TECHNICAL OVERVIEW

ACCELERATED COMPUTING AND THE DEMOCRATIZATION OF SUPERCOMPUTING

Accelerated computing is revolutionizing the economics of the data center. HPC enterprise and hyperscale customers deploy accelerated servers because GPUs deliver unprecedented cost savings for their data center.

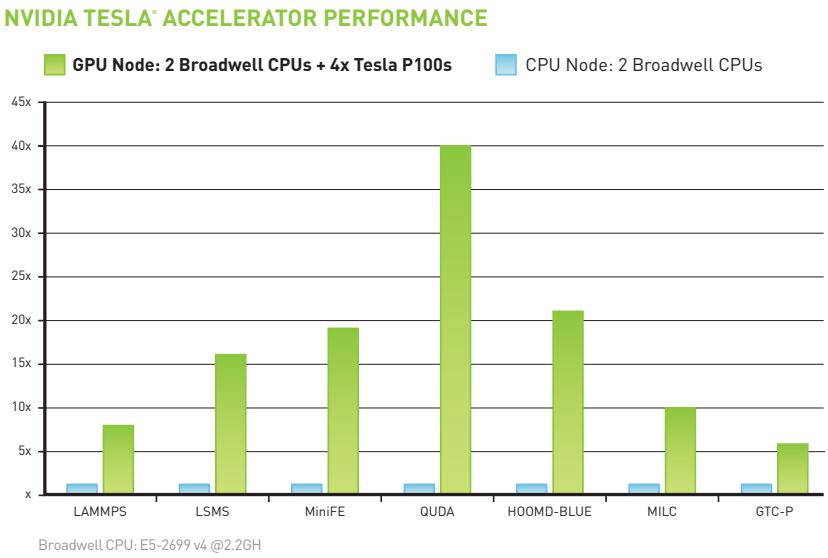
This whitepaper provides an analysis on how accelerators like the NVIDIA® Tesla® P100 can lower data center cost by up to 50%.

The New Data Center

Today, data centers are built by interconnecting lots of COTS (commercial off-the-shelf) technologies. Customers look to deploy the most cost-effective system by making trade-offs between commodity components like CPU, memory, and interconnect. However, cost savings are incremental.

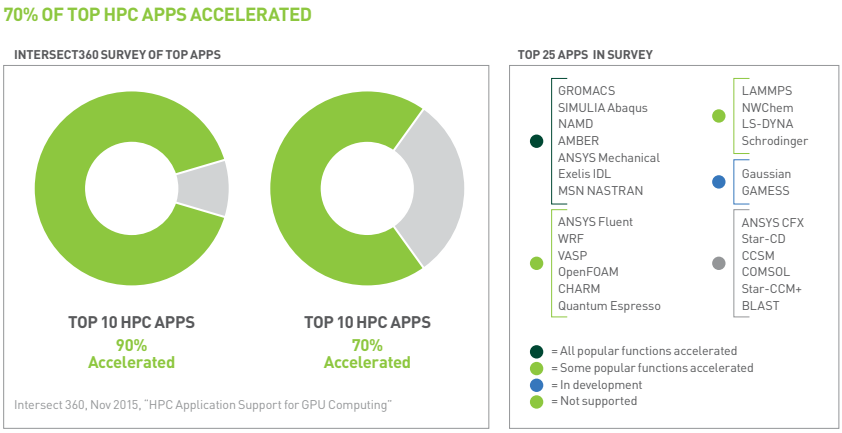
Accelerators fundamentally change the economics of the data center because application performance gains are no longer incremental. Applications typically speed up by 5-10X with accelerators. With Tesla P100 some applications experience over a 20X performance boost, as is shown below.

Figure 1: Accelerators deliver a 5-10X performance boost, fundamentally changing the economics of the data center.



A recent survey by Intersect360 Research, a leading HPC analyst firm, showed that 70% of the most popular HPC Data & Analytics applications support GPU acceleration today. With hundreds of accelerated applications, the question is no longer if GPUs should be deployed in the data center, but how many. The answer can result in dramatic improvements in cost savings.

