The Genius of NVIDIA Tech Strategy Beyond AI Innovation

Did you know that NVIDIA has an incredible market value of 1 trillion dollars in 2023?

On May 30, 2023, it became a member of the trillion-dollar club, joining elite companies such as Google's Alphabet, Apple, Microsoft, and Amazon.

It's the first chip company to accomplish this milestone, with this success coming from Wall Street's growing fascination with AI as more and more tech companies rush to integrate generative AI tools into their products.

What does NVIDIA have to do with generative AI in the first place?

Well, NVIDIA is a chip manufacturer, and it supplies chips to major companies like Microsoft and Google. These chips assist in training large language models for generative AI, which power advanced chatbots like **ChatGPT** and **Google Bard**.

So, how did NVIDIA become such a significant player in the tech space? How did it become so successful?

This is the story of how NVIDIA became a trillion-dollar company. -

[Main Content]

Chapter 1: NVIDIA Founders' Story and Company History (How NVIDIA *got here)*

Let's go back to the start, a relaxed evening in April 1993. At a modest Denny's roadside diner in East San Jose, three innovators set out on a journey that would transform computing history forever. Jensen Huang, an electrical engineer full of ideas, joined forces with Chris Malachowsky, who's passionate about tech from Sun Microsystems, and Curtis Priem, an experienced graphics chip designer.

Sitting over coffee and sharing their aspirations, they brought NVIDIA to life.

Their mission? To lead the way in accelerated computing, a world where graphics-based processing could tackle challenges beyond the reach of traditional computing.

Why? Because they saw opportunities where others saw limits. They realized that video games weren't just popular entertainment; they presented intricate challenges that could push computational boundaries in the future.

What would you do with $40,000 dollars? -

A luxury tour, a high-spec gaming PC ? or a Shaddy fixer-upper?

Well, If you're determined to do something, ***nothing can stop you***. -

If you need inspiring words, don't do it. - Elon Musk, CEO of SpaceX and Tesla

The NVIDIA H100 tensor core GPU, if we think about it today, costs around $35,000. However, when NVIDIA started, they only had $40,000 in their pockets, but they leveraged this humor budget, and of course, with their strong belief in their vision, they laid the foundation for what NVIDIA would become in the future. -

NVIDIA started in a tiny office in Sunnyvale, California. At first, they just used the letter "N" and letter "V" for their files, short for "next version." But when they needed a real name, they found "invidia," which means "envy" in Latin, and it totally matched their ambition. Hence the company name 'NVIDIA'

Chapter 2: Early Success

NVIDIA's path to success was paved with strategic product launches and advancements in graphics technology. A key moment arrived in 1998 when they launched the RIVA TNT, cementing their status as creators of top-notch graphics adapters. -

Yet, it was 1999 that truly changed the game. On January 22, they went public, marking a significant milestone. Shortly after, they released the **GeForce 256 (NV10)**, a game-changing product that revolutionized consumer-level 3D hardware. This was the first to integrate **on-board transformation and lighting (T&L)**, setting a new standard in the industry. -

Not long after, the company won a contract to develop graphics hardware for Microsoft's Xbox game console, earning NVIDIA a substantial $200 million advance. Still, NVIDIA's engineers continued to deliver, with products like the GeForce2 GTS in 2000 and the acquisition of 3dfx's intellectual assets in 2002, further solidifying the compa `ny's market position.

NVIDIA's success story didn't stop there. In 2004, they assisted Sony in designing the graphics processor for the PlayStation 3. The company also expanded its horizons by diversifying into gaming, automotive electronics, and even mobile devices.

NVIDIA's commitment to innovation was evident with the launch of the Tegra series tailored for mobile devices in 2011. Notably, collaborations with major players like Toyota and Baidu underscored their significant role in artificial intelligence and self-driving vehicles.

