What Is NVIDIA Kyber? Rack-Scale System Explained (2026)

Published July 15, 2026 31 min read

Executive Summary

NVIDIA Kyber is the company's fourth-generation rack-scale computing architecture, designed as the successor to the current "Oberon" rack design that underpins the NVIDIA GB200 NVL72 and GB300 NVL72 systems (Source: blogs.nvidia.com) (Source: www.datacenterdynamics.com). Kyber is built to house NVIDIA's forthcoming Rubin Ultra graphics processing units (GPUs), the flagship accelerator of the Vera Rubin platform, and its flagship configuration connects 576 Rubin Ultra GPUs in a single scale-up NVLink domain (Source: blogs.nvidia.com). Where prior racks such as GB200 NVL72 mount GPU trays horizontally and rely on cable cartridges, Kyber rotates compute blades vertically "like books on a shelf" and replaces cabling with a large printed-circuit-board midplane (Source: blogs.nvidia.com) (Source: www.servethehome.com).

Kyber is inseparable from a second architectural shift: an 800 volt direct current (800 VDC) power distribution system that NVIDIA says will support 1 megawatt (MW) racks starting in 2027, replacing the 54 VDC in-rack distribution used since the Hopper and early Blackwell generations (Source: developer.nvidia.com). NVIDIA states that full-scale production of 800 VDC data centers will coincide with Kyber rack-scale systems in 2027 (Source: developer.nvidia.com), citing benefits including a roughly 5% end-to-end efficiency gain, up to 70% lower maintenance costs, and up to a 30% reduction in total cost of ownership (Source: developer.nvidia.com) (Source: developer.nvidia.com).

The GPU inside Kyber, Rubin Ultra, roughly doubles the silicon of a standard Rubin GPU, packing four 800 mm² compute dies instead of two and pairing them with 16 HBM4e (fourth-generation high-bandwidth memory, enhanced) stacks totaling 1 terabyte (TB) of capacity (Source: www.heise.de) (Source: www.heise.de). A single Rubin Ultra package is projected to deliver roughly 100 petaflops of FP4 (4-bit floating point) compute, and a full 144-package Kyber rack is projected to reach 15 FP4 exaflops of inference throughput or 5 FP8 exaflops of training throughput (Source: www.heise.de) (Source: www.heise.de).

Kyber's release date is now contested. NVIDIA's own roadmap places Rubin Ultra, and by extension Kyber, in the second half of 2027 (Source: www.datacenterdynamics.com). However, research firm SemiAnalysis reported in early July 2026 that the Kyber NVL144 rack has slipped more than 12 months to 2028 because the PCB midplane at the heart of the design "remains challenging from a manufacturability standpoint," and that the larger NVL576 configuration is also likely delayed or limited to small volumes (Source: www.cnbc.com) (Source: www.cnbc.com). NVIDIA publicly rejected the report, stating "Our roadmap is intact" (Source: www.cnbc.com).

In the meantime, the current-generation Vera Rubin NVL72 rack, using standard Rubin GPUs rather than Rubin Ultra, entered full production in June 2026 with partner availability in the second half of 2026 (Source: vrlatech.com), delivering up to 3.6 exaflops of NVFP4 (NVIDIA's 4-bit floating point format) performance per rack, roughly five times the output of the GB200-generation systems it replaces (Source: redmondmag.com). Cloud providers including Microsoft Azure, Amazon Web Services, Google Cloud, Oracle Cloud Infrastructure, and CoreWeave are already building Rubin-ready facilities, and Microsoft's Fairwater "AI superfactory" sites in Wisconsin and Atlanta illustrate how quickly rack-scale generations must be absorbed into live data center infrastructure (Source: redmondmag.com) (Source: news.microsoft.com). This report explains what Kyber is, how it differs from the Oberon rack it replaces, what Rubin Ultra brings to the platform, why 800 VDC power delivery is a prerequisite rather than an optional upgrade, and how the reported manufacturing delay reshapes the competitive landscape against AMD's Helios rack and Google's TPU Ironwood pods.

Introduction and Background

Modern artificial intelligence (AI) training and inference workloads no longer run on individual servers; they run on entire racks engineered as a single computer. NVIDIA calls facilities built around this principle "AI factories," and it has shipped three successive rack-scale architectures since 2024: the GB200 NVL72, the GB300 NVL72, and now the Vera Rubin NVL72, each connecting dozens of GPUs into one non-blocking interconnect domain that behaves, from a software perspective, like a single massive accelerator (Source: developer.nvidia.com). NVIDIA founder and chief executive officer (CEO) Jensen Huang has framed this as an annual cadence: "With our annual cadence of delivering a new generation of AI supercomputers... Rubin takes a giant leap toward the next frontier of AI" (Source: nvidianews.nvidia.com).

NVIDIA Kyber is the next step past that cadence: a rack architecture purpose-built for the Rubin Ultra GPU, the higher-density successor to the standard Rubin GPU that is currently shipping inside the Vera Rubin NVL72. Where Vera Rubin NVL72 is built on NVIDIA's third-generation MGX (Modular GPU Accelerated Extreme-Scale) rack architecture (Source: developer.nvidia.com), Kyber represents a structural break from that lineage: it discards the horizontal server-tray layout that has defined NVIDIA racks since Hopper and replaces it with vertically mounted compute blades, a solid printed-circuit-board midplane, and a rear-mounted NVLink switch backplane(Source: www.heise.de).

