Intersect360 Research White Paper: Bright Computing: Scaling HPC and AI from Edge to Core to Cloud

EXECUTIVE SUMMARY

As we move into a new era of expanding applications and technology choices, the notion of scalability is changing. To meet the needs of today's IT organization, scalability needs to mean a lot more than adding more nodes or bytes. Enterprise workloads have evolved to include more types of scalable, high-performance applications. Traditional workloads still exist, but with the Big Data revolution, analytics entered the mix, and today's major initiatives are focused on artificial intelligence (AI). Today's IT infrastructure has to scale across different types of applications, with optimal utilization of resources and personnel.

This proliferation of high-performance enterprise workloads has corresponded to a related expansion in the technology components to solve them. The pendulum has swung from standardization back to specialization, fueled by AI and analytics. Incorporating these specialty components, and using them efficiently, is another hallmark of the new scalability.

Additionally, cloud computing is part of almost every enterprise IT environment, including High Performance Computing (HPC), analytics, and AI. Nevertheless, most high-performance workloads are run on-premises, for a variety of reasons, including data movement, software licensing, operational considerations, and most of all, cost. Complicating matters further, many applications leverage data from the ever-expanding "edge." Managing edge-to-core-to-cloud is a major initiative for enterprises implementing analytics and machine learning.

That's where Bright Computing comes in. With system management software spanning HPC, data analytics, and AI, Bright Computing is filling a necessary role in high-performance segments. Bright Cluster Manager sits across an organization's HPC resources, spanning core, cloud, and edge, and organizes them across workloads.

Bright Cluster Manager supports mixed environments: mixed processing elements, mixed architectures, and mixed operating systems. Administrators can track and manage the utilization of accelerators, including both GPUs and FPGAs, monitoring which workloads make the best use of each environment. Most importantly, the purview spans the hybrid cloud, which is an increasingly common model across enterprise computing.

Bright Computing addresses the challenges of the new scalability. More than bytes or cores or bandwidths, these are the fundamental hallmarks of the new scalability for high-performance enterprise environments.

White paper sponsored by Bright Computing. 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

MARKET DYNAMICS

New Dimensions of Scalability

Scalability is a concept as old as computing, with an ever-pressing need for stronger processors and more data. But as we move into a new era of expanding applications and technology choices, the notion of scalability is changing as well. To meet the needs of today's IT organization, scalability needs to mean a lot more than adding more nodes or bytes.

Enterprise workloads have evolved to include more types of scalable, high-performance applications. Traditional technical workloads are still in play, but organizations are increasingly incorporating machine learning and analytics into their high-performance datacenter environments. These new requirements are supported by a diversifying set of technologies. In order to drive multi-dimensional performance, systems may need to incorporate the latest technologies from a vast array of options in processors, accelerators, interconnects, and storage, keeping all of them optimally utilized. And those are just the on-premises considerations. Today's IT infrastructures need to scale outside organizational boundaries, with data and computing across hybrid environments from edge to core to cloud. Failing to meet any of these new notions of scalability—new workloads, new technologies, hybrid cloud, at full performance—limits an enterprise's ability to make the most of its IT initiatives.

Scaling Workloads: Incorporating Analytics and AI

Across both High Performance Computing (HPC) and enterprise computing environments, a few mega-trends have significantly altered the IT landscape over the past ten years. First the Big Data phenomenon swept through the industry, placing a greater emphasis on the volume and velocity of enterprise data and unlocking the value of analytics. Today the artificial intelligence (AI) has reraised the stakes, bringing a machine learning revolution to applications and algorithms.

In the *training* phase, machine learning applications work through known data examples for "successes" or "failures" of matching particular patterns: this picture is or is not a cat; this insurance claim is or is not fraudulent; this pattern of weather data does or does not predicate a hurricane following a certain path. In the inference phase, the same algorithm reacts to new data to predict outcomes. In many cases, the results of these predictions are fed recursively back into the algorithm for ongoing refinement.

This application revolution includes popular topics such as speech recognition, personalized medicine, and autonomous vehicles, but the full scope of AI's potential is expansive. Where there is sufficient data, machine learning has been found effective in augmenting existing applications, or in some cases replacing them. In some breakthrough areas, AI creates new categories of capabilities that didn't previously exist.

Failing to meet any of these new notions of scalability—new workloads, new technologies, hybrid cloud, at full performance—limits an enterprise's ability to make the most of its IT initiatives.

