Optimizing HPC workloads with Amazon EC2 instances

An increasing number of customers are choosing to run their high performance computing (HPC) workloads on AWS, because of the scale and flexibility the cloud offers. With access to virtually unlimited capacity, engineers, researchers and scientists can maximize their research efforts and reduce time to results when compared to a fixed on-premises HPC environment. HPC system owners can deploy a cloud-based cluster environment in minutes, incorporating a range of AWS services designed specifically for demanding HPC applications. This flexibility, and along with the on-demand nature of the cloud, allows for a different approach to meeting the needs of HPC users and applications.

To get the best performance for your workload, you should choose the right combination of services to meet the needs of your applications. AWS offers flexible compute, storage, networking and software to enable the creation of an HPC environment to match your specific requirements. A big part of this will be choosing the right Amazon Elastic Compute Cloud (EC2) instances from the wide range of EC2 instance types available.

Different families of EC2 instances are optimized for a range of workloads, allowing the selection of the correct CPU, GPU, FPGA, memory, storage and networking along with the correct operating system, based on the needs of your applications and the demands of your industry or vertical market.

To assist with identifying the key features and capabilities of these EC2 compute instances and then matching them to the appropriate workloads, we have described below a number of popular HPC workloads and the most appropriate EC2 instances for each one. This will ensure you have the best match of compute platforms to support your HPC requirements from the broad selection available in EC2.

Industry Specific Workloads and Applications

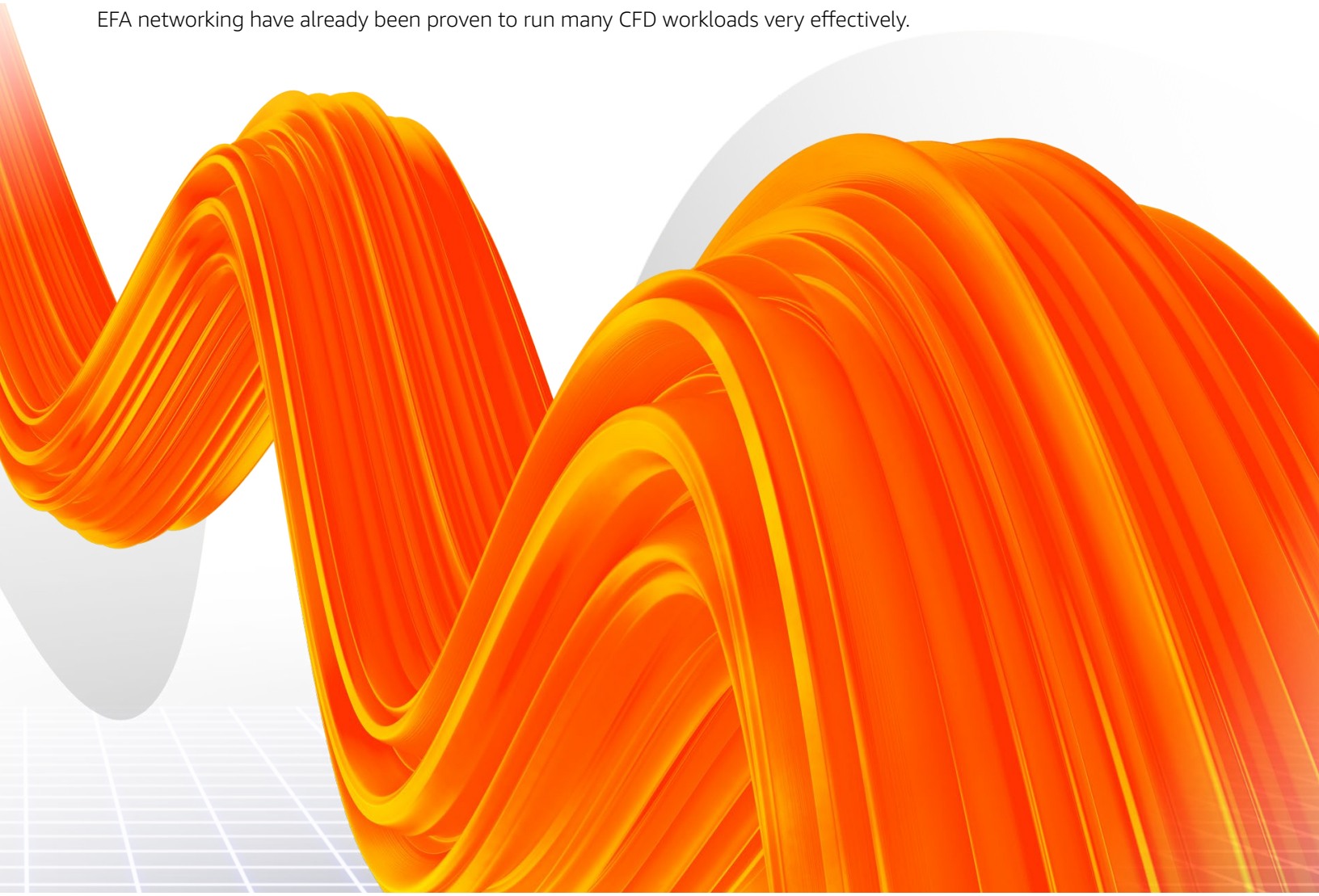
Manufacturing/Engineering/Automotive/Oil & Gas

