

Supply Chain Management
For complex supply chain management solutions, the same headwinds facing broad adoption exist as in the preceding paragraph, due to the technical challenge in dematerializing mobile, physical goods. However, if a simple, tightly integrated supply chain already exists between a goods producer and buyer, and they wish to migrate away from a paper-based chain of custody to a secure digital records management system, a simple permissioned blockchain is easy to implement. Each good being transferred can be represented as an NFT on the blockchain. This would allow for the parties to migrate to a more efficient records management system and ensure each time the good is transferred, the transfer is recorded immutably in a shared database. Given the relatively obvious benefit of this use case, and assuming the supply chain is small and therefore the complexity of coordination is low, this would be an example of a “Single Use” NFT solution, as seen in Figure 3.

These use-cases, both current and potential, show the type of promise NFTs have. More use-cases will likely emerge in the near future and we hope that the examples provided here will motivate managers and firms to start considering the transformative potential of NFTs. However, caution is warranted, as NFTs face significant obstacles in their path to mass adoption. In fact, in the short time since Beeple’s $69M USD digital art sale, NFT transaction activity has already slowed considerably. However, such a development is very common for new technologies. As Gartner and its Hype Cycle reminds us, new technologies become visible through a “technology trigger,” such as the very public attention drawn to the sale of Beeple’s art in 2021. Shortly after, the public’s anticipation of the new technology sharply increases, until it hits a “Peak of Inflated Expectations,” after which people realize that the new technology is not quite ready yet for wider adoption. Expectations fall into a “Through of Disillusionment.” This, we believe, is where NFTs are currently positioned. Soon, we predict, future developments will occur that address technical, legal, security, and market uncertainties. These will fuel NFTs towards a “Slope of Enlightenment,” where we will see many more use-cases being adopted as firms prepare for NFTs as part of their “Plateau of Productivity,” whether for single-use, localization, substitution, or transformation purposes. While it is unclear how long this will take, we predict that the future of NFTs will be very exciting. Eventually.

REFERENCES

ANDREW PARK is the incoming Assistant Professor of Management Information Systems, University of Victoria, Victoria, BC, Canada. His research interests include information systems, and innovation and technology management. Contact him at aparik@uvic.ca.

JAN KIETZMANN researches Management Information Systems and Innovation at University of Victoria, Victoria, BC, Canada. He is also an Associate Editor with Business Horizons. His research focuses on organizational and social perspectives related to emerging technologies. He received the Ph.D. degree in innovation and information systems from London School of Economics, London, U.K. Contact him at jkietzma@uvic.ca.

LEYLAND PITT is the Dennis F. Culver EMBA Alumni Chair of Business with the Beedie School of Business, Simon Fraser University, Vancouver, BC, Canada. He has authored or coauthored many journals such as Journal of the Association for Information Systems, Journal of Strategic Information Systems, MIS Quarterly Executive, Information Systems Research, Journal of Information Technology, California Management Review, Sloan Management Review, and MIS Quarterly (which he also served as Associate Editor). He is currently an Associate Editor of Business Horizons and the Journal of Advertising Research, and Editor of the Journal of Wine Research. Contact him at Leyland_Pitt@sfu.ca.

AMIR DABIRIAN is currently the Vice President of the Division of Information Technology and a Faculty Member with the Department of Marketing, California State University, Fullerton, CA, USA. He received the Ph.D. degree from the Department of Industrial Economics and Management, KTH Royal Institute of Technology, Stockholm, Sweden, with a research interest in employer branding. Contact him at amir@fullerton.edu.
Elliptic Curve Pairings

Joshua Brian Fitzgerald, Heliax AG

Elliptic curve pairings are a powerful tool and a popular way to construct zero-knowledge proofs, which are beginning to be used in blockchains as a way to provide privacy in the transaction ledger.

