Deloitte.

The Economics of High-Performance Computing

Cloud computing has yielded enormous value for enterprises.

The business case and patterns are clear for leveraging general purpose compute on the cloud for massive tasks such as enterprise resource planning (ERP), enterprise asset management (EAM), and back office functions. Doing so has allowed businesses to capitalize on the scale and security of the cloud as well as inherent new cloud capabilities while allowing them to remove distractions from non-core functions such as managing large groups of data centers for common tasks. Without question, the economics of general-purpose computing in the cloud are unmatched.

When it comes to the most demanding computational tasks, however, such as artificial intelligence (AI) and machine learning (ML) for things like research, customer experience, and supply chain management, a somewhat different approach may be required. In the case of high-performance computing (HPC), where the computing takes place impacts both what can be achieved and the cost to do so. The challenge is to think through which tasks are best suited for the cloud and which innovative and differentiating applications require some of the most sophisticated edge hardware.

Considering differentiating capabilities

When we look at the computing infrastructure landscape, we note three large categories of capabilities:

1. General purpose compute
2. Software-as-a-Service (SaaS) applications
3. Differentiating capabilities that require specialized and/or edge hardware

General purpose cloud computing is ideal for applications that enable processes within an industry or are somewhat standard across industries. Examples include common ERP and EAM solutions, where some latency is acceptable and increases in data flows can be easily accommodated with on-demand scale. These kinds of applications have been tested, proven out, and refined, and today, while they offer great value and are vital for business operation, capabilities for data processing and readout do not create true market differentiation. Indeed, these capabilities are available to all enterprises, and as such, they are effectively table stakes investments. With this, it makes economic sense to acquire them at the best competitive price and rely on cloud providers to manage scale and evolving AI capabilities.

Meanwhile, customer relationship management (CRM) functions for marketing, sales, service, customer care, and field service can directly impact customer and market perception and need to be carefully managed. These functions, however, are also common within and across industries. While the business processes, automation, and the human capital that manages it are necessary for competitive differentiation, these functions are sufficiently mature that it makes most business sense to leverage best-in-class SAAS applications.

High performance AI/ML for industry-leading or world-changing applications is different. The most demanding computational work in things like gene therapy, drug research, deep image analysis, and deep learning require specialized hardware, such as GPUs. There are also separate edge applications requiring specialized compute near the data sources. We see this in smart and secure spaces including theme parks and stadiums, cruise ships and resorts, schools and airports, and factories and transportation, to name a few. In these spaces, real-time decisions are required across thousands of sensors per facility, and this requires edge hardware enabled by GPUs.

For instance, a stadium could use 5,000 or more cameras for surveillance, and the data they capture is needed to make real-time decisions on access control, safety, way finding, and commerce. Each camera sensor puts out between 1Mb/s to 15Mb/s.

The collective data across all sensors of one facility could be in the range of 10-30Gb/s, and most of the data is useless or non-actionable. Edge GPU infrastructure can compress, filter, and elevate actionable intelligence to operational systems in near-real-time, avoiding costly and unnecessary network and cloud bottlenecks.

This table shows the potential for **6x reduction in training costs** and **5x reduction in training time** by using GPUs on the cloud, as opposed to using traditional CPUs on the cloud.

Type	# Instances	Hardware per Instance	Instance Type	Amazon EC2 Cost per Hour	Amazon EMR Cost per Hour	Training *(Minutes)*	Training Costs
GPU	16	4x T4	g4dn.12xlarge	$3.912	$0.27	6	$6.69
GPU	6	8 x V100	p3.16xlarge	$24.48	$0.27	5	$12.38
CPU	16	64 vCPU	r5a.16xlarge	$4.608	$0.27	33	$42.93

Source: https://aws.amazon.com/blogs/big-data/improving-rapids-xgboost-performance-and-reducing-costs-with-amazon-emr-running-amazon-ec2-g4-instances/

