INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

2017 EDITION

SYSTEMS AND ARCHITECTURES

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

Table of Contents

List of Tables

ACKNOWLEDGMENTS

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

Prof. Tetsuya Asai, Hokkaido University

Kirk Bresniker, HPE

John Carter, IBM

Tom Conte, Georgia Tech

Ty Garibay, Arteris

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

Dr. Yoshi Hayashi, Renesas Electronics

Bruno Michel, IBM

David Mountain, LPS

Shubu Mukherjee, Cavium

Kushagra Vaid, Microsoft

Marilyn Wolf

SYSTEMS AND ARCHITECTURES

1. INTRODUCTION

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

This section considers four different types of systems:

- 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.
- Cyber-physical systems (CPS) provide real-time control for physical plants. Vehicles and industrial systems are examples of CPS.
- Mobile devices such as smartphones provide communication, interactive computation, storage, and security. For many people, smartphones provide their primary or only computing system.
- Cloud systems power data centers to perform transactions, provide multimedia, and analyze data. Cloud systems represent a trend towards common design principles and methodologies between traditional enterprise, high performance scientific, and web native compute.

1.1. CURRENT STATE OF TECHNOLOGY

All four of these application areas are in general use: mobile devices number in the billions worldwide; cloud systems provide an increasing array of services; cyber-physical systems provide essential services; and Internet-of-Things networks perform important services in a range of applications.

These systems do not exist in isolation: mobile devices, IoT edge devices, and cyber-physical systems all provide data that is analyzed by cloud systems. Many complex systems exhibit characteristics of both IoT and CPS. Certain aspects of data centers and cloud systems, power management and thermal management for example, make use of cyber-physical and IoT techniques.

The volumes of data generated by IoT and cyber-physical systems is staggering. A fleet of 1000 vehicles generates four petabytes per day of data from their onboard sensors; that volume of data is equal to the total data volume handled by Facebook. Given the huge amounts of data generated, much raw data never makes its way to the data center. Efficiency breakthroughs allowing *in situ* analysis of raw data in IoT and CPS systems could provide extremely disruptive economic potential.

1.2. DRIVERS AND TECHNOLOGY TARGETS

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

1.3. VISION OF FUTURE TECHNOLOGY

Artificial intelligence (AI) has emerged as a critical technology in applications as diverse as smartphones and autonomous vehicles. Much AI-driven computation will occur in the cloud, but we expect mobile systems, IoT edge devices, and cyber-physical systems to all include AI components.

We expect augmented reality (AR) to emerge as an important application area, particularly for mobile and cloud systems. The large demands on computating, sensing, and display for AR will drive the development of mobile systems in particular. While multimedia has driven many aspects of mobile system development for many years, we have reached perceptual limits for many multimedia applications. We expect AR to take the place of multimedia as an important driver for mobile systems.

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

Security and privacy are critical requirements for virtually all computing systems and are necessary for all four of our roadmapped systems.

The notion of fog computing—computing resources between edge devices and the cloud—has emerged over recent years. We expect the deployment of fog computing to accelerate driven by the growth of IoT and cyber-physical systems. We will discuss this emerging trend in more detail in Section 4.1.

A key attribute for cloud systems is the radius of effective communication of data. Traditional architecture, deploying tens of cores on a system-on-chip (SoC), make use of low latency, high bandwidth direct attached memory. Modern applications such as social computing must operate on huge datasets that cannot be held in those types of memories. If data access times to other memory systems are long, 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. If delays become short enough, then programmers can treat that data as local leading to increased performance and reduced bug rate and maintenance due to simplified implementation. Optical communications promise to greatly expand the radius of fast data transfer over the next few years.

2. Scope of Report

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

3. Summary and Key Points

- We expect artificial intelligence and augmented reality to become important new drivers for the growth of all four system areas.
- Internet-of-Things and cyber-physical systems both generate vast quantities of data that will accelerate the growth of big data.
- Security and privacy are key system requirements for all four system areas.
- Optical interconnect and other advances in system interconnect provide the opportunity to provide an inflection point for both cloud system design and the design of applications that run on those cloud systems.

4. Emerging Trends

4.1. Fog Computing

The growing importance of internet-of-things (IoT) systems and cyber-physical systems (CPS) has led to the concept of *fog computing*, intermediate layers of physically distributed computational capacity. We expect the fog paradigm to grow in importance over the coming years. 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. Bandwidth and power constraints mean that we have to limit the distance the data travels. Some data will be processed within IoT nodes. Other data will be sent to a hub, located in the fog, for additional processing. Fog processing is particularly important for the correlation of multi-sensor data.

