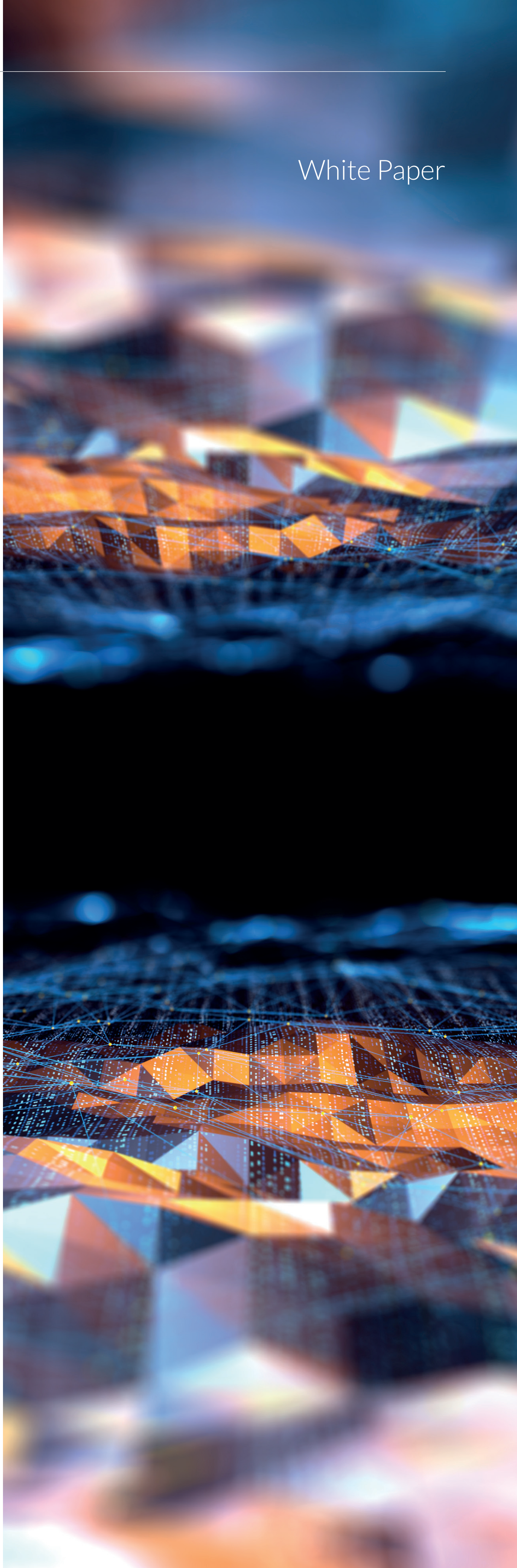


The New Compute Continuum is Powering the Future of IoT

Index

Introduction

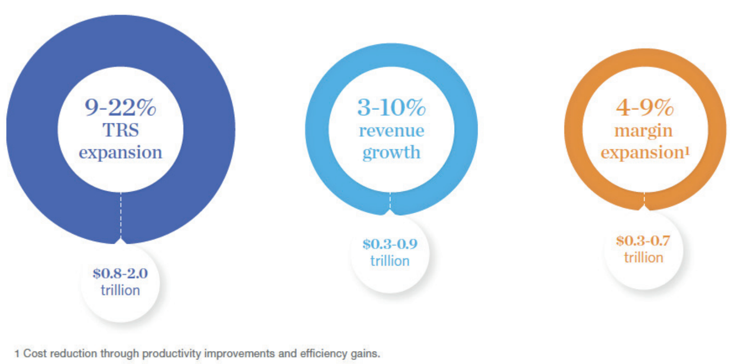
The accelerating growth of Internet of Things (IoT) devices and the data those devices and systems capture has spurred the industry to reconsider where it computes data. As a result, rather than continuing to rely on a traditional data center compute model, the industry has embraced the notion of a compute continuum: Putting the right compute resources at optimal processing points in the system – from cloud data center to edge systems and endpoint devices.

