IEEE
Advancing Technology for Humanity

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS™

2023 UPDATE

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.

IEEE COMPUTER SOCIETY | IEEE | IEEE rebooting COMPUTING | IEEE COMPUTER SOCIETY TFRC Task Force on Rebooting Computing | IRDS

Table of Contents

List of Tables

List of Figures

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: Kirk Bresniker, Terry Cox, Dr. Yoshi Hayashi, Ian O'Connor, Hiroyuki Kondo, Masaaki Kondo, Tsutomu Matsushita, Nobuhiko Nakano, Takura Norikatsu, Kentaro Sano, Mitsuhisa Sato, Toru Shimizu, Satoshi Takagi, Yoshiki Yamaguchi.

IRDS Systems and Architecture Chairs Kirk Bresniker and Stephen Dukes

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. In this chapter, we identify key challenges and technology requirements by observing the patterns in play in the systems and architectures anticipated during the period under study.

To make a complex domain more approachable, we have presented information as a series snapshots taken from specific viewpoints, upon individual areas of concern. We start by looking at drivers for change and imminent trends across the industry. We consider how societal change and customer demand reveal new high value problems (Application Pull) and then look at how new architectural capabilities create emerging trends (Technology Push).

In the second section, we present a series of application-specific views, each considering a different class of system, highlighting the goals and challenges specific to each area. We have resisted the urge to seek to classify the industry into what would be an impractically over-simplified hierarchy and instead urge the reader to consider each category as an isolated narrative intended to highlight the concerns that exist within a specified application area. In practical terms, there is considerable overlap between each class of system discussed and it will become immediately apparent that challenges illustrated in one section can be seen to repeat at different scales across many of the systems considered. Our intent is to illustrate important challenges in context, in the simplest form possible, for ease of communication.

A note on the 2023 edition: this is a minor edition update. Given the technology inflection points being driven from the top down by pervasive application of discriminative and generative AI and from the bottom up by heterogeneous integration we are focusing on the major update due in 2024. For a summary of the areas of investigation for the major update see section 7.

1.1. SUMMARY AND KEY POINTS

- All types of systems considered are intrinsically linked by the **lifecycle of artificial intelligence (AI)** and this has implications for the optimization of any given type.
- Building **anti-fragile supply chains** is critical for sustainable growth in the industry. This is a fractal pattern that impacts all points in the supply chain at all scales.
- We expect **artificial intelligence** and **augmented reality** to continue to act as important drivers for the growth of all four system areas, especially in supporting the **massive underlying data analytic flows**.
- Our ability to optimize supply chain capabilities with AI are constrained by our current commercial models with respect to training data. A **conceptual shift** is necessary to move to the next level.
- Internet-of-Things (IoT) and cyber-physical systems (CPS) 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 reduced instruction set computer instruction set architecture (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, performance and cost envelopes creating opportunity for non-conventional architectures and approaches including those tailored to harvested energy.

1.2. APPROACH TO HIGHLIGHTING KEY CHALLENGES

To best understand the challenges currently presenting, we categorize systems by *design envelope*, allowing us to consider the aspects of Space, Weight, and Power (SWaP) as well as performance demanded in each category.

In subsequent discussions, we consider the following *four different categories of systems*:

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. These are ubiquitous-scale objects that pervade the environment.
2. Cyber-physical systems (CPS) provide real-time control for physical plants. Vehicles and industrial systems are examples of CPS. These can be considered 'human analog'-scale objects that fill niches previously covered by people.
3. Personal Augmentation devices such as smartphones, wearables and AR devices provide communication, interactive computation, storage, and security. For many people, smartphones provide their primary or only computing system. These are typically human-scale objects intended to augment the capabilities of an individual. In prior revisions of this chapter, we tracked this category as Mobile Devices, but this is no longer sufficient to capture the increasing diversity in this category.
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. These are macro-scale, distributed systems.

Given the wide range of systems—ranging from self-powered monolithic very large-scale integration (VLSI) devices to industrial park-sized data centers—each has its own set of market drivers, challenges and opportunities, power and thermal considerations, technology targets and metrics as described in Sections 3 through 6.

With respect to the previously defined categories, we identify the constraints of power consumption for associated processors and devices as follows:

1. Processors with less than 10mW power consumption: processors in this category target sensing devices in IoTe systems.
2. Processors with 100mW-10mW power consumption: processors in this category are mainly used for embedded systems, aiming for high performance with lower than 100mW power budget.
3. Processor with 10W-100mW power consumption: processors in this category do not require cooling systems like heat sinks or cooling fans, and are used for personal augmentation and ADAS (Advanced Driver Assistance Systems) applications.
4. Processors for high-end systems including cloud, servers and supercomputers: since there are many variations in cooling methods such as water cooling, power supply, and installation conditions, mainly in data centers, the constraints of power consumptions are different. The constraint for power consumption of systems depends on the cooling technologies and the total power supply of the facilities. For example, in the cloud, as long as the power efficiency of processors is kept good enough such that the processors can be cooled, there is no power limitation per chip. Therefore, the main criterion is the power efficiency of processors.

For the modern advanced systems, the acceleration technologies such as graphics processing unit (GPU) and field programmable gate array (FPGA) are indispensable for high performance and efficiency. While this adds burden to the software environment, the SWaP efficiency they yield is increasingly required for viable business models balancing operational cost versus development and maintenance costs.

Increasingly, these four categories of systems are being combined into entire edge-to-cloud, large scale, and 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.

Although we are considering each of these categories independently, it is important to understand that the process of building AI solutions is a super-system that spans all the above types. Typically, data will be collected from personal augmentation and IoT devices at the edge and aggregated in the Cloud in order to support the efficient training of machine learning models at scale. These models are then packaged and deployed back to cyber-physical edge systems that can perform real time inferencing upon live data from the original personal augmentation and IoT devices. Thus we must consider the impact of both operational and training phases of these systems as they become the dominant architecture in future.

1.3. 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 lines 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-1. Application Benchmark/Systems and Architecture Cross Matrix

	Cloud	IoTe	CPS	PA
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

2. Drivers

In this section, we discuss both Societal and Application Pull (i.e., changing demands and applications which define the targets that systems and architectures should fulfill) as well as Architectural and Technology Push (i.e., how the underlying technologies - from components to software stacks - are opening up new opportunities for growth).

2.1. Societal and Application Pull

2.1.1. The Transition from Decision-Support to Decision-Making Systems

Historically, systems architectures have always been built upon the presupposition that many, independent sources of record provide information to support human decision-makers who integrate this information and execute appropriate actions. It is implicitly assumed that people are at the center of the architecture, reconciling disparate and often contradictory information.

There are, however, fundamental limits to scale for people-centric organizations, with most demonstrating an S-curve for growth. Nonetheless, at the forefront of the New Economy, a growing number of enterprises are transitioning to a business model where machine learning is leveraged to replace people in the decision-making process, facilitating more reliable replication and execution of business strategy, at much larger scale.

Given the significant commercial advantages conferred by this approach, it is anticipated that the majority of businesses will be forced to adopt this strategy to remain competitive. As a result, we must expect a fundamental conceptual shift in systems architectures, as demand moves from an annual or quarterly cadence of decision-support data that is human-readable, to a real-time stream of moment-by-moment information intended to be acted upon by AI models.

Societally, this is expected to introduce disruption larger in scale than the industrial revolution, with almost all industries, roles and products being transformed through new ways of looking at existing problems with these enabling technologies. Whilst this is likely to be extremely positive for the semiconductor industry in general, we should be mindful not to see the future as 'more of the same, but faster' and should anticipate a 'Cambrian explosion' of novel applications and products, some of which should be expected to disrupt the industry itself.

