

ACCELERATED COMPUTING AND THE DEMOCRATIZATION OF SUPERCOMPUTING

Table of Contents:

The Accelerated Data Center	3
Optimizing Data Center Productivity	6
Same Throughput with Fewer Server Nodes	7
Aren't GPU-Accelerated Servers More Expensive?	8
What About Operational Cost?	10
What if the CPU is Given Away at No Cost?	11
Maximizing Budget and Throughput	12
Lower Cost with Acceleration	13
Democratizing the Supercomputer	14

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® V100 can lower data center cost by up to 45%.

The Accelerated Data Center

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

Accelerators fundamentally change the economics of the data center, producing application performance gains that are no longer incremental. A single GPU-accelerated server can replace over 100 CPU-only servers. The number of CPU-only servers replaced by a GPU-accelerated server is called the node replacement factor (NRF). The NRF will vary based on the applications running on the server. One GPU-accelerated servers (with 4xV100) running a balanced workload allocation of these popular HPC applications replaces 31 CPU-only servers (NRF= 31X). However, the workload mix of supercomputing sites include more AI workloads which typically have an NRF of over 40X. [\[See the HPC Application Performance Guide for details\]](#)

Figure-1: A mixed workload of popular HPC applications has an NRF of 31X.

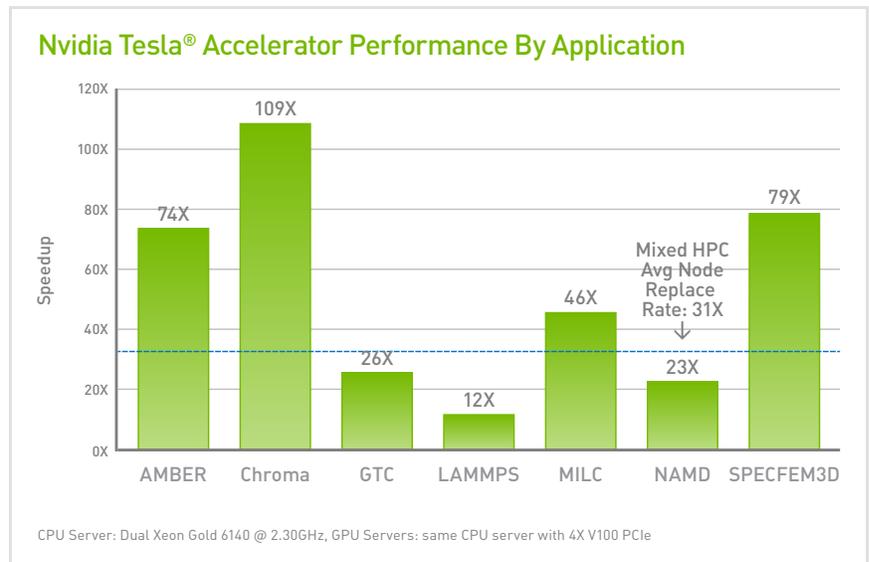


Figure-1

A recent survey by [Intersect360 Research](#), a leading HPC analyst firm, showed that 15 of the top 15, and 20 of the top 25 most popular HPC applications support GPU acceleration today.

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.