Understanding Kyber requires understanding the rack it replaces. NVIDIA's current GB200 NVL72 and GB300 NVL72 racks are internally codenamed "Oberon," a name NVIDIA itself has not used in official marketing but that has become the accepted shorthand across the trade press and independent hardware analysts covering the company's roadmap (Source: www.nextplatform.com) (Source: glennklockwood.com). Oberon-class racks have carried NVIDIA's AI infrastructure business through two GPU generations, Blackwell and Blackwell Ultra, and analysts at TechInsights estimate that this system-level integration strategy, rather than raw silicon gains alone, is central to NVIDIA's forecast that data center infrastructure revenue will reach $1 trillion by 2028 (Source: library.techinsights.com).

This report examines what NVIDIA Kyber is, how its architecture and 800 VDC power delivery differ from the Oberon rack, what the Rubin Ultra GPU contributes to the platform, how Kyber fits within the broader Vera Rubin platform and NVIDIA's rack-scale roadmap through 2027, and what the SemiAnalysis-reported manufacturing delay means for enterprises, cloud providers, and NVIDIA's competitors as of July 2026.

What Is NVIDIA Kyber? Definition and Rack-Scale Taxonomy

NVIDIA Kyber is a rack-scale server architecture, not a single chip. It refers to the physical rack, midplane, power delivery system, and NVLink switch fabric that together house and interconnect NVIDIA's Rubin Ultra GPUs and Vera CPUs (central processing units) at data center scale. In NVIDIA's own words, Kyber is "the next-generation MGX NVL rack design that will double the NVLink domain per rack to fit 144 GPUs" relative to the Vera Rubin NVL72 rack (Source: developer.nvidia.com).

Kyber exists in a taxonomy of scale-up configurations rather than as a single fixed product. At the smallest documented scale, a single Kyber rack is described as an NVL144 system, packing 144 Rubin Ultra GPU packages, a figure independently corroborated by GTC coverage describing "144 Rubin Ultra GPUs with 72 self-designed Vera processors" (Source: www.heise.de). At a larger scale, NVIDIA describes Vera Rubin Ultra NVL576, which combines "eight separate MGX NVL racks, each with 72 Rubin Ultra GPUs, all in a single 576-GPU NVLink domain with copper and direct optical connections" (Source: developer.nvidia.com), consistent with the figure most widely cited in trade press, that Kyber "connects 576 Rubin Ultra GPUs" (Source: blogs.nvidia.com). At the largest documented scale, NVIDIA describes Kyber NVL1152, which links eight full Kyber racks (8 x 144 GPU packages) into one 1,152-GPU all-to-all domain using direct optical interconnects, and NVIDIA states this configuration "provides the foundation for the next era of extreme scale-up AI computing using NVIDIA Feynman," the architecture slated to follow Rubin Ultra (Source: developer.nvidia.com).

Kyber sits inside a broader family that NVIDIA calls the third-generation MGX rack-scale architecture, which also includes the Vera Rubin NVL72 compute rack, the Groq 3 LPX inference rack, the Vera CPU rack for reinforcement learning, and the BlueField-4 STX storage rack, all sharing common mechanical, power, and cooling envelopes under what NVIDIA terms the Vera Rubin POD. Kyber is designed to extend that same ecosystem, built by more than 80 MGX partners, into the higher-density Rubin Ultra generation (Source: nvidianews.nvidia.com).

Inside the Kyber Rack: Architecture and Engineering

Kyber's defining engineering decision is orientation. NVIDIA racks from GB200 NVL72 onward have mounted compute trays horizontally, sliding in from the front like drawers in a filing cabinet. Kyber inverts this: compute blades are mounted vertically, and NVIDIA has described the effect as "rotating compute blades vertically, like books on a shelf," which the company says "enables up to 18 compute blades per chassis" while purpose-built NVLink switch blades sit at the rear, connected through a cable-free midplane (Source: blogs.nvidia.com).

Independent hardware coverage from the GTC 2025 unveiling described the same shift in more mechanical detail: "NVIDIA plans to ditch the cable cartridge and switch to a midplane design," moving fans and power supplies out of the rack entirely to raise compute density (Source: www.servethehome.com) (Source: www.servethehome.com). The 2025-era prototype midplane observed at that show had 18 columns and four rows of connectors joining the compute and NVLink switch blades (Source: www.servethehome.com), and independent analyst Glenn Klockwood, who has tracked NVIDIA's rack designs across multiple GTC events, estimated that this passive copper midplane "eliminates two miles of copper cabling" per rack compared to the cable-cartridge approach used in Oberon, and that a single Kyber rack carries "over 87,000 NVLink pins" across its midplane connectors (Source: glennklockwood.com) (Source: glennklockwood.com).

The GTC 2026 (2026) revision of Kyber shown by Jensen Huang described a three-layer physical structure: a front board carrying four Rubin Ultra GPUs and two Vera CPUs, a midplane behind it carrying power and data connections, and an NVLink backplane at the rear carrying the network switches that tie all boards together (Source: www.heise.de). Klockwood's contemporaneous notes on that same event describe a rack of "2x compute chassis per rack" and "18 compute sleds per chassis," for 36 sleds total, each sled carrying 4 solid-state drives (SSDs), 2 to 4 network interface cards (NICs), 4 GPU sockets, and 2 CPU sockets, in a ratio of 2 Vera CPUs to 4 Rubin Ultra GPUs per sled (Source: glennklockwood.com).

