High Performance Computing on AWS Redefines What is Possible

High Performance Computing (HPC) has been key to solving the most complex problems in every industry, and changing the way we work and live. From weather modeling to genome mapping to the search for extraterrestrial intelligence, HPC is helping to push the boundaries of what's possible with advanced computing technologies. Once confined to government labs, large enterprises, and select academic organizations, today it is found across a wide range of industries.

In this paper we will discuss how cloud services put the world's most advanced computing capabilities within reach for more organizations, helping them to innovate faster and gain a competitive edge. We will discuss the advantages of running HPC workloads on Amazon Web Services (AWS), with Intel® Xeon® technology, compared to traditional on-premises architectures. We will also illustrate these benefits in actual deployments across a variety of industries.

In 2017, the market for cloud HPC solutions **grew by 44%** compared to 2016.[i]

HPC FUNDAMENTALS

Although HPC applications share some common building blocks, they are not all similar. HPC applications are often based on complex algorithms that rely on high performing infrastructure for efficient execution. These applications need hardware that includes high performance processors, memory, and communication subsystems. For many applications and workloads, the performance of compute elements must be complemented by comparably high performance storage and networking elements. Some may demand high levels of parallel processing, but not necessarily fast storage or high performance interconnect. Other applications are interconnect-sensitive, requiring low latency and high throughput networking. Similarly, there are many I/O-sensitive applications that, without a very fast I/O subsystem, will run slowly because of storage bottlenecks. And still other applications such as game streaming, video encoding, and 3D application streaming, need performance acceleration using GPUs.

Today, many large enterprises and research institutions procure and maintain their own HPC infrastructure. This HPC infrastructure is shared across many applications and groups within the organization to maximize utilization of this significant capital investment.

Cloud-based services have opened up a new frontier for HPC. Moving HPC workloads to the cloud can provide near instant access to virtually unlimited computing resources for a wider community of users, and can support completely new types of applications. Today, organizations of all sizes are looking to the cloud to support their most advanced computing applications. For smaller enterprises, cloud is a great starting point, enabling fast, agile deployment without the need for heavy capital expenditure. For large enterprises, cloud provides an easier way to tailor HPC infrastructure to changing business needs and to gain access to the latest technologies, without having to worry about upfront investments in new infrastructure, or ongoing operational expenses. When compared to traditional on-premises HPC infrastructures, cloud offers significant advantages in terms of scalability, flexibility, and cost.

ON-PREMISES HPC HAS ITS LIMITS

Today, on-premises HPC infrastructure handles most of the HPC workloads that enterprises and research institutions employ. Most HPC system administrators maintain and operate this infrastructure at varying levels of utilization. However, business is always competitive, so efficiency needs to be coupled with the flexibility and opportunity to innovate continuously.

Some of the challenges with on-premises HPC are well known. These include long procurement cycles, high initial capital investment, and the need for mid-cycle technology refreshes. For most organizations, planning for and procuring an HPC system is a long and arduous process that involves detailed capacity forecasting and system evaluation cycles. Often, the significant up-front capital investment required is a limiting factor for the amount of capacity that can be procured. Maintaining the infrastructure over its lifecycle is an expensive proposition, as well. Previously, technology refreshes every three years was enough to stay current with the compute technology and incremental demands from HPC workloads. However, to take advantage of the faster pace of innovation, HPC customers are needing to refresh their infrastructure more often than before. And it is worth the effort. IDC reports that for every $1 spent on HPC, businesses see $463 in incremental revenues and $44 in incremental profit, so delaying incremental investments in HPC – and thus delaying the innovations it brings – has large downstream effects on the business.

There are other limitations of on-premises HPC infrastructure that are less visible and so are often overlooked leading to misplaced optimization efforts.

Stifled Innovation:

Often the constraints of on-premises infrastructure mean that use cases or applications that did not meet the capabilities of the hardware were not considered. When engineers and researchers are forced to limit their imagination to what can be tried out with limited access to infrastructure, the opportunity to think outside the box and tinker with new ideas gets lost.

