CASE STUDY

DRIVENETS

Bentaus

SUMMARY

Bentaus is an energy-first NeoCloud that develops, deploys, owns and operates AI infrastructure, hardware and software, and manages compute power in real time through its Ziani Systems platform.

CHALLENGE

- Maximize tokens per watt on a 2,560-GPU AMD Instinct cluster
- Keep performance as the cluster scales
- A validated design, not tuned from scratch
- Open, vendor-agnostic Ethernet

SOLUTION

DriveNets AI Fabric for the back-end and converged front-end and storage networks, on DriveNets 2500S switches and the DriveNets-AMD reference architecture, delivered turnkey by DIS.

BENEFITS

- 30% less energy per token with minimal impact to cluster performance, as measured by Bentaus
- Higher GPU utilization, more tokens per second, less power per million tokens
- Production-grade blueprint from day one
- One switch and one OS across every fabric

NEOCLOUD - CUSTOMER STORY

Optimal AI performance per watt for AMD Instinct clusters and DriveNets AI Fabric

Bentaus built its AMD Instinct™ MI325X NeoCloud cluster on DriveNets AI Fabric and cut energy per token by 30%.

AI clusters have become some of the densest loads on the electrical grid, and the grid is pushing back. For NeoClouds, power availability, not GPU supply, is becoming the main limit on how fast new capacity comes online.

Bentaus, an energy-first NeoCloud, was built for exactly this constraint, curtailing power in real time and providing power back to the grid by dynamically reducing GPU and CPU power consumption while keeping AI workloads running continuously. With power this tightly managed, every watt the cluster draws has to turn into tokens, which makes GPU utilization a first-order business metric and puts the network that keeps the GPUs busy at the center of the energy equation.

This case study examines Bentaus' deployment of DriveNets AI Fabric across the back-end, front-end and storage networks of its new 2,560-GPU AMD Instinct™ MI325X cluster in Boca Raton, Florida. Bentaus measured 30% less energy per token on this cluster, running AMD Instinct MI325X GPUs, DriveNets AI Fabric and Bentaus Ziani software, with zero impact to cluster performance.

Bentaus' Challenge

Bentaus' Ziani Systems platform, GPU-FLX™ and CPU-FLX™, manages compute power in real time, delivering grid-responsive, vendor-agnostic AI infrastructure built on open standards that maximizes tokens per watt and lowers cost per token.

For a NeoCloud that measures itself in tokens per watt, network efficiency is energy efficiency. A GPU waiting on the network still draws power while producing nothing. For its new 2,560-GPU AMD Instinct MI325X cluster, Bentaus needed a network that would:

- Drive GPUs to high utilization and preserve performance as the cluster scales
- Be validated end to end for AMD Instinct environments, not tuned from scratch
- Run on standard Ethernet, keeping the stack open and vendor-agnostic
- Carry front-end and storage traffic on the same switch platform

Why DriveNets Stood Out

DriveNets AI Fabric maximizes the performance of AMD Instinct-based clusters through full-stack networking optimization: from the switch, via the NIC, to RCCL collective communications and kernel-level optimization. The design deployed at Bentaus is based on a reference architecture jointly engineered by DriveNets and AMD, which gave Bentaus a production-grade blueprint from day one.

The reference architecture's published benchmarks on AMD Instinct MI350 series clusters show the DriveNets fabric delivering up to 5% higher multi-node throughput and 12-16% faster time to first token (TTFT) at scale than publicly available industry results, all on standard Ethernet compatible with any GPU, NIC and optics.

Bentaus' energy edge: Ziani on the DriveNets-connected cluster

These results were measured by Bentaus and reflect Ziani's compute power management.

