Running Engineering Simulations on HPC in the Cloud

Dr M Starr   
*(Amazon Web Services, United Kingdom);*

**Abstract**

The use of cloud computing technologies within High Performance Computing (HPC) for engineering simulation has grown considerably within the last few years. The complexity and scale that comes with cloud environments can make understanding the benefits and new capabilities that streamline tasks for HPC users and administrators harder to see. Amazon Web Services (AWS) is seeing a growing adoption of engineering customers specifically for their HPC services and abilities. Cloud compute allows customers flexibility, whether moving workloads completely off their traditional compute cluster, or using cloud compute capabilities to extend their on-premises capability by bursting engineering simulations out to the cloud as needed.

The pay-as-you-use nature of Cloud Compute means users can scale out simulations to compute sizes that would be difficult to achieve with a fixed size on-premises compute cluster. This drives design by allowing engineering simulation to increase in complexity or throughput as users get easy access to a large quantity of the latest compute hardware. With this comes cost optimisation as well as improved sustainable use of compute whereby machines don’t sit idle for periods of time. To this end, AWS are investing heavily in the latest technology - particularly Arm based Graviton - to improve the efficiency and sustainability of compute leading to a more sustainable simulation workflow.

This talk will describe these advantages to running engineering simulation workloads on Cloud Compute (with an emphasis on AWS specific services) as well as how this can help drive upcoming developments with Digital Twins, Machine Learning and Artificial Intelligence.

# High Performance Compute and The Cloud

As complexities of engineering challenges increase, so too does the fidelity of simulation needed to capture the physics involved. To gain a better understanding of physical effects, the simulation models get more complicated. It is common these days for simulations to require more computational resources than a single workstation can provide, to achieve a timely turn-around of results.

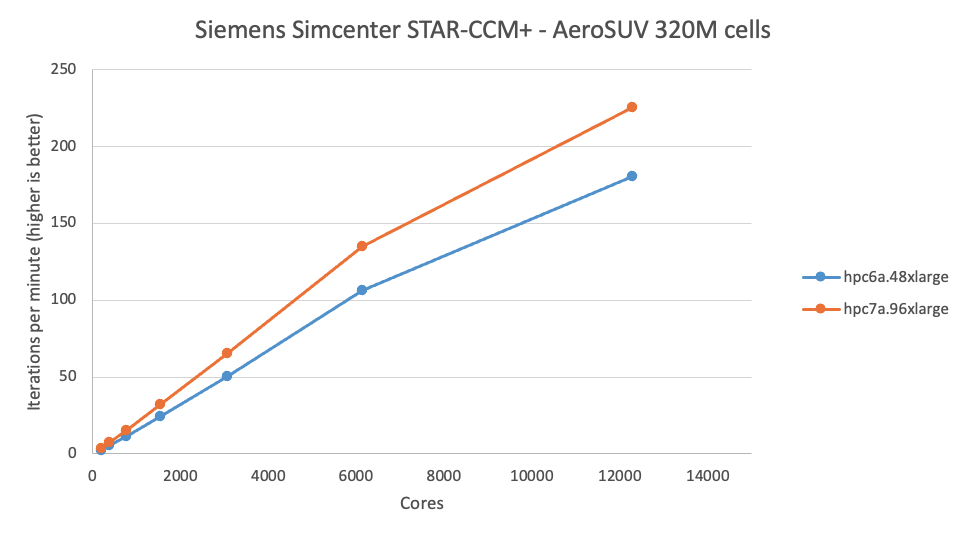
This is where High Performance Compute (HPC) becomes a tool in the workflow. Multiple computers linked with high-speed networking form a cluster that engineers can access to run simulations. The larger the cluster, the larger the simulation that can be run on it – or a greater number of jobs.

Historically HPC clusters would reside on company premises and would be refreshed periodically. Decisions would be made early in the procurement stage about the compute configuration and which type of simulation the hardware would be most biased towards. For example, some simulation types require many fast compute cores, while others need fewer cores but lots of Random Access Memory (RAM) and fast disk Input/Output capability. Once a decision was made on the size of compute cluster to be used, that decision would generally be fixed until the end of life of that cluster.

With HPC in the compute resources are consumed in an as-required approach. Workloads can be moved completely to the Cloud from on-premises hardware, or a hybrid approach is used for burst periods whereby Cloud compute extends the compute capacity available on-premises. A pay-as-you-use compute service – such as that offered by Amazon Web Services (AWS) – results in a more efficient use of compute resources and the ability to expand compute capability as and when needed. Rather than risking compute machines sitting idle between peak project simulation times, or long queues of jobs waiting to run, the capacity of the cloud can be drawn on when needed, and resources returned if unneeded.

# Flexibility

Use of Cloud Compute for HPC provides greater flexibility in terms of hardware upgrades compared to on-premises clusters. When a new generation of hardware is released in AWS, adding that hardware to an already-existing compute cluster can be as simple as a change of text on one line of a configuration file. This makes testing on, and permanently implementing new generations of hardware (Figure 1: below) very straightforward, as well as trying different hardware architecture.

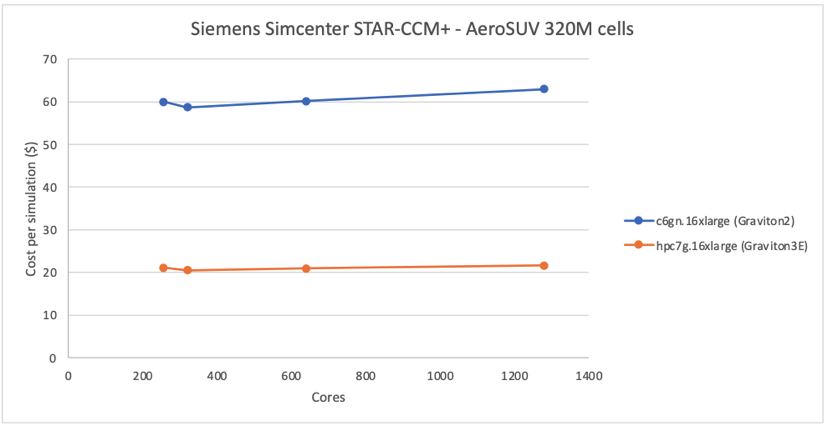


1. A graph of performance of Siemens Simcenter STAR-CCM+ on the AeroSUV 320M cell dataset. AWS’ Hpc7a outperforms Hpc6a up to 1.29x on a per core basis.

It is possible to mix and match hardware on one cluster such that there are instance types with many cores as well as instances with built in NVMe drives for fast data I/O. Different CPU architectures can also be mixed, as more simulation codes run on GPU and Arm based compute cores.

Simulation software that can scale well across many compute machines see cloud compute capacity bringing speed-up benefits. Considering only the cost of compute cores, a simulation that scales linearly can cost almost the same regardless of the number of cores that are used; yet as the core count is increased the job run time decreases (Figure 2: below). In essence, there are cases where simulations can be turned around quicker at minimal extra cost.

AWS are investing heavily in the latest technology – such as the Arm based Graviton processor - to improve the efficiency and sustainability of compute leading to a more sustainable simulation workflow. AWS is focused on efficiency and continuous innovation across our global infrastructure, with the aim of powering operations with 100% renewable energy by 2025.



1. Siemens Simcenter STAR-CCM+ – AeroSUV 320M – cost per simulation for Graviton3E is 2.8x better than for Graviton2