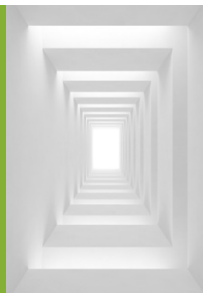
Juniper® Validated Design

# JVD Solution Overview: AI Data Center Frontend Fabric for Inference with HPE Juniper QFX switches, Apstra Data Center Director, and AMD MI300 GPUs



JVD-AICLUSTERDC-AIMLINF-01-01

## Executive Summary

Modern AI inference environments require predictable latency, scalable throughput, and efficient resource utilization to support high query volumes while maintaining a consistent user experience. As inference deployments transition from experimentation to production, the frontend network becomes increasingly important in enabling reliable communication between inference clients, load balancing services, and GPU-accelerated compute infrastructure.

The **AI Inference Network Design with HPE Juniper QFX switches, HPE Juniper Apstra Data Center Director, and AMD Instinct™ MI300X GPUs** demonstrates how a standards-based Ethernet frontend fabric can efficiently support AI inference workloads while maintaining predictable performance characteristics.

The architecture is designed to provide low latency, scalable bandwidth, and predictable request handling characteristics for production inference environments.

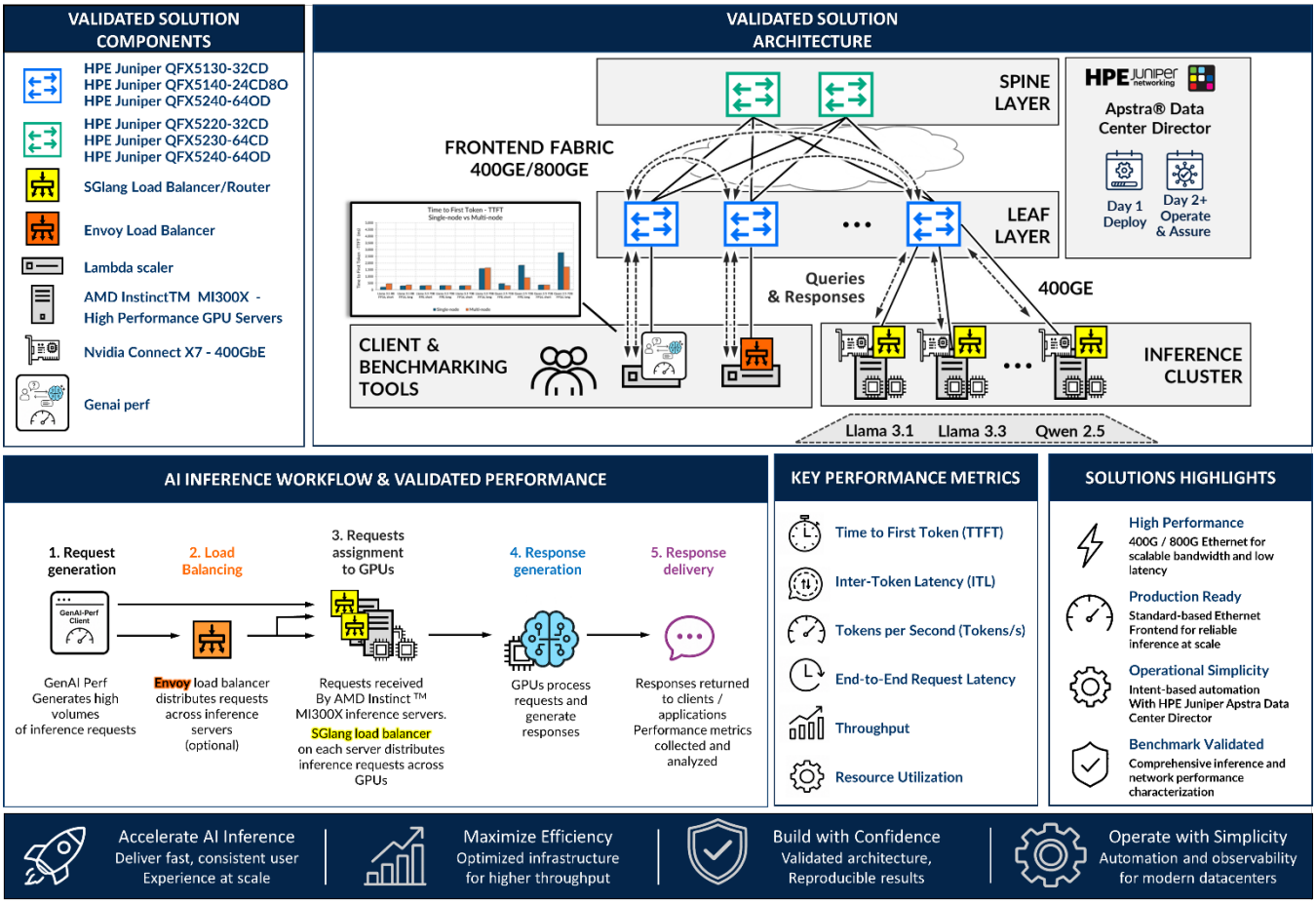
The solution document provides a benchmark-focused reference architecture for high-performance AI inference environments and demonstrates how modern Ethernet-based frontend networks can support production inference deployments.