4.2. Internet-of-things and Cyber-physical Systems

This roadmap provides separate analysis of IoT edge (IoTe) devices and cyber-physical systems. While both types of systems connect computing devices to the physical world, and there is some overlap in the usage of these terms, we

believe that considering them separately in this roadmap gives readers greater insight into the evolution of such systems. We can contrast CPS and IoT systems in several ways:

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

5. CLOUD

The term *cloud* refers to the engineering of data center scale computing operations. Cloud systems support a number of important applications: web service; media streaming, shopping and commerce; big data for social networking, recommendations, and other purposes; precision medicine; training of AI systems, and high performance scientific computation for science and industry.

Emerging applications for cloud have characteristics that differ from those of scientific computing in several ways. Scientific computation emphasizes numerical algorithms. Cloud applications, in contrast, emphasize streaming for multimedia and transactions for commerce and other database applications; these applications are also more input/output (I/O)-centric than are some scientific applications. Social networking is a prime example of the graph-oriented algorithms that cloud systems must execute on very large models.

5.1. MARKET DRIVERS

Market drivers for the cloud include direct services (multimedia, shopping, shared experience), big data and data analysis (social network analysis, AI, smart cities, smart industry, precision medicine). We note that these applications differ from traditional scientific computing applications that emphasize numerical methods.

5.2. CHALLENGES AND OPPORTUNITIES

Cloud applications present several challenges for system designers. Data centers are starting to take advantage of heterogeneous core types, much as embedded systems have done for many years. System architects need to balance the performance improvements for chosen applications provided by specialized accelerators against the utilization of these specialized cores. 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 optical networking to greatly enlarge useful locality radius over the next few years. Memory bandwidth is a constraint on both core performance and number of cores per socket. Stacked memories, which are starting to come into commercial use, provide higher bandwidth memory connections. Thermal power dissipation continues to be an important limit.

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. The term *hyperconvergence* is used to describe the point at which I/O speeds approach internal interconnect speeds.

5.3. POWER AND THERMAL CONSIDERATIONS

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

5.4. METRICS

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

Table SA-1 Difficult Challenges

	2017	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*	*2026*	*2027*	*2028*	*2029*	*2030*
# cores per socket	28	32	38	42	46	50	54	58	62	66	70	70	70	70
Processor base frequency (for multiple cores together)	2.50	2.75	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90	4.00	4.10
Core vector length	512	512	512	512	1024	1024	1024	1024	2048	2048	2048	2048	2048	2048
L1 data cache size (in KB)	32	36	36	38	38	40	40	42	42	44	44	44	44	44
L1 instruction cache size (in KB)	32	48	48	64	64	96	96	128	128	160	160	160	160	160
L2 cache size (in MB)	1	1	1	1.5	1.5	1.5	2	2	2	2.5	2.5	2.5	2.5	2.5
LLC cache size (in MB)	55	61	67	73	81	89	97	107	118	130	143	157	173	190
# of DDR channels	6 to 8	6 to 8	6 to 8	6 to 8	8 to 10	8 to 10	8 to 10	8 to 10	10 to 12	10 to 12	10 to 12	10 to 12	10 to 12	10 to 12
Peak DDR bandwidth (GB/s)	154	188	205	256	352	458	595	773	1005	1307	1699	2209	2871	3733
HBM ports	4	4	4	4	6	6	6	6	6	6	6	6	6	6
HBM bandwidth (TB/s)	2	2	2.4	2.4	6	6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6
PCI-e lanes	48	56	64	72	80	88	96	104	112	120	128	136	144	152
PCI-e per lane bandwidth (GT/s)	16	32	32	32	32	64	64	64	64	128	128	128	128	128
Socket TDP (Watts)	205	215	226	237	249	262	275	288	303	318	334	351	368	387

L1 = level 1 cache; LLC = last-level cache; TDP = total power dissipation.

6. MOBILE

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

6.1. MARKET DRIVERS

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

6.2. CHALLENGES AND OPPORTUNITIES

Mobile systems present several challenges for system designers. Multimedia viewing, such as movies and live TV, have driven the specifications of mobile systems for many years. We have now reached many of the limits of human perception, so increases in requirements on display resolution and other parameters will be limited in the future based on multimedia needs. However, augmented reality will motivate the need for advanced specifications for both input and output in mobile devices. Mobile device buyers demand frequent, yearly product refreshes. This fast refresh rate influences design methodologies to provide rapid silicon design cycles; it can also suggest the use of programmability to provide a broad range of models on a given platform. Financial transactions are now performed using mobile devices. We expect this trend to grow, particularly in developing nations, where financial technology will leapfrog. Security and privacy are key concerns for mobile devices, particularly for financial transactions.

