SECURELY DEPLOY AND OPERATE NVIDIA AI CLOUDS

Powered by BlueField-3 DPUs for NVIDIA AI systems

White Paper

| JANUARY 2024

Table of Contents

List of Figures

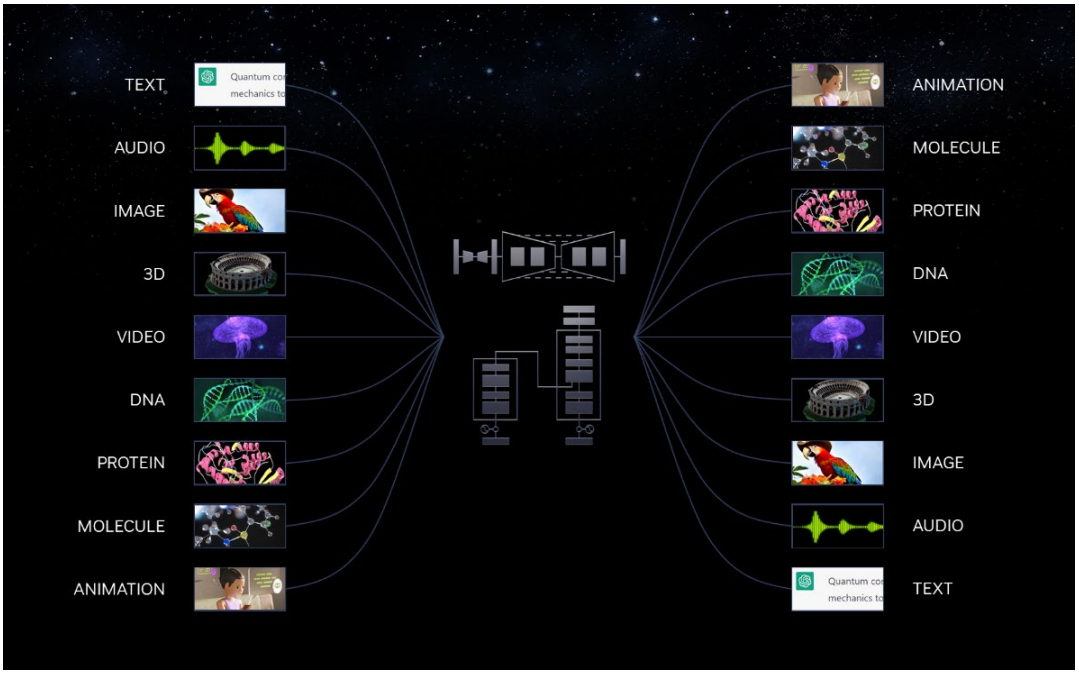
Abstract

Organizations are rapidly adopting generative AI to transform their product offerings and operations. This adoption presents vast opportunities but also brings challenges, such as navigating infrastructure complexity, ensuring security against evolving threats, and meeting strict data management requirements. Overcoming these challenges empowers organizations to maximize AI's potential. NVIDIA's accelerated computing platform is designed to operate at data center-scale and abstract the complexity from users, paving the way for AI transformation across all industries. The integration of BlueField-3 DPUs into NVIDIA's accelerated systems elevates performance, enhances efficiency, and increases security in AI cloud data centers. This white paper explores how NVIDIA's BlueField-3 DPUs bridge the gap between the demanding requirements of complex AI cloud infrastructure and the need for efficient operational implementation.

The Struggle to Operationalize Generative AI

Generative AI has created a sense of urgency for organizations to reimagine their products and business models. Goldman Sachs estimates the total addressable market (TAM) for generative AI software at $150 billion, a significant 22% portion of the global software industry's[1] $685 billion value. The proliferation of generative AI tools in business and society is projected to drive a 7% in global GDP, amounting to nearly $7 trillion, over the next decade.

Figure 1. Generative AI Enables Users to Create Unique and Captivating Content



[1] Source: Autor et al. (2022), Goldman Sachs Research

The surge in generative AI popularity, exemplified by applications like ChatGPT, has prompted a global race among organizations to leverage its transformative capabilities. Yet nearly every organization faces challenges to integrate AI into their operations. Key challenges include infrastructure complexity, modern security risks, and strict data management requirements. These hurdles prevent the deployment of the requisite data center infrastructure for generative AI, hindering its adoption or the ability to fully realize its potential.

Infrastructure Complexity

Deploying and operating accelerated computing as the infrastructure to support generative AI is highly complex and requires full-stack expertise. From powerful GPU computing, networking, and storage hardware to AI software, advanced AI models, and domain-specific applications—organizations need to integrate and fine-tune every layer of the stack to achieve optimal performance. Moreover, managing the infrastructure to accommodate the transient nature of AI workloads, and ensuring effective resource provisioning, is technically demanding and puts a heavy burden on IT operations.

Scaling cloud data centers to handle the extreme load of generative AI, while ensuring top performance and responsiveness, isn't easy. It involves deploying distributed computing frameworks, load balancing, and effective traffic management to provide real-time service to numerous users. With generative AI, the scale of the infrastructure often influences time-to-market (TTM) and market share growth. However, scaling the infrastructure to accelerate TTM can complicate deployment and operations, ironically delaying TTM. The solution is a well-integrated, properly designed and appropriately sized validated reference architecture. Such an architecture is essential for delivering an optimal infrastructure for world-class, scalable generative AI services.

Modern Security Risks

AI data centers face similar, if not greater, security threats as other data centers. An IBM report indicates that in 2023, the global average cost of a data breach rose to $4.45 million, a 15% increase over three years. The multi-tenant nature of cloud data centers and a growing cyber threat landscape compel organizations to continuously re-assess and refine their security postures. In conventional cloud infrastructures, both the policies of service providers and the applications of tenants share the same trust domain (the CPU), exposing organizations to risks and affecting operational agility.

As organizations embrace generative AI for various applications, they face fresh security vulnerabilities. Firstly, generative models can inadvertently generate sensitive, false, or inappropriate content, posing legal and reputational risks. Moreover, these models are susceptible to targeted attacks using modified input data to skew or poison the model's output. In addition, model training can inadvertently incorporate confidential, copyrighted, or inaccurate data.