A recent trend in cryptography is to leverage zero-knowledge proof systems that allow one party to convince another that they have information that satisfies certain conditions without revealing the information itself. This technology is particularly useful when multiple parties who may not trust each other nevertheless need to coordinate to produce some useful information. For example, the Zcash cryptocurrency uses zero-knowledge cryptography to hide the amounts and recipients in Zcash transactions while allowing an observer to verify the resulting balances have been adjusted correctly.

The heart of many of these zero-knowledge proof systems is the elliptic curve pairing, a useful function that can test if two points on an elliptic curve are related in the same way as another pair of points.

ELLIPITIC CURVES

An elliptic curve is a polynomial that can be written as $y^2 = x^3 + ax + b$. If the $y^2$ on the left side were simply $y$, then this would be a typical cubic polynomial, familiar from a high school algebra class. Squaring the $y$ on the left side turns this cubic into an elliptic curve with a pleasing symmetry across the x-axis.

Amazingly, the points on an elliptic curve can be used to define an algebraic group structure. Draw a straight line diagonally across an elliptic curve and, being cubic, it will intersect the line at up to three points. These intersection points “add up to zero” in the elliptic curve group structure. If $P=(x_1, y_1)$, $Q=(x_2, y_2)$, and $R=(x_3, y_3)$, then having a straight line $y = mx + b$ passing through these three points tells us that $P + Q + R = 0$ in the elliptic curve group. This can be rearranged as $P + Q = -R$, which gives us a formula for the sum of two points on an elliptic curve and a geometric method for calculating it. To add two points: draw a line through them and look for the third intersection point. The inverse of this point is the desired sum.

The identity element, or zero, of the elliptic curve group is an abstract point at infinity. Every line can be thought of as passing through this point as the line heads off the page and out to infinity. (The point at infinity can be made explicit using a projective coordinate system where it becomes a regular point with integer coordinates.)

The inverse of an elliptic curve point is its reflection across the x-axis. The line through these two points will be vertical and not intersect the curve proper in any other place. If $P = (x_1, y_1)$, then the vertical line $x = x_1$ passing through $P$ will also intersect the curve at $Q = (x_1, -y_1)$. The equation representing these intersections is $P + Q = 0$, which can be rearranged to show that $Q = -P$. Thus, the additive inverse of a point is its reflection across the x-axis.

Using some calculus, we can find tangent lines on an elliptic curve. A line that is tangent to a point $P$ on the elliptic curve can be thought of as intersecting the curve twice at that point. There will be a third intersection point, $Q$, showing that $P + P + Q = 0$ or, equivalently, $Q$, showing that $2P = -Q$.

Therefore, with some algebra and a little calculus, we can add two different points or double any single point on an elliptic curve. This gives us an efficient way to find anyth multiple of any point on the curve using a double-and-add style algorithm requiring $Q$, showing that $O (\log n)$ steps.

ELLIPITIC CURVES IN CRYPTOGRAPHY

An elliptic curve is often introduced using the real numbers $\mathbb{R}$ as the base field for which $x$ and $y$ are members. But an elliptic curve can be constructed over any field, including finite fields, such as the integers modulo a prime number. When constructed over a finite
field, an elliptic curve group will have a finite number of elements, and its group structure will be either cyclic or a product of two cyclic groups.²

Over a finite field, each point on an elliptic curve has finite order, so for any point $P$ there is some whole number $r$ such that $rP = 0$. All the points $P$ such that $rP = 0$ form a subgroup of the elliptic curve group called the torsion subgroup.

Now we can see how elliptic curve groups can be used in cryptography. Suppose you have a secret number $r$. Take some publicly known point $P$ and compute $nP$ using the efficient algorithm mentioned earlier. The point $nP$ is just a coordinate pair $(x, y)$ and can be published freely. Someone else who knows both $P$ and $nP$ can discover $n$ by naively multiplying $P$ by itself repeatedly until reaching $nP$, but this will take ages if $n$ is large. Faster algorithms for discovering $n$ exist, but all are exponential in the number of digits of $n$. This is the elliptic curve discrete logarithm problem (ECDLP) that is used as the basis for elliptic curve cryptography.

