Futurum
RESEARCH

After Huang, NVIDIA's AI Factory Blueprint Still Sets the Standard

NVIDIA's Reference Designs Span Power, Compute, Agents, and Robots, but Europe Shows the Limits No Blueprint Can Engineer Away

Analyst(s): Nick Patience
Publication Date: November 12, 2026
Document #: AIONP202611

What You Need to Know

- The Blueprint Is the Product: At GTC Berlin in October 2026, NVIDIA showed that it now specifies the mechanical, electrical, and software design of the entire AI factory, from DSX POD containers and 800 VDC power to the OpenShell agent runtime, handing partners a bill of materials to build.
- Software Monetizes the Blueprint: NVIDIA will license its proprietary agent software at about $2,000 per GPU per year, on top of a software business already worth more than $1 billion a year, and the growth market is AI factories outside the large clouds, the same buyers DSX POD targets.
- Europe Shows the Limits: Grid queues, expensive on-site power, and thin risk capital constrain the strategy in Europe. NVIDIA's answer is to divert compute to wherever it can be switched on fastest.
- Geography Decides Where the Blueprint Executes: Cheaper power puts the US ahead, Taiwan builds much of the blueprint, and Europe's best opportunity lies in the power and grid equipment layer, where it already has strong companies.

The Futurum View

NVIDIA has moved from selling components to writing the specification for the whole AI factory, and the resignation of Jensen Huang on November 12 makes that shift more significant. A strategy built on reference designs, from DSX POD containers and 800 VDC power to the OpenShell agent runtime, lives in specifications that partners already build to, so it depends far less on one executive's judgment than NVIDIA's earlier, founder-led bets did. We believe the blueprint makes NVIDIA's platform harder to displace than at any point in its history, because a rival now has to match a system and an ecosystem rather than a chip, although most of its costs fall on NVIDIA's partners rather than its customers. Energy and capital sit outside the blueprint – though NVIDIA has been deploying its own capital aggressively – and how much of each a region can supply will decide where AI factories get built.

Huang resigned days after acknowledging an affair with Alice Huang, an executive at NVIDIA Cloud Partner Megaspeed whom US intelligence agencies reportedly suspect of working for China's Ministry of State Security. No charges have been filed, and Megaspeed is the subject of a US Justice Department criminal inquiry into whether NVIDIA chips reached restricted Chinese buyers, which it denies. The episode leaves the architecture of NVIDIA's strategy intact but strains two of its foundations: the trust governments place in NVIDIA as a sovereign AI partner, and the diligence it applies to the neocloud partners that are the most natural buyers of DSX POD and its licensed software.

