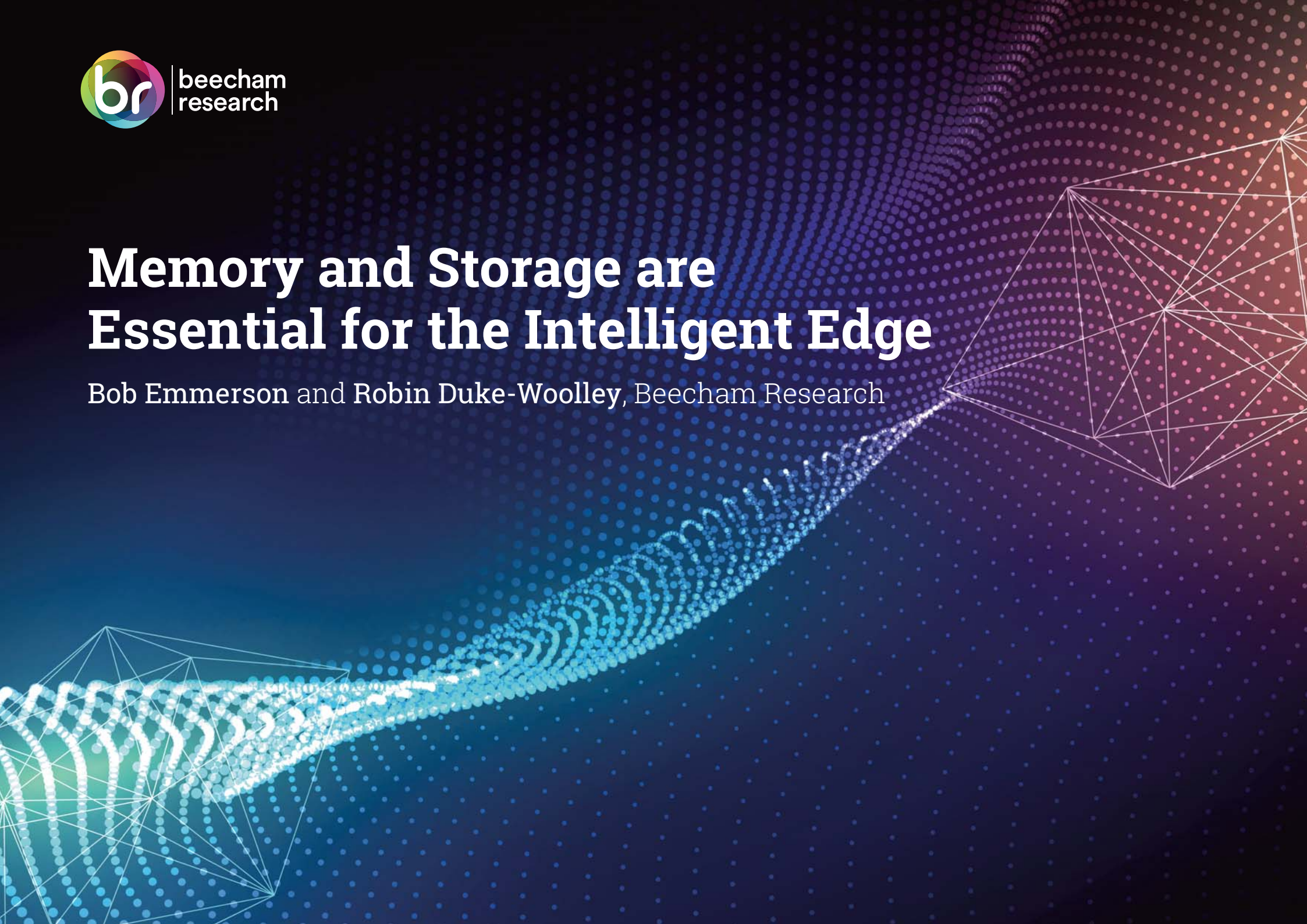


Memory and Storage are Essential for the Intelligent Edge

Bob Emmerson and **Robin Duke-Woolley**, Beecham Research

beecham research

Memory and Storage are Essential for the Intelligent Edge

The Intelligent Edge, also known as edge computing, is becoming increasingly essential for operational efficiencies and this change from a centralised model of computing is happening at speed. Shifting the intelligence from the cloud to the edge enables operational and management decisions on real-time events to be made at the local level. In fact, the Linux Foundation has projected a global edge computing infrastructure market to be worth $800B by 2028[1].

Realising that key benefit involves various edge-related challenges. Initially, the development was constrained by the computing resources of the edge hardware. However, advances in silicon chip technology enabled the creation of small, low-power chipsets that could be embedded in edge hardware to deliver the required performance.

