TECHNICAL OVERVIEW

DEPLOYING NVIDIA CLOUD-NATIVE ARCHITECTURE FOR BARE-METAL, SECURED SUPERCOMPUTING

OVERVIEW

High-performance computing (HPC) and AI have driven supercomputers into wide commercial use. They've become the primary data processing engines to enable research, scientific discoveries, and even product development. When designing a supercomputer architecture, the goal is twofold: to mitigate factors that impact performance, and to accelerate application performance whenever possible.

Supercomputing has become more democratized and leveraged by more users with a diverse set of applications. The challenge today is to bring cloud services into supercomputing infrastructures, both without adversely impacting application performance, and by delivering a bare-metal supercomputing experience.

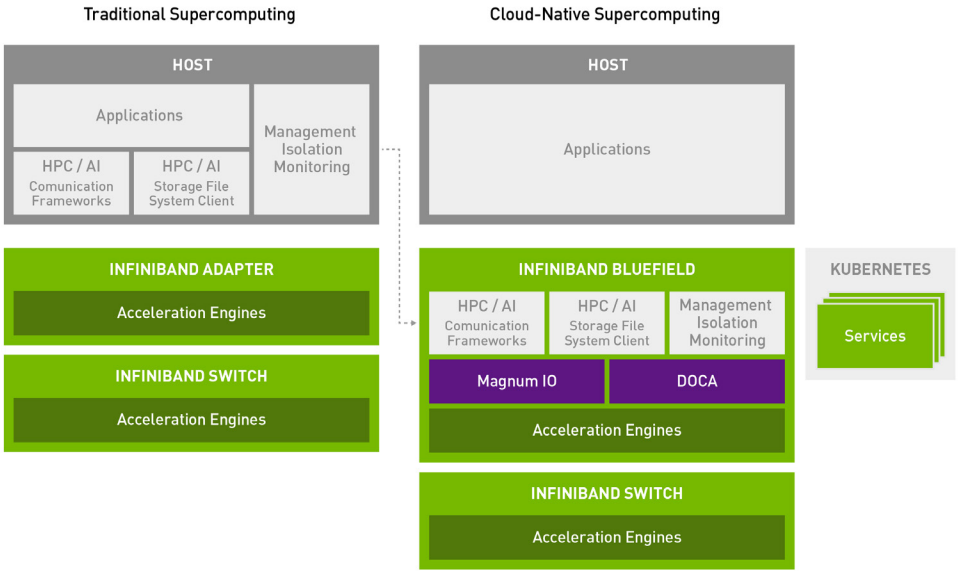
NVIDIA Cloud-native Supercomputing does all this and more by offloading the overhead and orchestration to achieve a secure multi-tenant environment, while providing performance isolations on shared computing resources. NVIDIA's cloud-native supercomputing solution enables the cloud to address the anticipated growth in users and HPC and AI applications.

NVIDIA CLOUD-NATIVE SUPERCOMPUTING

Cloud-native Supercomputing (CNSC) blends the power of high-performance computing with enhanced offloading, acceleration, security, and ease-of-use of cloud computing services.

The NVIDIA® CNSC platform leverages the BlueField® Data Processing Unit (DPU) architecture, and offers several advantages:

- High-speed, low latency NVIDIA Quantum InfiniBand networking to deliver unmatched bare metal performance.
- Additional application and middleware offload and acceleration capabilities.
- Secure infrastructure for data protection and tenant isolation.
- Remote data storage allocation for tenant workloads.
- NVIDIA DOCA™ software framework for software-defined Infrastructure and tenant services.
- NVIDIA UFM®: a comprehensive fabric management platform.



