Juniper® Validated Design

JVD Solution Overview: AI Data Center GPU Backend Fabric with Juniper RDMA-aware LB and BGP-DPF —Juniper Validated Design

sol-overview-JVD-AICLUSTERDC-UEFQPP-01-01

Executive Summary

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

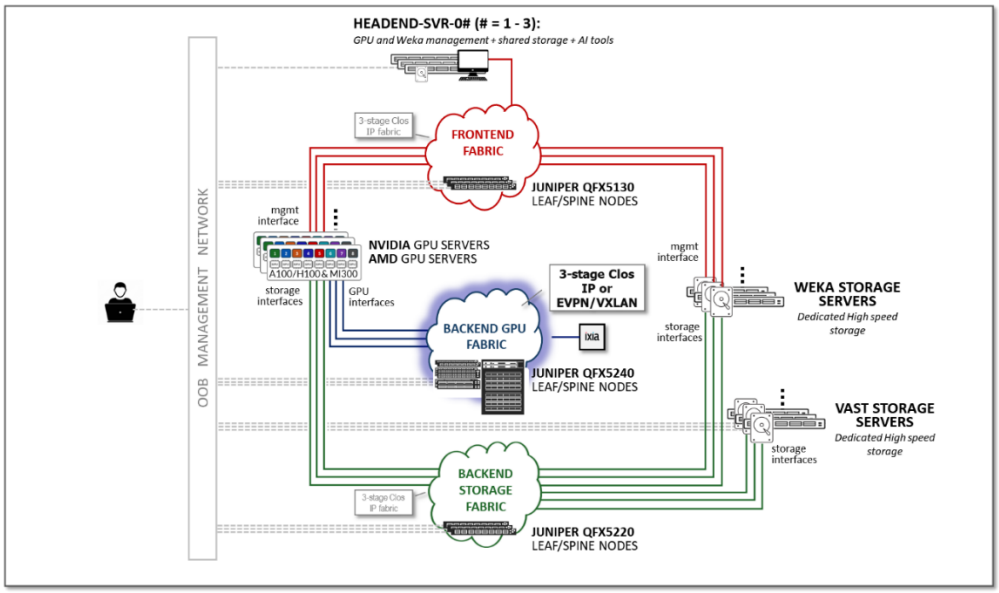
Juniper Networks addresses these challenges by delivering scalable, standards-based, and automated solutions tailored for AI clusters of all sizes. This Juniper Validated Design (JVD) describes the implementation of an AI cluster featuring Juniper **RDMA-aware Load Balancing (RLB)** and **BGP Deterministic Path Forwarding (BGP-DPF)** in the GPU backend fabric, replacing traditional, ineffective Equal-Cost Multi-Path (ECMP) load balancing implementations. This architecture is based on AI-optimized **Juniper QFX5240** switches and **NVIDIA H100 DGX systems**. The solution uses standard Ethernet, RoCEv2, and BGP protocols with no proprietary hardware or

Juniper's solution accelerates innovation, ensures design flexibility, and avoids vendor lock-in across backend, frontend, and storage fabrics. All designs are validated through rigorous testing by Juniper subject matter experts to ensure performance, reliability, and operational simplicity.

Solution Overview

Juniper's complete AI data center validated solution includes three key fabrics: the **Frontend Fabric**, **Storage Backend Fabric**, and **GPU Backend Fabric**, as shown in Figure 1.

Figure 1: AI JVD Reference Design



This Juniper Validated Design focuses on the implementation of a GPU backend fabric that leverages **Juniper's RDMA-aware Load Balancing (RLB)** and **BGP Deterministic Path Forwarding (BGP-DPF)**, along with **Stateless Address Autoconfiguration (SLAAC)** and **BGP auto-discovery** using **IPv6 Neighbor Discovery**, to deliver consistent performance at scale with minimal configuration.

Details on the **Frontend** and **Storage Fabric** architectures are covered in previous AI Data Center JVDs.

AI and ML workloads require predictable, low-latency, and high-throughput communication between GPU nodes. Traditional ECMP load balancing often fails to evenly distribute RDMA traffic across available fabric paths, leading to congestion, packet reordering, and inconsistent performance.

To overcome these limitations, Juniper has introduced more effective load balancing mechanisms, including **RDMA-aware Load Balancing (RLB)** combined with **BGP Deterministic Path Forwarding (BGP-DPF).** This transforms RoCEv2 RDMA traffic handling into a routing-based solution rather than relying on hash-based algorithms. This approach allows each flow to be steered intentionally using standard BGP route selection mechanisms, eliminating randomness and ensuring in-order delivery and consistent behavior. In the event of a link failure, the system automatically falls back to **Juniper's Dynamic Load Balancing (DLB),** allowing traffic to be rerouted and AI jobs to complete, with minimal disruption.

