CYBER-INSIGHT: Evaluating Cyberinfrastructure Total Cost of Ownership

PAUL FISCHER, WIM CARDOEN, DAVID RICHARDSON, CHONGHUAN XIA, DAO WHITE, JULIA HARRISON, and THOMAS E. CHEATHAM III, University of Utah Center for High Performance Computing 155 S 1452 E, RM. 405, USA

DANIEL A. REED, University of Utah 201 Presidents Circle, RM. 205, USA

The proliferation of commercial cloud computing platforms presents a potential financial cost-saving opportunity for research computing workflows which have traditionally relied on highly-capitalized on-premises datacenters and local support staff. Rigorously comparing the relative costs — i. e. total cost of ownership (TCO) — of commercial cloud vendors and self-managed, on-premises cyberinfrastructure requires an in-depth understanding of each platform. We present CYBER-INSIGHT, a web-based calculator to evaluate and compare the TCO of different computing platforms. CYBER-INSIGHT leverages a flexible data model that research computing experts may customize to their own needs and actual costs.

CCS Concepts: • **Hardware → Yield and cost modeling**; • **Computer systems organization → Cloud computing**; • **Software and its engineering → Cloud computing**; *Software performance*; *Software creation and management*.

Additional Key Words and Phrases: total cost of ownership, cyberinfrastructure, cloud computing, high-performance computing, research computing, performance, financial cost analysis

ACM Reference Format:
Paul Fischer, Wim Cardoen, David Richardson, Chonghuan Xia, Dao White, Julia Harrison, Thomas E. Cheatham III, and Daniel A. Reed. 2021. CYBER-INSIGHT: Evaluating Cyberinfrastructure Total Cost of Ownership. In *Proceedings of PEARC21: Practice & Experience in Advanced Research Computing (PEARC) Conference Series (ACM PEARC21).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Traditional research computing leverages on-premises high-performance computing (HPC) and data resources. On-premises datacenters are generally funded by a mix of institutional budget, grants, and pooled resources from participating research groups. Depending on the scale of the institution, annual expenditures on datacenter assets — servers, GPUs, interconnects, storage, power, cooling, racks, etc. — may range from tens of thousands to hundreds of millions of US dollars.

Commercial cloud providers, by contrast, allow customers to design and rent "virtual datacenters" that run on physical datacenters managed by these providers. Customers are billed for the provided services. Compared to large capital investments and the support staff's wages for on-premises HPC, the costs plus fees associated with utilizing such a datacenter are marginal. Additionally, cloud providers may offer access to specialized computational resources — application-specific integrated circuits (ASICs), tensor processing units (TPUs), quantum processing units (QPUs), etc. — that are not readily available to small and mid-size HPC centers.

Determining the parameters for which research computing on premises versus commercial cloud resources is more cost-effective (assuming performance parity) in running particular research computing workloads requires an in-depth study of the respective total costs of ownership (TCOs). To address this, we designed and developed CYBER-INSIGHT, a web-based TCO calculator for generic computing resources and workloads.

A version of this paper with additional figures available at: https://cyberinsight.chpc.utah.edu/#papers.

1.1 Limitations of the existing TCO calculators

Each major commercial cloud provider[1, 2, 4, 5] or broker[3, 6] offers his own version of a TCO calculator which also functions as a marketing tool to attract customers. Among these calculators none provided a satisfying framework to compare the respective TCOs for running research computing workloads on various computing platforms. Additionally, these calculators use rigid schemas to perform cost analysis: e. g. , they do not allow sufficient flexibility to encompass the often heterogeneous nature of research computing and on-premises HPC and data resources.

1.2 Concerns regarding cloud computing

Customers of cloud computing expect a similar level of computational performance as on similarly-configured on-premises datacenter resources. Thus, if the relative performance of a workload run on-premises exceeds its performance on a cloud platform by an order of magnitude or more, then the merit of comparing the platforms quantitatively on final TCO loses merit.

Diagnosing low-level performance problems on commercial cloud computing platforms is often hindered by limited customer access to the underlying infrastructure (e. g. , the high-performance interconnect (HPI)). In order to mitigate some of these issues, some cloud providers offer guidance which pertains only to their platforms.

Research computing is also highly heterogeneous with greatly varying demands on software, data protection, and general scalability and may be time-sensitive (e. g. , deadlines associated with grants).

2 SOFTWARE DESIGN

Our goal was to create a portable TCO calculation tool that allows administrators to collaborate and compare their TCO analyses. To this end, we chose to build a web application with a graphical user interface (GUI).

2.1 CYBER-INSIGHT's software components

The funding for the CYBER-INSIGHT project was contingent on building the tool as a Jupyter Notebook, which supports multiple programming languages. We implemented a Python library to define our TCO models and algorithms as well as a GUI library for Jupyter Notebook which depends on the `ipywidgets` package. The tool is self-contained and includes features such as data import/export, data modification, and data visualization, all reflected in user-friendly GUI widgets. Users familiar with Python can easily modify the underlying TCO model to fulfill their own needs.

2.2 Data Model

CYBER-INSIGHT's primary data model is called a `Resource`, which represents a static configuration for a computational or storage resource. The `Resource` model includes the following three sub-models:

- The `Attributes` list. It includes high-level fields for the resource, such as node count, core count, cloud-hosting fee (if applicable), average utilization (e.g. a system which is idle 5% of its time has an average utilization of 95%). The attributes list is static.
- The `Benchmarks` list. A resource may include any number of generic benchmarks, which follow a simplified schema. In the case of computational resources, benchmarks require the wall time, node count, and power consumption (if applicable) of a particular workload run.
- The `Cost` tree. It is a flexible/dynamic hierarchy which is representative of the resource's TCO over its lifetime. CYBER-INSIGHT users can design their own cost hierarchy to accomodate their financial reporting and resource configuration. The root node of the cost tree is a formula that partitions the cost into common domains such as hardware, datacenter infrastructure, etc.