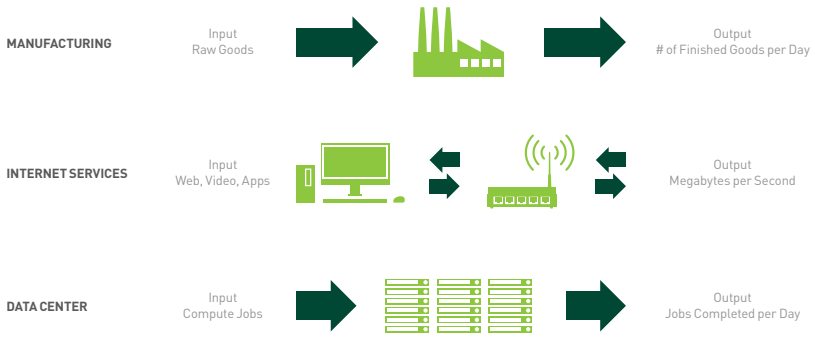
Figure 2: Many of the most popular applications in HPC and deep learning are GPU-accelerated to deliver unprecedented productivity and cost savings in the data center.



Ultimate Metric for Productivity

The productivity of an infrastructure is often measured by its throughput. In manufacturing, the primary metric driving profitability is the number of goods produced per day. In cloud services, users pay based on tiers of data throughput, measured in megabytes per second. The data center is no different.

Figure 3: Productivity of an infrastructure is often measured by its throughput, and the data center is no exception.



Data center throughput is measured by the amount of work completed in a given time (i.e. number of jobs per day or per month). While this architecture is complex, users typically abstract away all the system complexities into a simple working model; they submit job requests into a black box through a job scheduler and expect results soon after.

In high-performance computing, researchers rely on velocity of output of the data center for discoveries and insights. Higher throughput means more scientific discoveries are delivered to researchers every day. In web services, thousands of consumers using various types of devices may request live video streaming of a trending event. Higher throughput means a better user experience.

Throughput is the ultimate metric of data center productivity.

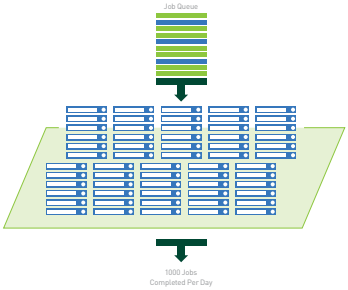


Figure 4: The CPU-Only data center has 1,000 CPU-only nodes processing 1,000 jobs per day.

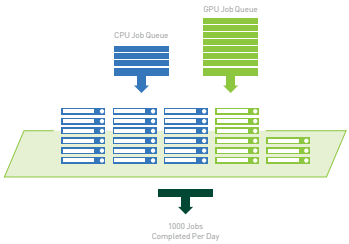


Figure 5: The accelerated data center has 300 CPU-only nodes and 70 nodes with Tesla P100 accelerators, delivering the same throughput as the CPU-Only data center with 63% less nodes.

Same Throughput with Fewer Server Nodes

To illustrate how throughput and cost savings are related, let's assume there are two data centers. The CPU-Only data center is comprised of traditional CPU servers and the accelerated data center is comprised of a mix of traditional CPU servers and GPU-accelerated servers. Each node is dual-socket CPU design, while GPU-accelerated nodes have two NVIDIA Tesla P100 accelerators attached to the node. In terms of workload profile for both data centers, we're assuming 70% of the jobs are based on applications that support GPU computing.

In this whitepaper, we'll assume that a single CPU node can process one unit of work, or job, per day. So the CPU-Only data center has 1,000 nodes capable of processing 1,000 jobs.

Let's take a look at the accelerated data center. Because 70% of jobs support GPU computing, 700 jobs in the queue can run on accelerated nodes while 300 jobs should run on CPU-only nodes. With a conservative assumption that GPU-enabled jobs run 10X faster on a Tesla P100 node compared to a CPU node, only 70 accelerated nodes are needed to deliver 700 jobs per day. 300 CPU nodes are required for the remaining jobs in the queue, for a total of 300 server nodes.

The accelerated data center delivers the same productivity with 63% less servers, racks, and networking equipment. This translates into tremendous saving in both acquisition cost as well as operation cost due to lower power and physical space requirements.

Aren't GPU-Accelerated Servers More Expensive?

Accelerators add cost to a node, so customers often make the mistake of concluding that the GPU-accelerated solution is more expensive. To analyze the cost impact of adding accelerators, let's start with an example breakdown of a server node.