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

Compelling examples exist across the HPC landscape. Consider, if AI can be trained to read an x-ray or an EKG, can it also be trained to look at seismic data to find likely oil reservoirs? If AI can learn the rules of a game such as poker or go to find unexpected winning strategies, can it learn a different "game," such as "optimize the airplane wing," subject to rules corresponding to structures and constraints? AI is already useful in exploring target molecules for drug discovery and virus inhibitors; the same molecular science can be tailored to the creation of new plastics and polymers.

The appeal of AI goes far beyond the science and engineering of HPC in its potential to revolutionize traditional enterprise IT challenges. Anywhere there's data, machine learning can look for patterns and insights, and data exists today like it never has before. Logistics optimization, product lifecycle management, defect detection, customer relationship management, fraud detection, staffing optimization, robotic process automation: in all these areas and more, machine learning can introduce enhancements from evolutionary to revolutionary.

The challenge for organizations pursuing these initiatives is universal. Traditional workloads still exist. With the Big Data revolution, analytics entered the mix, and today's major initiatives are focused on machine learning. No one tripled the IT budget. Today's IT infrastructure has to scale across different types of applications, with optimal utilization of resources and personnel.

Anywhere there's data, machine learning can look for patterns and insights, and data exists today like it never has before.

Scaling Technologies: New Processors, New Storage, New Interconnects

The proliferation of high-performance enterprise workloads has corresponded to a related expansion in the technology components to solve them. In previous generations, IT infrastructure was typically homogeneous, built from industry standards, with predictable roadmaps of improvement. Today the pendulum has swung from standardization back to specialization, fueled in great part by a diversity of new, high-profile applications. Incorporating these specialty components, and using them efficiently, is another hallmark of the new scalability.

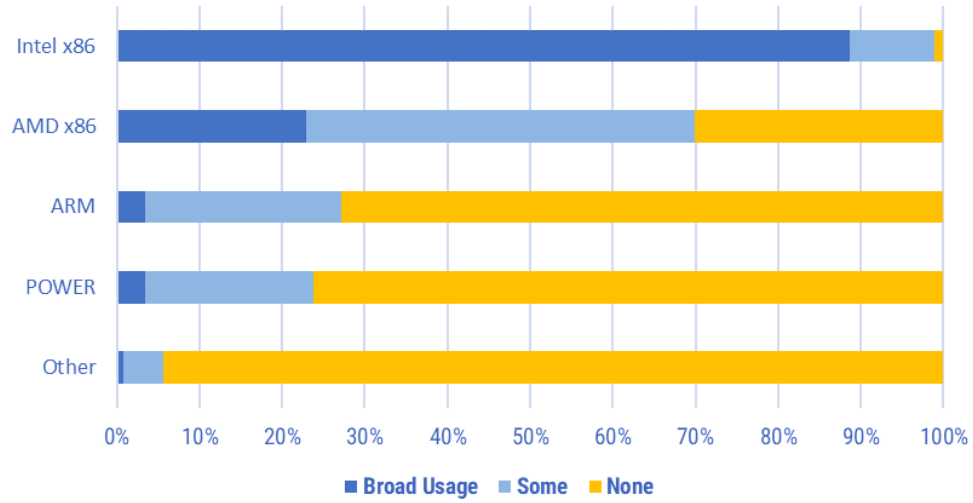
Nowhere is this more evident than in processing technologies. Recently this was nearly the sole domain of x86 architecture processors from Intel. A few alternates, such as IBM POWER processors, dotted the landscape, but most cluster nodes looked (and operated) the same. Today Intel is still Goliath—its x86 Xeon processors are in 98% of HPC environments—but that doesn't tell the whole story, and in fact, competition has heated up.

Within the x86 CPU sphere, AMD has made significant inroads with its EPYC processors, now based on AMD's third-generation Zen3 architecture. Endorsements are building, and AMD EPYC is now at least in trial in 70% of HPC datacenters, with widespread usage at more than one site in five. And these aren't the only CPUs. Processors based on ARM architecture are

also getting a significant look, now installed at more HPC sites than IBM POWER. (See chart below.)