In 2016, NVIDIA launched the GeForce 10 series, featuring the Pascal microarchitecture and simultaneous **multi-projection (SMP)** support, which enhanced multi-monitor and virtual reality rendering. NVIDIA jumped into AI by creating open-source ventilators during the global pandemic. This showed how much they cared about tackling critical healthcare problems. -

NVIDIA made strategic moves, like buying Mellanox Technologies in 2019, to grow in high-performance computing. A year later, they revealed the Ampere microarchitecture and bought Arm from SoftBank for $40 billion, showing their drive to dominate the industry. Fast forward to 2023, they teamed up with Getty Images to roll out features leveraging Generative AI, those features were powered by NVIDIA's state-of-the-art foundation model - Edify.

NVIDIA's financial journey ran parallel to its technological strides. In 2020, despite a 6.8% decline in annual revenue to $10.9 billion, NVIDIA still reported a profit of $2.8 billion. This dip was offset by the surge in demand due to the pandemic, leading to Q2 2020 sales of $3.9 billion, a 50% rise from the previous year. The effects of the pandemic accelerated the adoption of NVIDIA's technologies, especially laptops and virtual workstations, enabling remote work, crypto mining, and virtual collaboration.

Throughout these shifts, NVIDIA showcased resilience and adaptability. But what's the secret behind their continuous growth and how they become dominant in the market?

Chapter 3:

[A dramatic turn -> add Myself in a side window and narrate the following section]

NVIDIA is making waves in the tech world by seamlessly integrating its GPU and DPU hardware accelerators with Kubernetes, one of the most popular cloud-native platforms for managing and automating containerized apps.

The combination of Kubernetes and GPUs delivers unmatched scale for AI workloads. Kubernetes is designed to scale the infrastructure by pooling the compute resources from all the nodes of the cluster. GPUs are used for massive parallelism required for training and inference of complex deep learning models. -

Currently, 48% of organizations use Kubernetes for AI/ML workloads, and the demand for such workloads also drives usage patterns on Kubernetes more and more significantly.

You know, unlike CPUs that handle general computing tasks, GPUs specialize in graphics rendering and parallel processing, which is why they excel at tasks such as machine learning, deep learning, scientific simulations, as well as AI model training and inferencing, which is in high demand these days.

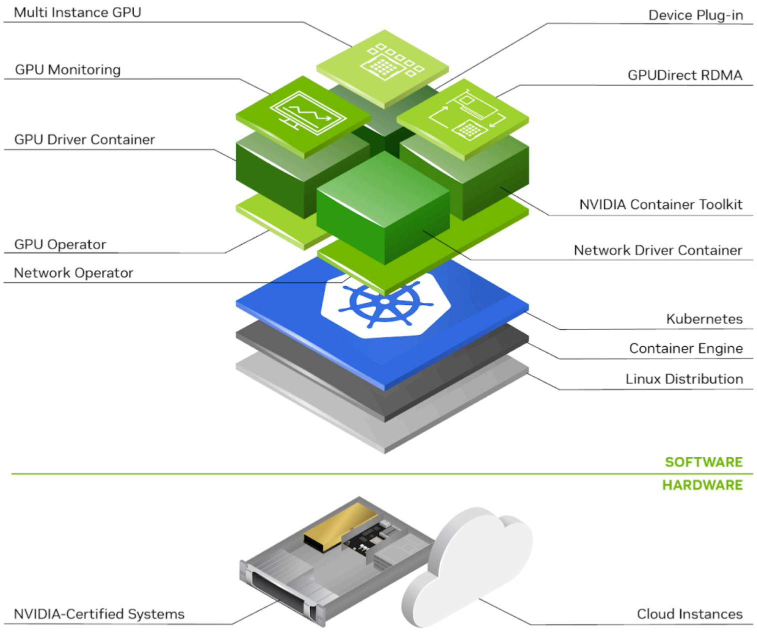


Figure - Kubernetes Integrations

By combining the scalability of Kubernetes with the parallel processing power of GPUs, NVIDIA has created an environment where AI workloads can thrive. Key platforms like NVIDIA DGX and EGX are already using Kubernetes as their orchestration layer, and NVIDIA is actively collaborating with platform vendors (*Microsoft Azure, Amazon Web Services, Google Cloud Platform, Oracle Cloud, and more*) to smoothly integrate its cloud-native GPU infrastructure with Kubernetes.