Make no mistake, the shift we are seeing is radical. It is also vital. In the traditional cloud model, enterprise data is directed straight to the cloud for processing as that is where the majority of the heavy compute intelligence is located. But, in the transformative data-driven era we now live in, this is increasingly not a viable long-term economic model due to the volume of data and new emphasis on security, safety, privacy, latency and reliability.

Today, data insights drive near real-time decisions directly affecting the operation of factories, cities, transportation, buildings and homes. To cope, computing must be fast, efficient and secure, and that generally means putting more compute firepower closer to the data source. It builds the case for more on-device endpoint computing, more localized computing with a new breed of network and private edge servers and sensible choices over which workloads need to remain in cloud data centers. The good news is that recent technology advances have converged to make this digital transformation economically viable.

The benefit of this new matrixed thinking around compute optimization is that it:

- Improves responsiveness by reducing decision-making latency
- Increases data security and privacy
- Requires less power
- Uses less network bandwidth
- Maximizes efficiencies, reliability and autonomy
- Reduces infrastructure and operational costs



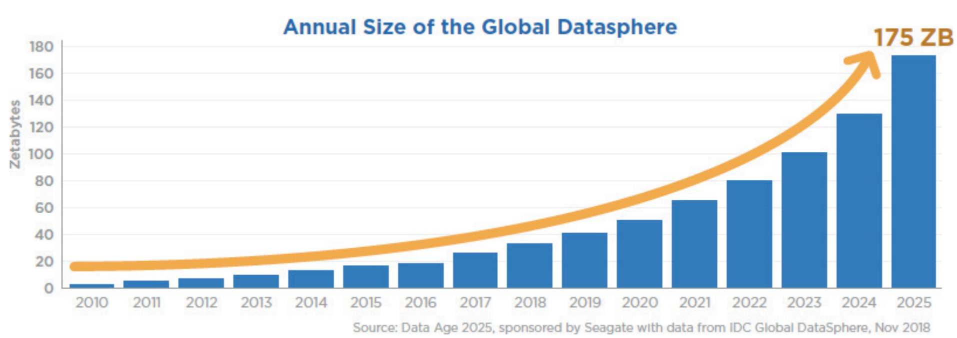
Successful embrace of the computing continuum could yield more than $1 trillion in benefits in productivity gains, revenue growth and total shareholder return (TRS). Source: Tech-enabled transformation: The trillion-dollar opportunity for industrials. *McKinsey & Company, September 2018*

The combination of these benefits ultimately delivers new computing capabilities and opportunities to organizations that embrace the transformation. In fact, the upside to industries is enormous. A recent McKinsey & Company study estimates that the industrial sector alone – including automotive, transportation, aerospace and defense, and the general factories subsector – will enjoy more than $1 trillion in new value (new revenue and expanded profit margins) by applying data, analytics and IoT technologies[1].

How to take advantage of the compute continuum

As we have stated, the traditional cloud-based model is running into scalability challenges due to the volume of data and the demands of new data insight-driven usage and business models. Unlike today, IT (information technology) and OT (operational technology, which has responsibility for automation and factory operations) will merge into one network and one mode of operation where IT data, OT data, social data and customer data are melded together to extract business insights.

That assertion is backed up by predictions that global internet protocol (IP) traffic will soar three-fold from 2017 to 2022, rising at an annual rate of 26 percent per year. Total IP traffic could be 396 exabytes (EB) a month by 2022, according to Cisco's Visual Networking Index. Global mobile traffic may rise even faster: The index suggests a seven-fold jump in the same period to a total of 77.5 EB a month of mobile traffic, from just 11 EB in 2017. Taking into consideration additional forms of traffic, the research firm IDC predicts that the Global Datasphere will grow from 33 Zettabytes in 2018 to 175 zettabytes by 2025[2].



The explosion of IoT-enabled data is straining the existing data center compute model. *Source: The Digitization of the World From Edge to Core. IDC, November 2018*

Simply put, building greater numbers of traditional data centers is not a long-term sustainable answer.

In this whitepaper, we lay out the definitions, trends and requirements, architecture and products and benefits of the new compute continuum. We think this is important as it will help provide a coherent definition of terms – particularly around the use of 'the edge,' and it will deliver a fresh understanding of the new compute capability available. Ultimately, we hope the information in these pages will help industry to rethink how products, systems and networks are designed, allowing companies to take best advantage of the coming data-driven business opportunities.

