IEEE
Advancing Technology
for Humanity

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

INTERNATIONAL
ROADMAP
FOR
DEVICES AND SYSTEMS™

2021 EDITION

SYSTEMS AND ARCHITECTURES

THE IRDS IS DEVISED AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

Roadmaps are inherently a team effort. Our Systems and Architectures team would like to thank our colleagues on the other teams for their information, input, and suggestions. We would also like to extend our appreciation to the IRDS leadership team for their insight and guidance. I would like to wholeheartedly thank our Systems and Architectures team members for their diligence and creativity:

Prof. Tetsuya Asai, Hokkaido University

Kirk Bresniker, HPE

John Carter, IBM

Tom Conte, Georgia Tech

Ty Garibay, Arteris

Prof., Yuko Hara-Azumi, Tokyo Institute of Technology

Dr. Yoshi Hayashi, Keio University

Bruno Michel, IBM

David Mountain, LPS

Shubu Mukherjee, Cavium

Kushagra Vaid, Microsoft

Kirk Bresniker – Hewlett Packard Labs, Stephen Dukes – Dreamerse

SYSTEMS AND ARCHITECTURES

1. INTRODUCTION

The Systems and Architectures section of the roadmap serves as a bridge between application benchmarks and component technologies. The systems analyzed in this section cover a broad range of applications of computing, electronics, and photonics. By studying each of these systems in detail, we can identify requirements for the semiconductor and photonics technologies that make these systems and applications possible.

This section considers four different types of systems, as follows:

1. Internet-of-things edge (IoTe) devices provide sensing/actuation, computation, security, storage, and wireless communication. They are connected to physical systems and operate in wireless networks to gather, analyze, and react to events in the physical world.
2. Cyber-physical systems (CPS) provide real-time control for physical plants. Vehicles and industrial systems are examples of CPS.
3. Mobile devices such as smartphones provide communication, interactive computation, storage, and security. For many people, smartphones provide their primary or only computing system.
4. Cloud systems power data centers to perform transactions, provide multimedia, and analyze data. Cloud systems represent a trend towards a synthesis of design principles and methodologies taken from traditional enterprise, high performance scientific, and web native compute. Increasingly these systems are utilizing artificial intelligence to continue to improve operational efficiency, becoming CPS in their own right.

Increasingly, these four categories of systems are being combined into entire edge to cloud large scale intelligent social infrastructure systems of complex interlocked information lifecycles. Each is continuing to demand ever greater capacity in diminishing space, weight and power envelopes, giving economic motivation to gaining as much as we can from conventional approaches as well as even greater potential for novel approaches. The implications of this trend are summarized here and more fully explored in the 2018 IRDS White Paper "Preparing for Data-Driven Systems and Architectures – Edge, Cloud and Core".

1.1. THE RISE OF SECURITY AS A FIRST ORDER DESIGN PRINCIPLE

While the performance gains from Moore's Law have allowed near continuous improvement on existing performance and efficiency on existing applications and the operating systems and libraries that support them throughout both the geometric scaling and equivalent scaling eras, this is not without ramifications. Some of the critical software infrastructure still in current use were conceived when physical security was the primary mechanism by which data integrity was maintained. The reliability, availability, and serviceability (RAS) of systems was of much greater concern that the security of systems which were often vertically integrated and turnkey.

The forces behind the evolution of the modern threat landscape have evolved over decades but are growing exponential in complexity. As the process roadmaps transitioned from geometric scaling and its clock speed dominated single threaded performance improvement cycle to the relatively modest frequency gains of equivalent scaling, performance and efficiency were improved by greater and greater levels of monolithic integration. This started with the integration of cache RAM, then multiple cores, memory controllers and I/O complexes eventually yielding complex system-on-chip (SoC) designs. The formula of "Moore (exponential scaling down of device size) – Dennard (scaling down device power with size, ended circa 2005) + Rock (exponential increase in cost of a fab with each process step)" [1] a recipe for consolidation both microscopically on die as well as macro-economically across the industry with reduction of commercially instruction set architectures (ISAs) to primarily Arm and x86, leading to a software monoculture. While Moore's Law continued, so did Amdahl's Law, that parallelism is challenging, which in turn lead to virtualization technology becoming crucial first to efficient utilization of the increasingly complex SoCs and then to the agility demanded of cloud native applications. At the edge, computational capabilities that would have required entire data centers are now commonplace in mobile and embedded devices. All of these effects have now converged with the rise of well-funded state and non-state actor exploitation research teams to yield the modern advanced persistent threat landscape that has evolved from exploiting programming bugs and the weak memory security models to sophisticated side channel attacks capable of stealing cryptographic keys via inference in timing variations due to shared substrates underneath virtualization layers. Attacks within the supply chain are growing increasingly sophisticated in moving from exploit hardware added after deployment to in-production attacks via hardware, firmware, and potentially within components themselves.

The response is to drive towards increasingly higher levels of cryptographic protection, authentication, and attestation. While these techniques are well established over communications networks, they are now being brought down to the level of component-to-component communications. In addition to cryptographic protection of data both at rest and in flight, which requires both cipher engines as well as key management, this zero-trust model will require authentication exchanges between components prior to utilization. Authentication will itself require management of certificates and may eventually need to be linked back to a physically (PUF) [2] of a silicon device itself. This will place an added demand on fabs to not only manufacture the components but to also provide the provenance of the components so that they can be authenticated prior to every use. Increasingly the ability to detect a massive, distributed advanced persistent threat will likely require artificial intelligence to pro-actively detect anomalous behaviors across complex edge-to-cloud infrastructure, but that in turn will require increases in computational efficiency and data analytics to establish the base lines and chains of evidence. Cybersecurity and artificial intelligence (AI) are co-dependent for continued advancement.

1.2. EVOLUTION OF EDGE TO CLOUD PLATFORMS TOWARDS PERVASIVE DATA ANALYTICS

All four of these following application areas are in general use: 1) mobile devices number in the billions worldwide—regardless of whether they are operated privately or for public consumption; 2) cloud systems are engendering new programming languages and methodologies and cloud-native computing; 3) cyber-physical systems provide essential services, and 4) Internet-of-Things networks perform important services in a range of applications.

These systems do not exist in isolation. Mobile devices, IoT edge devices, and cyber-physical systems all provide data that is analyzed by cloud systems. Many complex systems exhibit characteristics of both IoT and CPS. Certain aspects of data centers and cloud systems—power management and thermal management, for example—make use of cyber-physical and IoT techniques. Figure SA-1presents these associations across nine Space, Weight, and Power (SWaP) design envelopes ranging from embedded to exascale HPC data centers and how the IoT edge, CPS, and Data center categories overlap. A next generation social infrastructure solution, such as intelligent mobility or AR/VR augmented gaming will position the fourth category, mobile devices, to interact with all of these design envelopes to deliver a complete solution.

