NVIDIA Groq 3 LPU Explained: Vera Rubin Inference Chip

Published July 15, 2026 39 min read

Executive Summary

NVIDIA's "Groq 3 LPU" is a real, newly shipping product: it is the language processing unit (LPU) that resulted from NVIDIA's non-exclusive technology licensing agreement with AI chip startup Groq, announced on December 24, 2025, and productized at NVIDIA's GTC conference on March 16, 2026, as the seventh chip of the **NVIDIA Vera Rubin** platform (Source: groq.com) (Source: nvidianews.nvidia.com). The deal, reported at roughly $20 billion in cash, is NVIDIA's largest transaction ever and brought Groq founder Jonathan Ross (a co-creator of Google's original Tensor Processing Unit, or TPU) and Groq president Sunny Madra into NVIDIA to lead the effort, while Groq itself continues to operate independently under new CEO Simon Edwards and keeps its GroqCloud service running (Source: cnbc.com) (Source: groq.com).

An LPU (Language Processing Unit) is a class of AI accelerator, pioneered by Groq in 2016, that is purpose-built for inference (running an already-trained model) rather than training (Source: groq.com). Instead of relying on off-chip high-bandwidth memory (HBM) as GPUs do, an LPU keeps model weights in large pools of on-chip static random-access memory (SRAM) and uses a compiler-scheduled, deterministic "assembly line" architecture rather than dynamic hardware scheduling, which Groq says delivers substantially faster and, architecturally, up to 10 times more energy-efficient inference than GPUs (Source: groq.com). The new **NVIDIA Groq 3 LPU** (internally codenamed LP30, since Groq skipped a second generation) is manufactured with Samsung and ships in dedicated **Groq 3 LPX** racks of 256 interconnected LPU chips, each delivering 500 megabytes (MB) of SRAM, 150 terabytes per second (TB/s) of SRAM bandwidth and 2.5 TB/s of scale-up bandwidth, for a rack total of 128 gigabytes (GB) of SRAM, 40 petabytes per second (PB/s) of on-chip memory bandwidth, 640 TB/s of scale-up bandwidth, and 315 petaFLOPS of FP8 inference compute (Source: nvidia.com) (Source: developer.nvidia.com).

Paired with the **NVIDIA Vera Rubin NVL72** rack (72 Rubin GPUs and 36 Vera CPUs), the Groq 3 LPX system delivers up to 35 times higher inference throughput per megawatt for trillion-parameter models and up to 10 times more revenue opportunity for AI factories, according to NVIDIA's own projected benchmarks, which the company frames as opening a " premium tokens" business model where speed itself is monetized (Source: nvidianews.nvidia.com). Jensen Huang stated at GTC that Vera Rubin-with-LPX is in production and will likely ship in the third quarter of 2026, with broad availability from cloud partners including Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure in the second half of 2026 (Source: au.pcmag.com) (Source: nvidianews.nvidia.com).

Compared with GPUs and TPUs, the LPU trades flexibility and memory capacity for latency predictability: its SRAM-first design cannot hold a large model on a single chip (a single LPU holds far less data than a GPU's HBM pool), so it always requires many interconnected chips working as a coordinated "single processor," but in exchange it delivers highly deterministic, low-jitter token generation that GPUs struggle to match at high concurrency (Source: au.pcmag.com) (Source: nvidianews.nvidia.com). NVIDIA's own architects describe the design as complementary rather than competitive with GPUs: Rubin GPUs handle prefill and decode attention, while LPUs accelerate the latency-sensitive feed-forward and mixture-of-experts portions of decode, a split NVIDIA calls attention-FFN disaggregation (AFD) (Source: developer.nvidia.com). Wall Street's initial reaction was mixed to skeptical of the reported $20 billion price given Groq's estimated $90 million to $500 million in annual revenue, even as analysts acknowledged the deal strategically blunted the threat from Cerebras Systems, SambaNova Systems, D-Matrix, and Google's own TPU lineup (Source: barrons.com) (Source: fortune.com). Real-world GroqCloud customers, from sports-data provider Stats Perform (7 to 10 times faster inference) to fintech Stash (roughly 37% lower request latency and about 15 times more agent tasks per unit time) to Formula 1's McLaren racing team, already illustrate the commercial case for LPU-class inference ahead of the NVIDIA-branded rollout (Source: groq.com) (Source: groq.com).

