MLSys 2026 — Industry Track Benchmarks (Oral)

Paper(arXiv)

SAKURAONE:
An Open Ethernet-based AI HPC System And Its Observed Workload Dynamics in a Single-Tenant LLM Development Environment

Authors — Fumikazu Konishi, **Yuuki Tsubouchi**, Hirofumi Tsuruta
SAKURA internet, Inc.

SAKURA internet

AI infrastructure needs more than peak compute

AI infrastructure needs more than peak compute

What peak FLOPS misses

SUSTAINED CAPACITY
Month-scale 70B-class LLM development

LOSSLESS COLLECTIVES
Predictable, low-congestion GPU-to-GPU paths.

OPEN OPERATIONS
Vendor flexibility, and lifecycle control.

AI infrastructure needs more than peak compute

What peak FLOPS misses

SUSTAINED CAPACITY
Month-scale 70B-class LLM development

LOSSLESS COLLECTIVES
Predictable, low-congestion GPU-to-GPU paths.

OPEN OPERATIONS
Vendor flexibility, and lifecycle control.

SAKURAONE response

800 H100 GPUs
Headroom for repeated LLM development runs.

RoCEv2 Ethernet
Open ethernet-based RDMA, rail-optimized and separated storage I/O path.

SONiC / SAI
An open NOS–based fabric and disaggregation of the NOS and the switch ASIC.

AI infrastructure needs more than peak compute

What peak FLOPS misses

SUSTAINED CAPACITY
Month-scale 70B-class LLM development

LOSSLESS COLLECTIVES
Predictable, low-congestion GPU-to-GPU paths.

OPEN OPERATIONS
Vendor flexibility, and lifecycle control.

SAKURAONE response

800 H100 GPUs
Headroom for repeated LLM development runs.

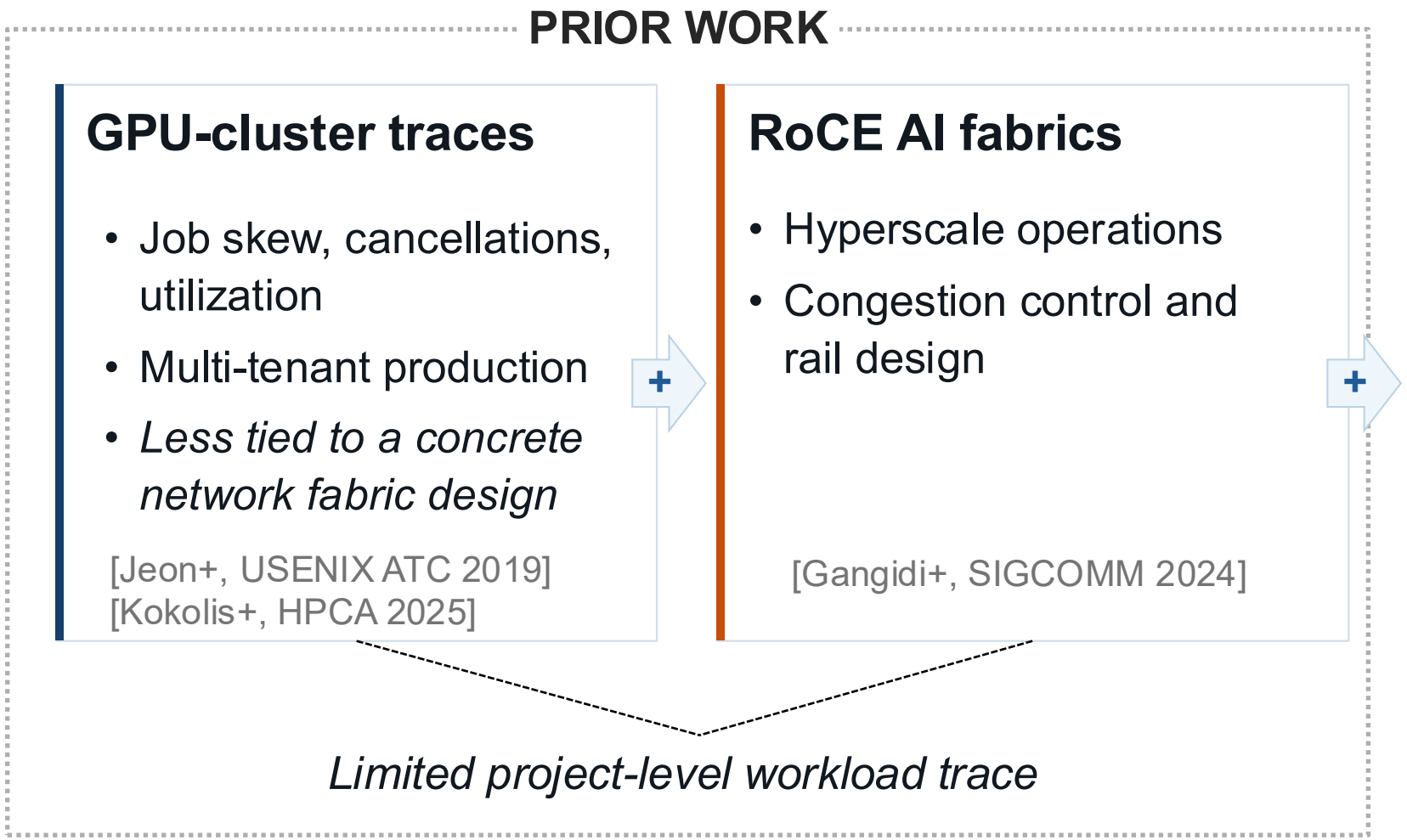
RoCEv2 Ethernet
Open ethernet-based RDMA, rail-optimized and separated storage I/O path.

