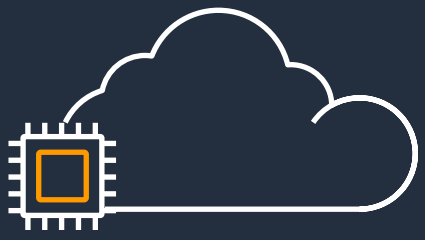
Running HPC Workloads on AWS

Why HPC on AWS

AWS has developed leading cloud computing services which are tailor-made for demanding HPC workloads. By selecting the right combination of cloud-based compute and storage resources, customers can meet the demands of the most challenging HPC applications. AWS provides an elastic and scalable cloud infrastructure, powerful orchestration tools and advanced services to make it easy to quickly deploy and operate a sophisticated cloud based HPC system. With Amazon EC2, you can choose from a broad range of compute instances allowing you to match the optimal instance with your particular workload's characteristics. You can then combine this with high performance storage and networking options built specifically for HPC workloads. This means engineers, researchers, and HPC system owners can innovate beyond the limitations of on-premises HPC infrastructure, enabling scale out applications to run without limits. Workloads from many different fields of science and industry run on AWS, including HPC applications such as genomics, computational chemistry, financial risk modelling, computer aided engineering, weather prediction, and seismic imaging, as well as emerging technologies such as quantum computing and machine and artificial intelligence/ deep learning workloads.



$507

revenue for every $1 invested in HPC[1]

The AWS advantage

HPC on AWS provides a number of key benefits when compared to an on-premises HPC environment, including:

- **Rapid deployment** – With AWS you can spin up an HPC cluster in minutes allowing you to react quickly to changing business demands.
- **Compute elasticity** – You can have an HPC system with as little as a single compute node stretching all the way to 1 million cores or more, rapidly reducing the time to results.
- **Flexibility of configuration** – Your cloud-based HPC can consist of a mixture of differing compute instances to meet the needs of a variety applications and workloads. This means there's no need to compromise with a one size fits all approach, and instead you can select the best compute profile to meet your HPC workload needs.
- **Purpose-built HPC tools and services** – AWS offers a range of services designed specifically to support HPC workloads such as AWS Batch, EFA low latency networking, FSx for Lustre and DCV, with tools such as Parallel Cluster for set up and operation of your HPC cluster in AWS.
- **No need for data centers** – Running HPC workloads in the cloud means no further need for expensive data center facilities, mains power circuits and air conditioning systems. It also shifts the burden of hardware procurement, maintenance, refresh cycles and OS software licensing to AWS.
- **Enables productive remote workplace** – Users can gain access to high performance remote Virtual Desktop computing sessions that can provide similar performance to that of workstation computers used in engineering or design offices. For staff working from home, a VDI session powered by NICE DCV or AppStream allows for demanding desktop applications to be used without compromising the performance or security of your assets and intellectual property.

AWS benefits

Why move from on-premises

AWS can provide a greater ROI and lower Total Cost of Ownership (TCO) when compared to operating a fixed infrastructure acquired through a capital purchase. You can easily size your HPC to meet changing demands without having to operate a fixed infrastructure sized for a workload peak. As an example, a spike in compute demands at the end of the working day, as found with investment banking risk workloads.

Why move to AWS

AWS has been named a leader in the Gartner Magic Quadrant for IaaS and PaaS, for both completeness of vision and ability to execute. AWS also has among the broadest and most complete range of services designed for HPC. This includes the widest range of compute options featuring processor architectures from Intel, AMD, NVIDIA and Arm, including the latest Arm Graviton2 based EC2 instances. Amazon EC2 compute instances benefit from hardware-based virtualization using AWS Nitro, which manages compute virtualization without placing the overhead on the instance, giving near bare metal performance. There are HPC specific instances such as the C5n, storage options such as FSx for Lustre and high-speed low latency EFA networking, along with various tools and services to cover virtually every application and use case. These include services such as AWS Batch and ParallelCluster. AWS Batch dynamically provisions the optimal quantity and type of compute, based on the volume and specific resource requirements of submitted batch jobs, allowing you to pay only for the compute you use. ParallelCluster on the other hand makes cluster deployment in the cloud simple and straight forward. Last but not least, if you need help in moving your HPC workloads to the cloud, AWS has an extensive network of AWS Partners who can provide services and support to help you as you migrate your applications to AWS.

