Intersect360 Research White Paper: THE FAST-CHANGING, SYMBIOTIC EVOLUTION OF HPC AND AI

MARKET DYNAMICS

The Convergence of HPC and AI

For decades, scientific discovery has been fueled by investments in High Performance Computing (HPC). Whether in research fields, such as climate modeling, astrophysics, or genomics, or commercial endeavors, such as engineering, oil exploration, or pharmaceuticals, HPC has been deployed to help scientists and engineers reach their next insights and innovations sooner, and to make the once-impossible possible.

While the need for scientific advancement does not wane, the computing tools that comprise HPC have evolved over multiple generations. Monolithic supercomputers gave way to modular clusters of servers, open-source Linux displaced proprietary versions of Unix, and cloud computing emerged to complement on-premises data centers. Applications themselves have also evolved: "Big Data" analytics introduced a new dimension of data-driven computing. HPC has been changed by all these trends, yet the need for HPC has persisted.

Today, artificial intelligence (AI) has transformed the landscape yet again, in a way that is thoroughly entwined with HPC. First, HPC technologies are essential in creating the complex systems to train AI models, and in many cases, to make real-time inferences based on them. Furthermore, AI feeds back into HPC, as many calculations can potentially benefit from the predictive methods of machine learning.

"Artificial intelligence" is a general term that has existed for decades in computer science (and science fiction), representing the idea that a computer program can learn or adapt based on information it is given. A more precise term for the current generation of AI is "machine learning," wherein algorithms are trained to recognize patterns in data and to apply what they have learned to new data. "Deep learning" is a specialized type of machine learning that continues to teach itself recursively with a neural network architecture, without requiring human interaction to continue to learn on an ongoing basis.

The internet-driven data connectivity of Hyperscale companies—such as Amazon, Alphabet (Google), Apple, Meta (Facebook), and Microsoft—ignited a new AI boom in the previous decade and remains a hotbed of machine learning innovation today. Because of their scale, these companies found themselves in possession of levels of data never before accumulated. By also investing in HPC technologies, they were able to build machine learning and deep learning architectures that ushered in a new era of AI. Thanks to this convergence, the

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

combined HPC-AI market has taken off, with $62 billion in worldwide data center spending in 2022, including a whopping $17.9 billion from Hyperscale companies worldwide.[1]

This new era of AI requires HPC to drive machine learning and deep learning. Merely possessing the data is insufficient. Gaining insights from it requires massive computation at scale, with parallel computational elements examining far-flung data connections. To serve web pages during a busy internet shopping day might require many servers, but one transaction has nothing to do with the next, so there is no burden to connect disparate web servers. But if an organization wanted to learn patterns in the transactions, there is a much greater requirement for connectivity and calculation. Fast networks and advanced computational processing elements are a necessary part of the equation. HPC feeds into AI.

Furthermore, AI feeds back into HPC. The idea that AI complements HPC without replacing it is implicit in an examination of the computational techniques at play. Historically, most scientific computing applications are *deterministic*, meaning they are used to solve complex equations in order to arrive at a specific answer. Presuming a mathematical error has not been made, the accuracy of the answer is reflective of how well the equations represent physical reality. A machine learning algorithm is *experiential*. It studies data patterns in existing data (the training phase) in order to make predictions or determinations about a new piece of data (the inference phase).

Consider how both these approaches can be used in a common problem, such as predicting the path and intensity of a hurricane in the days prior to landfall. A scientific computing approach takes a mathematically intensive stance to calculate the track and intensity based on a wealth of variables such as air temperature, water temperature, wind patterns, air pressure, etc. The more accurately the equations and variables represent reality, the better the prediction can be. A machine learning approach looks at how other storms have behaved in the past in similar or dissimilar situations in order to predict how this one might behave. The quality of the prediction hinges on how relevant the previous learning is to the new situation. Either way, lives on land depend on it.

The symbiotic relationship between HPC and AI is evident in the dualistic approach now taken by the majority of HPC users, over 80% of which now have initiatives centered on AI—or more specifically, machine learning—usually in close concert with traditional HPC. In a 2023 survey of HPC users, Intersect360 Research found a wide range of ways in which machine learning was being adopted to complement scientific computing. (See chart.)