SONiC / SAI
An open NOS–based fabric and disaggregation of the NOS and the switch ASIC.

Build the production platform first;
validate it with benchmarks and a single-project workload trace.

Where this case study fits in prior work.

Where this case study fits in prior work.



Where this case study fits in prior work.

PRIOR WORK

GPU-cluster traces

- Job skew, cancellations, utilization
- Multi-tenant production
- *Less tied to a concrete network fabric design*

[Jeon+, USENIX ATC 2019]
[Kokolis+, HPCA 2025]

+

RoCE AI fabrics

- Hyperscale operations
- Congestion control and rail design

[Gangidi+, SIGCOMM 2024]

+

Our Paper

- Mid-scale open-Ethernet system (SONiC / RoCEv2)
- Benchmarks plus single-project trace
- Lower cross-tenant confounding

Limited project-level workload trace

Contributions

Contributions

Finding 1

SONiC / RoCEv2 is competitive across HPC, AI, and storage benchmarks.

- TOP-500 in ISC2025

  Ranked 49th in HPL

- MLPerf Training v4.1

  Within 2–17% of NVIDIA Eos
  (DGX, Quantum-2 NDR InfiniBand)

Contributions

Finding 1

SONiC / RoCEv2 is competitive across HPC, AI, and storage benchmarks.

- TOP-500 in ISC2025

  Ranked 49th in HPL

- MLPerf Training v4.1

  Within 2–17% of NVIDIA Eos (DGX, Quantum-2 NDR InfiniBand)

Finding 2

A single-project trace still shows skew that mirrors multi-tenant clusters.

- Job count vs. GPU-occupied time invert
- CANCELLED jobs dominate occupancy
- Workload composition shifts across months

Sizing SAKURAONE from BLOOM-176B

Sizing SAKURAONE from BLOOM-176B

Public reference: BLOOM-176B

Jean Zay supercomputer, public training report [Le Scao+, arXiv 2022]

GPU COUNT	DURATION	COMPUTE HOURS
384 × A100	3.5 months	1.08M hours

A concrete benchmark for LLM training time and compute budget.

scale anchor

Sizing SAKURAONE from BLOOM-176B

Public reference: BLOOM-176B

Jean Zay supercomputer, public training report [Le Scao+, arXiv 2022]

GPU COUNT	DURATION	COMPUTE HOURS
384 × A100	3.5 months	1.08M hours

A concrete benchmark for LLM training time and compute budget.

scale anchor

Hopper 2–3× per-GPU throughput + operational headroom

SAKURAONE sizing target

Production LLM development target, not a one-shot demo

MODEL	TOKENS	DURATION
70B class	≈ 300B	≈ 4 months

Enough capacity for repeated, overlapping training cycles.

Sizing SAKURAONE from BLOOM-176B

Public reference: BLOOM-176B

Jean Zay supercomputer, public training report [Le Scao+, arXiv 2022]

GPU COUNT	DURATION	COMPUTE HOURS
384 × A100	3.5 months	1.08M hours

A concrete benchmark for LLM training time and compute budget.

scale anchor

Hopper 2–3× per-GPU throughput + operational headroom

SAKURAONE sizing target

Production LLM development target, not a one-shot demo

MODEL	TOKENS	DURATION
70B class	≈ 300B	≈ 4 months

Enough capacity for repeated, overlapping training cycles.