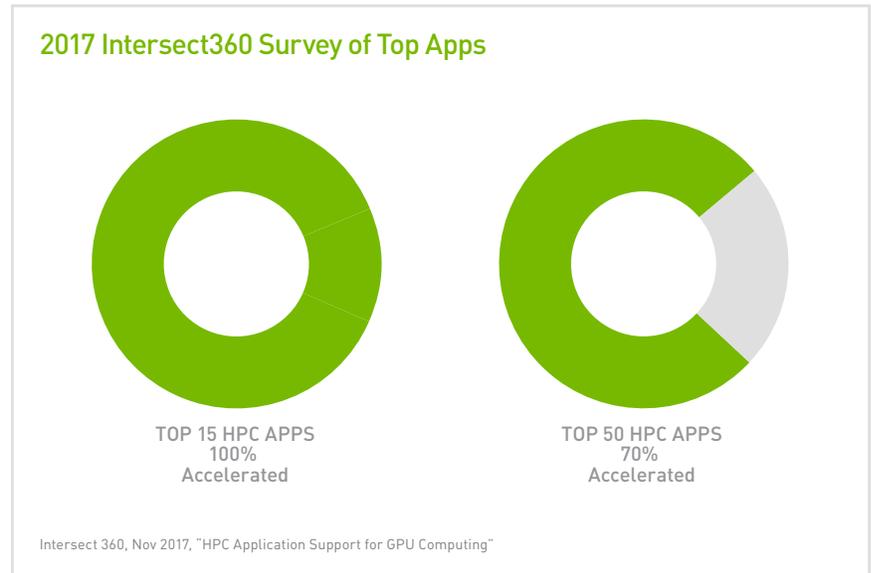


Figure-2

Now, with over 580 GPU-accelerated applications, and more applications being added every day, the question is no longer if GPUs should be deployed in the data center, but how many. This broad coverage of GPU-accelerated applications will help boost data center throughput and utilization, resulting in dramatic improvements in cost savings.

Figure-3: With over 580 GPU-accelerated applications, IT managers can optimize data center throughput and optimization.

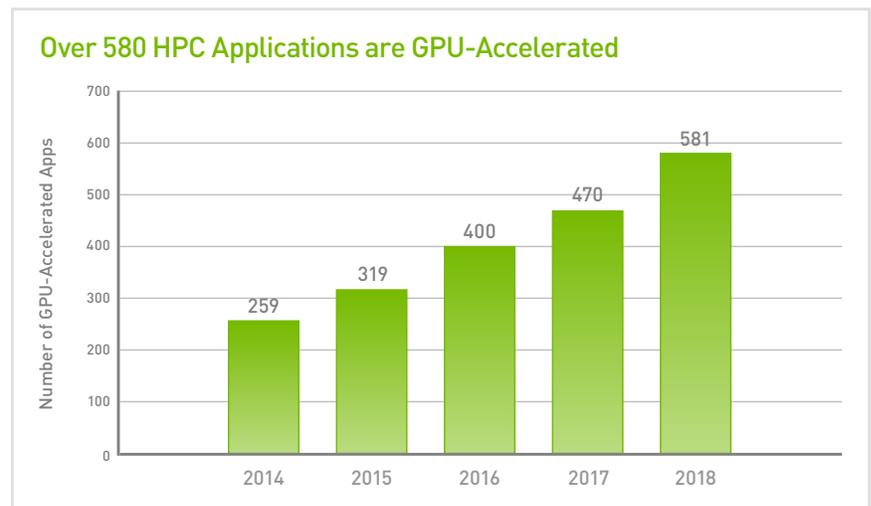


Figure-3

In addition to increasing the number of GPU-accelerated applications, NVIDIA continually optimizes each component of the software/hardware stack to improve end-to-end application performance over time. In Figure 4 we show the average NRF for a mixed-HPC application workload (shown in Figure 1) has increased from 6X to 31X in just 3 years.

Figure-4: Tesla platform performance improvement over time. Node Reduction Factor (NRF) increases from 6X to 31X in just 3 years

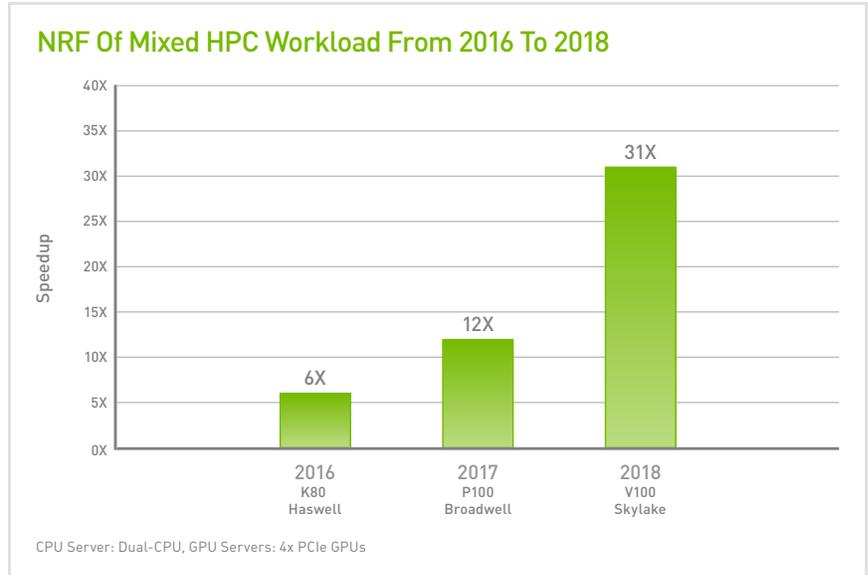


Figure-4

A manufacturing facility is evaluated based on its production or throughput, and the data center is no exception.