Notably, the density of this GTC 2026 display appears to be roughly half that of the original GTC 2025 prototype, which used four compute chassis of 18 blades each for 72 compute blades per rack and was described as a 600 kilowatt (kW) rack (Source: glennklockwood.com). Klockwood explicitly flags this as an open discrepancy: "the design at GTC26 appeared to have lost half its GPU density," and "it remains unclear if this new Kyber is now a 300 kW rack" (Source: glennklockwood.com) (Source: glennklockwood.com). This report treats the 600 kW figure as the vendor-stated design target for a single Kyber compute rack and flags the lower, unconfirmed figure as a possible interim demo configuration rather than a finalized specification, consistent with the general pattern of Kyber's specifications shifting between successive GTC keynotes.

NVIDIA Kyber vs Oberon: What Changed Between Rack Generations

"Oberon" is the working name for the rack architecture underneath both the GB200 NVL72 and GB300 NVL72 systems, the two rack-scale platforms NVIDIA has shipped in volume since 2024 (Source: www.nextplatform.com) (Source: glennklockwood.com). NVIDIA itself describes Kyber directly as Oberon's successor: "the move to rack server generation NVIDIA Kyber, the successor to NVIDIA Oberon, which will house a high-density platform of 576 NVIDIA Rubin Ultra GPUs by 2027" (Source: blogs.nvidia.com), a framing independently repeated by trade press: "Kyber, the successor to Nvidia Oberon, will house 576 Rubin Ultra GPUs by 2027" (Source: www.datacenterdynamics.com).

Four structural changes separate the two generations. **Orientation and cabling**: Oberon racks such as GB200 NVL72 use horizontal server sleds connected through rear-mounted cable cartridges, while Kyber replaces those cables with a fixed midplane and vertical blades (Source: www.servethehome.com). **NVLink domain size**: Oberon racks max out at a 72-GPU NVLink domain (Source: www.nvidia.com), whereas a single Kyber rack targets 144 GPU packages, and multi-rack Kyber configurations scale to 576 or 1,152 GPUs in one domain (Source: developer.nvidia.com) (Source: developer.nvidia.com). **Power delivery**: Oberon-generation racks distribute 54 VDC power inside the rack via copper busbars fed by up to eight power shelves (Source: developer.nvidia.com), whereas Kyber is engineered from the ground up for 800 VDC delivery, discussed in detail below. **GPU silicon**: Oberon racks house Blackwell and Blackwell Ultra GPUs (B200 and B300), while Kyber is purpose-built for Rubin Ultra (Source: www.nextplatform.com).

The two rack families are not backward-compatible. Independent analysis of NVIDIA's roadmap notes plainly that "the Kyber rack architecture is incompatible with traditional Blackwell NVL72 infrastructure and requires new facility power and cooling design" (Source: vrlatech.com). This is precisely why the reported manufacturing delay, discussed later in this report, matters beyond a single generation: operators who deployed Oberon-class GB200 or GB300 racks cannot simply upgrade in place to Kyber, they must plan an entirely new electrical and mechanical footprint.

Rubin Ultra: The GPU Silicon That Kyber Was Built to House

Rubin Ultra is the GPU accelerator Kyber exists to hold. It follows the same naming pattern NVIDIA used when it introduced Blackwell Ultra as a mid-cycle upgrade to Blackwell: an existing architecture pushed further before the next full architectural generation. At GTC 2026, NVIDIA doubled the silicon budget relative to standard Rubin, with a Rubin Ultra GPU package consisting of "four compute dies instead of two, each measuring 800 mm²,"

roughly doubling available compute (Source: www.heise.de). Memory scales alongside compute: each Rubin Ultra package is paired with 16 HBM4e stacks "which together have a capacity of one terabyte" (Source: www.heise.de).

By comparison, a standard Rubin GPU, the accelerator shipping now inside Vera Rubin NVL72, delivers 50 petaflops of NVFP4 inference compute per GPU with 3.6 TB/s of NVLink bandwidth (Source: nvidianews.nvidia.com). Rubin Ultra is reported to push single-package compute to roughly 100 petaflops of FP4, doubling standard Rubin's per-package figure (Source: www.heise.de), a figure corroborated by independent roadmap tracking that lists the Rubin Ultra package at approximately 100 PFLOPS (petaflops) of NVFP4, 1TB of HBM4e, and roughly 3.6 kW of thermal design power per package (Source: vrlatech.com). At rack scale, a complete 144-package Kyber system is projected to reach 15 FP4 exaflops of inference throughput, or 5 FP8 (8-bit floating point) exaflops for training workloads (Source: www.heise.de).

Rubin Ultra also brings a new NVLink generation. NVLink 7 is reported to deliver roughly a "6x improvement over NVLink 6" per GPU (Source: vrlatech.com), the generation used in standard Rubin, which itself already delivers 260 TB/s of aggregate bandwidth across a 72-GPU rack (Source: nvidianews.nvidia.com). Between racks, TechInsights independently describes the Vera Rubin Ultra NVL576 configuration as "an extreme-scale system with 4.6PB/s HBM4e bandwidth and a 115.2TB/s CX9 interconnect, delivering a 14x the performance of the GB300 NVL72" (Source: library.techinsights.com) (Source: library.techinsights.com).

