IDC ANALYZE THE FUTURE

IDC TECHNOLOGY SPOTLIGHT

Sponsored by: Rambus

This IDC Technology Spotlight highlights key, often hidden, memory and interface technologies that are enabling high-performance electronic systems to serve the disruptive trends of the next decade: IoT, 5G, and Artificial Intelligence.

Hidden Signals: The Memories and Interfaces Enabling IoT, 5G, and AI

February 2019

Written by: Shane Rau, Research Vice President, Computing Semiconductors

Introduction

The transformation of data into useful information requires data collection, processing, transmission, storage, and analysis. Data transformation is a journey, both from the physical (analog) world to the digital world and across the Internet system landscape, from endpoints to edge infrastructure, through communications infrastructure to the datacenter and back.

Not Just PCs and Servers Anymore

Data's journey used to be through a closed information technology (IT) loop, book-ended by human beings. People input data; PCs and servers processed, transmitted, and stored data; people analyzed the output. Otherwise, data moved hardly at all, isolated in disconnected, embedded machines enabled by proprietary operational technologies (OT).

But data no longer rests. In the past 10 years, several major developments have opened and extended the data transformation journey:

» Development: Ubiquitous Connectivity

- **4G Cellular** – Cellular phones transformed into smartphones once cellular 4G LTE air interfaces became common. Yet, even as cell phones have carried 4G into high volume, cellular technology has migrated across the Internet landscape, including many consumer, automotive, and industrial systems. While non-cell phone systems represented about 6% of all systems with cellular technology in 2014, non-cell phone systems will represent about 20% of all systems with cellular technology in 2023.
- **Wi-Fi** – Wireless Fidelity (Wi-Fi) has become the standard for local-area wireless connectivity in the home, on the road, and in the office. By 2023, the number of Wi-Fi-enabled systems will exceed 5 billion per year. From 2019 to 2023, most Wi-Fi-enabled systems will have shifted from the 1.3 gigabits per second (Gbps) throughput of the 802.11ac Wi-Fi standard (which the Wi-Fi Alliance now calls Wi-Fi 5), to the 3.5 Gbps throughput of the 801.11ax Wi-Fi standard (now called Wi-Fi 6).

AT A GLANCE

KEY STATS

» From 2014 to 2023, data created by electronic systems will grow 10x to 103 zettabytes.

» The average server will support over 670 gigabytes of DRAM in 2023.

WHAT'S IMPORTANT

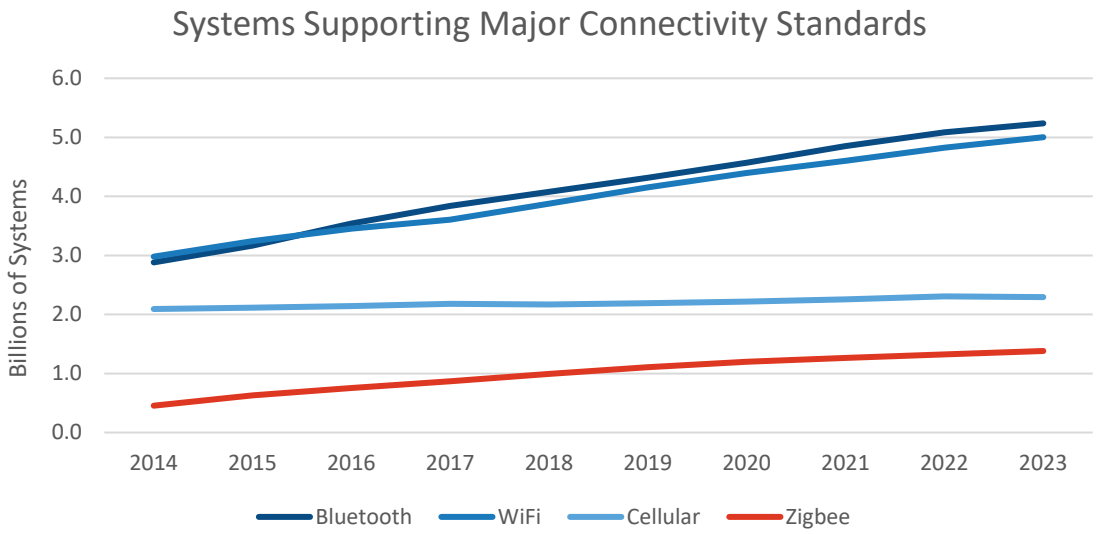
Connectivity between chips (interfaces) and where data is stored in real-time (memories) are as important as the connectivity between systems and the processors that compute the data.