2.1.2. Focusing upon Ensuring Supply Chain Anti-fragility[1]

In previous editions of the roadmap, we have focused upon aspects of reliability, availability, serviceability and security of individual systems, in line with the vertically integrated and turnkey nature of many solutions within the industry. Current

[1] *While fragile systems are harmed by volatility and dependable systems are tolerant of volatility, an "anti-fragile" system becomes stronger in the face of volatility. As external factors in the near future continue to increase volatility, anti-fragility is necessary in order to thrive in this context.*

thinking in this space has evolved rapidly, however, and it is important to recognize the intrinsically fractal nature of the challenges found in this problem domain. Recent events have placed our attention firmly upon our position in the global supply chain and helped to reveal that no matter at what scale you look, you find the same set of patterns repeating at each level.

As a result, it has become obvious that the aspects of reliability, availability, serviceability and security should be considered as holistic components of maintaining the supply chain. Rather than approaching the problem from a system component level, we should instead be mindful of how a security vulnerability in a dependency relied upon by one of our suppliers can be utilized by attackers to create reliability issues in our systems that in turn facilitate availability problems for our customers in ways that can be leveraged into adversarial attacks on a global scale.

Manipulation of the global supply chain has become a valid target for both political and financial gain in the early 21st Century. On top of this, we anticipate increasing volatility as a result of climate change, geo-political unrest and the disruption of the move to AI upon modern society. Success under these conditions requires not just a robust and resilient supply chain position, but one that should be strengthened by volatility rather than merely resistant to it.

It is important, therefore, that we view this problem as a network of systems of systems, somewhat akin to the way that the protocols underpinning the Internet were designed to automatically route around disruption and automatically recover from transient faults.

Basic tools still include the use of 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 uncloneable function (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 (traceability) 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 proactively detect anomalous behaviors across the 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 AI are co-dependent for continued advancement.

Beyond this, however, is a broader commercial problem that must be addressed if we are to fully realize the benefits of AI within the industry. The wider goals of supply chain optimization depend upon accurate and reliable machine learning models that require large volumes of data for training. In the context of the supply chain, this data is owned by many stakeholders and may contain detail that reveals proprietary IP or commercial liability. As a result, there is an inherent bias against data sharing at a level that can support the necessary modeling.

To move beyond this blocker, it will be necessary to find commercially acceptable forms of data sharing that permit improvements in supply chain whilst protecting business interests and confidential IP to the highest standards.

2.1.3. Evolution of Edge to Cloud Platforms towards Pervasive Data Analytics

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

These systems do not exist in isolation. Personal augmentation 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-1 presents these associations across nine SWaP design envelopes ranging from embedded to exascale high-performance computing (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, personal augmentation 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 circa 2018 connected ADAS vehicles generate four petabytes per day of data from their onboard sensors; that volume of data is equal to the total data volume handled at that time by Facebook (now Meta). 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 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 the investment of infrastructure and energy. At every scale of the design envelope from embedded IoT devices to exascale data centers, the numerator and denominator of this time-limited return on investment (ROI) can be affected by adoption of novel computational, memory and communication approaches from both the “More Moore” and “Beyond CMOS” roadmaps.

DESIGN ENVELOPE	SYSTEM CATEGORY				INTEGRATED TECHNOLOGIES
Beacon & Sensor Nodes	IoTe	Personal Augmentation Devices	CPS		Trusted data sources 2.5D/3D integration of sensors, memory, accelerators, computation, and comms Energy Harvesting with inducted power boost modes SRoT/Blockchain trust mechanisms
Access Point					Unified 5G/Wi-Fi access point for IoT sensor network Identity, Activity, Locality triangulation ML/AI augmented operation
Aggregation Point					Robust environmentals Edge local secure hosting of containerized workloads Static composition Smallest IT/OT (Information technology / Operational Technology) Blended Platform
Edge Hardened					Robust environmentals Legacy PXI/AXIe plus next gen modular FF Static composition Robust IT/OT Blended Platform target at several capacity points
Personal Sensor Network					Sensor Fusion Platforms for ADAS, SCM, SNS, etc. Gaming applications with milisecond or microsecond latencies Computation for security satisfying frameworks like GDPR
User Equipment					5G/6G Smartphones Wearables with ultra-low power consumption Graphics/AR/VR devices as human analog-scale objects
Single System Flex					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
Enclosure Composable				Cloud Systems	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
Rack Scale					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
Aisle/Pod Modular					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
Exascale HPC					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

Figure SA-1. SWaP Design Centers across Infrastructure System Categories

2.2. EMERGING TRENDS

Artificial intelligence [3] continues to be a primary driver for the demand for growth in capabilities. Increasing levels of AI-driven computation are being provisioned in the cloud, but many advantages exist to performing decision-making tasks in real time at the edge, so we expect personal augmentation systems, IoT edge devices, and cyber-physical systems to all include significant AI components.

In all cases, this move towards pervasive AI creates new demands on data analytics, both in the training of AI/machine learning (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 must be reflected in security and attestation features down to the lowest level devices.

We still expect AR/VR to emerge as an important application area, however there are significant challenges in developing suitably compact interface devices that will begin to drive social acceptance of this degree of obvious augmentation in social spaces. A primary challenge in this domain remains the need to synthesize data streams captured locally with geographically and contextually related live and pre-distributed data streams, within the perceptual limitations of the users, which places speed-of-light limitations for low latency computational turnaround. In the meantime, a core driver for personal augmentation systems is instrumentation, with health monitoring, safety alerting and sensory enhancement all being significant value added features to devices in this category.

The combination of low cost AR interfaces and low latency 5/6G networking, along with the mass exposure to remote working triggered by the pandemic open the path to many opportunities in telepresence. It is expected that advances in haptics, proprioception, and perhaps olfactory and gustatory transducers will be necessary to support this in comfort.

Even as IoT and CPS both continue their massive build outs, security and data privacy concerns remain a significant problem.

While we continue to describe IoT, CPS and Cloud as distinct classes, edge-to-cloud is emerging as a continuum, where a clear cycle of *'collecting data at the edge, moving it to the cloud for training, then exporting models to the edge for low latency inferencing'* is becoming visible. 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 dispersed 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, however this introduces significant challenges with respect to regulatory compliance. 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. There are emerging conflicts appearing between regional data protection and AI regulations that risk turning some regions into 'no-go zones' for CPS systems due to the impossibility of complying with proposed regulation. It is likely that we will see a gradual transition from current 'with-the-sun' Cloud based workloads to much finer grained distributed, in-network computing, which will in turn introduce a broad range of new commercial and regulatory challenges.

The radius of effective communication of data has become an increasing problem for Cloud and HPC systems. Modern data-driven applications must operate on huge datasets that cannot be held in traditional types of memories, nor be addressed directly by conventional microprocessors. Data access times to block storage can be orders of magnitude longer than access to limited, high speed, local storage, so 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 move to a chiplet based approach to integration of memory and I/O complexes into massive systems on chips (SoCs) increases the capacity of low latency storage but as a result also relegates application specific accelerators to the block mode, high-latency off-chip regime, limiting the scalability of external acceleration.

As the number of cores available within a package continues to increase, this class of problem expands into an ever wider application space, exposing a fundamental challenge for software developers who continue to have difficulty making the transition from classical, single-threaded processing to techniques that push the limitations defined in Amdahl's Law. AI is likely to be the dominant driver for demand for new devices as conventional developers struggle to conceptualize the structures and architectures necessary to utilize the full performance of modern, conventional processors.

The complex integration of cache, memory, and I/O blocks alongside the high core count in the current generation of processors continues to reveal the specter of multiple, advanced, persistent threat side-channel attacks which can be used to subvert the integrity of solutions.