Definitions

The cloud

Cloud data center computing is the relatively centralized availability of computer system resources, especially data storage and computing power, usually located in large internet-accessible data centers located great distances from the user. Here, vast arrays of servers can be allocated according to a user's needs to handle their elastic compute requirements, delivering coherence and economies of scale that help users manage compute costs.

Many companies across virtually all verticals have realized they can transform their business by mining vast new sources and amounts of data to make insightful decisions.

Edge computing

As a term, the 'edge' means many things to many people today. In this paper, we define the edge starting where endpoint devices plug into the network and ending at cloud data centers. The edge therefore encompasses everything between endpoint devices and the cloud. In many ways, the edge becomes an offload layer from the cloud. It's closer to the endpoint devices (so latency remains lower) but contains powerful enough compute to handle most workloads – including artificial intelligence (AI) model training and machine learning (ML) inference (comparing harvested information from devices with a trained AI model).

We might think of the edge as a new kind of data 'rapid-reaction force' built for today's and tomorrow's compute needs. Indeed, edge servers are enjoying enormous growth rates, as compute gets more distributed along the continuum to cope with modern needs.

For example, a state-of-the-art factory today may have thousands of sensor-driven endpoints. Each assembly line will have many machines and one compute service that is processing data from all those endpoints. The factory floor as a whole could have a 5G base station and an edge server to process data from all assembly lines in near real-time and efficiently. Different deployments will have different combinations of compute resources depending on needs. Some may only require endpoints and edge. Others may require endpoint, edge and cloud so sifted data can be shared between factory sites, and with a company's wider supply chain partner ecosystem. This task-specific computing strategy enables flexibility.

Endpoint

Endpoints are where data is harvested and primary compute frequently happens. Endpoint devices are becoming far more capable for many applications, ranging from healthcare to industrial to automotive and autonomous vehicles to mobile devices and more. The right type of compute is required on the endpoint to extract information locally from the harvested data while satisfying the operational requirements of digital transformation listed above. This is the fundamental reason we see more products in the industry with digital signal processing (DSP) and ML integrated next to general-purpose compute.

Why now?

Many companies across virtually all verticals have realized they can transform their business by mining vast new sources and amounts of data to make insightful decisions. They understand that doing so ahead of the competition leads to long-term success. To get there, they want to leverage more of their computing resources and create IoT networks to monitor shop floors or farm fields and so forth.

This transformation is underway. Gartner estimates that in 2018, nearly 90% of enterprise-generated data was processed in a centralized data center or cloud. The same study projects that by 2022, only 25% of data-processing will be centralized, with the remaining 75% occurring somewhere in the compute continuum outside the cloud[3]. Capturing all that data will be billions of new edge and endpoint devices, powered by continual improvements in compute technology that allow increasing processing power to be cost-effectively deployed. Mobile devices, for example, are expected to grow from 1.7 billion to 2.2 billion in the decade ahead, according to Strategy Analytics and Arm forecasts. Smart IP cameras will jump from 160 million to 1.3 billion, according to Gartner and Arm, and AI-enabled devices will rocket from 300 million to 3.2 billion, according to IDC's Worldwide Embedded and Intelligent Systems Forecast, 2017-2022.

This explosion of hardware means that IoT data volume is growing much faster than bandwidth is to the cloud. Consider just one use case: In the future, Arm anticipates 500 million high-definition (HD) image sensors will produce 300 EB of data per month. That will exceed available bandwidth to the edge network, let alone the cloud. A portion of that data (say, security cameras) contains more data than can be efficiently or cost-effectively computed on the endpoint device (the camera), but is much better handled in an edge device, near the endpoint device.