Elliptic curve cryptography is fast and uses smaller keys than other cryptosystems, such as Rivest-Shamir-Adleman. For these reasons, it is the primary public-key cryptography used to secure Internet communication today. It is the reason I can send my credit card number or email password over the public Wi-Fi at a coffee shop, and no one else in the vicinity, including the coffee shop owner or their Internet service provider, can snoop on it and see my information.

PAIRINGS ON ELLIPTIC CURVES

The elliptic curve group used as the basis for cryptography is not a general algebraic group. Elliptic-curve-based groups have additional structure and particularities that are specific to elliptic curves. This means that elliptic curves could have specialized attacks that exploit their extra structure to make the ECDLP easier to break.

One such structure is the elliptic curve pairing. An elliptic curve pairing is a nondegenerate bilinear map $e: E[r] \times E[r] \to \mu_r$, where $E[r]$ is the torsion subgroup of an elliptic curve group, and $\mu_r$ is the multiplicative group of $r$th roots of unity in the field $F_q^r$ ($q$ is the characteristic of the field over which the elliptic curve is constructed). The exponent $k$ is called the embedding degree and is an important security parameter for pairings.

Bilinearity means that $e(nP, mQ)_r = e(P, mQ)_r^n = e(nP, Q)_r^m = e(P, Q)_r^{nm}$, where the $r$ subscript means “in $\mu_r$.” Nondegeneracy means that $e(P, Q)_r \neq 1$ as long as neither $P$ nor $Q$ is zero (the point at infinity).

The existence of a pairing on an elliptic curve means that one can check to see if two pairs of points are related in the same way. Suppose we have a secret number $n$, and we also have four elliptic curve points: $P, nP, Q,$ and $nQ$. Using a pairing, we can test to see if $nP$ is the same multiple of $P$ as $nQ$ is of $Q$ by checking the following pairing equation:

$$e(P, nQ)_r = e(nP, Q)_r.$$

Checking this equation involves computing the resulting root of unity on each side and comparing the results. At no point do we discover $n$, but we can tell that both $nP$ and $nQ$ are $n$th multiples of $P$ and $Q$. This leaks a little bit of information about these points that somewhat weakens security.

This difference in security can be formalized using computational hardness assumptions. Some relevant hardness assumptions for elliptic curve pairings are the computational Diffie-Helman (CDH) problem and the decisional Diffie-Helman (DDH) problem as well as the discrete logarithm problem discussed earlier.³ The CDH problem asks whether it is feasible to compute $nmP$ only given $P, nP$, and $mP$. The CDH problem is infeasible for carefully chosen elliptic curves. (Notice that the CDH problem would be feasible if the discrete logarithm were also feasible to compute.)

The DDH problem doesn’t ask that $nmP$ be computed, only that it be recognizable. That is, given only
$P$, $nP$, $mP$, and $Q$, decide whether $Q$ is actually equal to $nPmP$. This is easy to do with a pairing by checking this pairing equation:

$$e(nP, mP)_r = e(P, Q)_r.$$

**COMPUTING A PAIRING**

A few different definitions of pairings exist, but the most commonly used pairings are based on Tate’s definition, which allows several optimizations. Tate’s pairing is defined as follows:

$$e(P, Q) = f_{r, P}(Q)^{q^k - 1}. $$

In this formula, $r$ (on the right-hand side) is the order of an $r$-torsion group to which $P$ must belong, $q$ is the characteristic of the field over which the elliptic curve is constructed, and $k$ is the embedding degree discussed in the section “Pairings on Elliptic Curves.”

Computing the pairing has two major phases: the Miller loop, where function $f_{r, P}(Q)$ is computed using a double-and-add style algorithm, and then the final exponentiation, in which $f_{r, P}(Q)$ is taken to the power of $(q^k - 1)/r$.