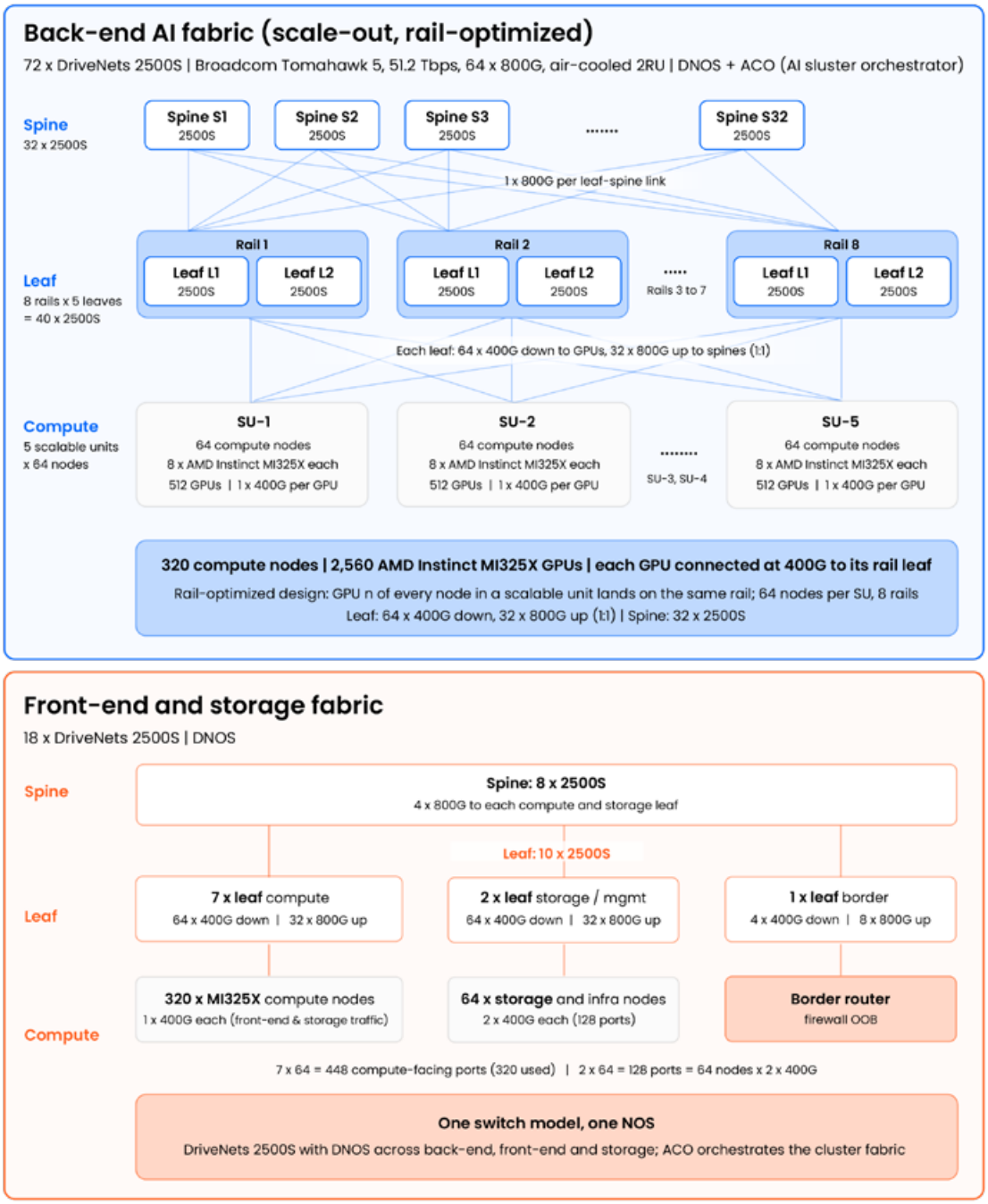
- Validated in Bentaus' AI cluster environment, the AMD Instinct MI325X cluster orchestrated under Ziani GPU-FLX™ kept 97% of its output on 30% less GPU power.
- In addition, the cluster achieved a 23% increase in token throughput (tokens per second) on AMD Instinct MI325X nodes orchestrated by Ziani Tune™.

Solution

Bentaus selected DriveNets AI Fabric for the back-end and converged front-end and storage networks of its new cluster, and deployed them with DriveNets Infrastructure Services (DIS) as a turnkey project: design, rack and stack, cabling, bring-up, integration & full-cluster benchmarking.