To address these challenges, organizations need to enforce robust safeguards, such as input validation, content filtering, and continuous monitoring, to prevent exploitation and attacks. Regular software updates and patches are essential to address emerging threats in the rapidly evolving AI security landscape.

Stringent Data Management Requirements

AI large language models (LLMs) are trained on vast amounts of data to learn patterns, relationships, and language understanding, where extensive and diverse training data is key to their impressive language capabilities. However, training LLMs using traditional storage technologies, which cannot provide the high throughput and low latency required for efficient data access during training, has proven inadequate. This leads to slow data retrieval that bottlenecks the training process, resulting in longer training times and inefficient use of computing power. As model sizes and training datasets expand, these traditional storage solutions fall short, failing to deliver the essential storage capacity and I/O performance. Moreover, they are not tailored to handle the parallel data access patterns required for distributed training.

Direct-attached storage (DAS) alleviates some challenges in LLM training, yet it presents drawbacks such as limited storage capacity and scalability, complex management and monitoring, and lower resiliency. Additionally, DAS can hinder collaboration among multiple researchers or data scientists working on the same LLM training project. The process of sharing, and syncing data across machines can become cumbersome and prone to errors. Successful training and inferencing of generative AI models require high-capacity, high-speed storage solutions such as NVMe/NVMe-oF, used in combination with distributed file systems or object storage to efficiently store and serve large models.

To ensure the seamless integration of generative AI into operations, organizations globally are seeking effective ways to address the infrastructure intricacies, security risks, and data management requirements when deploying and operating AI data centers at massive scale.

Securely Deploy and Operate NVIDIA AI Clouds

NVIDIA pioneered accelerated computing to solve problems beyond the capabilities of conventional computers. Spanning from chips and systems to software, acceleration libraries, and application frameworks, NVIDIA's accelerated computing is a full-stack, data center-scale platform that enables AI for every industry, forging new and improved business models.

Figure 2. NVIDIA's Accelerated Computing Stack for Generative AI

Modern AI data centers must accelerate every workload. This acceleration is essential to deliver the performance, scalability, and efficiency needed for powering the next wave of applications, while also optimizing energy efficiency. NVIDIA's accelerated computing is designed to abstract complexity for its users, thereby empowering businesses across various industries to integrate and benefit from modern AI capabilities. The foundation of NVIDIA's accelerated computing rests on several key pillars:

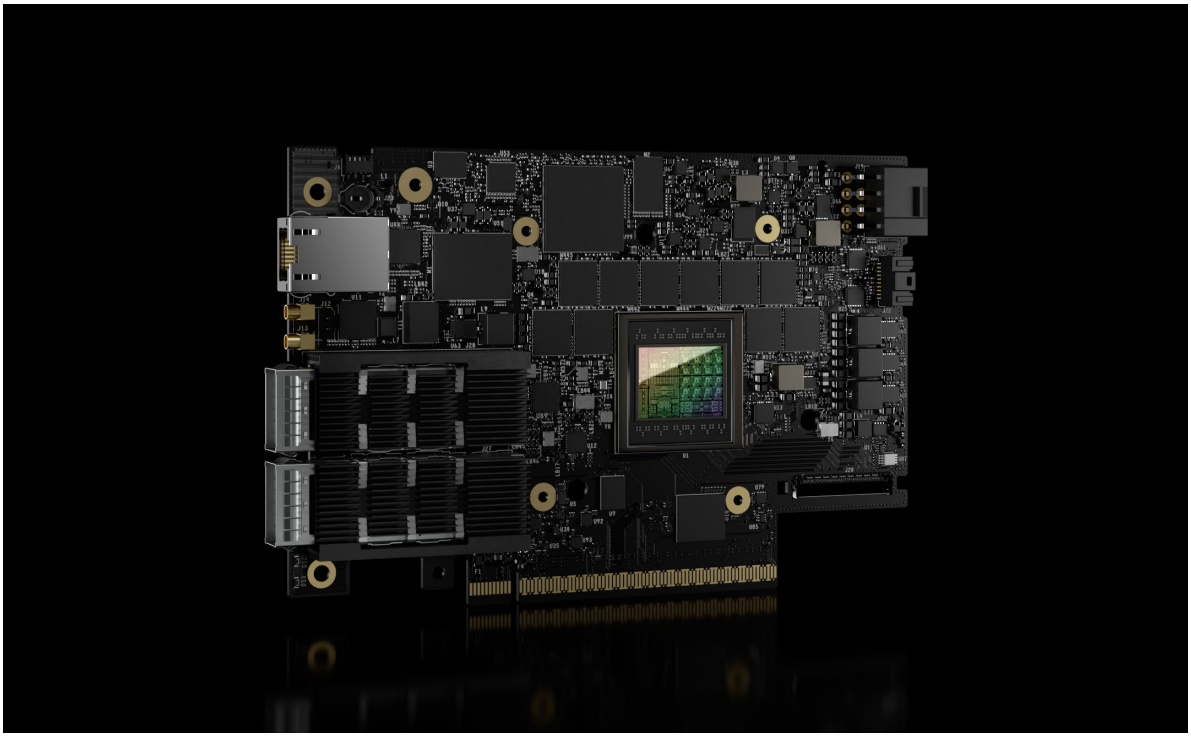
- CPUs that excel at parallel processing and are optimized for accelerating AI workloads.
- CPUs for serial processing and running hyperthreaded applications.
- DPUs, ideal for infrastructure computing tasks; used to offload, accelerate, and isolate data center networking, storage, security, and manageability workloads.
- Accelerated networking comprised of SuperNICs and network switches, together forming an optimal network fabric for AI computing.

BlueField Powers NVIDIA-Accelerated Systems

NVIDIA's accelerated systems are designed to meet the most rigorous performance requirements of a growing range of complex and diverse AI applications, while ensuring cloud manageability and security. Embodying these principles, NVIDIA has integrated BlueField®-3 DPUs and BlueField-3 SuperNICs across its range of accelerated computing systems, including NVIDIA HGX™ H100, and NVIDIA OVX™ L40S, among others.