**MILLER’S ALGORITHM**

Function $f_{r, P}$ is a rational function in $x$ and $y$ with a zero of multiplicity $r$ at $P$ (meaning $rP = 0$) and a pole of multiplicity $r$ at the point at infinity. This function can be constructed iteratively out of linear factors. At each point in this process, we need to know how many zeros and poles our function has and at which places. **A divisor** is the mathematical construction that takes care of this bookkeeping for us. Divisors are out of the scope of this article but are implicitly used in this section for counting multiplicities of zeros and poles.

Suppose, for some $n$, the line $ax + by + c = 0$ passes through points $P$ and $nP$. Because of the group law on elliptic curves, this same line will also pass through $-nP$. This line will have a zero of order 1 at each of $P$, $nP$, and $-nP$ as well as a pole of order 3 at the point at infinity. We will denote this line through $P$ and $nP$ as $v_{nP}$.

Suppose we also find the vertical line $x + d = 0$ passing through $nP$ and $-nP$. This line has zeros of order 1 at both $nP$ and $-nP$ and a pole of order 2 at the point at infinity. We will denote the vertical line through $nP$ as $v_{nP}$.

Dividing these gives us a rational function

$$g_{n, P}(x, y) = \frac{1}{v_{nP}} \frac{ax + by + c}{x + d},$$

which will have zeros of order 1 at $P$ and $nP$ but poles of order 1 at $(n + 1)P$ and the point at infinity. We can compute these functions for each $n$ from 1 to $r$. When multiplied together, these functions have a telescoping-like behavior, where the linear factor in the numerator of $g_{n, P}$ creating a zero at $nP$ cancels out the linear factor in the denominator of the previous function $g_{n-1, P}$ that had created a pole at $nP$.

The target function can be found by taking the product of these rational functions like so:

$$f_{r, P}(x, y) = \prod_{i=1}^{r} g_{i, P}(x, y),$$

where $g_{i, P}(x, y)$ is the rational function whose numerator is a line passing through $P$ and $iP$, and whose denominator is a vertical line through $iP$, as described previously.

This suggests an algorithm for computing $f_{r, P}(Q)$ by successively evaluating $g_{i, P}$ at $Q$ and accumulating the results by multiplying the current state by $g_{i, P}(Q)$ for each $i$ up to $r$. This algorithm’s runtime complexity is $O(r)$, and so it is not suitable for our purposes where large $r$ will be used.

However, we can make this into an $O(\log r)$ algorithm by using a double-and-add style approach. The preceding product formula suggests this iterative formula for computing $f_{m+1, P}$ from $f_{m, P}$:

$$f_{m+1, P} = f_{m, P} \frac{l_{nP}}{v_{nP}},$$

Similarly, we can also get a crucial formula for $f_{2m, P}$ from $f_{m, P}$:

$$f_{2m, P} = (f_{m, P})^2 \cdot \frac{l_{nP}}{v_{2nP}}.$$ 

Combining these gives us a double-and-add method for computing $f_{r, P}$ that runs in $O(\log r)$ time.

**THE FINAL EXPONENTIATION**

The only remaining piece of the pairing formula left to compute is the final exponentiation, in which the result of $f_{r, P}(Q)$ is taken to the power $(q^k - 1)/r$. In practice, the characteristic of the field $q$ is very large, say
256 bits, and the embedding degree \( k \) is greater than 1. (Two commonly used elliptic curve families, BarretoNaehrig (BN) and BarretoLynnScott, use \( k = 12 \) and \( k = 24 \), respectively.) A greater \( k \) increases security but makes the final exponent much greater and harder to compute. To make things even worse, a large \( k \) also means the field elements we are working with are members of a \( k \)th-degree extension field, where multiplication is more difficult than in the base field. “Pairing-friendly” curves have an embedding degree that is large enough to provide security but not too large that efficiency is compromised.

Thankfully, a few optimizations can help. For BN elliptic curves, the embedding degree will always be 12. Then, \( q^{12} - 1 \)/\( r \) can be factored as \( (q^6 - 1)(q^2 - 1)(q^4 - q^2 + 1)/r \). The two binomial factors can be computed easily with the Frobenius operation, leaving the remaining \((q^4 - q^2 + 1)/r \) part with a much lower exponent.

