USER STORY

Open Networking at Scale: How SAKURA internet Deployed a TOP500 GPU Supercomputer with SONiC

Organization

SAKURA internet Inc. is an internet company founded in 1996. Under the corporate philosophy of "Turning 'what you want to do' into 'what you can do,'" we develop a variety of services to meet customer needs and propose DX solutions that cater to various industries.

Since our founding, which began with the provision of shared server services, we have expanded to offer services such as "Koukaryoku" to support generative AI, and "SAKURA Cloud," which has been conditionally certified for use in government cloud systems. A key feature of our company is that we handle everything from development to operations in-house.

Overview

SAKURA internet is responding to the growing demand for computational infrastructure driven by the rapid adoption of generative AI by continuously procuring next-generation GPUs and strengthening reliable operational systems in our own data centers. As a digital infrastructure company contributing to the sustainable development of the digital society, our mission is to provide cloud services for generative AI.

To continuously meet the increasing demand, we recognized the necessity of resolving the following challenges:

- Ensuring vendor-neutral supply to mitigate risks
- Adopting technologies with high neutrality
- Accelerating delivery speed

Why SONiC?

We selected SONiC for our new 800-GPU cluster because it directly addressed these needs. SONiC provided:

- High transparency, as it is implemented on a Debian/Linux platform
- Active development of new features supported by the global community
- The ability to streamline operations by leveraging the same technologies used for Linux servers

In addition, SAKURA internet has a corporate culture of leveraging OSS and bottom-up initiatives. This culture supported our adoption of SONiC and enabled us to launch a GPU cloud infrastructure in a very short period of time.

Significance

Within just **four months**, SAKURA internet built a new GPU infrastructure powered by SONiC. After a period of trials and refinement, this infrastructure was subsequently commissioned as the GPU supercomputer SAKURAONE.

The service **achieved 49th place in the TOP500** ranking, demonstrating that an open implementation based on SONiC and open source software can deliver world-class results on the international stage. It's the **only system in the global top 100** (June 2025) running a fully open networking stack based on SONiC and Ethernet, a rare distinction that underscores the competitiveness of open, vendor-neutral technologies.

Deployment

Requirements for the GPU Infrastructure

To provide customers with a secure and reliable environment, our GPU infrastructure must support multi-tenancy to ensure security between users, as well as resilience against congestion specific to RDMA communication between GPUs.

To meet these requirements, we leveraged the following technologies:

- EVPN/VXLAN: Ensures tenant isolation and scalability through virtual networking
- RoCEv2 (ECN/PFC/CNP): Enables lossless RDMA communication between GPUs and provides congestion control
- Dynamic Load Balancing (Flowlet mode): Distributes traffic across large-scale parallel workloads

SONiC was chosen because it was among the first commercial distributions to support all of these features.

Network Architecture

To achieve the bandwidth and scalability required for large GPU clusters, we deployed a **Clos topology** with:

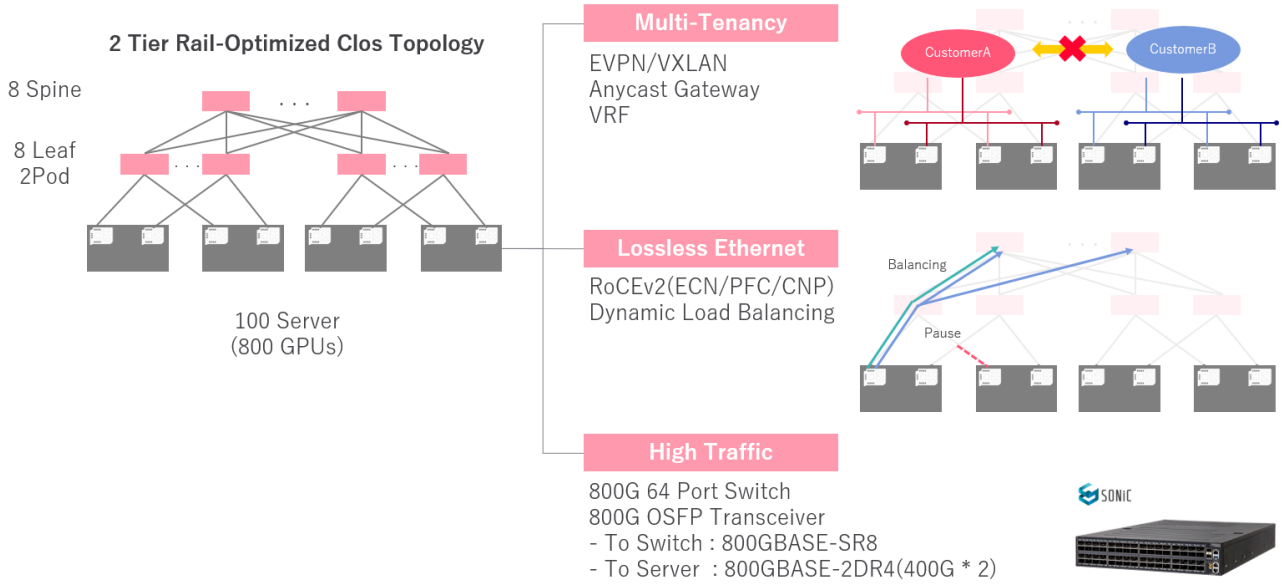

Figure 1. Overview network diagram.