Technological trends

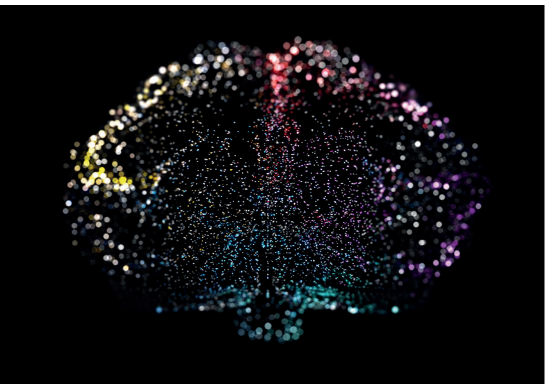
Let's now examine the technological trends that are contributing to the success of the compute continuum.

Compute

On the compute side, developers never had it so good. Successive generations of processors at the heart of IoT devices have delivered 32-bit and 64-bit low-power, general-purpose compute in addition to more recent signal processing and machine learning functions. Traditionally, stand-alone compute blocks were required for each type of compute, which restricted product development to a few experts. These days, all developers regardless of experience have easy access to the type of compute necessary to get the job done as close as possible to the data source. The emergence of 64-bit devices for the edge opened the door for traditional cloud applications to migrate to the edge. With a cloud-native development model, a developer now can map out the best point to locate the data compute along the continuum in the edge all the way up to the cloud.

Signal processing

When the ML wave began, few people envisioned that having low-power compute with signal processing features would have such an impact. The traditional approach to deploying ML at the edge is to let the neural network consume raw data. The side effects of this approach are that the network requires a large amount of memory and compute. A new school is emerging that does the ML processing in two steps. First, a signal conditioning or data clean-up phase, followed by the neural net computing. By doing a first pass of signal processing on the raw data, the network compute and memory requirements are dramatically reduced. The network now acts on clean data, which is a less complex task to process.



Taking AI to the cloud is inefficient today; we need to take the processing to the data, not the other way around.

The AI/ML influence

Another major contribution to the rise of compute continuum is artificial intelligence and machine learning. Because of relentless improvements and innovation with processors, AI and ML algorithms which once were confined to be computed in the data center cloud are now being handled at the edge and on even the smallest of endpoint devices. This in turn is changing how we think about AI.

AI grew up in the cloud. The existence of effectively unlimited computing in the cloud for almost nothing has led to the rapid evolution of algorithms that allow users to gain access and insight from data. But taking AI to the cloud is inefficient today; we need to take the processing to the data, not the other way around.

Today, over 95% of AI-enabled devices are in mobile, smart home and IoT market segments, with the remaining 5% in infrastructure, according to Arm estimates. From that data, we can conclude that over 90% of the AI-enabled devices being shipped today are based on the Arm architecture.

Forty-three percent of global telecommunications decision makers whose firms are using edge computing believe that it will give them the flexibility to handle present and future artificial intelligence demands.[4]

At the same time, AI and ML are being deployed already on CPUs, with more than 40 percent of respondents in a recent Arm engineering survey[5] indicating they're deploying algorithms there rather than in specialty or offload engines.

These demands require a new layer of edge processing infrastructure as well as the supporting software services to manage it:

- Advances in ML software techniques allow the ML operation to be run at the endpoint or the infrastructure edge in very small, energy-efficient modules ("Tiny ML") as opposed to the hyperscale data center.
- ML can now be executed on endpoint devices and edge devices, delivering faster real-time insights for a much better user experience and increased revenue.

- Heterogeneous compute
  - Various architectures – CPUs, GPUs, NPUs, etc. – can be combined in many different possible heterogeneous systems. For instance, always-on functionality might be provided by small and highly efficient processors, coupled with much more powerful processors and accelerators in the same physical device.
  - This is important because edge computing workloads vary greatly, from security (62%) to network traffic management (59%) to data caching (39%) to connected application acceleration (27%), according to TECHnalysis Research.[6]

Security

After a few years of discussions and explorations the industry has converged on the need for and the means to deploy secure IoT systems. The Platform Security Architecture (PSA) – a robust and standardized framework for security – enables devices to be born securely. Many governments have already issued security directives for IoT or are in the process of doing that. We are far ahead from when Arm released the IoT Security Manifesto a few years ago.