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 scale 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.

2.2.1. CAMBRIAN EXPLOSION OF ARCHITECTURES

As we move from 2D to 3D, we face a new inflection point represented by the opportunity for an explosion of new architectures that exploit the potential of stacking combinations of processing, memory, acceleration and I/O in denser and more efficient forms. This makes it difficult to project many elements of the roadmap forwards as we move from a period of familiar 'general purpose computing' devices to a proliferation of exploration into new ways to address current challenges.

It is anticipated that there will be a period of searching for commercial advantage through the application of specialized architectures that offer niche benefits in specific market segments as a result of these new physical device capabilities and there is likely to be significant product differentiation before we loop around to a commercial situation where 'modular IP within devices' becomes as common as the current 'modular devices on board' model.

There are likely to be two predictable trends in this period of innovation, however. Challenges currently experienced at a data center and HPC level will migrate down to a SoC scale as the problem becomes one of moving sufficient data at low enough latencies to fully utilize the number of cores available within a given device. Since this approach is only viable for certain classes of conventional software problems, we will also see multiple approaches to optimizing AI architectures which are a workload that can potentially benefit from intrinsic parallelization without the limitations inherent when requiring humans to design massively multithreaded architectures.

2.2.2. 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 the creation of very effective models based on statistical inference. Unsupervised learning techniques can trawl huge volumes of structured and unstructured data to find correlations independently 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 whom 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 rather by the degree to which we have engineered systems to admit as much data as possible into the training regime as our physics and legal and security systems will allow.

Practically, it is important to note that the creation of models represents only about 5% of the effort needed to commercialize an ML-based product — significant work is required to improve machine learning operations (MLOps) tooling and processes to cover all the challenges that exist in this emerging space, as detailed in the MLOps Roadmap [7].

2.2.3. FROM ONE PHYSICS TO MANY

Through the first two eras (geometric and equivalent) of semiconductor scaling, there have been incredible advances in the high-level aspects of computer science — algorithms, programming languages, storage and communications technologies. However, they were all 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 expected exponential growth characteristics they were not admitted.

Even the obvious defects in security source to the conceptual basis of software models based on the 1960s threat landscapes failed to be fixed at the source because of the 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, e.g. quantum computing, along with cryogenic computing, has emerged as a particular focus area. However, it is not the only one - other areas include novel switching technologies, such as:

- Faster, more efficient switches using carbon nanotubes or 2D materials in the switch channels;
- Bringing computing resources to where the data is stored to reduce communication overhead with In-memory computing;
- Using the rich and orthogonal properties of light to densify data representation as well as resonance and interference phenomena to implement complex mathematical operators in silicon photonic computing;
- Adiabatic and reversible computing that operate at the limits of thermodynamic information theory;

- Neuromorphic and brain-inspired computing that draw inspiration from biological systems but, much as with aerodynamics, use materials and energies not available to their biological analogs, and
- Networks of organic and inorganic materials whose behavior calculates desirable functions at SWaP breakthroughs; as systems created in our own image that are designed primarily to host intellect that offer computation as a byproduct of intelligence.

2.2.4. 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 born 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 personal augmentation devices, that ratio is expected to shift to as much as 75% of enterprise information will *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 other words, in all things intelligent).

There are two forces that keep data at the edge—physics and law. The exponential growth of recorded data, which currently doubles every two years, means that even with the advent of 5G communications and massive communication 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 never be (at least in the foreseeable future) 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[2] continue to offer the protections that they strive to, when the vast majority of data will never be in a data center, and 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 in proximity 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 cryptography 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 that 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.
- Specific examples of these challenges are addressed in more detail in the Factory Integration chapter.

2.2.5. FROM IMPERATIVE TO DECLARATIVE

Imperative control systems rely on the enumeration of conditionals and responses, i.e. the classic if-then-else diamonds of flowcharts. The problem with imperative control is that the systems that we are creating—social, technical and economic—are too 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. Consequently, this means overheads and the inefficient use of resources, whether it is compute resource allocation, spectrum allocation 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

[2] *General Data Protection Regulation*

learning and autocorrelation could naively, blindly discover those correlations humans discount because it cannot presume it knows better.

2.2.6. From Scarce Memory to Abundance

A decade after Alan Turing created the mathematical theory of computation, John von Neumann was realizing that theory as an operational feat of engineering in his 1946 outline of EDVAC[3]. 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 grow 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.

2.2.7. 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 can obtain a value of the state function of the enterprise, but it is at best days, most likely weeks old and represents 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 of fraud or inefficiency. But we can also extend 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").

2.2.8. 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

[3] *Electronic Discrete Variable Automatic Computer*

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.

2.2.9. 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-increasing levels 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 is 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 multiple foundries of relatively competitive logic processes, this again enfranchises an ever increasing number of innovators everywhere.

2.2.10. 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.

2.3. ARCHITECTURAL TRENDS

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

This roadmap provides separate analysis of 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.

2.3.2. 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.

2.3.3. EDGE TO CLOUD SERVICE MESHES

Whether public or private, cloud systems today offer compute, storage, networking infrastructure deployable via Application Programming Interfaces (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 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.

3. System Category—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 an on-demand, as-a-service consumption model. While they had their origins in web services; 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. Starting with commercial off-the-shelf (COTS) systems, cloud infrastructure progressed to custom loading of standard systems to purpose-built at the motherboard level. Today's cloud-native compute, storage and networking infrastructure features bespoke processor 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 AI/machine learning (ML) integrated into so many applications, the demand for accelerated floating point computation is more universal and all applications are being dominated by operational and capital costs of data movements at scale. In all cases, the general trend is also to utilize this 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.

3.1. 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 application domains.

3.2. 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, the breadth of traditional enterprise and 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.

3.3. Design Envelope Considerations

This category encompasses the largest and most power-hungry of systems, where we are more concerned with scalability than physical size as a constraint. At this macro level, we are concerned with geo-political issues regarding the siting of essentially immovable systems near to the resources needed to operate them, whilst mitigating the latency issues associated with large, distributed systems. Historically, we have tended to be less concerned with power issues as a driver within this envelope, however we must now recognize that the collective scale of these systems represent a global challenge. 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. More significantly, continuing growth in demand for Cloud resources means that we

are still on a trajectory that shows data center workloads consuming all of the world's current electricity generating capacity, within the next 15 years. A reduction in transactional energy cost (J/bit) of at least three orders of magnitude is required in order to push that out by another decade.

