Financial Case Study on the Use of Cloud Resources in HEP Computing

Shigeki Misawa[1,*], *Christopher* Hollowell[2], *Jerome* Lauret[1,**], *Tejas* Rao[2], and *Alexandr* Zaytsev[2]

[1]Scientific Data and Computing Center, Brookhaven National Laboratory, Upton, New York 11973
[2]Formerly with the SDCC, Brookhaven National Laboratory.

Abstract. An all-inclusive analysis of costs for on-premises and public cloud-based solutions to handle the bulk of HEP computing requirements shows that dedicated on-premises deployments of compute and storage resources are still the most cost-effective. Since the advent of public cloud services, the HEP community has engaged in multiple proofs of concept to study the technical viability of using cloud resources; however, the financial viability of using cloud resources for HEP computing and storage is of greater importance. We present the results of a study comparing the cost of providing computing resources in a public cloud and a comprehensive estimate for the cost of an on-premises solution for HEP computing. Like previous studies, the fundamental conclusion is that for the bulk of HEP computing needs, on premises is significantly more cost effective than public clouds.

1 Introduction

Since their inception, public clouds have revolutionized the deployment of computing services through a combination of financial (pay as you go) and technical innovations (infrastructure, platform and software as services). This has resulted in an accelerating migration of computing services from private (on premises) resources to infrastructure owned and operated by public cloud providers. The bulk of these services depends on the provisioning and deployment agility of resources in the public cloud. In contrast, "core" computing services, comprising the majority of resource consumption in enterprise data centers, have been slow to migrate to public clouds. The open question for all organizations, including those involved in high energy (HEP) and nuclear physics (NP) research, is whether or not it makes sense to move these core services to public clouds. An important consideration when evaluating such a move is its financial viability; migration for purely technical reasons is not sufficient.

2 Core Services

The primary mission of data centers supporting HEP and NP is to provide the computing and data storage resources required to extract science from theoretical and experimental investigations. High performance computing (HPC) systems with GPUs are typically favored for

*e-mail: misawa@bnl.gov
**e-mail: jeromel@bnl.gov

theoretical endeavors while experiments mostly rely on high throughput computing (HTC) systems without GPUs. These resources must be accompanied by data storage for home directories and working (or scratch) space where researchers can store programs and data being analyzed or generated. These four services (HPC, HTC, home directories and working storage) represent the bulk of on premises resources and costs borne by most HEP/NP data centers. Note that costs for near line "bulk" disk storage, with capacities in the 100 petabyte scale, and exabyte scale tape systems are not considered as few HEP/NP data centers possess these systems. Also calculating the cost for these services in the cloud is complicated due to the plethora of configuration options and the multitude of fees applied to them [1].

3 Comparing Costs

For this investigation a comparison was made between the cost of providing the four core services at the Scientific Data and Computing Center (SDCC) at Brookhaven National Laboratory (BNL) and at Amazon Web Services. Costs at other public cloud providers, e.g Google Cloud Platform and Microsoft Azure, were assessed to be similar to Amazon pricing. The on premises costs were taken from actual acquisition and operating costs of these services at the SDCC. Cloud costs were derived from the pricing information provided by Amazon Web Services for systems and services that most closely match what is provided by the SDCC [2]. As local policy prevents dissemination of absolute cost information, the comparison between on premises and public cloud is made through rounded cost ratios.

3.1 Baseline Assumptions

The cost calculation for both on-premise and cloud computing hinges on a myriad of environmental and institutional factors, extending beyond mere technical aspects such as resource capacity and performance prerequisites. The subsequent sections delve into the pertinent considerations concerning on-premise and public cloud deployments for this comparative analysis.

3.1.1 On Premises

Assumptions about on premises capabilities and operations are as follows. First, a modern, large scale, energy efficient data center building already exists on site. The building provides multiple megawatts of UPS protected, generator backed power for IT loads and space for ~100s of standard 19-inch racks. Second, the cost of building it is not factored into on premises costs as scientific programs typically do not bear those costs and it already exists, so there is no need to build it. Third, it is assumed that the organization has sufficient critical mass to support services on premises cost effectively and economies of scale exist to keep IT equipment and infrastructure support costs small relative to hardware costs.

In calculating costs, 100% resource utilization is assumed for compute resources. All costs are based on purchases at BNL and include overhead, power, and cooling. The cost of equipment includes installation services for compute and 5 years of vendor provided support (i.e., hardware warranty and firmware support). These costs may be different for other organizations. On premises labor cost for hardware support of compute resources are small relative to the total life cycle cost for these resources, a byproduct of the economies of scale. For this reason, they are not included in the on premises calculation. However, for data storage labor is a non trivial fraction of total cost and is therefore included in the on premises calculation.