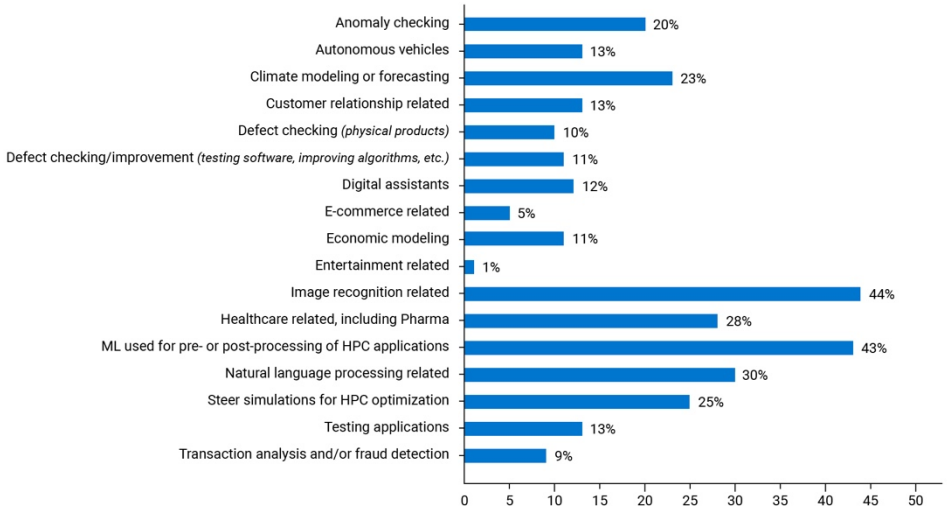
A more precise term for the current generation of AI is "machine learning," wherein algorithms are trained to recognize patterns in data and to apply what they have learned to new data. This new era of AI requires HPC.

Furthermore, AI feeds back into HPC, complementing HPC without replacing it.

[1] Intersect360 Research HPC/AI Market Sizing and Forecast Webinar, May 2023, https://www.intersect360.com/presentation/may2023/.

What Machine Learning is Used for at HPC Sites

Intersect360 Research, 2023



When respondents were asked how machine learning is being adopted to complement scientific computing, one of the most common responses was "pre- or post-processing of HPC applications," selected by 43% of respondents. This is to be expected among HPC users, and it also highlights one of the most straightforward ways in which AI can benefit HPC. One example of pre-processing is the notion of target reduction. Consider a pharmaceutical company that has 10,000 possible chemicals it might test for a new drug. It is prohibitive to send that many on to scientific simulation, let alone to synthesize them for physical testing. Machine learning might contribute by selecting the top one percent that are likeliest to bear results. Traditional HPC takes over for further analysis of the top 100.

Post-processing involves many types of analysis of results. In this case consider an oil company that has done seismic analysis of a potential oil field for development. Rather than having a geophysicist scanning every possible area, AI can do the first pass, highlighting areas of interest for the engineers to explore further.

This type of post-processing is related to image processing, which emerged as the most popular answer in the survey, selected by 44% of respondents. If AI can be trained to survey pictures of animals in order to identify cats, it can similarly be trained to look at scans from medical devices to detect anomalies, or to alert manufacturers to defects on a production line.

The third most common response, natural language processing, also has broad appeal. It is worth mentioning that this survey was completed before ChatGPT captured the world's imagination with generative AI. Future Intersect360 Research surveys will explore the widening range of applications of AI in HPC environments.

The symbiotic relationship between HPC and AI is evident in the dualistic approach now taken by the majority of HPC users, over 80% of which now have initiatives centered on AI.

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

Lurking just below these other responses is perhaps the most exciting potential combination of HPC and AI, the use of machine learning to "steer simulations for HPC optimization," currently being explored by one in four respondents. The idea of "engineer-in-the-loop" simulations, in which calculations are continually updated as an engineer explores design updates, has been pursued for decades, but it hasn't proven practical—largely due to human latency. AI brings this idea back in the form of computational steering. Given guidelines and goals, a well-trained machine learning algorithm has the potential to iterate through design ideas, studying HPC simulations as it goes. The engineer need only review the most promising ideas at the end.

