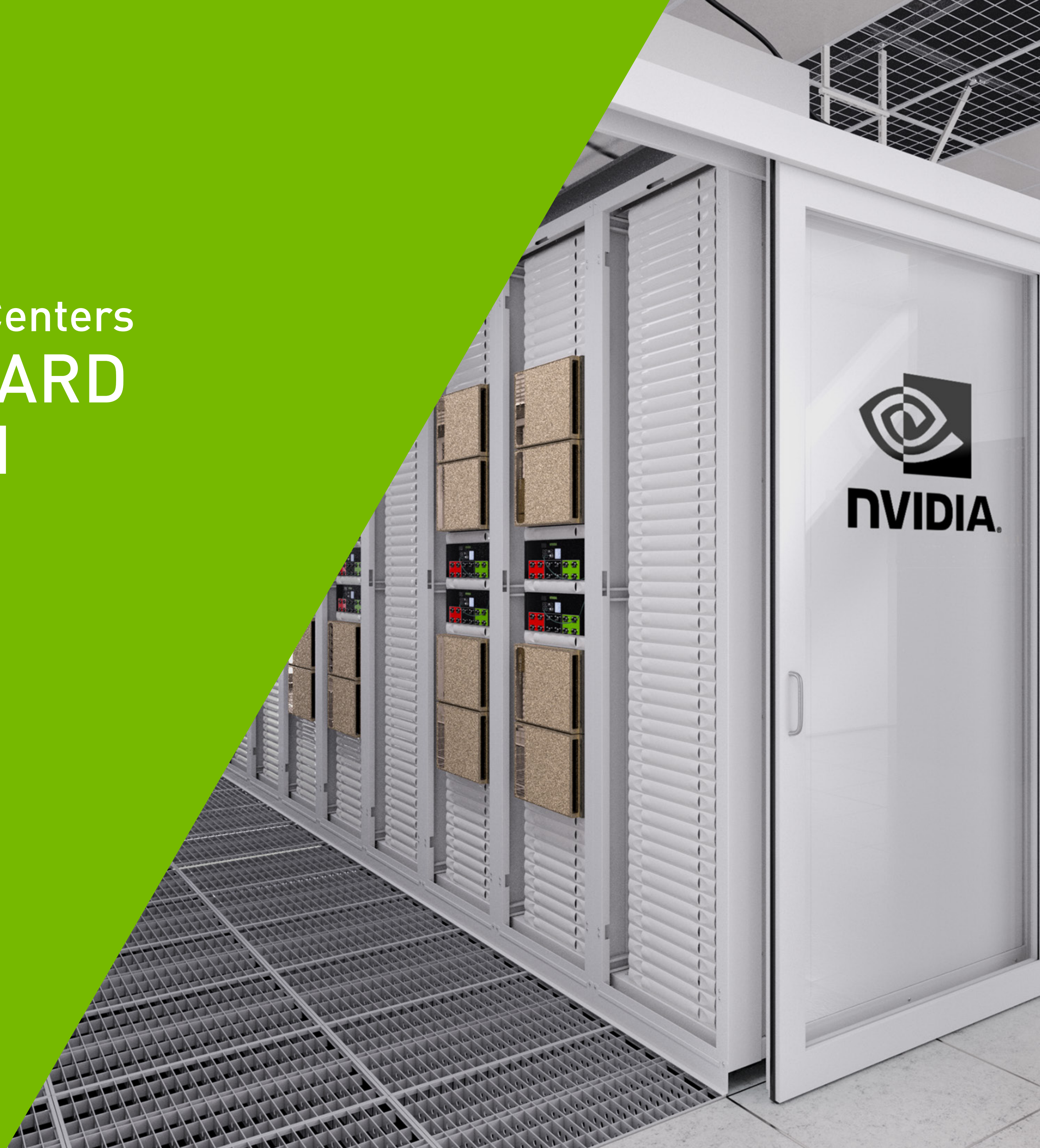
Accelerated Data Centers

FAST-FORWARD INNOVATION



CAPITALIZING ON KEY TRENDS AND DISRUPTIVE OPPORTUNITIES

Uncertainty is never the friend of any CIO, and times have become even more uncertain. Managing a company's information and compute technology is challenging even in a world that hasn't been disrupted by a global pandemic. In 2020, the level of complexity rose in tandem with the data volume and management issues created by a new and almost entirely remote workforce. In 2021, many of these issues will continue to press CIOs as they prepare post-pandemic data center strategies that consider exponential advances in software and hardware speeds, new technologies like 5G, data processing unit (DPU)-based computing, new and demanding workloads, and consumer-like expectations for business services.