Introduction and Background

The phrase "NVIDIA Groq 3 LPU" describes a genuinely new product category rather than a rebrand: it is the result of NVIDIA's non-exclusive licensing agreement with Groq, the AI inference chip startup founded in 2016 by former Google engineer Jonathan Ross, one of the original architects of Google's TPU program (Source: cnbc.com).Groq's chip class, the **LPU (Language Processing Unit)**, is a term Groq trademarked to describe processors purpose-built for large language model (LLM) inference rather than the general-purpose, training-and-inference workloads that graphics processing units (GPUs) handle (Source: groq.com). For nine years, Groq operated as one of NVIDIA's would-be challengers, selling inference capacity through its own GroqCloud service and competing for the same hyperscale and enterprise inference budgets that NVIDIA GPUs typically served (Source: cnbc.com).

That changed on December 24, 2025 (Christmas Eve), when Groq announced it had "entered into a non-exclusive licensing agreement with Nvidia for Groq's inference technology," under which Ross, Groq president Sunny Madra, and a large portion of Groq's engineering team joined NVIDIA, while GroqCloud and Groq itself continued operating independently under new CEO Simon Edwards (Source: groq.com). CNBC and Reuters both reported the deal's value at roughly $20 billion in cash, structured as a technology license and talent transfer rather than a formal acquisition, a structuring choice several analysts linked to antitrust sensitivity, since Groq had been considered one of NVIDIA's most credible inference-market rivals (Source: cnbc.com) (Source: reuters.com).

The product of that agreement arrived roughly three months later. At NVIDIA's GTC 2026 conference in San Jose on March 16, 2026, CEO Jensen Huang unveiled the **NVIDIA Vera Rubin platform**, NVIDIA's next-generation AI infrastructure architecture, and announced that it now comprised seven chips across five rack-scale systems, up from the six chips originally previewed in January 2026, with the seventh chip being the newly integrated **NVIDIA Groq 3 LPU** (Source: developer.nvidia.com) (Source: nvidianews.nvidia.com). This report addresses what TARGET_QUERY searchers are actually looking for: a clear explanation of what the NVIDIA Groq 3 LPU is, how it fits into the Vera Rubin platform, how it compares with GPUs and TPUs, what its specifications and pricing implications are, and what real deployments already show about LPU-class inference economics, all grounded in verifiable sources published between the December 2025 announcement and July 2026.

What Is a Groq LPU? Definition and Taxonomy

An LPU, or Language Processing Unit, is a category of AI accelerator chip that Groq describes as "a new category of processor," created specifically to run trained AI models (inference) rather than to train them (Source: groq.com). Groq's own architecture page frames the LPU around four design principles: **software-first design**, in which the compiler's architecture was finalized before any chip design work began; a **programmable "assembly line" architecture**, where data "conveyor belts" move instructions and operands between fixed-function execution units with no runtime synchronization required; **deterministic compute and networking**, meaning every operation's timing is knowable at compile time down to the clock cycle; and **on-chip memory**, in which both compute and memory live on the same die rather than being split between a processor and external high-bandwidth memory (HBM) chips (Source: groq.com).

That last design choice is the LPU's defining feature. Groq's SRAM-based on-chip memory delivers memory bandwidth "upwards of 80 terabytes/second," compared with roughly 8 TB/s for the off-chip HBM used on contemporary GPUs at the time of Groq's original LPU whitepaper, a bandwidth gap Groq credits with up to a 10 times inference speed advantage on its own architecture (Source: groq.com). The tradeoff is capacity: SRAM is dramatically faster than HBM but far smaller per chip, so a single Groq LPU can only hold hundreds of megabytes of weights, not the tens or hundreds of gigabytes a GPU's HBM pool can hold. DigitalOcean's technical explainer, published July 6, 2026, summarizes the taxonomy this way: "Groq's LPU trades capacity for predictable latency," while GPUs "trade peak efficiency for flexibility" and fixed-function chips like AWS Inferentia2 and Google's TPU "trade flexibility for lower cost per token on one fixed model" (Source: digitalocean.com).