It is a key issue at the intelligent edge where a lot of memory as well as computational power is now needed. Memory resources need to match computing resources. Both are needed to perform perception tasks locally, with high accuracy, low latency, and energy efficiency.

---

1 https://www.linuxfoundation.org/en/press-release/lf-edges-state-of-the-edge-2021-report-predicts-global-edge-computing-infrastructure-market-to-be-worth-up-to-800-billion-by-2028/

New challenges for the Edge

As computing moves closer to data sources, edge architectures will become more complex. Workloads increase and as a consequence more memory and storage are needed. This workload is expected to account for 74% of the data generated by IoT solutions by 2023[1] (place in footer: Source: IDC Global DataSphere, 2018). Storing code and data will become more critical as they increase from megabytes to terabyte levels to accommodate the growing deployment of use cases at the edge.

In addition, the relatively recent availability of small, edge AI chipsets has raised the intelligence bar. It has enabled advanced data analytics to be performed in IoT devices and other IoT-enabled products. Previously this task was conducted in data centers and the Cloud. The edge and the Cloud have always performed complementary roles; now those roles have changed significantly. In addition, real-time aggregation and compute will become increasingly critical, as new solutions will embed new AI chipsets and machine learning accelerators.

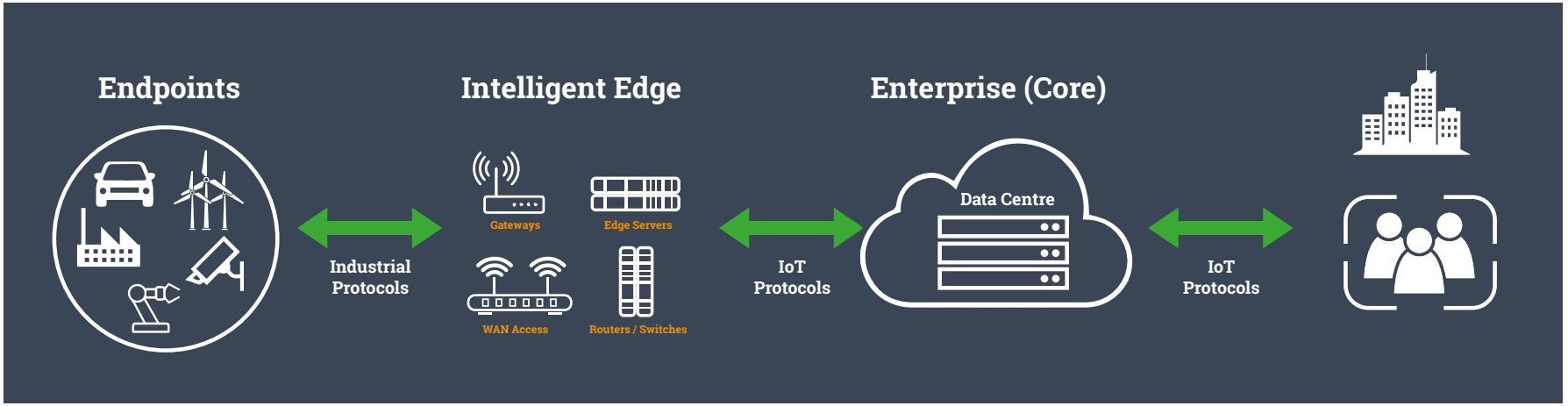


Figure 1. Advances in chipset technology has enabled intelligence, the processing and analysis of raw IoT data, to be provided in near real time at the edge of the network.

[Source: Beecham Research]

Figure 2. Results of a survey into the importance real-time data processing at the network edge

[Source: Beecham Research]