The volume of data generated by IoT and cyber-physical systems is staggering. The sensor fusion platforms of a fleet of 1000 conventional connected advanced driver-assistance systems (ADAS) vehicles generates four petabytes per day of data from their onboard sensors; that volume of data is equal to the total data volume handled by Facebook. While today the vast majority of in-vehicle data is discarded after it has been analyzed to provide immediate operational and safety benefits, efficiency breakthroughs allowing *in situ* analysis of raw data in IoT and CPS systems could provide extremely disruptive economic potential. What may evolve is the edge-to-cloud platform where today's hub-and-spoke model is replaced by complex and dynamic topologies where cloud as-a-service consumption models are extended out from the data center towards successively smaller edge device meshes. As increasingly sophisticated computation infrastructure is distributed towards edge devices, a new class of latency-sensitive distributed massive data analytic applications could emerge, such as: intelligent mobility systems, 5G and successive communications networks and advanced augmented reality/virtual reality (AR/VR) gaming applications are all examples of application classes where millisecond or microsecond latencies on complex data analytic and data synthesis workloads may demand several tiers of computational capacity trading off space, weight, power and performance against latency.

What admits data into economic activity is an information lifecycle—acquisition, assurance, analysis, insight and action—in which the analysis allows for timely action and for which the costs of analysis are outweighed by the benefits of action. Timeliness is the most important constraint, followed by the per cycle costs of analysis yielding a time limited return on investment. At every scale of design envelope from embedded IoT device to exascale data center, the numerator and denominator of this time-limited return on investment (ROI) can be affected by adoption of novel computational, memory and communications approaches from both the "More Moore" and "Beyond CMOS" roadmaps.

<table>
<tr><th>DESIGN ENVELOPE</th><th colspan="3">SYSTEM CATEGORY</th><th>INTEGRATED TECHNOLOGIES</th></tr>
<tr><td>Beacon
Sensor</td><td rowspan="4">IoTe</td><td rowspan="6">CPS</td><td rowspan="5"></td><td>Trusted data sources
2.5D/3D integration of sensors, memory, accelerators, computation, and comms
Energy Harvest with inducted power boost modes
SRoT/Blockchain trust mechanisms</td></tr>
<tr><td>Access point</td><td>Unified 5G/Wi-Fi access point
Identity, Activity, Locality triangulation
ML/AI augmented operation</td></tr>
<tr><td>Aggregation Point</td><td>Robust environmentals
Edge local secure hosting of containerized workloads
Static composition
Smallest IT/OT Blended Platform target</td></tr>
<tr><td>Edge
Hardened</td><td>Robust environmentals
Legacy PXI/AXIe plus next gen modular FF
Static composition
Robust IT/OT Blended Platform target at several capacity points</td></tr>
<tr><td>Single System
Flex</td><td rowspan="5"></td><td>OPC/Rack/Tower systems with next gen modular FF option bays and electrical/optical memory fabric (Gen-Z/CXL) expansion
Static fabric configurations between reboots
Low cost point-to-point expansion</td></tr>
<tr><td>Enclosure
Composable</td><td rowspan="4">Data-Center</td><td>Blade Enclosure augmented with next gen modular FF and memory fabric at the enclosure and rack level
Enclosure level switching of fabrics
Static/Dynamic fabric configurations</td></tr>
<tr><td>Rack Scale</td><td rowspan="3"></td><td>Dense next gen modular FF enclosures with integrated switching
Large Scale memory fabric enclosure as endpoint
Dynamic fabric configuration
Dematertialized and legacy free
Design for Flex Capacity, Co-Lo, aaS Consumption models
Containers on memory fabrics</td></tr>
<tr><td>Aisle/Pod
Modular</td><td>Dense next gen modular FF enclosures with integrated switching
ToR switch
Dynamic fabric configuration
Dematertialized and legacy free
Design for Flex Capacity, Co-Lo, aaS Consumption models
Petascale HPC and Petascale Enterprise in-memory DB/Analytics</td></tr>
<tr><td>Exascale HPC</td><td>DC scale memory-semantic fabric over photonics
All liquid/conduction cooling environmentals
Aisle/Pod modular for I/O nodes
2.5D/3D integrated CPU/GPU/Memory modules</td></tr>
</table>

Figure SA-1 SWaP Design Centers across infrastructure system categories

1.3. DRIVERS AND TECHNOLOGY TARGETS

As described above, this section of the chapter describes four types of systems: IoT edge devices, cyber-physical systems, mobile devices, and cloud systems. Each has its own set of drivers and technology targets as described in Sections 5 through 8. Given the wide range of systems—ranging from self-powered very large-scale integration (VLSI) devices to industrial park-sized data centers—we should expect each system area to merit its own description and metrics.

1.4. VISION OF FUTURE TECHNOLOGY

Artificial intelligence [3]has emerged as a critical technology in applications as diverse as smartphones and autonomous vehicles. Much AI-driven computation will occur in the cloud, but we expect mobile systems, IoT edge devices, and cyber-physical systems to all include AI components. In all cases, this move towards pervasive AI creates new demands on data analytics, both in the training of AI/ML models and in the value of inference of those models on novel data sources. For time critical inferencing, this will mean the desire to host increasingly complex models in decreasingly small SWaP footprints, including in energy harvested environments. For use cases in which the subject matter, such as natural language processing, is under continuous evolution, the models will need to be continually improved which at a minimum creates the need for secure over the air updates but also may require distributed attestation and training when data sets are prevented either by law or by economics from being centralized for continuous re-training. Since models are continuously derived from data, the provenance and security of the data flowing into these continuous integration and deployment regimes becomes paramount and may need to be reflected in security and attestation features down to the lowest level devices.

We expect augmented reality/virtual reality (AR/VR) to emerge as an important application area, particularly for edge to cloud systems. The large demands on computing, sensing, and display for AR will drive the development of mobile systems in particular. While multimedia has driven many aspects of mobile system development for many years, we have reached perceptual limits for many multimedia applications and the content delivery networks capable of globally staging desirable content at acceptable latencies yield excellent streaming performance which will only increase with the transition to 5G/WiFi6 wireless technologies. What is novel in the AR/VR systems is the need to synthesize data streams captured locally with geographically and contextually related live and pre-distributed data streams, potentially from many uses co-experiencing an event. All of this will need to take place within the perceptual limitations of the users, which places speed-of-light limitations for low latency computational turnaround. Thus we expect AR/VR to take the place of multimedia as an important driver for mobile systems as well as edge to cloud systems

IoT and CPS are both in widespread use and these systems will continue to expand in scope. We will discuss the relationship between the two in more detail in Section 4.2. Associated with these two types of systems is the increasing use of digital twins that provide computational models for real-world systems. Digital twins are used in both industry and healthcare to help drive analysis and control.