Around the world, business leaders are seizing this opportunity to transform their organization, leveraging technology to solve the world's most challenging problems and positioning themselves to take advantage of next-generation hardware and software products and services. But to be successful, enterprises need a modern, cohesive computing infrastructure—from the data center to the edge—that provides the functionality, performance, security, and scalability to run next-generation workloads.

This is an introduction to the key principles of the software-defined, hardware-accelerated data center of the future, created for those who are ready to pursue their digital transformation journey, explore a new architecture for hybrid cloud, and use breakthrough applications to drive their business forward.

> "AI and machine learning have quickly expanded from research labs to data centers in companies across virtually every industry and geography."
>
> — Jensen Huang, Co-founder and CEO, NVIDIA

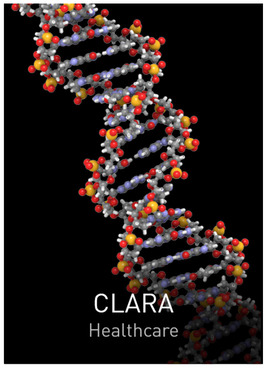
THE MODERN DATA CENTER: KEY TO FASTER AND DEEPER INNOVATION

Data centers are now the engines of the new economy, generating hundreds of billions in annual revenue for all kinds of businesses. They power every digital activity, now accelerated by the pandemic and remote work, and are the key to solving some of the world's greatest challenges.

The explosion of data has been happening for decades, and the global pandemic has only accelerated the digitization of society and its attendant onslaught of data. Business leaders see this as both an opportunity and an obstacle. The vast stores of data being created hold incredible potential—but they need to be mined to create actionable insights. And increasing consumer-like expectations for business services have placed extra pressure on enterprises looking to transform.

In 2020, more people worked from home and students transitioned to remote learning. Society relied heavily on technology to shop, visit doctors, and more, putting strains on services and revealing shortcomings in user experiences.

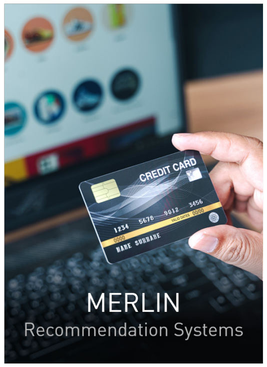
It's a data center-scale and full-stack challenge. Although data centers have the ability to tackle difficult problems and generate enormous value, solving these problems in finance, retail, healthcare, and many other industries requires CIOs to modernize their IT infrastructure. The exponential advances in software and hardware are key to enterprise innovation.

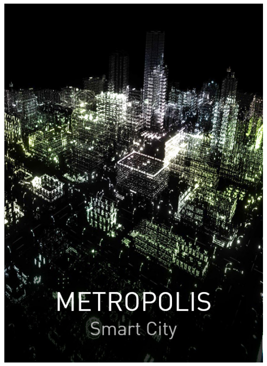

CLARA
Healthcare

DRIVE
Autonomous Vehicles

RIVA
Conversational AI

ISAAC
Robotics


MERLIN
Recommendation Systems


METROPOLIS
Smart City

A POWERFUL COMBINATION: ACCELERATED COMPUTING AND AI SOFTWARE

Advances in software enable next-generation applications like conversational AI, advanced audio and video streaming, recommender systems, and more to create engaging e-commerce experiences and opportunities for growth. For example, on some of the world's largest online commerce sites, recommender systems account for as much as 30 percent of revenue. Just a 1 percent improvement in the relevance of recommendations can translate into billions of dollars in revenue. And that's only one potential area of improvement.

