What a TCO analysis won’t tell you

Dig deeper to discover the true cost of your on-premises HPC investments

Executive summary

Organizations considering investments in high performance computing (HPC) should look deeply at the hidden costs of on-premises solutions. These factors, including lost productivity and missed innovation, can negatively affect other R&D investments that depend on HPC and thus lower revenues. Most organizations will find that cloud-based HPC solutions will deliver better ROI, even when a basic Total Cost of Ownership (TCO) analysis suggests continued on-premises investments. Furthermore, lower cost has not always been the primary accelerant for innovation when it comes to cloud adoption. Instead, the agility and flexibility offered by cloud-based HPC have helped organizations move forward unconstrained by the need to forecast future HPC workloads every few years.

INTRODUCTION

High performance computing (HPC) drives innovation and thus revenue in many modern industries—in fact, some could hardly exist without it. As the demand for computing resources increases to keep pace with business needs, we are starting to see more movement toward cloud-based HPC solutions, and factors related to cost drive most decisions. How can you get the most bang for your buck? Straightforward TCO analysis is bound to miss some key factors that are difficult to quantify but still have a profound effect on the bottom line. This paper will examine these hidden factors that should influence decisions about HPC investments:

- Lost productivity
- Missed innovation
- Technology refresh cost
- Increasing technical debt
- Longer time to results
- Risk management

It's important to remember that HPC systems are tools for innovation. Since modern innovation is an iterative process, faster iteration means faster production of meaningful discoveries. That means that the cost of all other R&D investments increases when they're backed up in a queue. Protecting total R&D investment over and above the cost of HPC (e.g. researcher and support staff salaries, wind tunnels, electron microscopes and other expensive equipment, etc.), which typically far outstrips the cost of HPC clusters, is vitally important. So even if the cost of cloud-based HPC exceeds the putative TCO of on-premises solutions, organizations often find that on-premises investments lead to substantial lost revenue thanks to lost output, slow innovation, and weaker products.

STANDARD TCO ANALYSIS OF ON-PREMISES HPC

Most enterprise-level HPC use on-premises solutions for some of their needs, and there are many good reasons for this. Specialized staff can offer dedicated support for well-understood HPC projects, and architecture can be tailored to spec for regularly repeated work, yielding long-term cost savings. For narrow tasks that are unlikely to evolve over time, there are fewer hidden costs to on-premises HPC.

And the ROI of HPC is unmistakable. IDC reports that for every $1 spent on HPC, businesses see $463 in incremental revenues and $44 in incremental profit (Hyperion Research, 2018). These eye-opening numbers make it clear that HPC is great for the bottom line—but they don't address how best to invest in these solutions.

> *"For every $1 spent on HPC, businesses see $463 in incremental revenues and $44 in incremental profit."*

For SMBs, the cost of acquiring on-premises HPC can be prohibitive, so the business case for cloud-based solutions is usually straightforward. But even large enterprises can find that their on-premises solutions aren't cost-effective at low-utilization. At the 2017 Internet2 Global Summit presentation "HPC in the Cloud," Omnibond CEO Boyd Wilson stated that the most likely scenario for a given on-premises HPC solution (85% server utilization and 50% power/cooling utilization) cost about $0.023 per core hour, nearly identical to the cost per core hour of a three-year reserved cloud HPC contract. But this doesn't include network or administrative labor and assumes a zero-cost building. That means on-premises HPC solutions can only be more cost-effective than cloud-based HPC when utilization rises far above 85%—a level which is very difficult to achieve in practice. And as we will see, there are many other factors that add to the cost of on-premises HPC.

HIDDEN COSTS OF ON-PREMISES HPC

On-premises HPC can't offer the flexibility of cloud-based solutions. For certain organizations—those with extremely reliable, unchanging HPC needs—this is irrelevant, but for most, that inflexibility brings with it a wide range of hidden costs. Furthermore, there are many other factors that are easily missed that should influence decision making, so even at consistently high utilization of on-premises resources, many enterprises will still find cloud-based HPC the most attractive option.

LOST PRODUCTIVITY

Underutilization can be a real problem with on-premises solutions—but what about overutilization? As utilization increases, engineers and researchers have to wait for their jobs to start, which leads to downtime and lost productivity. 72.8% of organizations using HPC reported some level of pent-up demand that was either delayed or cancelled due to lack of resources, and 29.2% reported that this pent-up demand exceeded 50% of their total yearly workload volume *(Hyperion, 2018)*.

