Challenging the Barriers to High Performance Computing in the Cloud

Breaking down common misconceptions about cloud-based HPC solutions

Executive summary

High Performance Computing (HPC) is crucial to organizations across industries, making it possible to drive innovation and revenue. While adoption of cloud-based HPC solutions is increasing, misconceptions about their cost, security and performance persist. It's crucial to challenge these beliefs and break down common barriers to cloud-based HPC to prevent organizations large and small from being held back by outdated, inaccurate information. HPC on AWS, powered by Intel® Xeon® Scalable processors, offers the most elastic, scalable cloud infrastructure to run HPC applications, and the range of services makes it easier than ever to get started quickly, securely, and cost-effectively.

INTRODUCTION

High Performance Computing (HPC) has transformed industries, from finance and genomics to autonomous driving and seismic imaging. The demand for these computing resources continues to rise, as organizations of all sizes strive to keep pace with innovation in the marketplace. In this dynamic environment, enterprises that turn to cloud-based HPC solutions are positioned to be more competitive and drive greater ROI. According to Hyperion Research in their study, "Cloud Computing Comes of Age," the proportion of all HPC sites that use public clouds has quintupled, from 13% in 2011 to 74% in 2018. In a 2019 Hyperion study, 40% of HPC cloud users believe all their HPC jobs could be run in the cloud, meaning there's substantial headroom for growth.

Cloud computing is helping to democratize High Performance Computing by putting powerful computational capabilities in the hands of researchers, engineers, and organizations who lack access to on-premises infrastructure or need more HPC resources to thrive. Its flexibility and scalability offer virtually unlimited capacity, eliminating wait times and long job queues. Access to new and evolving services and applications make it easy to evolve and modernize workflows, like incorporating machine learning with HPC.

With HPC in the cloud, organizations only pay for the capacity they use, and there's no risk of on-premises infrastructure becoming obsolete or poorly utilized. In addition, cloud-based services enable innovation without constraints by delivering faster results and improved flexibility. AWS gives organizations the power to create HPC clusters on demand, instead of waiting for equipment to be built—helping drive business insights and organization productivity.

Despite its advantages, some organizations remain hesitant to move their HPC workloads to the cloud. Whether it's overall cost, security, data transfer or performance, there are several perceived barriers to cloud-based solutions that prevent these teams from achieving their true potential. By exploring these barriers and demonstrating how cloud-based HPC answers each objection, this paper provides a path for organizations to begin realizing the benefits of cloud-based HPC.

PERCEIVED BARRIER 1: COST

For many organizations, the cost of running HPC in the cloud is a major concern. In a recent market survey conduced by a third party for AWS, almost half (49%) of participants said cost and cost-management were barriers. On-premises solutions are viewed as a known investment; most enterprise-level HPC use on-premises solutions for some of their needs, and basic TCO (total cost of ownership) analysis may suggest continued on-premises investments offer better returns. But hidden costs to on-premises HPC infrastructure can multiply.

Demand for on-premise HPC resources often exceeds capacity—by as much as 300%, according to Hyperion Research studies. Lost productivity due to an overutilized system has massive implications for organizations that place high value on the pace of innovation. Businesses also must consider the age of their on-premises hardware, which often cannot compete with the newest-generation cloud technology. By eliminating the need for periodic technology and infrastructure refresh cycles every three to five years, businesses using cloud-based HPC ensure innovation can move as fast as possible. Cloud-based HPC means organizations don't have to buy and support on-premises servers for pre- and post-processing services—costs never accounted for in basic TCO analysis. Also, once data is in the cloud, cost savings increase due to innate workflow advantages. Data can more easily move from one step to the next, or simulation outputs can be stored in the cloud and only post-processing visualization appears on desktop. Cloud-based solutions also make it possible for users who cannot invest the upfront capital and effort to acquire on-premises infrastructure to take advantage of HPC.

When analyses go beyond simple cost per core hour, cloud-based HPC can holistically improve ROI. It offers an elastic, scalable infrastructure with virtually unlimited capacity and access to an ever-expanding fleet of services, so engineers, researchers and system owners can innovate beyond the limitations of their on-premises resources.

