

NVIDIA: Where Should NVLink Fusion Draw the Platform Boundary?

Internal ID: S&F-STR-412

Topic: Customer Strategy, Platform Strategy, Operations Strategy

Industry: Semiconductors

Source: Strategy & Finance

Publication date: 5 October 2026

License: Not specified

Decision date and information cutoff: October 2, 2026, 4:00 p.m. ET

Decision scope: How broadly NVIDIA should scale NVLink Fusion over fiscal 2028-2029

Case convention: This is an author-created strategy exercise. Company disclosures are historical; the decision is hypothetical. No private deliberations are represented.

The decision

NVIDIA has announced a way for partners and customers to connect custom compute silicon to NVLink and NVIDIA rack-scale systems. It has also expanded its relationship with Marvell and invested \$2 billion in the company. The public record, however, does not disclose Fusion production deployments, customer volumes, revenue, unit economics, or the extent to which partner designs might replace NVIDIA GPUs.

The strategy question is how far to take the platform. NVIDIA could keep Fusion selective, supporting custom compute only where it fits within tightly governed NVIDIA racks and networking. It could broaden access into a widely available third-party compute platform. Or it could focus engineering and supply on NVIDIA-designed GPUs and integrated systems, limiting Fusion to announced commitments and narrow exceptions. Which path should it take over the next two years, and what evidence should trigger a change in course?

This is a choice about platform scope, access and interfaces-not a decision to abandon the GPU roadmap or a claim that Fusion has already reached commercial scale. The public evidence leaves open whether custom compute will expand total AI-infrastructure demand, retain NVIDIA content in specialized workloads, or substitute for NVIDIA accelerators.

A large growth engine and substantial commitments

For the quarter ended July 26, 2026, NVIDIA reported \$96.2 billion in revenue, including \$89.0 billion from Data Center and \$48.7 billion in hyperscale Data Center revenue. These categories are NVIDIA's reported classifications; they do not isolate GPU sales or identify ultimate end users. [P1]

Revenue exposure is concentrated. One direct customer represented 16% of quarterly revenue. Three direct customers represented 16%, 15% and 13% of first-half revenue. NVIDIA also estimated that an AI research and deployment company contributed meaningfully through purchases from its cloud customers. Direct customer percentages do not show the complete end-user picture: cloud channels can obscure who ultimately uses the compute, and direct and indirect exposures may overlap. [P1]

NVIDIA reported \$279 billion in supply and capacity commitments as of July 26, up from \$119 billion in the previous quarter. The commitments are primarily for memory and manufacturing capacity across current and future architectures. NVIDIA says some supplier agreements can be changed before firm orders, potentially at additional cost. The figure is not a Fusion budget, a shipment total, or a firm customer order book. The filing also describes supply constraints and risks that customers may delay purchases because of capital, land, power, construction or facility limits. [P1]

Management's August earnings-call transcript projected top-five hyperscaler capital expenditure of nearly \$800 billion in 2026 and \$1.3 trillion in 2027, and said customer growth forecasts exceeded what NVIDIA expected to supply. These are management projections, not realized spending or NVIDIA orders. [P7] If demand continues to outrun supply, a partner route could let NVIDIA participate in systems using specialized compute. But a larger addressable infrastructure forecast does not show which silicon customers will buy or whether facilities will be ready to use it.

Fusion: an announced route for custom silicon

In May 2025, NVIDIA described NVLink Fusion as a semi-custom AI infrastructure approach. Its announcement said partners could build custom compute that connects through NVLink and can be combined with NVIDIA rack-scale systems and networking. The release named MediaTek, Marvell, Alchip, Astera Labs, Synopsys and Cadence in the ecosystem, and said Fujitsu and Qualcomm planned custom CPUs to pair with NVIDIA GPUs. NVIDIA used broad compatibility language, including that customers could use "any ASIC"; this remains the company's announced capability, not independent proof of universal compatibility or deployed systems. [P2]

In March 2026, NVIDIA and Marvell announced an expanded partnership through Fusion and a \$2 billion NVIDIA investment in Marvell. Their release discussed custom silicon, optical connectivity, AI networking and AI-RAN. It did not disclose investment terms, customer deployments, volumes, returns or Fusion-specific economics. The investment signals strategic commitment, but not how many designs will reach production or what NVIDIA will earn from them. [P3]

Several strategic mechanisms could support Fusion. Custom accelerators might serve predictable workloads where a customer values specialized performance, efficiency or supply choices. NVIDIA could retain system, fabric, networking or software attachment even when it does not supply every compute chip. A broader ecosystem could create demand for the integrated platform and extend customer relationships. Yet each is a hypothesis: public materials do not quantify NVIDIA content per Fusion system or show whether such systems add demand or displace GPU purchases.

The alternative is to prioritize NVIDIA-designed GPUs and integrated systems. That may preserve product differentiation and focus scarce engineering, supply and customer-support capacity on NVIDIA's own roadmap. It could also leave customers seeking custom compute to build separate systems or deepen relationships with alternative platforms. Whether Fusion would ease a supply bottleneck is unknown: the filing does not separate NVIDIA-controlled from partner-controlled capacity, and data-center power and readiness remain constraints even when chip options expand. [P1]

Customers and rivals are developing their own silicon

The competitive picture is not a simple choice between merchant GPUs and custom chips. Google says its TPU fleet trains and serves Gemini, Veo, Imagen and Anthropic's Claude. Its November 2025 Ironwood announcement positioned the chip for training and inference, said general availability was expected "in the coming weeks," reported Anthropic's plan to access up to one million TPUs and described early customer testing. The source does not establish Ironwood general availability by this case cutoff or that the named models run specifically on Ironwood. Google's fleet-wide TPU use and its Ironwood-specific claims should therefore be kept distinct. [P4]

AWS says Trainium3 supports production workloads and reports customer cost benefits. Microsoft says its Maia 200 accelerator is deployed in Azure for inference and compares its performance per dollar with hardware in its fleet. These are company and customer claims, drawn from different workload and system contexts; they are not common-benchmark comparisons. [P5] [P6] They indicate that large cloud platforms are investing in alternatives, but do not establish comparative performance, total cost or market share on a like-for-like basis.

