AI and HPC: Cloud or On-Premises Hosting

The Benefits and Limits of Cloud Computing in the Fastest Growing Segment of IT

Introduction

Artificial Intelligence (AI) and High-Performance Computing (HPC) are both computationally-intensive workloads. They demand fast central processing units (CPUs), accelerators, very large data sets, and fast networking to support the high degree of scaling typically required. All this fast hardware can be difficult to manage and expensive. AI and HPC adopters must try to minimize costs while delivering the performance and agility demanded by the organization's mission. Chief among the decisions that must be made is whether to build and host the application on a public cloud or build an on-premises infrastructure. While the industry trend is clearly to move new applications to the cloud, AI and HPC workloads have performance, data requirements, and utilization characteristics that could lead one to go in the opposite direction.

This paper will look at some of the common factors that help determine where these applications should be hosted to optimize their complex set of requirements against the backdrop of ever-challenging budgets. As is often the case, the answer to the question is, "It depends." However, the sheer magnitude of computing resources required for these powerful applications may lead many serious adopters back to their own datacenters.

This paper assumes a general understanding of AI and HPC technologies. For those seeking a more comprehensive introduction to AI, we have prepared a white paper here and an assessment of Dell EMC in AI here.

The Cloud Computing Landscape For AI and HPC

Let's start by looking at the array of cloud computing infrastructures offered by the major cloud service providers (CSPs). Each cloud provider has standard and proprietary hardware offerings, but tend to differentiate more in the unique software they make available as a service ("Machine Learning as a Service" or MLaaS). While these solutions can be attractive, they can also present a lock-in barrier to subsequent migration to on-premises solutions or alternative cloud services.

Acceleration Hardware

From a hardware standpoint, the CSPs all support some degree of elastic instances to accelerate training and inference processing of various Deep Neural Networks (DNNs). Graphics processing unit (GPU)-equipped instances are typically priced by the hour. With that common baseline, Google adds their own tensor processing unit (TPU) accelerators for TensorFlow-based training and inference, while Amazon adds the Xilinx-based F1 field-programmable gate array (FPGA) instances to AWS EC2 for inference processing. Amazon recently announced their own application-specific integrated circuit (ASIC) for inference processing, but it is not yet available. Baidu also announced its own ASICs, dubbed Kunlun, for AI training and inference acceleration. Note, these ASIC accelerators do not support HPC applications at this time.

Software and Machine Learning as a Service

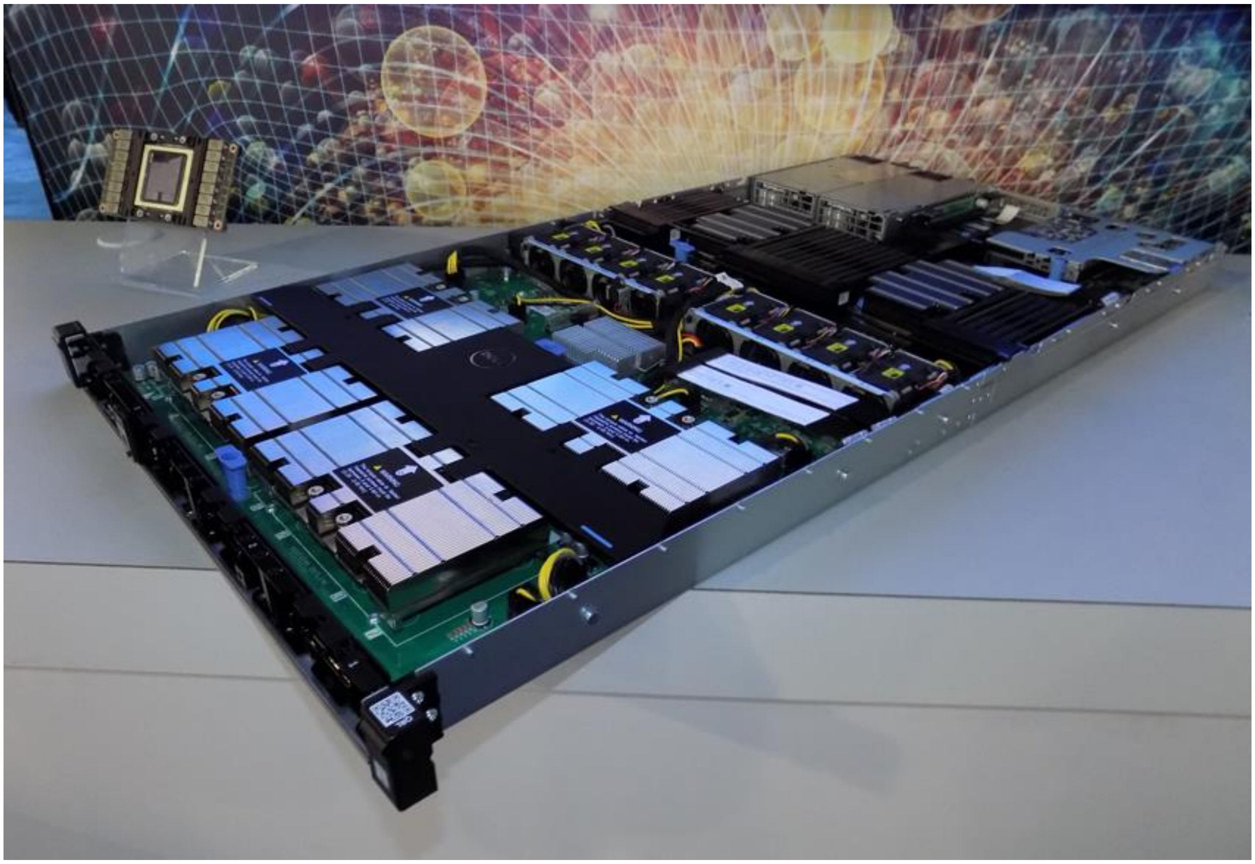
Cloud hardware instances are available with in-house optimized frameworks and management tools and with pre-built and customizable neural networks for image, voice, natural language processing, and a few specific application areas. These MLaaS offerings can potentially reduce the cost and time required to add AI features to an application and they offer a quick and easy path for experimentation.

