SAKURAONE: Open Ethernet-Based, AI-Ready HPC with 800GbE and SONiC

Fumikazu KONISHI[1)], Yuuki Tsubouchi[1)], Hirofumi Tsuruta[1)], Takashi Inoue[2)], Kiyohiro Kurosawa[2)]

1) Research Center, SAKURA internet inc. 2) SAKURA internet inc. , Japan

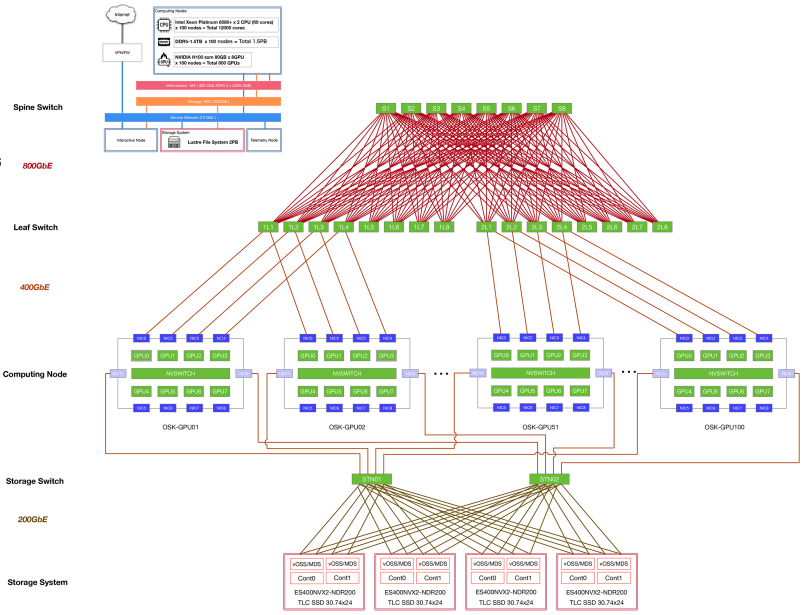
Background/Objectives

Japan's private sector requires production-grade AI computing on par with Big Tech, but with vendor-neutral infrastructure.

A key motivation is the use of open, disaggregated networking. This design improves vendor agility and supply chain resilience by decoupling NOS from the underlying ASICs, accelerates innovation through community-driven modular components, and enables lossless AI fabrics through RoCEv2 QoS (PFC, ECN).

Scalable overlays (EVPN/VXLAN) further support multi-tenant AI workflows, while open management frameworks provide automation and observability essential for transparent large-scale operations.

SAKURAONE, SAKURA internet's managed GPU supercomputer, was developed to:
(i) demonstrate the scalability of an all-Ethernet fabric with SONiC for AI/HPC,
(ii) provide reproducible AI software stacks for enterprises, and
(iii) evaluate performance using community benchmarks.
At ISC~2025, SAKURAONE ranked 49th on the TOP500 (HPL) and, to our knowledge, is the only Top-100 system built entirely on an open 800,GbE + SONiC networking stack.



System:

100 nodes; each: 2 Intel Xeon Platinum~8580+, 1.5TB DDR5, 8 times NVIDIA H100 SXM5 (80GB).

Network:

Rail-optimized leaf-spine; host links 400GbE; leaf-spine 800GbE; RoCEv2; SONiC on Broadcom Tomahawk5; NIC GPU affinity preserves NUMA locality for collectives.

Storage:

2PB all-flash Lustre on DDN ES400NVX2-NDR200; per-node dual 200GbE for storage.

Software:

Rocky Linux9, Slurm, CUDA~12.x, NCCL~2.2x, Apptainer/Singularity & Pyxis, curated conda ML stacks.

Benchmarks:

HPL, HPCG, HPL-MxP (HPL-AI class), IO500, MLPerf Training.

Table 1: SAKURAONE Computing Nodes

Name	Description
Chassis	Supermicro GPU SuperServer SYS-821GE-TNHR
CPU	Intel Xeon Platinum 8580+ x 2 CPUs
Core (per CPU)	120 (60)
GPU	NVIDIA H100 SXM 80GB x 8 GPUs
Memory (RAM)	DDR5-1.5TB (96GB x 16)
System storage (SAS)	372GB x 2
Data storage (NVMe)	7.68TB x 4
Local network interface	4GbE (2Gbps x 2)
Interconnect network interface	NVIDIA Mellanox MCX75310AAS-NEAT ConnectX®-7 400GbE/NDR x 8
Storage network interface	NVIDIA Mellanox MCX75310AAS-NEAT ConnectX®-7 400GbE/NDR x 2