Connectivity

System developers have the best choice ever in terms of connectivity. Bluetooth, WiFi, NB-IoT, LPWAN, mesh, whatever the application topology, range and power requirements may be, there is a radio to fit those needs. To make things even better, 5G has started to roll out on a global basis. 5G was designed to extend mobile networks for the IoT use case with support for billions of nodes with small packets at low latency and with support for quality of service. The convergence of connectivity, AI/ML and easy developer access to a variety of compute makes digital transformation possible for all industries and organizations.

Networking evolution

WirelessHeart, industrial ethernet (CC-Link IE, EtherCAT, Ethernet/IP, Modbus TCPJlkj, POWERLINK, PROFINET, and SERCOS III), and other manufacturing networking protocols have for years ruled the roost inside factories. The auto industry and others have introduced their own networking protocols to bridge the gap between the traditional networking world and particular industries. Networking professionals will need to set up interfaces, as cloud providers have done for application developers, to allow professionals to spin up and down networking services and connections as necessary.

Consider the example of networking in the connected-car space. The endpoint device is the engine control unit (ECU), which communicates with an edge gateway inside the vehicle for services such as voice and gesture control, vehicle information, ADAS and power train data. To access larger, more data-heavy information, the vehicle gets infrastructure services by communicating with a µMEC (multi-access edge computing), a small form factor hardware/software platform for especially smart city services. It can use 5G, WLAN or fiber connection and be installed on light poles, vehicles, etc. The µMEC proof-of-concept is based on LuxTurrim5G and open source components.

As these economic and technological drivers transform networking, they're changing enterprises as well. The edge and endpoint compute continuum collapse several stages of the traditional IoT chain into a converged box and brings data center and cloud compute resources on-premise. This in turn brings the IT world and OT worlds closer together.

Requirements such as latency, security and privacy and autonomy are among the drivers pushing compute from cloud to edge and endpoint.

Drivers and requirements

Latency

As the IoT grows and we propel more intelligence to edge and endpoint devices, latency becomes more important. Take platoon driving in autonomous vehicles. Each vehicle needs millisecond response times when it comes to braking or otherwise reacting to road and or other vehicle circumstances and can't wait for the cloud to analyze a real-time situation before acting. Additionally, all IoT applications where there is potential danger to human life have strict requirements on response/action latency. For example, a command to shut down a large motor has to be issued within milliseconds from detecting failures on the factory floor.

Data privacy and residency

Whether due to the sensitive nature of the data, company policies, or in response to regulatory requirements, many organizations may prefer to keep sensitive or more valuable data on premises or locally. Similarly, users prefer to keep as much personal information as possible constrained to their endpoint devices. The infrastructure edge and endpoints allow for the processing and storage of that data within the constraints of such policies or regulation.

Power consumption

Power consumption matters across the compute spectrum, whether it's a small battery-powered or energy-harvesting endpoint device or an edge server plugged into a central power source. When it comes to billions of new IoT devices, a few milliwatts per device, multiplied across all those devices adds up. In addition, transmitting data drains the battery more than computing the data, so the ability to transmit less data by performing local compute makes the battery last longer.

Bandwidth

Data transit is expensive, and IoT data volume is growing much faster than bandwidth to the cloud. A five-second sound clip in a voice-activated task sent to the cloud requires 330 giga operations/sec (GOPS) to perform the automatic speech recognition, transmitting the audio data to the cloud at 32 kbps. It requires 600 GOPS to perform natural language processing on the same clip sending it back to the endpoint device. This doesn't include the compute already performed on the device for voice detection, keyword spotting, beamforming, echo cancellation and noise suppression. So larger data volumes require larger networks which then cost more and consume more power.

Reliability/Functional safety

Edge and endpoint solutions enable enterprises to address crucial reliability and safety requirements. Managing mission-critical machinery and equipment requires data capture, processing, and control at the edge to reduce the potential for unanticipated downtime or malfunctions that can quickly add up to millions of dollars in repair costs and lost revenue — not to mention the safety of onsite workers. Any operation or system that may cause harm to humans directly or indirectly is usually designed with functional safety in mind. In the worst case, the system is allowed to fail, but under no condition is it allowed to fail in a manner that is harmful to humans.