COST	CPU-ONLY NODE (Dual Socket CPU)	ACCELERATED NODE (4x Tesla P100)
CPU	$2,000 (x2)	$2,000 (x2)
GPU	-	$5,500 (x4)
NIC, Memory, Misc. Cost	$4,000	$4,000
Total Node Cost	$8,000	$30,000

Table 1: Breakdown of CPU-only and GPU-accelerated node cost.

A single CPU socket costs $2,000 and other necessary components, like NICs and DDR4 memory, cost $4,000, totaling $8,000 for the node. By adding four Tesla P100 accelerators to the same node design, the node now costs $30,000.

While node cost is higher with GPUs, nodes cannot operate without other data center technologies like cables, switches, storage, and software—all which are significant adders to the overall cost. The typical breakdown of data center cost is as follows:

DATACENTER TECHNOLOGIES	% OF SYSTEM ACQUISITION BUDGET
Servers	60%
Networking	10%
Storage	10%
Software and Services	20%
Total Cost	100%

Table 2: Typical system budget allocation for the data center.

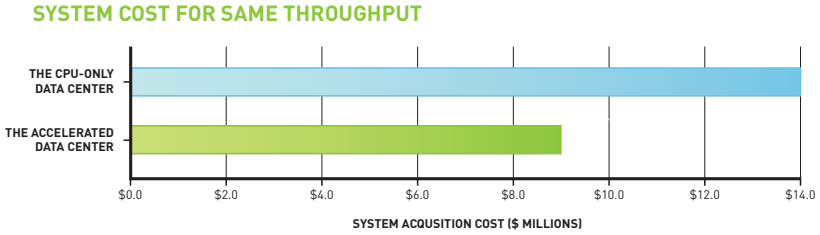
Using this breakdown, the CPU-Only data center would need to budget $8M for server nodes, $1.5M for networking and storage, and $3M for software and services. The accelerated data center needs a smaller budget for server nodes and networking, as there are less nodes to interconnect. Budget for software and services also decreases due to the lower number of nodes and smaller overall system budget. The budget for storage is the same as the CPU-Only data center.

COST	THE CPU-ONLY DATA	THE ACCELERATED DATA CENTER
CPU Nodes	$8,000 x 1000 Nodes	$8,000 x 300 Nodes
Tesla P100 Nodes	-	$30,000 x 70 Nodes
Servers	$8M	$4,5M
Networking	$1.5M	$1M
Storage	$1.5M	$1.5M
Software and Services	$3M	$2M
Total Data Center Cost	$14M	$9M

Table 3: The accelerated data center with Tesla P100 reduces system cost by 39% compared to the CPU-Only data center.

While the customer needs $14M to deploy the CPU-Only data center, they only need $9M when some of the nodes are accelerated as in the accelerated data center. The end result is 35% savings by choosing the accelerated data center. With a higher amount of workload accelerated the savings can be up to 50%.

Figure 6: 35% cost savings with the accelerated data center



What if the CPU is Given Away at No Cost?

In some situations, the CPU may be discounted in an attempt to stay competitive with a GPU-accelerated solution. So let's take an extreme case where the CPU is free for the CPU-Only data center, but full cost for the accelerated data center.

COST	THE CPU-ONLY DATA CENTER (Free CPU)	THE ACCELERATED DATA CENTER
CPU Nodes	$4,000 x 1000 Nodes	$8,000 x 300 Nodes
Tesla P100 Nodes	-	$30,000 x 70 Nodes
Servers	$4M	$4.5M
Networking	$1.5M	$1M
Storage	$1.5M	$1.5M
Software and Services	$3M	$2M
Total Data Center Cost	$10M	$9M
Total Data Center Throughput	1000 Jobs/Day	1000 Jobs/Day

Table 4: Even if CPUs were sold at zero-cost in CPU-Only data center, the accelerated data center with Tesla P100 is still saves 15%.

While the CPU-Only data center requires less budget to deploy 1,000 nodes, other cost factors in the data center remain the same. Node cost reduces in half to $4,000 and the overall data center cost reduces 29% to $10 million.

Even in this extreme, unlikely scenario, the accelerated data center is still lower in overall cost by 10%.