Groq's original chip, internally called a Tensor Streaming Processor (TSP), was first detailed publicly in 2019, when Groq announced an architecture capable of one quadrillion operations per second (one petaOp/s) on a single chip and up to 250 trillion floating-point operations per second (FLOPS), a claim the company said made it the first in the industry to hit that milestone (Source: nextplatform.com). By late 2020, Groq's first-generation GroqChip carried roughly 220 megabytes of on-chip SRAM (Source: nextplatform.com). Groq founder Jonathan Ross has explained the rationale for skipping general-purpose flexibility: because inference "scales with the number of queries or users" rather than with the number of machine learning researchers running training jobs, Ross argued the harder, larger-volume engineering problem in AI was never training but deploying trained models cheaply and reliably at scale (Source: nextplatform.com). This philosophy, that inference is a distinct engineering problem deserving distinct silicon, is the taxonomic root of every LPU generation that followed, including the chip now sold under the NVIDIA brand.

Within the broader AI accelerator landscape, LPUs sit alongside three other major categories: general-purpose **GPUs** (NVIDIA, AMD), which remain the default for both training and inference because they support hot-swapping models and precision formats without recompilation; **TPUs** (Google) and similar fixed-function accelerators like AWS's Inferentia2, which are compiled once per model and shape and then serve that fixed configuration at low cost per token; and emerging **dataflow chips** such as Tenstorrent's, an earlier-stage category with less mature production software (Source: digitalocean.com). Wafer-scale designs from Cerebras Systems, whose WSE-3 chip packs 4 trillion transistors, 900,000 AI-optimized cores, and 44 gigabytes (GB) of on-chip SRAM onto a single silicon wafer delivering 125 petaFLOPS, represent a related but architecturally distinct approach that also avoids off-chip HBM in favor of massive on-die memory (Source: cerebras.ai).

The NVIDIA-Groq Deal: From Rivals to a $20 Billion Alliance

For much of 2025, NVIDIA CEO Jensen Huang was publicly "sometimes dismissive" of rival inference-chip architectures, according to Reuters Breakingviews, even as NVIDIA quietly pursued a series of "acquihire" deals for AI talent and technology, including a more than $900 million agreement in September 2025 to hire Enfabrica's CEO and license its technology (Source: breakingviews.com) (Source: cnbc.com). According to Groq CEO-turned-NVIDIA chief software architect Jonathan Ross, the seeds of the Groq deal were planted quietly in early 2025, when NVIDIA opened its NVLink interconnect protocol to third-party accelerator makers: "It all started early in 2025 when Nvidia released NVLink and was going to allow partners to connect to it," Ross said at GTC 2026 (Source: eetimes.com). Groq's president Sunny Madra proposed a joint experiment splitting inference workloads between NVIDIA GPUs and Groq LPUs, an idea Ross initially resisted, "not because I didn't think it was a good idea, I just didn't know if it was going to work," before agreeing to let a small engineering team pursue it (Source: eetimes.com).

The proof-of-concept moved fast once approved. "We presented [our demo] to Jensen. Three days later, Jensen called up and said, 'Why don't we work more closely together?' Three weeks later, the deal was done," Ross recounted, adding that he began working at NVIDIA full time on December 25, 2025, the day after the licensing agreement was publicly announced (Source: eetimes.com). CNBC characterized the resulting transaction as Nvidia's largest purchase on record, describing it as an acquisition of "assets from 9-year-old chip startup Groq for about $20 billion," financed through cash NVIDIA held after building up $60.6 billion in cash and short-term investments by the end of October 2025, up from $13.3 billion in early 2023 (Source: cnbc.com). Groq itself described the arrangement more narrowly, as a licensing deal for its inference technology under which "Jonathan Ross, Groq's Founder, Sunny Madra, Groq's President, and other members of the Groq team will join Nvidia to help advance and scale the licensed technology," while Groq "will continue to operate as an independent company" (Source: groq.com).

The disclosed price tag drew scrutiny. Groq had been valued at just $6.9 billion in a $750 million funding round only three months earlier, in September 2025, with investors including BlackRock, Neuberger Berman, Samsung, Cisco, Altimeter, and 1789 Capital (Source: cnbc.com). Reuters Breakingviews calculated the implied $20 billion price as roughly 40 times Groq's revenue, a multiple it called an "awkward point of comparison" against the 29-times valuation Cerebras had targeted in its own aborted initial public offering (Source: breakingviews.com). Groq had reportedly been "targeting revenue of $500 million this year amid booming demand for AI accelerator chips," according to CNBC, and Seaport Research analyst Jay Goldberg, one of Wall Street's few Sell-rated NVIDIA analysts, wrote that "It is not clear what sparked the rapid, significant increase in value to Nvidia's $20 billion price" (Source: cnbc.com) (Source: barrons.com). D.A. Davidson analyst Alex Platt was similarly blunt, writing that "We find it hard to believe there aren't better assets in the same market for Nvidia to take a look at," pointing to the limited memory capacity of Groq's chips as a constraint on their applicability (Source: barrons.com). NVIDIA shares fell 1.6% to $187.45 in the trading session following the news, against a 0.4% decline in the Nasdaq Composite, though Truist Securities analyst William Stein noted the price represented "less than half of Nvidia's net cash and less than its expected free cash flow for the current quarter" (Source: barrons.com) (Source: reuters.com). Stein also argued the deal was strategically defensive: "Because Groq's leadership formerly worked on the TPU project, we believe Groq's LPU architecture is likely similar to that of the TPU, and designed for better latency and energy performance in large scale inference," and that acquiring the technology "could make Nvidia's capabilities more appealing to high volume inference customers" facing competition from Google's own increasingly cost-competitive TPUs (Source: barrons.com).