Reduced Productivity:

On-premises systems often have long queues and wait times that decrease productivity. They are managed to maximize utilization – often resulting in very intricate scheduling policies for jobs. However, even if a job requires only a couple of hours to run, it may be stuck in a prioritized queue for weeks or months – decreasing overall productivity and limiting innovation. In contrast, with virtually unlimited capacity, the cloud can free users to get the same job done, but much faster, without having to stand in line behind others who are just as eager to make progress.

Limited Scalability and Flexibility:

HPC workloads and their demands are constantly changing, and legacy HPC architectures cannot always keep pace with evolving requirements. For example, infrastructure elements like GPUs, containers, and serverless technologies are not readily available in an on-premises environment. Integrating new OS or container capabilities – or even upgrading libraries and applications – is a major system-wide undertaking. And when an on-premises HPC system is designed for a specific application or workload, it's difficult and expensive to take on new HPC applications, as well as forecast and scale for future (frequently unknown) requirements.

Lost Opportunities:

On-premises HPC can sometimes limit an organization's opportunities to take full advantage of the latest technologies. For example, as organizations adopt leading-edge technologies like artificial intelligence/machine learning technologies (AI/ML) and visualization, the complexity and volume of data is pushing on-premises infrastructure to its limits. Furthermore, most AI/ML algorithms are cloud-native. These algorithms will deliver superior performance on large data sets when running in the cloud, especially with workloads that involve transient data that does not need to be stored long term.

CLOUD IS A BETTER WAY TO HPC

To move beyond the limits of on-premises HPC, many organizations are leveraging cloud services to support their most advanced computing applications. Flexible and agile, the cloud offers strong advantages compared to traditional on-premises HPC approaches.

HPC on AWS, with Intel® Xeon® processors, deliver significant leaps in compute performance, memory capacity, and bandwidth and I/O scalability. The highly customizable computing platform and robust partner community enable your staff to imagine new approaches so they can fail forward faster, delivering more answers to more questions without the need for costly, on-premises upgrades. In short, AWS frees you to rethink your approach to every HPC and big data analysis initiative and invites your team to ask questions and seek answers as often as possible.

Innovate Faster with a Highly Scalable Infrastructure

Moving HPC workloads to the cloud can bring down barriers to innovation by opening up access to virtually unlimited capacity and scale. And one of the best features of working in a cloud environment is that when you solve a problem, it stays solved. You're not revisiting it every time you do a major system-wide software upgrade or a bi-annual hardware refresh.

Limits on scale and capacity with on-premises infrastructure, usually led to organizations being reluctant to consider new use cases or applications that exceeded their capabilities. Running HPC in the cloud enables asking the business critical questions they couldn't address before, and that means a fresh look at project ideas that were shelved due to infrastructure constraints.

Migrating HPC applications to AWS eliminates the need for tradeoffs between experimentation and production. AWS and Intel bring the most cost-effective, scalable solutions to run the most computationally-intensive applications on-demand. Now research, development, and analytics teams can test every theory and process every data set without straining on-premises systems or stalling other critical work streams. Flexible configuration and virtually unlimited scalability allow engineers to grow and shrink the infrastructure as workloads dictate, not the other way around. Additionally, with easy access to a broad range of cloud-based services and a trusted partner network, researchers and engineers can quickly adopt tested and verified HPC applications so that they can innovate faster without having to reinvent what already exists.

Increase Collaboration with Secure Access to Clusters Worldwide

Running HPC workloads on the cloud enables a new way for globally distributed teams to collaborate securely. With globally-accessible shared data, engineers and researchers can work together or in parallel to get results faster. For example, the use of the cloud for collaboration and visualization allows a remote design team to view and interact with a simulation model in near real time, without the need to duplicate and proliferate sensitive design data. Using the cloud as a collaboration platform also makes it easier to ensure compliance with ever-changing industry regulations.