At its core of this transformation is the **NVIDIA Container Toolkit**, an extension to Docker Engine and Containerd. This toolkit simplifies the developer experience by exposing GPUs to containers, reducing the complexities associated with configuring the GPU software stack. Developers can pull the CUDA container image without the need to install the entire stack.

By the way, CUDA is a parallel computing platform and programming model developed by NVIDIA for general computing on GPUs.

It's highly portable and scalable across various environments, from high-end GPUs to edge devices like Jetson Nano.

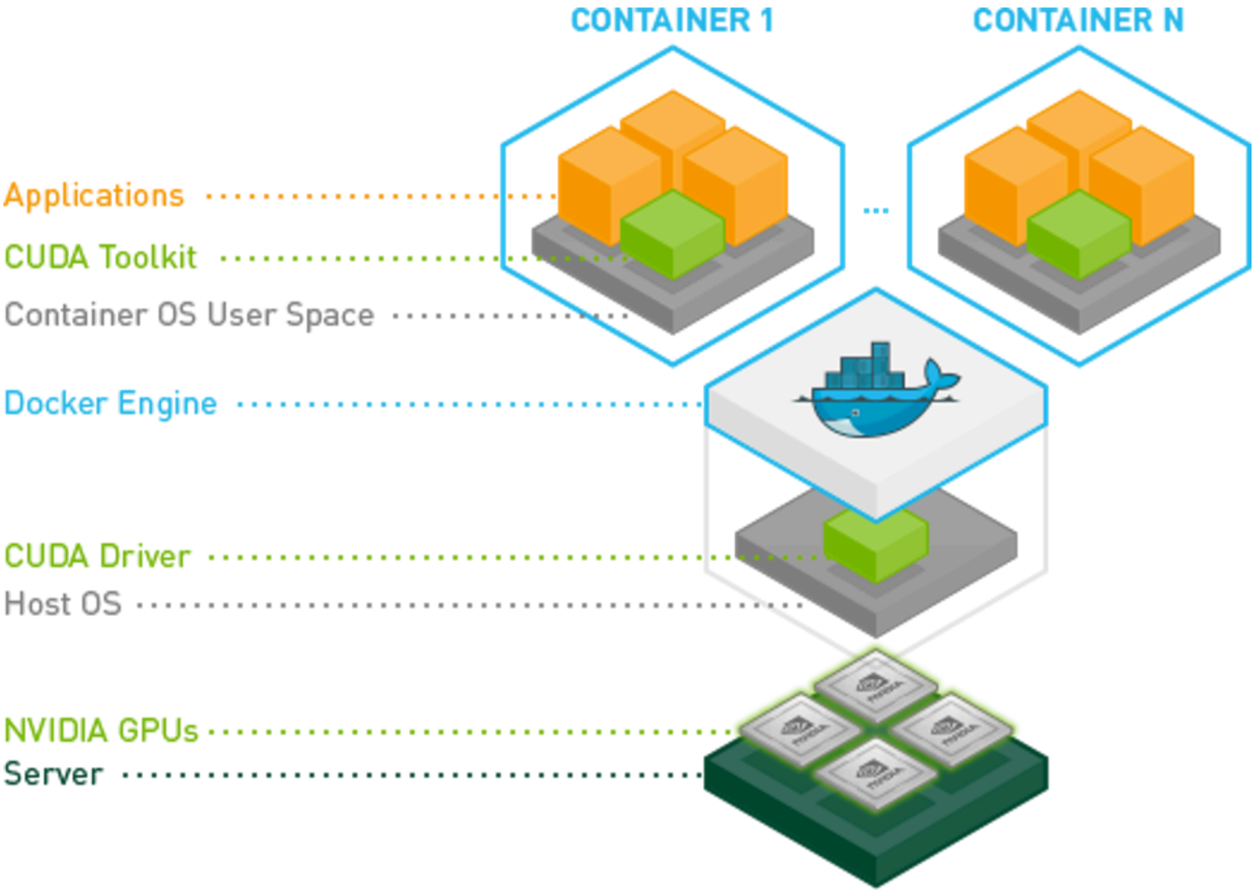


Figure - NVIDIA Container Toolkit

NVIDIA GPU Cloud (NGC) serves as a central hub for cloud-native, GPU-optimized AI resources. NGC also has its own container image registry, called NVCR - which stands for NVIDIA Container Registry. Just like container registries