While we continue to describe IoT, CPS and Cloud as distinct classes, these distinctions may become less useful over time. Edge to cloud is emerging as a continuum, where the same application program interface (API)-driven infrastructure-as-code that has come to define the data center computational environment grows outward towards the edge in search access to the disproportionately growing data. Industry analysts have predicted that by 2025 as much as 75% of enterprise generated data will never be housed in a traditional data center—public or private. That data and the computational platforms that will provide access to and analysis of that data will be increasingly geographically disbursed into communications, power, transportation and building systems. These systems will host both data and computational resources proximal to that data, all of which will be consumable on demand using the same consumption models as the hybrid public/private cloud. Again, security of both the data at rest in edge systems and the access to it will require both cryptographic protections as well as end-to-end zero trust attestation that will be continuous from edge to cloud.

A key attribute for cloud systems is the radius of effective communication of data. Traditional architecture, deploying tens of industry standard cores on a system-on-chip (SoC), making use of low latency, high bandwidth direct attached byte addressable memory and higher latency block mode access to shared I/O resources for storage and message passing, was the logical outcome of the second “equivalent” scaling Moore’s law era and was well suited to the general purpose enterprise applications of that time. The general purpose cores provided moderate performance improvement supplanted by increasingly complex cache, memory and I/O systems integrated into the SoC.

This approach is no longer viable. Modern applications such as massive data analytics and graph analytics must operate on huge datasets that cannot be held in those types of memories nor addressed directly by conventional microprocessors. If data access times to block storage can be orders of magnitude longer, programmers must use more sophisticated programming techniques to manage delay—techniques that are often rendered useless by algorithms that do not have predictable locality, such as graph analytics on time varying graphs. The integration of memory and I/O complexes into massive SoCs also relegates application specific accelerators to the block mode, high latency off chip regime. Also, the complex integration of cache, memory, and I/O blocks along with the high core count that supplanted the modest performance increases in the cores themselves has come under increasing attack by advanced persistent threat side-channel attack which can be used to subvert even hardware assisted virtualization.

For dense rack, aisle and data center scale systems, the convergence of open memory-semantic fabrics and photonics are re-shaping the moderate latency regime. When end-to-end latencies are between 300ns and 500ns, software designers can take advantage of relatively straightforward memory resource utilization mechanisms. Memory-semantic fabrics allow for the promotion of accelerators to first-class participants alongside general purpose cores, allowing each to scaling independently. Photonics allows data center distances to be traversed for the same energy cost as board to board distances and offers much greater physical design freedom and immunity from radio frequency interference (RFI) and emissions. When coupled with a high-radix switch, photonics and memory-semantic fabrics could offer affordable exascale memories at the rack scale, memory latencies at the aisle scale and unified message passing at the data center scale and potentially beyond.

1.5. TEN SYSTEM-LEVEL TECHNOLOGY INFLECTION POINTS

Adoption of Data-Driven systems in science, engineering, enterprise governance, and social infrastructure will be accelerated by the confluence of key technology inflection points.

1.5.1. *FROM PROGRAMMING TO TRAINING AND INFERENCE*

This shift is driven by the combination of open source software frameworks and the rise of AI machine learning frameworks capable of creation of very effective models based on statistical inference. Unsupervised learning techniques can comb over huge volumes of structured and unstructured data to find correlations independent of expert blind spots. Intelligence craves data and artificial intelligence is no exception.

This creates a shift in the economic potential from those who create code to those who create the data without which those code stacks are not useful. This also challenges us because the utility of these AI systems is limited not by the ingenuity of the human programmers but instead by the degree in which we have engineered systems to admit as much data as possible into training regime as our physics and our legal and security systems will allow.

Creation of models is only half of the challenge—deploying and utilizing the model, gathering anomalies from operation to fuel of continuous integration and continuous deployment also demand infrastructure and innovation.

1.5.2. *FROM ONE PHYSICS TO MANY*

Through the first two eras (geometric and equivalent) of semiconductor scaling, there have been incredible advances in the other aspects of computer science—algorithms, programming languages, storage and communications technologies all contributed, but they were fundamentally modulated by the CMOS transistor. Innovations were tested against the cost and performance improvements predicted by Moore's law and if they did not have the exponential growth characteristics they were not admitted.

Even the obvious defects in security source to the conceptual basis of software models based on 1960s threat landscapes failed to be fixed at the source because of dominance of architectures with the tailwind of CMOS advances.

Now, as CMOS advancement transitions from equivalent scaling to 3D Power scaling, novel computational approaches are increasingly competitive. The work that might spring most quickly to mind, quantum computing, along with cryogenic computation, emerged as a particular area of focus. However, it is not the only one. Other areas include novel switching technologies, such as:

- Carbon nanotubes;
- Adiabatic and reversible computing that operate at the limits of thermodynamic information theory;
- Neuromorphic and brain-inspired computing that draws inspiration from biological systems but, much as with aerodynamics, utilizes materials and energies not available to their biological analogs; and
- Networks of organic and inorganic materials whose behavior calculates desirable functions at breakthroughs in space, weight, and power; as systems created in our own image that are designed primarily to host intellect that offer computation as a byproduct of intelligence.

1.5.3. *FROM DATA CENTERS TO DATA EVERYWHERE*

Today, 90% of information that the enterprise, public or private, cares about is housed in a data center. By its very name, it describes the actions that we have undertaken. In order for data to enter into economic activity, it must be centered, either because it was created there or it had to be transported there. But, with the advent of so many rich, high definition sensors housed in the ever proliferating number of mobile devices, in as little as five years that ratio may shift drastically to as much as 75% of enterprise information never being housed in a data center.

It is not that the data center footprint will shrink, although it will continue to coalesce into clouds both public and private, but that data will grow exponentially and disproportionately at the edge, in distributed social infrastructure, in edge devices personal, public, and private, in all those intelligent things.

There are two forces that keep data at the edge—physics and law. The exponential growth of recorded data, currently a two-year doubling period, means that even with the advent of 5G communications and massive communications backbones, there will never be enough bearer capacity to centralize all the data and even if there was, Einstein's limit of the speed of light means that at even metropolitan distances our fastest communications will fail to meet the demands of autonomous vehicles or 5G communications.