COST-EFFECTIVE AWS SOLUTIONS

AWS delivers an integrated suite of services that provides everything needed to build and manage HPC clusters in the cloud, simply and cost-effectively. There are no upfront capital expenditures or lengthy procurement cycles, and the only cost is for capacity used. It offers flexible pricing models that provide significant cost savings for time-flexible, stateless workloads. AWS constantly delivers new services and features, like 2nd generation Intel® Xeon® Scalable processors with Integrated Deep Learning Boost to enable new capabilities, improved performance, and optimization for all current HPC frameworks. AWS offers cost management and analysis tools such as AWS Cost Explorer and AWS Budgets. Additionally, AWS partners like Ronin have built cost-control models on the platform.

Pricing tiers

AWS offers a free tier to gain hands-on experience, as well as three pricing options to help simplify budgets.

- On-Demand instances mean organizations pay for compute capacity by per-hour or per-second depending on which instances are run. On-Demand instances are recommended for users who don't want longer-term commitments or upfront payments, for applications with short-term, spiky or unpredictable workloads that cannot be interrupted, and for applications being developed or tested on Amazon EC2 for the first time.
- Spot instances make it possible to request spare Amazon EC2 computing capacity for up to 90% off the On-Demand price—ideal for applications with flexible start/end times, price-constrained applications, or users with urgent computing needs for large amounts of additional capacity.
- Reserved instances provide a significant savings (up to 75%) compared to On-Demand instance pricing—ideal for applications with steady-state usage, applications that may require reserved capacity, or customers that can commit to using EC2 over a one- to three-year term to reduce total computing costs.

Storage solutions

Storage options and costs are critical factors when considering an HPC solution. AWS offers flexible object, block or file storage for transient and permanent storage requirements.

- Amazon Elastic Block Store (EBS) provides highly available, consistent, low latency block storage for Amazon EC2. It helps tune applications with the right storage capacity, performance and cost.
- Amazon FSx for Lustre provides a high-performance file system optimized for fast workload processing, like HPC, machine learning, video processing, financial modeling, and electronic design automation (EDA). These workloads commonly require data to be presented via a fast and scalable file system interface, and typically have data sets stored on long-term data stores like Amazon S3.
- Amazon S3 Glacier is an extremely low cost, highly durable object storage service for long-term backup and data archive—ideal for customers looking for inexpensive solutions for infrequently accessed data.

Cost analysis tools

AWS offers a range of tools to help access, organize, understand, control and optimize AWS costs and usage.

- The AWS Cost Management dashboard shows the status of the month-to-date AWS expenditure, pinpoints the services that account for the majority of the overall expenditure, and makes it easy to understand how costs are trending at a high level.
- AWS Cost Explorer helps visualize, understand, and manage AWS costs and usage over time via an intuitive interface that makes it easy to create custom reports (including charts and tabular data) that enable cost and usage analysis, both at a high level and for specific requests.
- AWS Budgets offers the ability to set custom cost and usage budgets, and receive alerts when those thresholds are exceeded. It also offers daily, weekly or monthly budget portfolio updates via emailed AWS Budgets Reports.

Case Study: OpenEye Scientific

OpenEye Scientific, a provider of computational drug discovery software, was looking to help its customers accelerate research and cut costs—so they moved their software to the cloud. Now they use AWS to give their clients highly scalable, maintenance-free access to up to hundreds of thousands of processors to perform cloud-native computational chemistry for drug research and development. OpenEye Scientific is saving money by using unused Amazon EC2 capacity at a discounted cost. "By using Amazon EC2 Spot Instances, we saved $800,000 last year," says Craig Bruce, OpenEye's Head of Infrastructure.

"Our customers benefit from that cost savings as well, because EC2 Spot Instances give them the ability to be more flexible. Whether they need to generate images in milliseconds or perform complex chemistry operations taking many hours, they now have the cost flexibility they need."

– Craig Bruce, Head of Infrastructure, OpenEye

PERCEIVED BARRIER 2: DATA SECURITY

Concerns about cloud security are nothing new. Many industries that use HPC heavily have stringent security requirements, and it's a commonly cited obstacle to cloud-based HPC solutions. 43% of participants in the HPC market survey had concerns about data security and governance, and 42% also listed data privacy.

