6WIND
Virtualized Networking Software

Networking for AI Solution Brief

Offload | Accelerate | Isolate

Challenges

- *High packet processing rates*
- *Bursty, unpredictable traffic*
- *Strict compliance and isolation*
- *Dynamic service chaining*
- *Cost-efficient scaling*

Benefits

- *Boosted GPU Utilization*
- *Faster AI Services*
- *Zero-Trust Multi-Tenancy*
- *Elastic Scaling*
- *Lower Infrastructure Costs*
- *Smarter Cost Control*

Use Cases

- *Inference Service Chaining*
- *GPU-aaS Platforms*
- *Enterprise AI*
- *Hybrid & Sovereign AI*
- *Compliance-Driven Pipelines*

Executive Summary

Artificial Intelligence (AI) is transforming industries, from real-time assistants and multimodal inference to sovereign AI deployments and GPU-as-a-Service (GPU-aaS) platforms. But achieving AI's potential requires more than powerful GPUs—it demands **frontend networking** that is secure, scalable, and optimized for AI workloads.

6WIND's **Networking for A.I.** solution, developed in collaboration with NVIDIA, combines the **Virtual Service Router (VSR)** and **Host Networking Accelerator (HNA)**, designed for **BlueField-3 DPUs**, to deliver:

- **Low-latency connectivity** between users, APIs, data sources, and GPU clusters.
- **Multi-tenant isolation** with zero-trust enforcement.
- **Elastic scalability** for Kubernetes-based AI environments.

Challenges

Modern AI deployments face unique networking pressures that traditional architectures cannot address:

- **High packet processing rates** from multi-user inference APIs.
- **Bursty, unpredictable traffic** driven by prompt-based AI interactions.
- **Strict compliance and isolation** requirements for multi-tenant GPU environments.
- **Dynamic service chaining** for in-line security, logging, and policy enforcement.
- **Cost-efficient scaling** aligned to AI service economics.

Legacy CPU-forwarding paths and static hardware appliances become bottlenecks, leading to wasted GPU capacity and higher infrastructure costs.

Solution

6WIND Networking for A.I. brings hyperscaler-grade networking to any AI infrastructure without proprietary lock-in.

Virtual Service Router (VSR)

- Acts as the secure ingress/egress point for AI clusters.
- Provides IPsec encryption, NAT, firewall, and VPC interconnects.
- Integrates natively with Kubernetes networking for containerized AI workloads.

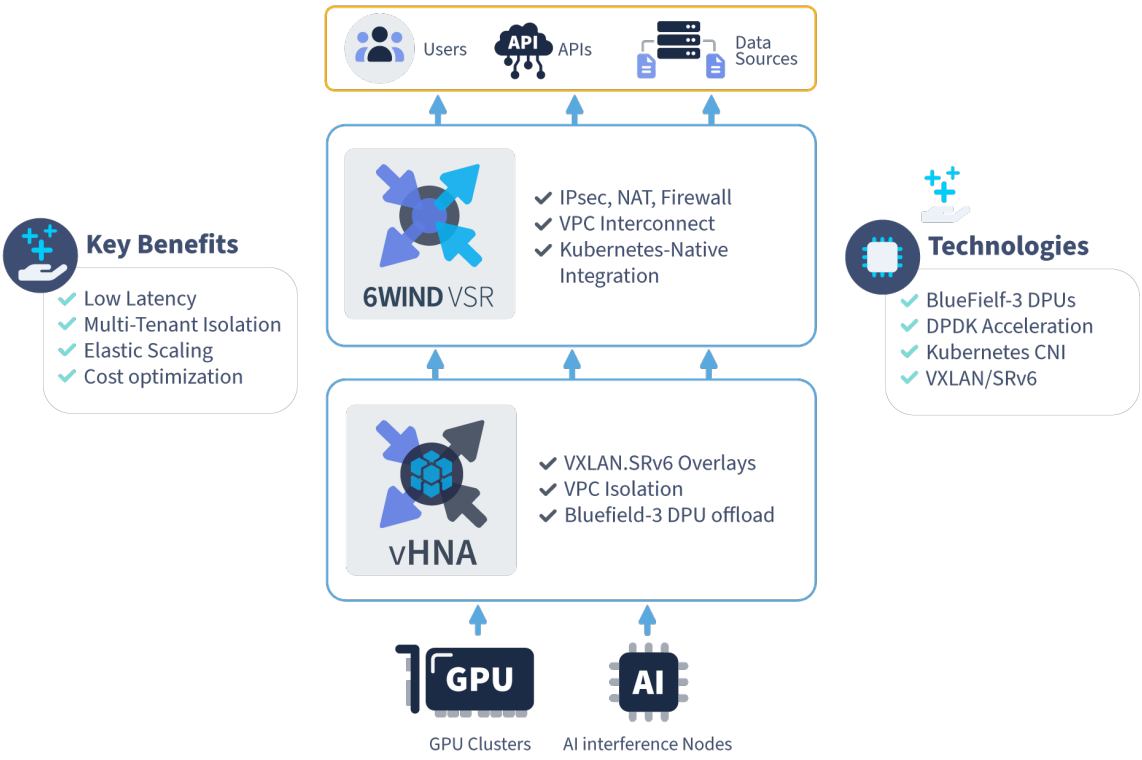
Host Networking Accelerator (HNA)

- DPDK-powered router with VXLAN and SRv6 overlay support.
- Ensures VPC-level isolation within Kubernetes clusters.
- Offloads networking tasks to NVIDIA BlueField-3 DPUs, freeing up to 20 CPU cores per server for AI workloads.

Kubernetes-Native Integration

- Deploys as a CNI plugin.
- Supports declarative, real-time orchestration of routing, ACLs, and segmentation policies.

6WIND Networking for A.I. – Architecture Overview



Here's the **architecture overview diagram** *showing the flow from* **Users/APIs** *through* **VSR → HNA → GPU Clusters**, *with side callouts for benefits and technologies*

Key Benefits

- Boosted GPU Utilization – Reclaims CPU resources for model training and inference.
- Faster AI Services – Sub-millisecond latency with predictable throughput.
- Zero-Trust Multi-Tenancy – VRF segmentation, ACLs, and L4 firewalling ensure tenant isolation.
- Elastic Scaling – Policies and routing adapt automatically to workload changes.
- Lower Infrastructure Costs – Software-defined, DPU-accelerated design replaces costly hardware appliances.
- Smarter Cost Control – Prompt-aware routing matches requests to the most cost-efficient or policy-compliant GPU resource.

Use Cases

- Inference Service Chaining – Apply security, logging, and policy controls before requests hit the GPU node.
- GPU-aaS Platforms – Deliver tenant-isolated, SLA-backed network performance.
- Enterprise AI – Dynamically segment AI resources between departments.
- Hybrid & Sovereign AI – Maintain compliance with data residency laws while delivering seamless performance.
- Compliance-Driven Pipelines – Inline encryption, telemetry, and audit logging for regulated industries.

Why 6WIND

- **Performance First** – Built for deterministic low-latency networking at scale.
- **Proven Integration** – Certified with NVIDIA BlueField-3, DGX systems, and Red Hat OpenShift.
- **Flexible Deployment** – Bare metal or virtualized, at edge, core, or cloud.
- **Security by Design** – Zero-trust architecture with full multi-tenant isolation.

Further Information

- 6WIND Virtual Service Router (VSR) – Product Datasheet
- 6WIND Host Networking Accelerator (HNA) – Product Datasheet
- 6WIND Networking for AI – Solution Page