Resulting scale: 100 nodes / 800 H100 GPUs

SAKURAONE System Overview

SAKURAONE System Overview

Compute plane

100 nodes · 800× NVIDIA H100 SXM

8 GPUs/node · 2× CPU · 2 TB DRAM

100 nodes

SAKURAONE System Overview

Compute plane

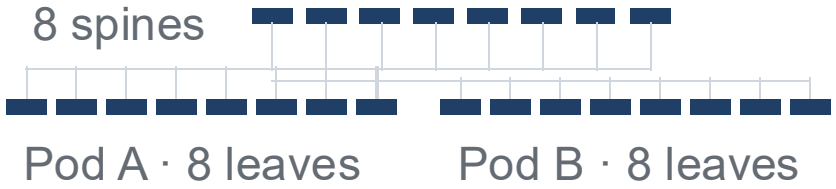
100 nodes · 800× NVIDIA H100 SXM

8 GPUs/node · 2× CPU · 2 TB DRAM

GPU interconnect fabric

8× 400 GbE NICs per node

NVIDIA ConnectX-7



SONiC · Broadcom Tomahawk 5 · RoCEv2

SAKURAONE System Overview

Compute plane

100 nodes · 800× NVIDIA H100 SXM

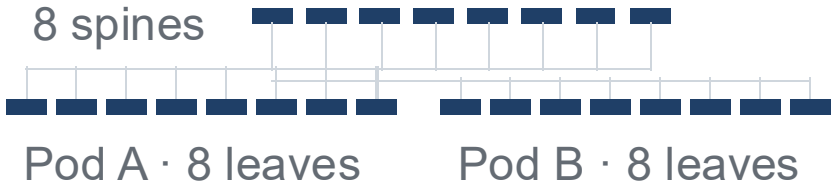
8 GPUs/node · 2× CPU · 2 TB DRAM

100 nodes

GPU interconnect fabric

8× 400 GbE NICs per node

NVIDIA ConnectX-7



SONiC · Broadcom Tomahawk 5 · RoCEv2

Storage plane

2× 200 GbE NICs per node

2 PB all-flash Lustre · 4× DDN appliance

DDN-1 **DDN-2** **DDN-3** **DDN-4**

SAKURAONE System Overview

Compute plane

100 nodes · 800× NVIDIA H100 SXM

8 GPUs/node · 2× CPU · 2 TB DRAM

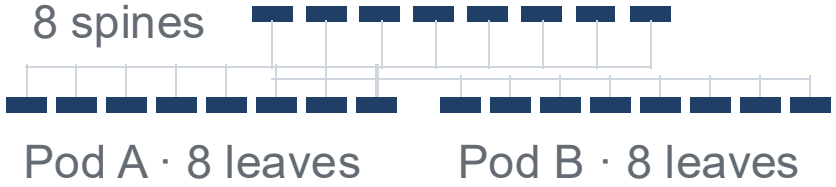
100 nodes

GPU node allocation

GPU interconnect fabric

8× 400 GbE NICs per node