Processors Installed at Surveyed HPC Sites

HPC Technology Survey: Processing Elements—CPUs and Accelerators, Intersect360 Research, 2021



Even if the CPUs weren't changing, there would still be a seismic shift in computing elements, driven primarily by the adoption of NVIDIA GPUs, used as application accelerators in conjunction with primary CPUs. The computational capabilities of graphics processors turned out to be well-suited for many categories of HPC applications. Intersect360 Research found that by 2017, the top 15 most-commonly cited HPC applications in our surveys all had some form of GPU computing support. Four years later, there has been a further GPU boom, as the accelerators have been well-suited to a wide range of machine learning applications as well. Today, 75% of commercial HPC users have GPU accelerators somewhere in deployment.

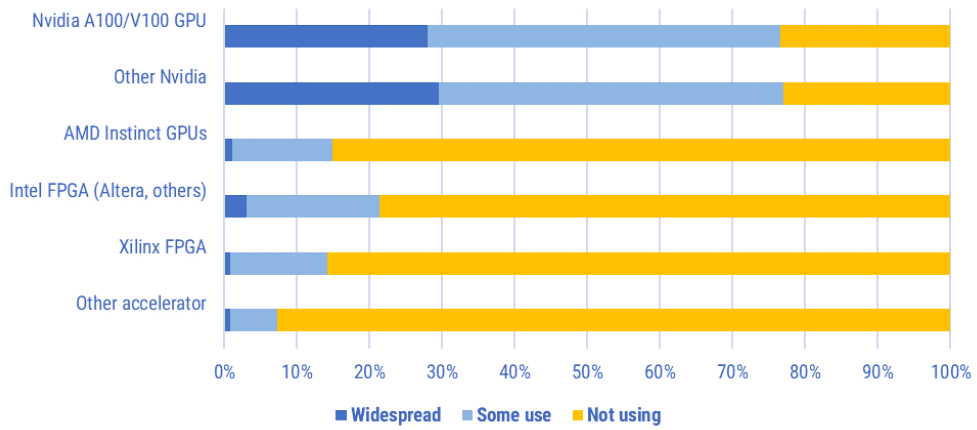
These are far from the only processing considerations. FPGAs (field programmable gate arrays) are also seeing increased usage, particularly for certain categories of machine learning. The field is also increasingly crowded by a buffet of specialty chips, each targeting a particular aspect of AI. These are reasonable considerations for enterprises whose workloads include the targeted applications. (See chart below.)

Storage and interconnects have similarly gotten more competitive. Analytics workloads tend toward data throughput, and flash storage is now a major part of most high-performance storage environments. It appears in network-attached all-flash arrays, as well as in hybrid flash-disk storage systems. Solid state disks (SSDs) are in local storage as well, creating another storage tier. With archives and cloud (see below) also in play, the storage hierarchy has gotten more complicated. And interconnects are still bifurcated between Ethernet and InfiniBand, notwithstanding an assortment of other proprietary, high-end supercomputing options.

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

Accelerators Installed at Surveyed HPC Sites

HPC Technology Survey: Processing Elements—CPUs and Accelerators, Intersect360 Research, 2021



Scaling to the Cloud: Hybrid Environments

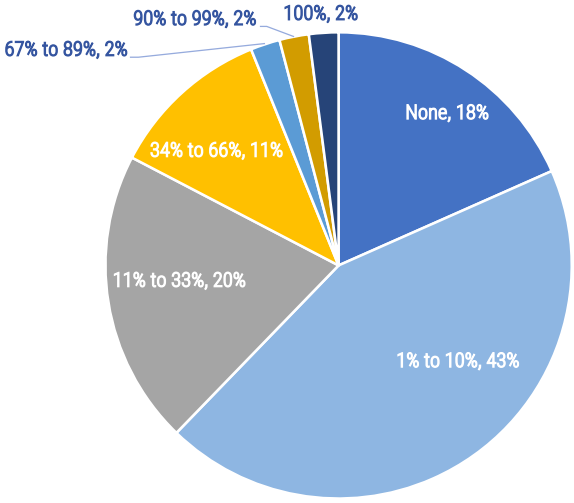
No discussion of the evolution of enterprise computing would be complete without looking at the dramatic effect of cloud computing. Cloud computing is part of almost every enterprise IT environment; it is well-established as a dynamic, scalable option for a wide range of workloads, including HPC, analytics, and AI.

Nevertheless, most high-performance workloads are run on-premises, for a variety of reasons, including data movement, software licensing, operational considerations, and most of all, cost. When applications can be run at a high level of utilization, most organizations find it more economical to buy in-house versus renting in the cloud. Over 80% of commercial organizations make at least some use of cloud for their HPC applications, but usually for only a minority of their total workload. (See chart below.)