The second force is law. There is no global standard on privacy and the relation and responsibility of the individual to the larger society, which means that there will not be a single regulatory regime that spans the globe. Just like citizens and goods today, data needs to obey the imposition of boundaries. Will frameworks like GDPR continue to offer the protections that they strive to when the vast majority of data will never be in a data center, when the very term data center will be an oxymoron?

The question to ask here is, "What will it take to admit as much data as possible into economic activity?" The first answer is to exploit the asymmetry of the query versus the data to be analyzed. Instead of moving the data to the compute, move the compute to the data. This requires us to understand where we position potentially shared computation resources proximal to the data—in sensors, edge devices, distributed edge compute enclosures, autonomous vehicles. The second requirement to admit data to activity is security in the broadest sense—protection, trust, and control. Protection: robust and energy-efficient cryptograph ensures that query and response are demonstrably safe and correct. Trust: provenance backed by secure supply chains, silicon roots of trust, and distributed ledger systems with low energy consensus functions ensures every byte flowing into an enterprise can be audited. Control: meta-data embedded unforgeably in the data ensures down to the byte and the access cycle all stakeholders in a computation can have their rights verified and protected.

1.5.4. FROM IMPERATIVE TO DECLARATIVE

Imperative control systems rely on enumeration of conditionals and responses, the classic if-then-else diamonds of the flowchart. The problem with imperative control is that the systems we are creating—social, technical and economic—are two complex to be enumerated. No matter how much time we spend, we never can catch the corner cases, there are always exceptions and that means we need to guard band and that means inefficient use of resources, whether it is spectrum allocations or transportation capacity.

Declarative management instead relies on systems that expose their operational state and control surfaces to goal seeking algorithms, such as reinforcement learning. Instead of enumerating all the "ifs" and "thens," we can set goals to be achieved and let the system strive to maximize those goals. This approach has the added benefit that it does not suffer from the human bias of presupposition of causality preventing us from finding correlations hiding in plain sight. A declarative system using unsupervised learning and autocorrelation could naively, blindly discover those correlations humans discount because it cannot presume it knows better.

1.5.5. FROM SCARCE MEMORY TO ABUNDANCE

A decade after Alan Turing created the mathematic theory of computation, John von Neumann was realizing that theory as an operational feat of engineering in his 1946 outline of EDVAC. What von Neumann noted then and what has remained true is that the fundamental limiter to computation is how reliably and cheaply the memory can be made that can keep up with computation.

Computation performance has always advanced faster than memory performance. But that is changing. As we enter the age of 3D power scaling, memory is advancing faster than computation. The regular rows and columns of memory; the inherent shared, redundant, and repairable structures of memory, and the low power dissipation of memory mean that it can growth in the Z axis in a way that may never be possible for the high power and random logic of computation.

With a structure of layers within a die, die within a module, and modules within a package, memories can scale. At that point, the switch to photonic communications can allow the scaling to continue at the enclosure, rack, aisle and data center scale. A second scalability of memory is scalability in energy. All of the novel memory technologies looking to replace the transistor memory, phase-change, resistive, spin torque, magnetic, all have a degree of persistence. They cost energy to write, they cost much less energy to read, but they cost no energy to maintain their contents. This is what can allow all of those zettabytes of data into unsupervised learning that we can now afford the energy to hold it all in memory. It also reintroduces a technology older than electronic computation— the lookup table.

The table of numerical functions used to be the constant companion of the scientist or engineer. Energy was expended to calculate numbers one time, to write those numbers one time, and then those costs could be amortized in perpetuity. From the 1970s onwards, it has been cheaper to recalculate a result than to remember and recall it. But with persistent memories applied to immensely complex calculations like machine learning routines, incredible volumes of information can be distilled into insights that can be taken to the most energy-starved environments like interplanetary space.

1.5.6. FROM HINDSIGHT TO FORESIGHT

If we consider all of the information technology infrastructure of a Fortune 50 company, the alphabet soup of HR, CRM, ERP, GL systems, we will find a system of hindsight knowledge. That is because what represents the state function of the enterprise—the operational data of all of those systems—is spread over petabytes in thousands of relational databases connected by hundreds of thousands of asynchronous updates, and much of that data would be copies. In order to evaluate the state function of the enterprise, we need to go through a ritual of reconciliation. We need to “close the books”, take a snapshot of all of those systems and painstakingly reconcile them. It is only then that a CEO/CFO executive leadership team have a value of the state function of the enterprise, but it is at best days, most likely weeks old and represented a single moment in time, the instantaneous close of the period.

If, instead, we were able to hold all of that operational state in a unified memory, evolving as a time varying graph, then we can achieve insight. The system function of the enterprise can be evaluated instantaneously and continuously, which means that we can also take its derivatives with respect to time and understand velocity and acceleration, gradient and curl. Now decision makers can ask any ad hoc question and the enterprise can answer. We have extended the concept of a digital twin from its origins in physical systems management and extended it to economic systems management. But what is more, we can unleash unsupervised learning and anomaly detection tools to audit and analyze the data, looking for the telltale signs fraud or inefficiency. But we can also extended the preventative maintenance concepts to this new economic model. While machine learning gives us powerful statistical inference tools to find in data the patterns we’ve seen before, techniques like graphical inference and belief propagation allow us to predict behaviors we haven’t seen.

From hindsight “what has been happening around here” we gain insight “what is happening right now” and then foresight “what most likely to happen next”.

1.5.7. FROM GENERAL PURPOSE TO BUILT-FOR-PURPOSE

“log2(X)*24 Traditionally, that is how long a point innovation has had to survive in months. If one expects an advantage of “X” times the state of the art today, then the log base 2 is how many doublings it will take to match. The Moore’s Law doubling period of 18~24 months has set the timeframe for innovation, especially when Dennard scaling was still available. Faster, cheaper to make and cheaper to use is a triple word score. Unfortunately since Dennard scaling ended 15 years ago the straightforward way to continue to reduce power and increase performance has been to make larger and larger die. We are at the point now of “dark silicon,” which means that we can make more transistors than we can deliver power to. If all the circuits on a die were active, the heat could not be removed fast enough and the chip would fail. Add one more law, Rock’s Law [4], the observation that each successive chip fab costs twice as much. “Moore – Dennard + Rock” is the recipe for consolidation at every level—the number of companies that can compete to the number of competitive architectures.

But during this transition period between equivalent scaling and 3D power scaling, may be a period when the tide will shift back to the economic value of novel accelerator design.

1.5.8. FROM PROPRIETARY TO OPEN

The Open Source development and collaboration model has proven incredibly effective in software, not only in the complexity of systems that can be delivered, but also in the diversity of those who are enabled to participate. This creates the virtuous cycle where internationalization and localization occur as primary efforts coincident with innovation rather than after the fact, creating greater diversity of representation that again fuels greater inclusion in the economic and social benefits of innovation.

