# The ROI of GPU-Accelerated Computing

2019

PENGUIN COMPUTING

## Overview

The potential power of artificial intelligence (AI) is drawing attention to how much untapped value sits in the vast quantities of data that organizations have accumulated in recent years.

Unfortunately, that abundance of information has hindered the extraction of business value from data. Processing such large quantities of data has historically required not only a great deal of computing power but also a great deal of time, exactly what organizations don't need when they seek competitive advantage. Previously, many organizations trying to analyze big data relied on costly, central processing unit (CPU) intensive infrastructure. With graphics processing unit (GPU)-accelerated computing, though, the information technology (IT) industry has a new, more effective, more efficient alternative.

While some still associate GPUs with video games, the same computing power that allows movie-quality gaming is also capable of putting an end to sluggish, painful data analytics and replacing it with near real time analysis. A single GPU can offer the performance of hundreds of CPUs for certain workloads, resulting in a profound paradigm shift.

Done properly, adoption of GPU-accelerated computing can offer a significant return on investment (ROI) today and pave the way to gain additional advantage from future technical developments.

Even better, as technology finally enables AI to reach the potential technologists have looked forward to for decades, GPU-accelerated computing helps pave the path for users to gain an increasingly larger advantage over competitors not deploying GPU-accelerated computing

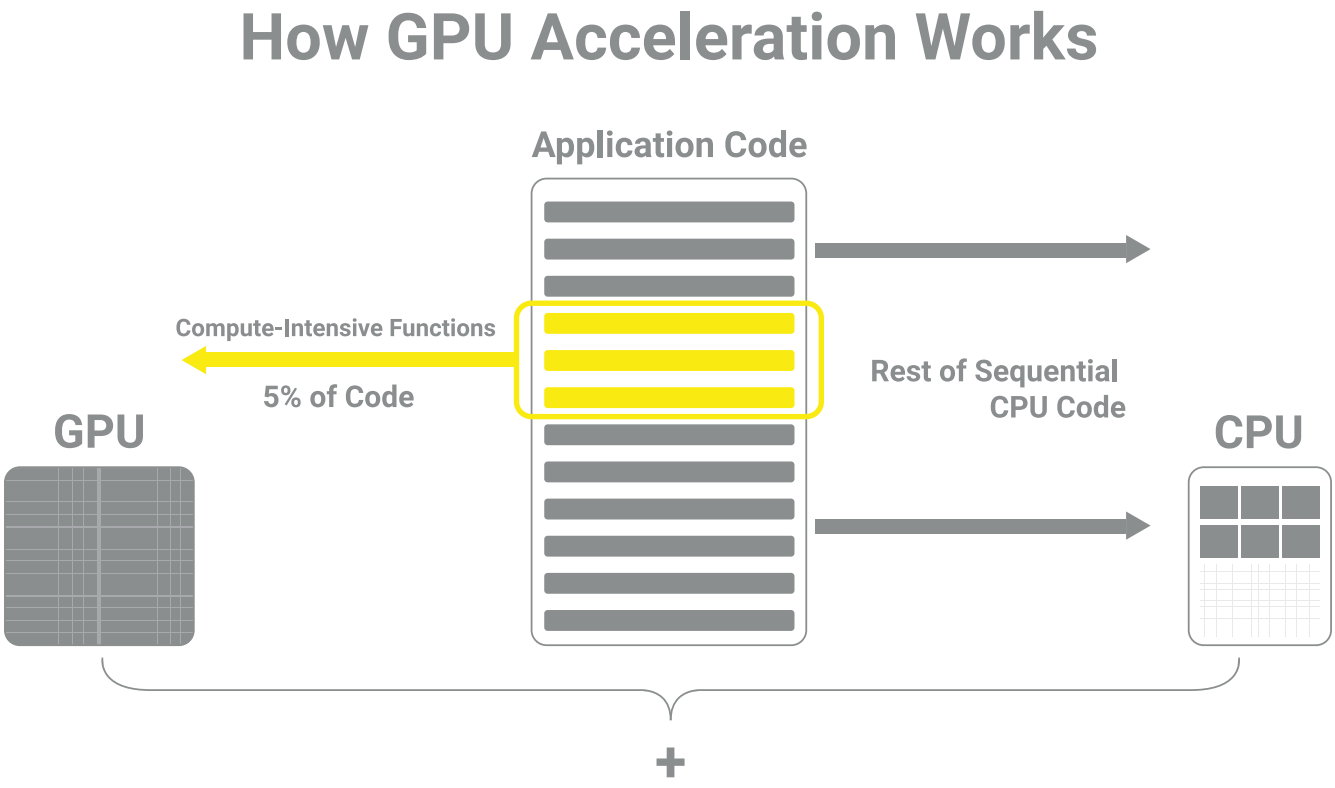
## ROI Versus CPU-Only

The modern CPU is exceptionally well-designed for general compute tasks, such as web surfing, word processing, and hosting databases. For years, computing hardware vendors focused on increasing CPU speed (in order to accomplish more computing in less time). But this has penalties, including higher power use and greater heat generation.

Multicore CPUs have made clock speed a bit less important, as each CPU can be thought of as being made up of multiple, smaller processors (typically, 2 to 8 for desktops and 10 to 32 for servers) that can work on some tasks in parallel. But these multicore CPUs still must support serial activities required for general computing activities, limiting what can be parallelized.

GPU-accelerated computing occurs when a GPU is used in combination with a CPU, with the GPU handling as much of the parallel process application code as possible. The GPU takes the parallel computing approach orders of magnitude beyond the CPU, offering thousands of compute cores. This can accelerate some software by 100x over a CPU alone. Plus, the GPU achieves this acceleration while being more power- and cost-efficient than a CPU[1].

Provided your system design team is experienced with building both CPU and GPU based systems and the storage technologies required for this level of data analytics, the outcome of moving to a GPU-accelerated strategy is superior performance by all measures, faster compute time, and reduced hardware requirements.