Over 80% of commercial organizations make at least some use of cloud for their HPC applications, but usually for only a minority of their total workload.

Proportion of HPC Workloads in Public Cloud at Surveyed Commercial HPC Sites

HPC Technology Survey: Cloud Computing, Intersect360 Research, 2021



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

Complicating matters further, many analytics and machine learning applications leverage data from the ever-expanding "edge." Edge devices are anything that gather or generate data that can be aggregated for later analysis, such as credit card readers, medical imaging equipment, cable TV boxes, or autonomous vehicles. Not only is data initiated at the edge, but this is also where machine learning inference happens, making decisions based on previously trained pattern recognition. For applications with recursive data analysis, some training can even be pushed to the edge, then distributed to other devices over the network. Managing edge-to-core-to-cloud is a major initiative for enterprises implementing analytics and machine learning.

Managing applications both on-premises and in the cloud has its benefits. The cloud can easily absorb "bursty" workloads that exceed on-premises datacenter capacity. It is also a veritable trove of computing options, making it a viable sandbox for trial of the latest components. On the other hand, workload management is made more complicated, as is the notion of data sovereignty and control. Professionally managing workloads and resources across a hybrid cloud environment, from edge to core to cloud is yet another aspect of the new scalability.

INTERSECT360 RESEARCH ANALYSIS

Bright Cluster Manager: Managing the New Scalability

The primary challenge in adapting to the new HPC—new workloads and new technologies—is a question of management. How can I manage disparate high-performance technologies, whether on-premise, in the cloud, or at the edge, in an efficient way? Clusters have always come with management tools, but mostly they are open-source point-products that were not developed to encapsulate such a wide array. There is an increasing need for a professional, supported cluster management tool that spans all these dimensions of the new scalability in HPC.

That's where Bright Computing comes in. With system management software spanning HPC, data analytics, and AI, Bright Computing is filling a necessary role in high-performance segments. The company's core product, Bright Cluster Manager, is the most-cited commercial system management package among surveyed HPC users.[1]

Bright Cluster Manager sits across an organization's HPC resources, spanning core, cloud, and edge, and organizes them across workloads. A choice of user interfaces—command-line or graphical—helps administrators monitor and optimize applications, both in terms of cost and time. Moreover, certain features of Bright Cluster Manager speak directly to the current trends in HPC.

There is an increasing need for a professional, supported cluster management tool that spans all these dimensions of the new HPC.

That's where Bright Computing comes in.

[1] Intersect360 Research, *HPC User Site Census: Middleware and Developer Tools*, 2019.

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

Bright Cluster Manager supports mixed environments: mixed processing elements, mixed architectures, and mixed operating systems. Administrators can track and manage the utilization of accelerators, including both GPUs and FPGAs, monitoring which workloads make the best use of each environment. This helps support an environment in which each workload gets assigned to its optimal resource, shortening time-to-solution and reducing system overhead. When applications need to be moved between resources, Bright Cluster Manager provides the underpinnings of Kubernetes, Docker, and Singularity containers, all managed through the same interface. (See image below.)

