Google TPUs, Custom Silicon, and EDA Tools

An expert with 24 years of semiconductor design experience and an investor discussed Google's TPU architecture, the competitive landscape of AI accelerators, and the EDA tools market. The expert's career spanned Sun Microsystems (2001), Apple's CPU design team for iPhones and iPads (2009-2017), building Google's mobile silicon team from scratch in India and the U.S., SiFive where he helped establish the India site, and currently serves as senior director in NXP Semiconductors' digital IP design team.

Google TPU Performance and Architecture

The expert explained that Google initiated the TPU effort in 2015 primarily to address internal compute demands for searches, YouTube, and other services. As the AI market expanded, TPUs became an offering through Google Cloud Platform (GCP). Unlike NVIDIA GPUs, which are more general-purpose and capable of graphics-related tasks, TPUs are custom-built specifically for AI and machine learning computations. This specialization yields superior performance on these specific workloads compared to NVIDIA GPUs.

According to published numbers, TPUs demonstrate performance advantages ranging from 25-30% to nearly 2X better than NVIDIA, depending on the use case. The expert emphasized this represents the fundamental difference between a highly customized design built for one task versus a more general-purpose design. TPUs excel at both training and inference workloads, though the latest version (V6, codenamed Ironwood) is specifically optimized for inference. Prior versions were designed for both large language model training and inference.

Market Adoption and Competitive Position

Based on publicly available information, the expert noted that Google sells TPUs through Google Cloud at hourly usage rates, with active customers using them for both training and inference. Market data indicates TPUs account for approximately 2-4% of all training deployments, while NVIDIA holds roughly 80% market share. This share is projected to increase by a couple of percentage points in 2025.

Regarding total cost of ownership, which includes power consumption and chip cost, Google TPUs demonstrate advantages over NVIDIA GPUs. However, the expert lacked specific performance metrics for inference workloads alone, though he could provide training-specific information.

The CUDA Moat and Software Ecosystem

The expert identified CUDA as the primary reason for NVIDIA's dominant moat in the AI ecosystem, calling it the de facto standard with extensive existing code and software libraries. This creates significant stickiness, making it extremely difficult for customers to migrate their infrastructure from NVIDIA GPU-based data centers to TPUs or other custom AI accelerators.

Google is attempting a similar vertical integration strategy with their own software stack. If customers actively engage with Google and establish their infrastructure on TPUs, they would face comparable difficulty migrating back to NVIDIA. However, since most of the market already uses NVIDIA, the migration to Google TPUs presents the greater challenge.

Competitive Landscape: Microsoft Maia and AWS Trainium

When comparing Google TPUs to Microsoft's Maia chip and AWS's Trainium chip, the expert noted that both Microsoft and AWS started their chip efforts later than Google and have been less aggressive in design optimization. His assessment was that Microsoft's chip is inferior to Google TPUs, though he had not examined detailed performance numbers for these comparisons.

Custom Silicon Efforts by AI Companies

The expert revealed that OpenAI is definitely pursuing custom silicon, having built a team led by a former senior director from Google's TPU group. However, they don't yet have working silicon and are either in the design phase or close to producing their first silicon prototype, based on information from friends in the industry. Anthropic, conversely, is not building their own silicon but is using TPUs as part of their machine mix, according to the expert's internet research.

Future of Merchant vs. ASIC Chips

Looking at the longer-term outlook, the expert discussed potential pathways for competition. One approach involves CUDA compatibility, which startups like Rivos are attempting. Rivos is building an AI accelerator with a RISC-V CPU core, with their value proposition being a software layer that seamlessly translates CUDA programming for their chip. The expert expressed skepticism about whether such CUDA-compatible approaches could succeed in training without violating NVIDIA's restrictions.

He identified inference as a different landscape altogether. The inference field is highly fragmented, spanning from complex chips on clouds to IoT devices selling for $1-$2. These chips must be extremely cost-sensitive in some cases and power-sensitive in others,

and the CUDA barrier exists to a lesser degree than in training. Inference represents the space where new entrants can potentially succeed.

The expert was uncertain whether the barrier to CUDA compatibility was business-related (high royalty payments to NVIDIA) or technical (NVIDIA not allowing it). Many people remain skeptical whether companies like Rivos can deliver CUDA-compatible chip software without violating NVIDIA's restrictions. He had heard about Baidu's Kunlun chip and Alibaba's chip but had not examined their details.

