Intersect360 Research White Paper: THE CASE FOR AI IN THE AZURE CLOUD

MARKET DYNAMICS

The Continuing Rise of Machine Learning

Through our research activities, Intersect360 Research interacts with thousands of real-world HPC users every year. A common thread that we're increasingly seeing in our surveys, at industry events, and in our work with clients, is that AI—specifically, machine learning—has become a major workload in nearly every industry sector.

Our most recent survey of HPC budget trends revealed that more than two thirds of HPC users have already implemented machine learning in their environments, with an additional 13% actively working toward machine learning implementation. Only a small fraction said that they were neither implementing nor investigating machine learning. (See chart.)

Usage or Investigation of Machine Learning at HPC Sites

HPC User Budget Map survey data, Intersect360 Research, 2022

Currently Running/Implementing Machine Learning	
We are currently using machine learning ML	67%
Currently investigating ML	17%
Not running now, but working to implement ML	13%
Not running or investigating ML	4%

The implications of AI and machine learning are evident throughout the industry. One example is in typical configurations for server nodes in HPC clusters. Over the past few years, the number of compute accelerators per node has risen steadily. In past years, the typical accelerated node would contain just one or two accelerators. This has changed markedly in the recent past, with many users saying that they are broadly using four accelerators per node. This is a direct result of users adding machine learning to their HPC workloads. Depending on the task and code, a compute accelerator like a GPU can increase throughput significantly, and GPU

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

architectures have been well-suited to machine learning. (See chart.)

Accelerators per HPC/AI server node

HPC/AI Technology Survey data, Intersect360 Research, 2022

Accelerators per HPC/AI server node			
	None	Some	Broad usage
1 accelerator per node	32%	42%	26%
2 accelerators per node	30%	48%	22%
4 accelerators per node	32%	38%	30%
8 accelerators per node	54%	36%	10%
10 accelerators per node	90%	8%	2%
16 accelerators per node	88%	9%	3%

Taking it to the Cloud

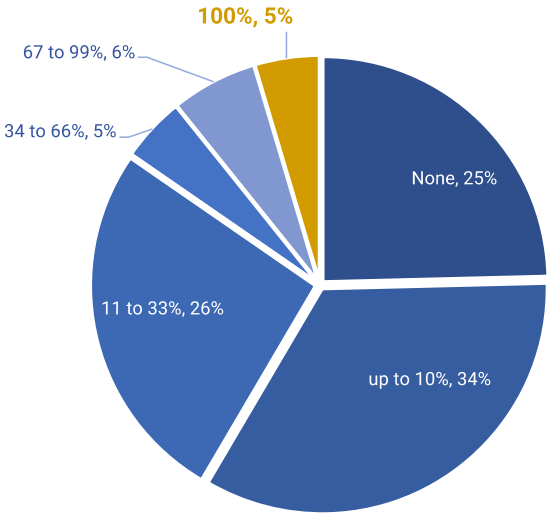
With a diversity of specialized computing required for the full range of HPC and AI workloads, HPC users are increasingly turning to cloud computing as a critical component of their overall capability. 75% of commercial/industrial HPC users leverage public cloud to run at least some of their HPC/AI workloads, and 16% are using cloud for over one-third of their total HPC/AI workload. 5% said they are now completely cloud-based for HPC and AI. (See figure below.)

Some of the use cases include overflow from on-premises workloads, special projects like machine learning model training, or testing out alternative configurations in order to ascertain and better tune application performance.

Not every organization has access to an HPC cluster for their AI/ML tasks and, even if they do, there are often situations where an on-premises cluster isn't the best solution. This paper explores three scenarios where running AI/ML workloads in a public cloud can be the most efficient and effective solution.

Percent of HPC Usage in Public Cloud, among Commercial HPC Users

HPC/AI Technology Survey data, Intersect360 Research, 2022



Scenario #1: The Need for Speed

Every industry and business faces emergency situations where it's "all hands-on-deck" in order to ensure business continuity and in some cases, even long-term survival. Let's take the example of an aerospace firm that suddenly hears of a mechanical failure severe enough to prompt discussion of a ground stop for all airliners using a particular component. This same type of situation could apply to an automotive or medical device manufacturer.

There are a number of questions that need to be almost simultaneously considered and evaluated. They include:

- Are we facing a design problem or was it a one-off extremely rare situation?
- Is our part the problem for sure, or is it how another mechanism interacts with ours?
- Was it truly a mechanical problem inherent with the part or perhaps a maintenance issue?
- Where did the components for our part originate? Was there a problem with the particular batch of parts?
- Have any other problems, even very minor problems, ever been reported concerning this part?

Every industry and business faces emergency situations where it's "all hands-on-deck" in order to ensure business continuity, and in some cases, even long-term survival.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

The answers to these questions (and many more) will determine the right course of action for the situation and dictate future steps to resolve the immediate situation.

