Democratizing High Performance Computing

Reducing barriers to entry with cloud-based solutions

Executive summary

High Performance Computing (HPC) has reshaped some industries and enabled others—but it has been difficult for smaller organizations to access because of large up-front costs and planning difficulties. But cloud-based solutions are helping to break down barriers to entry for even the smallest teams, and many different industries are finding it much easier to get started in HPC as a result. Between a broader range of flexible cloud solutions and an ever-growing set of third-party solutions to assist with onboarding, planning, and management, HPC resources are now within reach for almost everyone who needs them.

INTRODUCTION

High Performance Computing brings a powerful set of tools to a broad range of industries, helping to drive innovation and boost revenue in finance, genomics, oil and gas extraction, and other fields. For many smaller organizations, on-premises HPC infrastructure is too expensive to procure and maintain. They have been forced to make do with renting time on others' supercomputers, outsourcing design and engineering tasks, or running their applications on whatever computing hardware they can afford. Even within larger organizations that can afford to host their own HPC infrastructure, engineers and researchers must compete for scarce computing resources. But cloud-based HPC solutions are putting vast computational capabilities within reach of more and more organizations—and offering greater flexibility as well.

Using cloud-based HPC lets organizations get started quickly and start to realize benefits almost immediately. Many see faster innovation thanks to shorter turnaround times and improved flexibility, and collaboration is greatly enhanced between teams that might not be able to work together otherwise due to geographical or other logistical considerations. Cost optimization is also a key factor when considering cloud-based HPC—it is much simpler to predict and manage budget and resource use in the cloud.

Many startups and independent researchers who had not even considered buying and setting up their own HPC infrastructure because of perceived up-front costs are finding that it's now easier than ever—and much less expensive—to dive into cloud-based HPC. The ability to configure massive parallel computing clusters on demand in the cloud changes the rules—any team with a need for compute resources to solve a problem can start working on it in hours or days. As more organizations adopt cloud-based HPC, more applications, ISVs, and systems integrators are creating better and better solutions for a wider range of users.

BARRIERS TO ENTRY ARE ERODING QUICKLY

Many organizations, especially smaller ones, are held back by outdated beliefs regarding the cost and effort required to get started with HPC applications. Most of these are true enough for large on-premises HPC setups, but are no longer true for nimble cloud-based HPC solutions. As cloud-based HPC solutions have matured rapidly, they have become much easier to start using. Even small teams with limited resources are finding that they can test whether HPC can help them innovate faster, or get products to market faster, without taking huge risks with their budgets.

TRANSITIONING AND ONBOARDING

Until recently, organizations that switched from on-premises HPC to cloud-based solutions had to deal with transition issues like license management, or the need to revisit their systems to manage the use of elastic compute resources. As cloud-based HPC has matured, support ecosystems have developed around it to help make the transition simpler and less expensive. Of course, smaller organizations new to HPC won't have to deal with these issues and can take advantage of the support structures to jump-start their HPC efforts with cloud-born or cloud-native HPC applications.

Today, there are many options to help ease organizations through onboarding to first-time HPC use or transition from traditional on-premises HPC to the cloud. While internal change management is still up to the organization, most of the heavy lifting involved in getting started can be handled by the cloud provider or a third-party system integrator

Transitioning from on-premises HPC solutions is relatively simple now. Many on-premises applications are adding cloud-friendly licensing models, and new cloud-oriented ISVs are developing cloud-first applications to challenge industry leaders. In most cases, the cost of transitioning and decommissioning old hardware is more than offset by gains in productivity, innovation, and accelerated time-to-market.

And it's easier than ever to skip past traditional solutions and get started directly in the cloud. Small organizations can explore options with minimal investment and can get assistance from third-party vendors like Ronin, who develop portals that help small teams start doing their research without having to dive into the details of setting up HPC clusters.

Case Study: CADFEM

CADFEM UK and Ireland Ltd was founded in 1997, specializing in computer-aided engineering (CAE) for a range of sectors including renewable energy and aerospace, and also provides consultation, training, and High Performance Computing to help firms develop accurate models.

It was hard to compete with larger organizations when bidding on major contracts, where computational needs are extremely high. Small firms often don't have the budget to build large clusters, but as cloud-computing services started becoming more widespread and affordable, new opportunities arose.