Computational Fluid Dynamics

Computational Fluid Dynamics (CFD) is the science of modeling fluid flow in many different industrial and engineering settings, including virtual wind tunnel and aerodynamic simulations as well as hydrodynamics simulations for flood management, hydraulics, chemical processing and many others. It is commonly used in the automotive and aerospace industries, with every major car and airplane manufacturer using CFD to simulate aerodynamic effects to help them design and develop more efficient products that deliver better performance. Competitive sports such as yacht racing (INEOS Team UK), Formula 1 and other motorsports use CFD extensively to assist with the design and development of competitive and efficient designs.

Applications include Ansys Fluent, Siemens Star CCM+, OpenFoam, zCFD and many others.

CFD applications typically depend on MPI for parallel communications, so using EC2 instances that support low latency networking with Elastic Fabric Adapter (EFA) is highly recommended. Instance types such as the C5n, with 100 Gbps EFA networking have already been proven to run many CFD workloads very effectively.



Automotive/Engineering/Construction/Manufacturing

Finite Element Analysis

Finite Element Analysis (FEA) is the simulation of physical phenomenon, using a numerical technique known as the Finite Element Method. FEA uses advanced mathematics to calculate material behavior and structural stress and is used in a broad range of engineering and manufacturing sectors. FEA allows engineers to design and prototype new designs in a virtual environment without the need for physical and destructive testing. Industries that use FEA include aerospace and automotive manufacturers, who use it to help determine strength, behavior, durability and lifespan of a product.

As an example, automotive manufacturers use FEA to simulate impacts, meaning they can iterate quickly on a new vehicle design to develop and enhance its ability to protect its occupants and pedestrians during an impact. Architects and construction companies use FEA to test the designs of high-rise buildings to analyze behavior in situations such as the extreme wind loads experienced during a hurricane. Applications used to run FEA workloads include LS-Dyna, Simscale, Nastran, Pamcrash, Abaqus and COMSOL with many others also available.

FEA applications can be embarrassingly parallel, or tightly-coupled, meaning they depend on MPI and require access to a low latency network to enable maximum application performance. EC2 instances such as the C5n are particularly suited to FEA applications due to their balance of CPU, memory and network performance. With 100 Gbps EFA networking, a large number of CPU cores and memory capacity of up to 192GB, these instances provide all the elements necessary to run demanding FEA workloads.

However for FEA codes that demand a very large memory footprint, and are scaling out to only several nodes, customers should consider the Z1d, or the M5zn for applications that require high throughput, low latency networking.

Manufacturing/Engineering/Automotive/Financial Services

Monte Carlo simulations and high throughput workloads

Monte Carlo simulation and high throughput workloads are typically run in industries such as financial services, engineering and manufacturing, insurance and transportation. These applications are used to model risk and predict possible future outcomes, or process large volumes of data to gain insights. Investment banks use it extensively to model all forms of risk and it forms a fundamental part of calculating and reporting risk and exposure to industry regulators. Electronic Design Automation (EDA) and Seismic imaging are also typically categorized as high throughput workloads.