Ease of use

Running HPC in the cloud means no longer having to be responsible for hardware procurement, operation, maintenance, updates, operating system deployment and licensing. All AWS infrastructure is orchestrated from the console or command line. This enables rapid deployment of services, resources and automation, with features such as AWS CloudFormation Templates, AWS Step Functions and simple code execution using serverless technologies such as AWS Lambda. Alternatively, AWS Batch is a managed service that makes it even easier, as you don't have to manage the cloud infrastructure to run jobs, just submit them.

Security

Security is a fundamental element of AWS cloud, with data centers and a network architected to protect your information, identities, applications and devices. With AWS you can meet requirements such as compliance, data sovereignty and confidentiality using services and features built in and with those of our partners. Information is protected in the cloud with data stored in Amazon S3, which is designed to provide eleven 9s of durability.

Flexibility

The cloud offers far greater flexibility than a fixed on-premises environment, both in terms of size of the system and also the hardware and software profile. You can grow and shrink your cloud HPC environment based on your demand. This means you're only paying for what you use and not over-provisioning to meet a workload demand peak. You can also burst to the cloud from your on-premises HPC environment using a number of common 3rd party HPC schedulers including Univa Grid Engine, IBM Spectrum LSF, SLURM, Adaptive Computing Moab, or by using Bright Cluster Manager to automatically add and remove compute resources.

Faster results

With AWS you have access to virtually unlimited resources, meaning it's possible to deploy an environment of a few hundred cores to more than 1 million CPU cores, providing the ability to massively reduce the time to obtain results. One AWS customer, Western Digital, was able to run 3 weeks work in 8 hours by using 1 million cores, which has a major impact on time to market for one of their key products.

Industry solutions

Financial services industry

Industry challenges and painpoints:

The Financial Sector (FS) industry is highly regulated, placing reporting and compliance demands on banks and insurance companies. This translates into a high compute demand to enable the FS organizations to model and calculate risk, and provide data to the industry regulators. Customers typically run low latency grid schedulers such as IBM Spectrum Symphony, Tibco Data Synapse, or Windows HPC Pack.

AWS advantage:

Having access to on demand cloud capacity provides FS institutions with the additional compute resources needed to execute risk and compliance workloads for end of day reporting. All the leading grid schedulers run on AWS, ensuring high performance and throughput for demanding risk workloads.

AWS technologies for FSI:

EC2 (C5, C6, z1d, M5), S3, ParallelCluster, AWS Batch, EC2 Spot Instances

Examples:

- Bankinter
- Finra
- Pacific Life
- Standard Chartered

Life Sciences

Industry challenges and painpoints:

The Life Sciences sector is typically driven by project-based working, which means variability in the demand for HPC resources. On-premises HPC can be stretched one week and under-utilized the next, meaning it's difficult to have the right sized HPC environment. Life Sciences workloads can generate large amounts of data and so storing this securely, but being able to access and share it when necessary, is also challenging.

AWS advantage:

By shifting Life Sciences workloads to the cloud organizations only pay for the cores needed to complete the ever-changing needs of their jobs, with the ability to securely store and archive large amounts of data.

AWS technologies for Life Sciences:

EC2 (C5, z1d, M5), S3, FSx for Lustre, EnginFrame, DCV, ParallelCluster, AWS Batch, EC2 Spot Instances

Examples:

- AstraZeneca
- Genalice
- Fabric Genomics

Autonomous Vehicles (AV)

Industry challenges and painpoints:

Developing autonomous technology requires acquiring, processing, and storing petabytes (PB) of data in order to train and optimize deep learning models which operate the vehicle. To shorten time-to-results, AV companies leverage accelerated compute nodes. However, acquiring a large quantity of these platforms can be costly and requires a big upfront investment.

AWS advantage:

AWS provides infrastructure for PB-scale storage and access to the widest range of accelerated compute capacity in the cloud to handle massive scale, distributed deep learning workloads. A number of AV companies around the world rely on AWS for autonomous system development. Data can be stored in an AV data lake built upon Amazon S3. AWS's suite of accelerated EC2 instances includes technology from NVIDIA, Intel Habana, Xilinx, and AMD. AWS also provides custom hardware built specifically for training (Trainium) and inference (Inferentia) workloads. This broad range of technology, available on demand, ensures AV companies can efficiently develop self-driving technology without massive up-front investment.