While some perceive security and privacy benefits to on-premises HPC, they don't account for risk management issues like aging infrastructure that increase the security costs of maintaining compliance, or for expensive regulatory compliance and certifications often required for on-premises solutions. They also often don't sufficiently account for the complex landscape that security compliance has become and the benefits of a shared responsibility model where AWS helps relieve the customer's operational burden by operating, managing and controlling the components from the host operating system and virtualization layer down to the physical security of the facilities in which the service operates.

Leading cloud-based HPC providers now make an array of security measures available, from role-based access control (RBAC) to firewalls and private networks. They are constantly updating systems to the latest technology, and because associated costs get lower at scale, can offer much stronger security systems than an individual organization could afford for its own on-premises resources. In 2018, nearly half of all new HPC private clouds were supplied by cloud-based providers according to Hyperion—a testament to growing trust in cloud data security.

Additionally, cloud-based HPC providers may offer out-of-the-box templates designed to automatically keep up with changing regulation and compliance needs, helping simplify the creation of HPC clusters and saving time and money. That's potentially significant for smaller organizations who can avoid reinventing the wheel when getting started in the cloud.

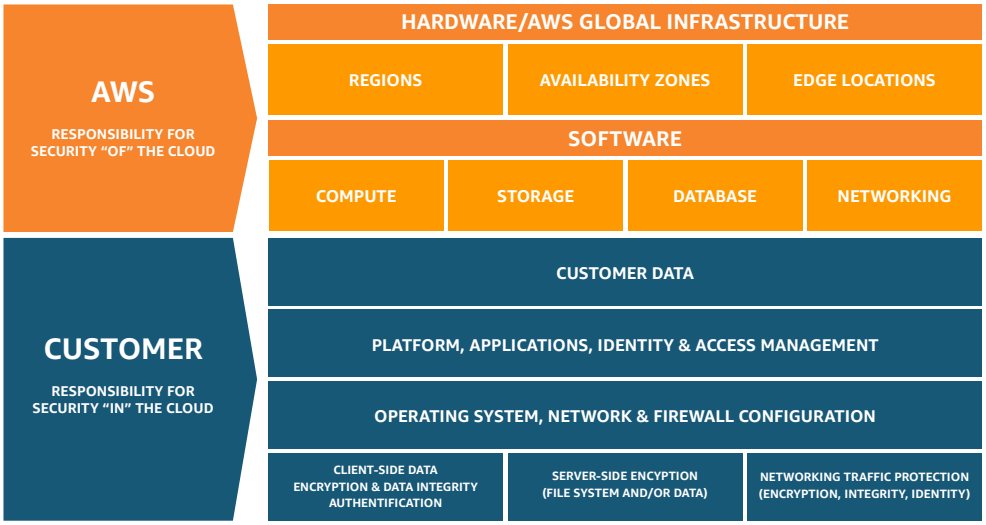
"My working assumption a year ago was that the cloud wasn't as secure as a brick data center. Now, I'm convinced it's more secure and there's less risk. We definitely get that from AWS."

– Adrian Heeson, Operations Director, British Gas

AWS SECURITY AND COMPLIANCE SOLUTIONS

Cloud security at AWS is the highest priority, and it offers several tools and services to ensure encryption, manage access, and secure regulated workloads. All data is stored in highly secure AWS data centers, and the network architecture is built to meet the requirements of the most risk-sensitive organizations. Additionally, customers maintain ownership and control of all content—they can select which AWS services can process, store and host content, determine where it will be stored, choose its secured state, and manage all access.

When a customer uses AWS services, they operate in a shared responsibility environment, where the secure functioning of an application on AWS requires action from both the customer and AWS. All institutions should explain the shared responsibility model to their stakeholders throughout the design, development, testing, and production phases of cloud adoption. Customers are responsible for security IN the cloud. They control and manage the security of their content, applications, systems, and networks. AWS manages security OF the cloud to protect infrastructure and services, maintain operational performance, and meet relevant legal and regulatory requirements.



Shared Responsibility Model

For customers storing data on AWS storage services or transiting on AWS networks, it's strongly recommended to encrypt data at rest and in transit. Encryption and data access control features are built into foundational service offerings like Amazon Simple Storage Service (Amazon S3), a highly scalable object storage service, Amazon Elastic Block Store (Amazon EBS), which provides network-attached storage to EC2 instances, and Amazon RDS, which provides managed database engines. These turn-key features provide documentation to help customers understand their data protection and the configuration options to customize system access and keys required to decrypt data.