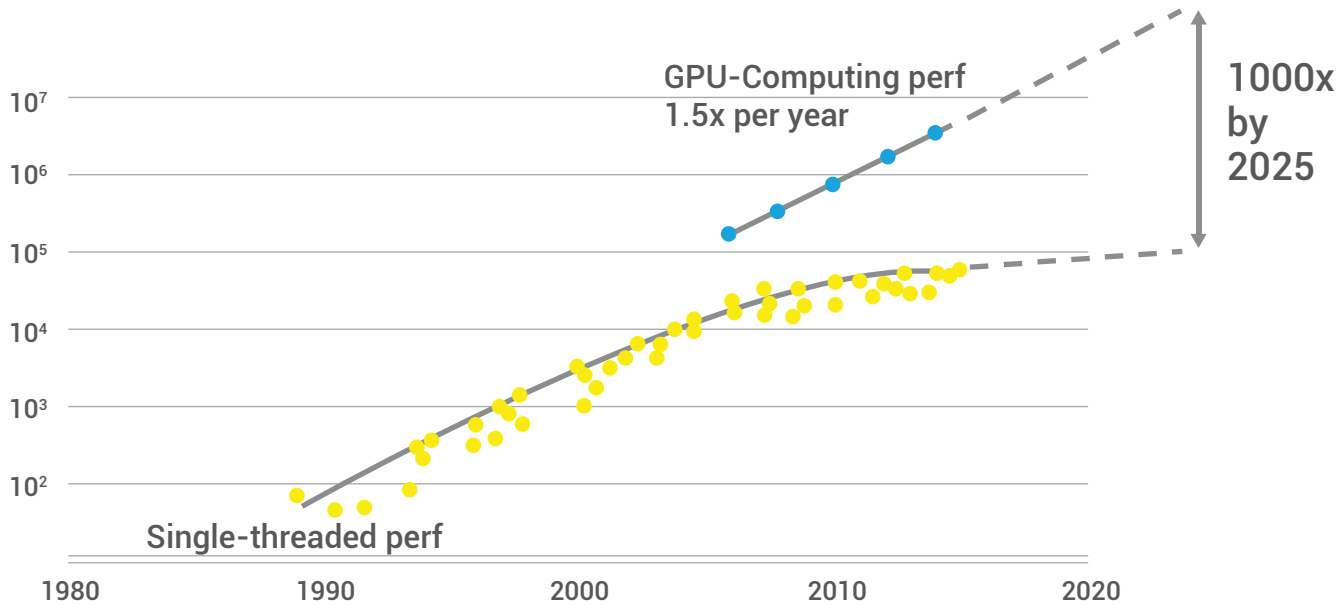


Some processes are inherently sequential and achieve best results with the CPU, and they will always be a part of any complete solution. Many other parallel application processes can benefit from GPU resources. For that reason, using CPUs and GPUs in combination takes advantage of the best of both technologies, tapping the impressive sequential processing power of the latest generation of CPU with the exponential capacity for parallel processing offered by top-performing GPUs.

A reduction in the level of improvement possible in the base CPU design - and increasing adoption of GPUs - leads NVIDIA to predict that GPUs will help provide a 1000X acceleration in compute performance by 2025[2].

This inevitable increase in the reliance on GPUs means that early adopters will enjoy not only greater computing power over time, but have a greater margin of difference over time than competitors who do not migrate to GPU-accelerated computing.

While these predictions may surprise some, consider that the NVIDIA® Tesla® V100, which is available in many data and compute servers, provides the performance of up to 100 CPUs in a single GPU.




40 Years of Microprocessor Trend Data

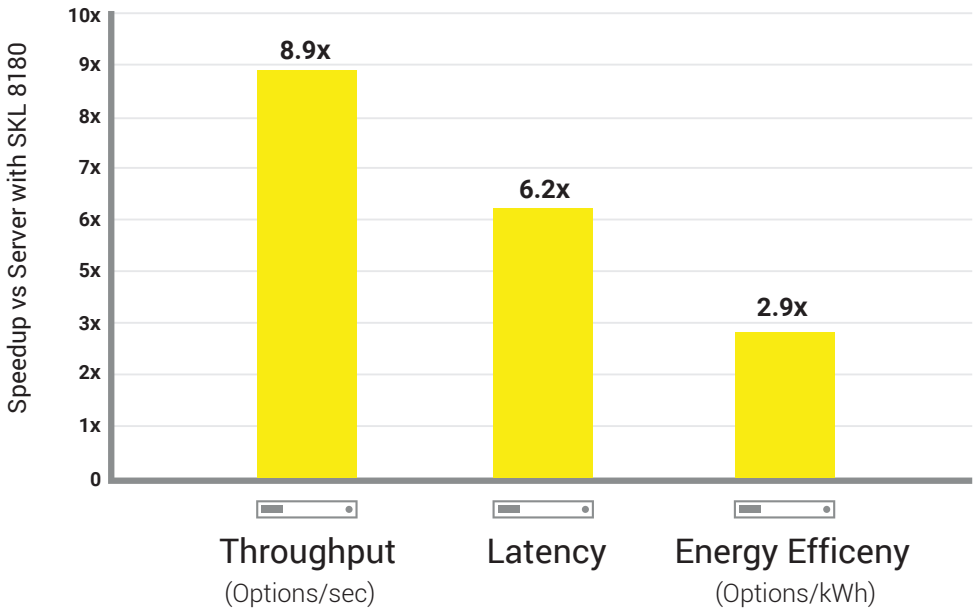
[1] "What's the Difference Between a CPU and a GPU?" Kevin Krewell, 2009

[2] https://www.nvidia.com/en-us/about-nvidia/ai-computing/

## Potential Costs and Cost Avoidance

A GPU-based supercomputer will occupy a much smaller physical footprint than an equivalent CPU-based supercomputer performing the same functions, resulting in potentially significant cost savings as well as positive environmental impact.