AWS technologies for Autonomous Vehicles:

EC2 (M5, C5, R5, G4, P3, P4), SnowBall, Direct Connect, S3, FSx for Lustre, EKS, ECS, AWS Batch, ParallelCluster

Examples:

- Lyft Level 5
- Mobileye
- Toyota Research Institute (TRI)
- Weride

Oil and Gas

Industry challenges and painpoints:

Oil and Gas exploration and extraction requires massive amounts of compute. Complex and demanding applications used for reservoir simulation require powerful instances with low latency networking to run effectively. Seismic analysis requires large scale compute resources and involves large volumes of data which need to be processed in order to identify potential hydrocarbon deposits. Users need access to high performance 3D workstations in order to view and interact with the output from the simulations being run.

AWS advantage:

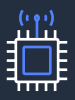
Having access to virtually unlimited amounts of compute resource allows for large scale processing of seismic survey data handling multiple projects in parallel, and provides a mechanism to assign costs on a per project basis. Reservoir modelling is possible using HPC instances such as the C5n with Elastic Fabric Adapter, supporting demanding applications used to optimize drilling and extraction. Using AppStream and DCV, users can access high performance virtual workstations capable handling demanding 3D applications.

AWS technologies for Oil and Gas:

EC2 (C5n,C6,R5,M5,G4, P3, P4), EFA, FSx for Lustre, AWS Batch, ParallelCluster, EnginFrame, NICE DCV

Examples:

- WoodSide
- BP
- Hess



Electronic Design Automation [EDA]

Industry challenges and painpoints:

The design, verification, and manufacturing of advanced semiconductor devices requires extremely large amounts of compute resource to deal with a very large volume of workloads generated by design and verification engineers. EDA customers may require HPC throughputs of over 1million jobs per day on their clusters, meaning throughput and utilization are critical factors in time to market for their products. Semiconductor design and verification throughput is critical to meeting product schedules, and for optimizing total engineering costs. Cost optimization for semiconductor design includes ensuring that engineering and verification staff are not sitting idle, waiting for simulation and analysis results, and that EDA software licenses are kept fully utilized. This is where AWS can help, by enabling rapid scale-up and scale-down of resources to meet workflow needs, and by allowing each application to be optimized using the best performing EC2 instance and storage options.

AWS advantage:

AWS provides access to virtually unlimited compute resources to handle large scale EDA workloads, providing additional capacity to supplement on-premises HPC resources or enable a cloud first approach to silicon design. To drive value for EDA ISV software licenses, AWS has instances such as the z1d and M5zn, which have clock speeds of 4.0 and 4.5 GHz respectively. AWS also supports all major CPU/GPU technologies from Intel, NVIDIA, AMD, Arm. Additionally, AWS Graviton processors are custom built by AWS using 64-Bit Arm Neoverse cores for maximum flexibility and compute price/performance.

AWS technologies for EDA:

EC2 (C5, C6, z1d, M5, R5), S3, ParallelCluster, AWS Batch, EC2 Spot Instances

Examples:

- Xilinx
- Innovium
- Mediatek

Manufacturing

Industry challenges and painpoints:

The main challenge facing this sector is the increasing demand for simulation capabilities to innovate faster and reduce the need for physical prototyping. This places continuous demands on HPC resources leading to oversubscription and long delays to projects and launch dates. Users are running demanding and very expensive applications, and so maximizing throughput and performance is a key requirement. HPC administrators have to try and allocate and schedule access to overloaded compute resources and in-demand application licenses, trying to prioritize access to HPC resources based on the business needs.

AWS advantage:

Leveraging HPC specific EC2 instances such as C5n with 100 Gbps networking EFA enables complex applications such as CFD and FEA to perform very well on the AWS Cloud. Customers can shift a backlog of HPC workloads, support extra project-based workloads while maximizing application license usage.

AWS technologies for Manufacturing:

EC2 (C5n,C6,R5,M5), EFA, FSx for Lustre, ParallelCluster, EnginFrame, DCV AV

Examples:

- INEOS
- Western Digital
- Boom Aerospace
- TLG Aerospace

Key products and services

Compute

AWS provides a broad range of compute services and instance types to suit HPC customers' needs and supports all major technology vendors including Intel, NVIDIA, Arm and AMD. Customers can select the compute instance to exactly match the profile and behavior of their HPC applications, including CPU optimized, memory optimized or GPU or FPGA accelerated, with options for low latency networking and high performance attached storage available on many instance types.