Autonomy

By leveraging compute power at the edge, vendors can continue to provide service when the internet backhaul, or other points, are temporarily unavailable and data can't reach the cloud. (And in some cases, constant connectivity may not be required for a given system). By strengthening the resilience of the larger architecture, the impact of outages is mitigated. Consider, for example, an oil rig at sea which needs to make safety and cost-critical decisions, whether or not it's connected to the cloud.

Cost

Reduction in network infrastructure costs (bandwidth, transmission and operations) is another key benefit to the compute continuum. An offshore drilling platform uses expensive satellite transmission to transfer data to the cloud, and therefore lots of data gets expensive quickly.

Security/node protection

Any discussion regarding IoT is incomplete if it does not address security. Edge and endpoint devices bring forward further challenges to security since they are not installed inside protected data centers. An endpoint – edge – cloud security architecture is required for all deployments to ensure reliable long-term operation.

Conclusion: The open road ahead

What can we expect to see over the next few years? Major digital transformations across all industries, the economy and life. Secure endpoint-edge-cloud IoT solutions deployed across factories, cities and homes. A fundamental transformation of the global economy to a more resource efficient, climate friendly and customer focused mode of operation.
This transformation will take place thanks to new players who, emboldened by the design flexibility and choice that compute continuum solutions provide, deliver new innovative products and build new consortiums to expand the compute continuum and coalesce around new standards to speed up and simplify deployments. We're seeing this type of collaboration already, as evidenced by groups such as MEC/Linux Foundation Edge/IIC/Edge Computing Consortium.

Rethinking how we deploy compute resources across this vast continuum of electronics, from data center to edge to endpoint, will fuel tremendous innovation and technological advancements in the coming years.

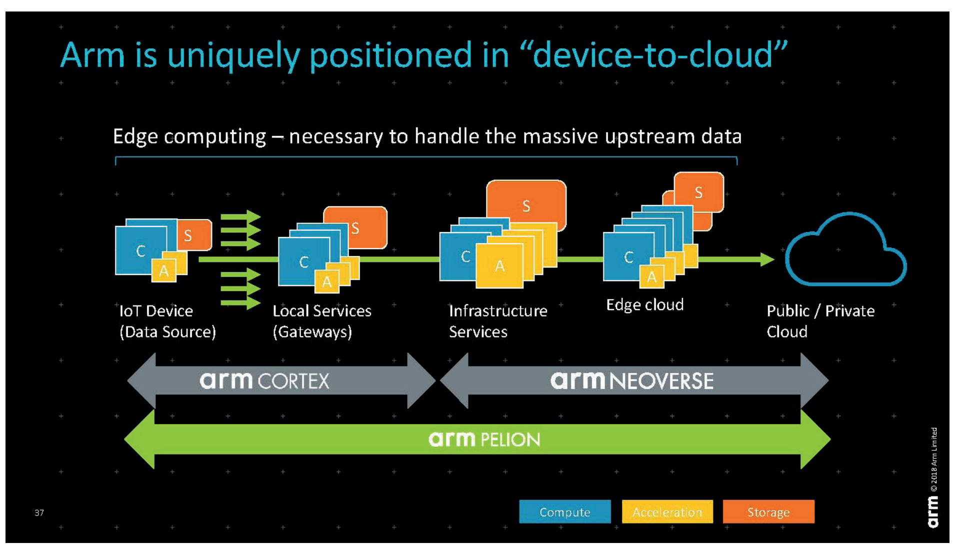
Endpoint and edge compute bring:

- Better responsiveness by reducing roundtrip latency
- Increased data security and privacy
- Reduced power consumption
- Better use of network bandwidth
- Efficiencies, reliability and autonomy to systems
- Lower deployment and operational costs.

By digitally transforming operations more completely with choice and solutions provided along the compute continuum, enterprises will see improved utilization and efficiency and new business models powered by actionable insights from converging new data sources.

The McKinsey study cited earlier notes that "successful transformation across the whole industrial sector" could improve return to shareholders by as much as $2 trillion, an increase of 9 to 22 percent. This includes up to nearly $1 trillion in revenue growth (an improvement of 3 to 10 percent), and potentially $0.7 trillion in margin expansion from efficiency gains (an improvement of 4 to 9 percent).