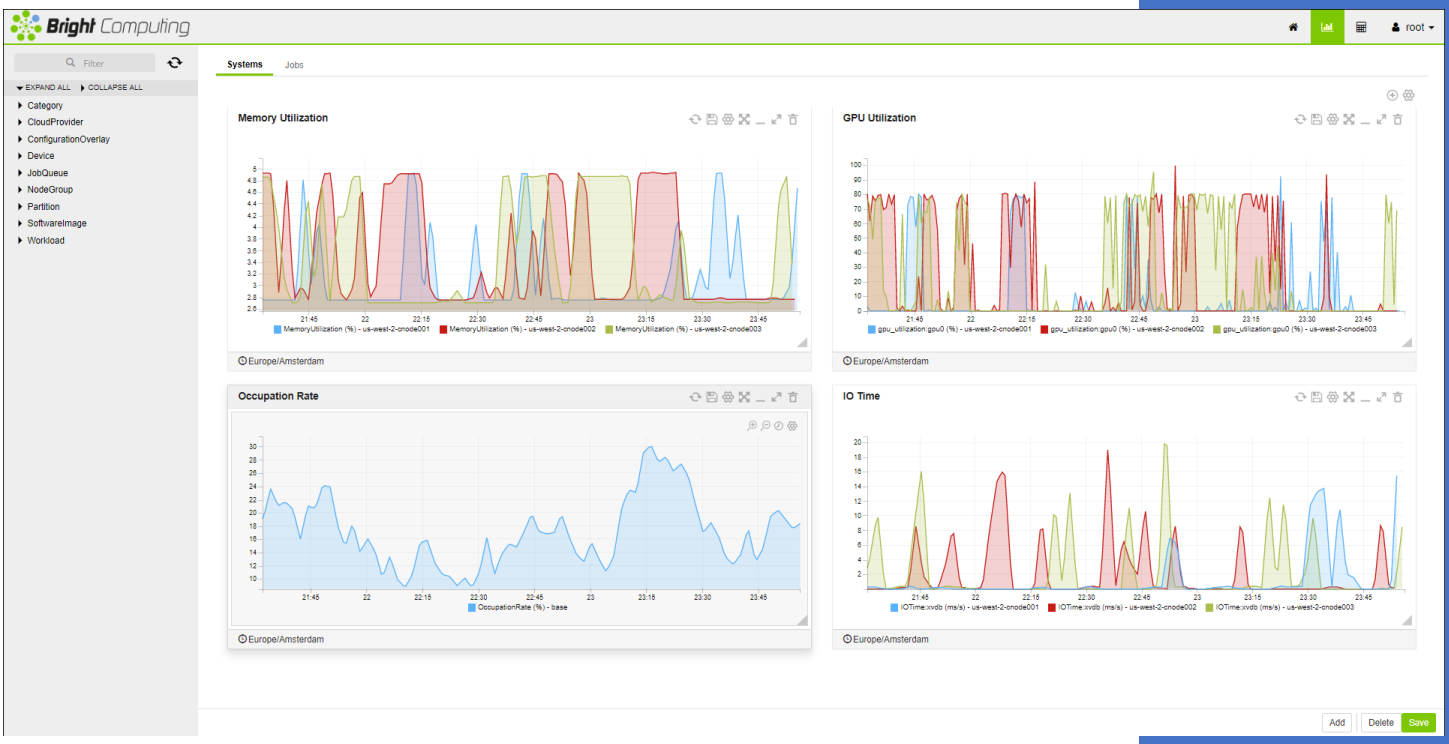


Image: Screenshot from Bright Cluster Manager (Bright Computing, 2020)

Most importantly, the purview spans the hybrid cloud, which is an increasingly common model across enterprise computing. Bright Cluster Manager can dynamically allocate instances from cloud providers like Amazon Web Services (AWS), Microsoft Azure, and cloud instances accessible from VMware's vSphere. The public cloud instances provisioned by Bright Computing are imaged identically to the those running on-premises, so that on-premises workloads can be migrated without modification. By managing these resources through the same user interface as on-premises systems, administrators can handle spikes in demand or shifts in deadlines while still optimizing total cost.

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

Edge to Core to Cloud

Incorporating Bright Cluster Manager addresses the fundamental challenge of matching resources to workloads. When an organization has multiple types of HPC systems with different operating environments—for example, a technical computing system with an open-source job manager such as Slurm, and a separate cluster with Kubernetes containers primarily for analytics or machine learning—there can be a question of which system has jurisdiction or priority for new jobs, particularly when reaching for other resources like cloud instances or edge computing data. Bright Cluster Manager sits atop all these platforms to organize resources across workloads.

This challenge is exacerbated when resources outside the datacenter are brought into the equation. How do you monitor the data collected or created at the edge? How are workloads and data sovereignty—the long-term control, stewardship, and access to data—managed in a hybrid cloud environment? In this case, more technologies and more suppliers have conspired to create a proliferation of tools, each proprietary to the corresponding organization or technology. The advantage of Bright Cluster Manager is that it is designed to span these diverse elements without locking into a particular vendor's path.

Fundamentally, Bright Computing addresses the challenges of the new scalability. HPC, analytics, and machine learning … Heterogeneous computing and storage architectures … Hybrid cloud environments … Managing workloads and data from edge-to-core-to-cloud. More than bytes or cores or bandwidths, these are the fundamental hallmarks of the new scalability for high-performance enterprise environments.

For more information about Bright Computing solutions for HPC, visit https://www.brightcomputing.com/product-offerings/bright-cluster-manager-for-hpc.

The advantage of Bright Cluster Manager is that it is designed to span diverse elements without locking into a particular vendor's path.

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