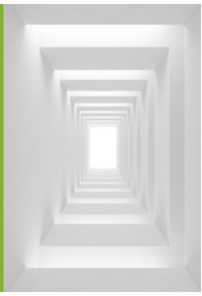
Juniper® Validated Design

JVD Solution Overview: AI Data Center Multitenancy with EVPN/VXLAN—Juniper Validated Design



sol-overview-JVD-AICLUSTERDC-EVPNType5-01-04

Executive Summary

Designing infrastructure for AI services introduces unique challenges, especially as the use of high-performance GPUs for AI training and inference drives massive data volumes and stringent performance demands. While traditional data center networking principles remain relevant, AI networks must now support large-scale traffic flows, minimize latency and packet loss, and enable predictable workload completion times. Meeting these requirements requires not only high-performance hardware and software, but also precise configuration and design.

Juniper Networks addresses these challenges by delivering scalable, standards-based, and automated solutions tailored for AI clusters of all sizes. This **JVD** provides validated designs and best practices for deploying EVPN/VXLAN fabrics that enable **GPU as a Service (GPUaaS)** and multitenancy in AI data centers. Juniper's architecture accelerates innovation, ensures design flexibility, and avoids vendor lock-in across backend, frontend, and storage fabrics. Advanced JVD testing by Juniper subject matter experts validates optimized designs that deliver high performance, reliability, and simplified operations.

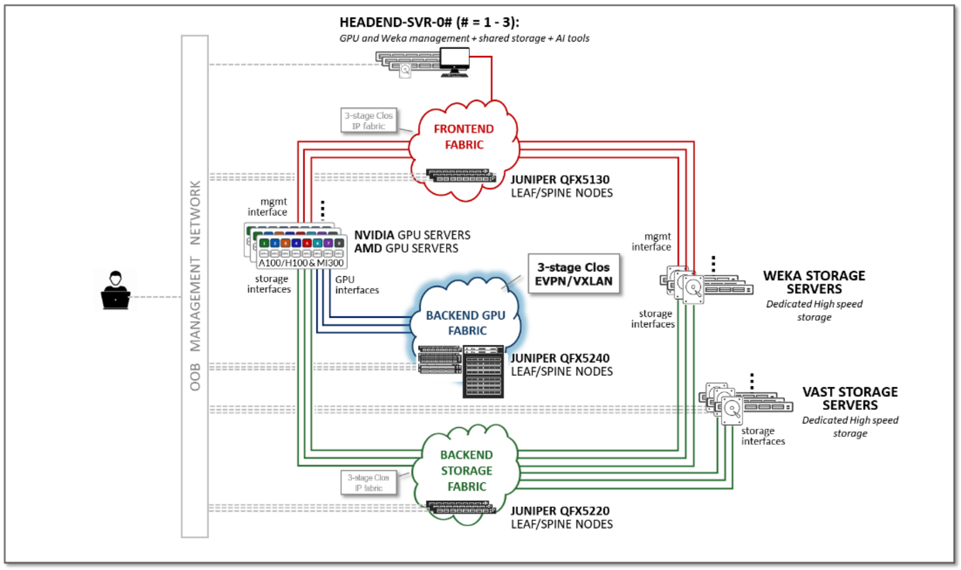
Solution Overview

The AI Data Center Multitenancy with EVPN/VXLAN JVD defines best practices, solution components, and configuration guidelines for deploying an **EVPN/VXLAN GPU backend fabric infrastructure**—based on **Juniper Networks QFX Series** high-performance switches—that supports **GPU as a Service (GPUaaS)** in AI clusters.

The complete AI Data Center solution consists of three key fabrics:

- Front-end Fabric
- Storage Backend Fabric
- GPU Backend Fabric

Figure 1: AI JVD Reference Design



This JVD focuses on the implementation of a **GPU backend fabric** specifically designed to support **GPU as a Service (GPUaaS)**

GPU as a Service (GPUaaS) provides on-demand access to GPU compute resources, allowing users and applications to dynamically allocate GPUs based on workload needs without managing the underlying infrastructure. By abstracting physical hardware, GPUaaS offers a scalable, flexible platform for AI training, data analytics, visualization, and more. It enables multiple teams or projects to efficiently share data center resources with centralized management and secure isolation for consistent performance.

GPU multitenancy enhances GPUaaS by allowing multiple tenants to share GPU resources within a common infrastructure. Instead of dedicating entire servers to a single user or team, GPUs can be flexibly assigned—even across servers—with each tenant operating in an isolated environment. This approach ensures efficient resource utilization, scalability, and cost-effectiveness, securely distributing GPU capacity across diverse workloads and scaling resources as needed.

To support GPU multitenancy, the GPU backend fabric must provide seamless communication between GPUs assigned to each tenant, while maintaining traffic separation between tenants, and ensuring optimal performance. The fabric must also support both Server Isolation multitenancy (where one or more servers are assigned to a tenant) and GPU Isolation multitenancy (where individual GPUs within a server are allocated to different tenants).