Barriers to Entry in Inference Silicon

The expert explained that building a chip for inference is not very difficult, particularly for designs targeting lower TOPS (trillions of operations per second) without aggressive power optimization. The architecture is fairly standard: matrix multiplication units, memory interface, a CPU scheduler to distribute tasks to the AI unit, and a DDR or external memory interface.

Since the field is extremely new, many companies are introducing innovative ideas. Significant optimization can be achieved through software to reduce the number of computations needed for inference questions. Many companies leverage a software-first approach. At the SoC architecture level, substantial improvements can be made to reduce back-and-forth traffic between execution units and external memory, which burns power and loses performance.

Various companies are exploring different approaches: near-memory compute so data doesn't travel far from memory (reducing power and latency), compute within the memory itself, and other relatively simple but effective changes. The AI and intelligent systems space on the edge shows considerable variation, with systems for robotics relying heavily on vision, other use cases depending more on audio inputs, and some being text-based. Startups are emerging to cater to very specific use cases within this varied landscape.

Google's Custom Silicon Team Organization

The expert described Google's smart organizational approach, maintaining a lean team by largely outsourcing design hardening to Broadcom. Google handles the chip architecture and writes the Verilog, then hands it to Broadcom for implementation. Google focuses on architecting the TPU itself, while Broadcom provides the skeleton of peripheral components needed to build a chip around the TPU—including inputs/outputs like PCIe, UCIe components, USBs, analog components, and interconnects. Google does not handle most of these peripheral elements.

This division of labor enables significant innovation despite not having as large a team as some merchant chip manufacturers. Additionally, because Google controls the full software stack, they can aggressively optimize hardware according to software—similar to NVIDIA with CUDA, but unlike Intel or AMD who may not control the complete end-to-end software stack.

Broadcom, MediaTek, and Marvell Comparison

The expert assessed Broadcom as having the most experience specifically in AI accelerators among these players. While MediaTek excels in the mobile space where they've built chips for a long time and competed well with Qualcomm, Broadcom leads in the AI space with proprietary in-house IPs. Their interconnect IP for chip-to-chip communication is particularly strong. The expert rated Broadcom as probably the best in this domain, with Marvell attempting similar work but likely not as close—Broadcom maintains a good head start over Marvell.

EDA Tools: Synopsys vs. Cadence Deep Dive

The expert provided extensive insight into the EDA tools landscape, noting that most companies use both Synopsys and Cadence rather than locking into one vendor. This enables benchmarking between tools, selecting the best option, and controlling prices through competitive tension.

Synthesis and Place-and-Route Tools

From Synopsys, the expert has used Fusion Compiler, their latest offering for physically aware synthesis followed by place and route. The downstream sign-off for timing is largely handled by PrimeTime across the industry, which is a Synopsys tool. Very few companies use Cadence's equivalent tool, Tempus. Synopsys claims that Fusion Compiler, coming from the same code base as PrimeTime, is very well correlated—meaning what you build and model closely matches actual sign-off results, theoretically yielding better end results.

Cadence offers Genus as the synthesis tool and Innovus for place and route, with Genus also capable of physically aware synthesis. However, since sign-off still uses PrimeTime, teams must ensure Genus plus Innovus correlates well with PrimeTime. For any new technology node, this correlation effort requires one to two months of work.

The expert observed that both tool suites (Cadence's Genus plus Innovus with PrimeTime sign-off, versus Synopsys's Fusion Compiler plus PrimeTime) are reasonably close in final results. However, for the most critical IPs, Cadence appears to have an edge. For CPUs specifically, Cadence engaged early with Apple, then Arm and Qualcomm,

becoming the EDA tool of choice for all three companies' CPU designs. CPUs push EDA tools to their limits because they demand maximum frequency while mobile chips require sensitivity to power and area. This intensive stress testing gave Cadence's tools substantial refinement. While Synopsys has started catching up, Cadence still maintains an edge for IPs demanding the most aggressive optimizations.

Historical Context: The Synopsys-Cadence Competition

The expert provided historical context from around 2013-2014 when the competitive dynamic shifted. Until then, Synopsys led due to an exclusive relationship with Intel. However, this exclusivity and market leadership bred overconfidence, and customer service standards declined. Meanwhile, Cadence aggressively partnered with key customers in the mobile space using Arm architecture.