CADFEM believes using cloud-based HPC to support simulation has "levelled the playing field" between small, specialized engineering firms and large enterprises, helping them stay ahead of the competition.

Uptake of CADFEM's simulation solutions has been very high, and they are easy to deploy internationally thanks to widespread coverage of data centers. They can give new customers a trial environment within 30 minutes, and once they've decided to buy, setting them up takes about a day, as opposed to two weeks with a physical workstation.

NEEDS ANALYSIS

Understanding needs is one of the oldest business problems, but it's also fairly straightforward once the initial trial-and-error period has led to solid results. Any organization considering HPC applications as part of their research or engineering programs needs to ask two big questions:

1. **What are our infrastructure requirements?** For on-premises HPC, infrastructure size is often dictated by budget, but the pricing and flexibility of cloud-based HPC means that a precise awareness of specific needs will be rewarded with reduced costs and less downtime for researchers.

2. **How much capacity will we need over time?** Correctly predicting need is a major driver of ROI in HPC, whether on-premises or in the cloud. Big capital expenditure items, like on-premises HPC infrastructure, have a 3-5-year procurement cycle. Organizations of all sizes usually struggle with predicting the capacity needed for the next 3-5 years. Buying based on an inflated expectation of growth leads to expensive, unutilized capacity. Pessimistic forecasts lead to oversubscribed resources and lower productivity. This can be especially challenging for smaller organizations or those new to HPC. Cloud-based HPC eliminates the need for long term forecasting, thanks to near-instant access to any required capacity and the latest technologies.

So many cloud-based use cases have been well defined by this point that it's become much easier for most organizations to find a starting point that matches their needs. System integrators can help with initial consultation or with managing needs throughout the lifespan of any given HPC project.

Of course, each organization understands its own needs best, but translating those needs into real-world compute resources does not need to be a cumbersome process. Smaller organizations would rather have their expensive engineering or research talent focus on what they do best, instead of figuring out their infrastructure needs.

Case Study: Metabiota

Metabiota helps its clients conduct risk analysis and assess the probability of human and financial losses caused by potential epidemics through its comprehensive infectious disease platform. They use hundreds of different data sources to build a range of models that run tens of millions of plausible event simulations, each representing a scientifically plausible, hypothetical scenario.

Metabiota initially sought to manage its custom code creation and simulation modeling while simultaneously managing and monitoring the HPC environment. Managing the underlying infrastructure meant Metabiota's data science team had less time to focus on research and model development. Furthermore, pinpointing the source of errors encountered while running millions of simulations became a drain on both Metabiota's time and its resources.

Rescale, a third-party service provider, has one key mission: to help organizations seamlessly run compute-intensive workloads of any size and scale by harnessing the power of cloud computing and the enterprise readiness of ScaleX, Rescale's HPC software as a service (SaaS) platform. Metabiota approached Rescale to understand how the Rescale platform could address its particular challenges. Rescale evaluated Metabiota's requirements, taking into consideration the team's need for data access on the platform to be strictly governed according to the different organizations involved. Metabiota found they could integrate proprietary code bases seamlessly, quickly identify errors within the scaling code, track each error down, and correct each error using Rescale much more quickly than on their own.

Today, Metabiota can focus more time on building robust models and allow Rescale to scale production up to tens of millions of simulations. This lets them be more client-focused and modify methods in order to suit client timelines better.

MANAGEMENT AND REPORTING

For smaller organizations moving into the HPC domain, managing infrastructure usually adds another layer of complexity. The need for managing licenses, tracking resource usage and mapping it to projects, and scheduling priorities is often an added responsibility for researchers and engineers in these firms.

While cloud-based HPC solutions can't eliminate these requirements entirely, they can and do streamline them and make them much easier for small, modestly funded teams to accommodate. Basic tracking management is baked into most cloud-based solutions, and it's easy to find robust third-party solutions built on existing cloud platforms that provide simple user-friendly interfaces to simplify infrastructure management.