such as **Docker Hub**, it operates a marketplace for container images, including pre-trained models, Jupyter notebooks, popular helm charts such as GPU operator, and Jarvis, which is the NVIDIA Platform for Conversational AI. This helps streamline the workflow and makes it easier for developers to build AI applications.

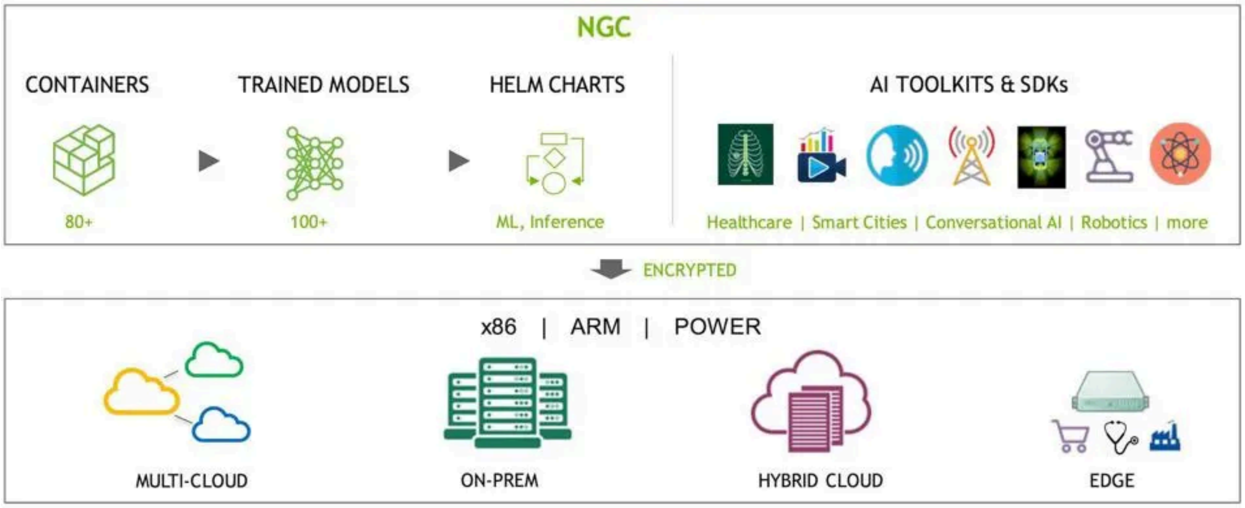


Figure - NGC

Kubernetes is great for optimizing resources and scaling worker nodes efficiently. However, it doesn't natively support scheduling and managing GPUs. To schedule GPUs in Kubernetes, we need something called a device plugin. NVIDIA, along with other providers like AMD and Intel, offers a device plugin for Kubernetes. This NVIDIA device plugin is a part of the Kubernetes device plugin framework. It lets the **kubelet agent** know how many GPUs are available on each node of the cluster and monitors the health of those GPUs.