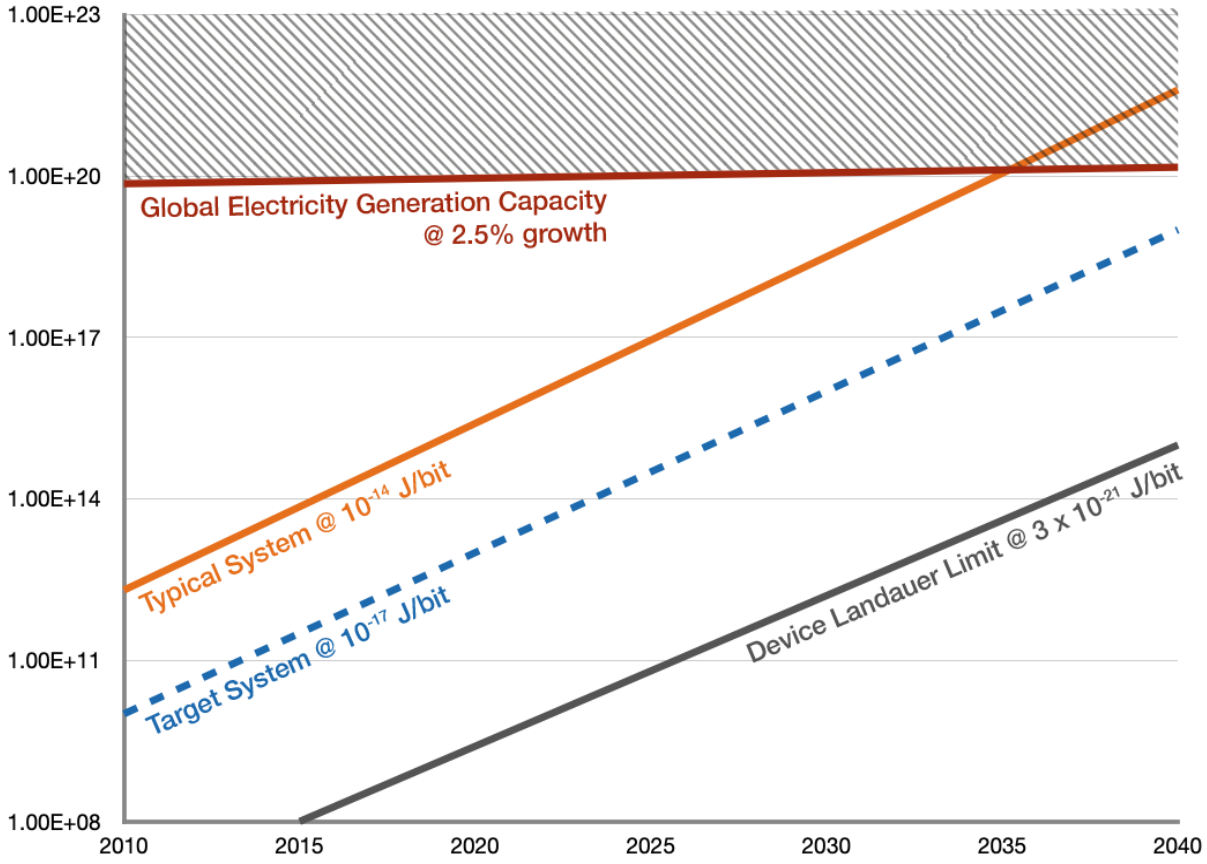

Figure SA-2. Cloud Energy Usage based upon BIT Transition Cost

3.3.1. TRENDS OF PROCESSORS FOR HIGH-END SYSTEMS

The power consumption of processors used for high-end systems such as cloud, servers, and high-performance computing ranges from tens of watt to a few hundreds of watt depending on the performance requirement of each processor chip and its cooling capacity. As these systems are composed of a cluster of multiple compute nodes, total system performance is usually a crucial metric rather than a single processor performance. Since the total system performance depends on the number of installed nodes which is mainly limited by total power and/or cooling budget of a data center facility, power efficiency, sometimes measured by performance per watt, is used to be a critical factor of the processors.

These types of processors have been and continue to be evolved in a different way to match with different workload characteristics: that needs lower latency, which needs throughput, and that needs compute capability. We assume three classes of high-end processors[4].

1. Latency sensitive processors: these processors are optimized for lowering any aspect of latency, usually prioritizing single thread performance. They have a rich out-of-order execution resources with advanced branch prediction and cache prefetching mechanisms, a deeper cache hierarchy with relatively larger a last level cache, and large main memory with standard double data rate (DDR) technologies to accommodate a large dataset. Typical workload of cloud and server systems with these processors includes network services, database processing, and graph processing.

[4] *This classification is originally from NGACI white paper (https://sites.google.com/view/ngaci/home)*

General purpose server class processors such as Intel Xeon, AMD EPYC, and Amazon Graviton are representative examples of these processors.

2. Throughput oriented processors: these processors are specially optimized for compute and data movement bandwidth, prioritizing parallel processing efficiency. They emphasize data level, thread level, and core level parallelism of the workload to achieve higher performance. In order to integrate more number of compute units, out-of-order execution resources and memory hierarchy are simplified. Integrated memory modules with high bandwidth memory (HBM) or graphics DDR (GDDR) are usually used to increase memory to processor bandwidth. Typical workloads of this type of processors are graphics and supercomputing applications. Representative examples of these processors are GPUs such as NVIDIA A100, A64FX used in Supercomputer Fugaku, and vector processors such as SX-Aurora Tsubasa.
3. Compute centric processors: these processors are specially designed for achieving very high compute throughput with less emphasis on main memory bandwidth. They usually have a large SRAM on-chip memory assuming dataset of the current workload fits in the on-chip memory. They can achieve very high compute density utilizing high bandwidth data supply from on-chip SRAM to functional units. Deep learning and inference applications are the main target of this type of processors. Representative examples of these processors include Cerebras WSE CS-2, Tesla D1, Esperanto ET, Graphcore Colossus MK2 GC200 IPU, Groq TSP, SambaNova SN10.

We forecast the technological trend of the latency sensitive processors and throughput oriented processors as shown in Table SA-2 and Table SA-3. In the forecast, we assume multiple dies are tightly connected in a package with chiplet technologies.

Table SA-2. Technology Trends of Latency Sensitive Processors

	2019	2022	2024	2026	2028	2030	2032	2034
# chiplet per socket	4-8	8	8	8	8-16	8-16	8-16	16-20
# core per chiplet	8	8	8-12	8-16	8-16	16-24	16-32	16-32
# core per socket (max)	64	64	96	128	256	384	512	640
Processor base frequency (GHz) (for multiple cores together)	2.2-3.0	2.5-3.3	2.8-3.4	3.0-3.5	3.2-3.6	3.4-3.7	3.4-3.7	3.4-3.7
Core total vector length	1024	1024	1024	1024	2048	2048	2048	2048
L1 data cache size (in KB)	36	40	40	42	42	44	44	44
L1 instruction cache size (in KB)	48	96	96	128	128	160	160	160
L2 cache size (in MB)	1	1.5	2	2	2	2.5	2.5	2.5
LLC cache size (in MB)	64-128	64-800	128-1024	256-1536	256-2048	512-4096	512-4096	512-4096
# of DDR channels	8 (DDR4)	12 (DDR4)	12 (DDR5)	12 (DDR5)	16 (DDR5)	16 (DDR6)	16 (DDR6)	16 (DDR6)
DDR bandwidth (TB/s)	0.20	0.31	0.61	0.76	1.02	1.1	1.2	1.2
DDR size per socket (in TB)	1.0	3.0	4.5	6.0	8.0	10.0	12.0	12.0
Socket max TDP (Watts)	280	300	400	450	600	600	700	700

L1=level 1 cache; LLC=last-level cache; Fabric=PCIe or new fabric (e.g. CXL); TDP=total power dissipation.

Table SA-3. Technology Trends of Throughput Oriented Processors

	2019	2022	2024	2026	2028	2030	2032	2034
# chiplet per socket	1	1	4	8-16	8-16	8-16	8-16	8-20
# SM core per chiplet	108	132	36	36	42	48	54	54
# core per socket (max)	108	132	144	288-576	336-672	384-768	432-864	432-1080
Processor base frequency (GHz) (for multiple cores together)	1.2	1.4	1.4	1.4	1.6	1.6	1.6	1.6
# of HBM ports	6 (HBM2)	6 (HBM3)	8 (HBM3)	8 (HBM3)	8 (HBM4)	10 (HBM4)	10 (HBM4)	10 (HBM4)
HBM bandwidth (TB/s)	1.6	3	4	4	6.6	8	10	10
HBM size per socket (in GB)	40	80	256	512	800	1024	1024	1024
Socket max TDP (Watts)	400	400-700	400-700	440-700	500-700	500-700	500-700	500-700

3.3.2. ACCELERATION TECHNOLOGY

3.3.2.1. GPU ACCELERATORS

The performance of HPC/DC systems is increasingly dependent on accelerators, such as GPUs, Tensor Processing Units (TPUs) and AI processors, rather than general-purpose CPUs. It is therefore necessary to project a roadmap of accelerators with their characteristics and performance, in order to project architectures of future systems in the HPC and DC categories.