The AWS cloud is compliant with the latest revisions of GDPR, HIPAA, FISMA, FedRAMP, PCI, ISO 27001, SOC 1, and other regulations. Encryption and granular permission features guard sensitive data without interfering with the ability to share data across approved users, and detailed audit trails for virtually every API call or cloud orchestration action means environments can be designed to address specific governance needs and submit to continuous monitoring and surveillance. With a broad global presence and the wide availability of Intel® Xeon® technology-powered Amazon EC2 instances, HPC on AWS enables engineers and researchers to share and collaborate efficiently with team members across the globe without compromising on security.

Optimize Cost with Flexible Resource Selection

Running HPC in the cloud enables organizations to select and deploy an optimal set of services for their unique applications, and to pay only for what they use. Individuals and teams can rapidly scale up or scale down resources as needed, commissioning or decommissioning HPC clusters in minutes, instead of days or weeks. With HPC in the cloud, scientists, researchers, and commercial HPC users can gain rapid access to resources they need without a burdensome procurement process.

Running HPC in the cloud also minimizes the need for job queues. Traditional HPC systems require researchers and analysts to submit their projects to open source or commercial cluster and job management tools, which can be time consuming and vulnerable to submission errors. Moving HPC workloads to the cloud can help increase productivity by matching the infrastructure configuration to the job. With on-premises infrastructure, engineers were constrained to running their job on the available configuration. With HPC in the cloud, every job (or set of related jobs) can run on its own on-demand cluster, customized for its specific requirements. The result is more efficient HPC spending, and fewer wasted resources.

AWS HPC solutions remove the traditional challenges associated with on-premises clusters: fixed infrastructure capacity, technology obsolescence, and high capital expenditures. AWS gives you access to virtually unlimited HPC capacity, built from the latest technologies. You can quickly migrate to newer, more powerful Intel® Xeon® processor-based EC2 instances as soon as they are made available on AWS. This removes the risk of on-premises CPU clusters becoming obsolete or poorly utilized as your needs change over time. As a result, your teams can trust that their workloads are running optimally at every stage.

AWS AND INTEL® DELIVER A COMPLETE HPC SOLUTION

AWS HPC solutions with Intel® Xeon® technology-powered compute instances put the full power of HPC in reach for organizations of every size and industry. AWS provides a comprehensive set of components required to power today's most advanced HPC applications, giving you the ability to choose the most appropriate mix of resources for your specific workload. Key products and services that make up the HPC on AWS solution include:

Data Management & Data Transfer

Running HPC applications in the cloud starts with moving the required data into the cloud. AWS Snowball and AWS Snowmobile are data transport solutions that use devices designed to be secure to transfer large amounts of data into and out of the AWS Cloud. Using Snowball addresses common challenges with large-scale data transfers including high network costs, long transfer times, and security concerns. AWS DataSync is a data transfer service that makes it easy for you to automate moving data between on-premises storage and Amazon S3 or Amazon Elastic File System (Amazon EFS). DataSync automatically handles many of the tasks related to data transfers that can slow down migrations or burden your IT operations, including running your own instances, handling encryption, managing scripts, network optimization, and data integrity validation. AWS Direct Connect is a cloud service solution that makes it easy to establish a dedicated network connection from your premises to AWS. Using AWS Direct Connect, you can establish private connectivity between AWS and your datacenter, office, or colocation environment, which in many cases can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

