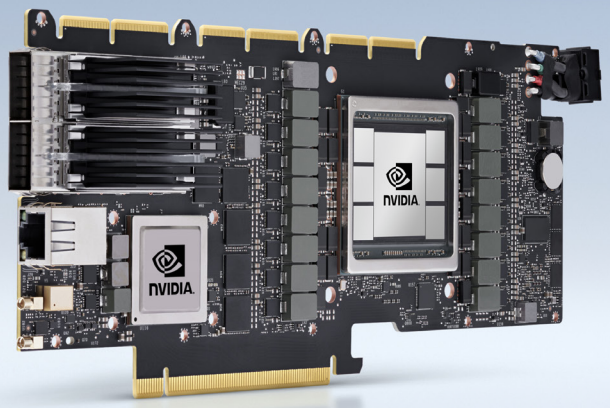
# CONVERGED ACCELERATORS

Networking and Compute, Unified



## A More Powerful, Secure Enterprise

NVIDIA converged accelerators combine the power of the NVIDIA Ampere architecture with the enhanced security and networking capabilities of the NVIDIA® BlueField®-2 data processing unit (DPU), all in a single high-performance package. This advanced architecture delivers unprecedented performance and strong security for GPU-powered workloads in edge computing, telecommunications, and network security.

## Better Performance

Because the NVIDIA Ampere architecture GPU and the BlueField-2 DPU are connected via an integrated PCIe Gen4 switch, there's a dedicated path for data transfer between the GPU and the network. This eliminates performance bottlenecks of data going through the host. It also enables much more predictable performance, which is important for time-sensitive applications such as 5G signal processing.

## Enhanced Security

The convergence of NVIDIA's GPU and DPU creates a more secure AI processing engine, where data generated at the edge can be sent across the network fully encrypted without traveling over the server PCIe bus, ensuring it's isolated from the host. This helps provide better protection for the host from network-based threats.

## Smarter Networking

The architecture of the NVIDIA converged cards allows GPU processing to be applied directly to traffic as it flows to and from the DPU. This enables a whole new class of applications that involve AI-based networking and security, such as data leak detection, network performance optimization and prediction, and more.

## Cost Savings

Because the GPU, DPU, and PCIe switch are combined together on a single card, customers can leverage mainstream servers to perform tasks previously only possible with high-end or purpose-built systems. Even edge servers can benefit from the same performance boost that's more typically found in specialized systems.

### KEY COMPONENTS

- NVIDIA A100 / A30 Tensor Core GPUs
- NVIDIA BlueField-2 DPU
  - NVIDIA ConnectX®-6 Dx
  - 8 Arm A72 cores at 2GHz
- Integrated PCIe Gen4 switch

### TOP USE CASES

- 5G VRAN
- AI-based cybersecurity
- AI on 5G

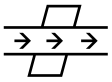
### SYSTEM SPECIFICATIONS

| Product | A100X | A30X |
|---|---|---|
| GPU memory | **80GB HBM2e** | **24GB HBM2e** |
| Memory bandwidth | **1,935GB/s** | **1,161GB/s** |
| Multi-Instance GPU (MIG) instances | **7** | **4** |
| Interconnect | **PCIe Gen4 (x16 physical, x8 electrical)**<br>**3x NVIDIA NVLink® bridge** | **PCIe Gen4 (x16 physical, x8 electrical)**<br>**1x NVIDIA NVLink® bridge** |
| Networking | **2x 100Gbps ports, Ethernet or Infiniband** | **2x 100Gbps ports, Ethernet or Infiniband** |
| Form factor | **Dual-slot full-height, full-length (FHFL)** | **Dual-slot full-height, full-length (FHFL)** |
| Max power | **300W** | **230W** |

| Use Case | Benefits Of Converged Accelerators |
|---|---|
| 5G vRAN | > Data no longer needs to go through the CPU and host PCIe system, greatly reducing latency.<br>> Higher throughput increases subscriber density per server. |
| AI-on-5G | > Reduced latency for 5G signal processing.<br>> High performance in a single card allows for compact and lower-cost servers. |
| AI-based cybersecurity | > GPU-based AI can be applied directly to network traffic with a high data rate.<br>> Data travels on an isolated path between the DPU and GPU. |

### Better Performance

Dedicated path for GPU network data transfer



### Enhanced Security

Avoid data flow across host PCIe systems

### Smarter Networking

Apply GPU processing directly to traffic flows

### Cost Saving

Acceleration power with mainstream servers

## Products

NVIDIA converged accelerators are available in two form factors.

### A30X

The A30X combines the NVIDIA A30 Tensor Core GPU with the BlueField-2 DPU. The design of this card provides a good balance of compute and input/output (IO) performance for use cases such as 5G vRAN and AI-based cybersecurity. Multiple services can run on the GPU, with the low latency and predictable performance provided by the onboard PCIe switch.

### A100X

The A100X brings together the power of the NVIDIA A100 Tensor Core GPU with the BlueField-2 DPU. It's ideal for workloads where compute demands are greater. Examples include 5G with massive multiple-input, multiple-output (MIMO) capabilities, AI-on-5G deployments, and specialized workloads such as signal processing and multi-node training

## Developer Ecosystem

NVIDIA converged accelerators expand the capabilities of the CUDA® and NVIDIA DOCA™ programming libraries for workload acceleration and offloading. CUDA applications can be run on the x86 host or on the DPU's Arm processor for isolated AI and inferencing applications. NVIDIA is also introducing a converged accelerator development kit, allowing selected partners and customers to receive a free A30X converged accelerator, early access to documentation, and sample GPU+DPU applications.

**Learn more**

To learn more about NVIDIA converged accelerators, visit
**nvidia.com/en-us/data-center/products/egx-converged-accelerator/**



NVIDIA