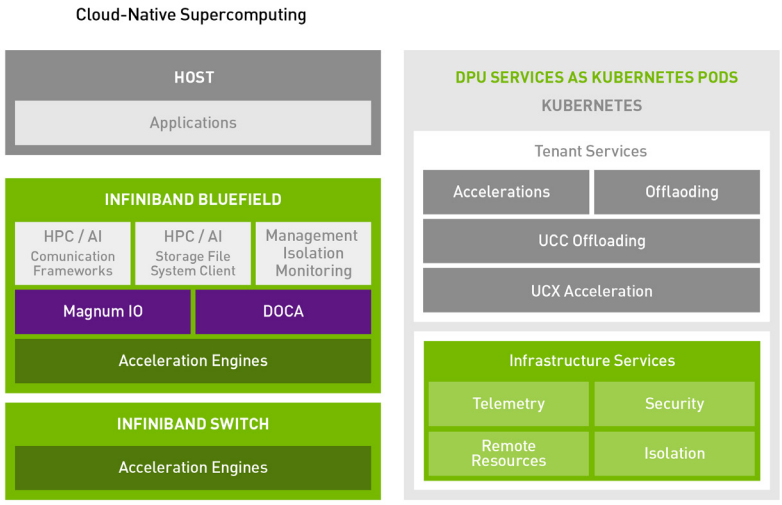
Traditional supercomputing offers bare metal performance for cloud IT managers, without any isolation or secure infrastructure. CNSC offers a flexible and secure infrastructure for multi-tenancy and provides CPU offload and advanced acceleration capabilities. The system administrator or cloud service provider can deploy multi-services as Kubernetes pods on the DPU that are unique per tenant request, or to provide infrastructure advantages such as telemetry, security, resource allocation, and more.

NVIDIA CNSC offers several advantages and offload capabilities, including:

- Acceleration and offloading of HPC communication on the NVIDIA DPU, such as Unified Collection Communication (UCC) collective offloads service on the DPU, or Unified Communication X (UCX) offload acceleration service on the DPU.
- Additional advantages include active messages, smart MPI progression, data compression on the DPU, and user-defined algorithms based on the DOCA API.

In terms of security, the rise of AI, 5G and smart devices has made it difficult to identify potential breaches and attacks against traffic that's moving through today's data centers. NVIDIA's CNSC provides a zero-trust architecture with security services such as host attestation, host introspection, and more. This zero-trust security platform allows developers to create fully secure environments that protect virtually every aspect of the cloud-native data center, in real time.

CNSC also provides storage virtualization for remote storage allocation as well as tenant Service Level Agreement (SLA), for example, rate-limiting, bandwidth-guaranteed, and reservation of network resources based on tenant isolation and requirements.



In the preceding diagram, note the two types of services: tenant services for acceleration and offloading, and infrastructure services for telemetry, security, remote resources allocation, and isolation. Both types of services should be managed and controlled by the system administrator or the cloud provider according to their needs or tenant SLA / requests.

USING KUBERNETES ON BARE METAL - TENANT ALLOCATION

On top of the bare metal (BM) service, the tenant can deploy and run the workloads and applications as Kubernetes containers.

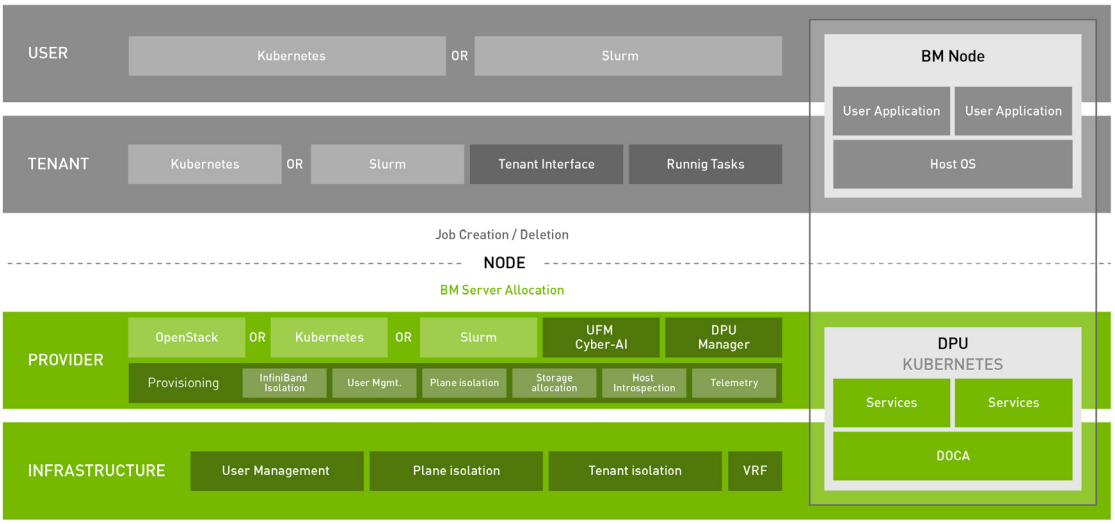
Currently, NVIDIA offers three container network interface (CNI) plugins and device plugins for InfiniBand:

1. SR-IOV Mode – To use a container, the IT manager creates virtual functions. Each container is mapped to a virtual function. With this CNI we can define QoS parameters and isolate virtual functions.
2. Shared mode – IT managers can create containers without any limitation on the bare metal machine, without any isolation or tenant SLA.
3. Host Device mode – in this mode, the number of containers is limited to the number of physical ports. Each container is mapped to a dedicated physical port; with this option isolation and SLA are available.