With the increase in remote work and video conferencing, accelerated computing platforms deliver powerful features that enhance the overall user experience and ability to collaborate. NVIDIA Maxine™ is a fully accelerated platform SDK for building and deploying AI-powered video conferencing features that use state-of-the-art models. Maxine provides a new AI framework to improve the video conferencing experience through a number of methods, such as modifying the focus of users to provide a more person-to-person interaction, providing active noise cancellation, real-time translation and subtitles, and even an avatar with natural facial expressions.

AI-driven services in speech, vision, and language present a revolutionary path for personalized, natural conversation, but they face strict accuracy and latency requirements for real-time interactivity. With the NVIDIA Riva conversational AI platform, developers can quickly build and deploy state-of-the-art AI services to power applications across a single unified architecture, delivering highly accurate, low-latency systems with little upfront investment.

CPU AND GPU COMPUTE AND WHAT THAT MEANS FOR ENTERPRISES

Today's enterprise data centers run a new wave of modern, accelerated applications, such as AI, machine learning, and data analytics, along with existing legacy applications. But this creates complications when using traditional data centers, storage architectures, and security technologies—often resulting in siloed infrastructure.

The data size and compute size of these new applications—as well as the problems they're trying to solve—are too large for CPU-only infrastructure. When accelerated computing software and systems are integrated, the same infrastructure can do much more work.

Originally created to redefine PC graphics, GPUs and their accelerated computing capabilities have since ignited a worldwide AI boom. They've become a key part of **modern supercomputing**. They've been woven into sprawling new **hyperscale data centers**. And they've become accelerators, speeding up many new applications, from encryption to networking to analytics and AI.

While GPUs are now about a lot more than the PCs in which they first appeared, they remain anchored in an established approach called parallel computing. And that's what makes GPUs so powerful. CPUs, to be sure, remain essential. Fast and versatile, CPUs linearly race through a series of tasks that require a lot of interactivity, for example, calling up information from a hard drive in response to a user's keystrokes. By contrast, GPUs break complex problems into thousands or millions of separate tasks and work them out at once—in parallel—to accelerate compute.

Architecturally, the CPU is composed of just a few cores with large cache memory that can handle a few software threads at a time. In contrast, a GPU is composed of hundreds of cores that can handle thousands of threads simultaneously. Another factor making all that power accessible is the parallel computing platform **CUDA®**. First released in 2007, it lets coders take advantage of the computing power of GPUs for general-purpose processing by inserting a few simple commands into their code.

In the **automotive industry**, accelerated GPU computing offers many benefits. It provides unmatched image capabilities for design and simulation, but it's also key to creating self-driving vehicles able to learn from—and adapt to—a vast number of different real-world scenarios.

In **healthcare and life sciences**, GPU computing is also ideal for accelerating the analysis of those images through deep learning. Deep learning can process medical data and help turn it into new insights and capabilities.

Innovative retailers and disruptive startups are using AI and accelerated computing to streamline logistics and store operations, prevent shrinkage, and deliver better shopping experiences both in stores and online.

In short, GPU computing has become essential, accelerating more and more areas where computing horsepower can make a significant impact.

DISRUPTIVE EFFICIENCY BY DESIGN

Servers, racks, and cooling units are starting to look vastly different thanks to GPU-accelerated computing, smart network interface cards (NICs), and AI-enabled software—all designed to deliver the modern accelerated data center.

Just a few years ago, the majority of the world's data centers relied on independent, physical servers tethered to arrays of hard disk drives (built on designs originating in the 1970s) with a 1G or 10G switch at the top of the data center rack. All applications had to fit into this fixed hardware model whether they did so efficiently or not. Most applications did not, and frequently, expensive computing and storage resources were only utilized at 15–35 percent. Systems configured to run just one application are expensive and rigid, as they can't be used for anything else and quickly become obsolete.

Applications today are far more fluid and dynamic and require varying amounts of compute, memory, storage, and networking.