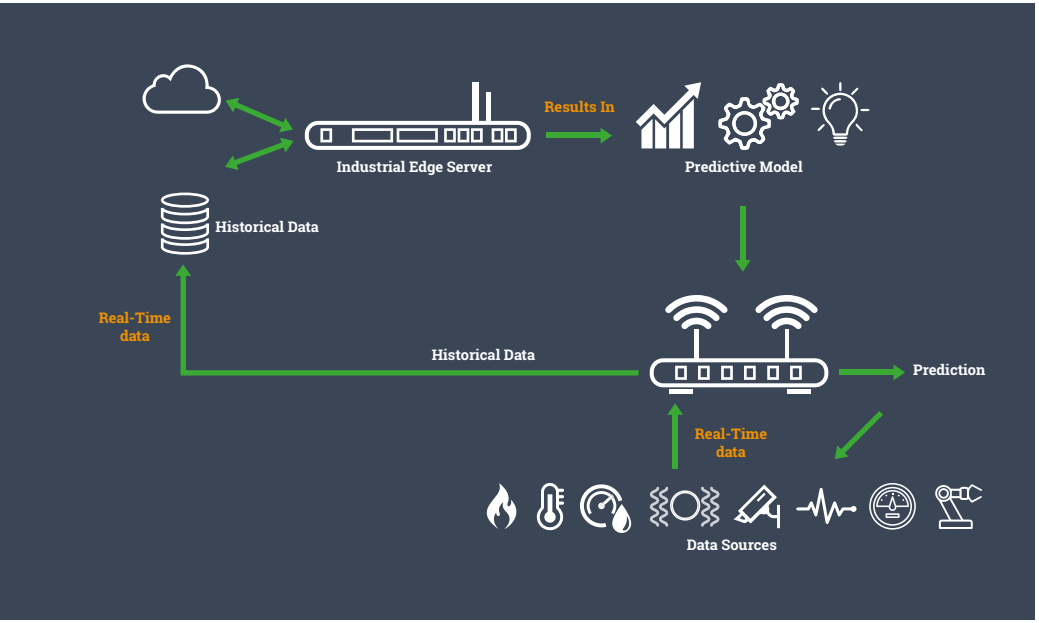


Figure 3. IoT gateways and edge servers in smart manufacturing.

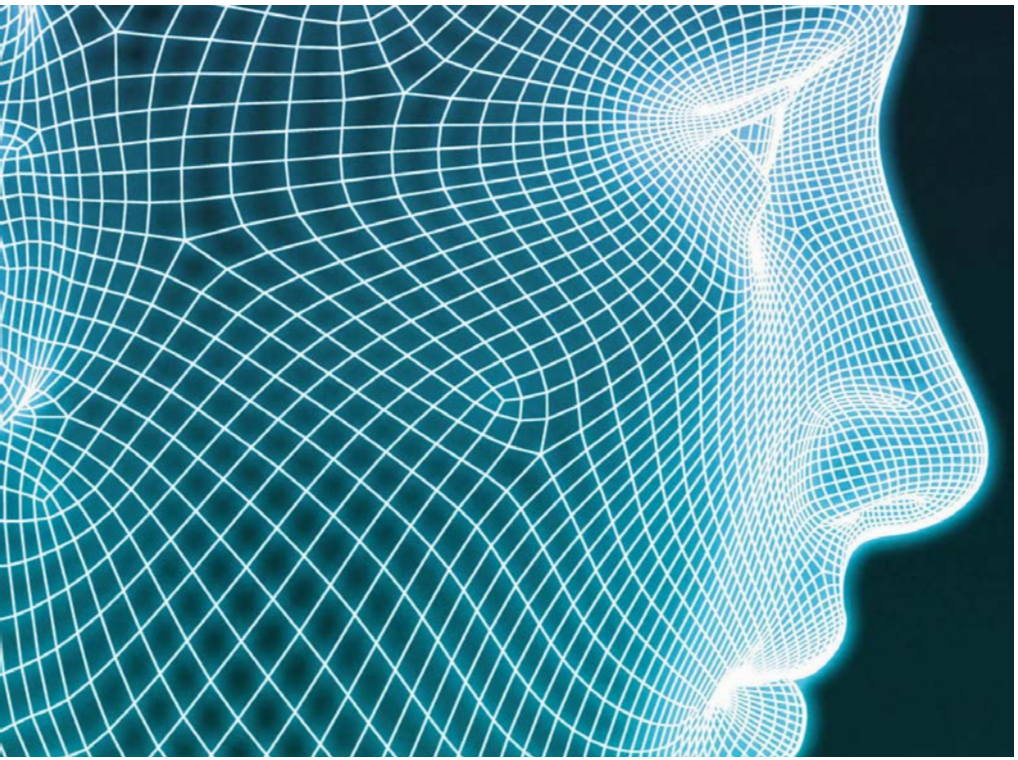
[Source: Beecham Research]

Reflecting this growing need, a survey in late 2020 conducted by Beecham Research among senior level operations professionals was drawn from a base of over 19,000 managers. One topic that it addressed was the importance of real-time data processing at the network edge. A high majority of 64% of respondents considered this would be important at some stage, with 50% saying this was either critical now or will be important in the next 12 months. This is an issue for now, not something that can be put aside.

Gateways and servers

Intelligent IoT gateways can offload computing tasks from smart devices by caching / storing information and acting as a localized cloud that delivers a layer of insight near the source of the data. It is more efficient to have the gateway act as the computing node to capture data on-premise and make analytics decisions locally. In addition, intelligent gateways can provide advanced features and functionality. There are environments in which it is more efficient to use a central local server to do this across multiple gateways. Edge servers can be used to create a scalable architecture in which more compute and storage units can be added to deal with increased complexity.