Table 2: Classification of NIC Usage and GPU Connectivity Characteristics

NIC	Device Name	Primary Usage	GPU Connectivity Type
NIC0	mlx5_0	High-speed inter-node communication	NODE (via GPU0 PCIe domain)
NIC1	mlx5_1	High-speed inter-node communication	NODE (via GPU1 PCIe domain)
NIC2	mlx5_2	High-speed inter-node communication	NODE (via GPU2 PCIe domain)
NIC3	mlx5_3	High-speed inter-node communication	NODE (via GPU3 PCIe domain)
NIC4	mlx5_4	High-speed inter-node communication	NODE (via GPU4 PCIe domain)
NIC5	mlx5_5	High-speed inter-node communication	NODE (via GPU5 PCIe domain)
NIC6	mlx5_6	High-speed inter-node communication	NODE (via GPU6 PCIe domain)
NIC7	mlx5_7	High-speed inter-node communication	NODE (via GPU7 PCIe domain)
NIC8	mlx5_8	Storage network (dedicated I/O path)	PXB
NIC10	mlx5_bond_0	Storage network (bonded for redundancy)	PXB (logical, multi-bridge path)
NIC9	mlx5_11	Management network (e.g., SSH)	SYS

This classification is based on `nvidia-smi topo -mp` output, analyzing PCIe connectivity and NUMA affinity between GPUs and NICs. NIC0–NIC7 are directly mapped to GPUs 0–7 using NODE-level PCIe paths for optimized inter-node communication. NIC8 and NIC10 are likely dedicated to storage traffic and are accessed via longer PXB paths or logical bonding. NIC9 is connected to the system via a SYS-level path that crosses NUMA domains.

Table 3: SAKURAONE Interconnect Network

Name	Description
Network technology	Gigabit Ethernet (GbE)
Ethernet switch speed grdade	800 GbE 2DR4 2 x 400G GbE
Protocol	RoCEv2 (RDMA over Converged Ethernet)
Network topology	Rail-Optimized Topology
Switch Chassis	Edge-core networks AIS800-64O
Switch Capability	51.2Tbps fullduplex
Software Stack	SONiC (Software for Open Networking in the Cloud)
Switch Chip	Tomahawk 5 chipset powered by Broadcom
Leaf Switch	16 chassis[1]
Spine Switch	8 chassis

Table 4: SAKURAONE Storage System

Name	Description	Count
Chassis	DDN ES400NVX2	4
Controller	Active Dual Controller	2
CPU	Ice Lake CPUs	2
NVMe	24 Drive (PCI Gen4)	24
Drive	TLC SSD 30.72TB	
Interface	200 GbE	4

Table 6: HPL Benchmark Summary

Item	Value
Matrix size (N)	2,706,432
Block size (NB)	1024
Process grid (P×Q)	16 × 49
Total processes	784
Total GPUs	784
HPL version	HPL-NVIDIA 25.4.0
Execution time (sec)	389.23
FLOPS	33.95 PFLOPS
FLOPS per GPU	43.31 TFLOPS
Max GEMM performance (single GPU)	55.34 TFLOPS
GPU SM count	132
GPU peak clock frequency	1980 MHz

Table 7: HPCG Benchmark Summary

Item	Value
Benchmark version	HPCG 3.1
Total distributed processes	784
Threads per process	16
Global problem dimensions (nx × ny × nz)	4096 × 3584 × 3808
Number of equations	55.9 billion
Number of nonzero terms	1.51 trillion
Total memory used (GB)	39,961.4
Memory for linear system and CG (GB)	35,169
Peak memory bandwidth (observed)	3.316 TB/s
Total GFLOP/s (raw)	437,361
GFLOP/s (with convergence overhead)	404,964
Final validated HPCG GFLOP/s result	396,295

Table 8: HPL-MxP Benchmark Summary

Item	Value
Benchmark version	HPL-MxP-NVIDIA 25.4.0
Matrix size N	2,989,056
Block size NB	4096
Process grid (P × Q)	24 × 32
Total processes	768
Peak clock frequency	1980 MHz
GPU SM version	SM 90
GPU SM count	132
Observed Rmax	3.3986e+08 GFLOPS
Rmax per GPU	442520.81 GFLOPS
LU-only	5.3919e+08 GFLOPS
LU-only per GPU	702,074.99 GFLOPS
Precision mode	Sloppy FP8 (sloppy-type = 1)
Validation result	PASSED (5.01e-05 < 1.6e+01)