NVIDIA ConnectX-7



SONiC · Broadcom Tomahawk 5 · RoCEv2

Storage plane

2× 200 GbE NICs per node

2 PB all-flash Lustre · 4× DDN appliance

DDN-1 **DDN-2** **DDN-3** **DDN-4**

Job scheduler

Slurm

Queue

Job submit

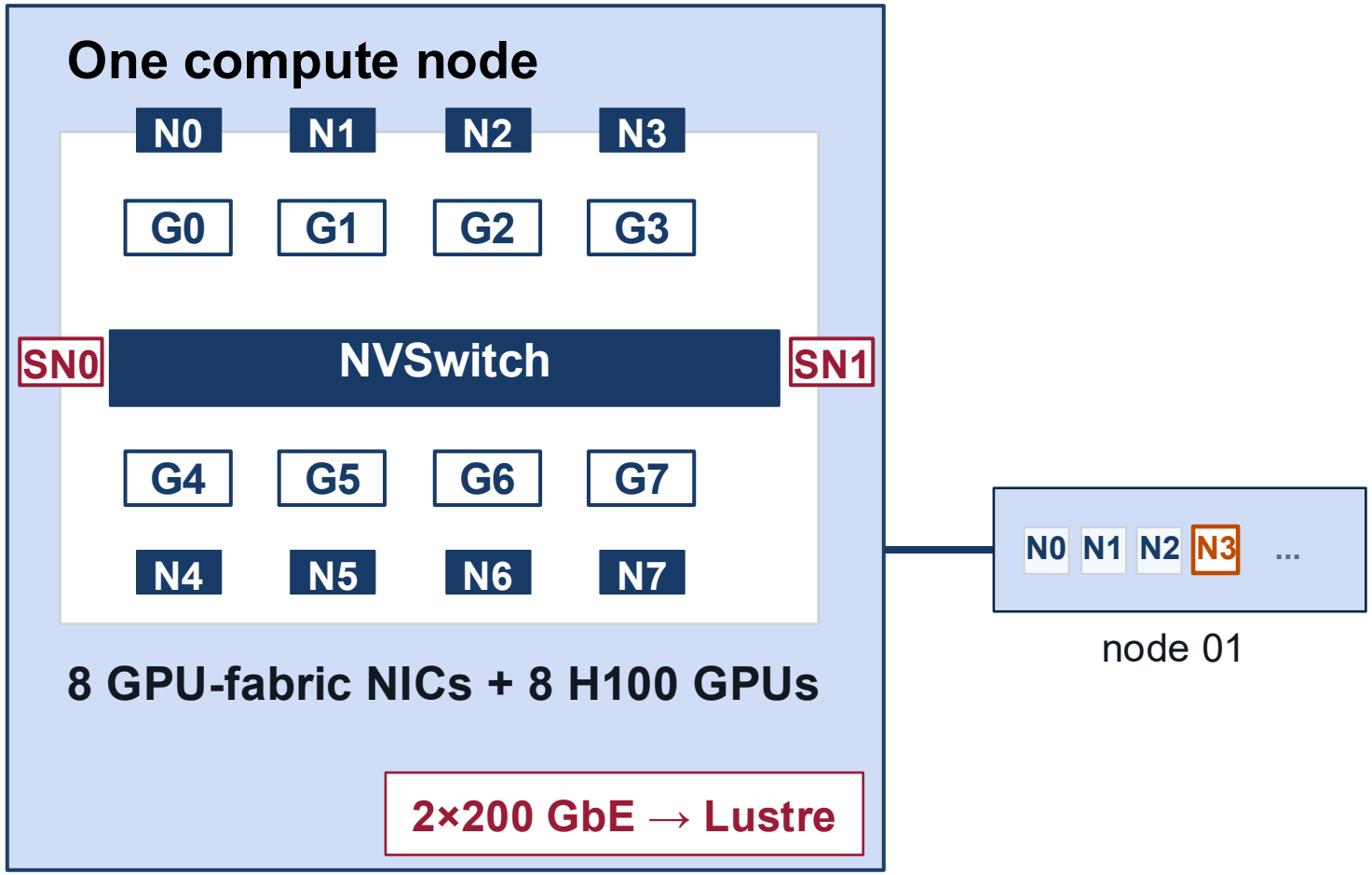
Users

GPU–NIC affinity and rail-optimized topology

Give predictable paths, and reduce contention

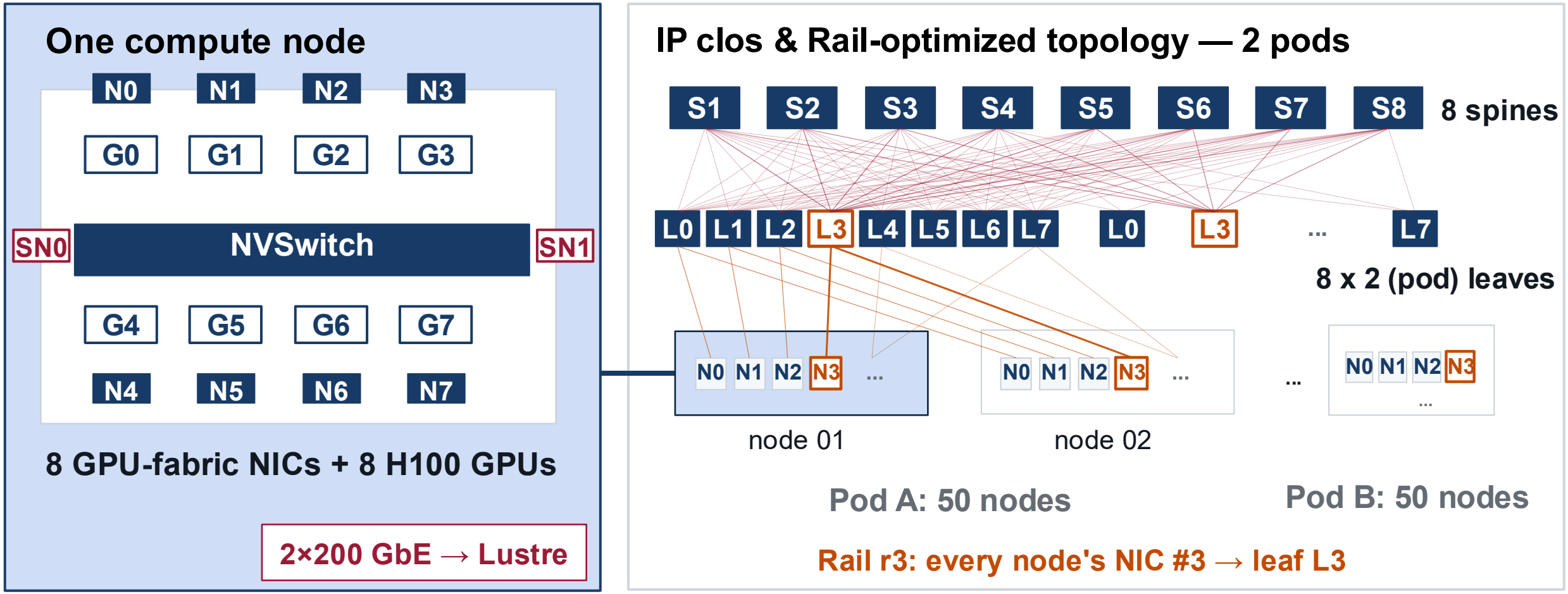
GPU–NIC affinity and rail-optimized topology

Give predictable paths, and reduce contention



GPU–NIC affinity and rail-optimized topology

Give predictable paths, and reduce contention



Contributions

Finding 1

SONiC / RoCEv2 is competitive across HPC, AI, and storage benchmarks.

- TOP-500 in ISC2025

  Ranked 49th in HPL

- MLPerf Training v4.1

  Within 2–17% of NVIDIA Eos (DGX, Quantum-2 InfiniBand)

Finding 2

A single-project trace still shows skew that mirrors multi-tenant clusters.

- Job count vs. GPU-occupied time invert
- CANCELLED jobs dominate occupancy
- Workload composition shifts across months

TOP-500 Benchmarks

Balanced validation across HPC, AI, and storage.