Physical infrastructure costs (from power to staffing support for more racks) may also be substantially reduced. Variable costs, such as direct hardware costs and implementation expenses, can also be reduced with a smaller physical deployment, such as that allowed by GPU-accelerated computing. Moving to GPU-accelerated computing also enables organizations to use virtual workstations for heavy compute environments and tasks, further reducing the amount of physical hardware required.




STAC-A2 Benchmark Performance Result

8x V100 GPU Server vs Dual Skylake Platinum 8180 Sever

Provided that you are working with a leading vendor who has significant experience designing both individual GPU-accelerated servers and full systems that can effectively support computing and/or data center requirements, you should get a close estimate of your final operating costs and performance before you even begin building the system.

In addition, due to the inherently flexible nature of GPU programmability, new algorithms are being developed and deployed quickly across a variety of industries. It's generally agreed that the most powerful GPU on the market today is the NVIDIA Tesla V100 accelerator and even it is still being optimized to support the increasing demand from organizations who see the value that GPU-accelerated computing can bring them.

According to NVIDIA, over 550 HPC applications are already GPU optimized[3]. In fact, an independent study by Intersect360 Research shows that 70% of the most popular HPC applications, including 15 of the top 15 have built-in support for GPUs.

Many applications have already been written for GPU-accelerated computing, including the following:

| | | | |
|---|---|---|---|
| **ANSYS DISCOVERY LIVE**<br>ANSYS | **ANSYS MAXWEL**<br>ANSYS | **ANSYS NEXXIM**<br>ANSYS | **ANSYS POLYFLOW**<br>ANSYS |
| **ANSYS WORKBENCH**<br>ANSYS | **HFSS SBR+**<br>ANSYS | **ANSYS FLUENT**<br>ANSYS | **ANSYS MECHANICAL**<br>ANSYS |
| **ARRI DE-BAYERING S...**<br>ARRI | **RAW CONVERTER**<br>ARRI | **ARYA.AI**<br>ARYA.ai | **AWP**<br>AWP |
| **TWINMOTION**<br>Abvent | **SEQNFIND**<br>Accelerated Technology... | **AXRTM**<br>Acceleware | **Storyteller**<br>Accuweather |
| **CINEMATIVE HD**<br>Accuweather | **HTMD**<br>Acellera | **ACEMD**<br>Acellera | **GPUGRID.NET**<br>Acellera Ltd |

[3] "Three Reasons To Deploy NVIDIA Tesla V100 In Your Data Center," 2018

However, if you must migrate code manually, several options exist.

OpenCL, a non-proprietary language option that is supported by AMD, Intel, NVIDIA, Apple, and others, provides portability across hardware platforms. OpenCL is currently the dominant open-source general-purpose GPU computing language and thus has a large developer community to tap into.
All tier 1 vendors support NVIDIA GPUs, which use the CUDA® parallel computing platform and programming model launched in 2006. Fortunately, CUDA is a mature language in its own right,

and is an extension of the C programming language. It includes a software development kit (SDK) and an API that allows the large community of CUDA developers to create parallel algorithms for proprietary GPUs.

Before undertaking code development, some chief information officers (CIOs) will want more specificity on potential ROI. Consider the type of application you want to run. High-performance computing (HPC) code can be classified as either compute-bound or memory-bound.

How can we quantify the HPC performance difference in comparing today's industry-leading GPUs and CPUs? For code that is memory-bound, the GPU delivers a roughly 5X enhanced performance. For compute-bound code, GPUs will boost performance from 5X in the case of code with optimized SSE (Server-sent Event) implementations, to up to 20X or more.[4]

[4] "Top 10 Objections to GPU Computing Reconsidered," By Dr. Vincent Natoli, 2011

## Opportunity Costs

Many firms are actively investing in new technology to further reduce the cost – and time – of analysis so they can gain competitive advantage. That's why many are already shifting to include GPU-accelerated computing as part of their long-term information technology (IT) strategy.