72.8%

report delayed or cancelled HPC jobs

This means that many organizations are unable to complete about 38 % of their desired HPC jobs. Given the massive ROI referred to in the introduction, that's a lot of potential revenue that never comes through. Though it can be tricky to pin down exact numbers for lost productivity, organizations that expect usage spikes or otherwise unpredictable demand should look more closely at cloud-based solutions.

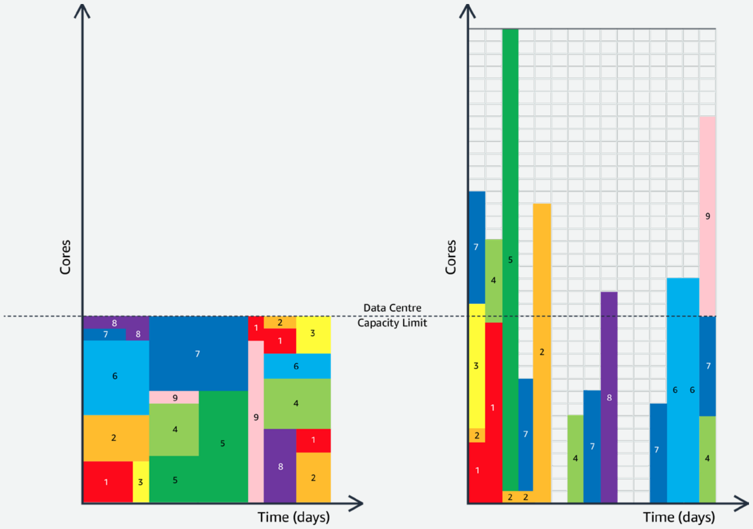

Figure 1: Fixed capacity vs. cloud. On the left is a traditional compute facility with limited capacity and thus limited availability. Rationing means most projects take much longer. The right-hand side shows a cloud-based solution, where scientists are free to work at an accelerated pace. In both cases, the same raw compute capacity is used, but the time to results for most projects is dramatically faster.

LOST INNOVATION

Lost productivity can also lead to lost innovation—though on-premises HPC can also limit innovation in other ways. There are many computing and storage options available in the cloud and many more in development, and it's simply not possible for any given on-premises system to offer every option. That means that even with unlimited access, engineers and researchers are constrained by the limits of their on-premises architecture. And, of course, on-premises systems age, whereas cloud-based systems are always leading-edge, so innovation can proceed at an accelerated pace aided by the latest technologies. The power of adapting HPC resources to evolving engineering and business needs is undeniable.

When innovation is lost, questions are left unasked, experiments are left undone, and potential revenue is left on the table. As with many of these hidden costs, it's difficult to estimate with any precision, but organizations that place high value on quick innovation or have changing needs must consider the advantages afforded by flexible, cloud-based HPC solutions.

Case Study: Celgene

Celgene is an American biopharmaceutical firm that manufactures drug therapies for cancer and inflammatory disorders. Using cloud-based HPC solutions, Celgene scientists have dramatically reduced the time it takes to complete HPC jobs needed for cancer drug research. "For our informatics researchers, computational jobs on AWS can be reduced to hours, compared to weeks or months on our on-premises HPC cluster," says Celgene's Associate Director of IT Lance Smith. As a result, researchers can run many more queries.

"By spinning up a few hundred nodes on AWS and getting results in less than a day, our scientific researchers have a lot more freedom to ask questions that weren't even possible before, questions they were afraid or unable to ask before because of hardware limitations or time constraints."

– Lance Smith, Associate Director of IT, Celgene

TECHNOLOGY REFRESH COST

On-premises HPC infrastructure is short-lived by design—they can only be patched or updated for a few years before they age into obsolescence. These migrations are expensive and risky compared with HPC cloud solutions, which are continuously refreshed. Storage must be refreshed every 4-5 years with a complete data migration, which usually also includes one or two quarters of costly overlap.

Cloud-based HPC investments eliminate the need for periodic technology and infrastructure refresh cycles. This helps keep HPC budgets predictable and generally less expensive than on-premises alternatives. Cloud-based HPC solutions keep up with hardware advancements every year, or even more often, compared to 3-5 years for most on-premises procurement. That means in many cases, on-premises solutions lag behind the cloud in access to the latest advances in computing and storage technologies, often leading to longer runtimes for newer, more complex algorithms and applications.