Here, we project technically feasible performance metrics of future accelerators in the cloud, DC and HPC domains, assuming that we will pursue computational performance to any extent possible in the category. We target high-end GPU and AI processors consuming 100W or more. In this projection, we predict the peak performance of an accelerator chip based on the same architecture as that of a baseline product in 2021. We assume that the peak performance is restricted by one of the scaling of transistor density, scaling the external memory bandwidth, and scaling of transistor's power consumption. Table SA-4 shows the estimated performances of high-end GPU with HBM[5] memories, for operations of FP64, FP32, BF16[6], FP8, and INT8.

[5] *High Bandwidth Memory (JEDEC standard)*

[6] *Brain floating-point format: number encoding format occupying 16 bits, equivalent to a standard single-precision floating-point value with a truncated mantissa field.*

Table SA-4. Performance Trends in High-End GPUs with High Bandwidth Memory (HBM)

	2022	2024	2026	2028	2030	2032	2034
FP64 [TFLOPS]	30	32	40	42	44	47	48
FP32 [TFLOPS]	60	70	80	84	89	94	96
FP16/BF16 [TFLOPS]	2000	2290	2640	2800	2960	3120	3200
FP16 [TFLOPS]	2000	2290	2640	2800	2960	3120	3200
INT8 [TOPS]	4000	4570	5280	5600	5920	6240	6400
HBM bandwidth (TB/s)	3	4	5	6.6	8	8	8

3.3.2.2. FIELD PROGRAMMABLE GATE ARRAYS (FPGA)

Programmable Logic Devices (PLDs) have developed from glue logic to the current Field Programmable Gate Array (FPGA). As the awareness of programmability and customizability grows, FPGAs have attracted more and more attention and are entering wider markets such as communication, storage, data center, sensing, and personal augmentation devices. Thus, the FPGA market has registered a compound average growth rate (CAGR) of around 4% from the beginning of the 21st century. FPGAs have evolved as custom-computing logic and are used in the replacement of glue logic and specialized computing, including rapid prototyping. In addition, high-performance computing is targeted because advanced semiconductor manufacturing has expanded the variety of FPGA architectures in this decade. The tangible examples in high performance and large-scale computing are Big Data, financial applications, high-frequency trading, network switching, large-scale AI applications, high-speed and low-latency large volume storage, etc.

The SA-4.5.2 shows the trends of high-performance FPGAs expect HBM and communication based on many parallel I/O pins. However, this will be affected by multi-die packaging technology. Future system design with HBM FPGAs may face a critical bottleneck because die-to-die interconnections strongly restrict the circuit design.

Key metrics include primary logic element (LUT), digital signal processing unit (DSP), on-chip distributed SRAM block (BRAM), off-chip RAM banks (HBM stacks) and the bandwidth (HBM bandwidth), and TDP (Thermal Design Power).

Table SA-5. Trends for FPGAs in High Performance Applications

	2022	2024	2026	2028	2030	2032	2034
#LUT (M)	8	12	12	17	17	22	22
#DSP (K)	10	15	15	22	22	30	30
BRAM size (Gb)	1	1.5	1.5	2.2	2.2	3	3
HBM stacks	2	2	2	4	4	4	4
HBM bandwidth (TB/s)	0.8	1.6	3.3	6.6	6.6	13.1	13.1
TDP (Watts)	133	126	179	174	174	180	180

3.4. METRICS FOR CLOUD PROCESSORS

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

Table SA-6. System Category Cloud—Key Processor Metrics

	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034
# cores per socket	64	64	96	96	128	128	256	256	384	384	512	512	640
Processor base frequency	2.5-3.3	2.5-3.3	2.8-3.4	2.8-3.4	3.0-3.5	3.0-3.5	3.2-3.6	3.2-3.6	3.4-3.7	3.4-3.7	3.4-3.7	3.4-3.7	3.4-3.7
Core vector length	1024	1024	1024	2048	1024	1024	2048	2048	2048	2048	2048	2048	2048
FLOPS per socket (TFLOPS)	5.1	5.1	8.6	8.6	12.3	12.3	52.4	52.4	83.6	83.6	111.4	111.4	139.3
HBM ports	6	6	6	6	6	6	6	6	6	6	6	6	6
HBM bandwidth (TB/s)	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	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	100	100	100	100	100
Socket TDP (Watts)	300	300	400	400	450	450	600	600	600	600	700	700	700

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

4. System Category—Personal Augmentation

Personal augmentation devices integrate computation, communication, storage, capture and display, and sensing. These systems are highly constrained in both form factor and energy consumption. As a result, their internal architectures tend to be heterogeneous. Cores in modern personal augmentation 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.

4.1. Market Drivers

Personal augmentation devices provide multiple use cases: telephony and video telephony; multimedia viewing; photography and videography; email and electronic communication; positioning and mapping, authenticated financial transactions, health and fitness monitoring, personal safety and environmental warning. Current and upcoming market drivers include: gaming and video applications; productivity applications; social networking; augmented reality and context-aware applications, and mobile commerce. Personal augmentation devices already make use of AI technologies such as personal assistants. Deployment of AI on and through personal augmentation devices will accelerate.

4.2. Challenges and Opportunities

Personal augmentation systems present several challenges for system designers. Multimedia viewing, such as movies and live TV, have driven the specifications of smartphone 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 personal augmentation 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 personal device to personal device and link to low-latency, distributed-edge compute infrastructure as well as multi-cloud global infrastructure. To date, personal augmentation 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 personal augmentation 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.

4.3. Design Envelope Considerations

This category is characterized by the requirement to be able to comfortably carry or wear the system for long periods of time. This constrains the size and heat output to fixed limits. Within that physical envelope, 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 personal augmentation devices may create thermal challenges that must be considered to ensure a comfortable experience for users.

4.3.1. Trends of Processors for Personal Augmentation

The processors used for personal augmentation systems need high processing performance, but cannot be equipped with cooling systems like heat sinks or cooling fans. The performance is restricted by power consumption limits due to thermal dissipation. Heat, generated by chips, must be radiated to ambient air via conduction through packaging. The power consumption of this type of processor must be in the range from 100mW to 10W. We assume that the power consumption limits of processors remain constant to forecast the trends. "DMIPS[7]" is used as an indicator of the performance of processors since "DMIPS" or "DMIPS/clock frequency" have been published for many central processing unit (CPU) cores, so it is easy to forecast for long term trends.

We forecast the trends by estimating DMIPS/Hz/Core of benchmark CPU, by extrapolating the trend observed before 2020 until 2034 , and estimate DMIPS trend for SoCs used in personal augmentation applications. Our assumptions are that, throughout this decade, (i) the number of Cores will increase from 8 to16 and to 32, and (ii) the clock frequency will be increased from 2.8 GHz to 3.0 GHz, 3.2 GHz, and to 3.4 GHz.