- 2-tier, rail-optimized topology
- Full bisection (Uplink/Downlink 1:1)
- Switch-to-switch connections: 800G (800GBASE-SR8)
- Server connections: 400G (800GBASE-2DR4 / breakout)

OSS-Driven Speed and Agility

Running Debian within SONiC allowed us to leverage our Linux expertise and reduce operational overhead with the following open source tools:

- Ansible: Automation of operations in combination with ZTP
- Prometheus: Flexible metrics monitoring using Exporter and our script
- Python: Development of additional commands as needed

We manage configurations through Infrastructure as Code (IaC) in GitHub, where pull request reviews ensure quality while reducing overhead. Looking forward, we are strengthening our CI/CD pipelines to enable continuous delivery and faster feature expansion.

Benefits

By leveraging SONiC and open source software, we rapidly built and deployed a GPU-focused cloud infrastructure entirely on open technologies. Key outcomes include:

- World-class performance: Achieved Linpack 33.95 PFlop/s and HPL-MxP 339.86 PFlop/s in the TOP500 benchmark, ranking 49th worldwide
- Open and scalable design: Implemented a GPU-specific cloud infrastructure through an open implementation powered by SONiC and OSS
- Cost efficiency and speed: Delivered a GPU infrastructure with exceptional cost efficiency and speed by utilizing OSS throughout the entire process

This deployment demonstrates that open technologies can deliver a highly competitive service combining performance, cost efficiency, and transparency — a milestone outcome for both SAKURA internet and the broader open networking community.

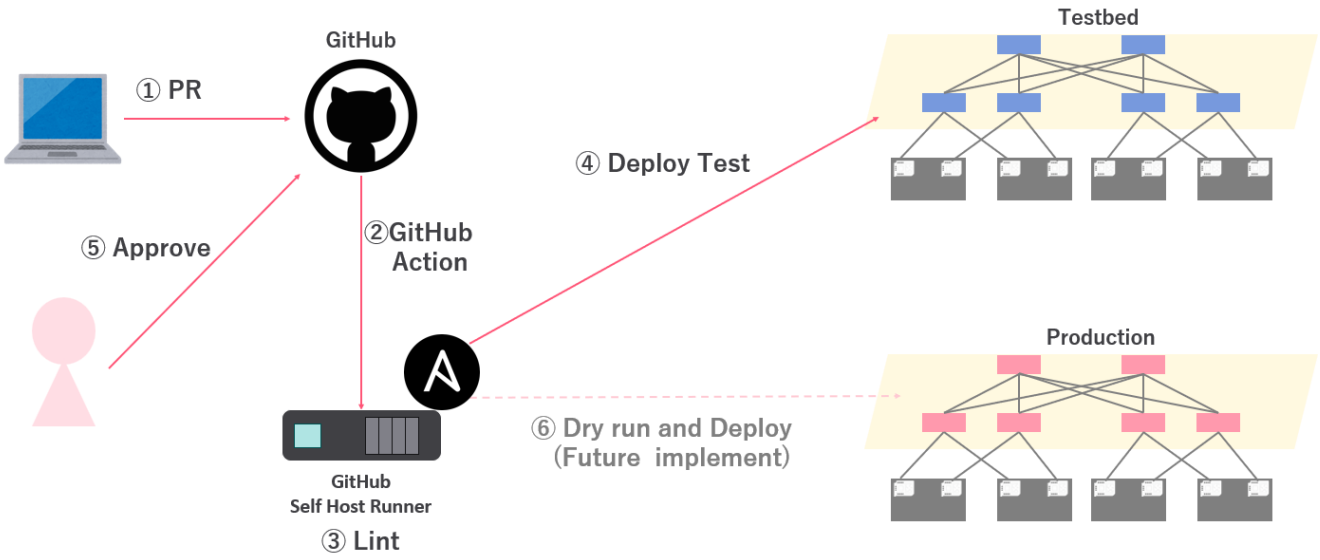

Figure 2. Automation using OSS

Lessons Learned & Future Plans

Our adoption of SONiC and open source software significantly accelerated the launch of our GPU infrastructure. Along the way, we identified several challenges:

- Configuration still requires awareness of specific hardware, which increases the complexity of automation
- Bug tracking can become complicated when multiple branches are involved
- As the product is still evolving, users themselves are sometimes responsible for implementing improvements or additional features

To resolve these challenges, we sometimes examined the source code directly. This ability to "look under the hood" is a unique strength of open source, and it reinforced a key lesson: for networking professionals, understanding the internals of the operating system is indispensable.

Looking ahead, we plan to share our experiences of building GPU infrastructures with SONiC with the broader community. We believe that close collaboration with both vendors and community will be critical to advancing SONiC and strengthening the ecosystem.

Conclusion

In just 4 months, SAKURA internet built an 800-GPU infrastructure powered by SONiC. After a period of trials and refinement, this infrastructure was officially commissioned as the cloud-based supercomputer "SAKURAONE", which ranked 49th in the TOP500, proving that an open source networking stack can deliver world-class performance at scale.

This achievement clearly demonstrates that even a fully OSS-based implementation can compete with the most advanced proprietary systems, combining transparency, neutrality, and performance.

Building on this success, we built a new cloud-based supercomputer and launched it as a commercial service. By using our service, AI researchers and industry players can instantly access generative AI infrastructure without the need to own large-scale GPU clusters themselves.

Through these efforts, SAKURA internet is accelerating the overall pace of AI research and development in Japan and to establish ourselves as a digital infrastructure company that contributes to the sustainable growth of the digital society.

Join SONiC

Become a SONiC member to collaborate, learn and shape the future of the Open Network Operating System.

sonicfoundation.dev/join-sonic