Optimizing Data Center 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.

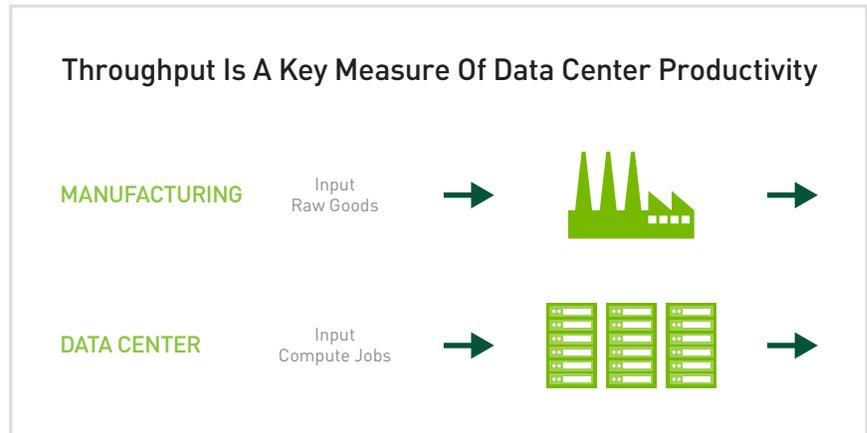


Figure-5

Data center throughput is measured by the amount of work completed in a given period of 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 data center throughput means more scientific discoveries are delivered to researchers every day. In a web services company, 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 a key measure of data center productivity.

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

The accelerated data center has 300 CPU-only nodes and 35 nodes with Tesla V100 accelerators, delivering the same throughput as the CPU-Only data center with 67% 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 V100 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 GPU-accelerated nodes while 300 jobs should run on CPU-only nodes. With a conservative assumption that GPU-enabled jobs run 20X faster on a Tesla V100 node compared to a CPU-only node, only 35 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 335 server nodes.

The accelerated data center delivers the same productivity with 67% less servers, racks, and networking equipment. This translates into tremendous savings in both acquisition cost as well as operation cost due to lower power and smaller 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 a comparison of server node cost.

SERVER COSTS	CPU-ONLY NODE (Dual Socket CPU)	ACCELERATED NODE (4x Tesla V100)
CPU (x2)	\$2,000	\$2,000
GPU (x4)	—	\$8,500
NIC, Memory, Misc. Cost	\$4,000	\$4,000
Core Networking Per Node	\$1,000	\$1,000
Total Node Cost	\$9,000	\$44,000

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

A single CPU socket costs \$2,000 with other necessary components, like NICs and DDR4 memory, and core networking components that cost \$5,000, is a total of \$9,000 per CPU-only node. By adding four Tesla V100 accelerators to the same node design, the node cost now totals \$44,000.

While node cost is higher with GPUs, nodes cannot operate without other data center technologies like storage, software, and services. The typical breakdown of data center cost is as follows:

DATACENTER TECHNOLOGIES	% OF SYSTEM ACQUISITION BUDGET
Servers	70%
Storage	20%
Software and Services	10%
Total Acquisition Cost	100%

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

Using this breakdown, the CPU-Only data center would need to budget \$9M for server nodes and approximately \$4.5M for storage, software, and services. The accelerated data center requires a smaller budget for server nodes and networking, since there are fewer nodes to interconnect. The budget for software and services also decreases due to fewer nodes and smaller overall system costs. However, we'll assume that the budget for storage is the same as the CPU-Only data center.

ACQUISITION COSTS	THE CPU-ONLY DATA	THE ACCELERATED DATA CENTER
CPU Nodes	\$9,000 x 1000 Nodes	\$9,000 x 300 Nodes
Tesla V100 Nodes	—	\$44,000 x 35 Nodes
Servers	\$9.0M	\$4.2M
Storage	\$3.0M	\$3.0M
Software and Services	\$1.5M	\$0.9M
Total Acquisition Cost	\$13.5M	\$8.1M

Table-3: The accelerated data center with Tesla V100 reduces system acquisition cost by 40% compared to the CPU-Only data center.