Compute

The AWS HPC solution lets you choose from a variety of compute instance types that can be configured to suit your needs, including the latest Intel® Xeon® processor-powered CPU instances, GPU-based instances, and field programmable gate array (FPGA)-powered instances. The latest Intel-powered Amazon EC2 instances include the C5n, C5d and Z1d instances. C5n instances feature the Intel Xeon Platinum 8000 series (Skylake-SP) processor with a sustained all core Turbo CPU clock speed of up to 3.5 GHz. C5n instances provide up to 100 Gbps of network bandwidth and up to 14 Gbps of dedicated bandwidth to Amazon EBS. C5n instances also feature 33% higher memory footprint compared to C5 instances. For workloads that require access to high-speed, ultra-low latency local storage, AWS offers C5d instances equipped with local NVMe-based SSDs. Amazon EC2 z1d instances offer both high compute capacity and a high memory footprint. High frequency z1d instances deliver a sustained all core frequency of up to 4.0 GHz, the fastest of any cloud instance. For HPC codes that can benefit from GPU acceleration, the Amazon EC2 P3dn instances feature 100 Gbps network bandwidth (up to 4x the bandwidth of previous P3 instances), local NVMe storage, the latest NVIDIA V100 Tensor Core GPUs with 32 GB of GPU memory, NVIDIA NVLink for faster GPU-to-GPU communication, AWS-custom Intel® Xeon® Scalable (Skylake) processors running at 3.1 GHz sustained all-core Turbo. AWS Auto Scaling monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost. Using AWS Auto Scaling, it's easy to setup application scaling for multiple resources across multiple services in minutes.

Networking

Amazon EC2 instances support enhanced networking that allow EC2 instances to achieve higher bandwidth and lower inter-instance latency compared to traditional virtualization methods. Elastic Fabric Adapter (EFA) is a network interface for Amazon EC2 instances that enables you to run HPC applications requiring high levels of inter-node communications at scale on AWS. Its custom-built operating system (OS) bypass hardware interface enhances the performance of inter-instance communications, which is critical to scaling HPC applications. AWS also offers placement groups for tightly-coupled HPC applications that require low latency networking. Amazon Virtual Private Cloud (VPC) provides IP connectivity between compute instances and storage components.

Storage

Storage options and storage costs are critical factors when considering an HPC solution. AWS offers flexible object, block, or file storage for your transient and permanent storage requirements. Amazon Elastic Block Store (Amazon EBS) provides persistent block storage volumes for use with Amazon EC2. Provisioned IOPS allows you to allocate storage volumes of the size you need and to attach these virtual volumes to your EC2 instances. Amazon Simple Storage Service (S3) is designed to store and access any type of data over the Internet and can be used to store the HPC input and output data long term and without ever having to do a data migration project again. Amazon FSx for Lustre is a high performance file storage service designed for demanding HPC workloads and can be used on Amazon EC2 in the AWS cloud. Amazon FSx for Lustre works natively with Amazon S3, making it easy for you to process cloud data sets with high performance file systems. When linked to an S3 bucket, an FSx for Lustre file system transparently presents S3 objects as files and allows you to write results back to S3. You can also use FSx for Lustre as a standalone high-performance file system to burst your workloads from on-premises to the cloud. By copying on-premises data to an FSx for Lustre file system, you can make that data available for fast processing by compute instances running on AWS. Amazon Elastic File System (Amazon EFS) provides simple, scalable file storage for use with Amazon EC2 instances in the AWS Cloud.

Automation and Orchestration

Automating the job submission process and scheduling submitted jobs according to predetermined policies and priorities are essential for efficient use of the underlying HPC infrastructure. AWS Batch lets you run hundreds to thousands of batch computing jobs by dynamically provisioning the right type and quantity of compute resources based on the job requirements. AWS ParallelCluster is a fully supported and maintained open source cluster management tool that makes it easy for scientists, researchers, and IT administrators to deploy and manage High Performance Computing (HPC) clusters in the AWS Cloud. NICE EnginFrame is a web portal designed to provide efficient access to HPC-enabled infrastructure using a standard browser. EnginFrame provides you a user-friendly HPC job submission, job control, and job monitoring environment.