3.1.2 Cloud

The assumptions made when calculating costs in the cloud revolve around a different set of issues. First, workloads currently running on site are move "as is" to the cloud, i.e. there is no re-engineering of workloads to reduce cloud costs. Second, the cost of migrating operations from on premises to public cloud aren't taken into consideration. Third, cloud pricing for single availability zone deployment of resources are used, matching the characteristics of on premises assets. Fourth, non-preemptable cloud pricing is used, as on premises resources provide guaranteed access to resources and "base load" demand must be satisfied. Fifth, pricing is assumed to be stable over a 5 year period and cost include local (BNL) procurement overhead. Sixth, cloud services that most closely match the capabilities and hardware resources of on premises services were chosen for comparison. Finally, network egress fees, i.e. per GB charges on data transfers out of the cloud are not included [3, 4]. As they only increase cloud costs and complicate the calculation, their omission doesn't materially affect the conclusions.

4 Costs

4.1 High Performance Computing

For HPC resources, a comparison was made between the SDCC's 2nd generation institutional cluster (IC) and a cluster of similar equipment at Amazon. The SDCC IC is an accelerated HPC cluster composed of dual socket compute nodes with 200 Gbps Infiniband (IB) cluster interconnect. Each node is outfitted with four Nvidia A100 GPUs, 4 TB of SSD storage and 1 TiB of memory. The Amazon EC2 instance type that is the closest match to the IC node, based on GPU count, is the p4de.24xlarge instance [5]. Each p4de.24xlarge node is roughly equivalent to two SDCC IC nodes. The configurations of the SDCC IC node and the Amazon EC2 node are summarized in Table 1.

HPC Node Comparison		
	On Premises	Amazon [5]
Node Equivalence	2 nodes	1 node
Node Type	bare metal	p4de.24xlarge
Cores	48[a]	96[b]
CPU	2 × Xeon 6336Y	Xeon P-8275CL
GPU	4 × Nvidia A100	8 × Nvidia A100
Memory	1 TiB	1 TiB
Storage	1 × 4 TB SSD	8 × 1 TB NVMe
Network	200 Gbps IB	400 Gbps ENA+EFA

[a] Physical cores
[b] Virtual cores

Table 1. HPC configuration, on premises vs public cloud. Two SDCC institutional cluster nodes are required to match the GPU performance of one Amazon p4de.24xlarge EC2 node.

SDCC IC cluster estimates assume a 5 year life and includes power and cooling costs in addition to the vendor support we noted earlier. As mentioned previously, local hardware support labor costs aren't included. Amazon EC2 costs, obtained from Amazon AWS pricing web pages, also assume a 5 year life and are based on the yearly rate for 3 year reserve

instances for 5 years. As mentioned previously, BNL overhead is included in the cost for both implementations and on the BNL side includes power and cooling costs. Under these conditions, as summarized in Table 2, the Amazon based implementation of a GPU accelerated HPC cluster ends up being about four times (4×) the cost of the on premises system [3]. If the cluster is deployed with spot instances cloud costs are still two and one half times (> 2.5×) on premise [6]. In addition, this assumes that the eviction of running jobs incurs no cost, including operational cost incurred by the end user.

HPC Cost Comparison		
	On Premises	Amazon
Relative Cost	1	≈ 4 (reserve)
Relative Cost	1	> 2.5 (spot pricing) [a]

[a] Amazon spot pricing ignores cost of eviction

Table 2. HPC cost comparison, on premises vs public cloud, where costs are normalized to the on premises cost.

4.2 High Throughput Computing

For HTC resources, a comparison was made between the SDCC HTC Linux farm and a similar configuration at Amazon. The newest nodes in the SDCC Linux farm are dual socket compute nodes with 10 Gbps Ethernet network connectivity. Each node is outfitted with 8 TB of SSD storage and 348 GiB of memory. The Amazon EC2 instance type that is the closest match to the SDCC farm node is the m6id.24xlarge instance [7]. Based on the HEP-SPEC06 benchmark each m6id.24xlarge node is approximately 20% faster than a SDCC farm node. The configurations of the SDCC farm node and the matching Amazon EC2 node are enumerated in Table 3.