Technologies like GDDR6, HBM, and memory buffers that move, accelerate, and store data are critical to enabling the IoT, 5G, and AI trends that will drive new markets for the next ten years.

- **New adaptive connectivity protocols** – Beyond cellular and Wi-Fi connectivity, wireless connectivity has diversified significantly in the past five years, adapting to suit cost, power, thermal and space needs of system types, workloads, and end-user needs. Adding to cellular, Wi-Fi, and Bluetooth technologies, IDC now tracks major wireless standards and 25 of their permutations. Figure 1 shows the significant growth in the number of systems supporting major connectivity protocols, including cellular, Wi-Fi, Bluetooth, and Zigbee (802.15.4).

FIGURE 1: ***Forecast for Connected Systems***

Ubiquitous Connectivity Means Increasing Data Flow Across the Internet



Zigbee is also referred to as 802.15.4

Source: IDC, 2019

IT Meets OT

Ubiquitous, secure, and adaptive connectivity among PCs and servers (IT) and embedded machines (OT) has led to major trends that have emerged over the past 10 years. Born from this connectivity were the Internet of Things (IoT) and the cloud. Notable is the virtual length between the locales of these entities, the IoT being a vast collection of interconnected, application-specific systems (industrial, automotive, healthcare, etc.), and general-purpose systems (datacenter servers). What happens when you connect formerly isolated, islands of data in those embedded systems to the Internet and their data flows into the datacenters?

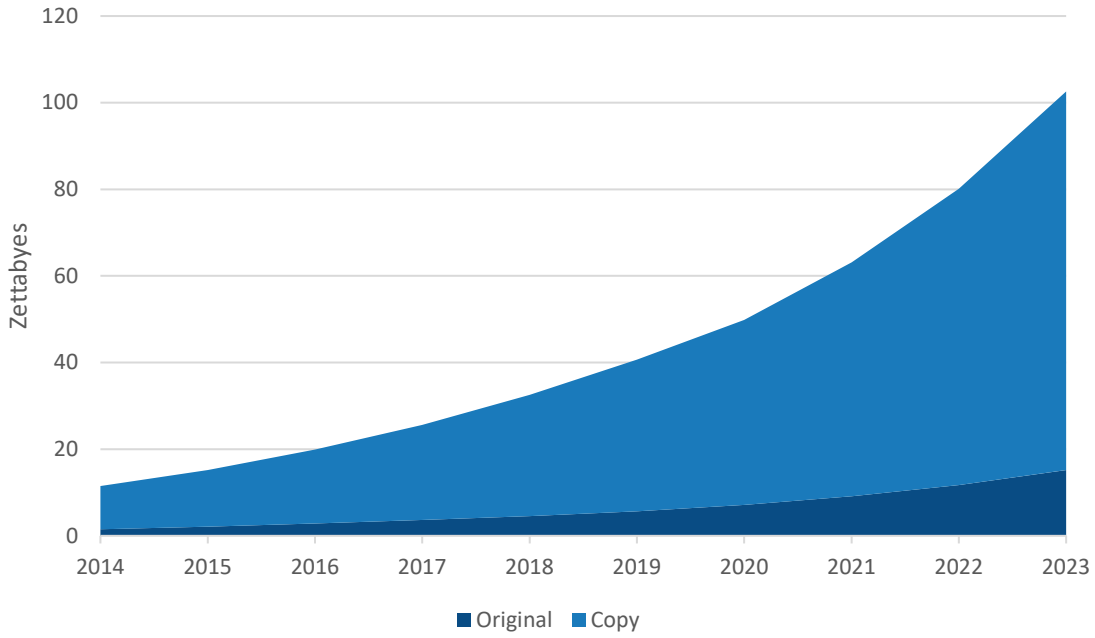
» The Data Deluge

- **Volume of data created** – The result of connectivity has been a deluge of data from these newly connected systems, industries, and users. Major characteristics of the data deluge include more people with more devices being used for longer periods and creating more data, and machine-to-machine connectivity, generating data of such volume that it surpasses the human ability to process. Figure 2 shows IDC's forecast for data creation. Compared to the data created by—largely

disconnected—systems in 2014, data created among—largely connected---systems in 2023 will be 103 zettabytes, or 10 times larger. IDC expects that, in 2023, more than half of all data created will come from endpoint IoT devices.