Operations & Management

Monitoring the infrastructure and avoiding cost overruns are two of the most important capabilities that can help an HPC system administrators efficiently manage your organization's HPC needs. Amazon CloudWatch is a monitoring and management service built for developers, system operators, site reliability engineers (SRE), and IT managers. CloudWatch provides you with data and actionable insights to monitor your applications, understand and respond to system-wide performance changes, optimize resource utilization, and get a unified view of operational health. AWS Budgets gives you the ability to set custom budgets that alert you when your costs or usage exceed (or are forecasted to exceed) your budgeted amount.

Visualization Tools

The ability to visualize results of engineering simulations without having to move massive amounts of data to/from the cloud is an important aspect of the HPC stack. Remote visualization helps accelerate the turnaround times for engineering design significantly. NICE Desktop Cloud Visualization enables you to remotely access 2D/3D interactive applications over a standard network. In addition, Amazon AppStream 2.0 is another fully managed application streaming service that can securely deliver application sessions to a browser on any computer or workstation.

Security and Compliance

Security management and regulatory compliance are other important aspects of running HPC in the cloud. AWS offers multiple security related services and quick-launch templates to simplify the process of creating a HPC cluster and implementing best practices in data security and regulatory compliance. The AWS infrastructure puts strong safeguards in place to help protect customer privacy. All data is stored in highly secure AWS data centers. 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 manages dozens of compliance programs in its infrastructure. This means that segments of your compliance have already been completed. AWS infrastructure is compliant with many relevant industry regulations such as HIPAA, FISMA, FedRAMP, PCI, ISO 27001, SOC 1, and others.

Flexible Pricing and Business Models

With AWS, capacity planning worries become a thing of the past. AWS offers on-demand pricing for short-term projects, contract pricing for long-term, predictable needs, and spot pricing for experimental work or research groups with tight budgets. AWS customers enjoy the flexibility to choose from any combination of pay-as-you-go options, procuring only the capacity they need, for the duration that it's needed, and AWS Trusted Advisor will alert you first to any cost-saving actions you can take to minimize your bill. This simplified, flexible pricing structure and approach allows research institutions to break free from the time- and budget-constraining, CapEx-intensive data center model.

With HPC on AWS, organizations can flexibly tune and scale their infrastructure, as workloads dictate, instead of the other way around.

AWS Partners and Marketplace

For organizations looking to build highly specific solutions, AWS Marketplace is an online store for applications and services that build on top of AWS. AWS partner solutions and AWS Marketplace lets organizations immediately take advantage of partners' built-in optimizations and best practices, leveraging what they've learned from building complex services on AWS. A variety of open source HPC applications are also available on the AWS Marketplace.

HPC ON AWS DELIVERS ADVANTAGES FOR A RANGE OF HPC WORKLOADS

AWS cloud provides a broad range of scalable, flexible infrastructure solutions that organizations can select to match their workloads and tasks. This gives HPC users the ability to choose the most appropriate mix of resources for their specific applications. Let us take a brief look at the advantages that HPC on AWS delivers for these workload types.

Tightly Coupled HPC: A typical tightly coupled HPC application often spans across large numbers of CPU cores in order to accomplish demanding computational workloads. To study the aerodynamics of a new commercial jet liner, design engineers often run computational fluid dynamics simulations using thousands of CPU cores. Global climate modeling applications are also executed at a similar scale. AWS cloud provides scalable computing resources to execute such applications. These applications can be deployed on the cloud at any scale. Organizations can set a maximum number of cores per job, dependent on the application requirements, aligning it to criteria like model size, frequency of jobs, cost per computation, and urgency of the job completion. A significant benefit of running such workloads on AWS is the ability to scale out to experiment with more tunable parameters. For example, an engineer performing electromagnetic simulations can run larger numbers of parametric sweeps in his Design of Experiment (DoE) study using very large numbers of Amazon EC2 On-Demand instances, and using AWS Auto Scaling to launch independent and parallel simulation jobs. Such DoE jobs would often not be possible because of the hardware limits of on-premises infrastructure. A further benefit for such an engineer is to use Amazon Simple Storage Service (S3), NICE DCV, and other AWS solutions like AI/ML services to aggregate, analyze, and visualize the results as part of a workflow pipeline, any element of which can be spun up (or down) independently to meet needs. Amazon EC2 features that help with applications in this category also include EC2 placement groups and enhanced networking, for reduced node-to-node latencies and consistent network performance.