For both IoT gateways and edge servers, memory selection is particularly important. Designers need to consider the total cost of ownership - not just to fulfill performance and storage requirements in the current solution, but the understanding that the use case will need to scale and evolve, where reliability and technology longevity will also be key.



According to IDC, 73% of enterprises view the edge as a strategic area of investment

AI at the edge

According to IDC, 73% of enterprises view the edge as a strategic area of investment and at least 90% of enterprise applications will embed AI by 2025. Deploying AI at the edge and enabling data analytic functionality that would earlier have been provided in the Cloud represents a significant challenge. AI systems are prediction machines. They ingress data, lots of data, and run algorithms at high-speed to make predictions that enable decisions and actions to be made in near real time. For example, machine vision is an AI application that can enable visual quality control inspection of identical items being transported on conveyor belts at high speed. AI apps at the edge need to perform analytics in a constrained environment that requires highly efficient compute and memory - operating at a lower power budget, validated to edge optimized processors, and the latest technologies to support complex AI compute demands.

AI accelerators

Traditional CPUs and microprocessors were designed for embedded applications that typically operate in predictable, sequential operations. But AI and machine learning frameworks need parallel compute methods to allow handling of significantly larger amounts of data in real-time. AI accelerators are specialized hardware designed to accelerate these basic machine learning computations and improve performance, reduce latency and reduce cost of deploying machine learning based applications.

High performance AI accelerators are purpose-built embedded hardware that offload compute intensive workloads. To execute neural networks these systems need to access memory to support the ingress of data at high speed. There are solutions that employ proprietary on-chip SRAM architectures, but these may be limited in scale and compute capabilities. Most solutions require off-chip memory and will need high-bandwidth memory to support machine learning inference models. Memory solutions such as DDR and LPDDR are the typical solution for edge applications, as they advance both in the number of IO and increasing transfer rates - while improving power efficiency and improved technology nodes.

Code and data storage will need to accommodate the increased demand in these new applications

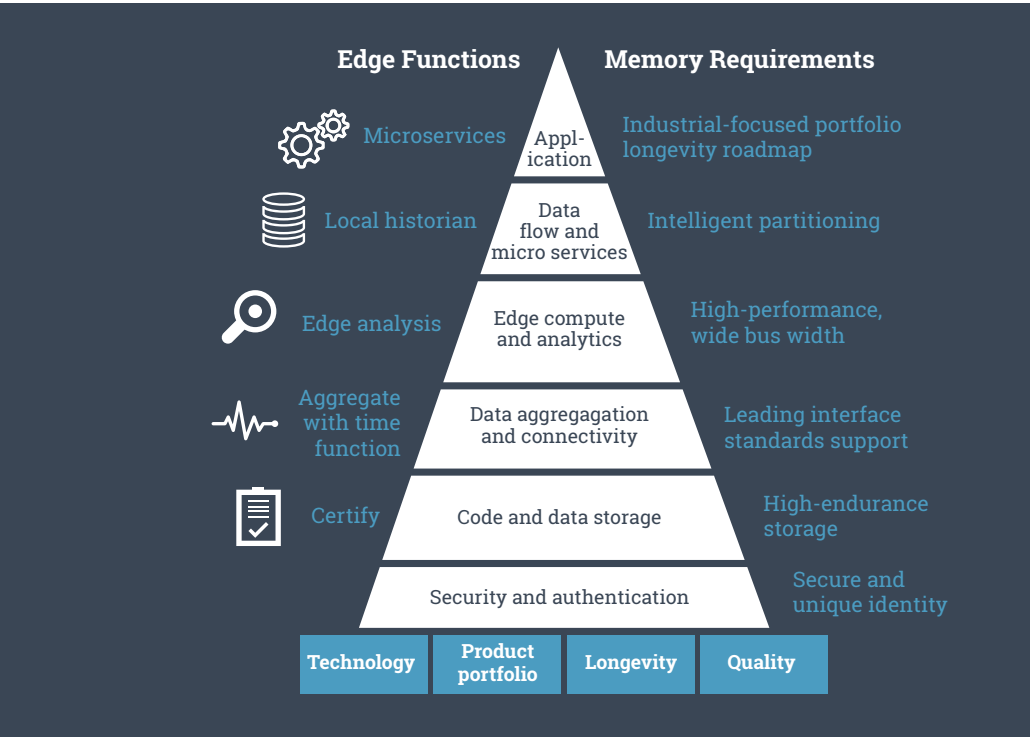


Figure 4. Memory resources need to align with the increased functionality

[Source: Micron]

Memory and storage at the edge

There will be a wide range of edge solutions that will vary based on the compute performance requirements as well as how the cloud/enterprise will integrate. The embedded resources within these solutions will vary as code size and storage requirement grows. Gateways will evolve to more sophisticated architectures that pack more compute and storage for on-site processing of information. This will entail:

- the ability to support multi-sensor data aggregation and connectivity in real-time
- multicore processor systems with hardware AI accelerators that support deep learning inference and higher compute requirements
- embedded local storage for on-premises data management
- new software middleware and APIs to support network virtualization and containers for microservices.