Since the exponent \((q^k - 1)/r \) is based on the parameters of an elliptic curve and known well in advance, precomputation can speed up the final exponentiation process.\(^4\)

Furthermore, occasionally multiple pairing values are computed and multiplied together according to some formula. The Miller loop portion of each pairing can be computed first, these results multiplied together using the formula, and then the final exponentiation can be applied once at the very end of the process.

Each transaction in this new cryptocurrency is public, so the amounts being transferred are viewable, and all account balances can be easily computed from the ledger.

However, using an elliptic curve pairing, we can transform this simple public ledger into a private one, where all amounts are encrypted, and balances cannot be computed from the information in the ledger. (The simple protocol that follows was chosen to illustrate the usefulness of an elliptic curve pairing in a blockchain environment, not for its security. This protocol suffers from a few security flaws that would need to be carefully addressed before implementing it safely.)

Our new private cryptocurrency will use elliptic curve cryptography to hide the amount of each transaction. The designers of this private protocol will choose a pairing-friendly elliptic curve \( E \) and a point \( P \) on the curve that generates a large prime-order subgroup. Then, when the sender wants to send the amount \( a \) in a transaction, it includes \( aP \) in the amount field rather than \( a \). Now, the amount of the transaction, \( a \), cannot be computed from the information in the transaction because of the ECDLP. Since elliptic curve arithmetic respects addition, an encrypted version of the balance, \( bP \), can be computed by adding all of the encrypted amounts together just as before.

But how can we check that the sender has enough funds in the account to complete the transaction? Because of the ECDLP, no validator can check directly that the sender’s balance exceeds the transaction amount. But if the sender includes some extra information, it ought to be able to convince a validator that the balance is high enough.

The sender chooses some point \( Q \) on the curve which it keeps to itself. It then computes \( b' = b - a \), the new balance after the transaction is complete, and computes its bit decomposition \((b'_0, b'_1, ..., b'_n)\). The sender includes \((b'_0Q, b'_1Q, ..., b'_nQ)\) in the transaction data.

From this encrypted bit decomposition, a validator can compute \( b'Q \) using the equation:

\[
b'Q = b'_0Q + 2b'_1Q + 4b'_2Q + ... + 2^n b'_nQ.
\]

By providing a bit decomposition of \( b' \), the sender can show that \( b' \) is positive by the simple fact that there aren’t enough bits to overflow the order of the elliptic curve subgroup we are working in.
Next, the validator chooses a random number $k$ and sends it to the sender. The sender computes $kQ$ and sends this back to the validator. The validator uses this elliptic curve pairing equation to check that the two versions of the balance are consistent with each other:

$$e(kP, b'Q) = e(bP, kQ).$$

Because of the bilinearity of the pairing, this equation shows that $b'k = bk$ in the order-$r$ target group for the pairing. With no foreknowledge of $k$, it is extremely unlikely that the sender can choose a bit decomposition that satisfies the pairing equation that is not the true decomposition.

Now that we’ve verified that the sender has enough to complete the transaction, the transaction data $T = (S, R, aP, (b_1Q, b_2Q, ..., b_nQ), k, kQ)$ are appended to the ledger, and the transaction is complete.

This protocol is easily checkable with pairing equations because of the highly arithmetical nature of the statement we are checking. Elliptic curve pairings can be used to check more complicated statements than this, however. Using a much more complex construction called azk-SNARK, a few pairing equations can check any statement in NP.

One major flaw in the protocol described previously is that the sender’s and recipient’s addresses are public, and anyone reading the ledger can see that the two parties are transacting, even if they can’t tell how much is being transferred. In Zcash, the sender and recipient addresses are also hidden, and a zk-SNARK is used to verify the addresses and link the amounts to the correct accounts without revealing the addresses.