The Blueprint Strategy

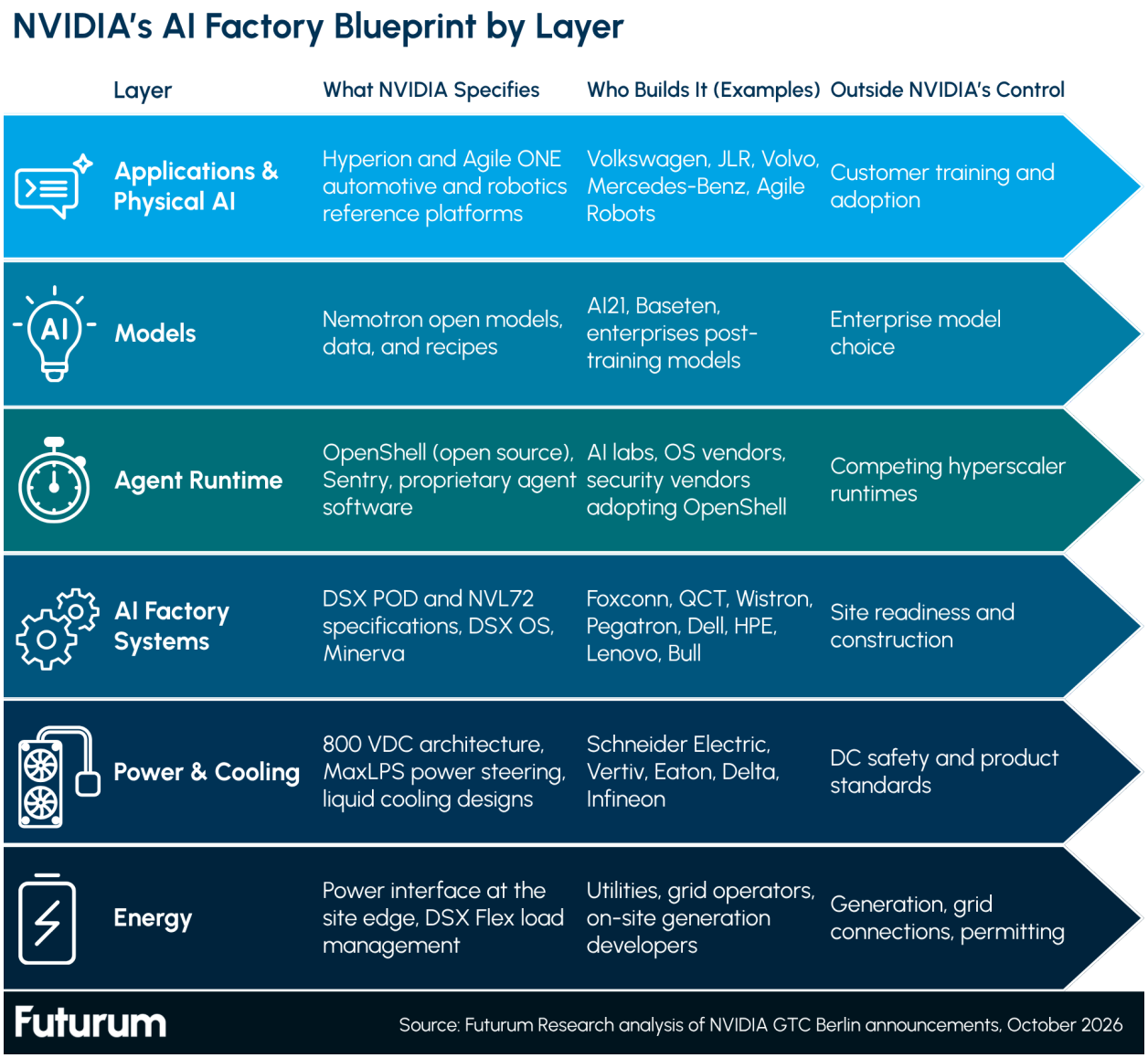
NVIDIA has built its business on reference designs for most of the past decade, with DGX systems giving the industry a standard AI server and NVL72 moving rack assembly from the customer's data center into the supply chain. DSX POD extends that logic to a shipping container holding seven Vera Rubin racks. Huang said at GTC Berlin that DSX POD's mechanical architecture, electrical architecture, and software APIs are already specified, with the cooling system stacked on top and power components built to a common standard so that everything is pluggable, and that the first shippable unit should arrive toward the end of 2026. NVIDIA describes its approach as vertically integrated but horizontally open, meaning partners are free to build and sell the systems as long as they meet its core specifications, with NVIDIA supplying a bill of materials for them to take to the customer.

The same pattern runs through every layer NVIDIA addressed in Berlin. In power and cooling, its 800 VDC power architecture sets the voltage and leaves others in the ecosystem such as Infineon, Schneider Electric, Vertiv, and Eaton to build the rectifiers, power racks, and converters to match. In the agent runtime, OpenShell defines how agents are contained and monitored, while in applications, Hyperion gives carmakers the computer, sensors, safety operating system, and driving software, and Agile ONE gives robot makers a pre-integrated humanoid starting point.

Figure 1 adapts Huang's five-layer cake of energy, chips, AI factories, models, and applications into six layers, separating power and cooling from energy, folding chips into AI factory systems, and adding the agent runtime that OpenShell now occupies, to show where NVIDIA's specification ends and its partners' work begins. NVIDIA specifies very little in the energy layer.

©2026 The Futurum Group, LLC. All rights reserved.