The expert personally experienced this transition while at Apple. Despite initial resistance from engineers, close collaboration between Apple and Cadence over one to two years dramatically improved the tools. Cadence's R&D team fully opened up to Apple with very close collaboration, significantly enhancing tool quality. This gave Cadence a substantial head start from that point forward.

Around 2018-2019, Synopsys released a new place-and-route tool called ICC2 (succeeding their original ICC). This tool proved very difficult to use and had numerous issues, setting Synopsys back with many customers. The industry expects year-on-year improvements in power, performance, and area through EDA tools, but this problematic Synopsys offering failed to deliver such improvements. This triggered additional customer migration to Cadence, at least for high-performance designs.

Current State and Stabilization

The expert reported that the situation has now stabilized, with Synopsys clawing back market share. They've established much better relations with Arm. Previously, when Arm developed a new CPU, they would test it on Cadence first, then provide it to Synopsys a year later. Now Synopsys engages much earlier with Arm. The Fusion Compiler tool is substantially better than ICC2, and the gap between the two vendors has reduced considerably. For AI-related blocks with significant parallel computation, Fusion Compiler sometimes actually outperforms Cadence tools.

Arm's Divestiture to Cadence

Regarding Arm's sale of their foundation IP to Cadence, the expert explained that Arm divested their Artisan library group because they're focusing increasingly on building actual SoCs. For a long time, Arm had not been making money from the memories and

standard cells they designed. The divestiture represents focusing on fewer things and executing them well. Cadence simply offered them a better deal. Synopsys already had in-house memories and standard cells, so they didn't need this acquisition. Cadence wanted something to compete with Synopsys in this area.

Hardware Emulation Tools

The expert has used both ZeBu (Synopsys) and Palladium (Cadence) hardware emulation platforms. He has not heard that one toolset is significantly better than the other. His experience suggests companies balance their financial engagement between the two vendors—if a company selects more Cadence tools for synthesis and place-and-route, they might choose Synopsys for emulation. However, the expert suspects Cadence's Palladium may be cheaper than Synopsys's ZeBu, based on observing more startups using Palladium. Cost could be a determining factor for resource-constrained companies.

Analog Tools

For analog design, the expert noted that companies he's worked with use Cadence Virtuoso. He has not seen people using Synopsys tools in this domain, though he couldn't speak definitively about the broader industry. This suggests Cadence maintains its historical lead in analog tool offerings.

Ansys Tools and the Synopsys Acquisition

The expert's teams use Ansys tools, particularly for electrical analysis including IR drop and electromigration. He believes the Synopsys-Ansys combination could make a significant difference because physics is Ansys's key strength. In electrical analysis, Ansys is by far the most commonly used tool. Cadence offers a competing tool called Voltus, but most customers use Ansys.

The acquisition enables Synopsys to offer an end-to-end suite. They already have PrimeTime for timing, PTPX (PrimePower) for power analysis, and now with Ansys they gain IR analysis capabilities. They also have physical verification tools. This complete package from synthesis through sign-off allows aggressive optimization so designs are "correct by construction."

At the chiplet level, Ansys has strong capabilities in interconnects, package modeling, and thermal effects. The expert views this as a big plus for Synopsys. For chip-level design, if customers want a single-vendor solution, preference may shift toward Synopsys. Under this scenario, the expert would rate Synopsys 8-9 out of 10, while Cadence would score a 6 due to their weaker sign-off tool suite. However, customers will likely continue using synthesis and place-and-route from Cadence with sign-off in

Synopsys, since Cadence's synthesis and place-and-route tools remain superior to Synopsys.

The expert clarified the typical workflow: synthesis and place-and-route with Cadence, then final verification with Synopsys. For Synopsys to leverage Ansys and gain market share in synthesis and place-and-route, they must successfully integrate sign-off tools within place-and-route itself. Currently, after final place-and-route, teams spend approximately three weeks fixing IR drop and electromigration issues. If this engine becomes part of place-and-route, performing these fixes during optimization, significant downstream effort would be saved, likely yielding better performance and power. Successfully integrating and demonstrating better results would provide Synopsys a major advantage and could potentially drive digital tool recapture among Arm, Apple, and Qualcomm.

