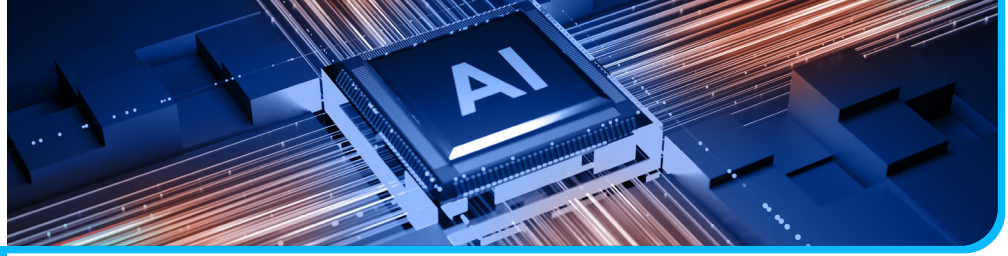
Virtualized Networking Software



6WIND AI-RAN Solution Brief

Accelerating Edge Intelligence with Distributed UPF and NVIDIA Grace Hopper

Challenges

- *Ultra-Low Latency Services*
- *Centralized UPF Bottlenecks*
- *Edge Compute Requirements*
- *Energy & ESG Compliance*

Benefits

- *Monetize Edge AI*
- *Lower CAPEX & OPEX*
- *Accelerate Time-to-Market*
- *Sustainability Gains*
- *Vendor Independence*

Conclusion

- *Push intelligence and packet processing to the edge*
- *Support demanding AI-native applications*
- *Reduce costs, latency, and carbon footprint*

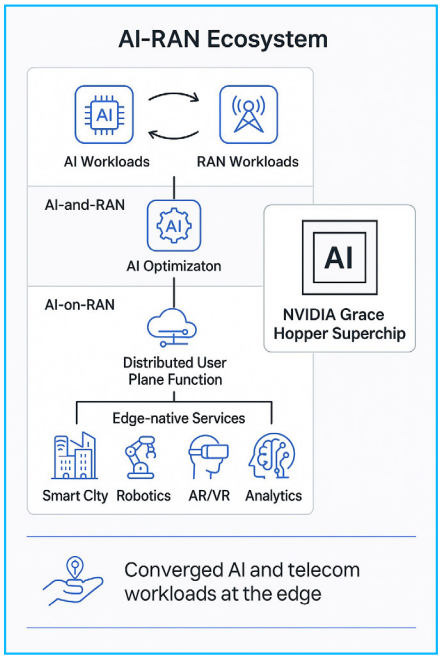
Executive Summary

AI-RAN is redefining the Radio Access Network by embedding Artificial Intelligence directly into the RAN architecture. This enables **real-time inferencing, context-aware automation, and edge-native services** with ultra-low latency.

In partnership with **NVIDIA**, 6WIND delivers a **Distributed User Plane Function (dUPF)** optimized for the **NVIDIA Grace Hopper Superchip** — converging telecom and AI workloads in a **single, compact, power-efficient edge platform**.

This joint solution delivers **sub-30µs latency** at **100Gbps** throughput using just **4 CPU cores**, allowing operators to:

- **Monetize AI at the edge** through inference-as-a-service, smart city analytics, and immersive applications.
- **Reduce TCO** by replacing centralized packet processing with distributed, software-based UPFs on COTS hardware.
- **Achieve sustainability targets** with 80%+ lower power per Gbps compared to legacy UPF appliances



Market Challenges

Operators face unprecedented demands as 5G and emerging 6G architectures evolve:

CHALLENGE	IMPACT
Ultra-Low Latency Services	AR/VR, autonomous robotics, and agentic AI require sub-millisecond performance.
Centralized UPF Bottlenecks	Creates backhaul congestion, increases latency, and raises transport costs.
Edge Compute Requirements	AI workloads demand high memory bandwidth, parallel processing, and minimal footprint.
Energy & ESG Compliance	Need to deliver higher performance with lower energy use and carbon footprint.

6WIND AI-RAN Solution Architecture

Distributed UPF (dUPF)

- **Cloud-Native & Flexible**: Runs on x86 or ARM, deployed on bare metal, VMs, or Kubernetes.
- **Standards-Compliant**: 3GPP N3/N4/N6/N9 interface support.
- **Performance-Optimized**: 100Gbps at <80W power draw with 4 data plane cores.
- **Edge Breakout**: Local breakout for real-time AI services reduces backhaul and improves responsiveness.
- **Security-First**: Integrated ACLs, VRF segmentation, IPsec tunneling.

NVIDIA Grace Hopper Superchip

- **Unified AI & Networking Node**: Combines GPU and CPU in one platform for co-located AI inference and packet processing.
- **High Memory Bandwidth**: Ideal for AI models requiring rapid data access.
- **Parallel Workload Execution**: Runs GTP-U forwarding, inference, and analytics simultaneously.
- **Compact & Efficient**: Optimized for space- and power-constrained edge sites.