They know that, despite upfront expenses and the necessary preparation, the opportunity cost of maintaining can be very steep. In industries where milliseconds can make the difference between success and failure, businesses that stay with a CPU-only strategy are at a real disadvantage.

Will competitors redirect funds saved by GPU-accelerated computing to a new project you can't afford to invest in at the moment? Will they use the insight extracted from the data to change business strategy in a way that leaves you behind? What will happen to organizations that don't prepare for AI and DL? There is no way to tell, only the certainty that change is inevitable. The question is, is change an obstacle or an opportunity for you?

The increased efficiencies of GPU computing will also likely lead the path for edge computing. As the coming improved networks enable a world of high-speed, low latency inference operations at the edge, the most powerful and power efficient platforms will naturally be selected for these applications. Organizations that are not ready to take advantage of this change will face an even greater hurdle to catch up as competitors move further ahead in this fast-paced space.

## The Pathway to Artificial Intelligence

The technology that makes GPU-accelerated computing desirable for current data analytics also makes it ideal for AI.

NVIDIA invented the first GPU in 1999. Soon, GPUs became commercially important in the video gaming space, where highly parallel graphics processing is required to create realistic, immersive, 3D environments. Linear algebra, that field of mathematics concerned with vectors and matrices, defines the operations that are essential to graphical processing. A graphical pane consists of a matrix (rows and columns) of pixels (picture elements), represented by numerical values, that render an image. The image can be manipulated in various ways through addition and multiplication by scalar values and other matrices. The GPU was designed specifically to address this functionality by efficiently implementing these linear algebra operations and doing so in parallel.

That same capability can also be applied to the emerging field of AI, which has strong potential. In the past, AI, and especially its subdiscipline of deep learning (DL), were slowed in earlier years by the lack of computing power. However, the GPU is changing this.
Deep learning focuses on training artificial neural networks (loosely based on how the brain is understood to process input), and then applying those trained networks to new input to derive an inference. The connection to GPUs comes from the fact that neural networks can be represented as matrices, and the same linear algebra operations that are so important to graphics processing are also used for DL training and inference.

Leading GPU designers are optimizing to support AI even further. For example, the NVIDIA Tesla includes new streaming multiprocessor (SM) architecture that is optimized for DL[5]. Also, with 640 tensor cores (providing tensor processing unit capabilities), Tesla V100 has earned the distinction of the world's first GPU to break the 100 teraFLOPS barrier for deep learning performance. In hyperscale server racks, the Tesla V100 delivers 47X higher inference performance than a CPU server[6].

This continuous improvement in GPU technology and the massive stores of data now available to improve algorithms will allow organizations already familiar with GPU-accelerated computing to more smoothly transition into AI.

There are relatively few system designers who have experience with both GPU-accelerated computing and AI, but those that do can help you build a system that takes full advantage of the capabilities of GPU acceleration. One way to differentiate among providers and understand their skill sets is to look for those who have experience with the full set of products from leading GPU vendors such as NVIDIA and its Tesla V100 GPU as well as the AI-focused NVIDIA® DGX™ platform.

[5] "NVIDIA Tesla V100 GPU Architecture," August 2017

[6] https://www.nvidia.com/en-us/data-center/tesla-v100/

**About Penguin Computing**

Penguin Computing specializes in helping startups, Fortune 500, government, and academic organizations with innovative high-performance computing (HPC) on-premise and in the cloud, artificial intelligence (AI), networking, and storage technologies, coupled with leading-edge design, implementation, hosting and managed services—including sys-admin and storage-as-a-service—and highly rated customer support. In the course of 20 years, Penguin Computing has served more than 2,500 customers in 40 countries across eight major vertical markets.

Because of its success building GPU-accelerated computing systems (including growing it's NVIDIA GPU business 778% over the past four years) Penguin Computing was awarded the Americas 2017 NVIDIA Partner Network (NPN) High Performance Computing (HPC) Partner of the Year Award.