FIGURE 2: ***Forecast for Data Creation***

Data creation and data duplication means data grows at 28% CAGR 2014-2023



Source: IDC, 2019

A major consequence of data creation is the need to store data in real time inside a system while it is being processed. The more data that's created, the more real-time memory a system needs. DRAM, due to its speed — enough to keep up with the processor in a system — is the most common real-time memory. Table 1 shows IDC's forecast for memory per major system categories, including IoT, PCs, graphics, and servers.

TABLE 1: ***Forecast for DRAM Main Memory per Major System Categories (Gigabytes per System)***

	2018	2019	2020	2021	2022	2023
Servers	304.69	375.52	446.72	523.52	599.68	670.72
Desktop PCs	6.20	6.99	7.86	8.70	9.59	10.55

Graphics	1.2	1.44	1.66	1.93	2.22	2.53
Industrial Automation Systems	2.04	2.17	2.26	2.32	2.44	2.58
Medical Imaging Systems	3.41	3.95	4.41	4.92	5.41	5.93
High-End Network Router	7.83	9.01	10.32	11.80	13.40	15.08

Source: IDC, 2019

- **Sophistication of data created** – Another result of connectivity has been a growing diversity of data types, such as data from factory machines that speak with specialized protocols, and data from consumer systems that tend to be very video-heavy. Whereas a factory system's architecture used to require specialized, even proprietary, processor and memories technologies optimized for its specialized data types (and the same for consumer systems), the mixing of different data types across networks defies old expectations. Processors and memory technologies, for example, can no longer assume one data type. Further, the human beings programming those systems can no longer assume traditional methods of data collection, processing, transmission, storage, and analysis.
- **Need for security –** Ubiquitous connectivity that opened IT and connected OT to the Internet landscape significantly increased the need to secure devices across that landscape, from the servers in the data center to the wearables and the endpoints. Not only the number of systems to secure increased by the billions but the tasks of securing each system increased due to the need to provision, monitor, update, and control the endpoint IoT devices and/or the network. Security must reach enterprise or government grade to realize the value in the billions of IoT systems shipping and connected annually.
- **Performance demands** – Connectivity that results in more data, more sophisticated data types and the need for security creates more processing and storage overhead in the system. The datacenter has been the leading edge of distributed computing and the industry's attempts to deal with the deluge of data. Notably, cloud service providers (CSPs) have been taking up the mantra of computing and, in doing so, investing in higher performance microprocessors and heterogeneous architectures.

  However, analysis across IDC's view of the Internet system landscape (IoT/Endpoints, Edge Infrastructure, Primary Clients like PCs and smartphones, Networking Systems, and Data Center) reveals that, over the next five years, the most opportunity is at the IoT/Endpoints and at the Edge Infrastructure in industries like automotive and industrial automation where AI is being deployed for

inferencing. ADAS, automotive control and smart home are the highest segments of growth in embedded and intelligent systems. Figure 3 illustrates the movement of compute performance across the Internet system landscape. Notably, there will be more collective compute power at the endpoints and edge than in the cloud and datacenter core.

FIGURE 3: ***Heat Map Forecast for Compute Capabilities***

As Average Compute Power per System Increases at Each Segment in the Internet System Landscape, Processors Will Need More and Faster Interfaces and Memories

