Overview:

The next- generation NVIDIA RTX Server delivers a giant leap in cloud gaming performance and user scaling. It packs 40 NVIDIA Turing GPUs into an 8U blade form factor that can render and stream even the most demanding games at GeForce RTX 2080 performance levels.
Cloud gaming performance and user scaling that is ideal for Mobile Edge Computing. From AAA games to virtual and augmented reality, the future is here.
NVIDIA vGPU technology enables up to 160 PC games to be run concurrently, with mobile games streamed at even higher concurrency rations using container technology. The NVIDIA RTX Server can run the most demanding and graphically intense games the world at high frame rates, which reduces latency.

Architecture:

The NVIDIA Container Toolkit is architected so that it can be targeted to support any container runtime in the ecosystem. For Docker, the NVIDIA Container Toolkit is comprised of the following components (from top to bottom in the hierarchy):

- nvidia-docker2
- nvidia-container-runtime
- nvidia-container-toolkit
- libnvidia-container

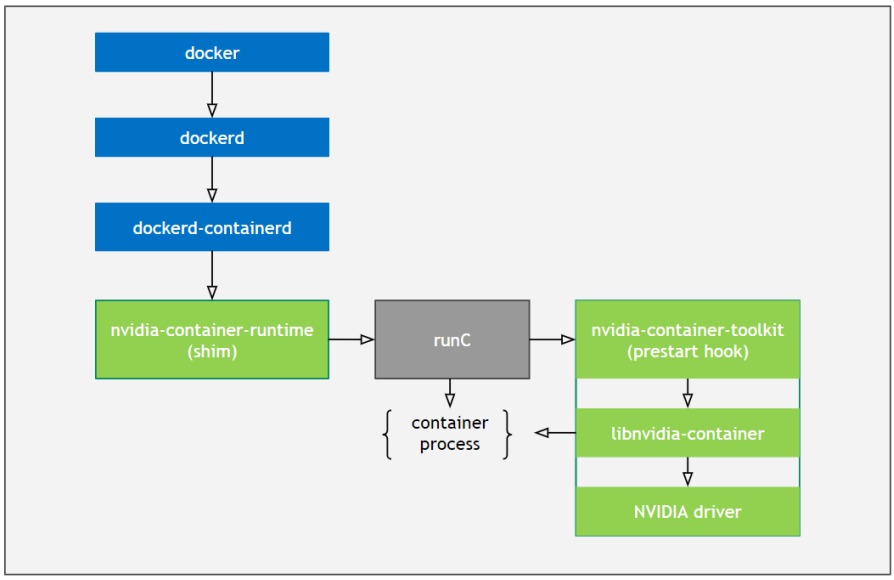

Figure 1: Architecture

- libnvidia-container- This component provides a library and a simple CLI utility to automatically configure GNU/Linux containers leveraging NVIDIA GPUs. The implementation relies on kernel primitives and is designed to be agnostic of the container runtime. libnvidia-container provides a well-defined API and a wrapper CLI (called nvidia-container-cli) that different runtimes can invoke to inject NVIDIA GPU support into their containers.
- nvidia-container-toolkit- This component includes a script that implements the interface required by a runC prestart hook. This script is invoked by runC after a container has been created, but before it has been started, and is given access to the config.json associated with the container (e.g. this config.json). It then takes information contained in the config.json and uses it to invoke the libnvidia-container CLI with an appropriate set of flags. One of the most important flags being which specific GPU devices should be injected into the container.
- nvidia-container-runtime- This component unsed to be complete fork of runc with NVIDIA specific code injected into it. Since 2019, it is a think wrapper around the navite runc installed on the host system. Nvidia-container-runtime takes a runc spec as input, injects the nvidia-container-toolkit script as a prestart hook into it, and then calls out to

the native runc, passing it the modified runc spec with that hook set. It's important to note that this component is not necessarily specific to docker.