To some extent, each CSP favors a specific (in-house) AI framework—Google promotes TensorFlow, AWS favors Apache MxNet, and Microsoft Azure aligns with their Cognitive Toolkit and Pytorch (as does Facebook). While all these frameworks are available as open source downloads for running in-house, the MLaaS application programming interface (API) offerings and tools such as AWS SageMaker or Google AutoML are strictly available from the hosting provider and only for apps developed and running in their cloud. For example, a developer will need to rewrite portions of an application if they later want to run the app in-house or move to an alternative CSP if the application is built using AWS SageMaker and Lex for dialog bots or AWS Polly for text to speech.

The Dell EMC Computing Portfolio For AI and HPC

Dell EMC has a rich portfolio of workstation, server, storage, and networking hardware for AI and HPC, including the latest GPUs, FPGAs, and CPUs. The PowerEdge C4140 Server and the Dell EMC Ready Solutions for AI are particularly important for this discussion.

FIGURE 1: THE DELL EMC POWEREDGE C4140



The Dell EMC PowerEdge C4140 is designed specifically to support the compute needs of HPC and Machine Learning with two Intel Scalable Xeon CPUs and four NVIDIA GPUs interconnected over NVLINK.

Source: Dell EMC

COMMON CONSIDERATIONS

SPECIFIC SITUATIONS: STARTUPS, ENTERPRISES, AND HPC

Most startups begin their AI journey using cloud-hosted services because it’s easier to spin up a GPU-equipped instance, upload training data, and begin to develop the neural network model than it is to plan, procure, and install the necessary hardware and software. However, many or even most startups will quickly outgrow this stage and reach the point where renting is no longer more affordable than owning the infrastructure.

Most enterprises already have a substantial IT organization with on-premises or co-located datacenters in place and typically conduct a total cost of ownership (TCO) analysis to determine the best approach. Many times, this analysis will hinge on the expected utilization rates of the GPUs and the scope and ramp of the organization’s AI journey. Since those are usually unknown factors in the early stages of research and development, many enterprises rightly choose to start their AI journey in the cloud and then move to their own hardware once they have production models and begin to keep the servers and GPUs busy. Note, inference services can still run on a public cloud to reach the largest audience even if the DNN is trained on-premises. However, startups must be careful to avoid paying for data movement or underutilized resources.

Lastly, HPC customers typically have an adequate user base and ongoing projects to tilt the TCO strongly in favor of on-premises infrastructure from the start. Also, many HPC shops already have GPUs for scientific applications and are eager to use them to solve new problems or develop models to make use of prior results.

Geographic Location

Obviously, organizations will need on-premises infrastructure if they work in locations where CSPs do not provide adequate services or where CSP pricing is high. Also, investigate the international availability of specific services and features, which may be limited to specific geographies — especially those in Beta or preview offerings.