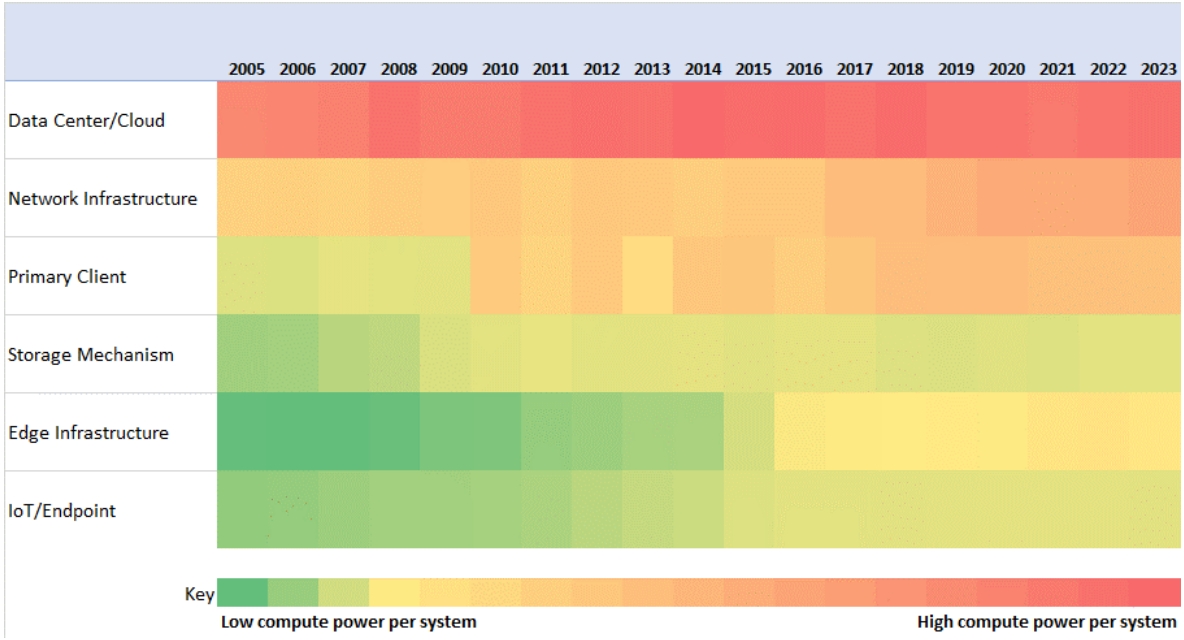


Figure note: IDC measures the compute power in each system type that we track according to the average number and type of data processing semiconductors contained in the system, including microprocessors, graphics processors, field-programmable gate arrays, digital signal processors, ASICs, and ASSPs.

Source: IDC, 2019

- **New services and business models** – Connecting IT and OT is forcing businesses to redefine business processes if they are to take advantage of cost savings and revenue opportunities. For example, a manufacturing business that connects its manufacturing systems to its IT systems cannot ignore the key performance indicators that come in from the factory floor and suggest ways to reconfigure systems for better safety, better waste control, and more efficient production.

Trends

Trends That Span IT and OT

These developments have stimulated the market to search for what should be the data path, the methods and route of the journey of transformation. IDC identifies three megatrends among industry enablers:

» **IoT** – Data begins its transformation from the physical world to the digital world with endpoint systems. The key enabling technologies of endpoint systems are sensors that capture data, usually in analog form, and data

processors, such as microcontrollers (MCUs) and microprocessors, that perform at least a minor level of compute. Increasingly, endpoint systems at all levels of sensing and computing capabilities are being connected directly or indirectly to the Internet and so are joining the Internet of Things and forming the edge of the Internet. IoT edge systems can be used in various roles such as sensing or actuator/control functions, interacting with the physical world as well as local user interface, and compute functions.

Of nearly 300 system categories that IDC tracks and forecasts, we categorize over two-thirds of them as IoT edge/endpoint systems that span 11 industries and billions of units. IoT edge/endpoint systems' collective volume and origins in low-end, microcontroller-based, disconnected platforms means they are the driving force in the massive transformation of traditional embedded systems to smart systems that require high-end enabling technologies, such as more memory, microprocessors, accelerators, and sensors.

- **5G** – As described, ubiquitous and reliable connectivity has been the most critical enabler of data transformation trends. Without high-performance LANs and WANs, there would be no cloud, no IoT, and no need for AI. So, we look towards the next major generation of connectivity — 5G — for the cadence of progress.

  5G intends to meet the forecasted data transmission requirements across the Internet landscape, from IoT endpoints through edge infrastructure and into core networks and the datacenter. Provisioning and managing IoT systems and the data coming in from trillions of sensors on those systems will place tremendous loads on the networks, as will video from mobile phones and tablets, and VR from PCs and game consoles. The 5G standard will be backward compatible with LTE, with the 5G overlay as a data and bandwidth booster. Significant changes in modulation, waveform, RF front-end MIMO (as much as 64 x 8), and support for 50-plus bands are expected.

  The industry will not see the first rollouts of 5G networks until 2020, with deployments continuing through 2025. 5G will be a point of disruption for multiple system categories, especially the enabling technologies within, including RF and modem semiconductors, but also any semiconductor that enables scalable compute performance, flexible power profiles, and robust signal integrity.

- **Artificial Intelligence** -- The huge amounts of data being generated, processed, transmitted, stored, and analyzed demands that primary clients (PCs, phones, tablets), endpoint systems, and datacenter systems (servers, storage systems, networking infrastructure) that are oriented toward the limited bandwidth and capacity of human management and control be reoriented toward the hyper-efficient bandwidth and capacity of machine management and control. This reorientation requires processors, interfaces, and memories optimized for artificial intelligence (AI).