Maximizing Budget and Throughput

If a customer has a fixed budget that must be spent, Tesla P100 offers unprecedented ROI by maximizing throughput. With 70% of the top applications already leveraging GPU acceleration, and more applications on the way, many customers choose to deploy more GPUs into the data center.

With the 35% savings that the accelerated data center made possible, the IT manager can shift the savings to purchase more GPU nodes. Let's call this new data center, the max-accelerated data center, which contains a mix of CPU-only nodes and more GPU-accelerated nodes compared to the accelerated data center. This assumes enough workload exists to use the additional GPU nodes.

COST	THE CPU-ONLY DATA CENTER	THE ACCELERATED DATA CENTER
CPU Nodes	$8,000 x 1000 Nodes	$8,000 x 300 Nodes
Tesla P100 Nodes	-	$30,000 x 220 Nodes
Servers	$8M	$9M
Networking	$1.5M	$1M
Storage	$1.5M	$1.5M
Software and Services	$3M	$2.5M
Total Data Center Cost	$14M	$14M
Total Data Center Throughput	1000 Jobs/Day	2200 Jobs/Day

Table 5: The accelerated data center with Tesla P100 delivers over 2x higher throughput compared to the CPU-Only data center.

IT managers can deploy 150 more GPU nodes with the cost savings made possible with Tesla P100 GPUs. The max-accelerated data center also enjoys lower networking, software, and services costs due to lower number of nodes. With 220 GPU nodes producing 10X throughput increase, the max-accelerated data center delivers 2200 jobs per day, or over 2X increase in throughput compared to the CPU-Only data center.

Lower Cost with Accelerations

IT managers care about cost. The budget is never big enough to cover all the programs and equipment required to keep the organization working smoothly, so any cost savings is a welcome relief. With Tesla P100, IT managers have slashed their data center costs.

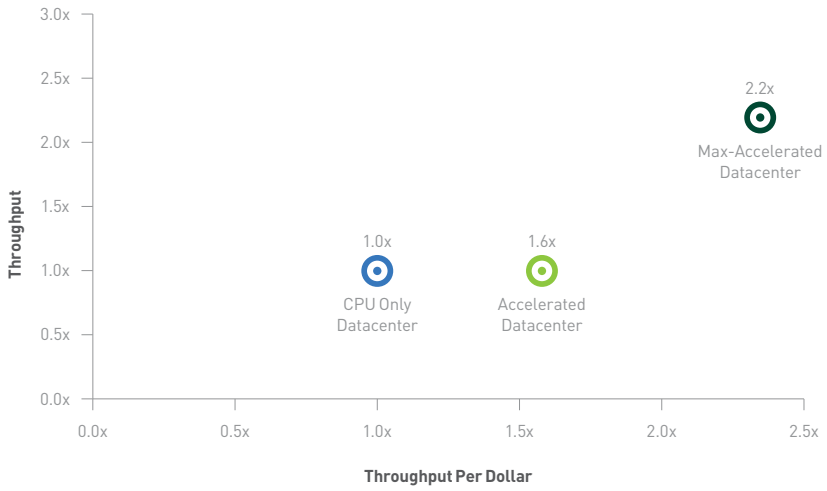


Figure 7: GPU-Accelerated data centers lower costs significantly and delivers the best throughput per dollar. (Data in above chart normalized to the CPU-Only data center)

In this whitepaper, we used three examples of data centers that customers deploy today. Compared to the CPU-Only data center, Tesla P100 in the accelerated data center results in 35% reduction in cost, saving $5 million in an overall budget of $14 million. For customers looking to maximize productivity, Tesla P100 in the max-accelerated data center delivers over 2X increase in overall productivity.

Democratizing the Supercomputer

Researchers and engineers in HPC, Hyper scale and Fortune 100 companies are dealing with massive amounts of data and their data center costs are growing. GPU-accelerated computing gives them not only better performance/throughput but also helps them manage their cost.

Accelerated computing democratizes supercomputing, making it affordable for more researchers, scientists and enterprise companies to deploy the system they need. Now, a university team focused on computing the cure to cancer, a research department looking to solve the mysteries of Big Bang, or a Fortune 100 company developing new business invitations can afford a computing system previously reserved supercomputing facilities.

In the era of accelerated computing, supercomputing is affordable and accessible.