Figure 1: NVIDIA's AI Factory Blueprint by Layer



Source: Futurum Research analysis of NVIDIA GTC Berlin announcements, October 2026

The PC Question

A journalist at GTC Berlin put the obvious objection to Huang directly, asking how the industry avoids the fate of the 1990s PC market, where one company defined the standards across every layer and the rest of the business stopped being profitable. Huang's answer was that a PC is like a hammer, finished and good enough when it leaves the store, whereas AI has to be taught domain skills by its owner, much as an apprentice becomes an artisan. Asked separately what would preserve NVIDIA's lead as rivals improve, he said the question had been put in "a very Intel way," because it used words such as control and competition, and argued that NVIDIA's real advantage is 25 years of consistency on one architecture. CUDA, he noted, runs wherever both Arm and x86 run, in every cloud and every neocloud.

©2026 The Futurum Group, LLC. All rights reserved.

Huang's apprenticeship argument has merit for customers, because the value an enterprise creates by training models and agents on its own data stays with that enterprise, although the PC comparison applies more directly to NVIDIA's partners. When NVIDIA specifies the mechanical, electrical, and software design of a system and hands partners a bill of materials, their room to differentiate shrinks, and system builders, OEMs, and infrastructure suppliers increasingly compete on execution, delivery, and service rather than on design. That is a healthy outcome for buyers and a demanding one for the ecosystem that NVIDIA depends on to reach them.

Software Is the Recurring Layer

Huang was unusually specific about how NVIDIA intends to make money from the software that runs on its AI factories. At the GTC Berlin press Q&A, he said that while all of the software is open, the parts that are NVIDIA proprietary will be licensed to enterprises at about $2,000 per GPU per year to run the software that operates agents on NVIDIA AI factories, and that software licensing will become an increasingly major share of the company's business. He added that NVIDIA's software business is already well over $1 billion a year, and that supporting AI factories outside the large clouds requires NVIDIA to provide and support the software that operates them.

Hyperscalers bring their own software, operations teams, and increasingly their own silicon, whereas neoclouds, enterprises, telcos, and sovereign AI factories need NVIDIA to supply much of the operating stack. Those are precisely the buyers DSX POD is designed for, so the hardware blueprint and the software license reinforce each other. NVIDIA added to that operating stack at GTC Berlin with Minerva, an AIOps platform for AI factory observability whose collector runs inside DSX OS, NVIDIA's operating layer for multi-tenant AI factories, with neocloud IREN as a launch partner. The license is material to NVIDIA but small for the customer, and a one-gigawatt AI factory with 500,000 GPUs would generate about $1 billion a year in license revenue for NVIDIA, against the roughly $40 billion a year in revenue Huang says such a factory should produce for the customer.

Huang predicted that OpenShell will be everywhere, just like Linux and the Chrome browser, and said it already has the support of more than 100 companies, including leading AI labs, operating system vendors, and cloud providers. Adoption of the free runtime creates demand for the strongest enforcement, Sentry running on NVIDIA's BlueField hardware, and for the proprietary agent software for which NVIDIA can charge.

Standardizing Infrastructure Does Not Guarantee Adoption

Further up the stack, NVIDIA's reach is thinner, and although its Nemotron open models ship with weights, training data, and recipes, in Futurum's AI Product Series survey of September 2026, only 6.7% of respondents said their company currently uses Nemotron for AI application development, against 74.2% for Anthropic Claude and 67.5% for OpenAI GPT (n=511). However, interest is building, with 18.0% planning to evaluate or adopt Nemotron, up from 13.8% in March 2026.

©2026 The Futurum Group, LLC. All rights reserved.

Physical AI shows the same gap between standard and market, as NVIDIA says humanoids are the most general-purpose robots but that only around 10,000 will ship this year, because robot intelligence is not yet good enough and there are too few real-world examples to train on beyond simulation, which is why it is releasing open models, GR00T, and the Agile ONE reference platform ahead of demand. In both cases, NVIDIA is attempting to set the standard before the market forms as it did with CUDA.

Europe as the Base Case

Europe was the setting for GTC Berlin, testing if the NVIDIA blueprint can work globally. Huang named three objections to AI growth in his Q&A (safety, energy, and capital) and offered NVIDIA technology as the answer to the first. NVIDIA reference designs are intended to unlock the other two yet face steep barriers to adoption globally.