The NVIDIA BlueField-3 DPU is an advanced infrastructure computing platform for data center infrastructure workloads. By offloading, accelerating, isolating networking, storage, and security, BlueField-3 DPUs enhance performance, optimize efficiency, and bolster security within AI data centers. BlueField-3 DPUs enable organizations to securely deploy and operate high-performance and efficient AI cloud data centers.

Figure 3. BlueField-3 DPU Platform

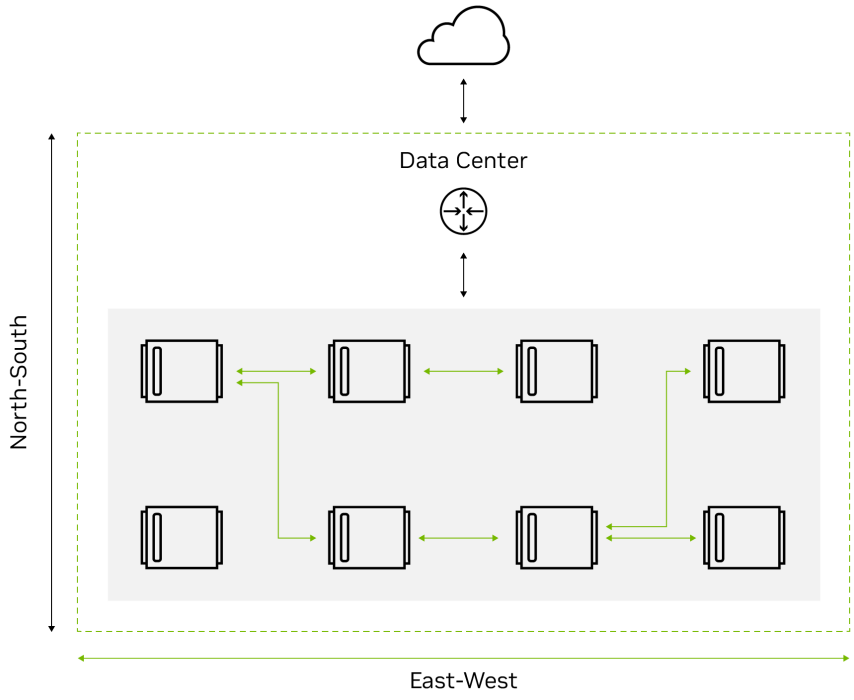


BlueField-3 SuperNICs are a new class of network accelerators, purpose-built for supercharging hyperscale AI computing workloads. Designed for network-intensive, massively parallel computing, BlueField-3 SuperNICs provide up to 400Gb/s of remote direct-memory access (RDMA) over Converged Ethernet (RoCE) network connectivity between GPU servers, optimizing peak AI workload efficiency.

The integration of BlueField-3 DPUs and SuperNICs within flagship NVIDIA accelerated systems capitalizes on NVIDIA's expertise in AI cloud data centers, optimizing the network traffic flow, accordingly:

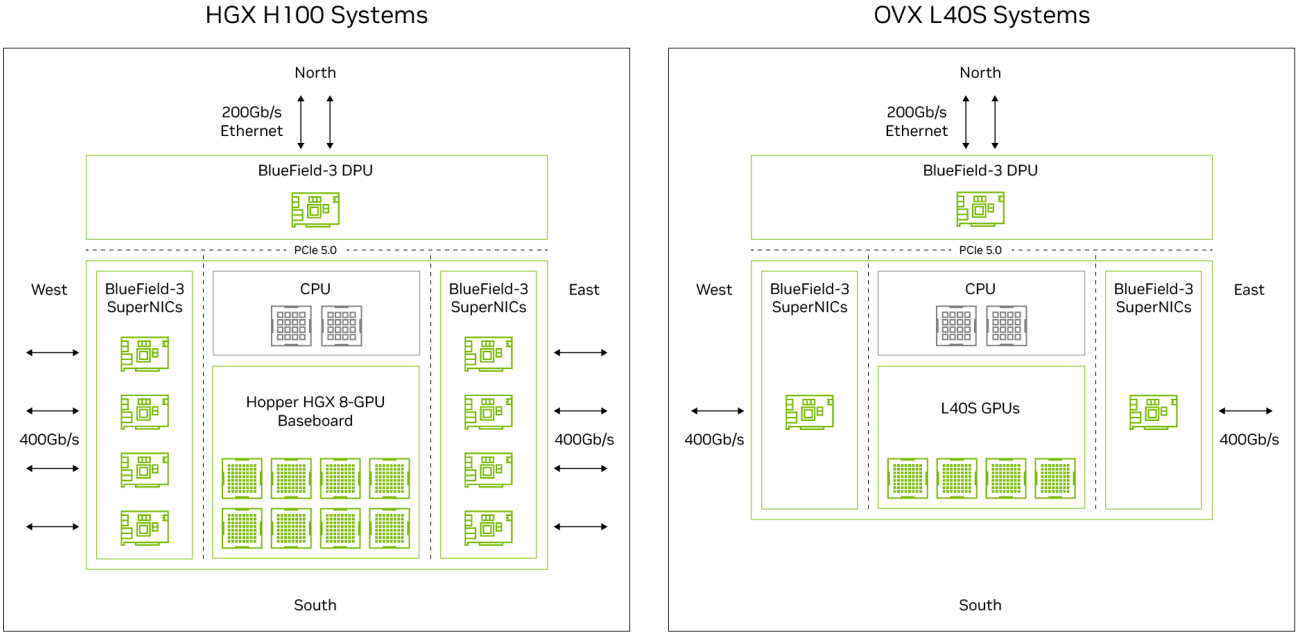
- **BlueField-3 DPUs for North-South (N-S) traffic:** Manages user traffic to the AI cluster, and from the cluster to external resources like cloud management systems, remote data storage nodes, and other data center environments, or Internet connections.
- **BlueField-3 SuperNICs for East-West (E-W) traffic:** Handles traffic between AI systems inside the cluster, typically used for distributed AI training, collective operations, and other AI computational tasks.