- **Compute Optimized** – C5/C5n/M5zn (Intel), C6gn (Arm/Graviton2) instance options for applications that require increased CPU performance.
- **Memory Optimized** – R5/R5n (Intel), R5a (AMD), R6g (Arm), X1/X1e/X1d/High Memory (Intel 4 socket) instances for customers that require larger memory requirements.
- **Accelerated Optimized** – P2/P3/P4 (NVIDIA GPU), G3, G4 (NVIDIA), Inf1 (Inferentia CPU), F1 (Intel high clock frequency) instance options for customers that require GPU instances for GPU workloads and remote visualization.
- **Storage Optimized** – I3/I3en/D2/H1 (Intel) various options for high performance local IO. Storage optimized instances are tuned for customers with large IO needs.
- **AWS Nitro System** – The Nitro system offloads virtualization layer to a dedicated Nitro card which enables AWS to provide near bare metal performance with the benefits of virtualization.
- **Quantum Computing Services** – AWS Braket provides a fully managed quantum computing service including developer framework, tools, simulators and access to physical quantum computers on a pay-as-you-go model.

Storage

AWS provides a range of solutions to help customers design the storage and IO fabric to suit the needs of their application. This includes locally attached scratch/temp using block or file storage, as well as high performance parallel file system options and long-term data storage in a data lake such as S3 or S3 Glacier.

- **EBS/Local IO** – Elastic Block Storage and ephemeral storage options (instance dependent) provide required storage capacity and performance needs for applications.
- **FSx for Lustre** – Provides a high performance parallel file system and enables multiple compute instances to share a performant storage cache for IO demanding workloads.
- **Amazon Elastic File Systems (EFS)** – Provides a simple, scalable, fully managed Elastic NFS filesystem.
- **Amazon S3** – S3 can be used as a foundational data lake to store research data.
- **AWS Direct Connect** – Provides customers with a dedicated network connection performance tuned to enable data transfer required for HPC workloads.

Networking

AWS provides options for high bandwidth low latency networking with Elastic fabric Adapter (EFA), which attaches to EC2 compute instances and delivers 100Gbit Ethernet connectivity coupled with AWS low latency drivers and OS bypass, enabling demanding MPI applications to run without constraint.

- **Elastic Fabric Adapter (EFA) –** Is a network interface custom built by AWS that enables HPC customers to run applications with low-latency, high-throughput inter-node communications at scale.
- **AWS Autoscaling –** Enables your HPC Cluster to grow and shrink on demand.
- **Cluster Placement Groups –** Ensures compute resources are deployed physically close to reduce network hops and therefore latency.
- **Enhanced Networking –** Enables you to configure your HPC environment with best practices for HPC deployments.
- **AWS VPC** – enables you to provision locally isolated sections of the AWS cloud so you can launch AWS resources in a virtual network that you define.

Automation and Orchestration

AWS provides a range of automation and orchestration tools to simplify the process of running HPC jobs in the cloud. AWS ParallelCluster is used to deploy and manage an HPC Cluster in the cloud, including setting up the head and compute nodes, scheduler, tools and libraries, storage and applications and monitoring the health of the cluster. AWS also provides a managed service known as AWS Batch, which manages all of the elements of a HPC system and allows users to submit workloads and only pay for the resources used to execute the jobs.

- **AWS Batch –** Enables running batch computing jobs on AWS, dynamically provisioning the optimal quantity and type of compute resources.
- **AWS ParallelCluster –** Is an AWS supported open source cluster management that makes it easy for you to deploy and manage HPC cluster on AWS.
- **NICE EnginFrame –** Is an advanced web front-end for accessing and managing HPC clusters and deploying applications.
- **Amazon EKS –** Enables Kubernetes applications in the AWS cloud or on-premises.
- **AWS Step Functions** – Enables complex HPC workflows to be managed and orchestrated.
- **AWS Lambda** – Is code execution as a service, providing serverless running.

Frameworks

- **Scale-Out Computing (SOCA)** – Provides a ready-to-use simple to deploy multi-user HPC environment that can be tuned for specific Industry workflows.

Visualization

To assist customers in accessing HPC resources and visualizing the output of HPC workloads, AWS offers Enginframe, a web-based portal providing access to your HPC systems in the cloud. It gives a simple to use interface to manage access to your applications, data and HPC jobs, meaning users don't have to learn complex command line syntax in order to take advantage of the HPC system.

- **NICE DCV** – Enables customers to run high fidelity remote desktop visualizations of simulations using high-performance remote display. Its protocols support 4K resolution when configured with an AWS G4 instance.
- **Amazon AppStream 2.0** – Streams your applications from AWS to any computer, including Chromebooks, Macs, and PCs.