Cloud-based HPC solutions like HPC on AWS offer in-depth dashboards and other tools to help clients manage and predict HPC usage. This makes it much simpler to stick to the budget and allocate resources as per business needs

PERFORMANCE CONCERNS

Many HPC practitioners feel that HPC simply cannot be performed adequately in the cloud. There are many reasons for this, but chief among them is the belief that the networking speed between compute nodes in the cloud is not fast enough for high performance. However, recent advancements like the Elastic Fabric Adapter from AWS have helped speed up cloud networking and trim latency to the point that all but the most resource-intensive HPC applications run just as well or better on the cloud than on on-premises infrastructure. On AWS, Amazon EC2 instances support enhanced networking that allow higher bandwidth and lower inter-instance latency compared to traditional virtualization methods. This enables users to run HPC applications requiring high levels of inter-node communications at scale. Another method used to reduce latency is to use placement groups, where all nodes in an HPC cluster are allocated within a single unit, for tightly coupled HPC applications that require low latency networking.

Modern cloud-based HPC performance is more than enough for most purposes—and considering the benefits of elasticity, and improved flexibility, it typically delivers a better ROI (Amazon Web Services, 2019).

SECURITY REQUIREMENTS

Concerns about digital security are as old as the internet, and cloud-based solutions of all kinds have had to work hard to prove themselves up to the task. Over time, most cloud-based HPC solutions have come to offer much stronger security than any individual organization can typically afford to provide itself, so the question of whether the cloud is secure is largely settled. Now the main questions revolve around control, monitoring, and backup. Who is in charge of setting and maintaining security policies? How is monitoring maintained and reported—and how quickly can incidents be resolved? How frequent and reliable are backups, and how much do they cost to maintain?

It turns out that the answers to these questions are easy to arrive at in partnership with a cloud provider, sometimes in conjunction with third parties that offer custom security solutions. Most teams should find the security they need to be both affordable and easily implemented.

Many industries that are heavy users of HPC in the cloud also have stringent security requirements. As an example, healthcare companies that do business in the cloud need to comply with regulations like HIPAA, and AWS already has "quick start" programs designed to automate the process of creating compliant environments even when complex security and privacy regulations apply. This means that a smaller startup in the same industry will never have to reinvent the wheel when they start using cloud-based HPC. Built-in monitoring, encryption, and backup can relieve staff and budget pressure that might otherwise keep a research or engineering project from getting off the ground.

Case Study: DNAnexus

DNAnexus provides an API-based platform for the sharing and management of data and tools that accelerate genomic research. This platform enables scientists and clinicians worldwide to speed up medical advances, improve patient care, and enhance R&D.

The global scope of research and clinical studies requires a secure and compliant environment within which researchers can share and collaborate on datasets and tools in real time. To accomplish this, DNAnexus developed a cloud-based genome informatics and data management platform. On AWS, the DNAnexus domain-specific data management platform offers tailored compliance and security and provides fine-grained data management for transparency, reproducibility, and data provenance for consistent bioinformatics pipelines and results.

Their customers can now pursue clinically compliant projects of enormous scale and scope, confident in the quality and efficiency of analysis and the security and ease of collaboration. AWS infrastructure and the DNAnexus platform controls and certified compliance lets them meet the demanding requirements of HIPAA, CAP/CLIA, GxP, and other privacy laws and regulations.

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 EC2instances 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.

CONCLUSION

HPC is vital to many industries and fields of research, including many that don't traditionally have the funding to build their own on-premises HPC solutions. Fortunately, cloud-based HPC solutions offer flexible, affordable compute power—but getting started can feel intimidating, especially for smaller organizations that may not have the expertise to set up and manage their own systems. But barriers to entry are fading away quickly, and those that still exist are easier to bypass than most people realize. We are entering a new phase of democratized cloud-based HPC solutions, where nearly anyone with a problem to solve can find the means to get their questions answered reliably, securely, and affordably.

AWS HPC offers accessible solutions for most business cases, from small research teams to large enterprise organizations looking for alternatives to maintaining their own large, expensive on-premises HPC solutions. Learn more about HPC on AWS at http://aws.amazon.com/hpc/

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

References

Amazon Web Services, "What a TCO analysis won't tell you: Dig deeper to discover the true cost of your on-premises HPC investments," 2019, https://d1.awsstatic.com/HPC2019/AWS%20HPC%20-%20What%20a%20TCO%20analysis%20wont%20tell%20you.pdf