- nvidia-docker2- This package is the only docker-specific package of the hierarchy. It takes the script associated with the nvidia-container-runtime and installs it into docker's /etc/docker/daemon.json file. This then allows you to run to automatically add GPU support to your container. It also installs a wrapper script around the native docker CLI called nvidia-docker which lets you invoke docker without needing to specify-runtime-nvidia every single time. It also lets you set an environment variable on the host to specify which GPUs should be injected into a container

NVIDIA GPU Operator:

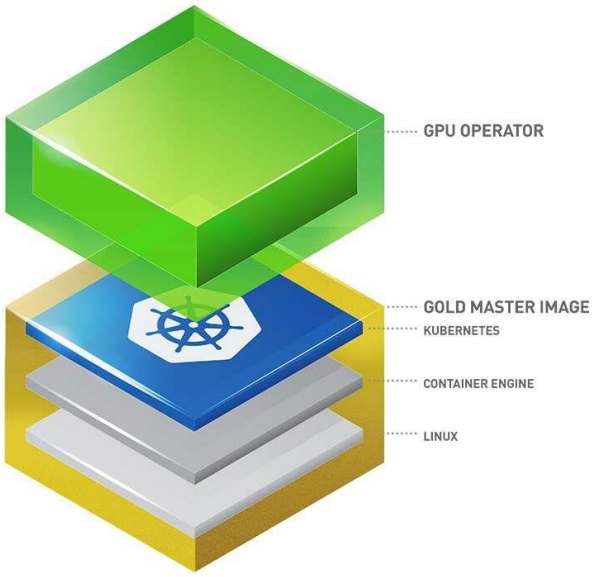

Figure 2: Nvidia GPU Operator

Kubernetes is an open-source platform for automating the deployment, scaling, and managing of containerized applications.

Red Hat OpenShift Container Platform is a security-centric and enterprise-grade hardened Kubernetes platform for deploying and managing Kubernetes clusters at scale, developed and supported by Red Hat. Red Hat OpenShift Container Platform includes enhancements to

Kubernetes so users can easily configure and use GPU resources for accelerating workloads such as deep learning.

The NVIDIA GPU Operator uses the operator framework within Kubernetes to automate the management of all NVIDIA software components needed to provision GPU. These components include the NVIDIA drivers (to enable CUDA), Kubernetes device plugin for GPUs, the NVIDIA Container Toolkit, automatic node labelling using GFD, DCGM based monitoring and others.

Kubernetes provides access to special hardware resources such as NVIDIA GPUs, NICs, Infiniband adapters and other devices through the device plugin framework. However, configuring and managing nodes with these hardware resources requires configuration of multiple software components such as drivers, container runtimes or other libraries which are difficult and prone to errors.

Services Offerings:

- Data Science Ideation Workshop- Every enterprise in the world has data science and AI on their agenda, but most don't know where to get started. This workshop will help customers create a roadmap to start AI in their business. We will show the art of the possible with NVIDIA technology.

Service hours may be applied to assist the customer with the following activities:

- Interviews, demos and discussion with key stakeholders to understand customer's goals.
- Analyze current workflows to understand if they can be AI accelerated.
- Map goals to activities and their prioritization.
- Develop services roadmap and engagement model.

- AI Onboarding Services for DGX- A multi-week engagement to activate your team on your NVIDIA DGX system. Enable your team to begin their DGX journey alongside experts who can answer questions and teach best practices.

Service hours may be applied to assist the customer with the following activities:

- Data science assessment & gap analysis
- Workflow assessments and interviews of stakeholders
- Recommend workflow optimization best practices
- Demo examples on AI at Scale workflows
- Education content on AI at Scale from software stack to data science model development and deployment
- Documenting assessment findings, recommended best practices, roadmap, and recommendations on customized training or future professional services engagements

- AI Hands-on-Training- Equip your team with hands-on training to optimally operate, develop and deploy AI at scale. Hands-on-trainings walk your team through best practices, tips and tricks to accelerate your AI applications with GPUs.