With encryption, the confidentiality of the customer's cryptographic keys is crucial. Security depends upon where the data was encrypted, who has access to, and who protects the keys. If data is encrypted by the customer prior to being ingested into the cloud, the cloud service provider has no ability to access keys or decrypt data—the customer has full control and responsibility. However, if a particular cloud service needs to decrypt the data in order to deliver its value, both the cloud service and the data owner are able to access the keys. Customers need assurance that the cloud provider only has access to decrypt data when the customer allows it. AWS Key Management Service (AWS KMS) is designed so that no one, including AWS employees, can retrieve customer plaintext keys and use them outside the service.

Intel®-based AWS instances deliver hardware-enabled security capabilities directly on the silicon to help protect every layer of the compute stack, including hardware, firmware, operating systems, applications, networks, and the cloud. Intel® Threat Detection (Intel® TDT) is also available on 2nd Generation Intel® Xeon® Scalable processors and delivers hardware-enhanced threat detection.

Compliance tools

AWS manages dozens of compliance programs in its infrastructure, meaning segments of necessary compliance have already been completed. These user-friendly compliance tools help customers spend less time on regulations and more time running their business.

- The AWS Artifact portal offers on-demand access to security and compliance documents, including Service Organization Control (SOC) reports, Payment Card Industry (PCI) reports, and certifications from accreditation bodies across geographies and compliance verticals that validate the implementation and operating effectiveness of AWS security controls.
- The AWS CloudHSM service helps meet corporate, contractual and regulatory compliance requirements for data security by using dedicated Hardware Security Module (HSM) appliances within the AWS cloud. CloudHSM makes it possible to control the encryption keys and cryptographic operations performed by the HSM.
- Amazon Inspector is an automated security assessment service to help improve the security and compliance of applications deployed on AWS. It automatically assesses applications for vulnerabilities or deviations from best practices and produces a detailed list of security findings prioritized by level of severity.

Management solutions

Controlling access to accounts, keys and other security services like firewalls is crucial to cloud-based HPC services, and AWS tools make it simple to manage these vital features.

- AWS Firewall Manager centrally configures and manages AWS WAF rules across accounts and makes it easy to bring new applications and resources into compliance with a common set of security rules from day one. It's a single service to build and manage firewall rules, create security policies, and enforce them consistently.
- AWS Identity and Access Management (IAM) provides a robust solution for managing users, roles, and groups that have rights to access specific data sources. Organizations can issue users and systems individual identities and credentials, or provision them with temporary access credentials using the Amazon Security Token Service (Amazon STS).
- AWS Secrets Manager makes it easy to rotate, manage and retrieve database credentials, API keys and other secrets throughout their lifecycle, eliminating the need to hardcode sensitive information in plain text.

Detection and protection

AWS has a range of services to help protect data from malicious attacks.

- Amazon GuardDuty offers a more accurate and easy way to continuously monitor and protect AWS accounts and workloads. It analyzes billions of events from multiple AWS log sources, using threat intelligence feeds to accurately detect threats.
- AWS Shield is a managed Distributed Denial of Service (DDoS) protection service that safeguards web applications running on AWS, providing always-on detection and automatic inline migrations that minimize application downtime. AWS Shield Standard is included for every AWS customer.

Case Study: FINRA

FINRA—the Financial Industry Regulatory Authority—is the largest independent regulator of financial markets in the U.S., and it's dedicated to investor protection and market integrity. To respond to rapidly changing market dynamics, FINRA moved about 90 percent of its data volumes to AWS to capture, analyze, and store a daily influx of 37 billion records. By migrating to AWS, FINRA created a flexible platform that can adapt to changing market dynamics while providing analysts with the tools to interactively query multi-petabyte data sets.

"We determined that cyber security is better in the cloud than it is in privately managed data centers."

– Steve Randich, Executive Vice President and Chief Information Officer, FINRA

PERCEIVED BARRIER 3: DATA TRANSFER

Running HPC applications in the cloud starts with moving the required data into the cloud, but this process can be an obstacle for many organizations. The market study found that 41% of those surveyed were concerned about getting data into—and out of—the cloud. Commonly cited data transfer barriers are time and money, but while it may seem easier in the long run to keep data in on-premises HPC infrastructure, the investment in moving data to the cloud is far outweighed by the benefits of more flexible, agile HPC.