The back-end fabric is a rail-optimized, two-tier design: each GPU connects at 400G to its rail leaf, and each leaf connects to the spines at 800G with a 1:1 ratio between downlink and uplink capacity. On top of the fabric, when the grid is under stress Ziani dials GPU and CPU power down in real time, and back up when the grid recovers, while AI workloads keep running continuously. The division of labor is simple: DriveNets makes sure the GPUs are never waiting on the network, and Ziani decides how much power the cluster draws and when.

Bentaus Cluster Architecture



"Everyone is racing to add GPUs. We built Bentaus to add compute while mitigating risks to the grid. Ziani flexes our clusters acknowledging grid conditions in real time, and DriveNets' fabric makes sure the GPUs we run are never waiting on the network. That is how you scale AI capacity now: more compute, delivered more efficiently, on power the grid can sustain"

Bob Davidoff, CEO & Founder, Bentaus

DEPLOYMENT DETAILS:

- **Storage**
  8 Weka storage nodes
- **Compute**
  2,560 AMD Instinct MI325X GPUs in 320 compute nodes (8 GPUs per node), in five scalable units of 64 nodes
- **Back-end fabric**
  72 DriveNets 2500S switches (32 spines, 40 leaves), rail-optimized 8-rail Clos; 400G per GPU, 800G leaf-to-spine, 1:1
- **Front-end and storage fabric**
  18 DriveNets 2500S switches (8 spines, 10 leaves: 7 compute, 2 storage, 1 border), serving 320 compute nodes at 400G & 64 storage and infrastructure nodes at 2 x 400G
- **Hardware**
  DriveNets 2500S, air-cooled 2RU, Broadcom Tomahawk 5, 51.2 Tbps, 64 x 800G OSFP ports
- **Software**
  DriveNets Network Operating System (DNOS); DriveNets AI Cluster Orchestrator (ACO) for automated provisioning and benchmarking; Bentaus Ziani Systems (GPU-FLX, Ziani Tune)
- **Services**
  DriveNets Infrastructure Services (DIS), Turnkey from design to performance acceptance
- **Site**
  Bentaus data center, Boca Raton, Florida



Results

With DriveNets AI Fabric running the back-end, front-end and storage networks of its AMD Instinct MI325X cluster, Bentaus achieved:

- **30% less energy per token**
  measured by Bentaus on the AMD Instinct MI325X cluster running DriveNets AI Fabric and Ziani software, with minimal impact to cluster performance. Same tokens, 30% less energy.
- **High GPU utilization**
  GPUs that are not waiting on the network, driving more tokens per second and less power per million tokens.
- **One platform for every fabric**
  the same 2500S switch and DNOS across back-end, front-end and storage, with ACO automating provisioning & benchmarking.
- **Performance preserved at scale**
  the efficiency the fabric creates holds across the full 2,560-GPU cluster.
- **Fast time to production**
  a validated DriveNets-AMD design & turnkey DIS delivery instead of component-level tuning.

Summary

The first AI race was to produce more tokens. The next race is to produce them on less power. For an energy-first NeoCloud, every watt has to turn into tokens. DriveNets AI Fabric gave Bentaus the network side of that equation: a full-stack-optimized Ethernet fabric on a validated AMD Instinct reference architecture that keeps GPUs productive and the stack open. On that foundation, Ziani dials cluster power down and up with the grid while workloads keep running, delivering AI capacity that is high-performing, efficient per watt and sustainable on the grid: more AI from every megawatt.

DRIVENETS

DriveNets builds high-performance networking solutions for large-scale AI infrastructure and service-provider networks. DriveNets' Network Cloud powers production networks at global operators such as AT&T and Comcast, while its AI Fabric solutions optimize the performance and time to first token (TTFT) in multi-vendor AI environments at foundation model labs, NeoClouds, and enterprises. DriveNets combines high-performance networking, end-to-end automation, and full-stack AI cluster optimization to remove network bottlenecks, maximize GPU utilization, accelerate AI cluster deployments, and improve token economics. More info at **www.drivenets.com**