In CNSC, the allocation of the resources is done by the cloud admin for the Bare Metal resources; therefore, a shared mode plugin is used for running workloads and applications as containers on X86 bare metal servers, without any limitation or isolation.

CLOUD NATIVE ARCHITECTURE

The NVIDIA Cloud Native architecture, as depicted in the following diagram, creates two (2) zones:



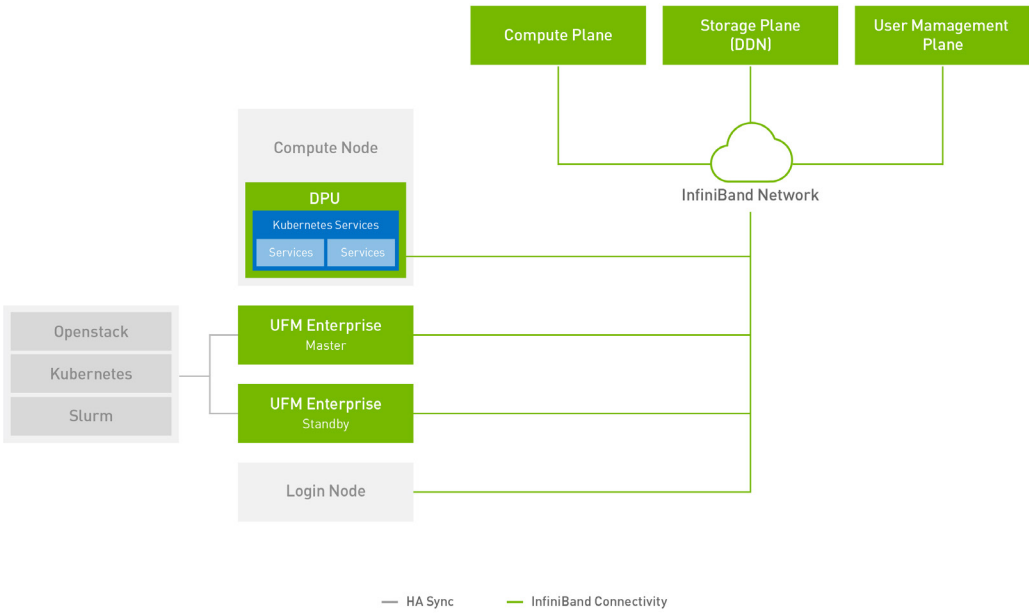
The Bare Metal server zone is provided to the tenant by the Cloud Admin and controlled by the tenant, while the DPU Zone is controlled by the system administrator, the cloud manager, or the cloud provider. The Bare Metal server zone contains two major planes—the Tenant Plane and the User Plane. Also, multiple users can be under the same tenant.

The Tenant receives the Bare Metal allocation from the provider and can decide which applications or workloads to run. The Tenant can also use any external tools to manage jobs between the users, for example, SLURM or Kubernetes.

The DPU zone has two major planes: the Infrastructure Plane and the Provider Plane. The former provides both the network infrastructure and the facilities to the providers; for example, an external network for Internet connectivity, with security and isolation for tenants, or connectivity to a remote storage allocation. The Provider Plane is responsible for all the dynamic allocations for infrastructure and tenant needs.

THE NVIDIA QUANTUM INFINIBAND NETWORK

The high-speed, low latency InfiniBand network is used for all cloud activities, including compute, storage, and user management.



Each compute node can support multiple DPUs, while NVIDIA Unified Fabric Manager (UFM) can be deployed in high availability mode in appliances or as a software deployment on servers with InfiniBand connectivity to the network.

UFM is used for InfiniBand management and orchestration and covers two critical aspects of this solution. Providing full backup and high availability for all the critical network services, including the Subnet Manager (SM), NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™[1] aggregation manager (AM), and UFM telemetry. Moreover, it exposes the software defined network (SDN) and cloud API for integration with 3rd party tools.