Figure 4. Optimized Networking for AI Clouds



The following figure illustrate BlueField-3 DPUs and SuperNICs design patterns on the N-S and E-W interfaces respectively of flagship NVIDIA systems:

Figure 5. BlueField-3 DPUs and SuperNICs in NVIDIA Accelerated Systems



Integrating BlueField-3 DPU into every AI system provides organizations with the following benefits:

- **Accelerated time-to-market with generative AI:** BlueField DPUs expedite AI workload deployment and operations, reducing provisioning time from months to days.
- **Scalable, elastic cloud computing:** BlueField DPUs transform AI data centers into elastic cloud platforms, enabling quick adaptation to changing AI workload demands.
- **Improved infrastructure performance:** Delivering a broad range of software-defined, hardware-accelerated services, BlueField DPUs enable infrastructure performance that is second to none.
- **Efficient data center deployments and operations:** BlueField DPUs streamline data center operations by simplifying server and network provisioning, particularly in EVPN network scenarios, reducing setup time and operational complexities.
- **Secure, multi-tenant cloud infrastructure:** BlueField DPUs adopt a zero-trust approach for multi-tenant AI clouds, ensuring they are secure and robust by offloading and isolating the data center control-plane.

Subsequent sections discuss the key BlueField-3 DPU capabilities and use-cases in NVIDIA AI cloud data centers. To learn more BlueField-3 SuperNICs, including how they accelerate AI computing, read to the white paper: *Next-Generation Networking for the Next Wave of AI*.

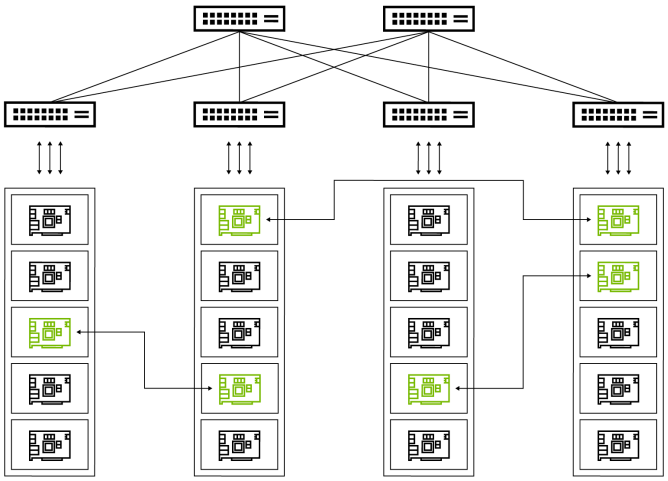
Accelerated VPC Networking

In the realm of modern AI cloud networks, there is a pressing need to sustain the intensive I/O associated with user interactions and data storage operations, particularly through the N-S network. These workloads require high-performance networking to meet the explosive user demand for generative AI applications and provide efficient data access during training operations.

BlueField-3 DPUs, featuring accelerated switching and packet processing (ASAP^2) technology, provide low-latency network connectivity at up to 400Gb/s speed, with zero CPU utilization. The NVIDIA ASAP^2 technology stack provides a range of network acceleration capabilities and full programmability through the NVIDIA DOCA SDK. BlueField offloads and isolates the SDN control plane software on the Arm cores and accelerates the networking data-plane in hardware. This provides CPU core savings along with better control and enhanced security for cloud operators in bare metal, virtualized, and containerized environments.

In a typical cloud environment, Virtual Private Cloud (VPC) networks play a pivotal role. They offer a flexible and secure way to connect a tenant's infrastructure resources, ensuring isolation between distinct tenant environments. This functionality is vital for building and operating scalable, reliable, and secure cloud computing infrastructure. BlueField DPUs accelerate VPC networking in AI clouds, delivering performance that is 8x faster than non-accelerated network environments.

Figure 6. Multi-Tenant Networking



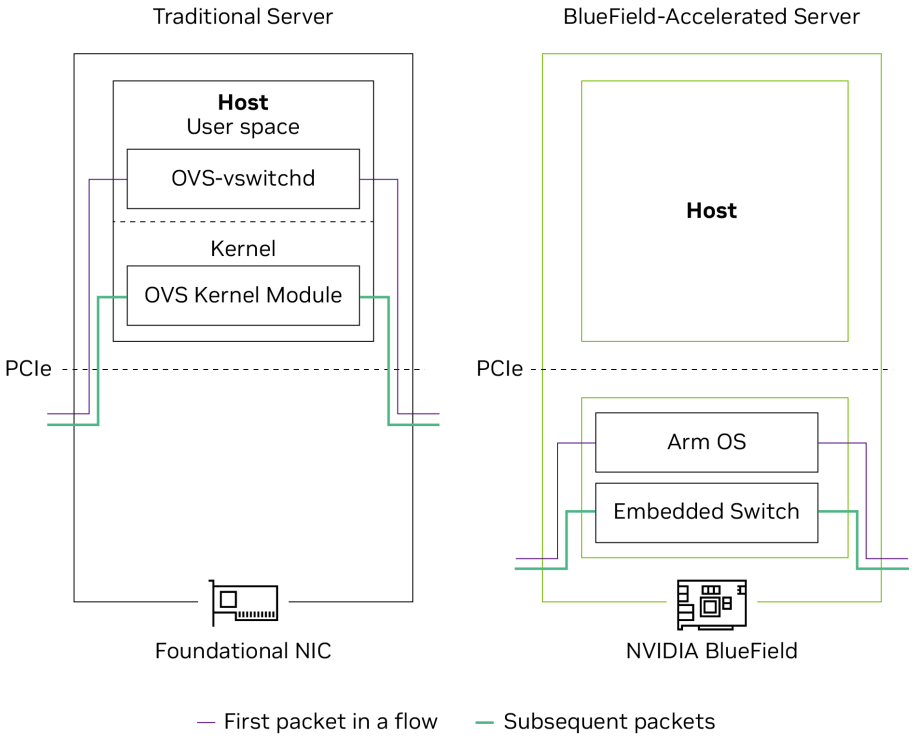
BlueField offers two out-of-the-box paths to create secure, multi-tenant, and high-performance AI cloud network environments:

- An OVS (Open vSwitch) / OVN (Open Virtual Network)-based SDN acceleration solution
- An VXLAN (Virtual Extensible LAN) - EVPN (Ethernet VPN)-based network solution

While both SDN and EVPN VXLAN aim to create multi-tenant networks, they employ different approaches. SDN centralizes control and abstracts network resources, whereas EVPN VXLAN distributes control with a BGP-based control plane coupled with MAC learning. NVIDIA BlueField DPUs fully offload, accelerate, and isolate the control and data-plane of both OVS/OVN and EVPN VXLAN.