Energy

Europe's main data center hubs, including Frankfurt, London, Amsterdam, Paris, and Dublin, face grid connection queues measured in years, against construction timelines measured in months. Ireland ended its de facto moratorium on new Dublin connections in December 2025 but now requires large new sites to bring matching dispatchable generation or storage, and the Netherlands restricts hyperscale development to two northern municipalities. The workaround common in parts of the US, generating power on site behind the meter, is available in Europe but far more expensive. In May 2026, European benchmark gas averaged around $16 per million British thermal units against less than $3 at Henry Hub in the US, before the EU Emissions Trading System adds a carbon cost to every megawatt-hour generated on site, and planning authorities are increasingly attaching renewable-fuel conditions to new generators.

Asked whether NVIDIA would take a bigger role in Europe's energy problem, Huang argued that AI demand is the best opportunity in a century to fund new energy generation of every kind, from fission and fusion to hydro, and that because NVIDIA's architecture is fungible, systems destined for a delayed site can be shipped to another customer. If compute flows to wherever it can be switched on fastest, grid queues cost Europe AI capacity as well as time, and capital is already moving; AWS chose Aragon in Spain for its €15.7 billion investment largely because grid capacity was available there. Moving AI factories outside Europe in search of cheaper power is often not an option either, because sovereignty and data residency requirements keep many public sector and regulated workloads within European jurisdiction, which makes the grid problem harder to route around.

Inside the AI factory, a second problem comes from NVIDIA's 800 VDC architecture, which converts grid power to direct current at the edge of the site. Safety and product standards for DC at that voltage are immature everywhere, and Europe faces an additional complication, since NVIDIA's design uses monopolar 800 VDC, the Open Compute Project backs bipolar ±400 VDC, and a European specification from ODCA and VDE sits at 750 VDC. That risks leaving European sites on a different standard from both NVIDIA and the hyperscalers.

©2026 The Futurum Group, LLC. All rights reserved.

Capital and Partners

Of the 17 DSX POD partners NVIDIA named, only Bull and Schneider Electric are headquartered in Europe, while six are Taiwanese, and Huang confirmed that NVIDIA will manufacture systems in Europe for European customers through Bull. He also told European audiences that the region needs a far more vibrant risk capital ecosystem, and his response to a question about delays at London-based neocloud Nscale was blunt, noting that customers who once accepted three years to build a data center now expect six months, and suggesting the company's employees try less sleep. Within days of GTC Berlin, neocloud shares fell sharply as credit tightened, a reminder that the buyers who most need NVIDIA's blueprint and its licensed software are also among the most financially exposed.

Megaspeed, the Singapore-based NVIDIA Cloud Partner at the center of the allegations, shows that the same channel also carries compliance risk. It reportedly bought more than $4.8 billion of NVIDIA equipment over two years and is the subject of a US criminal inquiry into whether chips reached restricted Chinese buyers, which it denies. As more of NVIDIA's growth depends on neoclouds and resellers rather than hyperscalers, the way it vets and monitors those partners will matter as much to customers and regulators as the blueprint itself.

Huang's advice to Europe was to start where it is strongest rather than everywhere at once, and we think the strongest candidate is the power and cooling layer of the AI factory. European companies, including Infineon, Schneider Electric, ABB, Siemens, and STMicroelectronics, are core partners in NVIDIA's 800 VDC ecosystem, supplying the power semiconductors, rectifiers, switchgear, and conversion equipment that every AI factory now needs, and Infineon timed a new collaboration with NVIDIA on high-voltage power conversion to coincide with GTC Berlin. That is a more realistic sovereignty opportunity than a European AI chip, because it builds on industrial strengths Europe already has.

Beyond Europe: The US and Asia-Pacific

The US

The blueprint executes fastest in the US, where cheap domestic gas makes behind-the-meter generation viable at scale, and where federal regulators, although now requiring co-located loads that rely on the grid to pay for transmission, have not prohibited co-location or dedicated generation. The largest buyers, however, are hyperscalers with their own standards and silicon, as Google, Meta, and Microsoft back the bipolar ±400 VDC specification, build custom accelerators, and will each want their own agent runtimes. NVIDIA's blueprint wins most easily where buyers lack the scale to write their own, which again points to neoclouds, enterprises, and sovereign customers.

©2026 The Futurum Group, LLC. All rights reserved.