Newer cloud-native HPC applications are designed to perform better on cloud-based elastic infrastructure. In part, this is because each application gets its own optimized environment, so they are untethered from each other's technical restrictions. Since performance can be as or more important than cost, organizations should think through how the applications they want to use will fare on the various HPC infrastructure options they are considering. AI and Machine Learning applications are particularly well suited to flexible cloud-based architectures, but many other applications will also see better performance using elastic cloud resources.

INCREASING TECHNICAL DEBT

Technical debt is created when actions that seem attractive and efficient in the moment create costly issues over time. On-premises HPC infrastructure is a notorious source of technical debt. As mentioned previously, organizations with high utilization of on-premises HPC infrastructure often find fixed configuration, on-premises HPC clusters logical and attractive. But evolving business needs, algorithms, engineering practices, and research concepts are often overlooked. Retooling and adapting these newer algorithms and engineering applications to meet the requirements of an existing infrastructure (instead of the other way around) results in delays, and below-par performance.

The additional work involved in making sure applications work on existing infrastructure is the interest payment on technical debt. The longer an organization sticks with its existing infrastructure, the higher those payments get. It doesn't take long for the increasing debt and interest burden to significantly limit innovation and impact competitiveness. With cloud-based HPC infrastructure, organizations have easy and early access to the latest infrastructure technologies, so they can employ the latest applications and algorithms to innovate faster.

LONGER TIME TO RESULTS

On-premises HPC solutions offer a limited number of cores available at any one time, but many loosely coupled applications deliver results more quickly when more resources are available. This linear scaling has been used to great effect in cloud-based genomic analysis and risk modeling. Any applications that use large data sets will also see faster results using cloud solutions. For these applications and many others where quick results are highly desirable, cloud-based HPC is the best choice.

Faster results generally lead to greater business agility and quicker innovation, and can also help businesses make these innovations available more quickly to gain first-to-market advantage. The costs of time spent waiting for results is difficult to calculate but very important to consider.

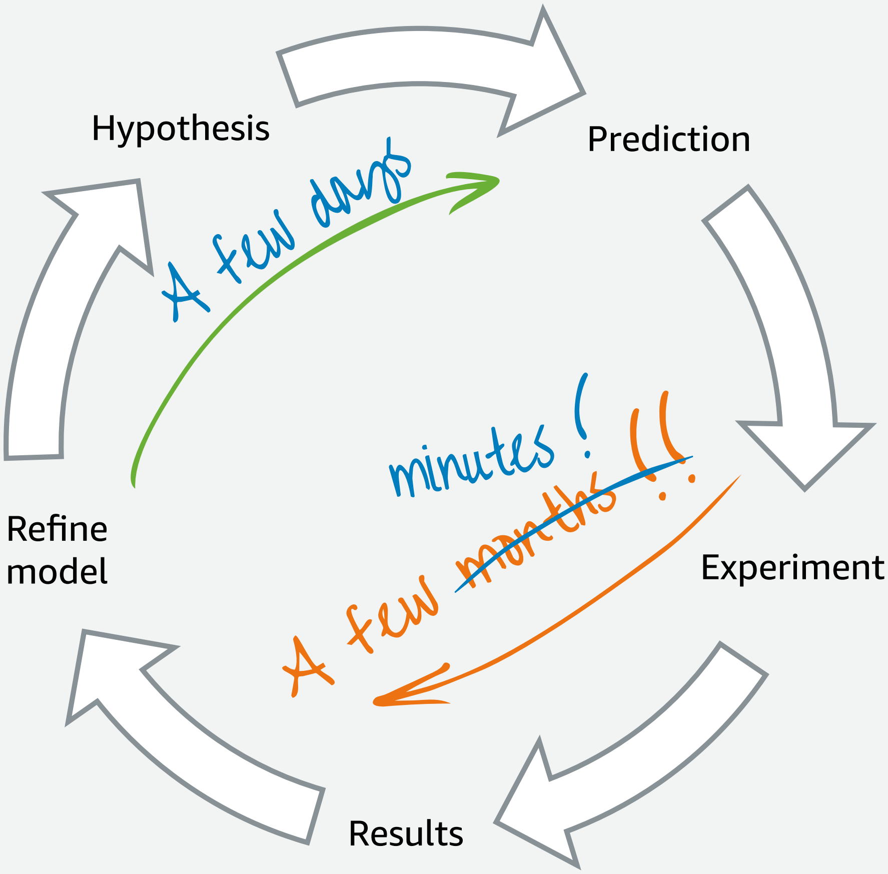

Figure 2: Innovation is an iterative process. Eliminating bottlenecks will speed time to innovation and measurably improve output.

