Special Analysis

A Thought Experiment on Accelerated HPC Cloud Growth

Alex Norton and Mark Nossokoff
January 2021

HYPERION RESEARCH OPINION

HPC in the cloud has continued on a strong upward trajectory over the last few years, fueled by a concerted effort from CSPs to address the technical capabilities needed to better run HPC jobs in cloud environments. Many CSPs have bolstered their HPC strengths through hiring HPC experts to speak with HPC customers on application-specific issues, such as porting and running hard HPC jobs in the cloud. Further, CSPs are continuously increasing the platforms offered for HPC with the addition of offerings such as high performance processors, access to bare metal, a variety of accelerator options, high-performance interconnects, multiple storage offerings, as well as software packages valuable to HPC customers. These steps have resulted in dramatic advances in cloud adoption over the past two years, including a major growth year in 2019.

Hyperion Research HPC cloud forecasts draw on years of user surveys and interviews to create an end-user spend perspective of running HPC workloads in the cloud. The previous forecasts anticipated a strong growth curve, but in 2019 the market grew beyond expectations in a stepwise fashion. 2019 was the first year the market showed a major growth bump, driven by a wide, albeit shallow, penetration of cloud computing. What the research showed was a move from HPC users running experiments with HPC applications in the cloud to running full workloads. Running new experiments to better understand what can and cannot run well in clouds also increased in 2019.

The economic struggles that resulted from the worldwide pandemic in 2020 have also been incorporated into the updated forecast. Elastic compute resources became more appealing to many buyers with the need to increase and decrease such resources on a moment's notice due to wild demand fluctuations. With this in mind, there is likely to be another major growth year, like 2019, within the next few years.

The timing of this next tipping point is difficult to predict. Given the changes in HPC users' behaviors over the course of the pandemic in the way they evaluated the cloud and adjusted their compute resources accordingly, there is strong anticipation for another major growth year within the next four years. This report will present a potential future tipping point scenario and will illuminate some of the drivers that could push the growth of HPC in the cloud even higher than the current forecast.

THE CURRENT 2020 HPC CLOUD FORECAST

Hyperion Research recently released the updated 2020 HPC cloud forecast including the impacts of covid-19, projecting HPC cloud spend out to 2024. The forecast is from the perspective of end-user

spending on cloud resources for HPC applications. For more information, please see the document titled: *2020 HPC CLOUD FORECAST*.

The updated cloud forecast projects user spending on external clouds for HPC workloads to reach almost $9 billion in 2024, in addition to the on-prem revenue traditionally tracked by Hyperion Research. The forecast is shown below in Table 1, highlighting a 17.6% CAGR from 2020-2024.

TABLE 1

HPC Cloud Forecast 2018-2024 ($M)

	2018	2019	2020	2021	2022	2023	2024	Five Year Forecast CAGR
NEW 2020 HPC Cloud Forecast	$2,466	$3,910	$4,300	$5,300	$6,400	$7,600	$8,800	17.6%
2019 HPC Cloud Forecast	$2,466	$3,910	$4,262	$5,135	$6,182	$7,418	-	24.6%

Note: CAGR for 2019 Forecast is for the time period 2018-2023

Source: Hyperion Research, 2020

Cloud growth has been driven by a few major factors over the past few years. The most common reason for using cloud for HPC has been the ability to address bursts in workloads and expand HPC resources to accommodate those surges. Further, the desire to run on hardware or software not currently available in an on-prem data center has pushed more users to the cloud, especially as AI adoption increased across the entire HPC landscape.

Finally, the cloud has become more cost effective for some HPC workloads, especially those that are embarrassingly parallel or those that utilize data sets already stored in the cloud, when compared with on-prem computing. This concept is two-fold. The cost-effectiveness improvements have been driven, in part, by CSPs' altering their pricing strategies to address this historical barrier to cloud adoption. In addition, users are beginning to reevaluate cost to include aspects of running HPC workloads typically not found on just the CSP bill. These included components such as:

- The ability to run at the proper scale for a workload instead of running on a smaller allocation of a system due to many users and high utilization
- The ability to run an application now versus waiting in a long queue time (and maximizing researchers' time and efforts)
- The ability to run on the most effective hardware and software platform rather than what is available on-prem
- The ability to run AI workloads where the data is stored or collected rather than moving large amounts of data