While the customer needs \$13.5M to deploy the CPU-Only data center, they only need \$8.1M when some of the nodes are accelerated as in the accelerated data center. The end result is acquisition cost savings of \$5.4M (40%) when choosing the accelerated data center.

Figure-6: 40% cost savings with the accelerated data center.

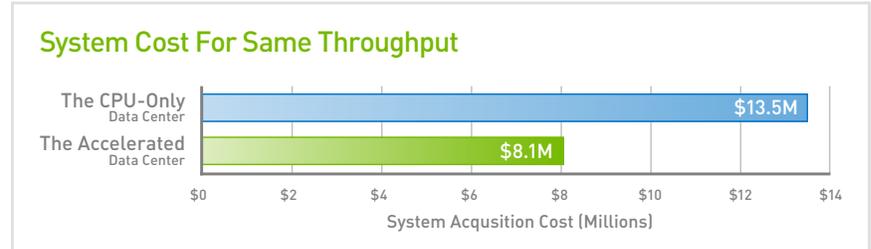


Figure-6

What About Operational Cost?

Once the servers are deployed in the data center they will also generate ongoing operational costs such as power consumption and cooling infrastructure. Today, many data center managers will assign a value that ranges from \$100-\$300/kw/Month to capture these operational costs. For this example we'll assume \$200/kW/Month for the 3-year life of each server.

OPERATIONAL COSTS	THE CPU-ONLY DATA CENTER	THE ACCELERATED DATA CENTER
CPU Nodes	1000 Nodes	300 Nodes
Tesla V100 Nodes	—	35 Nodes
Power Consumption per Node	600W	1,700W
3-year Operational Cost (\$200/kW*36)	\$4.3M	\$1.7M

Table-4: 3-year operational costs, the accelerated data center with Tesla V100 is still saves 60%.

TOTAL COSTS	THE CPU-ONLY DATA CENTER	THE ACCELERATED DATA CENTER
Acquisition Cost	\$13.5M	\$8.1M
3-year Operational Cost (\$200/kW*36)	\$4.3M	\$1.7M
Total Cost	\$17.8M	\$9.8M

Table-5: Total data center cost over 3 years, the accelerated data center with Tesla V100 saves 45%.

Based on this analysis, there is an additional \$2.6M (57%) savings in 3-year operational costs, resulting in a total 3-year cost savings of \$8.0M (45%).

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, unlikely case where the CPU is free for the CPU-Only data center, but full cost for the accelerated data center.

ACQUISITION COSTS	THE CPU-ONLY DATA CENTER (Free CPU)	THE ACCELERATED DATA CENTER
CPU Nodes	\$5,000 x 1000 Nodes	\$9,000 x 300 Nodes
Tesla V100 Nodes	—	\$44,000 x 35 Nodes
Servers	\$5.0M	\$4.2M
Storage	\$3.0M	\$3.0M
Software and Services	\$1.5M	\$0.9M
Total Acquisition Cost	\$9.5M	\$8.1M

TOTAL DATA CENTRE COSTS	THE CPU-ONLY DATA CENTER	THE ACCELERATED DATA CENTER
Acquisition Cost	\$9.5M	\$8.1M
3-year Operational Cost (\$200/kW*36)	\$4.3M	\$1.7M
Total Cost	\$13.8M	\$9.8M

Table-6: Even if CPUs were sold at zero-cost in CPU-Only data center, the accelerated data center with Tesla V100 is still saves 29%.

Node costs are reduced by 44% to \$5,000 and the overall acquisition cost is decreased by 30% to \$9.5 million. However, the CPU-Only data center still requires \$1.4M (17%) more than the accelerated data center. Additionally, when you include the operational costs discussed above, the accelerated data center is \$4.0M lower in overall cost (29%).

Maximizing Budget and Throughput

If a customer has a fixed budget that must be spent, Tesla V100 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 \$5.4M of acquisition cost savings generated by the accelerated data center, the IT manager can also decide to purchase more GPU nodes to introduce new GPU-accelerated applications such as Machine Learning and Deep Learning to the data center workload mix. Let's call this new data center, the "max-accelerated data center", which contains a mix of CPU-only nodes and additional GPU-accelerated nodes to run new AI workloads running over 40X faster on GPU-accelerated servers.