Just as cloud computing created vast new applications and services that delivered efficiencies on an unprecedented scale, the compute continuum will deliver even greater value. The convergence of various key technologies enables greater engineering productivity, faster innovation, more depth of insight, more possibilities for partner collaboration, new business models and new data-driven revenue streams that have not yet been identified. And as with previous industrial paradigm shifts, front runners stand to gain a proportionally larger economic benefit.



How Arm Technology Powers the Compute Continuum

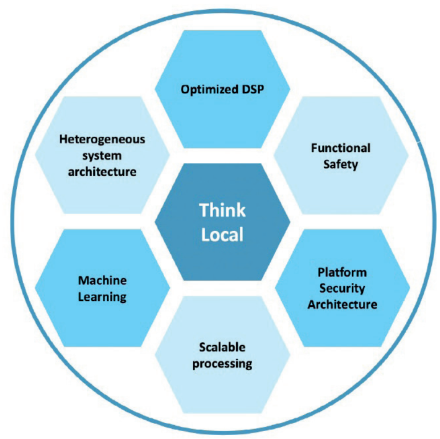
Scalable heterogenous compute

Each year, Arm invests heavily in expanding its offering of processor IP across the compute continuum to give partners the widest and most flexible options in designing their IoT systems. Arm processor IP scales from the extreme low power and high efficiency to the highest-performance devices.

At the high end, Arm Cortex-A devices have served as the foundation for the mobile revolution of the past 20 years, highlighted by 2019's introduction of the Arm Cortex-A77, a third-generation premium CPU based on DynamIQ technology.

Arm Cortex-M devices power the most energy-efficient solutions embedded devices today and are being increasingly deployed in endpoint configurations to leverage the power efficiency that many systems at that point in the continuum demand. For example, the Arm Cortex-M33 is optimized for cost and power-sensitive microcontroller and mixed-signal applications and is designed to address embedded and IoT markets, especially those that require efficient security or digital signal control.

Arm Cortex-R devices target real-time system requirements. They introduced virtualization at the highest security level while retaining the Protected Memory System Architecture (PMSA) based on a Memory Protection Unit (MPU).In October 2018, Arm launched the Arm Neoverse brand and product portfolio designed to deliver the cloud to edge infrastructure for a world of a trillion connected devices. At that launch, Arm promised a 30 percent performance uplift for each new product generation. In February 2019, Arm launched two new products in this portfolio: Neoverse N1 for cloud to edge compute and Neoverse E1 for data plane and blew away the promised performance improvements by achieving a 60 percent compute performance increase with even higher uplifts for some workloads.



The underlying principle of the compute continuum is the concept of "think local," which combines these key elements.

Security / Platform Security Architecture

Arm has also expanded its pioneering work in security with security solutions that support edge compute. The Platform Security Architecture, or PSA, was introduced by Arm in 2017. It is an open methodology for analyzing threats, architecting secure designs, and then implementing secure solutions. This architecture offers a route to improved security, helping partners to understand the requirements of designing, developing, and securing IoT devices at the endpoint, no matter their role.

Arm has hardware and software IP to address these four main threat types. These include technologies like Arm TrustZone help to protect against software threats; hardware IP like Arm CryptoCell helps protect against communication threats; Arm Cortex-M35P, SecureCore processors and Arm CryptoIsland products help protect against physical attacks. Arm's Kigen family of products helps protect against communication and mitigation attacks. You can find more details about these threats and the relevant IP on our website here.

Arm Corstone IP offers SoC designers a great solution to build secure designs faster. The product range addresses constrained, mainstream and rich IoT designs.

Endpoint-to-Cloud solution: Arm Pelion IoT Platform

One of the liberating benefits of this compute continuum is that heterogeneous, dispersed processing gives companies a chance to better leverage the data they're capturing, where it's captured, to help them transform their businesses. To capture the data, the endpoints need to be connected and managed in order for the data to be trusted. The Arm Pelion IoT Platform is a flexible, secure, and efficient foundation spanning connectivity, device, and data management. It accelerates the time to value of IoT deployments by helping you easily connect trusted IoT devices on global networks, invisibly administer them, and extract real-time data from them to drive competitive advantage.