Key Benefits

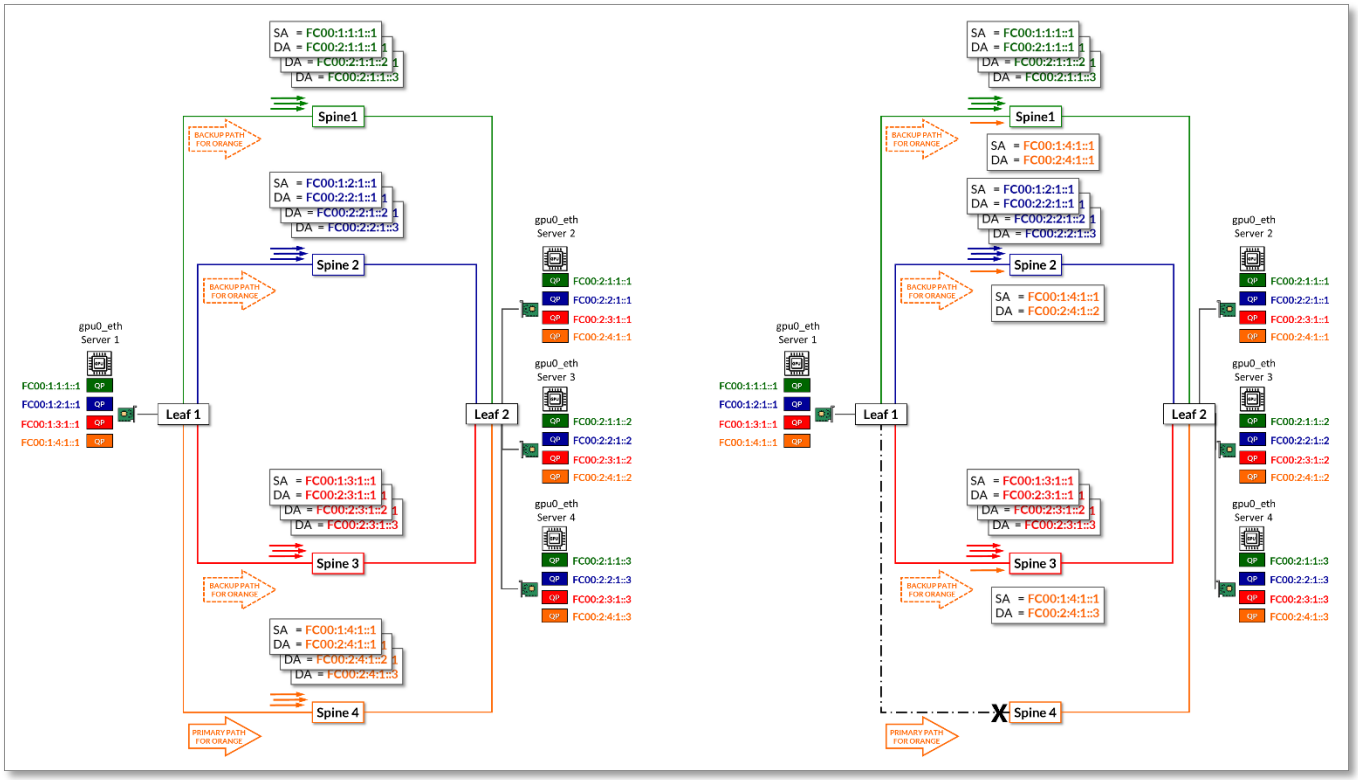
- Predictable performance
- In-order delivery
- Simplified troubleshooting
- Improved throughput for AI/ML workloads
- Hardware-agnostic, standards-based design

At the core of this design is the creation of multiple **RDMA Queue Pairs (QPs)** between the same GPUs, each mapped to different paths within the GPU backend fabric. Each GPU server NIC implements IPv6 SLAAC and automatically generates a set of IPv6 addresses, which are formed by combining a set of IPv6 prefixes received from the leaf nodes via Router Advertisements with the NIC's MAC address in EUI-64 format.

A Juniper-developed NCCL plug-in splits each RDMA transfer into multiple sub-flows, each associated with a different QP, and maps each QP to one of the SLAAC-generated IPv6 addresses.

In the fabric, leaf and spine nodes establish BGP sessions, each associated with a fabric color (e.g., green, blue, red). Routes are tagged using standard BGP extended communities that correspond to the color of the peer to which the routes are advertised. This essentially "colors" the paths across spine nodes, creating deterministic, color-aware, per-prefix forwarding across the fabric, controlled entirely through BGP routing decisions.

To enforce deterministic path selection, Juniper uses AIGP (Accumulated IGP), a standard BGP path attribute defined in RFC 7311. AIGP allows BGP to carry a metric similar to IGP cost. Routes for each IPv6 prefix are advertised with AIGP only to their preferred spine peer and without AIGP (treated as having infinite cost) to others. When leaf nodes receive multiple copies of the same route, the one with the lowest AIGP value is selected. This ensures both forward and return traffic follow the same path, enabling deterministic and symmetric flow steering across the fabric, without manual IP configuration or complex routing policies. When the preferred path fails, traffic is automatically distributed across backup paths using DLB.



To further improve scaling and simplify configuration, BGP sessions between leaf and spine nodes are also formed automatically using **Juniper's BGP auto-discovery,** also referred to as BGP auto-peering. When IPv6 is enabled on the interfaces, leaf and spine nodes use Neighbor Discovery to detect directly connected peers and dynamically form BGP sessions. No static neighbor IP configuration is required. Sessions are established using minimal configuration: the local AS number, a list of acceptable remote AS numbers, IPv6 ND,

Both BGP auto-discovery and SLAAC leverage Junos OS support for:

- RFC 4861: IPv6 Neighbor Discovery
- RFC 2462: IPv6 Stateless Address Autoconfiguration

Performance is further enhanced by combining by configuring **DCQCN (Data Center Quantized Congestion Notification**), which include **Priority-based Flow Control (PFC), Explicit Congestion Notification (ECN).**

Together, these technologies deliver a fabric that distributes traffic efficiently and predictably avoiding congestion, supporting in-order delivery, and keeping GPU-to-GPU communication on track. The result is a scalable, high-performance infrastructure purpose-built for AI and ML workloads, without requiring complex setup or specialized hardware.

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

Send feedback to: design-center-comments@juniper.net V1/250929