Through inference performance benchmarking and frontend network characterization, this solution provides practical guidance on infrastructure design, software frameworks, operational considerations, and validated deployment approaches for AI inference workloads. The solution introduces the **HPE Juniper QFX5140-24CD8O** as a frontend fabric leaf platform optimized for production inference deployments.

## Solution Overview

This solution characterizes AI inference performance using **AMD Instinct™ MI300X GPU systems**, modern inference frameworks, and industry-standard benchmarking tools including **NVIDIA GenAI-Perf**.

Figure 1. AI Inference Network Design with Juniper QFX switches and AMD Instinct™ MI300X GPUs solution summary




To characterize performance across multiple inference scenarios, the solution validates a range of commonly deployed Large Language Models (LLMs) with varying model sizes and inference characteristics, including:

- **Llama 3.1 8B** — representing smaller models optimized for lower latency and higher request concurrency.
- **Llama 3.3 70B** — representing larger-scale models with increased compute and memory requirements.
- **Qwen 2.5 72B** — representing alternative large-model architectures used for advanced conversational and reasoning workloads.

By validating multiple models across different parameter sizes, the solution demonstrates frontend fabric behavior under a range of inference conditions, helping customers understand the relationship between model scales, request concurrency, GPU utilization, and network performance.

The validated frontend fabric architecture follows a scalable 3-stage Clos Ethernet IP Fabric design with the following devices in the leaf and spine roles:

| VALIDATED LEAF NODES | VALIDATED SPINE NODES |
|---|---|

| | |
|---|---|
| HPE Juniper QFX5130-32CD | HPE Juniper QFX5220-32CD |
| HPE Juniper QFX5140-24CD8O | HPE Juniper QFX5230-64CD |
| HPE Juniper QFX5240-64OD | HPE Juniper QFX5240-64OD |

Connectivity between the leaf and spine nodes utilizes **400GbE** and **800GbE Ethernet links**, enabling high-bandwidth communication between inference clients, load balancing services, and GPU inference systems.

**AMD Instinct™ MI300X GPU servers connect to the fabric using 400GbE** links with Connect X7 NICs.

Benchmark testing and analysis evaluates the following Inference-serving metrics:

| Metric Category | Key Metrics | Purpose |
|---|---|---|
| Responsiveness | Time to First Token (TTFT) | Measures how quickly the system begins returning a response after a request is submitted. |
| Throughput | Output Tokens per Second (TPS) | Measures how efficiently the system generates output tokens. |

The benchmarking methodology uses **NVIDIA GenAI-Perf** to generate a high volume of inference requests towards **AMD Instinct™ MI300X GPU inference servers**.

While **NVIDIA GenAI-Perf** originates from the NVIDIA ecosystem, it is used in this solution as a vendor-neutral inference benchmarking framework and was selected due to its maturity, feature completeness, and successful operational integration within the validated Juniper lab environment.

Testing included sending inference requests directed to individual inference server, where the SGLand router distributes the requests to the GPUs on that system, or through an **Envoy load balancer** that distributes requests across multiple inference servers.

**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