[7] *Dhrystone MIPS (millions of instructions per second measured during the execution of Dhrystone benchmark)*

Table SA-7. Trends of SoC processors for Personal Augmentation Devices

	2022	2024	2026	2028	2030	2032	2034
Core performance (DMIPS/MHz/Core)	18.5	20.8	23.2	25.5	27.8	30.1	32.4
Number of Cores	8	16	16	16	32	32	32
Maximum frequency (GHz)	2.8	3.0	3.2	3.2	3.4	3.4	3.4
Performance (kDMIPS)	421	1,000	1,186	1,304	3,025	3,277	3,530
Process node (nm)	3	3	2.1	1.5	1.5	1.0 eq	0.7 eq

4.4. Metrics for Personal Augmentation Processors

The table SA-8 shows requirements for personal augmentation devices. 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-8. Personal Augmentation Technology Requirements

	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034
# CPU cores	8	8	16	16	16	24	24	24	32	32	32	32	32
Core performance (DMPIS/MHz/Core)	18.5	18.5	2.8	2.8	23.3	23.3	25.5	25.5	27.8	27.8	30.1	30.1	32.4
# GPU cores	32	64	64	64	128	128	128	256	256	256	512	512	512
Maximum frequency (GHz)	2.8	2.8	3.0	3.0	3.2	3.2	3.2	3.2	3.4	3.4	3.4	3.4	3.4
Number of cameras	4	4	4	6	6	6	8	8	8	8	8	8	8
Camera resolution (MP)	18	18	20	20	20	24	24	24	24	24	24	24	24
Number of sensors	10	10	12	12	12	12	12	16	16	16	16	16	16
5G/6G Max data rate (Gb/s)	5	7	7	7	10	10	10	20	20	20	50	50	50
Wi-Fi Max data rate (Gb/s)	30	30	30	30	30	30	30	30	30	30	50	50	50
Board power (mW)	5900	6190	6500	6830	7170	7530	7900	8300	8715	9150	9610	10090	10590

5. System Category—Internet-of-Things Edge Devices

An 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.

5.1. 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.

5.2. 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.

5.3. Design Envelope Considerations

IoTe must be designed to provide low total cost of ownership. This category is somewhat more abstract than the previous ones, as an IoTe device may or may not be portable, thus may or may not be constrained by size or lack of access to fixed power. Considering the sheer number of devices expected to be deployed, however, it should be expected that in the majority of cases, power budgets will be very low. 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.

5.3.1. Trends of Processors for Internet-of-Things Edge

The processors in this category are used for IoTe systems. From the viewpoint of power consumption, IoTe systems are classified into two types: standard systems with power consumption of 10 mW or less, and low power systems with power consumption of 1 mW or less. In IoTe, acquisition of sensing data is a functional requirement that is a "feature" explicitly requested by a client. Power saving, however, is a non-functional requirement that refers to all quality-related items such as usability, performance, scalability, and security. Device metrics are related to the elimination of the bottlenecks of these functional and non-functional requirements.

An IoTe system that acquires sensing data consists of sensor nodes that collect data, gateways that aggregate those data, clouds that convert information into big data, and display devices showing analysis results. The main bottleneck in the utilization of IoTe is the power required to operate the sensor node. Due to power constraints on the sensor node, the number of semiconductor devices, such as multi-cores and accelerators, is limited. In a low-power IoTe system, the number of CPUs and sensors is limited to the minimum number. The power consumption increases with data-measurement frequency (samples/second), data size (Bytes), and communication interval (transfers/second). In cases where wired power supply is difficult, wireless sensor nodes are used, but it is essential to reduce the frequency of wireless communication and power consumption of IoTe terminals.

A wireless sensor node mainly consists of a power-supply unit, sensor unit, micro controller unit (MCU) unit, and wireless unit. A battery and photovoltaic power-generation device and/or energy harvesting device are mounted on the power-supply unit together with a power-supply stabilizing circuit. In cases where the performance of the installed battery or energy harvester is insufficient, the range of IoTe utilization is limited. Therefore, the scope of IoTe application will be expanded by improving the performance of batteries and energy harvesters. The device metrics for the power-supply section of IoTe are improved battery life (B), battery hybridization (B+H) and energy harvesting (H), multimodal harvesting (H1 + H2), and zero battery harvesting.

Various cameras, and sensors for vibrations, currents, infrared rays, or chemicals are integrated in the sensor node. As the types of sensors increase, so the use cases of IoTe will expand. Hence, the number of sensors per device will increase over time in a standard IoTe system and smartphones with power consumption of 10 mW or more. In a low-power IoTe system however, low-power operation is prioritized, so the number of sensors is limited to one.

The MCU executes digital processing of sensing data. In a standard IoTe system, the number of CPUs used in the MCU will increase over time. However, the number of CPUs in a wireless sensor node is smaller than that in a wired sensor node. In a low-

power IoTe system therefore, as mentioned earlier, the number of CPUs will remain limited to one over time. The MCU is required to improve the frequency that meets the data-processing requirements of IoT but also reduce the leakage current that leads to the improvement of battery life. Conventional technologies, even with multi-threshold voltages and power-gating approaches, will not be sufficient. Therefore, it is expected that the application of SOI (Silicon on Insulator) devices (to enable dynamic body-biasing and flexible means of equating device performance with leakage requirements) and non-volatile memory (to enable aggressive power-down strategies and fast wakeup times, as well as novel non-volatile logic and In-Memory-Computing designs) will expand over time as an option to reduce leakage current.

A wireless sensor node is equipped with a wireless semiconductor chip used for communication with the gateway. In IoTe, Wi-Fi, LPWA, etc. are selected as low-power wireless specifications, and communication capacity is reduced under Tx/Rx energy (μW/bit) and constant power consumption. Because of cost priority, devices that are several generations behind are often used instead of cutting-edge devices. It is necessary to design a wireless sensor node so that the power consumption is equal to or less than the desired value. For that purpose, the current profile and energy consumption during operation of the entire wireless sensor node must be quantitatively evaluated through experiments.

5.4. 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-9. Internet-of-things Edge Technology Requirements

	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034
CPUs per device	2	4	4	4	4	6	6	6	8	8	8	8	8
Maximum CPU frequency (MHz)	310	315	320	325	330	335	340	346	351	360	363	369	375
Energy source (battery, 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
Tx/Rx power/bit (µW/bit)	0.085	0.057	0.038	0.025	0.017	0.011	0.0075	0.0050	0.0033	0.0022	0.0015	0.0015	0.0015
Battery operation lifetime (months)	9	9	9	12	12	12	12	18	18	18	18	18	18
Deep suspend current (nA)	32	27	23	20	17	14	12	10	9	8	7	7	7
Sensors per device	8	8	12	12	12	16	16	16	16	16	16	16	16

6. System Category—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 known as operational technology (OT), where ruggedness, extended environmental capabilities, low-cost have historically 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.

6.1. Market Drivers

Market drivers include automotive and aerospace vehicles, autonomous vehicles/unmanned arial vehicles (UAVs), 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.

6.2. 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 must flow out from 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.

Practical applications of machine learning represent highly asymmetric demands upon resources as very large volumes of data must be aggregated centrally for training, requiring very high bandwidth data pathways, but typical applications for inferencing with these models need very low latency operation in the field, requiring a very dense and energy efficient processing capability close to real time data sources.

It is anticipated that at least a two orders of magnitude increase in performance will be necessary to properly fulfill immediate aspirations in this segment, with a longer term goal being to meeting and exceed the characteristics of the human brain: greater than 10^{18} FLOPS in a volume of around 1 liter, with a mass of around 1kilogram and a power consumption of around 100W.

6.3. Design Envelope Considerations

By their nature, cyber-physical systems exist in the field, close to the real world applications that they interact with. This category is therefore physically constrained in package envelope, especially for systems that are required to be able to move autonomously. Power and heat envelope constraints are context dependent. 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. Cyber-physical systems typically perform roles that are of a critical nature and therefore also require uninterruptible power supplies that may constrain total power budget. Many cyber-physical systems present extreme temperature, electromagnetic and vibrational environments in which the electronics must operate.