Service hours may be applied to assist the customer with the following activities:

- Development of materials and trainings for training customization
- Delivering training content

Solution:

- Data analytics- Data analytics workflows have traditionally been slow and cumbersome, relying on CPU compute for data preparation, training, and deployment. Accelerated data science can dramatically boost the performance of end-to-end analytics workflows, speeding up value generation while reducing cost.
- Machine learning- Machine learning helps businesses understand their customers, build better products and services, and improve operations. With accelerated data science, businesses can iterate on and productionize solutions faster than ever before all while leveraging massive datasets to refine models to pinpoint accuracy.
- Entire Network Security and Visibility- The need for data is driving an increasingly distributed and complex network environment making data visualization and isolation essential security functionalities for every data center. Our Networking solutions enhance security, simplify data center automation and allows complete visibility to your network.

- Maximize Performance, Minimize Costs- GPU-accelerated data centers deliver breakthrough performance with fewer servers and less power, resulting in faster insights with dramatically lower costs. Train the most complex deep learning models to solve your biggest challenges with NVIDIA deep learning- NVIDIA deep learning solutions powered by NVIDIA® data center GPUs. Ideal for production-scale training and inference, NVIDIA's world-leading performance accelerates the most popular deep learning frameworks and over 550 high performance computing (HPC) applications.

- Conversational AI- Accelerate the Full Pipeline, from Speech Recognition to Language Understanding and Speech Synthesis. 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 NVIDIA's conversational AI SDK, developers can quickly build and deploy state-of-the-art multimodal AI services to power applications across a single unified architecture, delivering highly accurate, low-latency systems with little upfront investment.

- Prediction and forecasting- Prediction and forecasting are powerful tools to help enterprises model future trends. With NVIDIA accelerated data science, businesses can take massive-scale datasets and craft highly accurate insights to fuel data-driven decisions.

Platforms:

- CUDA-X AI-
    - TensorRT
    - Triton Inference Server
    - NeMo
    - NCCL
    - Optical Flow SDK
- Clara-

  - Clara Guardian
  - Clara Imaging
  - Clara Parabricks
- HPC-
  - HPC SDK
  - CUDA Toolkit
  - IndeX
- Drive-
  - Driver AGX
  - Driver Hyperion
- ISAAC-
  - Isaac SDK
  - Isaac Sim
  - Jetpack

Products:

- DGX
- NVIDIA- Certified
- Grace cpu
- Bluefield Dpus

Benefit:

- With NVIDIA GPUs on Google Cloud Platform, deep learning, analytics, physical simulation, video transcoding, and molecular modeling take hours instead of days.
- We can also leverage NVIDIA GRID virtual workstations on Google Cloud Platform to accelerate your graphics-intensive workloads from anywhere.
- With cloud-based GPU solutions, enterprises can access high-density computing resources and powerful virtual workstations at any time, from anywhere, with no need to build a physical data center.

- From virtual desktops, applications, and workstations to optimized containers in the cloud, data scientists, researchers, and developers can power GPU-accelerated AI and data analytics at their desks.
- GPU-accelerated data centers deliver breakthrough performance for compute and graphics workloads, at any scale with fewer servers, resulting in faster insights and dramatically lower costs. Sensitive data can be stored, processed, and analyzed while operational security is maintained.
- AI at the edge needs a scalable, accelerated platform that can drive decisions in real time and allow every industry to deliver automated intelligence to the point of action—stores, manufacturing, hospitals, smart cities.

Colocation Companies:

- Aligned Energy
- Digital Realty
- Vantage Data Centers
- Interxion
- NTT Communications
- Equinix
- NTT Data
- Data Dock
- Africa DataCentres

Partner Cloud:

- Alibaba Cloud
- Aws
- Oracle
- Tencent Cloud
- Google Cloud
- IBM cloud

- Microsoft Azure

Conclusion:

NVIDIA help creative and technical professionals maximize their productivity from anywhere by giving them access to the most demanding design and engineering applications from the cloud. With the latest NVIDIA data center T4 GPUs, users can enjoy the most advanced 3D graphics platform.

NVIDIA is the computing platform that transforms big data super-human Intelligence.

Reference:

https://www.nvidia.com/en-in/data-center/colocation-partners/

https://www.nvidia.com/en-us/networking/

https://www.nvidia.com/en-in/data-center/gpu-cloud-computing/

https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/arch-overview.html#arch-overview

https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/overview.html

https://www.nvidia.com/en-us/data-center/solutions/