Application Life Cycle: Experimentation, Development And Deployment

In the early phases of AI projects, the team typically does not have adequate information to plan a cost-effective infrastructure. In this phase, cloud services provide a one-stop shop for hardware, software, and ML services for model development and hyperparameter tuning. As early experiments begin to bear fruit and more model development projects are being contemplated, a TCO analysis can help determine when and if it makes sense to bring the work in-house. Assess the factors outlined below for a complete picture.

Once a Deep Learning (DL) model is deployed in a production environment, someone must continually monitor the performance, latency, and accuracy of inference predictions to ensure the quality of service remains in the targeted range. Here, the elastic nature of public cloud services can help manage the infrastructure through the ebb and flow of usage. However, the ongoing costs of inference processing, if heavily used, can become expensive in the cloud which leads many practitioners to migrate the work to on-premises infrastructure.

SETUP COSTS

Setup costs (which include the work to create the hardware and software infrastructure and everything else that needs to be done before the first training run) are typically lower for cloud-based AI and HPC development. For example, Amazon Deep Learning Amazon Marketplace Instances (AMIs) come completely configured and ready to use with GPUs and even FPGAs. Additionally, spinning up a few servers with GPUs is far easier in a cloud environment.

Dell EMC addresses the setup issue for on-premises AI with their Ready Solutions for AI. These are completely configured systems that are ready to use. The NVIDIA GPU Cloud can also help reduce software setup costs for popular AI frameworks and open-source HPC application stacks. Given these advances, the common perception that cloud offers an easier path to get started is perhaps overstated.

OPERATING AND CAPITAL COSTS

In the early phase of AI projects, the compute load may be somewhat light because it's dominated by experimentation with relatively sporadic server and GPU usage. In this phase, operating costs for training and inference work in the cloud are roughly comparable to on-premises hosting. However, more intensive usage in the cloud drives up costs significantly as an AI project begins to scale.

This is exacerbated by the common pricing model for GPUs, which is generally by the hour. Those hourly charges range from $0.18 to $24.48, depending on the generation of technology and size of instance.[1] For AI, ML, and DL, the more data that exists, the greater the opportunity for a more accurate prediction or result, which leads to demand for storage. Charges for public cloud storage vary based on zones, frequency of access, data transfer, acceleration, and replication rates.[2] When moving large datasets, bandwidth is more important than ever with T1 to T3 broadband ranging in cost from $300 to $3,000 per month plus set-up costs.[3] Consulting and other services are available via partners at an additional cost.

For heavily-loaded high-performance AI and HPC infrastructure (with many NVIDIA V100 GPUs, for example), the operating costs of hosting this work in the cloud can be

[1] https://aws.amazon.com/emr/pricing/for GPU instances
[2] https://aws.amazon.com/s3/pricing/
[3] https://www.costowl.com/b2b/office-internet-access-cost.html

higher than infrastructure in-house[4], even with a three-year cloud service contract. Moor Insights & Strategy advises organizations to undertake a detailed TCO analysis based on their specific situation to validate this assertion. For those who elect to keep their data-intensive workloads in the cloud, Moor Insights & Strategy also recommends implementing automated policies to help keep costs in check.

For most DL work, the software is predominantly open-source. The same is true for SPARK and other ML libraries, although many enterprises choose to leverage their contracts with major independent software vendors (ISVs) such as SAP, Oracle, and IBM, and use ML software from those vendors. In either case, the charges are constant with the number of users or CPUs licensed for use—whether on-premises or in a public cloud.

DATA AND APPLICATION LOCALITY

When considering locating new workloads, consider “data gravity” as it usually makes more sense to keep the application co-resident with the data sets which the application uses. If the data is already in the cloud, then the application should likely be located there as well. This is especially true for HPC and AI applications, since the training datasets typically span terabytes. Similarly, applications using the new AI features (or calling HPC libraries) should generally be co-located with the AI/HPC codes. This minimizes latencies, improves quality of service, and reduces costs. Of course, the application needs to reside there if the envisioned service is to be delivered via a public cloud.

The need to continually add and tag new data elements means the data’s velocity (or rate of change over time) must also be planned to transfer data to a cloud-hosted AI training service. In the case of Amazon AWS S3 storage, which is used by the AWS ML portfolio, transferring data into S3 is free, but there are costs associated with transferring data out of S3 or between regions and availability zones. Exporting data from AWS costs $0.09 per gigabyte (GB), while transfers between availability zones cost $0.01/GB.