The Vera Rubin Platform and the Road to Kyber

Rubin Ultra and Kyber are the top rung of a larger platform NVIDIA calls Vera Rubin, launched at CES in January 2026 as a family of six co-designed chips: the Vera CPU, the Rubin GPU, the NVLink 6 switch, the ConnectX-9 SuperNIC, the BlueField-4 DPU (data processing unit), and the Spectrum-6 Ethernet switch (Source: nvidianews.nvidia.com). NVIDIA claims this codesign delivers "up to 10x reduction in inference token cost" and a 4x reduction in the number of GPUs needed to train mixture-of-experts (MoE) models compared with the prior Blackwell platform (Source: nvidianews.nvidia.com). The Vera CPU itself is built with "88 NVIDIA custom Olympus cores" on the Armv9.2 instruction set (Source: nvidianews.nvidia.com).

The platform's rack-scale expression today is Vera Rubin NVL72, which NVIDIA describes as delivering "up to 4x better training performance and up to 10x better inference performance per watt, and one-tenth the token cost relative to NVIDIA Blackwell" (Source: developer.nvidia.com). The rack carries "nearly two times more transistors than NVIDIA GB200 NVL72" (Source: developer.nvidia.com), and independent roadmap tracking confirms Vera Rubin "entered full production at GTC Taipei on June 1, 2026," with partner availability following in the second half of the year (Source: vrlatech.com). Independent benchmark results cited by NVIDIA from third-party firm SemiAnalysis's InferenceMax suite show that "NVIDIA rack-scale systems deliver 50x better performance per watt and 35x lower cost per token" comparing GB300 NVL72 against the older H200 GPU generation (Source: developer.nvidia.com).

Positioning the generations against each other: GB300 NVL72, the current Oberon-class flagship, integrates 72 Blackwell Ultra GPUs and 36 Grace CPUs with 130 TB/s of NVLink bandwidth, delivering "1.5x more dense FP4 Tensor Core FLOPS and 2x higher attention performance" over standard Blackwell and up to a 50x overall increase in AI factory output versus Hopper-based platforms (Source: www.nvidia.com) (Source: www.nvidia.com) (Source: www.nvidia.com). Vera Rubin NVL72 supersedes it with 3.6 exaflops of NVFP4 inference per rack (Source: redmondmag.com). Kyber, running Rubin Ultra, is intended to supersede Vera Rubin NVL72 in turn, and NVIDIA's platform architecture explicitly nests all three tiers, NVL72, NVL144, and NVL576, as configuration options within the Vera Rubin Ultra generation, built around what independent GTC coverage describes as "144 Rubin Ultra GPUs with 72 self-designed Vera processors" per Kyber rack (Source: www.heise.de). At full POD scale, NVIDIA's current Vera Rubin deployment integrates "40 racks, 1.2 quadrillion transistors, nearly 20,000 NVIDIA dies, 1,152 NVIDIA Rubin GPUs, 60 exaflops, and 10 PB/s total scale-up bandwidth" across five specialized rack types working as one system (Source: developer.nvidia.com).

The 800 VDC Power Architecture Behind Kyber

Kyber cannot function on the power delivery scheme used by Oberon-class racks. Today's GB200 and GB300 NVL72 racks distribute 54 VDC inside the rack through bulky copper busbars, and NVIDIA's own engineers note that as racks exceed 200 kW, that approach "begins to hit physical limits": powering a 1 MW rack at 54 VDC would require "up to 200 kg of copper busbar," and scaled to a single 1 gigawatt (GW) data center, "the rack busbars alone... could require up to 200,000 kg of copper" (Source: developer.nvidia.com) (Source: developer.nvidia.com).

The jump in power density driving this constraint is already visible between Blackwell and Blackwell Ultra: NVIDIA reports that individual GPU power consumption "increased by 75%" between the two generations, while the growth of the NVLink domain to a 72-GPU system "resulted in a 3.4x increase in rack power density," in exchange for "a staggering 50x increase in performance" (Source: developer.nvidia.com) (Source: developer.nvidia.com) (Source: developer.nvidia.com). NVIDIA's response is a shift to 800 VDC distribution, converting medium-voltage AC directly to 800 VDC at the facility perimeter and eliminating most intermediate AC/DC conversion stages, a timeline NVIDIA describes as "leading the transition to 800 VDC data center power infrastructure to support 1 MW IT racks and beyond, starting in 2027" (Source: developer.nvidia.com), explicitly tied to Kyber's own rollout.

Inside a Kyber compute node, NVIDIA describes a "64:1 LLC converter" that steps 800 VDC directly down to 12 VDC immediately adjacent to the GPU in a single, space-efficient conversion stage (Source: developer.nvidia.com). At the rack and row level, NVIDIA claims 800 VDC busways transmit "over 150% more power... through the same copper" as a legacy 415 VAC or 480 VAC system (Source: blogs.nvidia.com), and forecasts efficiency and cost benefits including "up to 5% improvement in end-to-end power efficiency," maintenance costs "reduced by up to 70%," and a reduction in total cost of ownership "by up to 30%" (Source: developer.nvidia.com) (Source: developer.nvidia.com) (Source: developer.nvidia.com). NVIDIA states plainly that "full-scale production of 800 VDC data centers will coincide with NVIDIA Kyber rack-scale systems in 2027" (Source: developer.nvidia.com).