The same guiding principles of open source software development are being extended down the stack. As an example, Gen-Z [5]is a memory-semantic fabric driven by an industry consortium applicable to ever level of integration from embedded to exascale. It has been open for review by the open source software community during the entire draft period and lowers the barrier to innovation for novel computational, memory, and communications devices. Regardless of whether it maximizes the potential of conventional CMOS or enables new physics to accelerate a particularly onerous computation, lowering the barrier to innovation and breaking the cycle of improvement solely through consolidation is the antidote for today’s technical monoculture.

RISC-V [6] is an Instruction Set Architecture with an open governance model which fully embraces the open source development model in that it freely extensible and licensable. This is a unique new proposition which simultaneously allows for a sustained core software development model that also allows innovation and customization that can be realized in custom or programmable silicon. When coupled with the emerging capacity of from multiple foundries of relatively competitive logic processes, this again enfranchises an ever increasing number of innovators everywhere.

1.5.9. FROM CENTRAL AUTHORITY TO DISTRIBUTED SYSTEMS

Whether they are economic (cryptocurrency and public ledger), power (microgrids), or communications systems (mesh networks), distributed systems are more complex than centralized systems. But they are more sustainable, more available, more secure, and more equitable, which in turn makes them arguably more just.

1.5.10. FROM DATA AS COST BURDEN TO DATA AS OPPORTUNITY

From its inception, information technology has been dominated by the mechanical advantage and error reduction of automation of human calculations, affording an incredible increase in productivity. Coupled with this productivity increase is the inevitable desire to contain the associated costs. The combination of all of the other effects yield the more transformative effect—the shift of information technology from a cost center to a profit center by simultaneously increasing the return on processing information while reducing the cost of information. In fact, given the predictive capability of these systems and the efficiency at which ML/AI systems can operate themselves, everywhere there is data— every manufacturing step, every business operation, every customer interaction casts off information continuously— potentially at a greater level of return than the underlying process itself. The hypercompetitive business relentlessly and sustainably turns raw data to economic advantage via process improvement, investment strategy, customer satisfaction, market expansion, warranty reduction, and direct monetization.

2. SCOPE OF REPORT

This report describes four important types of systems: 1) Internet-of-Things (IoT) edge devices, 2) cyber-physical systems (CPS), 3) mobile systems, and 4) cloud systems. For each type of system, we discuss market drivers, challenges and opportunities, power and thermal considerations, and metrics.

3. SUMMARY AND KEY POINTS

- Security and privacy are key system requirements for all four system areas, both for protection of data at rest, in motion and in use as well as attestation of solution elements at every level down to low level devices.
- We expect artificial intelligence and augmented reality to become important new drivers for the growth of all four system areas, especially in supporting the massive underlying data analytic flows.
- Internet-of-Things and cyber-physical systems both generate vast quantities of data that will accelerate the growth of big data and create a continuum of edge to cloud systems.
- Advanced packaging is a key technology for enabling architectural diversity. Chiplets on 2.5D substrates, the wide variety of 3D technologies, and wafer-scale integration using fine pitch lithography can provide significantly increased local bandwidth.
- When coupled with photonics technology, fabric attached memory (both DRAM-based and non-volatile), and the recent emergence of the RISC-V ISA and other open source hardware initiatives, future architectures could become both more flexible and specialized, opening up new architectural dimensions of innovation. However, managing this extreme heterogeneity will present difficult application development and system software challenges.
- As data grows disproportionately at the edge, computation will follow it, with increasingly demanding workloads in increasingly challenging space, weight, power and costs envelopes creating opportunity for non-conventional architectures and approaches including those tailored to harvested energy.

4. EMERGING TRENDS

4.1. INTERNET-OF-THINGS AND CYBER-PHYSICAL SYSTEMS

This roadmap provides separate analysis of IoT edge (IoTe) devices and cyber-physical systems. While both types of systems connect computing devices to the physical world, and there is some overlap in the usage of these terms, we believe that considering them separately in this roadmap gives readers greater insight into the evolution of such systems. We can contrast CPS and IoT systems in several ways:

- Cyber-physical systems perform real-time control—the core control functions operate automatically and without user intervention. IoT systems put more emphasis on sensing: they are also more likely to provide data summaries to humans who adjust system operation based on those summaries.
- Many cyber-physical systems are, at their core, based on wired networks, although wireless sensors may be used in these systems. IoT systems are often deployed over larger areas and make more extensive use of wireless connections.
- Cyber-physical systems tend to operate at higher sample rates than do IoT systems. We choose for convenience of discussion a boundary of 1 second between cyber-physical and IoT systems. IoT systems are often organized as event-driven systems that either react to sensor activations or transmit data only when analysis indicates that a signal is of significant interest.

4.2. CONVERGENCE

The huge volume of data generated by IoT and cyber-physical systems means that within the next five years the majority and then the vast majority (as much as 75% by one estimate) will never reach traditional data centers. Even with the advent of increasing bandwidth from next generation 5G/WiFi6 wireless interconnects, data growth will out strip transmission capacity. Both transmission energy and costs as well as regulatory, security, and privacy burdens will keep data in edge devices. As edge systems become the majority of data resources, the desire to access them directly using the same cloud native APIs and continuous integration / continuous deployment software development methodologies will increasingly drive security and performance features and their enabling components into CPS and IoTe devices. This represents a convergence of the traditional operational technology (OT) components and methodology with their information technology (IT) equivalents. This represents a security and attestation challenge as many OT technology standards have been developed with lightweight security and little to no attestation mechanisms.

For this reason and for the need to provide additional low latency computation, cloud-native enabled IT computational footprint ranging from rack scale down to ruggedized small single servers designed for extended environmental conditions will become gateways stitching together the OT and IT worlds.

4.3. EDGE TO CLOUD SERVICE MESHES

Whether public or private, cloud systems today offer compute, storage, networking infrastructure deployable via APIs, infrastructure as code. They also allow data and application resources to be deployed via APIs as well, usually up to the physical extent of an extended high-availability zone. The trend within a zone is for greater and greater of levels of abstraction: data, applications, infrastructure are all abstracted as APIs and complex solutions are composed at scale and with high reliability and security without the developers having to understand, or have any access to, the lower level implementation details. This separation yields a degree of freedom on the cloud infrastructure designer to adapt novel technologies and to instrument the controls of these massive systems with AL/ML for operational efficiencies that human operators cannot achieve. However because of the lack of standards, compositing applications between zones of a single cloud provider, let alone across multiple providers, is extremely challenging. The disproportionate growth of data in edge systems coupled with the rise of low latency demanding applications such as AR/VR [3]may couple with the desire to compose solutions across the entire continuum of private to public cloud and edge to data center clouds in new constructs call service meshes. Service meshes may allow solution developers to balance latency, cost, reliability, security, privacy, availability and sustainability and re-introduce a counterforce to the consolidation of supply chain and lack of competition in current cloud data center providers. Key to service mesh construction is the adoption of ubiquitous zero trust endpoint security mechanisms rooted in physically uncloneable features in silicon and network independent name space resolution that can scale to a globally distributed edge to cloud ecosystem.