Asia-Pacific

Asia-Pacific, and Taiwan in particular, is where the blueprint is built. Six of the 17 DSX POD partners are Taiwanese, including Foxconn, Quanta, Wistron, Pegatron, GIGABYTE, and Delta, and Huang highlighted a 2.5x expansion in substrate capacity and the co-packaged optics NVIDIA developed with TSMC as central to scaling supply. Foxconn's Kaohsiung-1 AI factory already runs NVIDIA's 800 VDC architecture. Moving integration into the supply chain concentrates more of each AI factory's value, and more of its risk, in one region, which links back to Europe's sovereignty debate and to the resilience questions any buyer of DSX PODs will ask. We leave the wider APAC sovereign AI market, including Japan, Korea, and India, for separate coverage.

What to Watch

- OpenShell Governance: Whether NVIDIA moves OpenShell to a neutral foundation, which would strengthen its claim to be an industry standard, or keeps it NVIDIA-led, which would invite hyperscalers to back alternatives.
- Software Disclosure: Whether NVIDIA begins reporting software revenue separately as the agent license scales, and how many enterprises outside the large clouds pay the $2,000 per GPU fee, which NVIDIA's new leadership has yet to restate.
- Hyperscaler Counter-Standards: Progress on the ±400 VDC specification, custom accelerators, and competing agent runtimes, which together mark the competitive edge of NVIDIA's blueprint in the US.
- European Grid Reform: Connection-queue reform and flexible-connection rules for AI factories, the factor most likely to decide whether DSX PODs are deployed in Europe or elsewhere.
- Partner Economics: Margins and differentiation at OEMs, system builders, and infrastructure suppliers as NVIDIA specifies more of each system.
- Governance and Export Compliance: Who succeeds Huang, whether the board picks continuity or a reset, and whether the Megaspeed inquiry brings tighter scrutiny of NVIDIA Cloud Partners in Southeast Asia.

Other Insights from Futurum

At GTC Berlin, NVIDIA Puts the AI Factory in an Actual Container

NVIDIA Cosmos 3 and Open Agent Tools: Is Physical AI About to Leave the Lab?

Nokia's AI-RAN Platform Puts a Number on the Software-Defined RAN Bet

At GTC 2026, NVIDIA Stakes Its Claim on Autonomous Agent Infrastructure

©2026 The Futurum Group, LLC. All rights reserved.

About Us

About the Authors

Nick Patience is VP and Practice Lead for AI Platforms at The Futurum Group, where he covers AI infrastructure software, hyperscaler AI platforms, foundation model providers, and AI infrastructure companies. He has around 27 years of experience as a technology industry analyst and co-founded 451 Research, which was acquired by S&P Global in 2019. He can be reached at npatience@futurumgroup.com.

About The Futurum Group

Every day, The Futurum Group's analysts, researchers, and advisors help business leaders worldwide anticipate tectonic shifts in their industries and leverage disruptive innovation. Unlike traditional analysts, The Futurum Group works not only in analysis and research but also takes that insight and knowledge even further, engaging all the way through the go-to-market process.

Futurum Research provides in-depth research and insights on global technology markets using advisory services, custom research reports, strategic consulting engagements, digital events, go-to-market planning, and message testing. It also creates, distributes, and amplifies rich media content that all stakeholders read, watch, and listen to.

See more details on The Futurum Group at futurumgroup.com.

Copyright & Use License

Copyright Notice

Copyright ©2026 by The Futurum Group, LLC. All rights, including that of translation into other languages, are specifically reserved. No part of this publication may be reproduced in any form, stored in a retrieval system, or transmitted by any method or means, electrical, mechanical, photographic, or otherwise, without the express written permission of The Futurum Group futurumgroup.com. United States copyright laws and international treaties protect this publication. Unauthorized distribution or reproduction of this publication, or any portion of it, may result in severe civil and criminal penalties and will be prosecuted to the maximum extent necessary to protect the publisher's rights.

License Notice

This document may be distributed within the licensed organization only.

The following acts are prohibited:

Transmittal to others outside your immediate organization, including partners, resellers, external consultants, etc., in any media format

Posting on a website that is accessible to others outside your immediate organization

The possession or use within an unlicensed organization

Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. The Futurum Group shall have no liability for errors, omissions, or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve their intended results. The opinions expressed herein are subject to change without notice.