  Semiconductors optimized for AI workloads first launched a decade ago, starting as DSPs, but then also as FPGAs, ASICs, and GPUs. Next-generation solutions will focus on workload optimization. Notably, heterogeneous architectures will be instrumental in enabling developers to program once and leverage the most efficient solution for AI workloads. Deep learning is the lead growth area now, driven by the datacenter. Long term, however, the growth of the edge for inferencing tasks will be at the epicenter of market growth.

Definitions

- **5G** – 5G describes the next generation of mobile networks after 4G LTE mobile networks.
- **Artificial Intelligence** – Artificial Intelligence (AI) is the simulation of human intelligence by machines.

- **Cloud** – The cloud is not a physical entity, but a network of local and remote servers worldwide that are hooked together so they operate as a single, virtual system.
- **Graphics Double Data Rate (GDDR)** – GDDR is a form of DRAM but with higher bandwidth due to a wider memory bus than standard DRAM, such as DDR. For example, the latest version of GDDR, GDDR6, supports bandwidth up to 16 gigabits per second.
- **High-Bandwidth Memory (HBM)** – HBM stacks multiple DRAM chips into a single chip package to increase the amount of memory that a system can address (talk to) in a given unit of time.
- **Internet of Things (IoT)** - The IoT is a network connecting devices (either wired or wireless) or "things" that are autonomously provisioned, managed, and monitored.
- **Machine Learning (ML)** – ML is a form of AI; however, it focuses on the ability of machines to adapt to changing conditions on their own.
- **Memory Buffer** – Memory Buffers bring local intelligence to a DRAM memory module. By making the memory modules more intelligent, memory buffers expand the effective memory capacity of the module.
- **Serializer-Deserializer (SerDes)** – SerDes converts electrical signals from parallel into serial and back again to increase communications speed.

Memory and Interface Technologies and Applicability to Trends

As evolving connectivity has been critical to enabling the data transformation journey between systems, so has the evolving connectivity between chips inside systems been critical. The *interfaces* between high-performance data processors have been evolving because the market has recognized that monolithic, homogeneous microprocessors are insufficient to process and secure the amount and disparity of data flooding into high-performance systems. General-purpose microprocessors must connect to and share workloads with specialized processors, such as graphics processors (GPUs), field-programmable gate arrays (FPGAs), and custom application-specific integrated circuits (ASICs). Further, processors must be mixed and matched according to the needs of the application; *heterogeneous data types require heterogeneous processor types*. IDC forecasts that combined server GPU, FPGA, and AI ASIC revenue will grow from $4.2 billion (17% share of total server processing chips) in revenue in 2018 to $14.7 billion (37% share) in revenue in 2023.

CPUs, GPUs, FPGAs, and ASICs, however, share need for external *memories*—such as CPUs for main memory and GPUs for frame buffers—that hold data close to the processor. These needs thus put specific pressure on the interfaces between these processors and their memories. Thus interfaces, memories, and intimacy between the two are critical to making the often-hidden data signals flow efficiently; *heterogeneous processor types require heterogeneous interfaces and memories*.

Specific examples of enabling interfaces and memories that have emerged include:

- **Memory Buffers** – As illustrated in table 1, systems like servers will need increasing amounts of DRAM to function. However, such increasing capacities means increasing burden on the system microprocessor to manage that memory across the memory bus. Memory buffer chips reduce that burden by bringing local intelligence to the memory, on the memory modules. Buffer chips on Load-Reduced DIMMs (LR-DIMMs), for example, reduce the load

on the microprocessor's memory controller and across the memory bus by handling all data, command, and control signals sent to memory and handling all reads and writes to the DRAM chips.