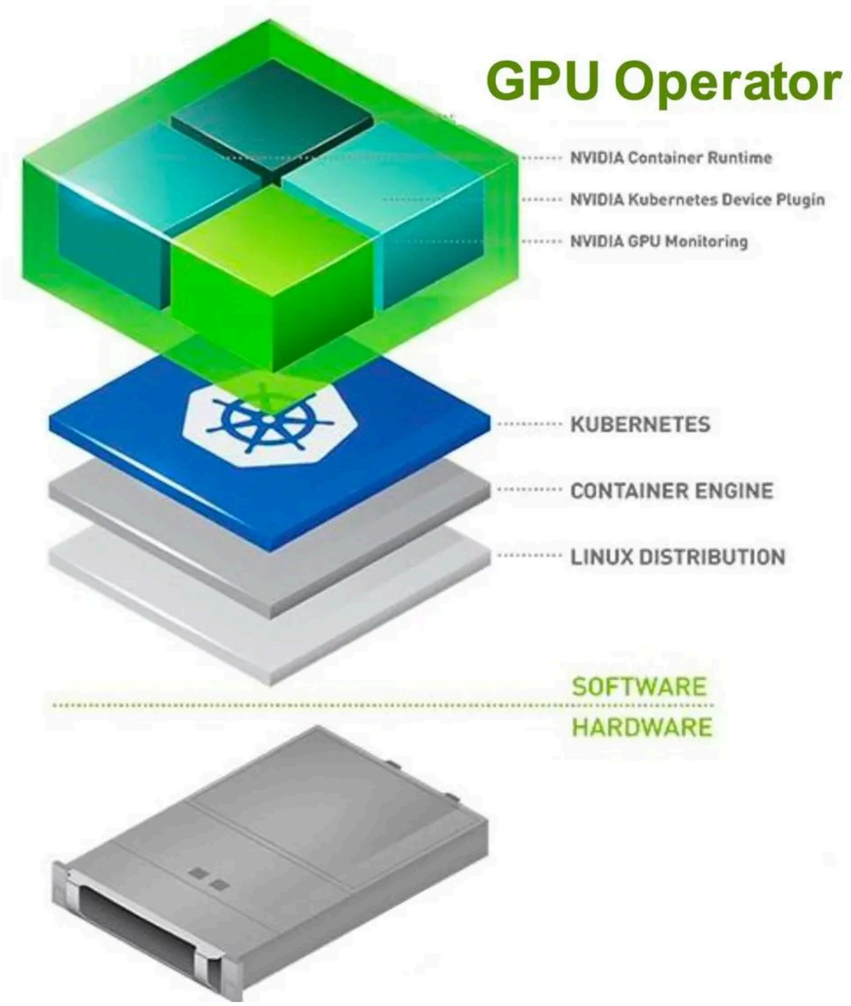


Figure - GPU Operator (NVIDIA DeepOps)

NVIDIA even took a step further by building a GPU Operator to simplify GPU management in Kubernetes. You see, to provision GPU worker nodes in a Kubernetes cluster, we always need some crucial components: the driver, container runtime, device plugin, and monitoring component.

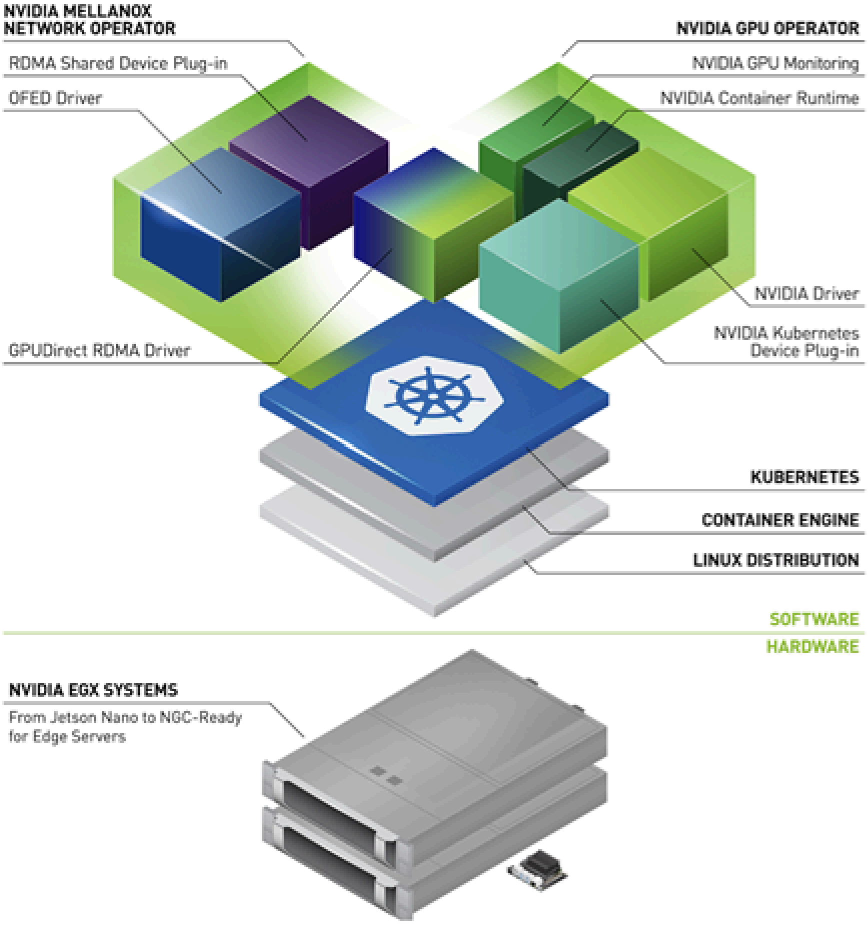
Using GPU Operators, we can automate the management of all NVIDIA software components needed to provision GPUs. For instance: NVIDIA drivers which are used to enable CUDA runtime, the Kubernetes device plugin exposes GPUs to Kubernetes worker nodes, the NVIDIA Container Runtime toolkit, automatic node labeling, and DCGM-based monitoring for managing and monitoring NVIDIA GPUs in large-scale, Linux-based cluster environments.-