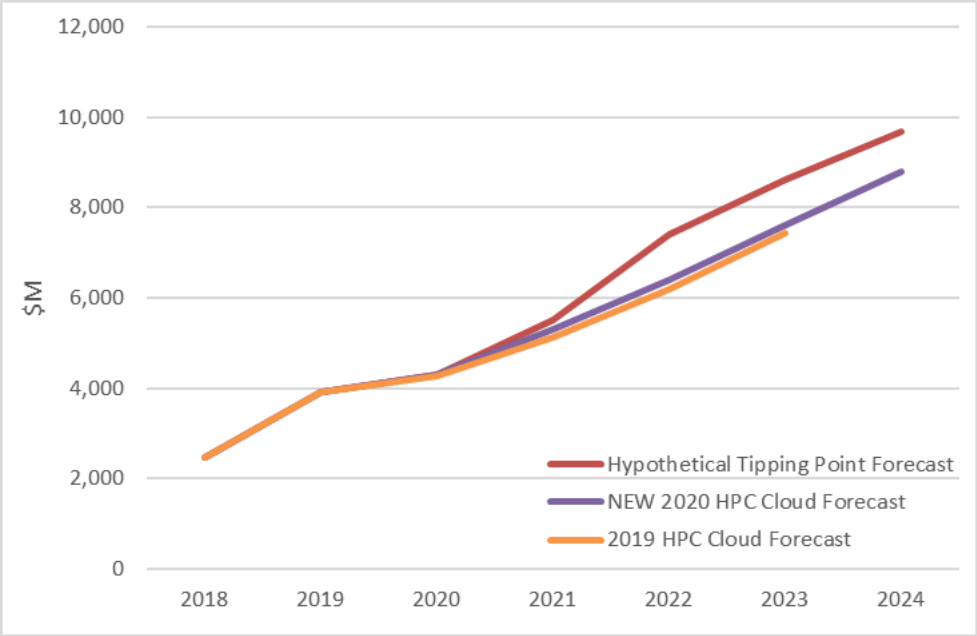
A HYPOTHETICAL CLOUD TIPPING POINT FORECAST

As mentioned earlier, Hyperion Research expects another major growth year to occur during the forecast period. Although it is very challenging, and in many cases impossible, to determine exactly when that unexpected growth may occur, it is possible to offer a hypothetical scenario. Below is a graphic that incorporates a theoretical uptick in the forecast in the 2021-2022 timeframe.

(*Note: This hypothetical forecast is not the official Hyperion Research 2020 HPC Cloud forecast and should not be treated as anything more than an exercise.)*

FIGURE 1

HPC Cloud Forecast 2018-2024 ($M) Including *Hypothetical* Tipping Point



Source: Hyperion Research, 2020

As can be seen in Figure 1, the theoretical growth in the 2021 to 2022 timeframe could send cloud spending on a more aggressive upward trajectory reaching almost $10 billion by 2024, $1 billion higher than the current 2020 Hyperion Research HPC cloud forecast. Below, in Table 1, is the supporting data table for Figure 1.

To reiterate, the *hypothetical* cloud forecast is not to be interpreted as a new forecast or an official forecast. It is merely created for the purposes of this thought experiment.

TABLE 2

HPC Cloud Forecast 2018-2024 ($M) INCLUDING Hypothetical Major Growth

	2018	2019	2020	2021	2022	2023	2024	Five-Year Forecast CAGR
2020 HPC Cloud Forecast with **Hypothetical** Tipping Point	$2,466	$3,910	$4,300	$5,500	$7,400	$8,612	$9,672	19.9%
2020 HPC Cloud Forecast	$2,466	$3,910	$4,300	$5,300	$6,400	$7,600	$8,800	17.6%
2019 HPC Cloud Forecast	$2,466	$3,910	$4,262	$5,135	$6,182	$7,418	-	24.6%

Note: CAGR for 2020 forecasts is of forecast time period: 2019-2024. CAGR for 2019 forecast is of forecast time period: 2018-2023.

Source: Hyperion Research, 2020

DRIVING FORCES FOR STRONGER CLOUD GROWTH

There are a number of factors that could lead to the theoretical forecast above, driven by technological advancements from both users and CSPs, as well as the need to address new workloads. None of these drivers alone would result in the major growth anticipated, but a combination of two or three, or the addition of some drivers not listed here, could push more users to increase their cloud utilization and draw non-cloud users to take advantage of the cloud.

Dramatic Additional Improvements in Running Hard HPC Jobs

CSPs have made focused efforts to handle the harder HPC jobs, which are applications that require high inter-processor communication, sometimes referred to as tightly coupled, but also applications that require extremely high bandwidth and low latency. These applications are typically run on-prem, given that the on-prem data centers have been tuned and optimized for these applications. The less demanding HPC jobs (i.e., those that parallelize easily) have already been running in the cloud effectively for some time. Most on-prem environments have highly optimized their hardware, system software, applications, storage, networking, etc. in order to have higher performance. The more complex HPC jobs typically involve heavy inter-processor communication and can struggle in virtualized environments, hence why they are usually run in on-prem data centers that have been configured specifically for that job. CSPs have responded to this situation by working to reduce their virtualization penalties through various hardware and software efforts, as well as increase bare-metal offerings to allow for harder HPC jobs to be run.

