Source: AI generated by GoogleGemini

Can Google really break Nvidia's grip on AI hardware?

Could Google's custom AI chips (or TPUs) dent Nvidia's dominance in the AI chip market? While Nvidia's dominance faces scrutiny, the AI market is vast. Whether rivals can scale fast enough to dent Nvidia's leadership remains an open question. Alphabet clearly benefits, but disruption is far from certain.

Jakub Dubaniewicz
Senior Equity Analyst
jakub.dubaniewicz@syzgroup.com

SYZ PRIVATE BANKING

Introduction

The question of whether Google's custom AI chips— Tensor Processing Units or TPUs— could meaningfully erode Nvidia's dominance has stormed back into the spotlight. Yet the debate over how long Nvidia can maintain its overwhelming market share is anything but new. Nor have TPUs suddenly materialised from nowhere; Google has used them for more than a decade, and the architecture is now in its 7th generation.

What makes this moment different is a rare intersection of catalysts. On one side, market conditions have shifted: the recent pullback in AI winners, persistent chatter about whether we are living through an AI bubble, and a rotation within mega-cap tech have all sharpened investor sensitivity to signs of disruption. On the other side, Alphabet has staged a powerful comeback as a market favourite, buoyed by a sequence of developments that collectively showcase the strength of its AI hardware strategy.

In quick succession, we learned of an expanded Google–Anthropic partnership, involving large-scale commitments to use TPUs; the successful launch of the latest Gemini model, trained entirely on TPUs; and reporting that Meta is in advanced discussions to spend billions deploying Google's TPUs in its own datacentres from around 2027. Together, these events underscore how valuable Google's custom silicon has become in powering the next phase of AI.

Alphabet clearly stands to benefit. But whether this constitutes a material threat to Nvidia is still heavily debated. Will Nvidia lose some market share? Even the most bullish analysts acknowledge it will. The market opportunity is so vast—and so lucrative—that multiple players are racing to develop accelerators capable of capturing slices of it. The key question up for debate is whether TPUs can be compelling enough, and scale quickly enough, to meaningfully dent Nvidia's leadership.

The longstanding effort to challenge Nvidia

The idea that Nvidia's 90%-plus share of AI compute is ripe for structural erosion has been around for years. AMD has pursued the opportunity relentlessly but ultimately falling short of the developer traction, ecosystem depth and sheer pace of iteration that Nvidia delivers. Broadcom has made no secret of its ambitions either, outlining aggressive custom silicon goals over the past twelve months. Huawei, meanwhile, has pushed forward with its Ascend line, gaining domestic Chinese momentum but struggling to scale globally.

In this context, Google's TPUs in partnership with Broadcom are not a sudden arrival, but another well-resourced entry in a broader competitive pattern. The difference is that Google has the combination of scale, real-world deployment and software stack maturity that many challengers lack. Its internal footprint for TPUs is massive, proven and commercially relevant. And unlike newcomers, Google's architecture is already battle-tested across models as heavy as Gemini.

So, the competitive dynamic is not a story of "TPUs arrive and upend the hierarchy overnight", but rather a continuation of a decade-long trend: credible alternatives appear, generate incremental pressure, and force Nvidia to move even faster.

Why are TPUs now gaining more attention?

1. Their technical profile has matured. Google's seventh-generation TPU is materially more capable than early iterations and is designed around large-scale model training and inference at hyperscaler scale. Performance per watt is strong. Networking fabric is tightly integrated. And critically, the stack is deeply aligned with the needs of Google's own workloads.
2. The economic proposition can be attractive for partners. Anthropic's decision to deepen its reliance on Google reflects not only TPU performance, but favourable economics tied to broader cloud commitments.
3. Meta news is strategically significant. Meta has historically been aligned with Nvidia, building AI infrastructure around its GPUs with little incentive to diversify. The fact that Meta is even considering deploying TPUs at material scale speaks to a genuine desire for supply diversity, better cost predictability and insulation from GPU shortages that have defined the past two years.

These developments highlight a simple truth: Google's TPUs are not niche, experimental or untested. They are a viable, scalable accelerator platform that will increasingly be used by more players in the ecosystem.

Why companies such as Meta are interested in TPUs, and custom chips in general:

- Supplier diversification – Large hyperscalers, such as Meta, cannot rely on a single chip vendor in a supply-constrained market. It makes sense that Meta is looking to develop its own custom chip (in partnership with Broadcom) as well as explore possibility to use Google's own custom chip.
- Vertical cost optimisation – Where silicon can be co-designed with a chip expert such as Broadcom or Marvell, such custom AI chip can materially outperform Nvidia's GPU on cost performance.
- Economics at extreme scale – For firms training multi-trillion-parameter models, even marginal efficiency gains are meaningful. This is why the largest hyperscalers explore alternatives to Nvidia's GPUs.

This explains the interest from Meta and Anthropic. The aim is not to replace Nvidia wholesale, but to supplement it with a credible second source.