Loosely Coupled Grid Computing: The cloud provides support for a variety of loosely coupled grid computing applications that are designed for fault-tolerance, enabling individual nodes to be added or removed during the course of job execution. This category of applications includes Monte Carlo simulations for financial risk analysis, material science study for proteomics, and more. A typical job distributes independent computational workloads across large numbers of CPU cores or nodes in a grid without high demand for high performance node-to-node interconnect, or on high-performance storage. The cloud lets organizations deliver the fault-tolerance

these applications require, and choose the instance types they require for specific compute tasks that they plan to execute. Such applications are ideally suited to Amazon EC2 Spot instances, which are EC2 instances that opportunistically take advantage of Amazon EC2's spare computing capacity. Coupled with Amazon EC2 Auto Scaling, and jobs can be scaled up when excess spare capacity makes Spot instances cheaper than normal. AWS Batch brings all these capabilities together in a single batch-oriented service that is easy to use, container-focused for maximum portability and integrates with a range of commercial and open source workflow engines to make job orchestration easy.

High Volume Data Analytics and Interpretation: When grid and cluster HPC workloads handle large amounts of data, their applications require fast, reliable access to many types of data storage. AWS services and features that help HPC users optimize for data-intensive computing include Amazon S3, Amazon Elastic Block Store (EBS), and Amazon EC2 instance types that are optimized for high I/O performance (including those configured with solid-state drive (SSD) storage). Solutions also exist for creating high performance virtual network attached storage (NAS) and network file systems (NFS) in the cloud, allowing applications running in Amazon EC2 to access high performance, scalable, cloud-based shared storage resources. Example applications in this category include genomics, high-resolution image processing, and seismic data processing.

Visualization: Using the cloud for collaboration and visualization makes it much easier for members in global organizations to share their digital data instantly from any part of the world For example, it lets subcontractors or remote design teams view and interact with a simulation model in near real time, from any location. They can securely collaborate on data from anywhere, without the need to duplicate and share it. AWS services that enable these types of workloads include graphics-optimized instances, remote visualization services like NICE DCV, and managed services like Amazon Workspaces and Amazon AppStream 2.0.

Accelerated Computing: There are many HPC workloads that can benefit from offloading computation-intensive tasks to specialized hardware coprocessors such as GPUs or FPGAs. Many tightly-coupled and visualization workloads are apt for accelerated computing. AWS HPC solutions offer the flexibility to choose from many available CPU, GPU or FPGA-based instances to deploy optimized infrastructure to meet the needs of specific applications.

Machine Learning and Artificial Intelligence: Machine learning requires a broad set of computing resource options, ranging from GPUs for compute-intensive deep learning, FPGAs for specialized hardware acceleration, to high-memory instances for inference study. With HPC on AWS, organizations can select instance types and services to fit their machine learning needs. They can choose from a variety of CPU, GPU, FPGA, memory, storage, and networking options, and tailor instances to their specific requirements, whether they are training models or running inference on trained models.

AWS uses the latest Intel® Xeon®Scalable CPUs which are optimized for machine learning and AI workloads at scale. The Intel® Xeon®Scalable processors incorporated in AWS EC2 C5 instances along with optimized deep learning functions in the Intel MKL-DNN library provide sufficient compute for deep learning training workloads (in addition to inference, classical machine learning, and other AI algorithms). In addition, CPU and GPU optimized frameworks such as TensorFlow, MxNet, and PyTorch are available in Amazon Machine Image (AMI) format for customers to deploy their AI workloads on optimized software and hardware stacks. Recent advances in distributed algorithms have also enabled the use of hundreds of servers to reduce the time to train from weeks to minutes. Data scientists can get excellent deep learning training performance using Amazon EC2, and further reduce the time-to-train by using multiple CPU nodes scaling near linearly to hundreds of nodes.