These questions are too complex to be answered by humans simply running database queries**.** AI and ML routines can run through the vast permutations involved with much more precision and orders of magnitude more quickly.

While the organization probably has AI/ML systems in their datacenter, are they configured correctly for the problems so that they can provide the quickest time to solution? Do they have enough, and the right type, of compute accelerators? Can they be scaled large enough and quickly enough?

This type of problem is tailormade for the cloud. The cloud gives users the ability to immediately set up compute instances that are configured correctly and have the right type and numbers of accelerators per node. The instances can be scaled up and down nearly instantaneously and even entirely new instances to answer other questions can be created on the fly.

Scenario #2: Quality of Solution

While many organizations have large datacenters, they often don't have the right systems to handle AI or ML tasks, particularly when these tasks aren't everyday requirements. Many organizations find themselves in a situation where they need to train a new ML model in order to keep their edge in business or research, but the rapidly growing size of the model means that their in-house systems will be tied up for days or weeks just with the training workload.

There is also a technology issue: Very few datacenter infrastructures can keep up with the relentless pace of new technology development. New CPUs, GPUs, and other accelerators are constantly being introduced while most major datacenter hardware refreshes only take place every three or so years.

New technology will absolutely reduce the time to solution, but it also has the ability to radically improve the quality of the solution. More variables can be added to models and more layers added to neural networks. The end results are solutions that can greatly improve business or research outcomes.

These questions are too complex to be answered by humans simply running database queries.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

The public cloud is a great alternative in this situation. Due to their size and buying power, cloud vendors often have an inside lane with tech manufacturers and usually get the latest and greatest technology, like CPUs and GPUs, ahead of even large system vendors.

This means that you can stand up cloud instances populated with the very best CPUs, GPUs, interconnects, and storage for your ML job and only use it (and pay for it) while you're using it for your short-term training task.

Scenario #3: We've Hit the Wall

The typical HPC/AI infrastructure is highly utilized, oftentimes at more than 95% on average, with jobs that can span weeks. As we've mentioned above, most organizations don't acquire HPC/AI system capacity very often, they typically make a big procurement once every few years.

With such a highly utilized infrastructure it is very difficult to schedule workloads that are unanticipated, brand new or have rapidly grown – like many ML and AI workloads. Regardless of how important these applications might be to the organization, the datacenter can't miracle resources into existence or quickly acquire new hardware when the budget is bare. Even a small cluster that can handle AI/ML could cost as much as $200,000 when you factor in the cost of the servers plus GPUs with prices starting at $10,000 each.

As we've all seen, AI and ML have the ability to revolutionize the way we solve problems and to provide solutions that we intuitively wouldn't consider. However, for AI/ML to have an impact on an organization, they have to be piloted and tested. This is where the problem comes in – which project are you going to cancel or delay in test a new ML model, particularly when you need to run it at scale?

This is where the cloud comes in. The huge variety of instances available in public clouds today ensures that you will be able to find an instance that mimics what you currently have in your datacenter or a new configuration that you might acquire when you buy new systems. With public clouds, you also have the ability to run your

AI/ML tasks at scale, meaning you can quickly evaluate how well your applications scale and correct problems right away.

INTERSECT360 RESEARCH ANALYSIS: NOT ANY CLOUD. THE RIGHT CLOUD.

Not all public clouds are created equally. They're designed to serve different markets and segments, some clouds are built to handle highly transactional enterprise loads while others are better at serving up streaming media, for example.

For HPC/AI workloads, the performance requirements are radically different than what is needed for enterprise applications. One of the cloud providers who seems to best understand this is Microsoft's Azure.

In looking at Azure we were impressed with the range of instances and solutions they offered. Azure has a flexibility that is missing from many cloud provider offerings, giving users a wide array of options for building their cloud cluster.

Azure offers x86 based instances with either Intel or AMD processors and even has an Arm option based on the Ampere Altra Arm CPU.

Users also need choice when it comes to accelerators, this is critical when it comes to time to solution and quality of solution for ML workloads. Most applications support CUDA and many support OpenCL acceleration and others have support for FPGAs. While nearly every cloud provider offers NVIDIA GPU instances today, they don't all offer an array of current and past models, plus several still don't offer the most current A100 GPU.

A vital component, but sometimes overlooked by competitors, is the interconnect that ties a cluster together. Microsoft's Azure features Mellanox HDR 200 Gb/s InfiniBand interconnects to ensure that users can get maximum performance and throughput from their HPC/ML systems.

AI and ML have the ability to revolutionize the way we solve problems and to provide solutions that we intuitively wouldn't consider.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Azure has the flexibility to meet nearly any need when it comes to HPC/ML processing. While no cloud vendor can offer true mix and match capabilities today, we believe that Azure, through their wide range of HPC/ML oriented instances, comes the close to giving users the ability to craft exactly the cluster they need to match their workload.

White paper sponsored by Microsoft Azure. Full neutrality statement at https://www.intersect360.com/features/neutrality-statement.
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com