HPL — dense FP64

33.95 PFLOP/s

78.3% per-GPU GEMM

compute-bound throughput

49th *

HPL-MxP — mixed precision

339.86 PFLOP/s

539.19 PFLOP/s LU-only

tensor-core throughput

43rd *

HPCG — sparse / comm

396.295 TFLOP/s

784 processes

memory- and communication-bound

12th *

IO500 — storage I/O

214.09

96 nodes · total score

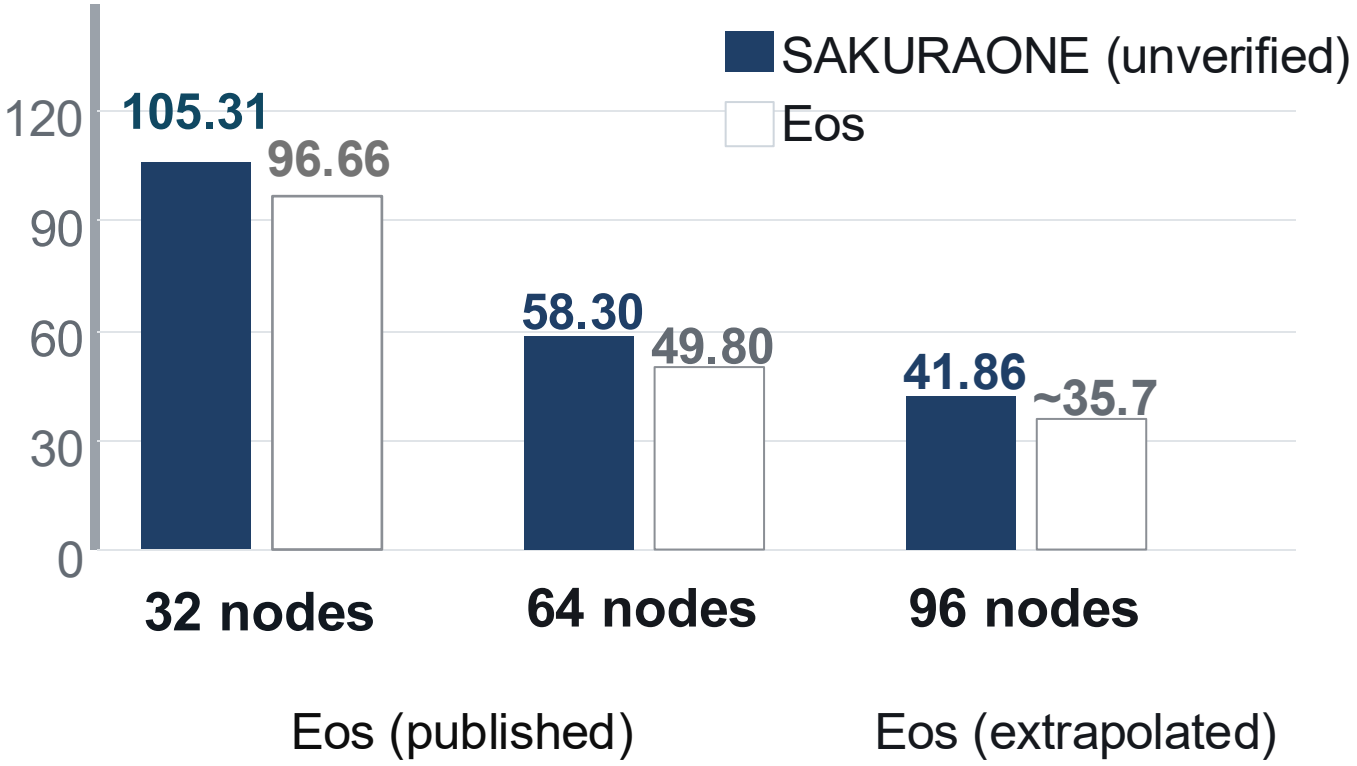
metadata + bandwidth on 2 PB Lustre

9th *

* Results of TOP 500 in ISC2025

MLPerf Training Benchmarks

GPT-3 175B Continues pretraining



vs. NVIDIA Eos

DGX, Quantum-2 InfiniBand, rail-optimized

9 – 17% gap

competitive on open Ethernet

DISCLAIMER
SAKURAONE MLPerf test runs, unverified.

Profiling for MLPerf GPT-3

Profiling for MLPerf GPT-3

Table 9· MLPerf Training (GPT-3) Benchmark

Item	32 N	64 N	96 N
Total GPUs	256	512	768
Data Parallelism	4	8	6
Tensor Parallelism	4	4	8
Pipeline Parallelism	16	16	16
Virtual Pipelines	6	6	6
Global batch size	1024	1536	2304
Micro batch size	2	2	6
Time-to-train (min)*	105.31	58.30	41.86
MFU (%)	38.3	41.2	35.9
Tokens/s/GPU	707.62	758	714.23
TFLOPS/GPU	757.13	815	710.73

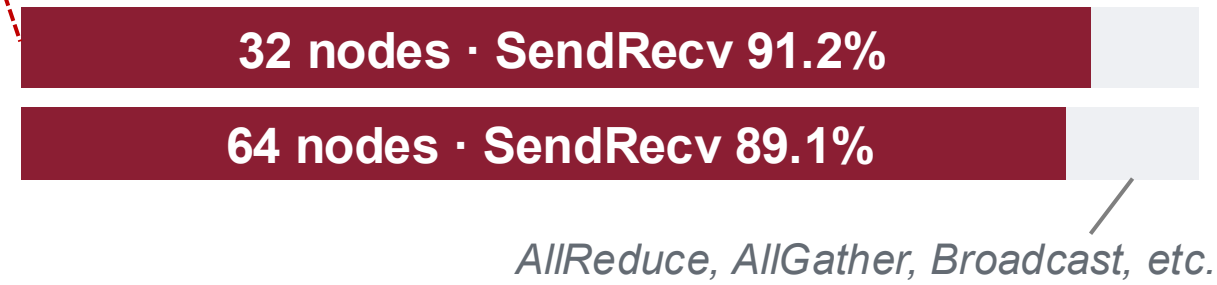
*Unverified. 8 GPUs/node; CP = 1, SP enabled for all configs.

Profiling for MLPerf GPT-3

Table 9· MLPerf Training (GPT-3) Benchmark

Item	32 N	64 N	96 N
Total GPUs	256	512	768
Data Parallelism	4	8	6
Tensor Parallelism	4	4	8
Pipeline Parallelism	16	16	16
Virtual Pipelines	6	6	6
Global batch size	1024	1536	2304
Micro batch size	2	2	6
Time-to-train (min)*	105.31	58.30	41.86
MFU (%)	38.3	41.2	35.9
Tokens/s/GPU	707.62	758	714.23
TFLOPS/GPU	757.13	815	710.73

*Unverified. 8 GPUs/node; CP = 1, SP enabled for all configs.