These examples leave an important boundary question. Hyperscalers can use custom accelerators for their own workloads while continuing to buy NVIDIA systems, and may make custom compute available to third parties. For NVIDIA, the strategic issue is whether Fusion becomes a governed route into those systems, a broader platform whose participants include potential competitors, or a distraction from integrated NVIDIA products.

What platform research can-and cannot-tell NVIDIA

Platform research separates firm scope, who may participate and which interfaces are open. Those are linked choices, but "open versus closed" can obscure the real design variables: NVIDIA could permit selected partners while retaining control over NVLink, rack integration, software and support. [A1]

Boudreau's study of handheld-computing platforms found an inverted-U association between licensing access to hardware developers and new-device introductions. Intermediate access policies were associated with the highest rate; the study does not establish a profit-maximizing policy for NVIDIA. Its historical handheld sample and product-launch outcome differ from AI infrastructure and GPU economics. [A2]

Gawer and Henderson's qualitative study of Intel describes the tension a platform owner faces when it both encourages complementary innovation and competes in adjacent markets. Separation and IP-sharing practices were among Intel's approaches. The case is evidence of a possible complementor-trust problem, not proof that a specific NVIDIA governance structure will succeed. [A3]

Adner and Kapoor's study of semiconductor lithography suggests that integration's value can vary with where bottlenecks sit and how mature a technology generation is. The study's historical, single-industry estimates should not be applied to NVIDIA; the practical question is whether the binding constraint is within NVIDIA-controlled components or among external complements such as power, facilities and customer capital. [A4]

The choice

Choose a path for NVLink Fusion over fiscal 2028-2029:

1. **Selective, governed access:** Work with chosen partners and customers on workload-specific custom compute while retaining defined control over NVLink, rack systems, networking, software and support.
2. **Broad platform access:** Invite a wider set of silicon partners and customers to build around Fusion, accepting greater variety and partner autonomy to increase the chance of ecosystem growth.
3. **GPU and integrated-system focus:** Prioritize NVIDIA-designed accelerators and complete systems, keeping Fusion narrow and limiting new partner enablement.

Set out the assumptions behind your choice. Identify how you would distinguish new demand from GPU substitution; what NVIDIA content and economics must remain in a system; which interfaces and support obligations should be shared; and how partner trust, customer concentration, supply commitments and infrastructure readiness affect the decision. Propose evidence gates for expanding, pausing or narrowing Fusion. Do not infer commercial results from announcements, treat supply commitments as shipments or a Fusion budget, or use vendor performance claims as comparable benchmarks.

Sources and conventions

All company facts are limited to information public by October 2, 2026, 4:00 p.m. ET. Reported results, management estimates, forecasts, announced partnerships and vendor claims are distinguished. No later operating or deployment outcomes are included. Academic sources provide concepts and mechanisms; no coefficients, profit forecasts or NVIDIA-specific effects are assumed.

- **P1:** NVIDIA Q2 FY2027 Form 10-Q, quarter ended July 26, 2026, filed Aug. 27, 2026; MD&A pp. 24-29 for revenue, supply constraints, demand risks and customer concentration; Note 10, p. 17 for supply and capacity commitments. [SEC filing](#)
- **P2:** NVIDIA, “NVIDIA Unveils NVLink Fusion for Industry to Build Semi-Custom AI Infrastructure With NVIDIA Partner Ecosystem,” May 18, 2025, full release, especially paras. 2-5. [NVIDIA Investor Relations](#)
- **P3:** NVIDIA, “NVIDIA AI Ecosystem Expands as Marvell Joins Forces Through NVLink Fusion,” Mar. 31, 2026, paras. 1-6. [NVIDIA Newsroom](#)
- **P4:** Google Cloud, “Announcing Ironwood TPUs General Availability and new Axion VMs to power the age of inference,” Nov. 6, 2025; opening and Ironwood availability/customer sections. [Google Cloud Blog](#)
- **P5:** Amazon, “Frontier agents, Trainium chips, and Amazon Nova: key announcements from AWS re:Invent 2025,” updated Dec. 4, 2025; Trainium3 and AI Factories sections. [About Amazon](#)
- **P6:** Microsoft, “Maia 200: The AI accelerator built for inference,” Jan. 26, 2026. [Microsoft Blog](#)
- **P7:** NVIDIA Q2 FY2027 earnings call, Aug. 26, 2026, corrected transcript, pp. 2-3. Figures are management forecasts and statements. [Transcript PDF](#)
- **A1:** Gawer, A. (2020), “Digital platforms’ boundaries: The interplay of firm scope, platform sides, and digital interfaces,” *Long Range Planning*, 53(2), 102045. [DOI](#)
- **A2:** Boudreau, K. J. (2010), “Open Platform Strategies and Innovation: Granting Access vs. Devolving Control,” *Management Science*, 56(10), 1849-1872. [DOI](#)

- **A3:** Gawer, A. & Henderson, R. (2007), "Platform Owner Entry and Innovation in Complementary Markets: Evidence from Intel," *Journal of Economics & Management Strategy*, 16(1), 1-34. [DOI](#)
- **A4:** Adner, R. & Kapoor, R. (2010), "Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations," *Strategic Management Journal*, 31(3), 306-333. [DOI](#)