NVIDIA's even committed further by building an entire stack of Infrastructure automation tools called NVIDIA DeepOps for Kubernetes and Slurm clusters with NVIDIA GPUs. Which allows customers to get a fully configured, cloud-native, and GPU-optimized infrastructure in a matter of minutes.

When it comes to training large language models though, the truth is that large machine-learning jobs would need to span many Kubernetes nodes to take full advantage of

GPU power. So, NVIDIA also developed GPUDirect for direct communication with the NIC or NVLink for cross-communication with the GPU to maximize efficiency.

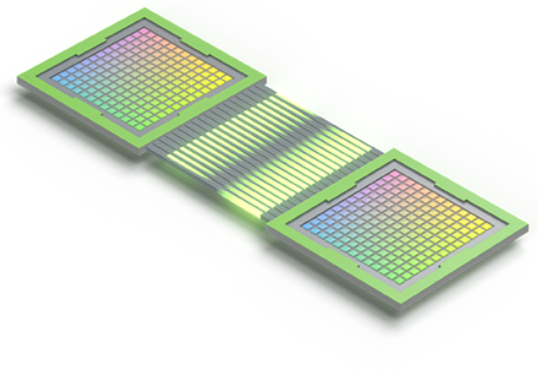
They even developed what is called NVIDIA Network Operator to automate the deployment and configuration of the software required for accelerated networking. As a matter of fact, OpenAI is using these technologies to train large AI models, including GPT4, DALLE, and Sora. - B-Roll 23B