**REFERENCES**


**JOSHUA BRIAN FITZGERALD** is a cryptographer and protocol developer for Heliax AG, Zug 6300, Switzerland. Contact him atjoshuabfitzgerald@gmail.com.
IEEE Computer Society conferences are valuable forums for learning on broad and dynamically shifting topics from within the computing profession. With over 200 conferences featuring leading experts and thought leaders, we have an event that is right for you. Questions? Contact conferences@computer.org.

### JUNE

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>07 June</td>
<td>CAI (IEEE Conf. on Artificial Intelligence)</td>
<td>Santa Clara, USA</td>
</tr>
<tr>
<td>10 June</td>
<td>ICHI (IEEE Int’l Conf. on Healthcare Informatics)</td>
<td>Houston, USA</td>
</tr>
<tr>
<td>12 June</td>
<td>WoWMoM (IEEE Int’l Symposium on a World of Wireless, Mobile and Multimedia Networks)</td>
<td>Boston, USA</td>
</tr>
<tr>
<td>17 June</td>
<td>CVPR (IEEE/CVF Conf. on Computer Vision and Pattern Recognition)</td>
<td>Virtual Conf.</td>
</tr>
<tr>
<td>21 June</td>
<td>CHASE (IEEE/ACM Conf. on Connected Health: Applications, Systems and Eng Technoloyies)</td>
<td>Orlando, USA</td>
</tr>
<tr>
<td>22 June</td>
<td>CBMS (IEEE Int’l Symposium on Computer-Based Medical Systems)</td>
<td>L’Aquila, Italy</td>
</tr>
<tr>
<td>23 June</td>
<td>ICIS (IEEE/ACIS Int’l Conf. on Computer and Information Science)</td>
<td>Wuxi, China</td>
</tr>
<tr>
<td>24 June</td>
<td>LICS (ACM/IEEE Symposium on Logic in Computer Science)</td>
<td>Boston, USA</td>
</tr>
<tr>
<td>26 June</td>
<td>BSC (Biennial Symposium on Communications)</td>
<td>Montreal, Canada</td>
</tr>
<tr>
<td></td>
<td>COMPSAC (IEEE Computers, Software, and Applications Conf.)</td>
<td>Torino, Italy</td>
</tr>
<tr>
<td></td>
<td>JCDL (ACM/IEEE Joint Conf. on Digital Libraries)</td>
<td>Santa Fe, USA</td>
</tr>
</tbody>
</table>

### JULY

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>02 July</td>
<td>CLOUD (IEEE Int’l Conf. on Cloud Computing)</td>
<td>Chicago, USA</td>
</tr>
<tr>
<td></td>
<td>EDGE (IEEE Int’l Conf. on Edge Computing &amp; Communications)</td>
<td>Chicago, USA</td>
</tr>
<tr>
<td></td>
<td>ICDH (IEEE Int’l Conf. on Digital Health)</td>
<td>Chicago, USA</td>
</tr>
<tr>
<td></td>
<td>ICWS (IEEE Int’l Conf. on Web Services)</td>
<td>Chicago, USA</td>
</tr>
<tr>
<td></td>
<td>QSW (IEEE Int’l Conf. on Quantum Software)</td>
<td>Chicago, USA</td>
</tr>
<tr>
<td></td>
<td>SERVICES (IEEE World Congress on Services)</td>
<td>Chicago, USA</td>
</tr>
<tr>
<td></td>
<td>SSE (IEEE Int’l Conf. on Software Services Eng.)</td>
<td>Chicago, USA</td>
</tr>
<tr>
<td>03 July</td>
<td>EuroS&amp;P (IEEE European Symposium on Security and Privacy)</td>
<td>Delft, Netherlands</td>
</tr>
<tr>
<td></td>
<td>CSF (IEEE Computer Security Foundations Symposium)</td>
<td>Dubrovnik, Croatia</td>
</tr>
<tr>
<td></td>
<td>ICALT (IEEE Int’l Conf. on Advanced Learning Technologies)</td>
<td>Orem, USA</td>
</tr>
<tr>
<td></td>
<td>ICME (IEEE Int’l Conf. on Multimedia and Expo)</td>
<td>Brisbane, Australia</td>
</tr>
</tbody>
</table>