None of this happens without an ecosystem. NVIDIA has recruited more than 20 partners spanning power semiconductor makers such as Analog Devices, Infineon, and Texas Instruments, power system component suppliers such as Delta and Vertiv, and facility-level power system providers including ABB, Eaton, Schneider Electric, and Siemens, to build the surrounding 800 VDC supply chain. At the facility level, Foxconn has already begun construction on an 800 VDC-ready site: NVIDIA and trade press describe "a 40-megawatt Taiwan data center, Kaohsiung-1, being built for 800 VDC" (Source: blogs.nvidia.com), one phase of a larger Foxconn-NVIDIA project that Reuters reported is "targeted to have 100 megawatts of power," built in stages starting with 20 MW (Source: www.reuters.com) (Source: www.reuters.com).

NVIDIA's Rack-Scale Roadmap Through 2027 and Beyond

NVIDIA's publicly stated cadence runs Blackwell (2024) to Blackwell Ultra (2025) to Vera Rubin (H2 2026) to Vera Rubin Ultra with Kyber (H2 2027) to Feynman (2028) (Source: vrlatech.com). NVIDIA has reaffirmed Rubin Ultra's H2 2027 window as recently as GTC 2026 in March and Computex 2026, and Kyber has always been the rack architecture attached to that date.

That schedule is now in dispute. On July 5 and 6, 2026, CNBC reported that research firm SemiAnalysis found "Nvidia's next marquee product, the Kyber rack-scale architecture designed to house its 2027 Rubin Ultra chips, has been delayed by more than 12 months to 2028." The bottleneck is a component-level manufacturing problem: "Kyber NVL144 rack architecture has been delayed to 2028 as the PCB midplane remains challenging from a manufacturability standpoint," referring to the specialized multi-layer circuit board that replaces cable cartridges in the design (Source: www.cnbc.com) (Source: www.cnbc.com). SemiAnalysis reportedly found that the larger NVL576 configuration linking eight racks via optical connections "is also likely delayed or limited to small volumes" (Source: www.cnbc.com), and that an interim fallback plan, bolting two current-generation racks together for similar scale-up capacity, "has since been cancelled due to heavy pushback from CSPs [cloud service providers] and hyperscalers over its odd design and heavy operational burden" (Source: www.cnbc.com).

NVIDIA disputes the report directly, stating "Our roadmap is intact" (Source: www.cnbc.com). Separately from the Kyber dispute, CNBC reports that NVIDIA's current-generation Rubin systems "are in full production and begin shipping this fall to eight cloud partners, including Amazon Web Services, Microsoft Azure and Google Cloud," and that SemiAnalysis itself "projects Nvidia's data-center compute revenue will run 20% above Wall Street consensus in the second half of fiscal 2027" (Source: www.cnbc.com) (Source: www.cnbc.com), a detail that complicates a simple "NVIDIA is behind schedule" reading: the base Rubin platform appears on track even if the more ambitious Kyber rack has slipped. Beyond Kyber, NVIDIA has already disclosed the outline of its next architecture, Feynman, which Kyber's NVL1152 configuration is explicitly designed to support, reported for a 2028 launch window.

Data Analysis and Evidence

Comparing NVIDIA's four most recent rack-scale generations side by side clarifies how much ground Kyber is meant to cover relative to Oberon. *Table 1* below summarizes the publicly disclosed specifications for each generation, drawing on NVIDIA's own product pages, its technical blog, and independent semiconductor-industry tracking.

Table 1: NVIDIA Rack-Scale Generations at a Glance

RACK (FAMILY)	GPU GENERATION	GPUS PER NVLINK DOMAIN	RACK-LEVEL NVLINK BANDWIDTH	RACK GPU MEMORY	STATUS AS OF JULY 2026
Oberon: GB200 NVL72	Blackwell (B200)	72	130 TB/s (Source: www.nvidia.com)	13.4 TB HBM3E, up to 576 TB/s bandwidth (NVIDIA GB200 NVL72 datasheet)	Shipped in 2024, first Oberon-class rack (Source: developer.nvidia.com)
Oberon: GB300 NVL72	Blackwell Ultra (B300)	72	130 TB/s (Source: www.nvidia.com)	20 TB fast memory, up to 576 TB/s GPU bandwidth (Source: www.nvidia.com)	Shipped in 2025 (Source: developer.nvidia.com)
Vera Rubin NVL72	Rubin	72	260 TB/s (Source: nvidianews.nvidia.com)	Each GPU offers 3.6 TB/s of NVLink bandwidth (Source: nvidianews.nvidia.com)	Full production since June 2026; partner availability H2 2026 (Source: vrlatech.com)
Kyber NVL144 (single rack)	Rubin Ultra	144 GPU packages (576 compute dies) (Source: developer.nvidia.com)	NVLink 7, roughly 6x NVLink 6 per GPU (Source: vrlatech.com)	Up to 1 TB HBM4e per GPU package (Source: www.heise.de)	Planned H2 2027 per NVIDIA; SemiAnalysis reports delay to 2028 (Source: www.cnbc.com)
Vera Rubin Ultra NVL576 (multi-rack)	Rubin Ultra	576 (8 racks x 72) (Source: www.datacenterdynamics.com)	4.6 PB/s HBM4e bandwidth, 115.2 TB/s CX9 interconnect (Source: library.techinsights.com)	Combines 8 Vera Rubin NVL72-class racks into one domain (Source: developer.nvidia.com)	Reportedly "delayed or limited to small volumes" (Source: www.cnbc.com)