6.3. POWER AND THERMAL CONSIDERATIONS

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

6.4. METRICS

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

Table SA-2 Mobile Technology Requirements

	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032
# CPU cores	8	10	10	12	12	18	18	18	25	25	25	28	28	28	30
# GPU cores	16	16	32	32	32	64	64	64	128	128	128	256	256	256	512
Maximum frequency (GHz)	2.5	2.75	3.025	3.3275	3.66025	4.026275	4.428903	4.871793	5.358972	5.894869	6.484356	7.132792	7.846071	8.630678	9.493746
Number of cameras	2	3	3	3	4	4	4	6	6	6	8	8	8	8	8
Camera resolution (MP)	12	12	15	15	18	18	20	20	20	24	24	24	24	24	24
Number of sensors	4	6	8	8	10	10	12	12	12	12	12	16	16	16	16
Cellular data rate (Mb/s)	12.5	21.6	21.6	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Wi-Fi data rate (Mb/s)	850	850	850	7000	7000	7000	28000	28000	28000	28000	28000	28000	28000	28000	28000
Board power (mW)	4900	5096	5350.8	5618.34	5899.257	6194.22	6503.931	6829.127	7170.584	7529.113	7905.569	8300.847	8715.889	9151.684	9609.268

7. INTERNET-OF-THINGS EDGE DEVICES

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

7.1. MARKET DRIVERS

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

7.2. CHALLENGES AND OPPORTUNITIES

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

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

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

7.3. POWER AND THERMAL CONSIDERATIONS

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

7.4. METRICS

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

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

	2018	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*	*2026*	*2027*	*2028*	*2029*	*2030*	*2031*	*2032*
CPUs per device	1	1	2	2	2	4	4	4	4	6	6	6	8	8	8
Maximum CPU frequency (MHz)	255	276.675	300.1924	304.9955	309.8754	314.8334	319.8707	324.9887	330.1885	335.4715	340.839	346.2925	351.8331	357.4625	363.1819
Energy source (B = battery, H = energy harvesting)	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H	B+H
Tx/Rx power/bit (μW/bit)	0.608	0.372096	0.227723	0.139707	0.08571	0.05714	0.038093	0.025396	0.01693	0.011287	0.007525	0.005016	0.003344	0.00223	0.001486
Battery operation lifetime (months)	6	6	9	9	9	9	9	12	12	12	12	18	18	18	18
Deep suspend current (nA)	61	52	44	38	32	27	23	20	17	14	12	10	9	8	7
Sensors per device	4	4	4	8	8	8	12	12	12	16	16	16	16	16	16

8. CYBER-PHYSICAL SYSTEMS

Cyber-physical systems are networked control systems. These distributed computing systems perform real-time computations to sense, control, and actuate a physical system. Many cyber-physical systems are safety-critical.

8.1. MARKET DRIVERS

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

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

8.2. CHALLENGES AND OPPORTUNITIES

Several challenges present themselves to cyber-physical system designers. Cyber-physical systems must be highly reliable at all levels of the design hierarchy. Security and privacy are critical to trusted operation. 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.

Artificial intelligence has enabled recent leaps in the capabilities of autonomous vehicles. We expect the use of AI for cyber-physical systems to continue to escalate.

A fleet of only 1,000 vehicles generates 4 petabytes per day of sensor data. That volume of data is equal to the total daily data collection of leading social media sites. Sensor data from cyber-physical systems can be used for big data applications and emerging products such as automated diagnosis and repair dispatch. The interaction between CPS, IoT, fog computing nodes, and cloud data centers presents an ongoing challenge and opportunity.

8.3. POWER AND THERMAL CONSIDERATIONS

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

8.4. METRICS

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

Table SA-4 Cyber-physical Systems Technology Requirements

	2018	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*	*2026*	*2027*	*2028*	*2029*	*2030*	*2031*	*2032*
Number of devices	64	64	64	64	64	128	128	128	128	256	256	256	512	512	512
CPUs per device	4	4	4	8	8	8	8	12	12	12	12	16	16	16	16

9. CROSS TEAMS

The Systems and Architectures roadmap team interacts with several other roadmap focus teams. The Application Benchmarking team provides application data that informs our system architecture analysis. The Outside System Connectivity team provides insight into the development of communication technologies for cloud systems.

10. CONCLUSIONS AND RECOMMENDATIONS

We summarize conclusions for each of our system areas:

- Cloud systems operate on huge data sets. The architecture of these systems influences the design of application software.
- Mobile systems have emerged as a key computing device for many consumers, as well as their original communications functions. Augmented reality and financial transactions are two examples of important emerging applications for mobile systems.
- Internet-of-things edge devices must provide sensing, computation, and communication at extremely low power levels. We expect energy harvesting to become more common in this class of devices.
- Cyber-physical systems perform real-time computations to control physical systems. Reliability is a key design requirement for CPS.

We have identified several recommendations:

- Security and privacy are critical to all our system areas.
- Energy harvesting is a key technology to enable the growth of IoT edge devices.
- Devices in all system areas should be designed to provide relevant AI features.
- Augmented reality will create further demand for computation, communication, sensing, and display on mobile devices.
- Cloud system architectures should take advantage of advances in interconnect to provide simpler programming models for cloud application programmers.