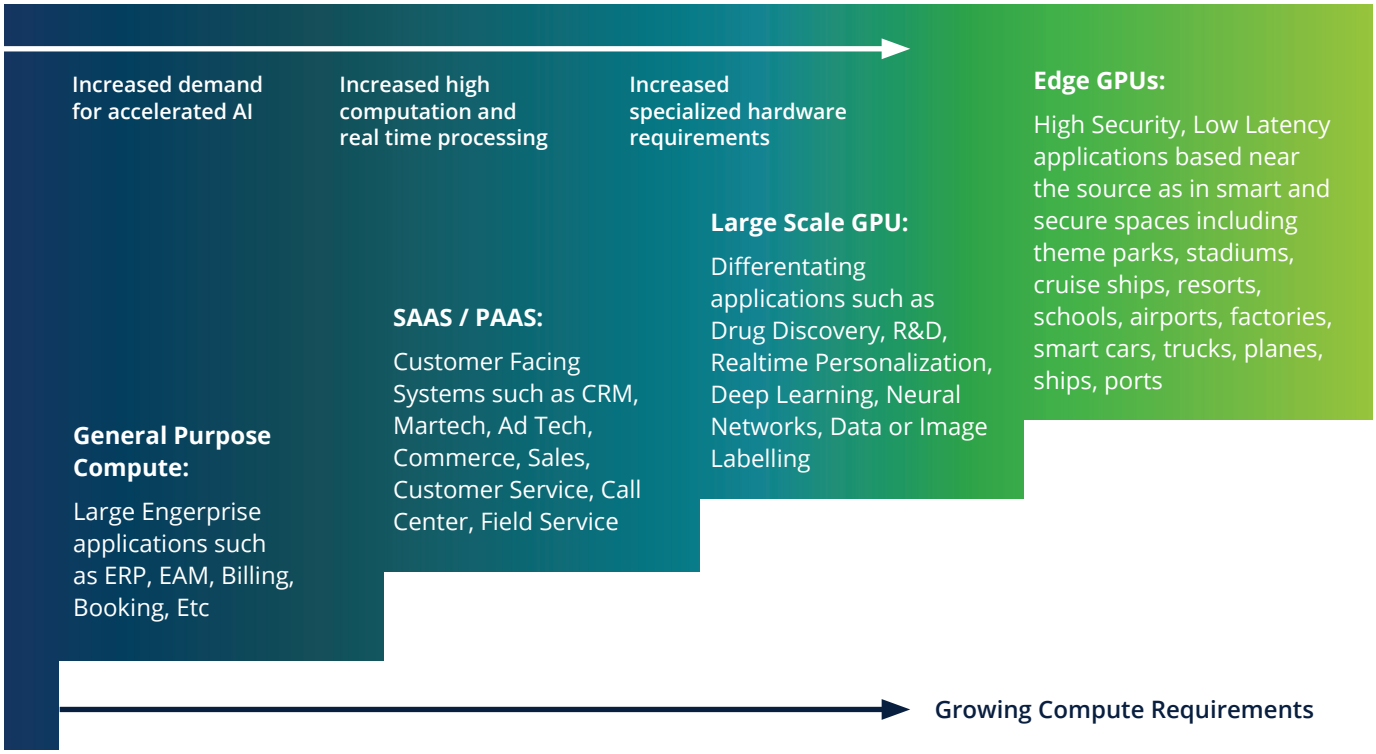
When it comes to differing capabilities that require HPC or edge hardware, the cloud economics that work for general purpose computing begin to be challenged.

Data Interchange / Day (GB)	1,000	50,000	500,000
Cost / GB	$0.12	$0.10	$0.08
Annual Cost	$43,800	$1,825,000	$14,600,000
5 Year Cost	$219,000	$9,125,000	$73,000,000

Consumption	Small	Medium	Large
Cost /GB(0.08-0.12)	$0.12	$0.10	$0.08
Data interchange / day in GB	1,000	10,000	100,000
Daily Cost	$120	$1,000	$8,000
Annual Cost	$43,800	$365,000	$2,920,000
5 Year Cost	$219,000	$1,825,000	$14,600,000

Depending on the volume of cloud transfer (ingress/egress), the cost of compute could be significant. As advanced models use images, streaming video, and real-time conversation, the data transfer will increase. Within some industries (e.g., life sciences), the cost of research and development infrastructure will grow.

Computational work involving large data flows and massive parallel processing present significant costs for connectivity, which includes networking costs as well as egress (approx. $0.05/GB) charges that cloud providers charge in addition to compute costs. Several industries (i.e., Life Sciences, Health Care and Technology, Media and Telecom) are dependent on significant data compute due to complexity of solutions (vision, video, drug discovery). These industries are transitioning this level of computation to an on-premise platform for cost savings and reduced latency. Over several years, an on-premise platform could be 50% lower in costs compared to cloud compute for similar data loads. A comparison of like-for-like GPU capabilities on cloud vs. on-premise shows that the cost of HPC compute on-premise is approximately 50% of the compute cost in the cloud, over a three-year period. Adding on high performance networking and cloud egress charges makes the business case substantially harder to justify.



Consider some of the HPC use cases that we have encountered that make the most economic and technical sense when performed at the edge.

Real-time Threat Detection

A large industrial plant may use thousands of cameras to monitor the equipment and property for threats. Each camera produces 10-30Gb/s, and most of the data is nonactionable. Feeding all this data to the cloud would require significant bandwidth, increase latency compete with mission critical connectivity requirements, and entail significant cloud storage at commensurate cost. Threat detection requires zero latency and dependable connectivity for immediate data inference and action orientation. Whether it is identifying a threat to a critical asset or spotting trespassers in secure areas, data latency can make all the difference in taking action. The ideal solution is that data is pre-processed with GPUs and integrated with enterprise solutions to instigate action (e.g., a security call). Model training can be enhanced using federated analytics from other industrial settings and activity.