### AUGUST

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 August</td>
<td>IRI (IEEE Int’l Conf. on Application-specific Systems, Architectures and Processors)</td>
<td>Porto, Portugal</td>
</tr>
<tr>
<td></td>
<td>ICCP (IEEE Int’l Conf. on Computational Photography)</td>
<td>Madison, Wisconsin, USA</td>
</tr>
</tbody>
</table>

---
Information Reuse and Integration for Data Science, Bellevue, WA, USA
14 August
• FiCloud (Int’l Conf. on Future Internet of Things and Cloud), Marrakesh, Morocco
21 August
• PST (Int’l Conf. on Privacy, Security & Trust), Copenhagen, Denmark
23 August
• HotI (IEEE Hot Interconnects Symposium), virtual
25 August
• SmartIoT (IEEE Int’l Conf. on Smart Internet of Things), Xining, China
27 August
• HCS (IEEE Hot Chips Symposium), Palo Alto, California, USA
30 August
• MIPR (IEEE Int’l Conf. on Multimedia Information Processing and Retrieval), Singapore

SEPTEMBER
4 September
• ARITH (IEEE Int’l Symposium on Computer Arithmetic), Portland, Oregon, USA
• RE (IEEE Int’l Requirements Eng. Conf.), Hannover, Germany
16 September
• SmartCloud (IEEE Int’l Conf. on Smart Cloud), Tokyo, Japan
25 September
• ACSOS (IEEE Int’l Conf. on Autonomic Computing and Self-Organizing Systems), Toronto, Canada
• IC2E (IEEE Int’l Conf. on Cloud Eng.), Boston, USA
• MASS (IEEE Int’l Conf. on Mobile Ad-Hoc and Smart Systems), Toronto, Canada
• STC (IEEE Software Technology Conf.), virtual

OCTOBER
2 October
• ICCV (IEEE/CVF Int’l Conf. on Computer Vision), Paris, France
• LCN (IEEE Conf. on Local Computer Networks), Daytona Beach, USA
9 October
• e-Science (IEEE Int’l Conf. on e-Science), Limassol, Cyprus
18 October
• SecDev (IEEE Secure Development Conf.), Atlanta, USA
22 October
• VIS (IEEE Visualization and Visual Analytics), Melbourne, Australia
31 October
• CLUSTER (IEEE Int’l Conf. on Cluster Computing), Santa Fe, USA

NOVEMBER
4 November
• ICEBE (IEEE Int’l Conf. on E-Business Eng.), Sydney, Australia
12 November
• SC (Int’l Conf. for High-Performance Computing), Networking, Storage, and Analysis), Denver, USA

DECEMBER
1 December
• ICDM (IEEE Int’l Conf. on Data Mining), Shanghai, China
5 December
• BIBM (IEEE Int’l Conf. on Bioinformatics and Biomedicine), Istanbul, Turkey
• RTSS (IEEE Real-Time Systems Symposium), Taipei, Taiwan
11 December
• ISM (IEEE Int’l Symposium on Multimedia), Laguna Hills, USA
14 December
• BCD (IEEE/ACIS Int’l Conf. on Big Data, Cloud Computing, and Data Science Eng.), Ho Chi Minh City, Vietnam
18 December
• iSES (IEEE Int’l Symposium on Smart Electronic Systems), Ahmedabad, India

Learn more about IEEE Computer Society conferences
computer.org/conferences
YOU HAVE ENORMOUS RESPONSIBILITY.

Protect yourself from risk.

1-800-493-IEEE (4333)
To learn more*, visit IEEEinsurance.com/IEEEPL

Program Administered by AMBA Administrators, Inc.
In CA d/b/a Association Member Benefits & Insurance Agency
CA License #0196562 | AR License #100114462
99804 (5/23) Copyright 2023 AMBA. All rights reserved