This impacts all industries. Processing a credit card transaction may require looking up a name (only a few data bytes) in an enormous database—a tiny compute load but a huge disk storage job. Engineering simulations are compute and memory intensive but may require very little storage capacity. Movie streaming is networking intensive but not demanding on compute and memory resources. Many workloads have unique computing characteristics, but all need to share the same data center resources.

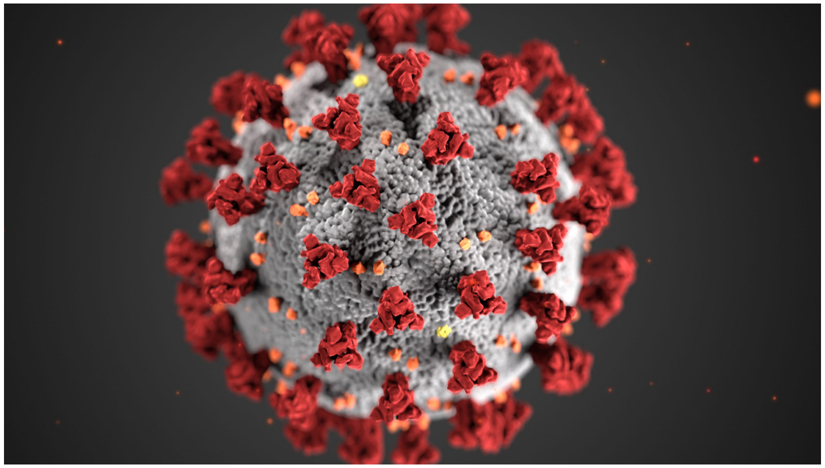
The new accelerated computing model strives to make all the compute-storage-networking resources dynamically configurable or "composable" to the specific needs of the applications running on it. The industry has begun transforming from dedicating specific systems for each application to a model of running any application on any system, as needed, with just-in-time provisioning.

FAST-FORWARD BUSINESS INNOVATION

Investing in accelerated data centers can fast-forward business innovation and monetize services to deliver new revenue streams. As AI continues to disrupt traditional business models, its impact can be felt by organizations across industries, from enhancing customer engagements to enabling scientific breakthroughs to driving more accurate predictive maintenance.

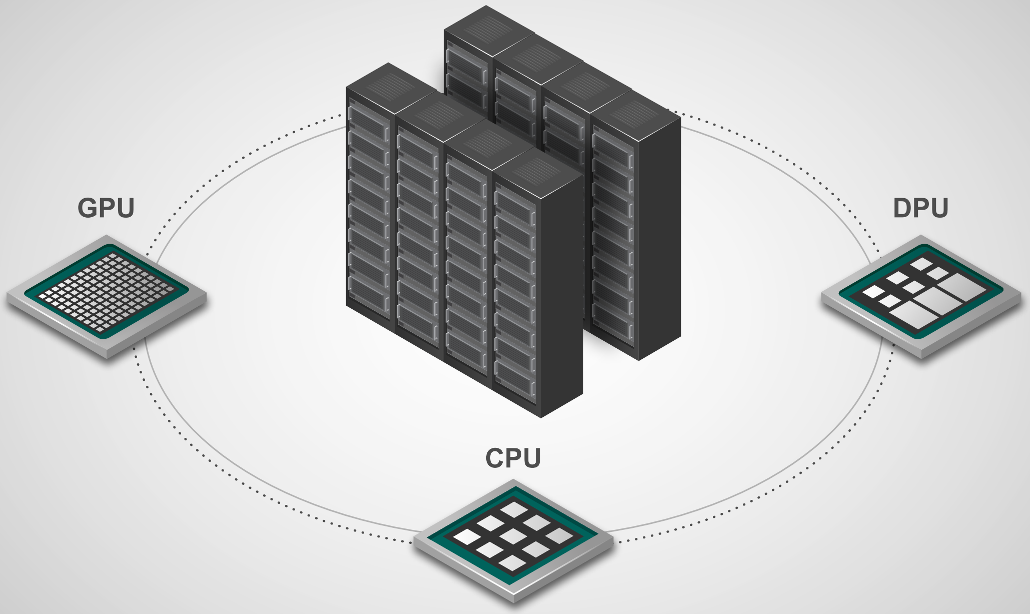
Consider **drug discovery** as a model, which has been critical to responding to the COVID-19 crisis. To help save lives and economies around the world, the healthcare industry has sped up research and operations, creating solutions in record time. It's taken less than a year to bring vaccines and treatments to patients—processes that often took a decade or more in the past.