HTC Node Comparison		
	On Premises	Amazon [7]
Node Equivalence	≈ 1.2 nodes	1 node
Node Type	bare metal	m6id.24xlarge
Cores	96[a]	96[a]
CPU	2 × Xeon 6336Y	Xeon 8375C
Memory	384 GiB	384 GiB
Storage	4 × 2 TB SSD	4 × 1.425 TB NVMe
Network	10 Gbps Ethernet	37.5 Gbps

[a] Virtual (logical) cores

Table 3. HTC node configuration - On premises vs public cloud. Note one Amazon m6id.xlarge EC2 node is approximately 20% faster than one SDCC HTC node based on the HEP-SPEC06 benchmark [8, 9].

Calculations of the cost of the SDCC and Amazon instances of the HTC farm follows the basic recipe previously outlined for HPC resources. A 5 year life span is assumed, with BNL costs adjusted for the approximately 20% higher performance of Amazon hardware and includes power, and cooling. As with HPC, the yearly rate for 3 year reserve instances

for a 5 year period is used for Amazon EC2 estimates. BNL overhead is included in the cost for both implementations. Under these conditions, as shown in Table 4, the Amazon implementation of an HTC farm ends up being approximately six times (6×) the cost of the SDCC system [3]. Moving the cluster to spot instances only reduces cloud costs to more than five times (> 5×) that of the on premises system, again assuming job eviction has no cost [6]. However, unlike HPC compute jobs, HTC jobs at the SDCC typically do not checkpoint. This means that eviction is likely to be costlier for HTC jobs compared to HPC jobs.

HTC Cost Comparison

	On Premises	Amazon
Relative Cost	1	≈ 6 (reserve)
Relative Cost	1	> 5 (spot pricing)[a]

[a] Amazon spot pricing ignores cost of eviction

Table 4. HTC cost comparison, on premises vs public cloud, where costs are normalized to the on premises cost.

4.3 Home Directories

Although a very small component of the SDCC infrastructure, both in cost and equipment, home directory service is a critical service. Without home directories, both the HPC and HTC resources would be unusable. Home directory access is characterized by small block and high concurrency accesses, as a result storage systems used for home directory service must be optimized for random access. Low write latency and good metadata performance (metadata operations per second and low latency metadata access) are also important.

At the SDCC, home directories are provided by a flash based NAS appliance with data compression enabled. A relatively stable compression ratio of 12:1 has been observed over the years for SDCC home directories. The comparable Amazon storage services is Amazon FSx NetApp ONTAP. In addition to data compression, ONTAP provides storage tiering that can substantially reduce the overall cost in the cloud. For the purposes of the comparison, it is assumed that 20% of the data is active. This percentage is quoted by Amazon as being derived from "industry research and customer analysis" on their "Amazon's FSx for NetApp ONTAP" pricing page [4]. Table 5 summarizes the characteristics and relative cost of home directory service on premises and at Amazon.

The calculation of storage costs is identical to the procedure for compute, except that local labor costs are included for on premises storage. For cloud storage, the cost savings gained through the use of tiered storage are also included. Under these conditions, the cost of on premises home directory services is more than two times (2×) the cost in the cloud. Although cloud is significantly less expensive than on premises, the absolute size of the savings is small relative to the cost of compute services and working file system storage. As a result, it has little impact on the total cost of core services, either in the cloud or on premises.

It should be noted that the lower cost of cloud storage for home directories is primarily due to storage tiering, i.e. the ability to move inactive data from high cost (SSD) to lower cost (HDD) storage. The difference disappears if tiering isn't effective. Tiering could be added to the on premises system, but lack of scale on premises significantly reduces any potential cost savings that might be achieved.

	Home Directory Comparison	
	On Premises	Amazon
Primary Media	SSD	SSD
Secondary Media	N/A	HDD
Compression	Yes	Yes
Compression Ratio	12:1	12:1
Storage Tiering	No	Yes
Active Data Percentage	100%	20%
Relative Cost	> 2	1

Table 5. Home Directory Service. Storage system configuration and cost comparison with costs normalized to the cost in the cloud.

4.4 Working File Systems

The final service that is examined is working storage for data generated or consumed by users. Working storage is characterized by high capacity (multiple petabytes) and high bandwidth (GBs/sec) and is optimized for large block I/O. At the SDCC this services is provided by a hard disk (HDD) based Lustre [10] file system. The Amazon equivalent is the lowest performance tier (12MB/s/TiB) HDD based Amazon FSx for Lustre configuration [11].