Moving data and HPC to the cloud improves efficiency by freeing up valuable financial and staff resources, and reduces business risks by storing data in a more resilient, secure environment. In addition, cloud-based HPC enables customers to utilize AI, machine learning and deep learning to mine all the data available from HPC simulations, narrowing the range of simulations required—meaning cheaper, faster HPC workload execution. Since newer, cloud-native HPC applications were designed to perform better on cloud-based elastic infrastructure, improved performance in the cloud can deliver better ROI.

AWS DATA TRANSFER SOLUTIONS

AWS has helped thousands of organizations migrate to the cloud, and has developed a complete, proven methodology to help simplify and accelerate migrations—whether it's one workload or thousands.

- AWS Snowball and AWS Snowmobile are data transport solutions that physically migrate petabyte- and exabyte-scale data sets into and out of AWS. They use devices designed to be secure to address common challenges with large-scale data transfers including high network costs, long transfer times, and security concerns—and can cost as little as one-fifth the cost of transferring data via high-speed internet.
- AWS DataSync is a data transfer service that makes it easy to automate moving data between on-premises storage and Amazon S3 or Amazon Elastic File System (EFS). It handles many of the transfer tasks that can slow down or burden IT operations, like encryption, scripts, network optimization, and data integrity validation.
- AWS Direct Connect is a cloud service solution that helps establish a dedicated network connection from on-premises resources to AWS, which can reduce network costs, increase bandwidth, and provide a more consistent network experience.

Case Study: DigitalGlobe

DigitalGlobe, now part of Maxar, is one of the world's leading providers of high-resolution earth imagery, data and analysis. In order to meet the growing demand for commercial geo-intelligence, they went all-in with AWS, migrating their entire 17-year imagery archive to the cloud. By using AWS Snowmobile, they moved 100 petabytes of data without requiring large file-transfer protocols and delivery workflows.

DigitalGlobe also uses Amazon SageMaker to handle machine learning at scale. By training algorithms to find relevance in image files hundreds of gigabytes large, they were able to call up required images from AWS Glacier to allow almost instantaneous access while keeping the remainder of their data in cost-effective long-term storage.

"DigitalGlobe was the first customer for AWS Snowmobile—a big data center on wheels—and we moved 17 years worth of data in a single, very cost-effective operation. It was a huge step up from our legacy tape plus disk."

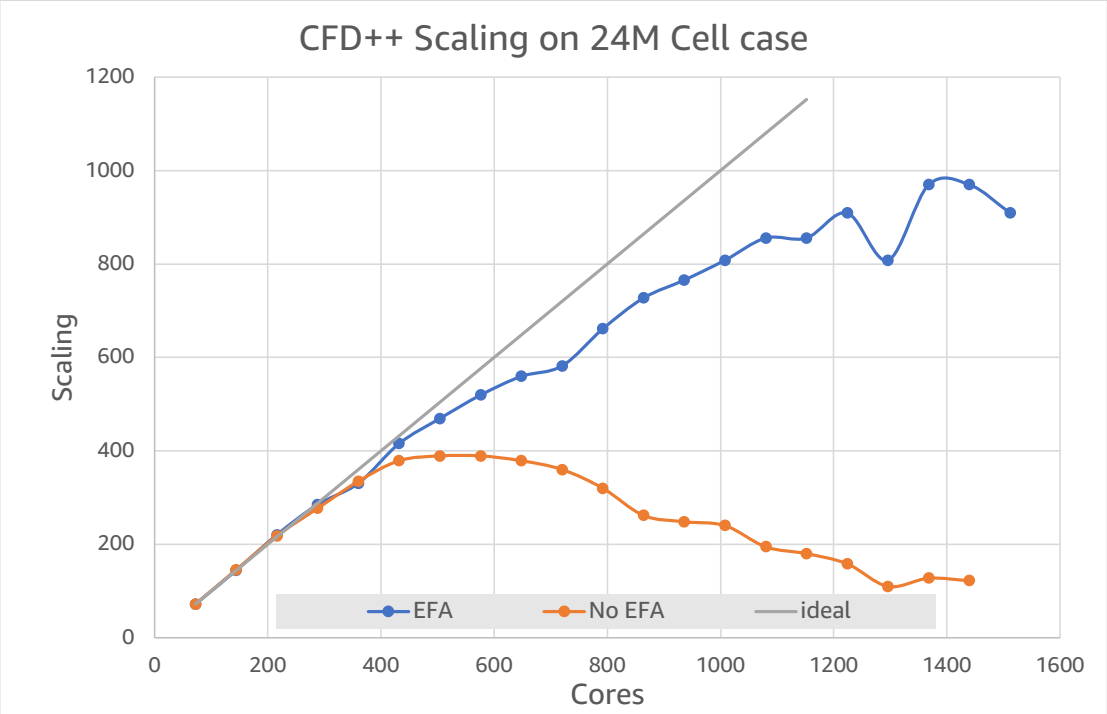
– Dr. Walter Scott, CTO and Founder, DigitalGlobe