Design Velocity and AI Automation

Many companies are aggressively leveraging AI to speed up chip design processes, with initial benefits already visible. The expert expects substantial benefits to become visible within one to two years. He anticipates sizable reductions in turnaround time from architecture specification to tape-out, along with decreased staffing requirements. While the process won't become as automated as software development with prompt engineering, significant efficiencies are achievable.

The expert cited startups publishing tools that can take architecture specifications and generate SoC RTL. Much of this is doable because interfaces are standardized. If blocks are well-defined in the architecture spec, manual RTL writing becomes unnecessary. He expects at least 25% speed-up in the design phase.

However, he noted that design represents only about two-thirds of the overall time-to-market from architecture conception. The chip then goes to manufacturing, where foundries take four months to return the first chip, followed by post-silicon verification. This portion of the cycle will persist, somewhat limiting the overall benefit.

On the design side, companies will require fewer people for designs. Many companies today don't produce as many chips as desired due to cost or inability to hire sufficient skilled people. With more automated designs, the number of chips being designed will increase, with more fine-tuning and custom-tailoring. For edge AI specifically, the next two to three years should see newer applications and novel approaches to reduce power and improve latency, resulting in diverse chip launches.

Additionally, semiconductors are becoming strategic for countries. China, India, and possibly the European Union will encourage their local semiconductor industries to

design more chips. This could result in a more fragmented market but with significantly more chips being designed overall.

Siemens EDA (Mentor Graphics)

Calibre from Siemens is the most dominant physical verification tool in the industry. Synopsys comes in a distant second, and Cadence's tool Pegasus is an even more distant third. Calibre is the industry standard for physical verification.

Beyond Calibre, Siemens offers tools covering synthesis, place-and-route, and power analysis, though the expert was uncertain if they have a static timing analysis tool. They acquired a startup called ATopTech that was building synthesis and place-and-route tools. Broadcom had been using those tools before Siemens acquired Mentor Graphics, which brought Calibre into the Siemens portfolio.

The synthesis and place-and-route tool originally developed by ATopTech was good seven to eight years ago, but the expert questioned whether they've kept pace with the industry. He hasn't seen prominent customers using their tools for synthesis and place-and-route, suggesting limited market traction beyond their dominant Calibre offering.

Foundry IP Landscape

The expert was familiar with standard cells and SRAMs from foundries but less so with analog IPs. He explained that TSMC's offerings are adequate for generic SoC designs—their memories and IPs—but don't provide enough choices for designs aggressively optimizing for frequency or power. Their design set is rather limited. If customers actively engage with TSMC, they have teams designing high-performance IP, but these come at a cost and may not be available to every customer.

The standard set TSMC offers is inferior to what Synopsys provides. Arm stopped developing their libraries after 16nm, so the expert had no information beyond that node. At 5nm, Synopsys is considerably better than TSMC.

Interface IP Market

The expert noted that Synopsys's biggest IP portfolio is in the interface space, including PCI Express, Compute Express Link, USB, and Ethernet. For DDR, continuous innovation is definitely happening. However, for interfaces like PCIe, innovation may not be as aggressive. Some peripherals are reasonably standard. The innovation is happening more on interconnects providing chip-to-chip connection, such as UCIe IPs and CXL that facilitate chiplet architectures.

Regarding whether Synopsys competes with Arm through customized IP blocks, the expert didn't think so. For high-value IPs like CPUs, Synopsys is not building IPs with Arm architecture. They're attempting something on RISC-V, though he was uncertain how aggressively they're pursuing it. While some competitive dynamics may exist, Synopsys is not competing in Arm's core bread-and-butter business.

Final Assessments

When asked to rate Google TPU on a 1-10 scale for adoption beyond internal workloads, the expert differentiated between technical merit and practical adoption. For performance per watt and product quality, he rated TPUs 8-9 out of 10. They undergo continuous improvements with full-stack integration between software and hardware, making them superior to NVIDIA solutions in that aspect. However, from an end customer's perspective where ease of use matters, TPUs score 5 out of 10 due to NVIDIA's CUDA moat.

For the EDA vendors, with the Ansys acquisition enabling single-vendor solutions, the expert rated Synopsys 8-9 out of 10, while Cadence received a 6 due to their weaker sign-off tool suite. However, he noted that customers will likely continue using synthesis and place-and-route from Cadence with sign-off in Synopsys, as Cadence's synthesis and place-and-route tools remain superior to Synopsys offerings.