Table 1 shows a consistent doubling pattern in NVLink domain size at each generational step, from 72 GPUs in the Oberon and Vera Rubin NVL72 generations to 144 in a single Kyber rack and 576 to 1,152 across multi-rack Kyber configurations, while HBM4e-driven memory bandwidth climbs roughly an order of magnitude between Vera Rubin NVL72 and the full Vera Rubin Ultra NVL576 configuration. The rightmost column also captures the central open question covered in this report: whether Kyber's headline configurations arrive on NVIDIA's stated H2 2027 date or slip into 2028 as SemiAnalysis has reported.

That delay risk becomes more consequential in light of the competitive field NVIDIA is racing against. *Table 2* situates Kyber alongside the two most frequently cited rack-scale or pod-scale alternatives as of mid-2026: AMD's Helios rack and Google's TPU Ironwood pod.

Table 2: Competitive Rack-Scale and Pod-Scale Landscape, Mid-2026

SYSTEM	ACCELERATOR	CHIPS PER RACK/POD	STATUS AS OF JULY 2026
NVIDIA Kyber NVL144	Rubin Ultra	144 GPU packages (576 compute dies) (Source: www.heise.de)	H2 2027 planned per NVIDIA; SemiAnalysis reports 2028 delay (Source: www.cnbc.com)
AMD Helios (Open Rack Wide v3)	MI400 series: MI450, MI430X, MI455X (Source: www.nextplatform.com)	64, 72, or 128 GPUs per system (Source: www.nextplatform.com)	Engineering samples/low volume H2 2026; mass production ramp by Q2 2027 per SemiAnalysis (Source: www.nextplatform.com); AMD disputes any thermal delay (Source: www.nextplatform.com)
Google TPU Ironwood pod	7th-generation TPU (Ironwood)	Up to 9,216 liquid-cooled chips per pod (Source: blog.google)	Generally available since 2025; ICI network spans nearly 10 MW at full pod scale (Source: blog.google)

Table 2 underscores that NVIDIA is not the only vendor pursuing rack- or pod-scale integration. AMD's Helios platform, built on the Open Rack Wide v3 specification co-developed with Meta, uses its MI450, MI430X, and MI455X accelerators in configurations of 64, 72, or 128 GPUs per system (Source: www.nextplatform.com), and SemiAnalysis has separately reported that the mass-production ramp for AMD's own flagship MI455X UALoE72 rack will not reach volume shipments until Q2 2027 (Source: www.nextplatform.com), a claim AMD's data center solutions general manager Forrest Norrod publicly disputed, stating "we have no significant thermal issue" and that AMD is "highly confident of ramping Helios in high volume in the second half of the year" (Source: www.nextplatform.com) (Source: www.nextplatform.com). Google, meanwhile, has been shipping rack- and pod-scale TPU systems in production for years longer than either NVIDIA or AMD; its Ironwood TPU pod scales to 9,216 chips delivering "a total of 42.5 Exaflops," which Google says is "more than 24x the compute power of the world" record-holding El Capitan supercomputer, each Ironwood chip offering 192 GB of HBM at 7.37 TB/s of bandwidth (Source: blog.google) (Source: blog.google) (Source: blog.google). If Kyber's reported 2028 slip is confirmed, it extends the window in which both AMD and Google can compete for the highest tier of scale-up AI infrastructure deals before NVIDIA's next flagship rack is broadly available.

Case Studies and Real-World Examples

Microsoft Fairwater: Building for Rubin Before Kyber Arrives

Microsoft's Fairwater program illustrates how cloud providers are staging their infrastructure investments around NVIDIA's rack-scale cadence rather than waiting for Kyber specifically. Microsoft's first Fairwater "AI superfactory" datacenter is in Wisconsin, and a second went live in Atlanta in October 2025, sharing the same design and featuring "NVIDIA GB200 NVL72 rack-scale systems that can scale to hundreds of thousands of NVIDIA Blackwell GPUs" (Source: news.microsoft.com). The two sites are linked by a dedicated fiber network: Microsoft has "deployed 120,000 miles of dedicated fiber for the network," a 25% increase in overall fiber mileage in a single year, so that the two Oberon-class sites can train models "together as an AI superfactory" rather than as isolated facilities (Source: news.microsoft.com). Microsoft executive vice president Scott Guthrie has framed the underlying philosophy as being about "building the infrastructure that makes them work together as one system," not simply accumulating GPU count (Source: news.microsoft.com).