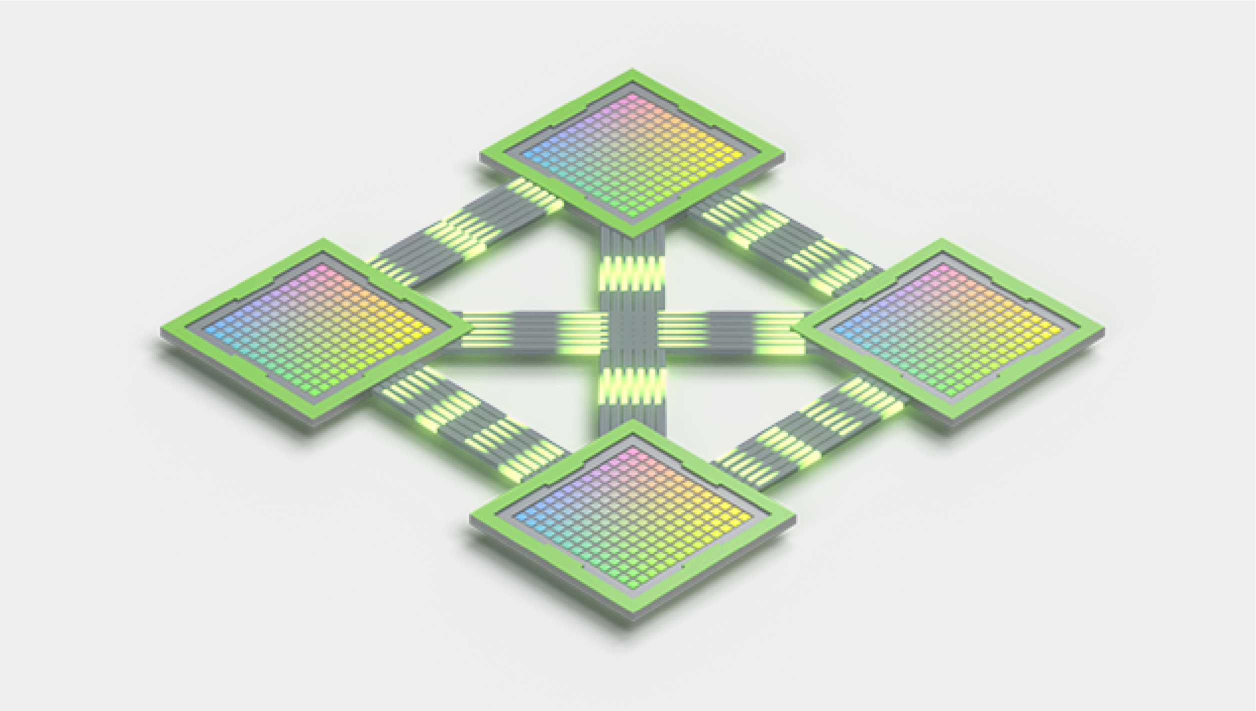
Network operator

NVIDIA also developed a Cloud-Native Stack to make the best use of compute resources. It's like is a collection of software to run cloud-native workloads on NVIDIA GPUs. NVIDIA Cloud Native Stack is based on Ubuntu, Kubernetes, Helm and the NVIDIA GPU and Network Operator. With the Cloud Native Stack, developers can create, test, and run containerized applications faster with GPUs.

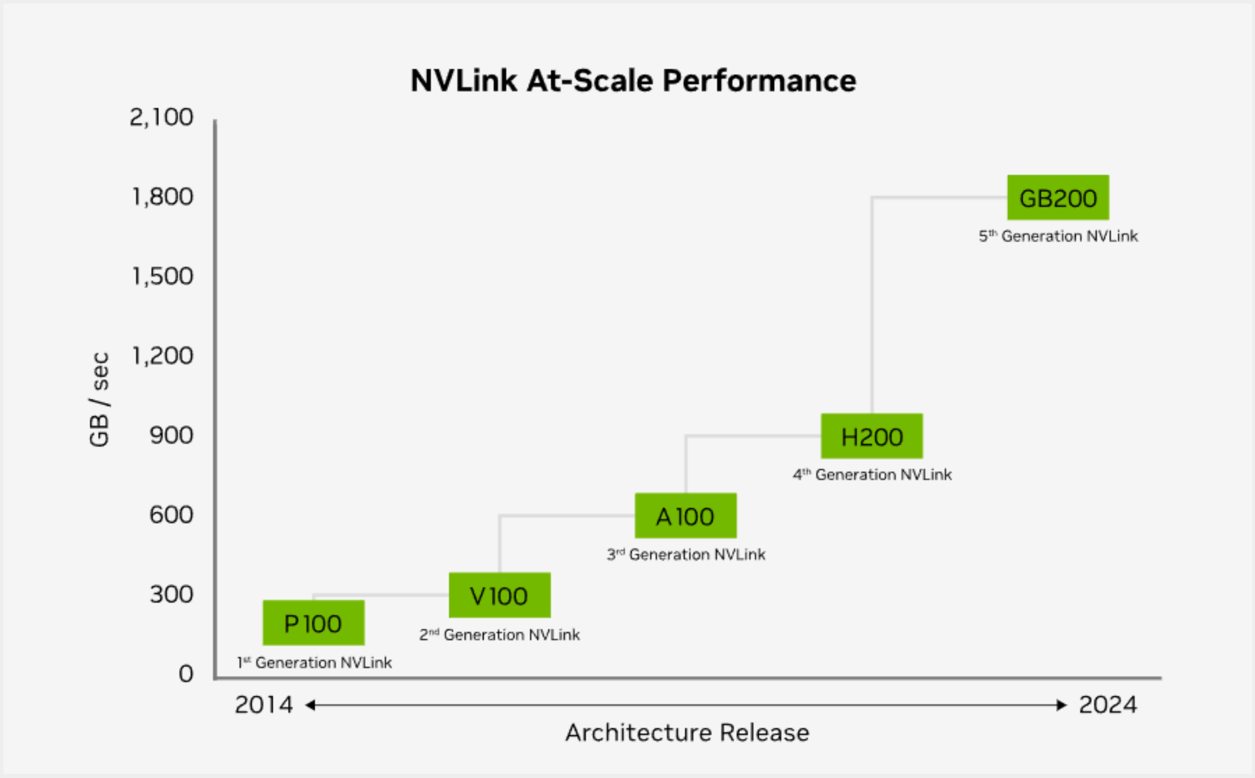
NVLink & NVSwitch: Fastest HPC Data Center Platform | NVIDIA