Functional safety

Arm has long advocated a holistic approach to functional safety, to look at the 'big picture' and ensure that there are no loose ends or cracks in designs that often contain hundredsof different hardware and software elements that need to comprehensively deliver the safety functions. Arm's Functional Safety Partnership Program showcases a range of functional safety partners who specialize in areas of software and tools, design services and training services.

Edge ecosystem and Project Cassini

To ensure a cloud-native experience across a diverse and secure edge ecosystem, Arm in 2019 announced Project Cassini. Project Cassini, an initiative with our ecosystem, will develop platform standards and reference systems that establish a cloud-native software experience for a secure infrastructure edge.

Leveraging the ecosystem to drive success

Arm is at epicenter of one of the most extensive edge partner ecosystems in the world today. Arm's silicon partners have shipped nearly 150 billion chips since Arm's founding founding in 1990. More than 50 billion of those have been in the last four years, and we expect another 100 billion in the next four years. The global Arm ecosystem delivers a healthy choice of competing cutting-edge products at affordable prices, as evidenced by the many use cases mentioned earlier. It also functions as a vibrant open source community especially around software and Arm Mbed OS. Here are some examples:

Transportation

A Silicon Valley data company wanted to detect delays in Caltrain, the regional train system that links the Silicon Valley with San Francisco. They set up a system monitoring the trains running on a route to detect Caltrain among a host of other trains like freight, light rail vehicles etc. To detect Caltrain in this system, a classification of different trains was needed which was achieved by deploying TensorFlow on Arm to create this system using a Raspberry Pi 3. The system cost $130 to build and required little knowledge of TensorFlow to deploy. This decision was made in order to keep the system affordable and, because of the absence of high speed internet, avoid sending images to a central server, which would cause massive delays.

Healthcare

Respiro from Amiko uses an Arm processor to power a sensor designed to fit inside the space and memory constrained inhalers that asthma patients are already using. The module uses machine learning to monitor inhalation technique and capture parameters such as inhalation flow rate, flow acceleration, inhalation volume, inspiration time and more. The intelligence powering this capability, the machine learning, is happening on the Arm processor in the device without the need to connect to the cloud.

Industrial

Machbase offers a unified database platform that includes three products: Edge, Fog and Cluster/Cloud. It works on the entire spectrum of Arm products and their licensees, including many edge examples, such as NVidia Jetson, Raspberry Pi 3, and ARTik. Arm's low-power, high performance technology enables Machbase to run hundreds of databases on Arm processors, from processing on smart factory floors to a utility's distribution infrastructure to an automotive ecosystem.

Smart cities

Neurotechnology's Face Verification SDK provides biometric identity verification for large-scale, high-security applications. It allows face verification to be performed offline, simplifies integration of facial authentication applications for PCs and mobile devices, performs liveness detection to reinforce anti-spoofing measures. All biometric data is kept on-device.

Smart homes

Intuition Robotics' ElliQ is a social companion robot that helps the elderly. The company's partner, Brodmann17, worked with them to speed up the performance of the robot.
They have a software solution that speeds up Arm CPUs. The SoC is a Qualcomm Snapdragon 820 which is Armv8-A. The robot sits on table and interacts with people so effectively that it can find where a person is speaking from and can turn its head to the person's direction.

Smart buildings

AquaSeca's device attaches to water pipes to check for leaks and other anomalies. The next generation of this device will have ML being done right on the device, making it much more efficient and scalable with most data staying on the device.

Infrastructure

Arm is enabling seamless connection to give a cloud-like experience across all of edge IoT through enablement of products such as HPE's Procurve switch line, which will be based on NXP running Arm Cortex-A72/A53.

Footnotes

1 Tech-enabled transformation: The trillion-dollar opportunity for industrials. McKinsey & Company, September 2018

2 The Digitization of the World From Edge to Core. IDC, November 2018

3 Technology Insight: Edge Computing in Support of the Internet of Things. Gartner, July 2017

4 Global Business Technographics Mobility Survey. Forrester Research, September 2018.

5 What's Powering Artificial Intelligence? Arm research, May 2019.

6 Computing on the Edge: A Survey Report. TECHnalysis Research, November 2017.