Table 9: Comparison of IO500 Results: 10 Nodes vs 96 Nodes

Benchmark	10 Nodes	96 Nodes
ior-easy-write (GiB/s)	262.91 (340.96 s)	198.80 (355.68 s)
mdtest-easy-write (kIOPS)	204.44 (347.00 s)	256.64 (363.81 s)
ior-hard-write (GiB/s)	15.84 (354.60 s)	24.61 (491.75 s)
mdtest-hard-write (kIOPS)	120.84 (340.05 s)	151.59 (332.84 s)
find (kIOPS)	1976.05 (56.39 s)	2637.17 (54.01 s)
ior-easy-read (GiB/s)	365.71 (245.15 s)	305.86 (231.13 s)
mdtest-easy-stat (kIOPS)	358.75 (197.40 s)	463.13 (200.14 s)
ior-hard-read (GiB/s)	205.64 (31.23 s)	255.31 (48.82 s)
mdtest-hard-stat (kIOPS)	262.43 (157.19 s)	408.13 (124.54 s)
mdtest-easy-delete (kIOPS)	168.19 (422.12 s)	198.91 (468.91 s)
mdtest-hard-read (kIOPS)	205.39 (200.53 s)	310.87 (162.97 s)
mdtest-hard-delete (kIOPS)	92.29 (445.91 s)	111.28 (453.80 s)
Bandwidth Score (GiB/s)	**133.03**	**139.80**
IOPS Score (kIOPS)	**248.74**	**327.84**
Total IO500 Score	**181.91**	**214.09**

Table 10: MLPerf Training (GPT-3) Benchmark Summary

Item	32 Nodes	96 Nodes
Total GPUs	256	768
Data Parallelism	4	6
Tensor Parallelism	4	8
Pipeline Parallelism	16	16
Context Parallelism	1	1
Sequence Parallelism	True	True
Global batch size	1024	2304
Micro batch size	2	6
Time-to-train (min)	105.31 (unverified)	41.86 (unverified)
Model FLOPs Utilization (%)	38.3	35.9
Training Throughput (tokens/s/GPU)	707.62	714.23
Computational Throughput (TFLOPS/GPU)	757.13	710.73

Table 11: MLPerf Training (Llama2 70B LoRA) Benchmark Summary

Item	1 Node	8 Nodes	64 Nodes	96 Nodes
Total GPUs	8	64	512	768
Data Parallelism	2	8	64	96
Tensor Parallelism	4	4	4	4
Pipeline Parallelism	1	1	1	1
Context Parallelism	1	2	2	2
Sequence Parallelism	True	True	True	True
Global batch size	8	8	64	96
Micro batch size	1	1	1	1
Time-to-train (min)	28.44 (unverified)	4.79 (unverified)	1.94 (unverified)	1.26 (unverified)

Results

The system achieved a sustained HPL (FP64) performance of 33.95 PFLOP/s (Rmax) on 100 nodes equipped with 800 GPUs. For HPCG, it delivered 396.295 TFLOP/s.
On HPL-MxP (FP8), the system reached 339.86 PFLOP/s, corresponding to approximately 442.5 TFLOP/s per GPU, and 539.19 PFLOP/s in the LU-only phase.
In IO500 (production), the system scored 181.91 on 10 nodes, showing strong metadata scaling. Despite higher latency than InfiniBand/Slingshot, the rail-optimized 800 GbE + RoCEv2 + SONiC fabric sustained competitive AI/HPC throughput, confirming open-stack scalability at Top500 scale.
For MLPerf Training, the system achieved scores of 105.310 unverified (32 nodes) and 41.862 unverified (96 nodes) on GPT-3, while on Llama 2 70B-LoRA it achieved 28.444 unverified (1 node), 4.789 unverified (8 nodes), 1.941 unverified (64 nodes), and 1.256 unverified (96 nodes), respectively.
Result not verified by MLCommons Association.

Conclusions

SAKURAONE shows that an open, SONiC-based 800 GbE fabric can deliver Top500-class FP64 and competitive AI performance at national scale while preserving vendor neutrality and cost efficiency.
Future work includes
(i) NCCL/MPI tuning with rail-aware routing and NUMA pinning, and
(ii) topology-aware placement with adaptive congestion control.

Reference

Fumikazu Konishi. Sakuraone: Empowering transparent and open ai platforms through private-sector hpc investment in japan, 2025. URL https://arxiv.org/abs/2507.02124.