Finally, on the data front, keep in mind that cleaning and curating the datasets for training is a time-consuming and labor-intensive process. It is often cited as the biggest unplanned expense and schedule factor in the entire process. If the dataset is to be

[4] For examples, see https://determined.ai/blog/cloud-v-onprem/ and https://www.dellemc.com/en-us/collaterals/unauth/analyst-reports/products/ready-solutions/dell_emc_big_data_tco_validation.pdf

used by other applications over time, think through those use cases. If these uses involve geographically dispersed teams, it may make more sense for that data and the AI it feeds to reside in a public or private cloud service.[5]

Security and Privacy

While Moor Insights & Strategy believes CSPs have done a reasonably good job of ensuring privacy and security for their users' data, some organizations may still be uncomfortable with the thought of sending highly confidential data outside their firewalls and facilities. Also, companies in certain industries such as healthcare and financial services need to comply with government regulations. These needs may be better met with secure on-premises IT.

In a study conducted by RightScale Cloud Management[6], 29 percent of respondents said that security represents a significant challenge to their organization in the cloud, and another 48 percent felt security was somewhat a challenge. While the risk of a security breach may be small, the impact of compromised data and credentials could be huge and even debilitating. Since AI and HPC typically involve massive proprietary datasets, security concerns can override the potential savings and ease of use benefits of hosting these new applications in the cloud. Note this is a corollary of the data gravity aspect discussed above. If the data is too sensitive to safely store in the cloud, then the new applications need to stay on-site.

Bursty Workloads

"Bursting" compute cycles from a private cloud or datacenter to public cloud services can be a cost-effective strategy when an organization needs to plan for a significant, but temporary increase in computing capacity. Bursting was pioneered in HPC seismic processing and helps reduce the capital costs of owning and managing more compute capacity than what is needed for steady-state operations. In fact, the principle can be used to respond to many on-demand surges in capacity (e.g., seasonal businesses during the holiday season). Training of DNNs might also be deployed in this fashion in the future, though currently it's hosted either on-premises or in the cloud.

Performance Requirements

Since nearly all public cloud providers now offer accelerators in their MLaaS infrastructure, the compute bandwidth in the cloud is roughly equivalent to what is

[5] https://aws.amazon.com/s3/pricing/
[6] https://www.rightscale.com/lp/state-of-the-cloud?campaign=7010g000001YbXx

possible with on-premises infrastructure. Cloud and on-premises compute instances can be configured to run on CPUs, GPUs, and even FPGAs (on AWS only, at this time). As for response times, cloud latency for inference queries is typically very low, although tests have shown latency can degrade significantly (up to 50 percent) when accessing cloud services across large distances. Latency generally increases by about one millisecond for every 60 miles, which can be a substantial hit for remote locations.

NEED FOR SERVICES

Many, if not most organizations, struggle with ML and DL because they lack the necessary skills and experience. This is why consulting services and training are often essential to support an organization’s early forays into AI. While some business partners offer these services to cloud users, few can match the expertise available from a full-service solutions vendor like Dell EMC.

CONCLUSIONS AND RECOMMENDATIONS

The migration from using privately owned datacenter hardware to cloud computing services is a reality. Spinning up a server or two with GPUs is incredibly easy to do in a public cloud infrastructure and the major CSPs have assembled an impressive suite of software and pre-trained neural networks to further simplify the on-ramp to their massive datacenters. In other words, starting with the cloud may make a lot of sense if an organization wants to experiment with AI and begin building DNNs.

However, many organizations will eventually need significant computing infrastructure for AI and HPC as their applications begin to run at scale. This, along with data transfer and throughput fees, begins to tip the cost balance in favor of building on-premises infrastructure as the organization matures in AI. The need for compute, storage, and networking speed is further magnified when AI training runs begin to demand tens (or even hundreds) of servers and GPUs. At this point, the benefits of easy cloud startup are dwarfed by the costs of dedicated cloud IaaS.

The MLaaS features in AWS, Google Compute, and Microsoft Azure provide a tempting platform for early learning and experimentation. For applications and data that are already hosted on those clouds, MLaaS can provide an attractive long-term solution for adding AI features like voice, translation, and image processing. However, those features may be inadequate for more serious AI endeavors and can present lock-in obstacles when an organization decides to shift to in-house computing services to save costs. For this reason and others, such as security and data movement costs, Moor

Insights & Strategy generally recommends that organizations plan to eventually host their AI and HPC workloads in private clouds and on-premises infrastructure.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR

Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by Dell. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.