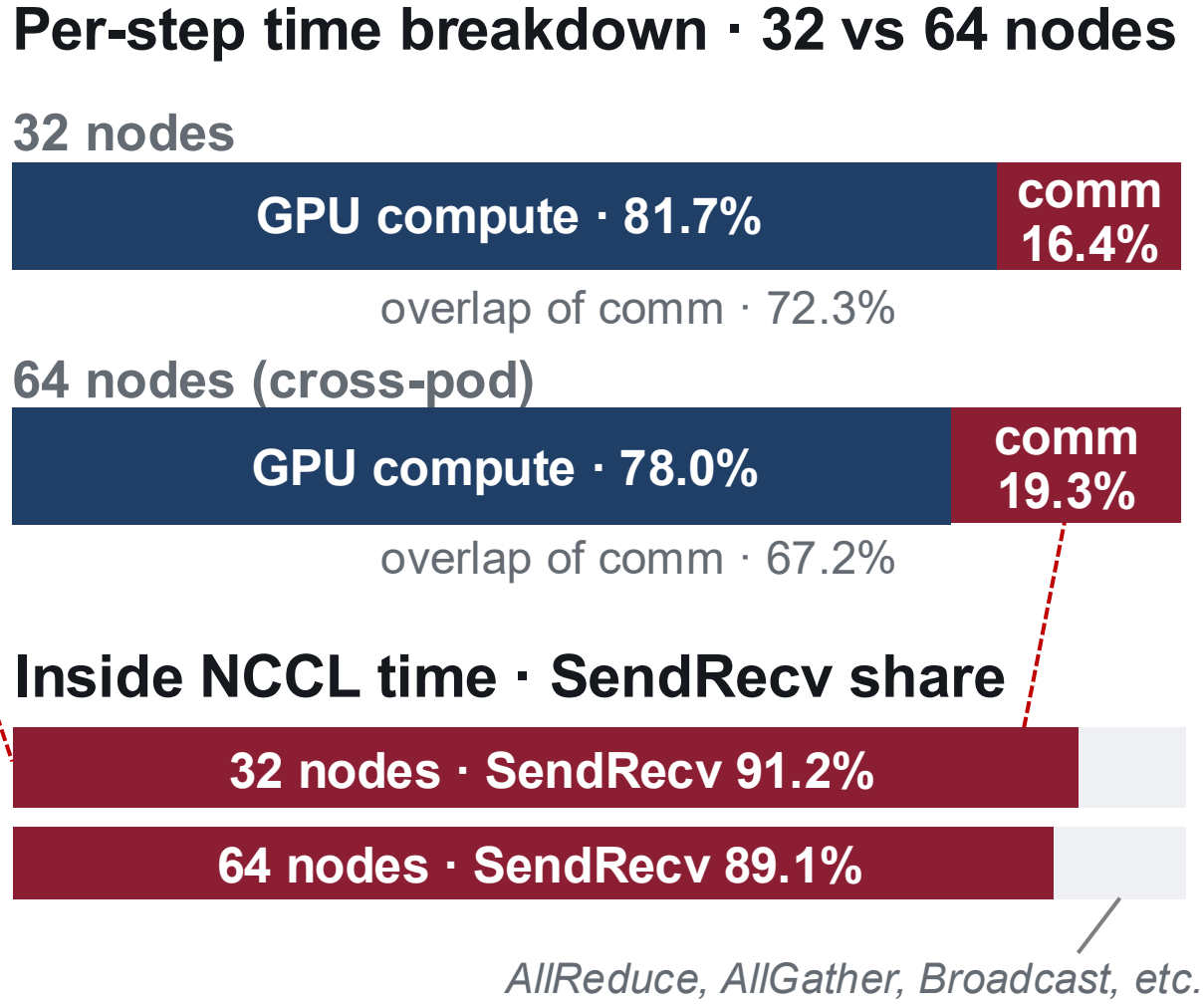
PP=16, VP=6 makes SendRecv dominant; cross-pod placement remains a bounded hypothesis.

Profiling for MLPerf GPT-3

Table 9· MLPerf Training (GPT-3) Benchmark

Item	32 N	64 N	96 N
Total GPUs	256	512	768
Data Parallelism	4	8	6
Tensor Parallelism	4	4	8
Pipeline Parallelism	16	16	16
Virtual Pipelines	6	6	6
Global batch size	1024	1536	2304
Micro batch size	2	2	6
Time-to-train (min)*	105.31	58.30	41.86
MFU (%)	38.3	41.2	35.9
Tokens/s/GPU	707.62	758	714.23
TFLOPS/GPU	757.13	815	710.73

*Unverified. 8 GPUs/node; CP = 1, SP enabled for all configs.



PP=16, VP=6 makes SendRecv dominant; cross-pod placement remains a bounded hypothesis.

Contributions

Finding 1

SONiC / RoCEv2 is competitive across HPC, AI, and storage benchmarks.