5. CLOUD

The term *cloud* refers to the engineering of data center scale computing operations—compute, storage, networking engineered for scale and for continuous resource redeployment and reconfiguration via APIs. Whether they are operated publicly or privately, they offer on-demand, as-a-service consumption model. While they had their origins in web service; media streaming, shopping and commerce; they are increasingly broadening their applications base to big data for social networking, recommendations, and other purposes; precision medicine; training of AI systems, and high-performance scientific computation for science and industry.

Cloud infrastructure has undergone several waves of optimization from its initial deployment of industry standard rack servers, storage and compute at data center scale: commercial off-the-shelf to custom loading to purpose-built at the motherboard level to today's cloud-native compute, storage and networking that can feature bespoke processors designs, networking interface and switch ASICs, and workload specific accelerators via FPGAs or ASICs.

The traditional differences between high-performance scientific computation and the first generations of web-scale applications are diminishing. Scientific computation traditionally emphasizes numerical algorithms whereas cloud applications, in contrast, emphasize streaming for multimedia and transactions for commerce and other database applications. Now with the AI/machine learning (ML) integrated into so many applications, the demand for accelerated floating point is more universal and all applications are being dominated by operational and capital costs of data movements at scale. Also, in all cases the general trend is to utilize as-a-service consumption model to foster independence of the user from not only a particular piece of hardware infrastructure but from one particular architectural approach. This is a critical enabler for introduction of novel computational approaches from either the "More Moore" or "Beyond CMOS" roadmaps.

5.1. MARKET DRIVERS

Market drivers for the cloud include direct services (multimedia, shopping, shared experience), big data and data analysis (social network analysis, AI, smart cities, smart industry, precision medicine). We note that while these applications have differed from traditional scientific computing applications that emphasize numerical methods, this distinction is becoming less important as data movement and storage costs come to dominate both applications domains.

5.2. CHALLENGES AND OPPORTUNITIES

The cloud data center, public or private, is no longer a homogeneous footprint of commercial off-the-shelf (COTS) compute, storage, and networking. The continued demand for efficiency and the both the breadth of traditional enterprise and high-performance computing (HPC) applications being migrated to hybrid public/private clouds as well as the new cloud-native applications are admitting bespoke silicon solutions in compute, storage and networking, analogous to the advantage of heterogeneous core types employed by embedded systems for many years. The huge scale of problems in social networking and AI, for example, means that algorithms run at memory speed and that multiple processors are required to compute. The *radius of useful locality*—the distance over which programmers can use data as effectively local—is an important metric. We expect the combination of increasingly integrated high-radix photonic switches and open memory-semantic fabrics to greatly enlarge useful locality radius and diversity of compute and memory endpoints over the next few years. Memory bandwidth is a constraint on both core performance and number of cores per socket. Three dimensional scaling of memory at every level—layers-in-die, dice-in-stack, stacks-in-package or stacks-on-ASIC will contribute greater local and fabric attached bandwidth. Thermal power dissipation continues to be an important limit, and may need to be addressed down to inter-die and intra-die cooling.

Cloud systems present significant challenges. Heterogeneous architectures can provide more efficient computation of key functions. Novel memory systems, including stacked memories, offer high performance and lower power consumption. Advances in internal interconnect may create tipping points in system architecture.

5.3. POWER AND THERMAL CONSIDERATIONS

We face fundamental physical limits on our ability to deliver power into and extract heat out of industrial park-sized data centers. Thermal effects limit performance and may affect rack-level utilization. Power and thermal limitations have implications at all levels of the design hierarchy: building, rack, board, and chip.

5.4. METRICS

Key metrics for cloud systems include number of cores or core equivalents per socket (cores may include any type of computational element, including central processing units (CPUs), graphics processing units (GPUs), or accelerators), base frequency, vector length, cache size, memory characteristics [double data rate (DDR), high-bandwidth memory (HBM)], PCI-e connectivity, and socket thermal power dissipation. L1 = level 1 cache; LLC = last-level cache; TDP = total power dissipation.

Table SA-1 Difficult Challenges

	2019	***2020***	***2021***	***2022***	***2023***	***2024***	***2025***	***2026***	***2027***	***2028***	***2029***	***2030***	***2031***	***2032***	***2033***	***2034***
# cores per socket	38	42	46	50	54	58	62	66	70	70	70	70	70	70	70	70
Processor base frequency (for multiple cores together)	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90	4.00	4.10	4.2	4.3	4.4	4.5
Core vector length	512	512	1024	1024	1024	1024	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048
L1 data cache size (in KB)	36	38	38	40	40	42	42	44	44	44	44	44	44	44	44	44
L1 instruction cache size (in KB)	48	64	64	96	96	128	128	160	160	160	160	160	160	160	160	160
L2 cache size (in MB)	1	1.5	1.5	1.5	2	2	2	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
LLC cache size (in MB)	67	73	81	89	97	107	118	130	143	157	173	190	200	200	200	200
# of DDR channels	6	8	8	10	10	12	12	12	12	12	16	16	16	16	16	16
HBM ports	4	4	6	6	6	6	6	6	6	6	6	6	6	6	6	6
HBM bandwidth (TB/s)	2.4	2.4	6	6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6
Fabric lanes	64	72	80	88	96	104	112	120	128	136	144	152	152	152	152	152
Per lane (GT/s)	56	56	56	56	56	56	56	56	56	56	56	100	100	100	100	100
Socket TDP (Watts)	226	237	249	262	275	288	303	318	334	351	368	387	387	387	425	425

L1 = level 1 cache; LLC = last-level cache; Fabric = PCIe or new accelerator fabric (CXL/Gen-Z/openCAPI/CCIX); TDP = total power dissipation.

6. MOBILE

Mobile devices integrate computation, communication, storage, capture and display, and sensing. Mobile systems are highly constrained in both form factor and energy consumption. As a result, their internal architectures tend to be heterogeneous. Cores in modern mobile units include: multi-size multi-core CPUs, GPUs, video encode and decode, speech processing, position and navigation, sensor processing, display processing, computer vision, deep learning, storage, security, and power and thermal management.