6.3.1. TRENDS OF PROCESSORS FOR CPS

The processors used as MCU's for ADAS or automotive infotainment systems, such as Level-2 ADAS electronic control unit (ECU) or Car Navigation systems need high processing performance, but often cannot be equipped with effective cooling systems like deep heat sinks or strong cooling fans. The power consumption must be in the range from 100mW to 10W. The processors targeted to Fully Autonomous Driving systems belong to the different category. We assume that the power consumption limits of processors remain constant, and "DMIPS[8]" is used as an indicator of the performance of processors since "DMIPS" or "DMIPS/clock frequency" have been published for many CPU cores, so it is easy to forecast for long term trends. We forecast the trends by estimating DMIPS/Hz/Core of target CPU, by extrapolating the trend observed before 2020 until 2034. And we assume that throughout this decade, CPU cores used for personal augmentation applications will be applied to automotive applications 5 years later, with lower clock frequency, and the number of cores will increase from 8 to 16, and the clock frequency will be increased from 1.0 GHz to 1.2 GHz.

Table SA-10. Trends of MCU processors for ADAS

	2022	2024	2026	2028	2030	2032	2034
Core performance (DMIPS/MHz/Core)	9.5	15.1	17.3	19.7	22.0	24.3	26.6
Number of Cores	8	8	8	8	16	16	32
Maximum frequency (GHz)	1	1	1	1.2	1.2	1.2	1.2
Performance (kDMIPS)	76	121	139	189	422	467	1,023
Process node (nm)	10	7	5	3	2.1	2.1	1.5

The processors used for embedded systems operate on a commercial power source but consume as little power as possible. On-chip flash memory microcomputers in this category are used with commercial power supplies, and may operate for several weeks with a large capacity battery. The power consumption is usually in the range of 10mW to 100mW. Since microcomputers are used in embedded devices, the performance of the processor itself does not lead to product value, so there is a tendency to select one with sufficient performance to establish an application from the viewpoint of power consumption and cost. Therefore, state-of-the-art processes are not always used. Power consumption is adjusted by scaling the operating frequency.

For embedded systems, on-chip non-volatile memory (e.g., Flash, MRAM, RRAM, FeRAM) is required. The process (technology node) is determined by cost. Even if emerging memory is included, nonvolatile memory (NVM) scaling laws are harder to predict. Utilized processes are limited to 65, 55, and 40 nm. Recently, an on-chip flash memory using a 28 nm process was introduced. In this analysis, even considering the scenario of a future disruptive NVM technology, the limit of miniaturization will not go beyond the 14 nm process in the next 10 years. This is several generations behind the current personal augmentation processor process. Taking this scenario, it is projected that the category of embedded processors will not evolve significantly.

6.3.2. CPS DEVICES

Cyber-physical systems should be considered with particular attention in future system architectures. This section considers the device requirements for CPS.

CPS constantly rotates the sense/think/act loop and handles a large amount of data. Ultimately, power saving and scalability of 10,000 times or more are required. A large number of leading-edge semiconductor devices are used in advance in the CPS system. For energy-saving, neuromorphic computing is applied using a model inspired by the fact that the human brain requires less than 0.0001% of the amount of energy consumed by a supercomputer. The neuromorphic semiconductor chips will evolve from the original digital type to the analog type. Furthermore, in order to increase the scalability, attention is focused on the annealing technology for the overall optimization of supply chains, social infrastructure, and smart cities. Annealing chips for non-silicon quantum computing and silicon quantum computing will be developed in parallel, and the range of applications for scheduling like personnel shift, and optimization of mobility operation management will be expanded. 3D chiplet technology that integrates them heterogeneously at the package level will be developed for clouds and edges used in CPS.

CPS is applied to various applications such as digital transformation (DX), green transformation (GT), Well-being, FinTech, and supply-chain management (SCM), starting with the integration of indirect operations such as accounting, procurement, and personnel. In the physical space, evolution occurs in the collection of information not only from things but also people. In

[8] *Dhrystone MIPS (millions of instructions per second measured during the execution of Dhrystone benchmark)*

cyberspace, the evolution to the metaverse will occur. Some of the computing for CPS can be done by simply improving software processing. However, to achieve mission-critical control that requires real-time performance and high position accuracy, the number of cases in which new hardware is required will increase over time.

CPS can be regarded as systems of systems (SoS), i.e. combinations of various systems. Focusing on power consumption, they consist of the low-power and standard IoTe systems, semiconductor chips of 100 mW to 10 W or less, and HPC (High Performance Computing) of 10 W or more. Therefore, it is necessary to consider the power-reduction requirements of various devices and systems in an integrated manner. If personal augmentation devices like wearables and implantable devices are used in metaverse, they may be components of CPS. In a device that comes into contact with a person, there is an additional requirement that the power consumption be 5 W or less to suppress the generated heat that cannot be air-cooled.

In CPS, it is important to assume distributed computing through cooperation between the edge and cloud as the main solution to suppress the power explosion caused by the increase in information. To streamline distributed computing, OT and IoTe devices are placed in the physical space, which will be connected to clouds and HPCs by wired or 5G/6G wireless networks.

The main device metrics required to enable the diversified edge-cloud cooperation in CPS are the number of devices (CPUs, GPUs, and FPGAs) and cores per device. The number of devices is defined as the number used for the assumed SoS. CPUs per device is defined as the number of CPUs used in the SoC (System of Chips).

The number of dedicated CPUs and dedicated accelerators customized for individual applications is also an important metric. Examples where these are used are 4K/8K and point cloud image capture and processing for 3D images and XR free viewpoints. Biometric for IDaaS is an important aspect of CPS security. If online games are associated with metaverse, it may be a component of CPS. As these CPS applications increase, the number of dedicated chips will increase. There are various requirements for the number of dedicated accelerators for the edge, and the rate of increase is larger than that for the cloud. Dedicated devices for the edge can be classified as 10 mW or less (IoTe) and 100 mW to 10 W or less (personal augmentation and ADAS) from the viewpoint of power consumption. The growth rate of the latter is larger than that of the former.

The number of dedicated devices for meeting non-functional requirements such as security is also important. Security is achieved by a combination of various technologies from camera authentication to biometric authentication. In addition to speeding up and improving image recognition and detection, security chips are required to have functions for visualizing results and obtaining customer satisfaction.

6.4. Metrics for CPS Processors

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

Table SA-11. Cyber-physical Systems Technology Requirements

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

7. CONCLUSIONS, RECOMMENDATIONS AND FUTURE PLANS

Significant focus, this year, has been placed upon Supply Chain. Whilst the industry as a whole has arguably the best supply chain processes of any, it is clear that a series of unanticipated global events have caused significant problems, with ramifications that have impacted all industries downstream. It would be unwise to view this experience as a transient event followed by a return to business as usual, however. As we move further into the 'vertical' phase of Moore's law growth, rapidly increasing change and volatility are to be fully expected, with the associated socio-political change that new technologies always bring. Consideration should be given to volatility as a regular and ongoing problem.

A large part of this change will be driven by the move to AI-based decision-making as a path to increased scale, bypassing the limits of conventional business structures. The challenges surrounding the ownership and control of the data necessary to train these models will likely outweigh the technical challenges in implementing the systems themselves, so thought is necessary to resolve the commercial concerns in this space.

Personal augmentation systems have emerged globally as a new driver for consumer technology, with uptake progressing beyond the original communications functions of mobile phones to encompass a wide range of enhancements to human capabilities, including augmented reality, the extension of vision into infrared, the provision of accurate contactless temperature and distance measurement, 3D scanning, sensing of multiple health parameters, fall detection and alerting etc. Provision of secure financial transactions continues as an important emerging application for personal augmentation systems.