Management

Using foundational AWS services you can manage an HPC environment with true granularity enabling assessment, audit, governance, compliance and monitoring of all resources consumed in the cloud. Tools such as AWS Cloud trail and Cloud Watch provide usage tracking and monitoring, and AWS Budgets assists with planning and cost control, ensuring your cloud usage is matched with your budgets and spending.

- **AWS CloudTrail** – API usage tracker for governance, compliance, operational auditing and risk.
- **Amazon CloudWatch** – Monitoring and observability service.
- **AWS Config** – Enables you to assess, audit and evaluate the configurations of your AWS resources.
- **AWS Cost Management** – A range of tools to help you manage your run cost of HPC environments and workloads.
- **AWS Budgets** – Improve planning and cost control with flexible budgeting and forecasting.
- **AWS Identity and Access Management** – Securely controls access to AWS services, and enables you to centrally manage users, security credentials, and permissions.

Customer Services/Assistance

AWS offers a range of support plans that provide access to tools and expertise that support the successful deployment and ongoing operational health of your AWS solutions. All AWS support plans provide 24/7 access to customer service, documentation, technical papers, and support forums. You can choose a support plan that best aligns with your AWS use case, providing you resources to plan, deploy, and improve your AWS environment.

AWS Partner Network

The AWS Partner Network (APN) is a global community of partners with skills and experience in designing, deploying, migrating and supporting solutions running in the AWS cloud. AWS Partners can provide additional tools, services, support and training in order to maximize your effectiveness and performance when running workloads in AWS.

HPC Application ISVs

AWS Partners that provide application software for workloads such as Computational Fluid Dynamics (CFD), Molecular Modeling, Reservoir Modeling, and Weather Modeling.

Partners: Ansys, Siemens, Altair, Cadence, Synopsys, OpenEye Scientific, CFD Direct, OnScale, S-Cube

HPC Management

AWS Partners that provide a fully managed cloud HPC environment and provide solution such as end-to-end cluster provisioning, deployment, management, and support for customers to deploy their HPC workloads on AWS.

Partners: Rescale, Altair, IBM, SchedMD, Zenotech, Ronin, TotalCAE, Core Scientific

Foundational Technology

AWS Partners that provide enabling technologies such as processors, accelerators, and operating systems for customers to run their HPC workloads on AWS.

Partners: Intel, Nvidia, AMD, ARM

Consulting partners

System integrators and strategic consultancies, that help customers of all types and sizes accelerate their HPC journey to the cloud.

Partners: Six Nines IT, Ronin, NTT (Flux 7), Aneo, OCF, NAG, BioTeam, Clovertex

Case Studies

- University of Sydney's Wildlife Genomic group protects wildlife using AWS and Ronin
- Amazon Prime Air's drone takes flight with AWS and Siemens
- Nissan and Rescale: Innovation that excites
- Astera Labs uses AWS, Intel and Six Nines to accelerate chip development
- OpenEye Scientific helps Beacon Discovery improve drug discovery process
- Woodside Energy uses AWS and S-Cube to slash the time taken to deliver Seismic data to the desktop from weeks to hours

Whitepapers, eBooks, and Infographics

- Whitepaper: Lowering Time-to-Results with Elastic Fabric Adaptor
- Whitepaper: The Cloud Steps up to Tightly Coupled HPC Codes
- Whitepaper: Challenging Barriers to High Performance Computing in the Cloud
- Analyst Spotlight: Computational Evaluation of Commercial Cloud HPC with a Global Atmospheric Model
- Whitepaper: Democratizing High Performance Computing
- Achieving optimal price/performance for your HPC workloads on AWS
- Hyperion Research Technology Spotlight: Smart Orchestration Speeds HPC Workflows in the Cloud
- Whitepaper: HPC on AWS Redefines what is possible
- Whitepaper: What a TCO Analysis won't tell you

Getting Started Projects

- Create an Elastic HPC Cluster using AWS ParallelCluster
- Create an End-to-End HPC Environment
- Getting Started with Amazon FSx for Lustre
- Getting started with Elastic Fabric Adapter(EFA)

Start your AWS journey for running HPC workloads with the following technical resources, contact your AWS account team or use the "Contact Us" page to reach us directly.

https://aws.amazon.com/hpc/

1 https://www.hpcuserforum.com/ROI/ - November 2020