PERCEIVED BARRIER 4: PERFORMANCE

Organizations using High Performance Computing expect high performance—and many don't believe the cloud can compete with on-premises data centers. 35% of participants in the market survey mentioned concerns about network performance and inter-connect latencies, and 29% mentioned broader performance concerns.

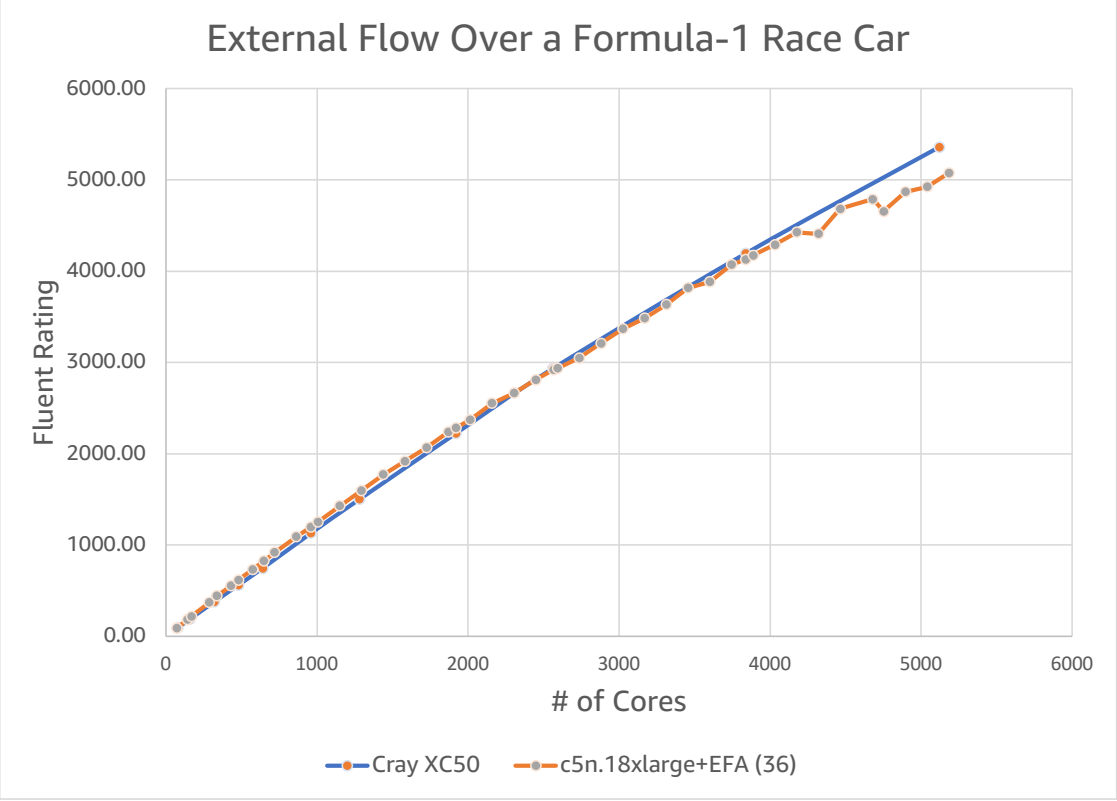
But the belief that the networking speed between compute nodes in the cloud is not fast enough for high performance is outdated. Recent advancements have helped speed up cloud networking and trim latency to the point where all but the most resource-intensive HPC applications run just as well or better on the cloud than on on-premises infrastructure.