In contrast to home directories, no data compression is assumed for working storage as the bulk of experiment data is precompressed. With this difference compared to home directories, working storage in the cloud is more than six times (6×) the price of on premises. Table 6 summarizes the characteristics and relative cost of working file system service on premises and at Amazon.

	Working Storage Comparison	
	On Premises	Amazon
Primary Media	HDD	HDD
Compression	No	No
Relative Cost	1	> 6

Table 6. Lustre based working storage service. Storage system configuration and cost comparison with costs normalized to the on premises cost.

5 Other Considerations

In assessing the total cost of hosting services, either on-premises or in a public cloud, it is imperative to contemplate factors extending beyond merely sustaining existing capabilities. This encompasses accounting for the cost associated with cloud migration, particularly if resources are currently provisioned on-premises. Moreover, ensuring service portability to avert cloud provider lock-in is crucial whether this entails the capability to seamlessly transition to alternative cloud providers or to revert services back on-premises. Service portability might pose a challenge to uphold since public clouds furnish easy access to a vast array of

services, many of which are proprietary and devised to entice users. Additionally, it's vital to be cognizant of potential unforeseen expenses related to data access fees within cloud services, as they could significantly influence the overall cost analysis.

Network egress fees, i.e. per gigabyte charges on data transferred out of the cloud, is one example of an access fee. Amazon FSx for NetApp ONTAP charges for SSD IOPS beyond a default 3 IOPS per GB of SSD storage, as well as access bandwidth and number of reads and writes to data in the capacity tier [4] are other examples. Complex fee schedules are also problematic, which is particularly evident with Amazon S3 and S3 Glacier, potential cloud alternatives to on premises near-line "bulk" disk storage and tape, that aren't considered in this analysis.

Lastly, an aspect not considered in this analysis is the potential course of action beyond the 5 year lifecycle of the equipment. In the case of on-premises infrastructure, the equipment can continue to operate beyond this period at roughly the same operational cost, given that the acquisition cost has been amortized over the initial five years. However, this comes at the expense of occupying additional space and consuming more power compared to newer equipment, alongside a potentially higher rate of hardware failures.

6 Conclusion

For organizations boasting substantial economies of scale and possessing an updated, modern data center, this analysis illustrates that hosting the core services necessitated by Nuclear and High Energy Physics is considerably more costly in a public cloud compared to on-premises. Nevertheless, juxtaposing the costs for services on-premises and in a public cloud is a complex undertaking. It's heavily contingent on the nature of services being transitioned, the methodology employed for the transition, and the distinctive circumstances of the organization orchestrating the move. Lastly, the multifaceted nature of cloud pricing, characterized by a plethora of options and a wide spectrum of fees for services, renders accurate cost estimation heavily reliant on a thorough comprehension of operational requisites.

References

[1] Amazon Web Services, *AWS S3 Pricing*, https://aws.amazon.com/s3/pricing/ (2023)

[2] Amazon Web Services, *Amazon Pricing*, https://aws.amazon.com/pricing/ (2023)

[3] Amazon Web Services, *Amazon EC2 On-Demand Pricing*, https://aws.amazon.com/ec2/pricing/on-demand/ (2023)

[4] Amazon Web Services, *Amazon FSx for NetApp ONTAP Pricing*, https://aws.amazon.com/fsx/netapp-ontap/pricing/ (2023)

[5] Amazon Web Services, *Amazon EC2 P4 Instances*, https://aws.amazon.com/ec2/instance-types/p4/ (2023)

[6] Amazon Web Services, *Amazon EC2 Spot Instances Pricing*, https://aws.amazon.com/ec2/spot/pricing/https://aws.amazon.com/ec2/spot/pricing/ (2023)

[7] Amazon Web Services, *Amazon EC2 M6i Instances*, https://aws.amazon.com/ec2/instance-types/m6i/ (2023)

[8] M. Michelotto, M. Alef, A. Iribarren, H. Meinhard, P. Wegner, M. Bly, G. Benelli, F. Brasolin, H. Degaudenzi, A.D. Salvo, Journal of Physics:Conference Series **219**, DOI 10.1088/1742-6596/219/5/052009 (2010)

[9] HEPIX Benchmarking Working Group, *HEP-SPEC06 (HS06)*, https://w3.hepix.org/benchmarking/HS06.html (2009)

[10] Lustre.org, *Lustre*, https://lustre.org/ (2023)
[11] Amazon Web Services, *Amazon FSx for Lustre Pricing*, https://aws.amazon.com/fsx/lustre/pricing/ (2023)