These workloads generally involve many largely independent work items, meaning they do not depend on a high-speed network for optimum performance. For workloads where high single threaded performance is most important, consider the M5zn with a clock speed of 4.5GHz (up to 192GB of memory) or z1d with a clock speed of 4.0GHz (up to 384 GB of memory). For other applications that prioritize lower cost per core over single threaded performance select the C5, M5, or R5 (based on memory requirements).

Life Sciences

Computational Chemistry

Computational chemistry is used in a range of industries including life sciences and pharmaceuticals, manufacturing, chemicals and petroleum and many other research disciplines. It includes fields such as quantum chemistry and molecular dynamics, and involves a number of methods that are used to study the structure, properties and behavior of molecules and atoms.

Many applications in this field can run on both CPUs and GPUs, including GROMACS, AMBER, NAMD and LAMMPS, NW-CHEM, and CP2K. Depending on the specifics of your calculations, you may find that using GPU-accelerated instances such as the P4d with NVIDIA A100 GPUs is most efficient. When running with only the CPU we recommend the C5n. Both the P4d and C5n offer access to EFA for low latency and high bandwidth communication.

A number of codes would be suitable to run on GPU enabled EC2 instances such as the P4d including MOLPRO, NW-CHEM, and CP2K. Other codes used in molecular dynamics such as AMBER, GROMACS, NAMD and LAMMPS use MPI and so having access to a low latency network such as that offered by the C5n are key to unlocking application performance.

Biology and Bioinformatics

Accelerating drug discovery, treatments and therapies couldn't be more important during a global pandemic, with many leading research centers and pharmaceuticals companies focused on studying and developing new treatments to fight diseases such as cancer and COVID-19. Areas such as drug discovery, screening, genome sequencing, DNA and genome analysis and clinical trials are all being accelerated using HPC to reduce the time to results and insights.

