VAUGHAN NELSON

September 2026

Winners & Warnings – Volume 5

Artificial Intelligence: More Compute, Fewer Architectures

Why the Custom Application-Specific Integrated Circuits (ASIC) Boom May Be the Real Bubble

Introduction

There are few irrefutable truths in investing. Prudent capital allocation, however, is almost always at the heart of success. Companies, industries, and countries that invest in high-return opportunities, while others hesitate, consistently outperform.

The Winners & Warnings series examines these dynamics in real time, identifying the conditions that shape tomorrow's leaders. Guided by the Capital Allocation Framework (CAF), we assess not only what companies do, but when and why. By tracking industry investment levels and returns on capital, we aim to capture the inflection points where leadership changes.

AI may not be the bubble. The belief that everyone needs their own AI chip might be.

Investors continue to debate whether the enormous amount of capital flowing into artificial intelligence represents a bubble. We think that may be the wrong question. AI demand continues to grow, and the returns on AI infrastructure may ultimately justify a much larger buildout. The excess may instead be developing in a narrow part of the ecosystem: the assumption that every major technology company needs its own semiconductor architecture.

Custom AI chips have proliferated across Google, Amazon, Microsoft, Meta, OpenAI and others. Industry forecasts increasingly assume that proliferation continues, with MediaTek estimating that the AI ASIC market could reach approximately $70–$80 billion in 2027.

That outlook embeds two important assumptions. First, today's period of architectural experimentation must become tomorrow's permanent industry structure. Second, the semiconductor companies enabling those custom chips must retain enough of the economics to justify investor expectations.

Both assumptions may prove too optimistic.

Custom ASICs Solved a Procurement Problem

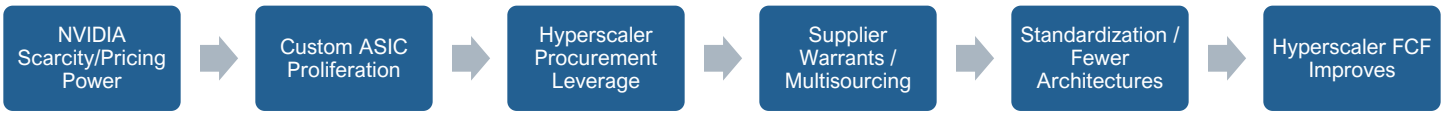
The decision to develop custom silicon was rational. AI demand accelerated faster than semiconductor supply could respond. Leading GPUs were expensive, difficult to obtain and concentrated around a dominant supplier. For hyperscalers, spending tens of billions of dollars on infrastructure, developing an alternative offered lower costs, greater supply diversity and negotiating leverage.

Viewed through this lens, custom silicon is as much a procurement strategy as a technology strategy. A custom ASIC does not need to replace NVIDIA to create value. It may only need to lower the cost of a particular workload or provide a credible alternative that improves procurement economics.

Some programs will clearly justify themselves. Amazon, for example, believes Trainium can generate substantial capital-spending savings at sufficient scale. The argument is not that custom ASICs cannot work. It is that achieving those economics requires scale, utilization and continued investment across successive generations.

The market may therefore be extrapolating a rational response to scarcity into a permanent industry structure in which every proprietary architecture reaches economic scale.

How the Custom ASIC Cycle Could Evolve



The Hyperscalers Are Capturing the Economics

Importantly, the hyperscalers' procurement strategy has not stopped with creating alternatives to NVIDIA. They are increasingly applying the same pressure to the companies enabling their custom chips.

Recent warrant agreements provide an unusually visible example. AMD has granted equity warrants tied to large AI purchase commitments from Meta and OpenAI. Marvell has now entered a similar structure with Google as part of a major custom-silicon relationship. The terms differ, but the economic principle is similar: when the hyperscaler creates enormous incremental revenue for the semiconductor supplier, it increasingly expects to participate in the resulting equity value.

These agreements can be thought of as **massive volume rebates paid in equity**.

They do not mean the chips are technologically weak. They do suggest that bargaining power may sit with the customer. A highly differentiated supplier normally captures value through price and margin. In these agreements, suppliers are instead willing to share part of the value creation in order to secure the volume.