Ahead of Rubin's broader rollout, Microsoft confirmed its Azure infrastructure is engineered for the transition: "Azure's AI datacenters are engineered for the future of accelerated computing," said Rani Borkar, president of Azure Hardware Systems and Infrastructure, adding that Fairwater sites in "Wisconsin and Atlanta were designed with the power, cooling and networking capacity" needed for Rubin-class systems without major rebuilds (Source: redmondmag.com) (Source: redmondmag.com), with each Vera Rubin NVL72 rack "providing up to 3.6 exaflops of performance," roughly five times the output of the GB200-based systems already installed there (Source: redmondmag.com). Sustainability was also a design constraint: Fairwater Atlanta's cooling loop's "initial fill is equivalent to what 20 homes consume in a year," reflecting a closed-loop liquid cooling system engineered specifically for GPU-dense racks (Source: news.microsoft.com).

Foxconn Kaohsiung-1: An 800 VDC Data Center Built Ahead of Kyber

Foxconn's Kaohsiung-1 facility in Taiwan is one of the earliest named 800 VDC deployments tied directly to NVIDIA's Kyber-era power architecture. NVIDIA and Foxconn describe it as a "40-megawatt Taiwan data center, Kaohsiung-1, being built for 800 VDC" (Source: blogs.nvidia.com), part of a broader partnership that Foxconn chairman Young Liu described at Computex 2025 as targeting "100 megawatts of power" built out in phases, "start[ing] with 20 megawatts... then add[ing] another 40" (Source: www.reuters.com) (Source: www.reuters.com). At the same event, Jensen Huang

framed the facility as a shared national resource, saying NVIDIA was building "an AI factory right for you (Foxconn) to use, for me to use, and for Taiwan the entire ecosystem to use," noting NVIDIA has 350 partners in Taiwan (Source: www.reuters.com). Foxconn's decision to build 800 VDC infrastructure ahead of Kyber's confirmed shipping date illustrates how facility-level electrical planning for gigawatt-scale AI factories must begin years before the racks that will occupy them are finalized, a lead time that makes the Kyber delay reported by SemiAnalysis operationally significant for site planners, not just for NVIDIA's product marketing.

Rubin-Class Systems for Open Science: Los Alamos, NERSC, and LRZ

Beyond hyperscale cloud deployments, national laboratories are adopting the Vera Rubin platform for open science computing ahead of the Kyber generation. Los Alamos National Laboratory (LANL) "has selected NVIDIA Vera Rubin, Vera CPU and NVIDIA Quantum-X800 InfiniBand for its next-generation Mission, Vision and Veritas systems," to be delivered by Hewlett Packard Enterprise (HPE) using HPE Cray supercomputing technology, with the systems supporting national security, open science, and agentic AI research respectively (Source: nvidianews.nvidia.com). At the Leibniz Supercomputing Centre (LRZ) in Germany, the forthcoming Blue Lion system will be "delivering approximately 30x the computing power of LRZ's current system" when it comes online in 2027, supporting astrophysics, environmental science, and life sciences research (Source: nvidianews.nvidia.com). NVIDIA reports that Vera Rubin's scientific-computing configuration delivers "more than 7 exaflops of AI for science, 5 petaflops of native FP64 support" with up to 144 GPUs, enough to place a single rack on par with systems on the TOP500 list of the world's most powerful supercomputers (Source: nvidianews.nvidia.com), and states that "NVIDIA Vera Rubin NVL4-based systems are expected to be available from global system manufacturers in Q4 this year," referring to Q4 2026 (Source: nvidianews.nvidia.com).

CoreWeave: A Cloud-Native Path to Rubin and Beyond

CoreWeave, a specialized GPU cloud provider, illustrates the customer-facing side of NVIDIA's rack-scale transitions. NVIDIA states that "CoreWeave will integrate NVIDIA Rubin-based systems into its AI cloud platform beginning in the second half of 2026," operated through the company's Mission Control software layer so that customers can move between GPU architectures without re-architecting their own deployments (Source: nvidianews.nvidia.com). CoreWeave's position as one of the earliest Rubin-generation cloud partners, alongside AWS, Google Cloud, Microsoft Azure, and Oracle Cloud Infrastructure (Source: nvidianews.nvidia.com), means it is also among the operators most exposed to any slip in the follow-on Kyber generation, since its business model depends on offering customers the newest available NVIDIA silicon on a predictable timeline.

Implications and Future Directions

If the SemiAnalysis-reported 2028 delay holds, the practical consequence is a longer runway for the current Vera Rubin NVL72 generation, and for Oberon-class GB300 NVL72 systems still being deployed, than NVIDIA's public roadmap implied as recently as GTC 2026. That extended runway cuts two ways. For enterprises and cloud operators, it reduces the risk of stranding capital in soon-to-be-obsolete Oberon infrastructure, since the incompatible, higher-density Kyber rack will not arrive to make that infrastructure look dated as quickly as previously expected. For NVIDIA's competitors, it is, as CNBC's sourcing suggests, "a rare technical opening at the high end of the market" for AMD's Helios platform and for Google's in-house TPU systems, both of which already have shipping or near-term rack- and pod-scale products of their own (Source: www.cnbc.com).

The 800 VDC transition underlying Kyber is itself a multi-year industrial undertaking independent of any single rack's ship date. NVIDIA has already recruited chipmakers, power system integrators, and facility electrical contractors across the value chain, and hyperscale operators such as Foxconn, CoreWeave, Lambda, Nebius, Oracle Cloud Infrastructure, and Together AI are designing new sites around 800-volt distribution regardless of exactly when Kyber racks ship (Source: blogs.nvidia.com). That means the electrical retooling of the AI data center industry will likely proceed on schedule even if the specific rack that was meant to be its first flagship application slips a year.