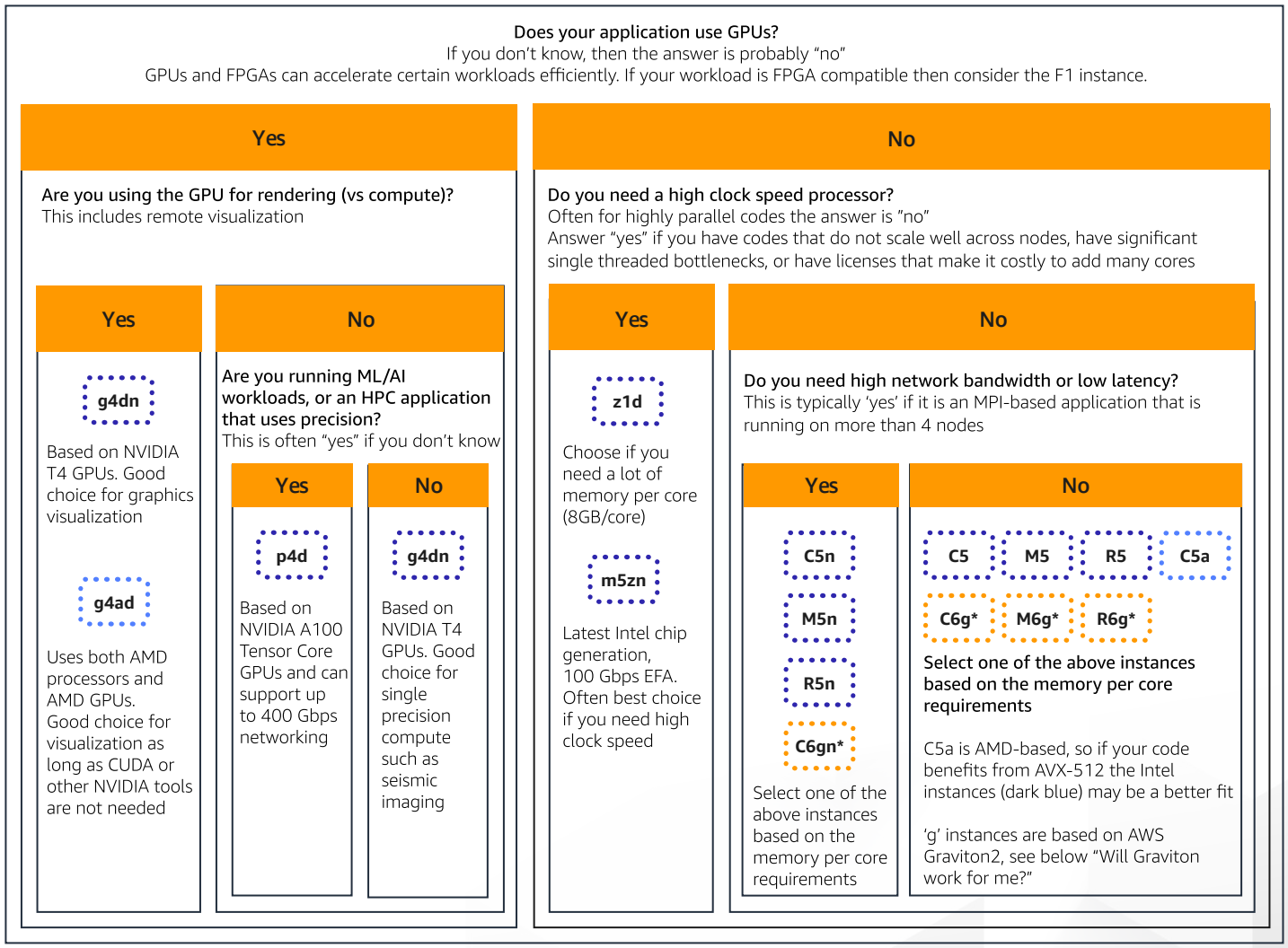
Applications used in this field include Cell Ranger, which is used to process RNA sequence outputs and gene expression analysis, and are typically serial workloads meaning CPU performance and memory bandwidth are key factors in overall application performance. Compute instances such as the C5 family and M5 family would be particularly suited to these types of workloads.

Advanced Research including Earth Sciences, Astrophysics, Climate Science

Leading research centers around the world are driving innovation and scientific discovery in a range of disciplines including Earth sciences such a weather forecasting, climate modeling, and astrophysics. Scientists are also studying the cosmos and the formation of the universe, and physics, where researchers are investigating materials sciences, particle and solid state physics, and more.

Many of the applications used in these fields of research such as WRF for weather forecasting and climate modeling, and AMReX/Castro used in astrophysics depend on MPI to handle interprocess communications. This means low latency networking is a key requirement to enable this to perform effectively. Instances such as the C5n with EFA are designed specifically to meet the needs of this type of workload.

Choose the right Amazon EC2 instance for your HPC workloads



*Will Graviton work for me?

If you are able to build your application for Arm architecture, you can benefit from the price performance offered by Arm-based AWS Graviton2 processors. AWS Graviton2 based C6g and C6gn instances are optimized for compute intensive workloads. This may include popular open source software like numerical weather forecasting models and fire dynamics simulations or other open source applications.

Graviton2-based Instances (ARM) **Intel-based Instances (x86)** **AMD-based Instances (x86)**

Workloads and instance types

The broad range of potential HPC workloads have individual characteristics that place different demands on the compute element of an HPC environment. By focusing on specific workloads and the industries that run them, we aim to help in selecting the right EC2 instance to meet your needs. The first section below describes the range of workload characteristics that differing HPC applications have, and matches them to the most appropriate EC2 instance type. We have also highlighted workloads and methods by industry, to assist you if you know the type of application you are running. This will help if you aren't completely sure of the exact behavior of your preferred software.