System administrators or cloud providers can integrate their own software platforms with the UFM API, and alternatively, can use open-source platforms, such as OpenStack, Kubernetes, or SLURM.

[1] NVIDIA SHARP technology improves upon the performance of MPI operation, by offloading collective operations from the CPU to the switch network and eliminating the need to send data multiple times between endpoints.

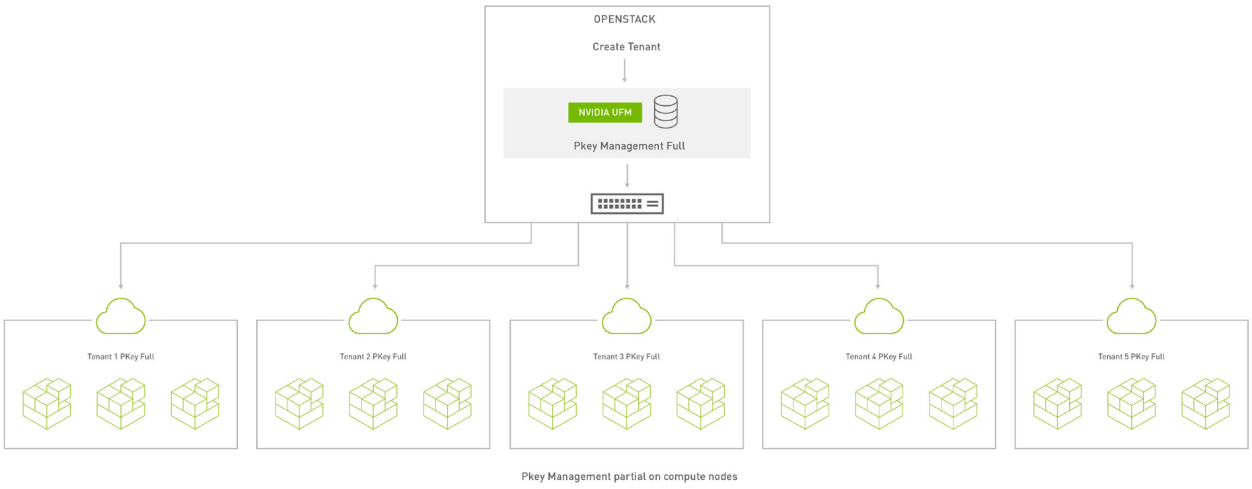
CLOUD-NATIVE SOLUTION BUILDING BLOCKS

NVIDIA uses OpenStack—the most widely deployed open-source cloud operating system, to orchestrate and provision the entire secured multi-tenant supercomputing cloud.

- To provision the InfiniBand fabric, NVIDIA uses NVIDIA UFM, which is fully integrated with the OpenStack networking platform Neutron, that provides 'networking as a service' using a modular layer 2 (ML2) mechanism driver. For managing the InfiniBand cluster, UFM seamlessly provisions tenant networks over the InfiniBand fabric and isolates resources between tenants.
- For DPU provisioning, NVIDIA uses the DPU Manager—a set of tools and services, such as Forman, Ansible, and AWX, that allows for the deployment and configuration of NVIDIA BlueField DPUs at scale, as part of the cloud-native supercomputer.

OPENSTACK INTEGRATION

With regard to OpenStack integration, The NVIDIA UFM ML2 plugin interacts with the UFM REST API to create a tenant, while also isolating it from other tenant resources. The following diagram illustrates how Partition Key (PKey) isolation limits traffic to a tenant scope, blocking any tracking between Tenants 1 and 2. It also depicts 15 nodes, isolated by OpenStack and UFM integration, into five tenants, each with unique PKeys.



With the basic default PKey configuration each node can interact only with UFM per UFM request. By assigning a bare metal resource to a tenant PKey, each tenant cluster node can interact only with the resources of the same tenant; it can not interact with other tenant resources.

> UFM

UFM offers customers and partners a robust REST API (also known as RESTful API) for integration and automation, or to build custom integration with any third-party software. UFM also offers several job scheduler and cloud platform plugins, which interact with the NVIDIA UFM Enterprise REST API; for example, Kubernetes CNI, OpenStack ML2, Slurm, and LSF.

> OpenStack SDN ML2 plugin

While creating a new tenant, the OpenStack SDN ML2 plugin sends a notification to the UFM REST API. The UFM reserves and isolates the ports on the bare metal resources, only to the selected tenant.