DRIVING INNOVATION ACROSS INDUSTRIES

Every industry tackles a different set of challenges. AWS HPC solutions, available with the power of the latest Intel technologies, help companies of all sizes in nearly every industry achieve their HPC results with flexible configuration options that simplify operations, save money, and get results to market faster. These workloads span the traditional HPC applications like genomics, life sciences research, financial risk analysis, computer-aided design, and seismic imaging, to the emerging applications like machine learning, deep learning, and autonomous vehicles.

Life Sciences and Healthcare

Running HPC workloads on AWS lets healthcare and life sciences professionals easily and securely scale genomic analysis and precision medicine applications. For AWS users, the scalability is built-in, bolstered by an ecosystem of partners for tools and datasets designed for sensitive data and workloads. They can efficiently, dynamically store and compute their data, collaborate with peers, and integrate findings into clinical practice—while conforming with security and compliance requirements.

For example, Bristol-Myers Squibb (BMS), a global biopharmaceutical company, used AWS to build a secure, self-provisioning portal for hosting research. The solution lets scientists run clinical trial simulations on-demand and enables BMS to set up rules that keep compute costs low. Compute-intensive clinical trial simulations that previously took 60 hours are finished in only 1.2 hours on the AWS Cloud. Running simulations 98% faster has led to more efficient, less costly clinical trials—and better conditions for patients.

> "The time and money savings are obvious, but probably what is most important factor is we are using fewer subjects in these trials, we are optimizing dosage levels, we have higher drug tolerance and safety, and at the end of the day, for these kids, it's fewer blood samples."
>
> Sr. Solutions Specialist, Bristol-Myers Squibb

Financial Services

Insurers and capital markets have long been utilizing grid computing to power actuarial calculations, determine capital requirements, model risk scenarios, price products, and handle other key tasks. Taking these compute-intensive workloads out of the data center and moving them to AWS helps them boost speed, scale better, and save money.

For example, MAPRE, the largest insurance company in Spain, needed fast, flexible environments in which to develop sales management insurance policy applications. The firm was looking for a cost-effective technology platform that could deliver rapid analysis and enable quick deployment of development environments in remote installations sites. Its on-premises infrastructure simply could not support these needs. The company turned to AWS for high-performance computing, risk analysis of customer data, and to create test and development environments for its commercial application.

> "The on-premises hardware investment for three years cost approximately €1.5 million, whereas the AWS infrastructure cost the company €180,000 for the same period, a savings of 88 percent."
>
> MAPFRE

KEEPING PACE WITH CHANGING FINANCIAL REGULATIONS

AWS customers in financial services are preparing for new Fundamental Review of Trading Book (FRTB) regulations that will come into effect between 2019 and 2021. As part of the proposed regulations, these financial services institutions will need to perform compute-intensive "value at risk" calculations in the four hours after trading ends in New York and begins in Tokyo.

The periodic nature of the calculation, along with the amount of processing power and storage needed to run it within four hours, made it a great fit for an environment where a vast amount of cost-effective compute power is available on an on-demand basis.

To help its financial services customers meet these new regulations, AWS worked with TIBCO (an on-premises market-leading infrastructure platform for grid and elastic computing) to run a proof of concept grid in AWS Cloud. The grid grew to 61,299 Spot instances, with 1.3 million vCPUs, and cost approximately $30,000 an hour to run. This proof-of-concept is a strong example of the potential for AWS to deliver a vast amount of cost-effective compute power on an on-demand basis.

Design and Engineering

Using simulations on AWS HPC infrastructure lets manufacturers and designers reduce costs by replacing expensive development of physical models with virtual ones during product development. The result? Improved product quality, shorter time to market, and reduced product development costs.