AWS performance exceeds the needs of almost every HPC use-case in terms of scalability, elasticity and raw performance, and typically delivers better ROI. EFA's unique OS bypass networking mechanism provides a low-latency, low-jitter channel for inter-instance communications. This enables tightly coupled HPC or distributed machine learning applications to scale to thousands of cores, so applications run faster—a 4x improvement in scaling over ENA for a standard CFD simulation as shown below.



In another performance benchmark exercise, Amazon EC2 C5n instances were compared to a Cray XC50 when running a standard CFD use case. Engineering simulation software provider ANSYS publishes ANSYS® Fluent® benchmarks of "External Flow Over a Formula-1 Race Car." This case has around 140-million Hex-core cells and uses the realizable k-epsilon turbulence model as well as the Pressure-based coupled solver and the Least Squares cell-based, pseudo-transient solver. Running the same benchmark using Amazon EC2 C5n instances and Elastic Fabric Adapter is a simple way to benchmark the performance of the solver on AWS and compare it against traditional HPC infrastructure.

The plot below shows the rating of a Cray XC50* and C5n.18xlarge with EFA. ANSYS defines† this rating as "the primary metric used to report performance results of the Fluent Benchmarks. It is defined as the number of benchmarks that can be run on a given machine (in sequence) in a 24-hour period. It is computed by dividing the number of seconds in a day (86,400 seconds) by the number of seconds required to run the benchmark. A higher rating means better performance."



The plot shows C5n.18xlarge with EFA got a higher rating up to 2,400 cores and is essentially on par up to about 3,800 cores.

* https://www.ansys.com/de-de/solutions/solutions-by-role/it-professionals/platform-support/benchmarks-overview/benchmarking-terminology

† https://www.ansys.com/solutions/solutions-by-role/it-professionals/platform-support/benchmarks-overview/ansys-fluent-benchmarks/ansys-fluent-benchmarks-release-19/external-flow-over-a-formula-1-race-car)

AWS PERFORMANCE SOLUTIONS

HPC on AWS delivers high performance for almost every class of HPC application. There are several solutions that make it easy and cost-effective to monitor and improve performance.

- Amazon EC2 instances, powered by Intel® Xeon® Scalable processors, deliver the most elastic and scalable cloud infrastructure to run HPC applications. With integrated Intel® AVX-512, EC2 instances can offer accelerated application performance with 2x more FLOPS than previous generation technologies. Intel® Deep Learning Boost brings new embedded performance acceleration for Artificial Intelligence workloads in the 2nd Generation Intel® Xeon® Scalable Processors available in EC2 C5 instances. C5 instances deliver up to 2x performance improvement for inference workloads compared to previous generation EC2 C4 instances.
- Elastic Fabric Adapter is a network interface for Amazon EC2 instances that enables customers to run tightly coupled applications requiring high levels of inter-node communications at scale. As a result, they gain the high application scalability and performance of on-premises HPC clusters with the on-demand elasticity and flexibility of the AWS cloud.
- Integrating Intel® message-passing interface (Intel MPI) with Elastic Fabric Adapter delivers end-to-end high performance and scale to distributed, parallel computing clusters.
- AWS Auto Scaling monitors applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost. This service makes it easy to set up application scaling for multiple resources across multiple services, from thousands to millions of cores in minutes.
- Amazon CloudWatch is a monitoring and observability service that offers data and actionable insights to monitor applications, respond to system-wide performance changes, optimize resource utilization, and get a unified view of operational health.

CONCLUSION

HPC is an essential function for many industries, but misconceptions about cloud-based HPC may prevent organizations from realizing the benefits of these powerful systems—like quicker time to market, new business insights, unprecedented agility and scalability, and more. When comparing on-premises infrastructure to cloud-based HPC, it's important to consider factors beyond a simple cost-per-core-hour analysis and look at the holistic business impact. Factors like personnel productivity, cutting-edge technology, and innovation acceleration are critical in the new digital economy—the difference between leading the industry or playing catch-up.

The ability to create bespoke compute clusters in the AWS cloud enables cost-effective HPC for most business cases, from small research teams to large enterprise organizations. And AWS offers a suite of integrated products and services that keep data private and secure, make it easy to migrate and transfer data, and deliver consistent high performance.

Get started with HPC on AWS at
http://aws.amazon.com/hpc