Bottlenecks and constraints to TPU adoption?

Despite growing interest in TPUs, they are not exempt from the same industry-wide supply constraints shaping the entire AI accelerator market:

- **Advanced node manufacturing capacity** – Only TSMC, Samsung and Intel can manufacture high-end AI chips. Even with aggressive expansion plans, their capacity is expected to grow around 14% CAGR, reaching roughly 69% above 2024 levels by 2028. But demand is growing even faster, ensuring a persistent structural shortage.

- **Advanced packaging limits (CoWoS and equivalents)** – The highest-performance accelerators, including Nvidia's H100–B200 range, AMD's MI300 and Google's TPUs, all rely on sophisticated 2.5D and 3D packaging. This remains a major bottleneck, with all foundries struggling to expand throughput quickly enough.
- **HBM shortages** – High Bandwidth Memory (HBM) remains supply-constrained. Memory vendors are scaling output, but demand from all major players continues to exceed supply.

None of these shortages are driven by weak demand – quite the opposite. Hyperscaler appetite remains extremely strong across all vendors. This is precisely why TPUs can gain share without meaningfully hurting Nvidia: the limiting factor is supply, not demand. The market is expanding faster than anyone can manufacture chips.

TPU-specific constraints - echo other attempts to challenge Nvidia

Where TPUs differ from GPUs is in their tight coupling to Google's infrastructure. Their performance depends not just on silicon but on Google's interconnect topology, software tooling, orchestration layers and optical-switching fabric. Reproducing this carefully finetuned environment externally is possible but complex.

Large-scale TPU adoption by external clients like Meta faces challenges. Migrating workflows built outside the Google ecosystem is costly and complex, requiring major engineering effort to adapt models, pipelines and tools—while relying on a smaller software ecosystem than CUDA. There's also strategic unease in depending on a rival's proprietary hardware. With Meta TPU deployment expected around 2027, it will be competing directly against Nvidia's Rubin and subsequent architectures, raising the competitive bar even further. This parallels patterns seen with other Nvidia challengers:

- AMD produced competitive hardware with the MI300 series, but software tooling (ROCm) remained a limiting factor.
- Huawei built powerful Ascend accelerators, but ecosystem limitations constrained adoption outside China.

TPUs are far more mature and have deeper internal optimisation than previous challengers. But the lesson holds: competing with Nvidia requires more than a fast chip—it requires a full-stack ecosystem.

Meta's reported interest reflects this nuance. The company is likely to test TPUs within Google Cloud before making any call about deploying them in its own datacentres, where replicating Google's tightly integrated system would require major engineering investment.

Why is overtaking Nvidia so difficult?

Across all challengers—whether AMD, Huawei or Google—three obstacles consistently emerge.

The software moat: CUDA remains the industry's dominant foundation. Two decades of optimisation give Nvidia a lead that is extraordinarily difficult to replicate. TPU software is powerful within Google's environment, but the broader ecosystem remains far smaller.

The system-level advantage: Nvidia sells complete platforms—networking, interconnect, software orchestration and tightly optimised servers. Rivals must compete not with one component, but with the entire stack. TPUs come closest to matching this coherence but are still tuned primarily for Google's infrastructure.

The pace of execution: Nvidia has moved to an annual architecture cadence, accelerating from Hopper to Blackwell to Rubin. Even fast-moving companies struggle to match this. By the time a rival reaches volume deployment, Nvidia is often preparing its next generation. TPUs will continue to grow, but this fast cadence limits how much ground rivals can gain in any given cycle.

Is Nvidia at risk? Only at the margins

The most important point is often forgotten: this is not a zero-sum game.

The demand for AI compute is exploding. Supply remains constrained. Even meaningful TPU adoption does not displace Nvidia in practice, because there is more demand than supply for every credible accelerator class. Analysts already assume some loss of share—that is not the risk.

The risk would be Nvidia missing revenue or margin expectations. Today, there is little evidence of that. Nvidia still leads in systems, software, performance and cadence. TPU adoption will rise, but Nvidia is on track to remain the market leader for the foreseeable future.

Conclusion – both are likely winners in a rapidly expanding market

The TPU debate deserves fresh attention. Alphabet's silicon efforts are proven and growing in relevance, with Meta's interest adding weight. Yet Nvidia's moat remains notoriously tough to challenge.

TPUs are the strongest alternative platform developed to date. They will gain share. Alphabet will monetise this. But Nvidia's leadership remains intact.

This is not a zero-sum contest. It is a rapidly expanding market with room for multiple winners—and right now, the two companies best positioned to win are Nvidia and Alphabet.

Welcome to Syzerland®

For further information

Banque Syz SA
Quai des Bergues 1
CH-1201 Geneva
T. +41 58 799 10 00
syzgroup.com

Jakub Dubaniewicz
Senior Equity Analyst
jakub.dubaniewicz@syzgroup.com