Architectures for HPC and AI

Both traditional scientific computing and machine learning rely on scalable, computationally intense workloads, served by high-performance clusters of servers. Most HPC users look to their internal infrastructure first before bursting out to public cloud, both due to skills and cost. In an Intersect360 Research survey, 77% of HPC users said they "have the in-house skills to take on machine learning."[2] When the infrastructure exists or can be implemented on-premises, it is generally viewed as being much more cost-efficient. 69% of HPC users say that "using public cloud is more expensive than using our on-premises systems."[3]

Still, machine learning can have different requirements than traditional HPC applications, and this is most notable in the processing elements that best suit the applications. Machine learning tends to rely heavily on GPU accelerators as added computational elements. Originally designed for high-performance graphics (GPUs are "graphics processing units"), GPUs have been widely deployed for over a decade in HPC strictly for their computational capabilities.

The types of calculations where GPUs shine have been a good fit for machine learning, and they are increasingly popular in HPC and AI. Presently, 89% of HPC users leverage accelerators—usually GPUs—for at least some portion of their workloads, and on average, 32% of deployed server nodes have GPUs installed. Users project this proportion to grow to 45% in two years. A server node might have only a single GPU accelerator, or as many as eight or more. The most common configuration today is four GPUs per node.[4]

87% of HPC users leverage accelerators—usually GPUs—for at least some portion of their workloads.

[2] Intersect360 Research, "Machine Learning in HPC: Workloads, Frameworks, Data Types, Configurations, Cloud Usage, Business Outcomes," September 2023.

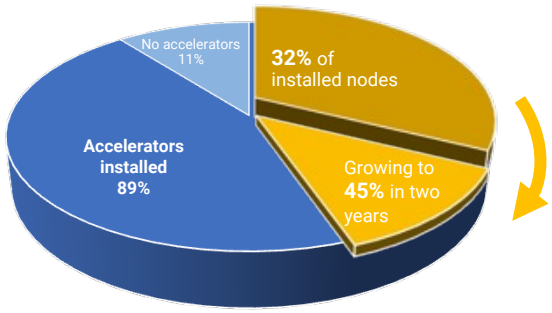
[3] Intersect360 Research HPC and AI Technology Survey, 2023.

[4] Intersect360 Research HPC and AI Technology Survey, 2023.

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

Growth in Accelerator Usage for HPC and AI

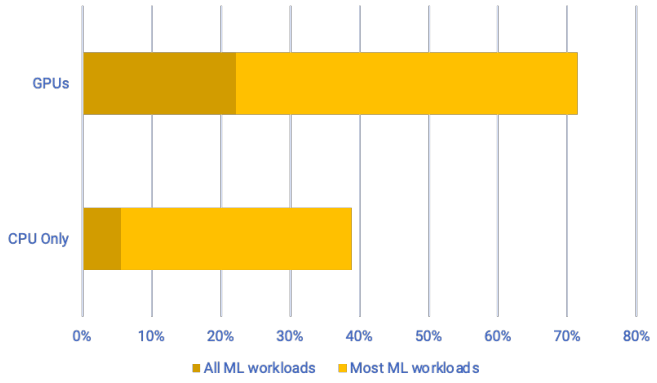
Intersect360 Research Machine Learning survey, 2023



These configuration trends point to the clear influence of AI in HPC. Prior to AI, usage of GPUs was common, but on fewer nodes, and at lower densities. The rise of nodes with four or more GPUs per node points to the extent to which HPC users are mixing AI into their HPC environments.

Processing Elements Used at HPC Sites for Machine Learning

Intersect360 Research Machine Learning survey, 2023



Of course, this comes at a cost. GPUs are expensive components, and it behooves organizations not to buy more of them than they need. In an Intersect360 Research survey, 16% of users reported "moderate" or "major" performance issues relative to expectations; when these occurred, "hardware-application mismatch" was seen as a culprit 49% of the time.

Furthermore, not all machine learning needs to be done on a GPU. There is an increasing trend to do more machine learning on high-performance CPUs when possible, saving on both expense and power consumption. (See chart.)

According to Intel, many machine learning algorithms perform better on Intel Xeon CPUs than they do on competing GPUs. As more machine learning–oriented features get built into the