6.1. MARKET DRIVERS

Mobile devices provide multiple use cases: telephony and video telephony; multimedia viewing; photography and videography; email and electronic communication; positioning and mapping, and authenticated financial transactions. Current and upcoming market drivers include: gaming and video applications; productivity applications; social networking; augmented reality and context-aware applications, and mobile commerce. Mobile devices already make use of AI technologies such as personal assistants. We expect the deployment of AI on and through mobile devices to accelerate.

6.2. CHALLENGES AND OPPORTUNITIES

Mobile systems present several challenges for system designers. Multimedia viewing, such as movies and live TV, have driven the specifications of mobile systems for many years. We have now reached many of the limits of human perception, so increases in requirements on display resolution and other parameters will be limited in the future based on multimedia needs. Content delivery networks (CDNs) pre-positioning relevant content globally addresses the need for low-latency unidirectional flow from content providers to consumers. However, augmented reality will motivate the need for advanced specifications for both input and output in mobile devices and promote the development of much more complex interactive topologies than today’s CDNs. Future ad hoc mobile mesh communities focused on live events, AR/VR multiparty gaming, or cooperative AR work environments will connect mobile to mobile and mobile to low latency distributed edge compute infrastructure as well as multi-cloud global infrastructure. To date, mobile device buyers demand frequent, yearly product refreshes, but this trend may not be sustainable. This fast refresh rate has influenced design methodologies to provide rapid silicon design cycles; if it attenuates then the push towards differentiation in the connected infrastructure may be the next location for innovation in devices and systems. Financial transactions are now not only routinely performed using mobile devices, they are preferentially performed on the devices due to the ability to add biometric and geographic identity confirmation. We expect this trend to grow, particularly in developing nations, where financial technology will leapfrog.

6.3. POWER AND THERMAL CONSIDERATIONS

Users want long battery life even with active use cases. However, battery chemistry improves slowly. Furthermore, given the high energy densities of modern batteries, we may see regulatory limits on battery capacity and the uses of high-capacity batteries. The high performance of modern mobile devices may create thermal challenges that must be considered to ensure a comfortable experience for users.

6.4. METRICS

Key metrics include CPU and GPU compute power, communication bandwidth, camera count, and sensor count. Augmented reality applications motivate more cameras as well as other types of sensors.

Table SA-2 Mobile Technology Requirements

	2019	***2020***	***2021***	***2022***	***2023***	***2024***	***2025***	***2026***	***2027***	***2028***	***2029***	***2030***	***2031***	***2032***	***2033***	***2034***
# CPU cores	10	10	12	12	18	18	18	25	25	25	28	28	28	30	30	30
# GPU cores	16	32	32	32	64	64	64	128	128	128	256	256	256	512	512	512
Maximum frequency (GHz)	2.8	3.0	3.3	3.7	4.	4.4	4.9	5.3	5.9	6.5	7.1	7.8	8.6	9.5	10.4	11.5
Number of cameras	3	3	3	4	4	4	6	6	6	8	8	8	8	8	8	8
Camera resolution (MP)	12	15	15	18	18	20	20	20	24	24	24	24	24	24	24	24
Number of sensors	6	8	8	10	10	12	12	12	12	12	16	16	16	16	16	16
5G Max data rate (Gb/s)	1	5	5	5	7	7	7	10	10	10	20	20	20	50	50	50
Wi-Fi Max data rate (Gb/s)	5	9.6	30	30	30	30	30	30	30	30	30	30	30	50	50	50
Board power (mW)	5100	5350	5620	5900	6190	6500	6830	7170	7530	7900	8300	8715	9150	9610	10090	10590

7. INTERNET-OF-THINGS EDGE DEVICES

An IoT edge (IoTe) device is a wireless device with computation, sensing, communication, and possibly storage. The device may include one or more CPUs, memory, non-volatile storage, communication, security, and power management. It may be line powered, battery powered or utilize energy harvesting.

7.1. MARKET DRIVERS

Market drivers for IoT include the following: smart cities; smart homes and buildings; medical devices; health and lifestyle; manufacturing and logistics, and agriculture.

7.2. CHALLENGES AND OPPORTUNITIES

IoTe devices must satisfy several stringent requirements. They must consume small amounts of energy for sensing, computation, security, and communication. They must be designed to operate with strong limits on their available bandwidth to the cloud.

Many IoT devices will include AI capabilities; these capabilities may or may not include online supervision or unsupervised learning. These AI capabilities must be provided at very low energy levels. A variety of AI-enabled products have been introduced. Several AI technologies may contribute to the growth of AI in IoTe devices—convolutional neural networks; neuromorphic learning; stochastic computing.

IoT edge devices must be designed to be secure, safe, and provide privacy for their operations.

7.3. POWER AND THERMAL CONSIDERATIONS

IoTe must be designed to provide low total cost of ownership. Given the high cost of pulling wires to IoT devices, as well as the cost of changing coin cell batteries, this means both wireless communication and energy harvesting. Many IoTe devices operate in harsh physical environments, putting additional strain on their thermal management systems.

7.4. METRICS

Key metrics for IoT include CPU count and frequency; energy source (battery or energy harvesting); communication energy per bit; battery operation lifetime; deep suspend current, and number of sensors. Tx = transmit, Rx = receive.

Table SA-3 Internet-of-things Edge Technology Requirements

	2019	***2020***	***2021***	***2022***	***2023***	***2024***	***2025***	***2026***	***2027***	***2028***	***2029***	***2030***	***2031***	*2032*	*2033*	*2034*
CPUs per device	1	2	2	2	4	4	4	4	6	6	6	8	8	8	8	8
Maximum CPU frequency (MHz)	257	300	305	310	315	320	325	330	335	340	346	351	360	363	369	375
Energy source (B = battery, H = energy harvesting)	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H
Tx/Rx power/bit (µW/bit)	0.372096	0.227723	0.139707	0.08571	0.05714	0.038093	0.025396	0.01693	0.011287	0.007525	0.005016	0.003344	0.00223	0.001486	0.001486	0.001486
Battery operation lifetime (months)	6	9	9	9	9	9	12	12	12	12	18	18	18	18	18	18
Deep suspend current (nA)	52	44	38	32	27	23	20	17	14	12	10	9	8	7	7	7
Sensors per device	4	4	8	8	8	12	12	12	16	16	16	16	16	16	16	16

8. CYBER-PHYSICAL SYSTEMS

Cyber-physical systems are networked control systems. These distributed computing systems perform real-time computations to sense, control, and actuate a physical system. Many cyber-physical systems are safety-critical. They interface to the systems they control via both standard and proprietary interconnects broadly know as operational technology (OT), where ruggedness, extended environmental capabilities, low cost have been paramount over considerations such as security and attestation. As these systems are increasingly connected edge-to-cloud, this will present an increasing attack surface, either for data theft, false signal injection, systems commandeering, or as a back door into the IT domain.