Integration & Orchestration

- Kubernetes-native deployment with Helm charts for rapid provisioning.
- Open APIs for integration with telco orchestration (ONAP, Nephio) and CI/CD pipelines.
- Real-time telemetry with gNMI, Netconf, and streaming analytics.

Key Technical Advantages Over Centralized UPF

METRIC	CENTRALIZED UPF	6WIND DUPF ON GRACE HOPPER
Latency	3–10 ms	<0.03 ms
Throughput Scaling	Requires scaling central core	Scales horizontally at the edge
Transport Cost	High (backhaul AI traffic)	Low (local breakout)
Power per Gbps	2–4x higher	80% lower
Deployment Time	Weeks–Months	Hours–Days

Business Benefits

- **Monetize Edge AI**: New services such as video analytics, industrial robotics, and AR/VR training can be sold with premium SLAs.
- **Lower CAPEX & OPEX**: Reduce server count, transport costs, and power usage with compact, distributed deployments.
- **Accelerate Time-to-Market:** Deploy AI-enabled services at the edge in hours with containerized UPF + AI workloads.
- **Sustainability Gains**: 80%+ lower energy per Gbps and reduced cooling requirements meet ESG and regulatory goals.
- **Vendor Independence**: COTS hardware + software-based UPF avoids proprietary lock-in.

Use Case Scenarios

USE CASE	VALUE DELIVERED
Edge Agentic AI	Real-time, context-aware inference for automation and decision-making at the edge.
Autonomous Systems	Low-latency connectivity for UAVs, connected vehicles, and robotics.
Smart Cities	Decentralized AI for traffic, environmental, and security analytics.
AR/VR & Immersive Media	Consistent, ultra-low latency experiences for entertainment and training.
Private AI-RAN	Secure, scalable enterprise 5G/6G with localized AI.
AI Data Lakes at the Edge	Pre-process and store data locally to cut cloud costs and improve real-time insights.

Conclusion

6WIND and NVIDIA have created a **blueprint for AI-powered RAN architectures** that meets the performance, scalability, and sustainability demands of modern telecom networks.

With **6WIND's dUPF on NVIDIA Grace Hopper**, operators can:

- Push intelligence and packet processing to the edge.
- Support demanding AI-native applications.
- Reduce costs, latency, and carbon footprint.

The result: a future-ready AI-RAN platform capable of enabling the **6G era** with distributed intelligence at its core.

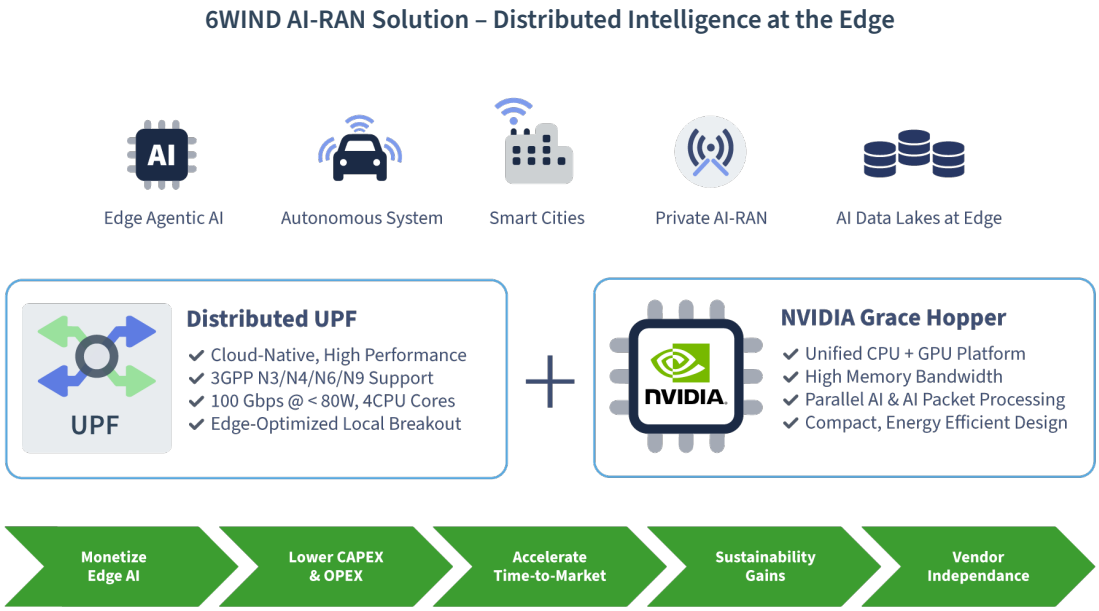


Diagram: 6WIND AI-RAN architecture showing dUPF + NVIDIA Grace Hopper integration for co-located AI inference and telecom packet processing at the RAN edge.