NVIDIA H100 PCIe avec connexion NVLink GPU vers GPU



NVIDIA H100 avec connexions NVLink GPU vers GPU



NVIDIA is making strategic moves to pave the way for a future where NVIDIA's integrated Kubernetes platform becomes the foundation for AI and high-performance computing (HPC) infrastructure. With this forward-thinking approach, this is how NVIDIA transforms the landscape of modern AI development. -

By the way, we have made a playlist called "Kubernetes in 30 days" about this fascinating technology, check them out if you're interested! -

[Outro and CTA]

Do you think that NVIDIA's cloud-native strategy is driving their AI innovation? I want to know your thoughts, in the comment section below. If you're interested in receiving more content on this topic, you can subscribe to our newsletter (links in the description). I will see you in the next one!

You show the GitHub repository https://github.com/NVIDIA/gpu-operator and the github address

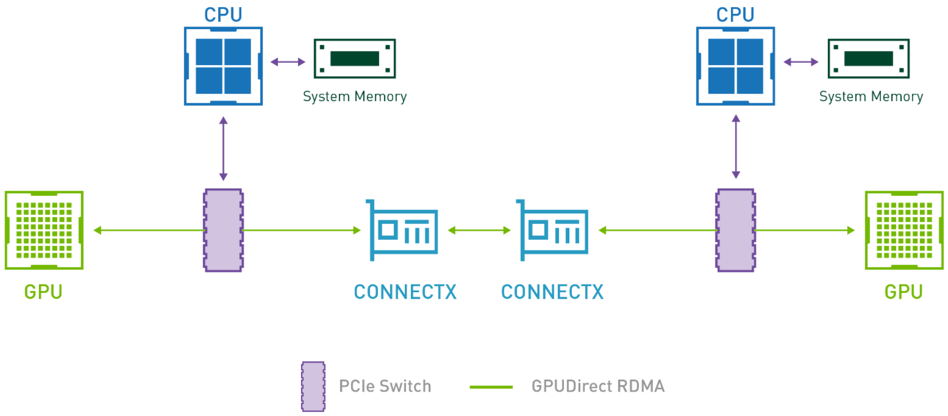
NVIDIA container runtime toolkit use this web page and web address
https://github.com/NVIDIA/nvidia-container-toolkit

DCGM-based monitoring use this web page and web address
https://developer.nvidia.com/blog/monitoring-gpus-in-kubernetes-with-dcgm/

Add NVIDIA DeepOps github repo https://github.com/NVIDIA/deepops and the GitHub address

add motion text one by one ‘in a matter of minutes’

for GPU direct use this picture on this page
https://developer.nvidia.com/gpudirect


CPU
System Memory
CPU
System Memory
GPU
CONNECTX
CONNECTX
GPU
PCIe Switch
GPUDirect RDMA