8.1. MARKET DRIVERS

Market drivers include automotive and aerospace vehicles, autonomous vehicles, medical systems and implantable devices, and industrial control.

Cyber-physical systems may make use of wireless interconnects, but critical functions are generally performed on a wired network. While an existing physical layer, such as ethernet, may be used for the fabric, the communication protocol is designed for real-time operation. Time-triggered architectures, for example, divide bus access into time slots; hard real-time functions are assigned fixed slots while soft-real functions may arbitrate for access to shared time slots.

8.2. CHALLENGES AND OPPORTUNITIES

Several challenges present themselves to cyber-physical system designers. Cyber-physical systems must be highly reliable at all levels of the design hierarchy. Physical security and isolation have traditionally been part of the design of these systems, but that is becoming a greater challenge as edge to cloud connected design becomes the dominant methodology. Wireless sensors are increasingly used in cyber-physical systems to reduce installation effort and weight; the challenging temperature and electromagnetic interference environments of the physical plants require much stronger component requirements than is the case for typical consumer applications.

Security and safety are critical for cyber-physical systems. Although security and safety have traditionally been handled separately in the design process, cyber-physical systems cause interactions that require safety and security to be handled holistically. Traditional safety practices are sufficient to address security concerns; similarly, computer security approaches are inadequate to handle many safety issues. Privacy is also a key concern for the data generated by cyber-physical systems. We expect the use of AI for cyber-physical systems to continue to escalate, and this again challenges the traditional isolation for safety and security of these systems as either increasingly complex compute must be incorporated into the edge endpoints for in situ inference and anomalous data for must flow out to the edge to cloud training infrastructure, either distributed or centralized.

The sensor fusion platform of today’s connected automobiles are capable of terabytes per day of raw data, almost all of which is utilized only over the very short term to optimize passenger safety and vehicle operations. But, like many of today’s isolated CPS platforms, the potential for the sensor data from cyber-physical systems for big data applications and emerging products such as automated diagnosis and repair dispatch or cooperative sensing is huge. The interaction between CPS, IoT, and edge to cloud infrastructure presents an ongoing challenge and opportunity.

8.3. POWER AND THERMAL CONSIDERATIONS

Some cyber-physical systems, such as vehicles, are powered by generators. In these systems, available power for the computational engine is determined by the capabilities of the generator and the electrical load presented by the physical plant. Many cyber-physical systems present extreme temperature environments in which the electronics must operate.

8.4. METRICS

Key metrics include the number of devices on the bus and number of CPUs per device.

Table SA-4 Cyber-physical Systems Technology Requirements

	2019	**2020**	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**	**2034**
Number of devices	64	64	64	64	128	128	128	128	256	256	256	512	512	512	512	512
CPUs per device	4	4	8	8	8	8	12	12	12	12	16	16	16	16	16	16

9. CROSS TEAMS

The Systems and Architectures roadmap team interacts with several other roadmap focus teams. The Application Benchmarking team provides application data that informs our system architecture analysis. The Outside System Connectivity team provides insight into the ongoing interplay of photonics and fabrics, which at the rack, aisle, and data center scale is blurring the line between compute, storage, networking infrastructure. More Moore and Beyond CMOS provide the novel computation, memory, and communications devices which are being increasingly required at the extremes of edge and exascale.

Table SA-5 Application Benchmark / Systems and Architectures Cross Matrix

	Cloud	*IoTe*	*CPS*	*Mobile*
Big Data Analytics	Y	Y	Y	Y
Artificial Intelligence	Y	Y	Y	Y
Discrete Event Simulation	Y			
Physical System Simulation	Y	Y	Y	
Optimization	Y`		Y	
Graphics/AR/VR	Y			Y
Cryptographic Codes	Y	Y	Y	Y

10. CONCLUSIONS AND RECOMMENDATIONS

We continue to track four system areas of design, but they are growing increasingly interdependent and in future it may be advantageous to delineate edge to cloud as a continuum rather than as four discrete system design centers.

Cloud systems are simultaneously engineered at the data center scale down to bespoke silicon in compute, storage and networking. Cloud native applications are being co-designed in tight loops with infrastructure hardware.

Mobile systems have emerged globally as key computing device for many consumers, far beyond as their original communications functions. Augmented reality and financial transactions are two examples of important emerging applications for mobile systems.

Internet-of-things edge devices must provide sensing, computation, and communication at extremely low power levels. We expect energy harvesting to become more common in this class of devices.

Cyber-physical systems perform real-time computations to control physical systems. Reliability is a key design requirement for CPS. As the union of IoT and CPS cross over to host the majority of global data they will be increasingly need to securely host a converged OT/IT function and cloud native applications will grow to span hybrid multi-clouds along both the public/private dimension as well as the edge to data center dimension.

We have identified several recommendations, as follows:

- Holistic security and privacy are critical to all our system areas—this will drive requirements for features and services back through the supply chain to provide provenance and provable attestation from end user back to system manufacturing, semiconductor fabrication, and original design engineering.
- AI/ML are data driven practices and will create the need for new, complex topologies of data flow from edge to cloud. As CI/CD practices follow the integration of AI/ML into every system design type, IT security and performance features will follow towards edge systems.
- Energy harvesting is a key technology to enable the growth of IoT edge devices.

- Augmented reality will create further demand for computation, communication, sensing, and display on mobile devices and will be a class of applications which will span ad hoc low latency mobile to mobile and mobile to edge infrastructure.
- Cloud system architectures should take advantage of advances in interconnect to provide simpler programming models for cloud application programmers.

11. REFERENCES

[1] K. Bresniker, "World Economic Forum," 17 September 2018. [Online]. Available: https://www.weforum.org/agenda/2018/09/end-of-an-era-what-computing-will-look-like-after-moores-law/. [Accessed 13 April 2021].

[2] Y. Gao, S. F. Al-Sarawi and D. Abbott, "Physical unclonable functions," *Nature Electronics,* vol. 3, pp. 81-91, 2020.

[3] IRDS, "IRDS Application and Benchmarking," 2020. [Online]. Available: https://irds.ieee.org/editions/2020/application-benchmarking.

[4] Wikipedia, "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Moore%27s_second_law. [Accessed 2021 13 April].

[5] Gen-Z Consortium, "Gen-Z Consortium," [Online]. Available: https://genzconsortium.org/. [Accessed 2021 13 April].

[6] RISC-V Foundation, "RISC-V Foundation," [Online]. Available: https://riscv.org/. [Accessed 13 April 2021].