COST	THE CPU-ONLY DATA CENTER	THE ACCELERATED DATA CENTER
CPU Nodes	\$9,000 x 1000 Nodes	\$9,000 x 300 Nodes
Tesla V100 Nodes	—	\$44,000 x 143 Nodes
Servers	\$9.0M	\$9.0M
Storage	\$3.0M	\$3.0M
Software and Services	\$1.5M	\$1.5M
Total Acquisition Cost	\$13.5M	\$13.5M
Total Data Center Throughput	1000 Jobs/Day	5327 Jobs/Day

Table-7: The max-accelerated data center with Tesla V100 delivers over 5x higher throughput compared to the CPU-Only data center.

OPERATIONAL COSTS	THE CPU-ONLY DATA CENTER	THE ACCELERATED DATA CENTER
CPU Nodes	1000 Nodes	300 Nodes
Tesla V100 Nodes	—	143 Nodes
Power Consumption per Node	600W	1,700W
3-year Operational Cost (\$200/kW*36)	\$4.3M	\$3.0M

Table-8: 3-year operational costs, the Max-accelerated data center with Tesla V100 still saves \$1.3M.

IT managers can deploy 108 more GPU nodes with the cost savings made possible by Tesla V100 GPUs. The max-accelerated data center also enjoys lower operational costs due to lower number of nodes. With 108 new GPU nodes producing a 40X throughput increase for AI applications, the max-accelerated data center delivers roughly 5300 jobs per day, an over 5X increase in throughput compared to the CPU-Only data center. That adds up to a 3-year operational cost savings of \$1.3M.

Lower Cost with Acceleration

IT managers care about cost savings. 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 V100, IT managers have slashed their data center costs.

Figure-7: GPU-Accelerated data centers reduce costs significantly and deliver the best throughput per dollar. [Data in chart normalized to the CPU-Only data center]

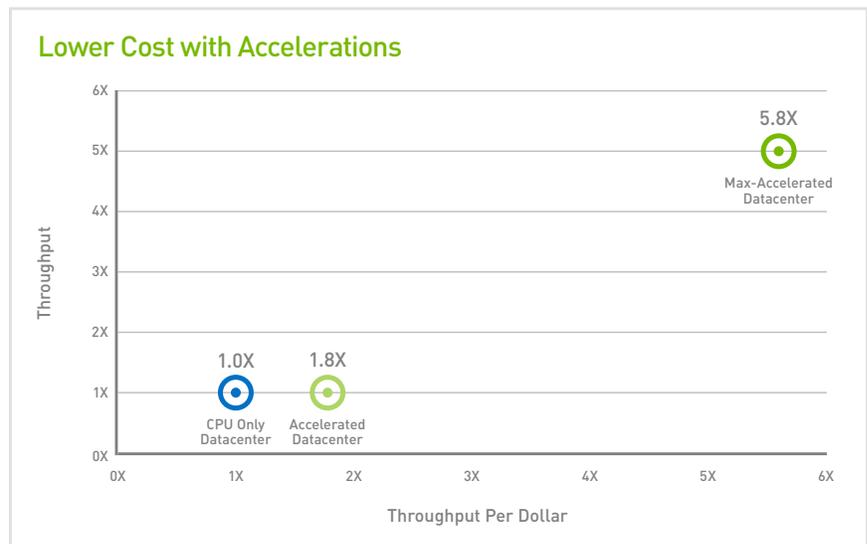


Figure-7

In this whitepaper, we used three examples of data centers that customers deploy today. Compared to the CPU-Only data center, Tesla V100 in the accelerated data center results in 40% reduction in acquisition costs, saving \$4.8 million in an overall budget of \$13.5 million. For customers who want to maximize productivity, Tesla V100 in the max-accelerated data center delivers over 5X increase in overall productivity for the same \$13.5 million acquisition budget.

Democratizing the Supercomputer

Researchers and engineers in HPC, Hyper scale, and Fortune 100 companies work with massive amounts of data as data center costs are growing. GPU-accelerated computing gives them better performance/throughput while also helping manage their costs.

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 finding a cure for cancer or a Fortune 100 company committed to innovation can each afford computing power previously reserved for supercomputing facilities.