Improving Public Safety

Law enforcement agencies worldwide have begun adopting body-worn cameras for police officers. Currently, footage from these cameras is transferred to servers at the end of a shift, which deprives agencies and the public from capturing real-time intelligence from the cameras. It also prevents the officer from using real-time AI assistance in their work. By adding edge AI hardware at a spoke location, such as a police car or fire truck, images processed in real time can fuel AI analysis that gives the officers insights, recommended actions, and other suggestions, all of which can enhance public trust and make a significant difference for officer safety. The idea is to allow the bodycam to be an intelligence-based "third-eye," and not just a recording device.

Connectivity for Self-driving Vehicles

Auto manufacturers are making serious investments in autonomous vehicles (AV), and competition is ticking up worldwide. One challenge to address is maintaining safe autonomous control in cases where wireless connectivity is interrupted (e.g., in a tunnel or a rural area). Given the impact on driver safety, consumer satisfaction, and differentiation in the marketplace, computing at the edge using HPC hardware in the vehicle is already making the long-term vision for AVs possible.

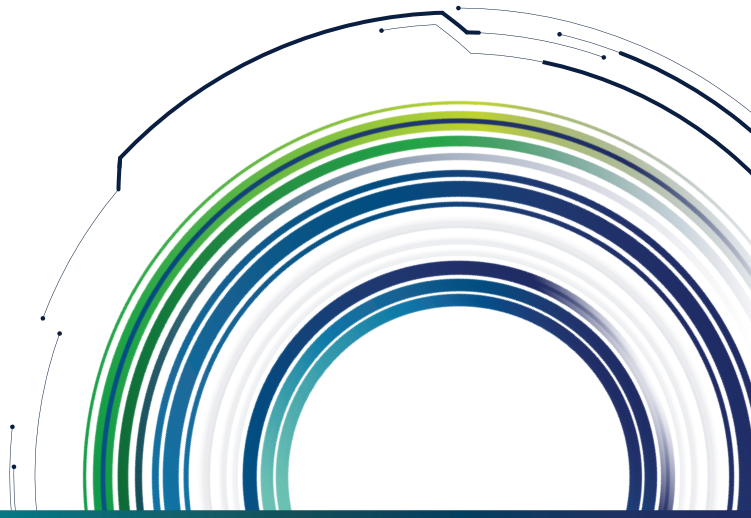
Cyberattack Detection and Prevention

AI has valuable applications for identifying malicious activity in a network. If those tools are hosted entirely in the cloud, however, a cyberattack on local hardware could prevent the AI from being used as intended. The organization needs to access the cloud to address the cyber vulnerability, and if the local network is taken offline, the organization is unable to address the issue. Computing locally helps mitigate the risk. Importantly, this does not mean AI for threat detection can never be hosted in the cloud. Instead, this example reveals how particular HPC use cases require a hybrid approach to computational hardware.

Developing a HPC strategy

Ultimately, HPC is suited for novel applications to enhance real-time human interpretation of massive datasets. These applications require intense data collection and rapid computation, and because they are new and groundbreaking, they require independent strategic consideration and customization for use cases today and tomorrow. A one-size-fits-all strategy for HPC and edge compute is insufficient.

Deloitte is a market leader in technology consulting, helping clients strategically develop differentiating solutions today and sustain that lead into the future. Our vertical specialization with cross-solution application in AI and HPC architecture ensures HPC integrates with your platform and embeds AI as a transformational shift in devising and using market leading solutions. We take a technology agnostic stance to help you identify and implement the tools that precisely fit your business challenge. Drawing on our expertise in technology infrastructure and cost optimization, we help ensure the enterprise has the optimal solution and architecture. Our holistic approach compares options across multiple variables, including infrastructure availability, infrastructure costs, opportunity costs, IT management costs, and data transfer costs, which provides a complete view of the considerations prior to spend. With our expertise in change management and human capital, we can help your enterprise stand up AI centers of excellence and drive sustainable change with an innovation mindset.



Get in touch

Christine Ahn
Principal
Deloitte Consulting LLP
chrisahn@deloitte.com

Brandon Cox
Principal
Deloitte Consulting LLP
brandoncox@deloitte.com

Goutham Belliappa
Managing Director
Deloitte Consulting LLP
gbelliappa@deloitte.com

Tanuj Agarwal
Snr. Manager
Deloitte Consulting LLP
tanuagarwal@deloitte.com