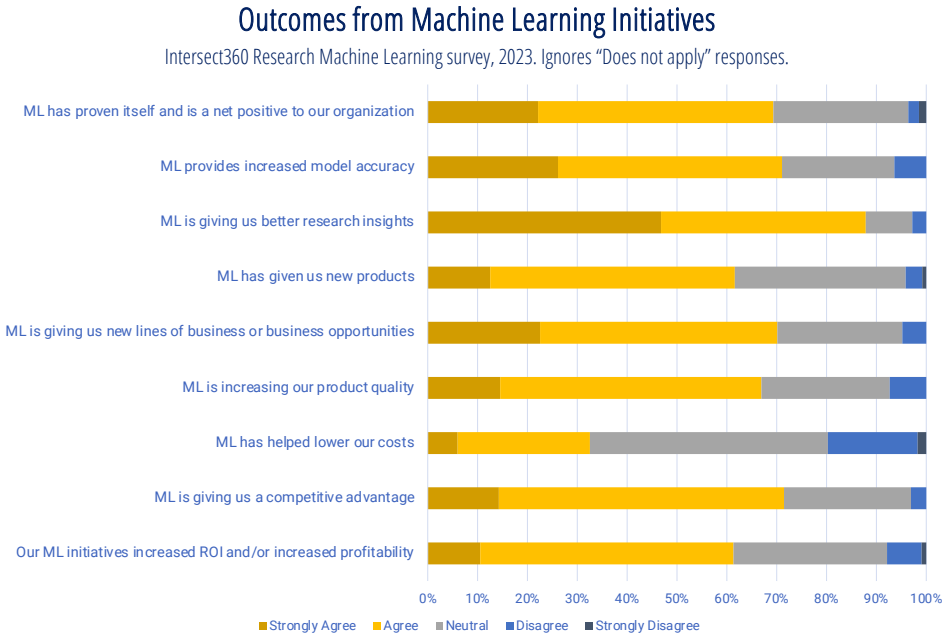
CPU, this could become even more prevalent. Intel says that performance on PyTorch, the most-common machine learning framework,[5] can be up to 10x higher with Intel's newest built-in features for mixed-precision performance.[6] Going forward, the ideal platform may have a balance of CPUs and GPUs that can be applied to any type of high-performance workload.

Driving Better Outcomes with AI

Regardless of how it is implemented, organizations are bullish on how AI can help their organizations. In our survey, respondents highlighted the many ways in which their organizations are benefiting from machine learning today:

- 88% say machine learning "is giving us better research insights."
- 71% say it "is giving us new lines of business or business opportunities."
- 62% say their machine learning initiatives led to "increased ROI or increased profitability."
- 67% say it "is increasing our product quality."
- And 69% say machine learning "has proven itself and is a net positive to our organization."

69% of surveyed users say machine learning "has proven itself and is a net positive to our organization."



[5] Intersect360 Research, "Machine Learning in HPC: Workloads, Frameworks, Data Types, Configurations, Cloud Usage, Business Outcomes," September 2023.

[6] Intel claims in this paragraph: https://www.intel.com/content/www/us/en/artificial-intelligence/processors.html.

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

INTERSECT360 RESEARCH ANALYSIS

AI is already a significant part of most HPC environments, and proof points, use cases, and enthusiasm are expanding. AI is contributing to growth across the HPC market, and the synergy between HPC and AI is accelerating the path of discovery and innovation for the organizations that pursue it.

The biggest challenge organizations face is knowing exactly what configurations will best suit their particular set of HPC and AI workloads, without oversubscribing on expensive components or falling into mismatch between hardware and application workloads. This creates a need for a trusted vendor that can tailor HPC-AI solutions to a particular vertical market.

Today, no vendor serves more organizations in HPC and AI than Dell Technologies, the leader in total HPC-AI solution revenue.[7] Dell Technologies was the most-cited vendor in a survey of HPC-AI users in 2023, both in number of sites and in total usage.[8] (See chart.) Dell Technologies offers "Dell Validated Designs" for HPC, analytics, and AI[9] that are specialized to target markets such as manufacturing (computer-aided engineering),[10] life sciences (genome sequencing),[11] financial services (risk management),[12] or government research (multi-use systems across HPC, analytics, and AI).[13]

Internally, Dell Technologies pursues advancements in HPC and AI through its HPC & AI Innovation Lab, where the company's engineers test and optimize new generations of technologies for processing, networking, and storage, including solutions based on Intel processors.[14] Globally, Dell Technologies also operates its HPC & AI Centers of Excellence, which showcase the latest solutions and provide community collaboration opportunities with the wider HPC community[15], and its Worldwide Customer Solution Centers, which offer remote access capabilities for testing and optimizing customer-specific workloads in collaboration with solution specialists.[16]

No vendor serves more organizations in HPC and AI than Dell Technologies, which offers Dell Validated Designs for HPC, analytics, and AI that are specialized to specific target markets.