- **GDDR6 – GDDR** exemplifies a specialized interface and memory for a specialized processor. Graphics double data rate memory provides GPUs with a wider memory bus, more throughput per clock, and more flexible power management schemes than traditional DDR memory. GDDR6 will succeed GDDR5 and increase transfer speeds from up to 12Gbps to up to 16Gbps, increase capacity from up to 8GB to up to 16GB, and reduce power draw from 1.5V to 1.3V. Notably, as GPUs have expanded from drawers of pixels on a screen to high performance processors of highly parallelized data, so too has GDDR memory. For example, GPUs are seeing new applications in automotive and AI applications and so IDC expects GDDR memory to follow.
- **HBM2** – High-bandwidth memory stacks multiple DRAM chips in a 2.5D package with a wider interface and lower clock speed than DDR4 to create a small form factor, higher bandwidth-per-watt solution for high-performance computing.
- **HBM Gen2** – supports total bandwidth of 256 gigabytes per second (GBps) across eight 128-bit independent channels and supports stacks of two, four, or eight DRAMs.
- **Chip-to-chip Interconnects** – For when high-performance chips need to connect to each other, chip-to-chip interconnects are also critical for maintaining high speed and signal integrity across variable physical distances. SerDes chips, for example, that convert electrical signals from parallel into serial and back are important in high-performance communications systems because, otherwise, system designs would face a choice of bloated, high-cost, high-power consumption chip interconnects on their boards or low performance; the fastest SerDes chips today drive speeds of 112 gigabits per second.

Conclusion

IoT, 5G, and AI are the transformative trends of our time. Though the ubiquitous connectivity that has created the data deluge and the processors that will compute that data get the most attention, the interfaces between the processors and the memories that keep data local to the processors are as critical as the processors themselves across the Internet system landscape, from endpoints to edge infrastructure and to the datacenter.

The interfaces between the processors and the memories that keep data local to the processors are as critical as the processors themselves.

Cloud service providers are building significant datacenter capacity to take the leadership in the global cloud services competition. Communications service providers are transforming their infrastructure to 5G to enable massive expansion of capacity and meet the data deluge. IoT is widely penetrating in the everyday consumer's life. Artificial intelligence is burgeoning across major industries, including industrial, retail, healthcare, automotive, and transportation. The technologies analyzed here may have independent purposes but are united as key enabling technologies through their economy of scale (standardization), scalability (to meet diverse constraints and requirements), and creators of new capabilities.

Considering Rambus

Dedicated to making data faster and safer, Rambus creates innovative hardware, software, and services that drive technology advancements from the datacenter to the mobile edge. The company's architecture licenses, IP cores, chips, software, and services span memory and interfaces, security, and emerging technologies to positively impact the modern world. Rambus collaborates with the industry, partnering with leading chip and system designers, foundries, and service providers. Integrated into tens of billions of devices and systems, the company's products and technologies power and secure diverse applications, including Big Data, Internet of Things (IoT) security, mobile payments, and smart ticketing. For more information, visit *rambus.com*.

Challenges

- **Technology challenges** – Technology providers are moving to support AI, 5G, and IoT with existing architectures and market solutions. However, IT and OT solutions vendors are at different points in the transformations of their businesses and what they need from their technology partners. While their needs are distinct, what their demands have in common are long-term solution roadmaps and not single, point solutions. They all will require memories and interfaces that meet various combinations of cost, power, performance, and signal-integrity needs.
- **Market inhibitors** – The enormous amount of data being generated, processed, transmitted, stored, and analyzed demands that system companies and technology suppliers move beyond features and specs to building flexible platforms that can be adapted for specific workloads and applications such as image recognition, natural language cognition, pattern recognition, and predictive analytics. However, customers cannot jettison their installed legacy platforms — especially in the quantities of billions of endpoint-to-datacenter system quantities discussed here — until the ROI of the next-generation platforms is proven and even newer, more complex platforms are cheaper than the platforms before.
- **Domain expertise** – Technology providers can no longer design solutions without understanding how those solutions will be used. Heterogeneous system architectures that mix and match general-purpose and specialized processors, interfaces, and memories in solutions optimized for certain applications will thus require technology providers to start from the end users and understand their applications and system workloads.
- **Rambus challenges** – IP companies like Rambus tend to be dwarfed by their customers and the ecosystem around them, placing them in a rather tenuous position of funding R&D for industries that have more funds than they do. As Rambus researches memory and interface technologies for licensing to the market, it will have to maintain a balance between standardization that enables scale and reasonable cost, and innovation that enables differentiation and margin.

About the analyst:

Shane Rau, Research Vice President, Computing Semiconductors

Shane Rau leads IDC's computing semiconductor research covering microprocessors and SoCs, discrete graphics processors (GPUs), FPGAs, and artificial intelligence (AI) accelerators in systems across the Internet, including in the datacenter, in PCs, at the edge, and at endpoints.

IDC Corporate USA

5 Speen Street
Framingham, MA 01701, USA

T 508.872.8200

F 508.935.4015

Twitter @IDC

idc-insights-community.com

www.idc.com