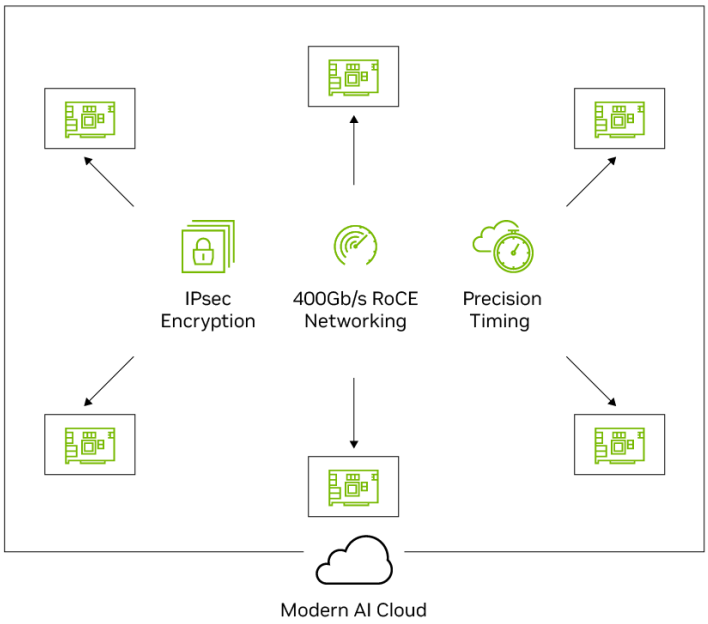
Notably, the software stack runs exclusively on the DPU. HBN is a DOCA networking service designed for creating EVPN VXLAN fabric on the BlueField platform. It enables organizations to deploy it at runtime, gaining the advantage of BlueField-accelerated, multi-tenant cloud network connectivity. Alternatively, organizations using custom networking software can seamlessly integrate BlueField's acceleration into their environments by leveraging DOCA Flow. This networking SDK empowers developers to create accelerated network applications on the BlueField DPU platform.

Figure 7. Software-defined, Hardware-Accelerated



BlueField's VPC network acceleration can also be enhanced with additional functions including IPsec inline encryption, RDMA, and precision timing. This provides secure and robust GPU access to remote storage and secure web application delivery, among other advanced use-cases.

Figure 8. Secure and Robust Networking



NVIDIA's BlueField-3 DPUs accelerate VPC networks within AI cloud environments. These DPUs are adept at supporting the two leading methodologies for multi-tenant networking. By doing so, they deliver solutions for data center connectivity that are not only high in performance, but also efficient in terms of CPU usage.

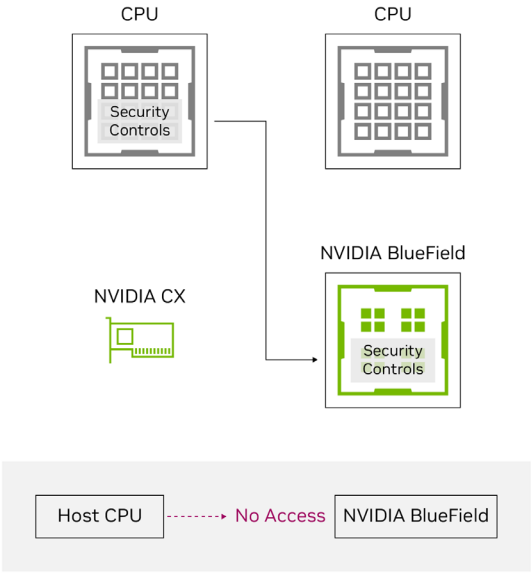
Zero-Trust Security

The rise of GPU-accelerated AI, extending beyond scientific domains, is revolutionizing the future of industries. AI-minded organizations are continuously navigating the security risks in modern data centers. To address these, many are adopting a zero-trust strategy, which involves deploying cybersecurity tools like firewalls and micro-segmentation agents on every server. This shift towards more comprehensive, distributed security strategies is increasing the need for BlueField-3 DPUs in each server.

BlueField, equipped with innovative hardware engines, accelerates security across the entire stack, securing data centers for generative AI workloads. Its built-in isolation restricts access from the host, making BlueField-powered AI systems fundamentally more secure than traditional systems.

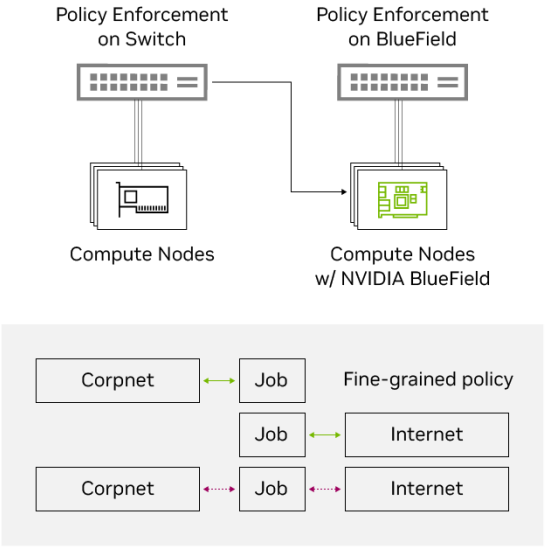
BlueField-3 DPU's comprehensive, zero-trust security approach extends from the data center perimeter to the edge of every server. Operating BlueField in zero-trust mode fortifies the data center security control-plane. If a host is compromised, the isolation between the security control-plane on BlueField, and the compromised host adds another layer of protection. This zero-trust approach makes it easier to isolate the threat and reduces the risk of an exploit spreading, preventing widespread damage within the data center.