Longer term, Kyber's NVL1152 configuration is already positioned by NVIDIA as the launch platform for its next architecture, Feynman, expected in 2028 (Source: developer.nvidia.com), meaning any Kyber slip has a knock-on effect on the timeline for the generation after it as well. Given that TechInsights projects data center infrastructure revenue reaching $1 trillion by 2028 on the back of exactly this kind of system-level integration (Source: library.techinsights.com), how quickly NVIDIA resolves the PCB midplane manufacturability problem SemiAnalysis identified will materially affect the pace of gigawatt-scale AI factory construction industry-wide, not merely NVIDIA's own product cycle.

Conclusion

NVIDIA Kyber is best understood as two simultaneous bets: a mechanical bet that vertical, midplane-connected racks can pack far more GPUs into one coherent NVLink domain than the horizontal, cable-based Oberon design ever could, and an electrical bet that the entire AI data center industry must move from 54 VDC to 800 VDC power distribution to support megawatt-class racks at all. Both bets are tied to a single GPU, Rubin Ultra, which

doubles the silicon and memory of standard Rubin and is projected to push a full Kyber rack to 15 FP4 exaflops of inference throughput.

The rack's significance extends beyond its own specification sheet. It is the infrastructure NVIDIA is counting on to extend its rack-scale integration strategy, the strategy TechInsights credits for the company's forecast of $1 trillion in data center infrastructure revenue by 2028, into a new architectural generation. Whether that infrastructure arrives on NVIDIA's stated H2 2027 date or, as SemiAnalysis reports and NVIDIA disputes, slips into 2028, will shape how quickly Microsoft, Foxconn, national laboratories, and cloud providers such as CoreWeave can deploy their next wave of AI factories, and how much competitive room the delay opens for AMD's Helios rack and Google's TPU Ironwood pods in the interim. As of July 2026, the underlying 800 VDC power transition and the Vera Rubin platform beneath Kyber remain on schedule regardless of the outcome; only the flagship Kyber rack itself, and the precise date enterprises can expect to deploy it, remains genuinely contested.

Frequently Asked Questions (FAQs)

What is NVIDIA Kyber? Kyber is NVIDIA's fourth-generation rack-scale server architecture, the successor to the "Oberon" design used in the GB200 NVL72 and GB300 NVL72 systems, built to house Rubin Ultra GPUs in configurations ranging from a single 144-GPU rack up to a 1,152-GPU multi-rack domain (Source: blogs.nvidia.com) (Source: www.datacenterdynamics.com).

When will NVIDIA Kyber be released? NVIDIA's official roadmap places Kyber, alongside Rubin Ultra, in the second half of 2027. Research firm SemiAnalysis reported in July 2026 that the Kyber NVL144 rack has slipped more than a year to 2028 due to a PCB midplane manufacturing problem, a report NVIDIA has publicly disputed (Source: www.cnbc.com) (Source: www.cnbc.com).

What is the difference between NVIDIA Kyber and Oberon? Oberon racks, used for GB200 NVL72 and GB300 NVL72, mount GPU trays horizontally, connect them with cable cartridges, and distribute 54 VDC power internally, topping out at a 72-GPU NVLink domain. Kyber mounts blades vertically, replaces cables with a PCB midplane, runs on 800 VDC power, and scales to 144, 576, or 1,152 GPUs per NVLink domain (Source: www.servethehome.com) (Source: www.heise.de).

What is NVIDIA Rubin Ultra? Rubin Ultra is the GPU accelerator housed inside Kyber racks. It roughly doubles standard Rubin's silicon, using four 800 mm² compute dies and 1 TB of HBM4e memory per package, with a single package projected to deliver approximately 100 petaflops of FP4 compute (Source: www.heise.de) (Source: www.heise.de).

Why does Kyber need 800 VDC power? At megawatt-scale rack densities, traditional 54 VDC in-rack distribution would require excessive copper and multiple inefficient conversion stages; NVIDIA states a 1 MW rack at 54 VDC would need "up to 200 kg of copper busbar" per rack, whereas 800 VDC delivers "over 150% more power... through the same copper," enabling megawatt-class racks without unmanageable copper volumes (Source: developer.nvidia.com) (Source: blogs.nvidia.com).

Is Kyber compatible with GB300 NVL72 infrastructure? No. As detailed above, Kyber requires new facility power and cooling design and is not backward-compatible with Blackwell-generation NVL72 racks such as GB300 NVL72, per independent roadmap analysis of NVIDIA's architecture transition (Source: vrlatech.com).

How does Kyber compare to GB300 NVL72? GB300 NVL72 connects 72 Blackwell Ultra GPUs with 130 TB/s of NVLink bandwidth, as detailed in Table 1 above; a single Kyber rack targets 144 Rubin Ultra GPU packages, and TechInsights projects the full Vera Rubin Ultra NVL576 configuration will deliver roughly 14 times the performance of GB300 NVL72 (Source: library.techinsights.com).

Tags: nvidia kyber, rack-scale computing, rubin ultra, vera rubin platform, 800 vdc power architecture, oberon rack, gb300 nvl72, nvidia gpu roadmap, ai data center infrastructure, nvlink