That distinction matters because custom ASIC forecasts are generally expressed in revenue. But revenue growth and economic value creation are not the same thing. Even if the custom ASIC market reaches $70–80 billion, hyperscalers may use their purchasing concentration to demand lower pricing, multiple suppliers, continuous cost reductions and participation in supplier upside.

Custom ASICs were designed in part to reduce NVIDIA's pricing power. The irony may be that they create an industry in which the hyperscalers hold the pricing power over the custom ASIC suppliers.

NVIDIA Is Selling an Architecture, Not a Chip

Custom ASIC suppliers may therefore be surrendering economics to win hyperscaler volume at the same time the standardized merchant alternative is becoming more valuable.

NVIDIA is no longer simply selling a GPU. Its platforms increasingly combine GPUs, CPUs, networking, rack-scale infrastructure and software into an integrated computing architecture. CUDA may be the most important part of that advantage.

A proprietary accelerator must continuously support its own software, compilers, libraries and engineering. NVIDIA benefits from a much broader ecosystem of researchers, model developers, infrastructure companies and independent developers continuously improving performance on its platform. Innovation outside NVIDIA can therefore improve the economics of installed NVIDIA hardware long after the chip has been purchased.

Developments such as TileRT illustrate the point. Performance gains can increasingly come from software, kernels, scheduling and model optimization without designing another accelerator. The hurdle for custom silicon is therefore not simply whether an internal chip is cheaper or faster at launch. It is whether that advantage can survive everything the broader software ecosystem discovers over the useful life of the hardware.

Scale in AI may therefore be measured not only by how many chips an architecture ships, but by how many people are working to make that architecture better.

Elon Musk's recent decision at SpaceX provides an interesting real-world example. Musk has extensive experience developing proprietary AI silicon and a long history of vertical integration. Yet as SpaceX prepares for a major expansion in AI compute, Musk has said the company intends to build exclusively around NVIDIA's Vera Rubin architecture.

The decision is notable precisely because Musk has the capability and willingness to build custom chips. It suggests an important distinction: the ability to build a custom chip does not mean building one is the optimal capital-allocation decision.

NVIDIA may not even need custom ASICs to disappear. NVLink Fusion allows custom accelerators to operate within NVIDIA-based infrastructure. The eventual outcome could therefore be multiple chips existing inside far fewer independent computing architectures.

The Hyperscaler Can Win Either Way

The hyperscaler therefore has multiple paths to better returns.

If custom silicon succeeds, the hyperscaler can use its purchasing scale, multiple suppliers and warrant structures to capture more of the economics. If the incremental return on a proprietary architecture begins to fall, it can slow the roadmap, extend an existing generation, narrow the chip to specific workloads or shift incremental demand toward a standardized merchant platform.

Neither outcome requires the hyperscaler to become less bullish on AI.

This is important because proprietary architecture requires ongoing research and development, engineering, software development, tape-outs, manufacturing commitments and supporting infrastructure. The next phase of the AI investment cycle may not be about spending less on AI. It may instead be about generating more compute from every dollar invested.

That creates the possibility of an unusual divergence: **hyperscaler free cash flow could improve at precisely the moment custom ASIC suppliers begin to disappoint.**

The Warning

The investment risk is not that artificial intelligence fails. It is that AI succeeds while the number of economically relevant architectures becomes more concentrated and the hyperscalers capture a larger share of the economics.

The custom ASIC thesis therefore requires investors to be right twice: **the architectures must reach scale, and the suppliers must retain the economics.**

Some programs will succeed. Google may sustain TPU. Amazon may sustain Trainium. Other custom accelerators may prove highly valuable for specific workloads. But AI compute can grow dramatically without every proprietary architecture reaching scale, and custom ASIC revenue can grow without the suppliers capturing the value implied by that growth.

The custom ASIC market does not need to disappear for today's expectations to prove excessive.

It only requires investors to have overestimated how many architectures will reach economic scale—and how much of the value their suppliers will be allowed to keep.

AI may not be the bubble. The belief that everyone needs their own AI chip might be.