Figure 2: GPU as a Service – Server isolation

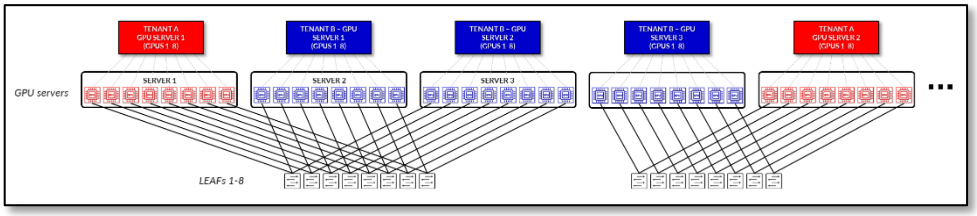
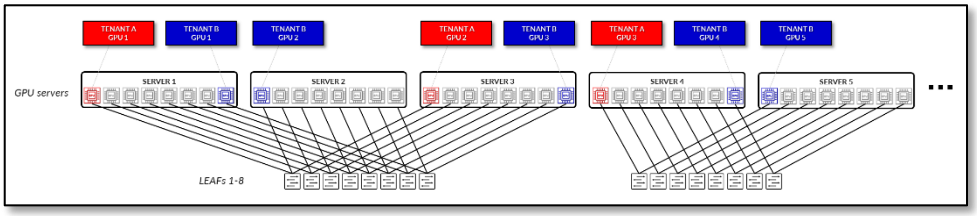

TENANT A GPU SERVER 1 (GPUS 1 8)
TENANT B - GPU SERVER 1 (GPUS 1 8)
TENANT B - GPU SERVER 2 (GPUS 1 8)
TENANT B - GPU SERVER 3 (GPUS 1 8)
TENANT A GPU SERVER 2 (GPUS 1 8)
GPU servers
SERVER 1
SERVER 2
SERVER 3
...
LEAFs 1-8

Figure 3: GPU as a Service – GPU isolation

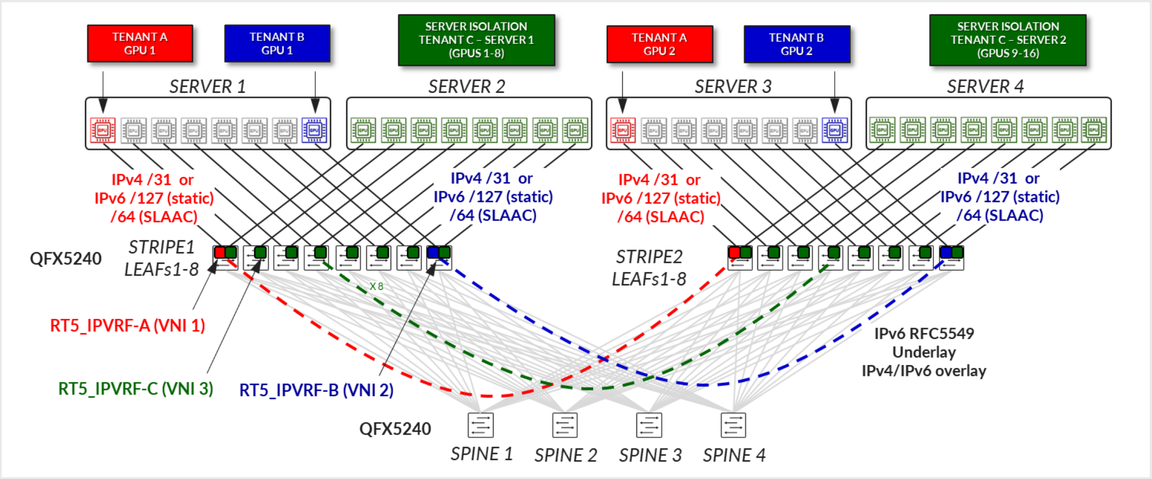


EVPN-VXLAN serves as the foundation for scalable multitenant environments in the GPU backend fabric, supporting two primary design approaches: pure Type 5 services with IP-VRFs only, and VLAN-aware services with MAC-VRFs and symmetric IRB.

While both models provide tenant isolation, the Type 5 model is often better suited for large-scale AI training environments, where jobs span multiple servers and GPUs operate as a single, tightly coordinated unit. These workloads communicate directly over IP using high-throughput, low-latency paths, making Type 5's streamlined IP routing and avoiding MAC learning ideal for performance and operational simplicity.

The EVPN-VXLAN Type 5 solution enables flexible segmentation to support both GPU-level isolation and server-level isolation. It allows operators to map network isolation boundaries directly to the underlying resource allocation model—whether GPUs are shared across tenants on the same server, or entire servers are dedicated to a single tenant.

Figure 4: EVPN/VXLAN type 5 solution for multitenancy


TENANT A GPU 1
TENANT B GPU 1
SERVER ISOLATION TENANT C - SERVER 1 (GPUS 1-8)
TENANT A GPU 2
TENANT B GPU 2
SERVER ISOLATION TENANT C - SERVER 2 (GPUS 9-16)
SERVER 1
SERVER 2
SERVER 3
SERVER 4
IPv4 /31 or IPv6 /127 (static) /64 (SLAAC)
IPv4 /31 or IPv6 /127 (static) /64 (SLAAC)
IPv4 /31 or IPv6 /127 (static) /64 (SLAAC)
IPv4 /31 or IPv6 /127 (static) /64 (SLAAC)
QFX5240
STRIPE1 LEAFs1-8
STRIPE2 LEAFs1-8
X8
RT5_IPVRF-A (VNI 1)
RT5_IPVRF-C (VNI 3)
RT5_IPVRF-B (VNI 2)
IPv6 RFC5549 Underlay IPv4/IPv6 overlay
QFX5240
SPINE 1
SPINE 2
SPINE 3
SPINE 4

Performance is further enhanced by combining a **Rail-Optimized Stripe Architecture** with advanced congestion management and load balancing mechanisms, including **DCQCN (Data Center Quantized Congestion Notification)**, **Priority-Based Flow Control (PFC)**, **Explicit Congestion Notification (ECN)**, and **Dynamic Load Balancing (DLB)**.

Corporate and Sales Headquarters
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or +1.408.745.2000
Fax: +1.408.745.2100
www.juniper.net

APAC and EMEA Headquarters
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: +31.207.125.700
Fax: +31.207.125.701