Code and data storage will need to accommodate the increased demand in these new applications. Popular IoT edge platforms typically require from 100 Mbytes up to 250 Mbytes of code storage at its base without any messages. As new services are required, on-board data storage will vary from managed NAND solutions such as e.MMCs, up to TB level industrial/automotive PCIe NVMe SSD embedded solutions in micro-datacenter systems to support the edge data ingested.

Figure 4 illustrates the essential role that memory plays in edge solutions. From high performance DRAM with wide bus width and efficient throughput, to managed NAND and SSD solutions, to the need for a portfolio of industrial-focused solutions. All these need to be part of design tradeoffs and use case considerations. IoT applications that support real-time analytics are driving distributed computing and increased data storage at the end node. This increases the need for high-performance memory solutions in smaller and smaller packages.

In addition, there are a growing number of chipsets and ASICs that embed AI accelerators but their compute performance is limited by the memory access operation. Edge compute systems that need to support machine learning inference up to GPU-level compute performance will require high performance dynamic RAM (DRAM) solutions. This is not just in terms of megatransfers per second (MT/s) throughput but the need to incorporate more efficient memory bank usage and therefore improve overall effective bandwidth. For example, Micron compared DDR4 technology at an equivalent data rate of 3200 MT/s vs their latest DDR5 system-level simulation and indicated an approximate performance increase of 1.36x greater in effective bandwidth. At a higher data rate, DDR5-6400 will have an approximate performance increase of 2.22x – double the bandwidth as compared to DDR4-3200.

Micron memory and Storage are the Heart of the Intelligent Edge

The key to creating more intelligence in connected edge devices lies in enabling data to be stored, moved, processed and secured efficiently. Micron's expansive portfolio of memory and storage products helps solve data challenges.

Low-power DRAM
High-performance memory quickly processes data in battery-powered devices so they can return to an energy-efficient sleep mode.

NOR flash memory
Low-latency high-endurance memory delivers reliable code and data storage for applications requiring fast boots, random-read access and low-density data storage.

Automotive-grade memory
Multichip packages and low-power DRAM are suitable for advanced driver-assistance systems (ADAS) and levels 4 and 5 autonomous vehicles.

MicroSD cards
Consumer- and industrial-grade storage cards deliver blazing fast file transfers with high endurance and video recording redundancy.

Micron Authenta™ technology
Chip-level security guards against unauthorized access and malware.

Mainstream DRAM
Affordable memory provides high-speed data retrival for edge devices.

3D managed NAND
High-capacity and high-performance storage holds massive amounts of data where it is created

Figure 5. Micron's comprehensive portfolio of memory and storage products help solve tomorrow's data challenges.

[Source: Micron Technology, Inc.]

Micron delivers the Intelligent Edge essentials

Micron is a world leader in memory and storage solutions. The company has pioneered most of the significant technology advances over the last 40 years and today it offers a comprehensive portfolio, that starts with foundational DRAM, NAND, and NOR Flash memory, and extends to SSDs, MCPs, and other semiconductor systems. This capability has enabled Micron to overcome the challenges outlined earlier: the space restraints of edge products, the demanding memory requirements of edge-based AI and the ability to anticipate tomorrow's data challenges.

Summary

The intelligent edge is based on an architecture that brings compute and analytics closer to the source of data. Primarily to be able to react to real-time events but also to reduce the cost of sending data to the cloud. Network systems that were once used as data aggregators and protocol translators are being replaced or supplemented with more intelligent solutions with on-premises compute and storage to bridge between the real-time systems and the IT domain.

AI at the edge will continue to evolve and we will see new technologies in hardware as well as incorporating edge-purposed software platforms. These new workloads may leverage AI accelerators and as such will need to consider memory transfer rates, amount of storage, as well as overall efficiency in constrained environment.

Memory has at times been seen as a commodity, a resource that is always available. This report highlights that this is not the case, and that memory and storage are a mission-critical consideration in intelligent edge architectures.