Figure 9. Zero-Trust Architecture



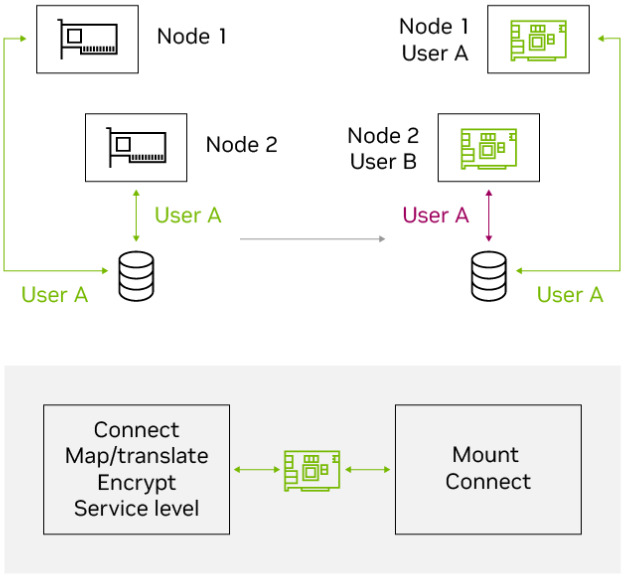
BlueField-3 DPUs also address scenarios where deploying security control agents on a host is impractical or unwanted, a common situation in multi-tenant, bare-metal cloud environments. Enforcing security policies at the top-of-rack switch is very limited and operationally complex. BlueField offers a solution by allowing the deployment of security controls on the DPU in every AI system, enabling distributed, fine-grained security policy enforcement on virtualized, containerized, and bare-metal servers. BlueField enables policy enforcement at the workload level (on the virtual function connecting the virtual machine/container POD), as opposed to the port or network levels.

Figure 10. Distributed, Fine-Grained Security



BlueField DPUs can further enhance data center security by blocking unauthorized user access to data. Integrating BlueField with the workload scheduling system allows it to become aware of user workload placement, enabling it to detect and block fraudulent data access requests.

Figure 11. Data Security



Beyond monitoring and blocking unauthorized access, BlueField's security engines provide robust protection against a myriad of threats. Specifically, they help protect AI cloud data centers from modern cybersecurity threats, by addressing the following functions:

- Enhancing platform security
- Accelerating IPsec and TLS encryption and decryption at line speed
- Enforcing software-defined security policies in hardware
- Performing stateful packet filtering
- Storing and managing keys in-hardware and accelerating PKI exchanges
- Accelerating storage encryption and decryption
- Detecting malicious code and mitigating attacks

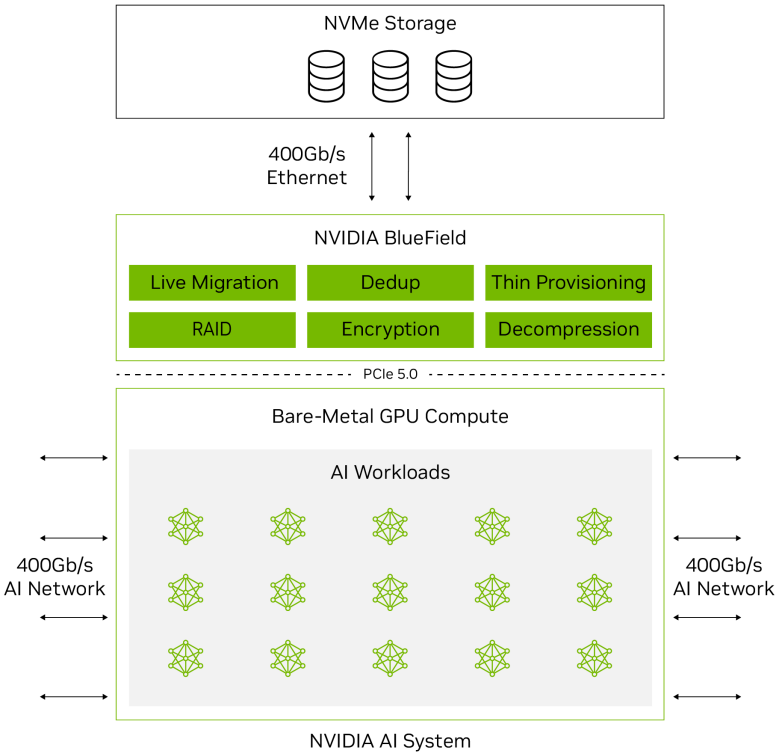
Composable Storage Infrastructure

Organizations integrating AI into their operations often struggle with the complexities of data management. When designing a high-performance AI cloud environment, a delicate balance is required. On one hand, there's a need to optimize tenant workload performance with cloud operators' priorities for composability and protection. This

balance typically involves a choice between over-provisioned, directly attached storage for performance, and thinly provisioned networked storage for flexibility and security.

Recognizing this challenge, NVIDIA has introduced the BlueField SNAP technology and DOCA SNAP Service, allowing organizations to provision networked storage to each AI system without sacrificing application performance. Specifically, BlueField DPUs efficiently offload storage and network tasks from the host CPU, accelerate data movement, and enable intelligent storage functions. By emulating directly attached storage using networked storage, it presents an NVMe or VirtIO-blk device on the PCIe interface. As a result, this setup enables the host to perceive the storage as a standard block storage device, oblivious to its network-based nature, thereby creating a win-win situation. Here, tenant workloads can access a high-performance storage infrastructure at whatever capacity while cloud operators can dynamically provision networked storage to each server as needed.