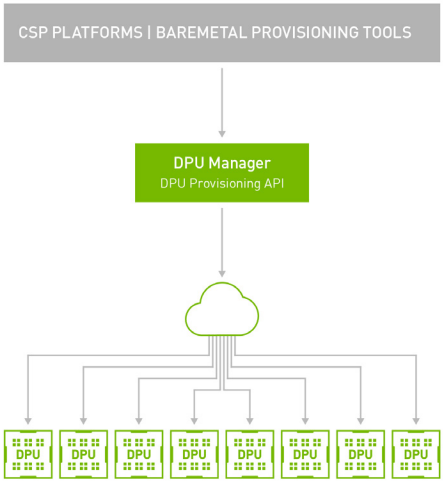
The ML2 Plugin and UFM Integration provides the following capabilities:

- Network provisioning per tenant.
- Full isolation between tenants.
- Allocation of network resources per tenant.
- Tenant SLA based on PKey isolation, such as rate limit, bandwidth guarantee (BWG), and SHARP resources allocation.

> The DPU Manager

The DPU Manager, which today is based on the NDO, provides several capabilities, all of which are exposed over the DPU provisioning API:

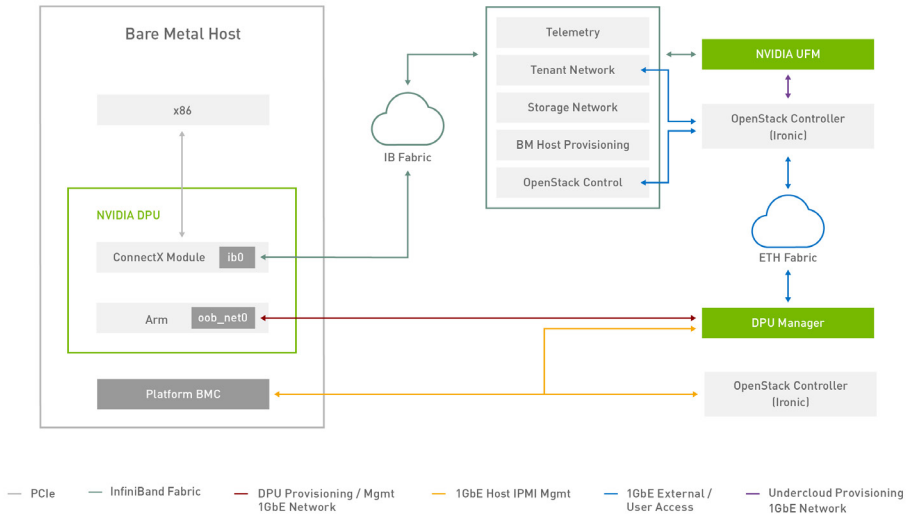
- DPU provisioning.
- DPU lifecycle management.
- Services deployment.
- Services lifecycle management.



Customers and partners can easily integrate 3rd party tools into the NVIDIA DPU Manager provisioning API.

COMPUTE NODE CONNECTIVITY

The following diagram shows the network that is used for boot cycles and bare metal machine provisioning. The 1 GbE platform Baseboard Management Controller (BMC) interface connects to the OpenStack ironic Controller and also to (actually through) the DPU Manager.



The 1G interface on the DPU connects to the DPU Manager for DPU provisioning. The network port connects to the InfiniBand cluster for compute and storage networking. It's important to note that with this solution, everything is done over the InfiniBand cluster – only the DPU provisioning takes place over the 1G interface.

READY TO TAKE THE NEXT STEP? LET'S BUILD AND DEPLOY.

NVIDIA CNSC delivers secured multi-tenant supercomputing while delivering a bare-metal supercomputing experience. Now that you're familiar with the fundamental concepts of NVIDIA Cloud Native Supercomputing architecture and the various components, you're almost there! Be sure to check out the **GTC Session: Build and Deploy a Cutting-edge Cloud-native Supercomputer**, and additional resources, which further showcases the architecture, diving deeper into the unique benefits of cloud-native supercomputing from NVIDIA.

Ready to Get Started?

Learn more about NVIDIA Cloud-Native Supercomputing at:
nvidia.com/en-us/networking/products/cloud-native-supercomputing

Learn more about NVIDIA UFM at: **docs.nvidia.com/networking/category/ufm**

© 2022 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP), and NVIDIA UFM are trademarks and/or registered trademarks of NVIDIA Corporation and/or its affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. 2289251. JUN22