- TOP-500 in ISC2025

  Ranked 49th in HPL

- MLPerf Training v4.1

  Within 2–17% of NVIDIA Eos (DGX, Quantum-2 NDR InfiniBand)

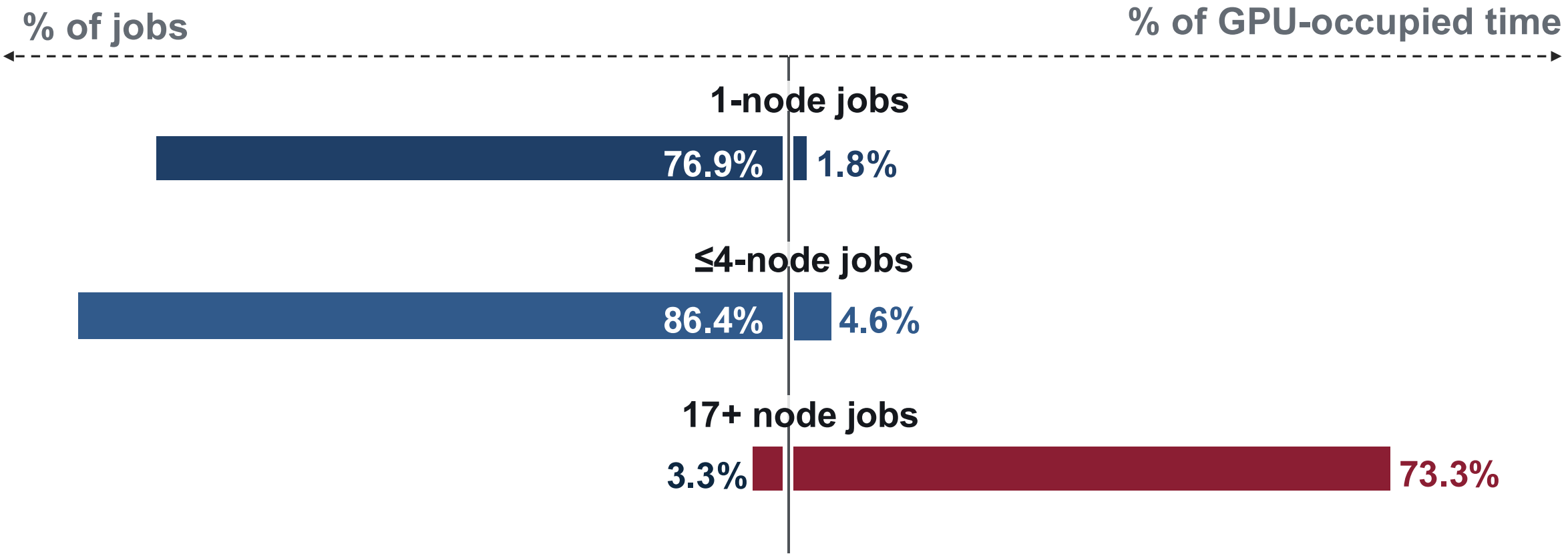
Finding 2

A single-project trace still shows skew that mirrors multi-tenant clusters.

- Job count vs. GPU-occupied time invert
- CANCELLED jobs dominate occupancy
- Workload composition shifts across months

Job Count vs. GPU-occupied Time

Small jobs dominate count; large jobs dominate GPU-time.



Seeing the same skew without cross-tenant workload mixing.

GPU-occupied time by terminal job state

CANCELLED	**73.5% GPU-time · 9.5% of jobs**
COMPLETED	**26.2% GPU-time**
FAILED	**0.3% GPU-time · 16.9% of jobs**
OTHER	< 0.1%

INTERVIEW

Users stop long-running jobs after inspecting loss curves or validation behavior.