Enterprises know they must embrace this transformation and speed or risk losing out to competitors who get there first. For example, 84 percent of surveyed executives fear missing their growth objectives if they don't scale their artificial intelligence efforts. Yet, 76 percent of them cite their struggle to achieve this goal.[1] Among other factors, they face challenges with using existing data center infrastructure to power these applications.



> "NVIDIA is putting its best technology to use in fighting COVID-19"
> — Forbes

[1] Accenture: "AI: Built to Scale, From Experimental to Exponential." 2019.



THE NVIDIA SOLUTION

The NVIDIA accelerated computing platform provides a way for customers to run diverse traditional and modern applications on a single high-performance, cost-effective, and scalable infrastructure. It brings together compute acceleration and high-speed secure networking in enterprise data center servers, built and sold by NVIDIA partners. This platform is supported by a vast suite of software that enables users to become productive immediately and can be easily integrated into existing industry-standard IT and DevOps frameworks, allowing IT to manage, deploy, operate, and monitor their infrastructure.

The NVIDIA accelerated computing platform is increasingly being used to accelerate next-generation workloads. To boost performance further, a software-defined, hardware-accelerated stack was developed, complete with a programmable network. The platform's data processing unit (DPU) accelerates networking, storage, security, and management applications.

DATA PROCESSING UNITS (DPUs): BUILD SECURE, PROGRAMMABLE, SOFTWARE-DEFINED DATA CENTERS

NVIDIA has developed a system on a chip (SoC) called a data processing unit (DPU) to offload data management and security functions, which have increasingly become software functions, from the main server CPU to an intelligent NIC. The DPU is paired with the NVIDIA Data Center Infrastructure on a Chip (DOCA) SDK. The combination delivers a programmable data center platform for the software-defined data center.

A DPU combines three key elements:

> An industry-standard, high-performance, software-programmable, multi-core CPU, typically based on the widely used Arm architecture, tightly coupled to the other SoC components

> A high-performance network interface capable of parsing, processing, and efficiently transferring data at line rate to GPUs and CPUs

> A rich set of flexible and programmable acceleration engines that offload and improve application performance for AI and machine learning, security, telecommunications, storage, and more

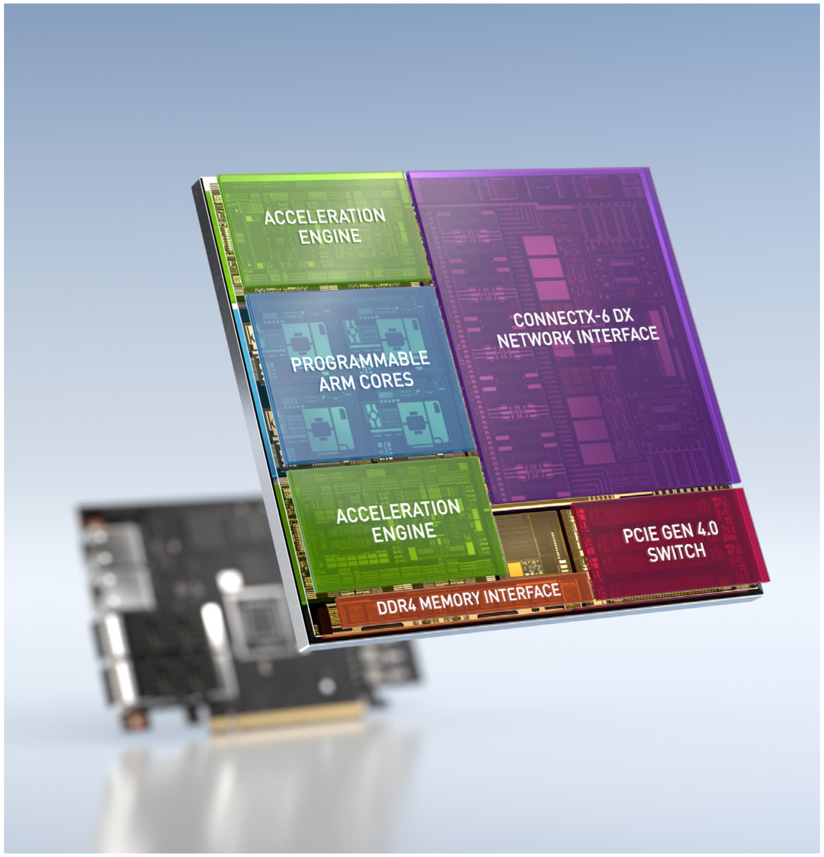
These DPU capabilities are critical to enabling isolated, bare-metal, cloud-native computing that will define the next generation of cloud-scale computing.