---

7 Combined HPC-AI system and storage revenue. Intersect360 Research HPC market model and forecast data, 2023.

8 Intersect360 Research, HPC-AI Technology Survey, 2023.

9 https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/briefs-summaries/validated-designs-it-modernization-infographic.pdf.

10 https://www.dell.com/content/dam/digitalassets/active/en/unauth/briefs-handouts/products/ready-solutions/dell_ansys_solution_brief.pdf.

11 https://www.dell.com/content/dam/digitalassets/active/en/unauth/briefs-handouts/products/ready-solutions/dell-genomics-parabricks-solution-brief.pdf.

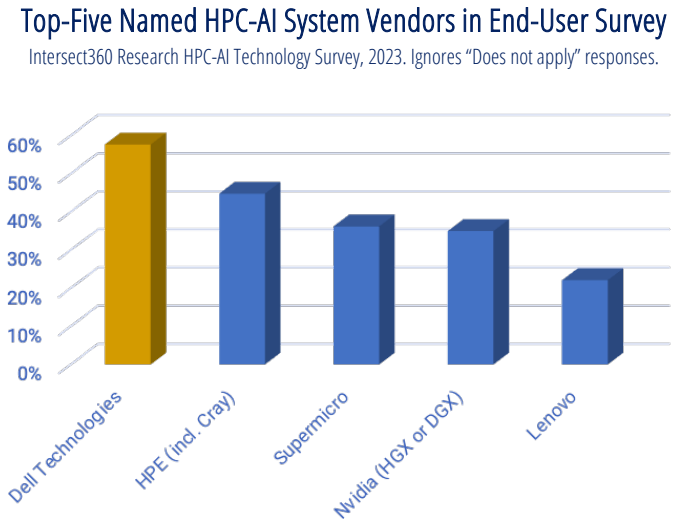
12 https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/briefs-summaries/dell-validated-designs-for-hpc-risk-assessment-solution-brief.pdf.

13 https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/briefs-summaries/solution-brief-dvd-for-govt-hpc-ai-da.pdf.

14 https://delltechnologies.com/innovationlab.

15 https://delltechnologies.com/coe.

16 https://delltechnologies.com/csc.



Advancing HPC and AI with Dell Technologies and Intel

Across multiple generations of innovations in HPC, the most consistent driver of performance year after year has been Intel. Even systems that leverage GPU accelerators typically have Intel processors at the heart of them. 93% of organizations rely on Intel CPUs in their HPC-AI data centers today.[17]

And some of Intel's biggest innovations are arriving on the market today. With the release of Intel Data Center GPU Max Series, HPC-AI users can now have both Intel Xeon CPUs and Intel Max GPUs in their high-performance data centers, both supported by the Intel OneAPI development environment. This combination allows users to leverage both an established processor and a new high-performance accelerator in the same system, selecting CPU or GPU as appropriate, and protecting their investments from previous generations of Intel-based optimizations.

Dell Technologies already addresses HPC and AI with Intel Xeon processors in its Dell PowerEdge XE9680 server. The Intel Max GPU is added in the Dell PowerEdge XE9640. Both will be integrated into Dell Validated Designs for HPC and AI.

Most importantly, Dell Technologies focuses on delivering value to its customers, basing solutions on the specific needs of a given environment. The Dell HPC-AI Community[18] forum completes the loop, allowing users to hear from experts in multiple fields, to share ideas, and to provide feedback on the features, technologies, and services that matter most.

93% of organizations rely on Intel CPUs in their HPC-AI data centers today.

With the release of Intel Max GPUs, HPC-AI users can now have both Intel Xeon CPUs and Intel Max GPUs in their high-performance data centers, both supported by the Intel OneAPI development environment.

[17] Intersect360 Research, HPC-AI Technology Survey, 2023.

[18] https://www.dellhpc.org.

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

AI is changing the world, and it is changing how we look at HPC. HPC technologies are needed for AI, and in turn, AI is helping to push the boundaries of HPC. For the organizations investing in it, this mutually beneficial feedback loop is kicking innovation and discovery into high gear. But it needs to be done intelligently; just throwing GPUs at a concept can lead to wasted money and wasted effort. With its leadership and experience across HPC and AI and Dell Validated Designs for specific workloads, Dell Technologies offers both the technology and the expertise to help organizations not only to deploy AI, but to be transformed by it.

To learn more, please visit www.dell.com/hpc.

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