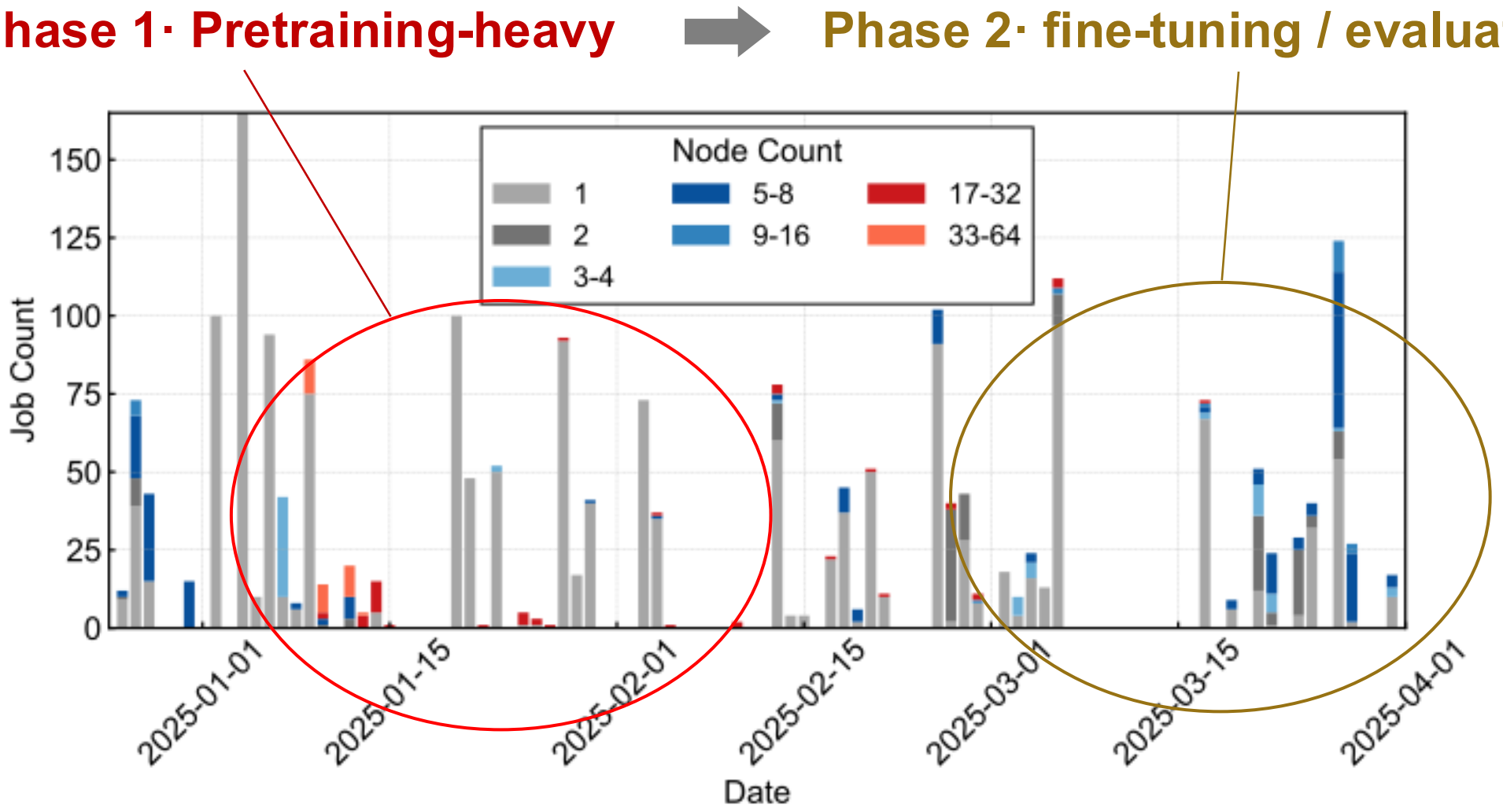
COUNTER-EVIDENCE

Fast failures are caught quickly, not run to completion.

CANCELLED jobs expose early stopping.

Daily job submissions by node count

Resource utilization shifts from large- to medium-scale jobs as the project progresses.

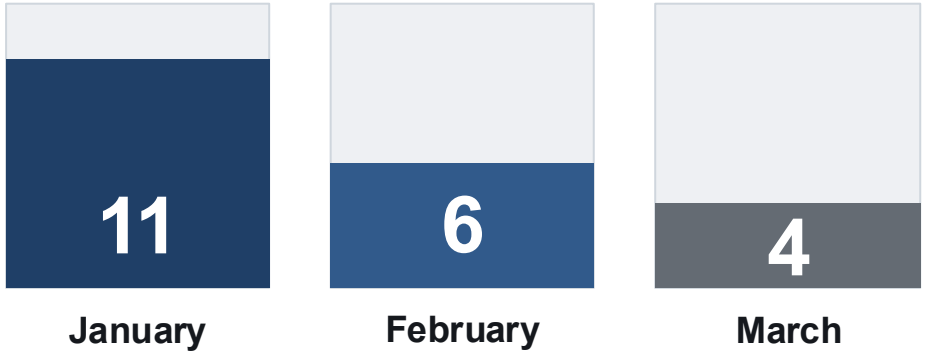


Fault Analysis

GPU-related faults are the most frequent failure mode

Fault Component	Count	Share (%)
GPU (ECC / HW error / unresponsive)	9	42.9
NVLink / NVSwitch / PCIe switch	4	19.0
NIC / transceiver	1	4.8
Interconnect switch (leaf/spine)	5	23.8
Storage switch	1	4.8
Misconfiguration	1	4.8
Total	**21**	**100.0**

Fault events by month



READING

January concentration is consistent with an early burn-in period.

More in the Paper

Implementation details

- Compute-node, NIC-affinity, and storage-system tables
- Software stack: Rocky Linux, containers, Slurm, monitoring

Extended benchmark tables

- HPL, HPCG, HPL-MxP, and IO500 problem sizes/results
- GPT-3 parallelism/MFU plus Llama 2 70B LoRA results

Discussion and limitations

- RoCE ECN/PFC tuning, single-tenant limits, and future telemetry/energy work

Conclusion

Paper(arXiv)

See the poster (next session)

Conclusion

Paper(arXiv)

1. **Sustained LLM development**
   - SONiC / RoCEv2
   - Separate GPU-to-GPU and storage paths
   - Rail optimized topology with Clos and GPU-to-NIC affinity

See the poster (next session)

Conclusion

Paper(arXiv)

1. Sustained LLM development

- SONiC / RoCEv2
- Separate GPU-to-GPU and storage paths
- Rail optimized topology with Clos and GPU-to-NIC affinity

2. TOP-500, and MLPerf benchmarks

- Ranked 49th in HPL, 9 - 17% gap in MLPerf GPT-3
- SONiC / RoCEv2 can be competitive to the proprietary ones

See the poster (next session)

Conclusion

Paper(arXiv)

1. Sustained LLM development

- SONiC / RoCEv2
- Separate GPU-to-GPU and storage paths
- Rail optimized topology with Clos and GPU-to-NIC affinity

2. TOP-500, and MLPerf benchmarks

- Ranked 49th in HPL, 9 - 17% gap in MLPerf GPT-3
- SONiC / RoCEv2 can be competitive to the proprietary ones

3. Job and fault analysis

- GPU-time skew, cancellations-heavy, and phase shifts
- GPU-related faults, not fabric-related, are dominant

See the poster (next session)