2.3 TCO evaluation

The TCO of a benchmark b on a compute resource r is calculated as follows:

$$\mathbf{cost}_{\mathrm{b}} = \frac{\mathbf{duration}_{\mathrm{b}}}{\mathbf{lifetime}_{\mathrm{r}} * \mathbf{efficiency}_{\mathrm{r}}} * \mathbf{cost}_{\mathrm{r}} + (\mathbf{powerUsed}_{\mathrm{b}} * \mathbf{powerCost}_{\mathrm{r}}) + (\mathbf{duration}_{\mathrm{b}} * \mathbf{hostingCost}_{\mathrm{r}})$$

This model enables the comparison of workloads across both cloud and on-premises resources. Usually, it will only be useful to compare workloads that use the same application with similar parameters. However, it is possible that centers may want to compare different workloads across computing platforms due to qualitative differences or limitations in either platform.

3 RESULTS

The CYBER-INSIGHT application is located at https://cyberinsight.chpc.utah.edu/. The current version of the tool includes five modules: import, create & edit, analyze, export, and a Python sandbox.

We haved used CYBER-INSIGHT to perform a TCO analysis on data from the University of Utah's Center for High Performace Computing (CHPC) on-premises datacenter, as well as on data harvested from the AWS, Azure, GCP, and Rescale cloud platforms. Our limited testing of research computing workloads showed that on-premises resources outperformed commercial cloud resources in cost effectiveness. Appendix A details the methods and challenges of gathering this data in a real HPC environment.

4 CONCLUSION

The tool is functional for running TCO analyses. For CHPC, on-premises HPC will remain relevant for research computing for simple workloads. We are working to make this tool relevant to more complex scenarios.

Research computing workflows may be "episodic" and involve periods of significant utilization of network and storage resources in addition to computation. We plan to extend the resource model to support the aforementioned resources, and develop a framework for combining benchmarks from different resources into single episodic workflows. For example, a workflow might involve a large ingress of data, some parallel computation, and then an egress of data. Each phase of this workflow incurs costs, especially in a cloud environment. An accurate representation of these costs is required to have a better understanding of the TCO across platforms.

We will plan to add additional technologies to support collaborative usage of CYBER-INSIGHT between institutions. We plan to host a central database where timestamped TCO results can be recorded and compared. Due to technological advances and thus changing market conditions, it is also of interest to record the TCOs for the different compute platforms as a function of time. Finally, we are improving our tutorial for CYBER-INSIGHT.

ACKNOWLEDGMENTS

We are grateful for the original University of Iowa team (Steve Fleagle, Kang-Pyo Lee, Ben Rogers, and Spencer Kuhl) who developed a preliminary prototype of the CYBER-INSIGHT tool.

This material is based upon work supported by the National Science Foundation under grant no. OAC-1812786.

REFERENCES

[1] Amazon Web Services. 2021. AWS Pricing Calculator. https://calculator.aws/

[2] Google Cloud. 2021. Google Cloud Platform Pricing Calculator. https://cloud.google.com/products/calculator

[3] HashiCorp. 2021. Cost Estimation. https://www.terraform.io/docs/cloud/cost-estimation/index.html

[4] Linode. 2021. Total Cost of Ownership (TCO) Cloud Calculator. https://www.linode.com/lp/tco-calculator/

[5] Microsoft Azure. 2021. Total Cost of Ownership (TCO) Calculator. https://azure.microsoft.com/en-us/pricing/tco/calculator/

[6] Numerical Algorithms Group. 2020. HPC Total Cost of Ownership (TCO) Calculator. https://www.nag.com/content/hpc-total-cost-ownership-tco-calculator-0

A CYBER-INSIGHT USAGE EXAMPLE

A.1 On-premises data collection methods

We determined the full-cost of on-premise computation by initially determining all costs incurred by our center. First, we determined the costs that could be directly assigned to computation, excluding the hardware costs. These costs include staff time, and any supplies that were directly purchased in support of HPC.

Our overhead falls into four categories: general overhead, networking overhead, user services overhead and datacenter overhead. General overhead was allocated based on the percentage of direct HPC costs to total direct costs across all of our cost centers. Consulting costs were then applied to the costs centers based on the total % of HPC (direct costs plus general overhead costs). Subsequently, we allocated networking to all costs centers.

Finally, we allocated all the datacenter costs (excluding power) based on the percentage of rack space to the total taken by HPC. The power consumed by a job, and the cost of hardware actually used in the jobs are then added using the tool.

A.2 Cloud data collection methods

As a medium-sized traditional academic HPC center we had very little in-house expertise in the use of commercial cloud environments. A large portion of the project was spent on developing this expertise and interacting with representatives of cloud providers. We also noticed that commercial cloud vendors had in general limited support for HPC research computing. We presume that this gap will shrink over time as both sides cooperate more closely.

While there are many commonalities between the major cloud platforms, there are notable differences. For example, there is no uniform approach to extract the financial cost information on a per-workload basis. Therefore, we extracted the appropriate costs from monthly cloud bills. For details on our approaches to each of the major cloud vendors, please contact the authors.