

CHIP-J Semiconductor Study Group Event Report

July 26, 2026 by T. Debun

Event Overview

Date: Thursday, July 24, 2026, 7:00 PM - 9:00 PM

Venue: Niterra North America (Mountain View, CA)

Admission: Free

This study group session featured two presentations: Katayama on "CPU Tidbits and Tales," and Eric on "Democratizing Chip Development." Both talks were followed by lively Q&A sessions with engaged audience participation.

Presentation 1: CPU Tidbits and Tales (Katayama)

Drawing on years of experience in CPU product definition, Katayama covered a wide range of topics from CPU history to cutting-edge packaging technologies.

He began with Intel's Ocotillo campus in Arizona and explained the "Tick-Tock" development cycle, where Tick introduced a new process node and Tock introduced a new architecture. This clockwork rhythm, designed to help PC makers launch products at CES every January, worked smoothly through SkyLake. The cycle broke down when Intel's 10nm transition stalled, resulting in five consecutive products on 14nm from KabyLake through RocketLake.

Katayama shared stories of two "phantom CPUs." CannonLake was Intel's first 10nm product but shipped only to Apple with GPU disabled due to poor GPU yield. Lakefield adopted ARM's big.LITTLE concept with Foveros 3D stacking, but had a brief life from 2020 to 2021. Its hybrid architecture and Foveros technology lived on in AlderLake and beyond.

The evolution from Intel 4004 (about 2,000 transistors in 1971) to PantherLake (3 billion transistors across 5 chiplets) was traced, including the little-known fact that the CPU concept originated from Japan's Busicom calculator company, which proposed the idea to Intel and sent engineers for joint development.

On GPUs, Katayama explained their original purpose for display rendering, including media processing, 4K rendering (calculating colors for 8.29 million pixels 30-120 times per second), multi-monitor support, and digital signage. GPUs became essential for AI because their massive parallel processing capability, originally designed for pixel computation, proved ideal for deep learning matrix operations. He distinguished between NVIDIA's SIMT approach and Intel/AMD's SIMD approach.

Yield and binning received detailed treatment, using AlderLake-N product definitions as examples. Higher-end products had significantly lower yield, while products with certain features disabled achieved much higher yield. "Yield depends on how you define the product" was the key takeaway, offering a product planning perspective that differs from the manufacturing viewpoint.

The presentation concluded with chiplet technologies, covering SoC, MCM, MCP, EMIB, CoWoS, and Foveros, explaining how disaggregating chips into smaller tiles improves yield and cost efficiency.

Presentation 1 Q&A Highlights

Q: How do large wafer-scale chips like Cerebras compare to chiplet-based packaging for AI applications?

A: Large chips are impressive but carry high business risk from defects. NVIDIA will likely move toward chiplets in the future, implementing functions across multiple tiles connected by virtual silicon.

Q: How do CPU lifespan differences between data center and laptop designs reflect in engineering?

A: Lifespan depends on signal pass counts and operating conditions (temperature), following a failure curve. Lowering frequency extends lifespan; raising it shortens it. Data center products assume 24/7 operation but use hot-swapping, so product-defined lifespan is similar to PCs (about 5 years), though total durability requirements are much higher.

Presentation 2: Democratizing Chip Development (Eric)

Eric, who previously researched FPGA virtualization at the University of Toronto, presented his vision of making hardware development accessible through cloud FPGAs and Kubernetes.

He contrasted software development, which is cheap, fast, and accessible to anyone, with hardware development, which requires hundreds of millions of dollars and years of work. While this gap stems from software dealing with "information" versus hardware requiring physical manufacturing, Eric challenged whether it truly cannot be changed.

In the post-Moore's Law era, applications increasingly require both software and hardware. Hyperscalers are building custom chips for AI, video transcoding, and cybersecurity. Data center operators who wanted only CPUs a decade ago now actively seek specialized hardware.

Software development benefits from Agile and CI/CD methodologies that enable rapid user feedback loops. Hardware development lacks this entirely. With FPGA design taking about a year, ASIC manufacturing two to three years, and integration several more months, developers wait three to four years before receiving their first user feedback.

Chiptip Technology's approach creates a small iteration loop using cloud FPGAs managed by Kubernetes. Developers implement designs on cloud FPGAs, deploy them publicly for real user feedback, and iterate rapidly. Only after gaining sufficient confidence do they proceed to ASIC manufacturing, eliminating the need for multiple expensive ASIC cycles.

A key technical differentiator is that Chiptip connects FPGAs directly to the network, unlike AWS's architecture where data must pass through the CPU via PCIe. This results in approximately 30x lower latency, making the platform suitable for large-scale hardware applications.

The FPGA cloud service is planned for launch in late 2026, in partnership with Emotion X (Japan) and QSD (US). Application partners include cybersecurity and 5G/6G communication companies. An FPGA-to-ASIC migration service is planned for 2028.

Presentation 2 Q&A Highlights

Q: How does this differ from traditional FPGA-based simulation acceleration services (Cadence, etc.)?

A: Traditional services are for hardware emulation only and are not accessible to end users. Chiptip's approach deploys applications in the cloud where actual users can access them, enabling the crucial feedback loop.

Q: What about I/O beyond network interfaces (analog interfaces, etc.)?

A: Currently the platform supports PCIe and network interfaces. Target applications are therefore network-oriented, such as cybersecurity and 5G/6G communications.

Q: If the business succeeds, how do you prevent Google or AWS from entering the same market?

A: A fundamental patent on FPGA virtualization in the cloud, filed during research at the University of Toronto about 10 years ago, serves as a barrier to entry. However, eventual acquisition by a major player is acknowledged as a possibility.