Figure 12. Cloud Storage Acceleration

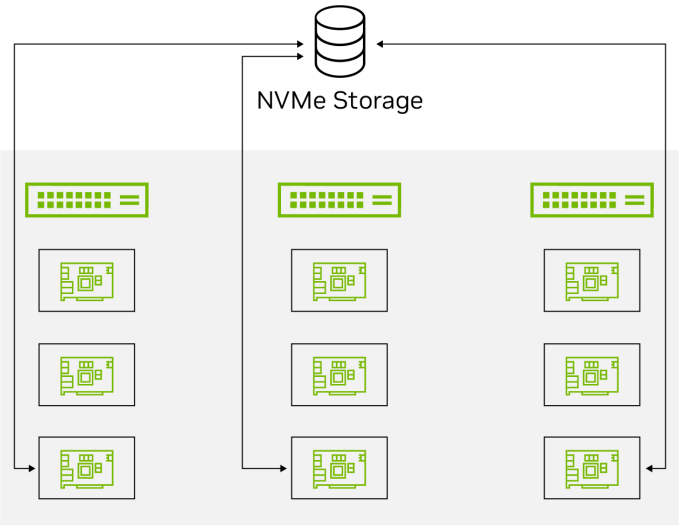


Additionally, BlueField SNAP accelerates GPU compute access to the cloud data store, facilitating AI workloads in efficiently fetching and storing data at astonishing speeds, exceeding 10 million IOPs. In addition, AI servers continue to use their standard operating system's NVMe PCIe driver, minimizing performance degradation while increasing efficiency. With BlueField DPUs, storage is virtualized, thinly provisioned, and consistently backed up. Moreover, this storage can migrate between servers as needed, providing savings in both Capex and OpEx.

Diving deeper into its capabilities, BlueField DPUs enable seamless scalability of storage instances and optimized data operations, offering unparalleled flexibility and performance. Through BlueField SNAP, cloud operators can effortlessly scale storage capacities for various workloads almost instantly. This technology eliminates the need for time-consuming manual processes, such as disengaging a node from the cluster to physically installing hardware. BlueField DPUs also allow compute nodes to boot directly from a remote storage volume, eliminating the reliance on local storage. This enhances flexibility to swap images across multiple nodes, while ensuring data integrity and consistency throughout the entire data center.

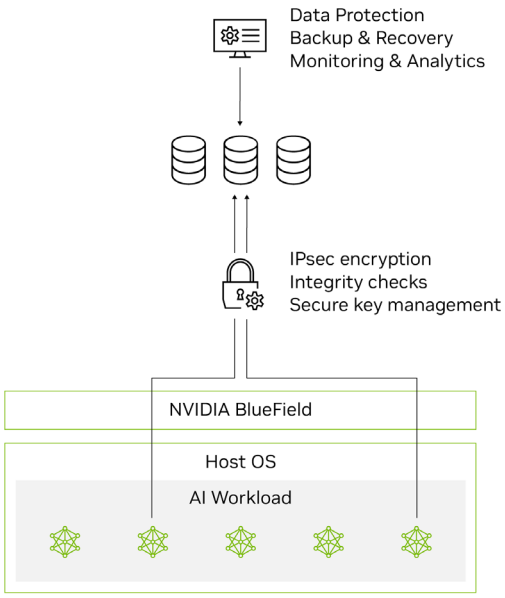
Furthermore, with the incorporation of BlueField DPU into every AI system, an AI cloud can seamlessly deploy a storage infrastructure and centrally serve it without compromising application performance. Centralized storage offers numerous operational advantages. For instance, with data stored centrally, operational teams can easily manage and safeguard data, implement modern backup and recovery methods, and monitor its usage over the network via the DPU. Central block storage is partitioned into smaller, logical drives, for booting a node or local storage, constituting a pool of logical drives that are allocated upon new tenancy and undergo cleanup in the background before being returned to the pool.

Figure 13. Efficient Data Operations



In addition to providing operational excellence, BlueField DPUs also provide a comprehensive suite of data security features, including encryption for both data-in-flight and data-at-rest, coupled with solid encryption key management, and data integrity checks. Rather than providing hosts with unfiltered access to the drives, they access them through a restricted interface, enhancing overall security. This design safeguards firmware, configuration, and data. Encryption keys are managed by agents within secure DPUs, separating them from the host.

Figure 14. Secure Data Fabric



The realm of AI workloads necessitates rapid and efficient data access and storage capabilities. BlueField SNAP technology integrates the advanced functionality of software-defined storage (SDS) with performance that surpasses that of directly attached storage, enabling a high-performance and secure data platform for AI data centers.

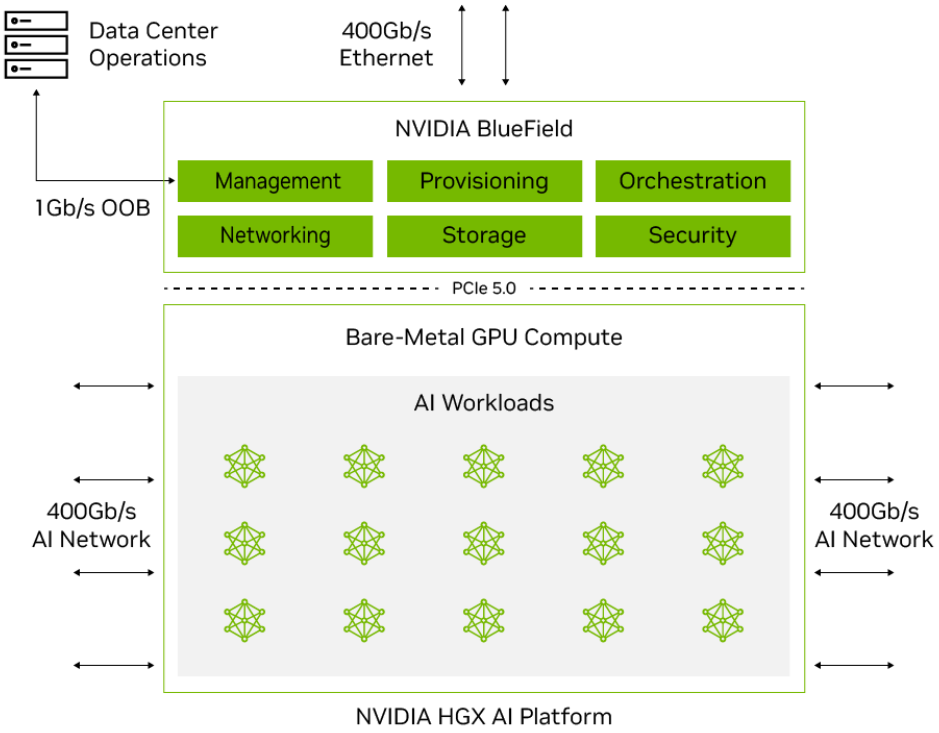
Elastic GPU Computing

GPU-accelerated computing presents a full-stack challenge, and, the transient nature of AI workloads, which requires a lot of computing power for brief periods, adds to this complexity. As AI workloads proliferate, the need for an elastic computing infrastructure that can accommodate changing workloads becomes increasingly urgent.