Workload Characteristics

Compute and Network Intensive

Compute and network intensive workloads place demands on all the main elements of a compute instance, including CPU, memory and network. The C5 family includes the C5n based on Intel Xeon Scalable 8000 Platinum CPUs, or the C5a featuring AMD 2nd Gen EPYC architecture CPUs. These instances are designed to address many common workloads including computational fluid dynamics (CFD), computer aided engineering (CAE), materials science and reservoir simulation. The C5n also includes 100 Gbps EFA networking, to support bandwidth and latency sensitive applications. Applications that demand high-levels of inter-instance communications using MPI can benefit from high network bandwidth and low latency should use the C5n.

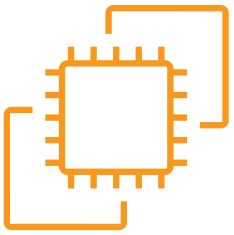
Single Threaded Performance

For applications where the high cost of commercial HPC applications licenses is the driving factor in seeking the best performance, instances such as the Z1d or M5zn would be a good match. The Z1d instance offers 4.0GHz clock speed and a memory/vCPU ratio of 8GB per core, making it a good fit for HPC applications that require high single threaded performance and high memory capacity. With the M5zn's maximum 4.5Ghz clock speed and memory/vCPU ratio of 4GB per core, it is a good fit for applications like finite element analysis using implicit methods that need high bandwidth low latency networking and ultra-fast processors, but don't need the high memory/vCPU profile offered by the Z1d.

Accelerated Computing

The additional threads and cores of a GPU can be harnessed to deliver results faster. HPC applications and codes that use CUDA and ACC can experience significant performance improvements when using accelerators. The P4d instance, with NVIDIA A100 GPUs, provides the highest performance for machine learning (ML) training and accelerated HPC applications such as natural language processing, object detection and classification, seismic analysis, and molecular dynamics. CUDA and OpenACC applications or rendering workloads can benefit from the NVIDIA T4 GPUs offered on the G4dn instances.

Customers that are able to optimize specific elements of their computation using hardware accelerators such as FPGAs can enable customized hardware acceleration for applications including genomics, analytics and financial risk modeling. FPGAs are featured in the F1 instance, and can offer over 100x acceleration when compared to CPUs for a range of use cases. To support the use of the F1 instances, AWS provides a range of tools and resources for developers to rapidly harness the performance of FPGAs.



Processor Agnostic Applications

Certain codes and applications can be compiled to run on a range of different CPU architectures. For customers that have the freedom to do so, options such as the C6g and C6gn Arm Graviton2 based EC2 instances provide a lower cost alternative for running your own codes. This may include popular open source software like numerical weather forecasting models, high performance analytics and Monte Carlo modeling as well as fire dynamics simulations and other open source applications.

Storage Options

Once you have selected the most appropriate EC2 instances to meet the needs of your workload, your next consideration will be which storage option is going to best meet your requirements, both for storage attached directly to your chosen compute instances, and storage presented as a file system attached to your cloud based HPC cluster. AWS provides options for storage including Elastic File System (EFS) or Elastic Block Store (EBS), which can be added directly to EC2 compute instances, providing local scratch storage or access to a POSIX file system should this be required by your applications. Alternatively, should your HPC simulations require a high performance file system to handle large volumes of data, AWS offers FSx for Lustre, which provides a scalable parallel file system which is attached to your HPC cluster in the cloud.

SUMMARY

Amazon EC2 instances provide the broadest range of options for running high performance workloads in the cloud, enabling you to choose the right compute platform with the right specifications to match the needs of your codes and applications precisely. AWS offers nearly limitless configuration of instances and solutions to cater to the demands of every HPC user's individual applications and workloads, and provides virtually unlimited infrastructure, latest generation processors and accelerator technology with optional high speed low latency networking. Whether you are migrating your HPC applications to the cloud or wish to burst to the cloud using a hybrid HPC model, AWS provides a broad range of scalable, flexible infrastructure and services that you can select to match your workloads and tasks.

As more and more customers consider running HPC workloads in the cloud, AWS is committed to helping customers harness the full potential of HPC in the cloud. Customers who run HPC workloads on AWS leverage the flexibility, scale and performance of the cloud to iterate quicker, innovate faster and reduce time to results.

Get Started with AWS

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