Case Study: Bristol-Myers Squibb

Biopharmaceutical giant Bristol Myers Squibb found that with the capacity and computational power offered by AWS, they could run simulations 98% faster and engineer more efficient and less costly clinical trials and improve the patient experience. In a particular clinical trial, they were able to reduce the number of subjects from 60 to 40 and the number of blood samples from 12 per subject to 5 per subject. Running simulations 98% faster has led to more efficient and less costly clinical trials and better conditions for patients.

RISK MANAGEMENT

The security costs associated with on-premises HPC solutions increase over time as the infrastructure ages, just like any other on-premises computing resource. Cloud-based HPC features consistent processes and costs that get lower at scale. Similarly, regulatory compliance and expensive certifications are required for many on-premises solutions. Cloud-based HPC infrastructure can help in keeping up with constantly changing regulatory compliance needs, so you can avoid spending time and resources to ensure compliance. These costs can be significant for organizations working in partnership with governments, handle personally identifiable information (PII), or that have special security requirements, so they should investigate these costs in depth before reaching a decision.

THE TRUE TCO OF ON-PREMISES HPC

When comparing cloud-based and on-premises HPC solutions, direct comparison of obvious costs like hardware, software, and monthly fees is just the starting point. It’s very important to examine all hidden costs carefully and quantify them as accurately as possible to arrive at the closest approximation of the true TCO of these solutions and then make the most informed decision.

Every organization considering on-premises HPC investment should explore the factors mentioned above as well as all the unique qualities that affect their use of HPC resources, estimate the effects on TCO, and add them to the total. Given the range of options and the relative transparency of cloud-based HPC pricing, this broader analysis will support the business case for cloud-based HPC surprisingly often.

Some applications, like AI/Machine Learning and genomics, should be deployed in the cloud whenever possible to take advantage of scale and speed that just can’t be achieved by most reasonably priced on-premises HPC solutions. And given the 6-12-month procurement time for on-premises infrastructure (losing up to 20% of the useful life of the hardware), it usually makes sense to get started right away with cloud-based HPC solutions.

Case Study: Openeye

OpenEye, a provider of computational drug discovery software, sought to help its customers accelerate their research and cut costs, so it decided to move its software to the cloud. Now they use AWS to give their clients highly scalable, maintenance-free access to up to hundreds of thousands of processors to perform cloud-native computational chemistry for drug research and development. OpenEye Scientific is saving money by taking advantage of Amazon EC2 Spot Instances to use unused Amazon EC2 capacity at a discounted cost. “By using Amazon EC2 Spot Instances, we saved $800,000 last year,” says Craig Bruce, OpenEye’s Head of Infrastructure.

> *“Our customers benefit from that cost savings as well, because EC2 Spot Instances give them the ability to be more flexible. Whether they need to generate images in milliseconds or perform complex chemistry operations taking many hours, they now have the cost flexibility they need.”*
>
> **– Craig Bruce, Head of Infrastructure, Openeye**

FLEXIBLE PRICING AND BUSINESS MODELS

AWS HPC lets organizations bypass capacity planning worries, so most jobs can be done much more quickly. 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 can choose any combination of these pay-as-you-go options, paying only for the capacity they need, for the time that it’s needed. AWS Trusted Advisor alerts customers to any cost-saving actions that can minimize costs. This simplified, flexible pricing structure lets research institutions 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 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 modern industries—but evaluating all available options can be a complex undertaking. When comparing on-premises investment with cloud-based solutions, relying solely on a TCO analysis can be deceptive. By taking the time to consider all the hidden costs of on-premises solutions and reflecting on HPC's role in powering the longer value chain that is the innovation engine, organizations can arrive at a more accurate assessment about the returns they should expect on their HPC investments. With virtually unlimited capacity and the largest range of instance-types and services in the cloud industry, AWS offers comprehensive, cost-effective HPC solutions for everyone.

Digging a bit deeper into the real costs of fixed-capacity, aging, on-premises infrastructure will help organizations make the right choice. 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

Hyperion Research, "HPC ROI Research Update: Economic Models For Financial ROI And Innovation From HPC Investments," 2018, https://www.hpcuserforum.com/ROI/

Internet2, "Cloud vs. Datacenter Costs for High Performance Computing (HPC): A Real World Example," 2017, https://www.internet2.edu/blogs/detail/14114