TLG Aerospace in Seattle, Washington put these capabilities to work to perform aerodynamic simulations on aircraft and predict the pressure and temperature surrounding airframes. Its existing cloud provider was expensive and could not scale to handle more performance-intensive applications. TLG turned to Amazon EC2 Spot instances, which provide a way to use unused EC2 computing capacity at a discounted price. The solution dramatically decreased simulation costs and can scale easily to take on new jobs as needed.

> "We saw a 75% reduction in the cost per CFD simulation as soon as we started using Amazon EC2 Spot instances. We are able to pass those savings along to our customers–and be more competitive."
>
> TLG Aerospace

Energy and Geo Sciences

Reducing run-times for compute-intensive applications like seismic analysis and reservoir simulation is just one of the many ways the energy and geosciences industry has been utilizing HPC applications in the cloud. By moving HPC applications to the cloud, organizations reduce job submission time, track runtime, and efficiently manage the large data-sets associated with daily workloads.

For example, using AWS on-demand computing resources, Zenotech, a simulation service provider, can power simulations that help energy companies support advanced reservoir models.

> Using the resources available within a typical small company, it would take several years to complete a sophisticated reservoir simulation. Zenotech completed it at a computing cost for AWS resources of only $750 over a 12-day period.

Media and Entertainment

The movie and entertainment industries are shifting content production and post-production to cloud-based HPC to take advantage of highly scalable, elastic and secure cloud services, to accelerate content production and reduce capital infrastructure investment. Content production and post-production companies are leveraging the cloud to accelerate and streamline production, editing, and rendering workloads with highly scalable cloud computing and storage.

One design and visual effects (VFX) company, Fin Design + Effects, needed the ability to access vast amounts of compute capacity when big deadlines came around. Its on-premises render servers had a finite capacity and were difficult and expensive to scale. Fin started by using AWS Direct Connect to scale its rendering capabilities by establishing a dedicated Gigabit network connection from the Fin data center to AWS. Fin is also taking advantage of Amazon EC2 Spot instances. Fin now has the agility to add compute resources on the fly to meet last-minute project demands.

"We are reducing our operational costs by 50 percent by using Amazon EC2 Spot instances."

Fin Design

AI/ML and Autonomous Vehicles

The AI revolution which started with the rapid increase in accuracy brought by deep learning methods, has the potential to revolutionize a variety of industries. Autonomous driving is a particularly popular use case for AI/ML. Developing and deploying autonomous vehicles requires the ability to collect, store and manage massive amounts of data, high performance computing capacity and advanced deep learning frameworks, along with the capability to do real-time processing of local rules and events in the vehicle.

AWS's virtually unlimited storage and compute capacity and support for popular deep learning frameworks help accelerate algorithm training and testing and drive faster time to market.

Developing and deploying autonomous vehicles requires the ability to collect, store and manage massive amounts of data, high performance computing capacity and advanced deep learning frameworks.

SUMMARY AND RECOMMENDATION

Technology continues to change rapidly, and it's clear that HPC has a critical role to play in enabling organizations to innovate faster, and enable them to adopt other leading-edge technologies like AI/ML and IoT. AWS puts the advanced capabilities of High Performance Computing in reach for more people and organizations, while simplifying processes like management, deployment, and scaling. Accessible, flexible, and cost effective, it lets organizations unleash the creativity of their engineers, analysts, and researchers from the limitations of on-premises infrastructures.

Unlike traditional on-premise HPC systems, AWS offers virtually unlimited capacity to scale out HPC infrastructure. It also provides the flexibility for organizations to adapt their HPC infrastructure to changing business priorities. With flexible deployment and pricing models, it lets organizations of all sizes and industries take advantage of the most advanced computing capabilities available. HPC on AWS lets you take a fresh approach to innovation to solve the world's most complex problems.

Learn more about running your HPC workloads on AWS at http://aws.amazon.com/hpc

i "HPC Market Update, ISC18," Intersect360 Research, 2018