The self-similar patterns of security and confidentiality can be seen to repeat at all scales from the smallest of IoT devices to the interactions that flow through our supply chain itself. A consistent and reliable approach to these aspects universally will unlock value in many areas.

As Cyber-physical systems continue to proliferate in scenarios where they are required to control physical systems, the reliability of these systems will increasingly fall into the purview of government regulation. As this regulation is currently being driven by demands very different from the concerns of the industry, attention should be given to ensuring that incompatible regulation does not become problematic for some markets.

We have identified several recommendations, as follows:

- Holistic security and privacy remain critical to all our system areas — this drives ongoing 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.
- Safely unlocking the data necessary to train AI solutions within the industry is a commercial challenge that conflicts with our existing approaches to protecting IP and new thinking will be required in order to realize the value promised by CPS and anti-fragile supply chain.
- AI/ML, IoT and CPS all present drivers for massive increases in data flow from edge to cloud. As this grows, the problems of information transfer bottlenecks within systems will transition from the HPC domain down through all the design envelopes. Utilizing the performance available in new devices will become harder as increasingly complex software approaches are necessary to face these challenges so it will be necessary to disseminate appropriate knowledge and tooling to allow customers to take advantage of these capabilities.
- Ownership of production AI/ML/CPS systems introduces new challenges and so consideration, and ongoing development of best known methods in MLOps is an area for further pre-competitive collaboration that will lift all boats.
- Energy harvesting is a key technology to enable the growth of IoT edge devices. Continued work is required.
- Augmented reality will create further demand for computation, communication, sensing, and display on personal augmentation devices and will be a class of applications which will span *ad hoc* low latency personal device to personal device and personal device to edge infrastructure.

Looking forward to the 2024 edition, the Systems and Architecture team are planning a major update which investigate the following themes:

- Accelerators are a cross-cutting technology impacting all system categories and design envelopes - IotE, CPS, Cloud, Personal Devices. Despite the continuous increase in design, validation, and manufacturing costs associated with leading edge processes nodes, the ability of accelerators to solve for size, weight, power constraints and yield sufficient return on investment when accounting for capital and operational cost is increasing across all system categories. These are no longer application specific functions in standards CMOS process, an increasingly wide variety of accelerators – in-memory, analog, quantum, and photonic approaches are being explored alongside GPU,

FPGA, DPU (data processing unit), APU (accelerated processing unit) CMOS based approaches. For the increasingly exotic accelerators technologies, the physical integration (which might need to accommodate cryogenic temperatures, hard vacuums, or electromagnetic isolation), the cyber-physical control integration and finally the productivity integration need to be considered. New industry standards such as Compute Express Link (CXL) as well as proprietary equivalents are reducing the design cost for high performance integration of accelerators both in direct connection to CPUs and GPUs as well as increasingly complex fabric topologies.

- Driven by a convergence of societal, technical, regulatory, and economic forces, traditional isolated goals for power efficiency are evolving into more holistic sustainability practices which incorporate whole-of-stack analysis and optimization. This often represents a fundamental shift in approach where the traditional focus on power as measured as the instantaneous rates of energy consumption in Watts is replaced by integration over time to yield total energy consumption and its associated sustainability impact measured in Joules. This will demand changes in design methodology, operational telemetry, and control systems.
- At the exascale and beyond, networks are growing increasingly complex and the endpoints they support are growing increasingly heterogeneous. These networks need to support an increasingly diverse set of communications access patterns and virtualized isolation all while, in some applications, being the fundamental scaling limiter. With the increasing complexity of both electronic and photonic physical interfaces and the latency from computational cores in CPUs, GPUs, and accelerators from the networking interfaces, in-network computation is a promising approach that can offload some workload and control plane tasks to positions in the network interfaces and the switches and affect in-flight data transformation. With the advent of photonic accelerators, this could also permit in-flight data transformation without resorting to repeated optical/electrical conversion with its associated energy costs and potential security vulnerabilities.
- Issues of Trust, Privacy, Security, and Authenticity continue to escalate in importance and attacks on the supply chain continue to grow in sophistication and depth of penetration into hardware, firmware, and software design and manufacturing cycles. All of this is occurring while the transition to post-quantum cryptography is under active review by the US NIST and could move from NIST recommendation to FIPS regulation within the next several years. Satisfying these increasing deep requirements will require solutions that reach back to factory infrastructure and data integration to provide the basis for provenance and attestation, but that basis will need to be extended through the supply chain to measurable confidence in outcomes.

Beyond these trends, in terms of the information tracked in next edition, we will be evaluating the following:

- More detailed quantitative measures to SWaP design envelopes to better characterize the fitness of novel acceleration, energy, and heterogeneous integration techniques across the system categories and design envelopes.
- Further reflect shift from the general purpose microprocessor-centric tables to more heterogeneous and diverse architectural approaches including industry standards at the heterogeneous integration substrate level and the intra-system and inter-system levels including the trade-offs between electronic and photonic signaling.

8. ACRONYMS AND ABBREVIATIONS

TERM	DEFINITION
2D	2 Dimensional
3D	3 Dimensional
5G	Fifth Generation
ADAS	Advanced Driver Assistance Systems
AI	Artificial Intelligence
API	Application Programming Interface
APU	Accelerated Processing Unit
AR	Augmented Reality
ASIC	Application-Specific Integrated Circuit
BRAM	On-chip distributed SRAM block
CAGR	Compound Average Growth Rate
CDNs	Content Delivery Networks
COTS	Commercial Off-the-Shelf
CPS	Cyber-physical System
CPU	Central Processing Unit
CXL	Compute Express Link
DDR	Double Data Rate
DMIPS	Dhrystone MIPS (Million Instructions per Second)
DPU	Data Processing Unit
DSA	Domain Specific Application
DSP	Digital Signal Processing Unit
DX	Digital Transformation
ECU	Electronic Control Unit
EDVAC	Electronic Discrete Variable Automatic Computer
FeRAM	Ferroelectric Random Access Memory
FIPS	Federal Information Processing Standards
FPGA	Field-Programmable Gate Array
GDDR	Graphics Double Data Rate
GDPR	General Data Protection Regulation
GHz	Gigahertz
GPU	Graphics Processing Unit
GT	Green Transformation
HBM	High Bandwidth Memory
I/O	Input/Output
IEEE	Institute of Electrical and Electronics Engineers
IOE	Internet of Everything
IoT	Internet of Things
IoTe	Internet-of-things edge
ISA	Instruction set architecture
IT	Information Technology
LUT	Primary Logic Element
MCU	Micro Controller Unit
ML	Machine Learning
MLOps	Machine-learning Operations
MRAM	Magnetoresistive Random-Access Memory
NIST	National Institute of Standards and Technology
NPU	Network Processing Unit

TERM	DEFINITION
NVM	Non-Volatile Memory
OT	Operational Technology
PLDs	Programmable Logic Devices
PUF	Physically Unclonable Function
RAM	Random Access Memory
RFI	Radio Frequency Interference
RISC-V	Reduced instruction set computer instruction set architecture ISA)
ROI	Return on Investment
RRAM	Resistive Random Access Memory
SCM	Supply Chain Management
SOC	System on Chip
SOI	Silicon on Insulator
SRAM	Static Random Access Memory
SWaP	Space, Weight, and Power
TBPS	Terabytes per Second
TDP	Thermal Design Power or Total Power Dissipation
THz	Terahertz
TPUs	Tensor Processing Units
Tx/Rx	Transmit and Receive
UAV	Unmanned Aerial Vehicles

9. 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 uncloneable 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].
[7]	Continuous Delivery Foundation “MLOps Roadmap”, [Online]. Available https://bit.ly/3rBg9hQ [2021 Edition]