NVIDIA's BlueField-3 DPUs are revolutionizing the deployment and operation of AI data centers at every scale. BlueField accelerates and streamlines server and network provisioning, speeding-up operations from bring-up to AI training from months to days. Moreover, compared to BlueField's support for booting over networked NVMe block storage traditional server provisioning, which involves iPXE network installation, is much slower and prone to error.

The adoption of BlueField at the data center level saves many hours in initial server setup. Furthermore, BlueField DPUs considerably simplify network provisioning, specifically in EVPN network scenarios. Unlike traditional approaches that require configuring network switches with MLAG/Multihoming, subnets, and gateway addresses, provisioning EVPN on BlueField DPUs instead of network switches eliminate the need for layer 2 configurations on switches. This reduction eliminates the need for hundreds of command lines per switch, streamlining the entire provisioning process.

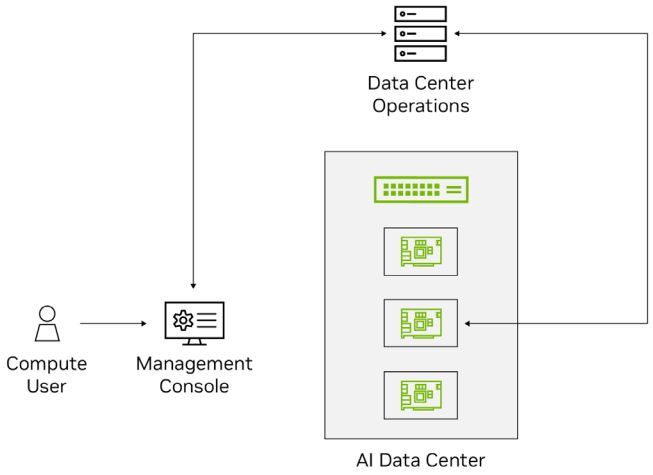
Figure 15. Rapid Provisioning



In the context of modern organizational needs, a flexible and resilient infrastructure capable of adapting quickly to evolving demands is crucial. BlueField DPUs help transform the AI data center infrastructure into an elastic cloud compute platform that can be easily scaled, repurposed, and allocated to tenants in hours instead of weeks.

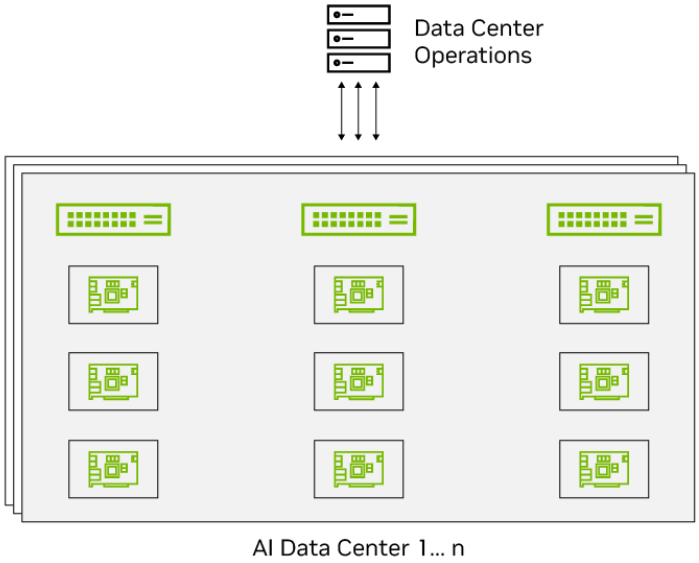
Typically, large enterprises juggle a number of AI initiatives and users that span different domains, teams, or application owners. Importantly, each infrastructure user typically comes with a unique set of infrastructure requirements based on their specific workloads. These requirements can range from the compute type (bare-metal, virtual machines, K8s containers), to operating-system types and versions, number of nodes, storage capacity, and more.

Figure 16. Elastic, Fungible Capacity



In a cloud IaaS environment, BlueField DPUs efficiently manage the provisioning and allocation of the underlying infrastructure, including computing, storage, and networking. This agility leads to faster turnaround times in responding to user demands. It facilitates the rollout of new workloads or seamless scaling of existing ones, translating into significant savings in time and effort for the organization.

Figure 17. Limitless Scaling



Modern AI data centers must be adaptable to rapid changes. BlueField DPUs facilitate this by enabling the elastic provisioning of GPU resources, allowing organizations to operationalize AI faster. NVIDIA has integrated BlueField DPUs into its reference architecture, providing a turn-key AI data center solution that combines world-class computing, advanced software tools, expertise, and the seamless delivery of ongoing innovation.

Conclusion

Organizations universally face challenges in integrating AI into their operations. NVIDIA's accelerated systems, equipped with BlueField-3 DPUs, provide the optimal infrastructure to power generative AI applications. The incorporation of BlueField DPUs into every system enables organizations to securely deploy and operate NVIDIA AI cloud data centers. This integration not only accelerates the development and deployment of generative AI solutions but also reduces time-to-market, offering a clear pathway to monetization. BlueField-3 DPUs are equipped with a range of essential capabilities, including accelerated VPC networking, state-of-the-art zero-trust security, composable storage solutions, and elastic GPU computing. These integrated functionalities collectively empower the creation of secure, high-capacity, multi-tenant AI cloud data centers.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. Neither NVIDIA Corporation nor any of its direct or indirect subsidiaries and affiliates (collectively: "NVIDIA") make any representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, and BlueField, 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.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com