Even with these improvements, many hard HPC jobs run more efficiently and effectively on-prem. If the pattern of improvement on inter-processor and inter-node communication-heavy workloads and the virtualization penalty continues to diminish, that could push more users to start running some of their harder HPC jobs in the cloud.

Sizeable Improvement in Ease of Use and Deployment

One key attribute of using the cloud, other than performance, price and security, that is often overlooked is the ease of deployment. Many users still do not fully understand how to effectively run jobs in the cloud, specifically how to alter their own code to run in a highly performant manner. A common theme Hyperion Research has uncovered in surveys and interviews with users is that there exists a learning curve to choosing what types of instances or platforms are required for their specific HPC workload and ultimately the deployment of the application. With an improved ease of use and ease of deployment for HPC jobs, which comprises of two main components (education and ease of use on the CSP side), the uptick for running more workloads in the cloud could be higher.

More Cost Effectiveness and Transparency for Large HPC Jobs

Cost has consistently been a point of concern for many HPC users when evaluating the cloud, whether it is the cost of the compute, the cost of the data movement, or the cost of additional resources on top of their on-prem budget. Furthermore, many CSPs have pricing models that start out quite inexpensive for small, experimental runs, but increase dramatically as the data set sizes and scaling grow. Some users have been shocked at their bills after scaling out an experimental workload in the cloud, highlighting the non-linear fashion of the pricing models. The higher cost for running in the cloud, is sometimes due to the lower performance, resulting in a job taking many more CPU hours to complete with run in a cloud.

On the other hand, recent studies have shown that some HPC workloads can be competitive in cost-effectiveness in terms of on-prem compared with the cloud, and there are even some HPC workloads that are more cost effective to run in the cloud versus on-prem. This group of workloads includes those that are highly parallelized, such as many genomic applications. Some of this is due to the continued technological improvements of the cloud for certain workloads.

Users are starting to factor into their cost-effectiveness analysis the non-monetary aspects of running an application in a cloud. This includes the fact that the cloud can have much shorter, and in many situations non-existent, queue times, as well as the ability to scale out beyond what an on-prem system may provide to their users. A continued effort to make the cloud more cost effective compared with on-prem for more HPC workloads, especially the harder HPC jobs mentioned previously, would result in further growth above what Hyperion Research is currently forecasting.

Major Increase of Running AI Workloads in the Cloud

The final driver behind the potential for a potentially higher growth year in the forecast is the influence of AI (including ML and DL) on HPC users and the continued trend of HPC users running AI workloads in the cloud. Based on recent studies, users indicate that they run a higher portion of their HPC-enabled AI workloads in the cloud compared with the portion of their traditional HPC modelling and simulation workloads in the cloud. When asked to project out the next year, users indicate both portions should increase, but the AI portion will grow more, highlighting a push to run more HPC-enabled AI workloads in the cloud. Covid-19 has accelerated cloud growth this year, and as AI continues to become more integral to HPC users' application portfolios, cloud growth may increase drastically to reflect the supplemental adoption, pushing the end user spend close to $10 billion by 2024.

CONCLUDING THOUGHTS

This exercise is designed to hypothesize the potential impacts of major changes to running HPC workloads in the cloud, including traditional modelling and simulation workloads, AI workloads, highly parallelized workloads, and inter-processor communication-driven workloads. Cloud adoption is growing for HPC, especially as the CSPs have made directed efforts to bridge the gap in addressing a wide variety of HPC workloads, especially those that have been thought of for years as unable to run in a cloud as effectively as on-prem. There are still users who will not run workloads in the cloud, for a variety of reasons, and some who never will. The same goes for workloads, as some workloads still do not run as effectively, cost or performance-wise, in the cloud when compared with on-prem solutions.

The state of the cloud market for HPC right now can be related, in a way, to the classic problem of two trains travelling from two different locations toward a common point. Users continue to experiment and work to understand how to take best advantage of the cloud, as well as learn how to modify their current applications to run more effectively in cloud platforms. That is one train. On the other hand, the CSPs continue to address the issues on their side, making their platforms more performant and friendlier to a wider and wider breadth of HPC applications. This includes the addition of more HPC-centric hardware and software, offering bare metal instances, the reduction of virtualization penalties, and better cost-effectiveness for running HPC workloads at scale. That is train number two. Each train is moving at different velocities towards the same point: where most workloads could potentially run cost-effectively and performant in a cloud platform as well as on-prem. That spot is attainable but will take some time and work to reach that point. The drivers highlighted in this paper will accelerate the respective user and CSP velocities when implemented.

For any questions, comments or suggestions, please email Alex Norton: anorton@hyperionres.com.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