ACCELERATING CHANGE IN THE DATA CENTER

In the modern enterprise data center, application performance optimization is at the forefront of most transformation initiatives. Accelerated applications will be offloaded from CPUs to GPUs, and SmartNICs based on programmable DPUs will accelerate data center infrastructure services such as networking, storage, and security. The addition of these GPUs and DPUs will deliver expanded application acceleration to all enterprise workloads and provide an extra layer of security. Virtualization and workload scalability will be faster, while CPUs will be freed up to run traditional applications and services.

The new data center architecture leverages software-defined, hardware-accelerated virtualization, which provides manageability, security, and flexibility. It also supports containers, which ease adoption and management of AI frameworks.

End-user spending on global data center infrastructure is projected to climb to $200 billion in 2021, up 6 percent from 2020, according to the latest forecast from Gartner. GPUs, DPUs, and relevant software will be a part of this new, modern data center and will result in significant performance, efficiency, and security improvements, helping businesses to fast-forward innovation and create new revenue streams.



NVIDIA AND VMWARE ARE ENABLING NEXT-GEN HYBRID CLOUD DEPLOYMENTS

NVIDIA and VMware are partnering to deliver an end-to-end enterprise platform for AI, expanding and simplifying the adoption of AI in the enterprise. This new platform integrates the NVIDIA AI Enterprise software suite with VMware vSphere, VMware Cloud Foundation, and VMware Tanzu, making it easier to deploy and manage AI. Every industry—from financial services to healthcare to manufacturing—will be able to deploy AI workloads using containers and virtual machines on the same platform.

Turning data into insights through data science and then putting that insight into production in an AI model is a journey with many steps. By intelligently accelerating these steps, enterprises can put their data to work to transform their industry.

A PATH TO AI THAT DELIVERS BREAKTHROUGH PERFORMANCE AT BREAKNECK SPEEDS

Whether creating quality customer experiences, delivering better patient outcomes, or streamlining the supply chain, enterprises need infrastructure that can deliver AI-powered insights. The NVIDIA accelerated computing platform delivers the world's leading solutions for enterprise AI infrastructure.

And for the fastest path to AI innovation and industry-proven results, NVIDIA has developed a first-of-its-kind infrastructure solution that enables any organization to operationalize AI at scale: NVIDIA DGX SuperPOD™ Solution for Enterprise.

This turnkey solution—built on the NVIDIA DGX SuperPOD reference architecture—gives businesses leadership-class infrastructure that can be rapidly and confidently deployed. It's a full-stack platform that includes industry-leading computing, storage, networking, infrastructure management, and data science workflow tools optimized to work together and provide maximum performance at scale, along with a white-glove implementation service and end-to-end-lifecycle support that ensure smooth deployment and operation. With its foundation in the NVIDIA accelerated computing platform, the NVIDIA DGX SuperPOD Solution for Enterprise is the turnkey hardware, software, and services offering that removes the guesswork from building and deploying AI infrastructure.

To learn more about the NVIDIA accelerated computing platform, visit **www.nvidia.com/data-center**

To learn more about NVIDIA DGX SuperPOD Solution for Enterprise, visit **www.nvidia.com/dgx-superpod**