The deal also reshaped the competitive landscape for other AI chip startups. Cambrian-AI Research founder Karl Freund noted that Microsoft-backed rival D-Matrix, which raised $275 million at a $2 billion valuation the month before the Groq news, was suddenly viewed more favorably: "I'm sure D-Matrix is a pretty happy startup right now," Freund told Fortune, adding "I suspect their next round will be at a much higher valuation" (Source:

fortune.com). Cerebras CEO Andrew Feldman posted that the deal validated inference-specialized silicon as a category, arguing that Nvidia's move reflects "a growing industry reality, the inference market is fragmenting, and a new category has emerged where speed isn't a feature, it's the entire value proposition. A value prop that can only be achieved by a different chip architecture than the GPU" (Source: fortune.com).

Inside NVIDIA Vera Rubin: The Seven-Chip Agentic AI Platform

The NVIDIA Vera Rubin platform is NVIDIA's next-generation AI infrastructure architecture, described by the company as "AI infrastructure for the era of agents," designed as a "multi-rack POD-scale system that brings together five purpose-built rack-scale systems into one massive, coherent AI supercomputer" (Source: nvidia.com). At GTC 2026, NVIDIA announced the platform had grown from six chips (as previewed in a January 5, 2026, technical blog) to seven, with the newly integrated Groq 3 LPU joining the NVIDIA Vera CPU, NVIDIA Rubin GPU, NVLink 6 Switch, ConnectX-9 SuperNIC, BlueField-4 DPU, and Spectrum-6 Ethernet switch (Source: developer.nvidia.com) (Source: nvidianews.nvidia.com). Jensen Huang called the platform "a generational leap, seven breakthrough chips, five racks, one giant supercomputer, built to power every phase of AI," and declared that "the agentic AI inflection point has arrived with Vera Rubin kicking off the greatest infrastructure buildout in history" (Source: nvidianews.nvidia.com).

Table 1 below summarizes the five rack-scale systems that make up the Vera Rubin platform, drawing on NVIDIA's official platform documentation and press materials.

RACK SYSTEM	CORE HARDWARE	PRIMARY FUNCTION
Vera Rubin NVL72	72 Rubin GPUs, 36 Vera CPUs, NVLink 6 switch, ConnectX-9 SuperNICs, BlueField-4 DPUs	General-purpose training and inference workhorse; trains mixture-of-experts (MoE) models with one-fourth the GPUs and delivers AI inference at one-tenth the cost per million tokens versus Blackwell (Source: nvidianews.nvidia.com)
Vera CPU Rack	256 Vera CPUs, Spectrum-X Ethernet	Reinforcement learning and agentic environment simulation; delivers results "twice as efficiently and 50% faster than traditional CPUs" (Source: nvidianews.nvidia.com)
Groq 3 LPX Rack	256 NVIDIA Groq 3 LPUs (LP30)	Low-latency, deterministic decode acceleration for agentic and premium-tier inference; 128 GB SRAM, 40 PB/s bandwidth, 640 TB/s scale-up bandwidth per rack (Source: nvidianews.nvidia.com)
BlueField-4 STX Rack	BlueField-4 DPU (Vera CPU + ConnectX-9 SuperNIC)	AI-native shared storage extending GPU memory for key-value (KV) cache; boosts inference throughput up to 5x versus general-purpose storage (Source: nvidianews.nvidia.com)
Spectrum-6 SPX Ethernet Rack	Spectrum-X Ethernet or Quantum-X800 InfiniBand switches with co-packaged optics	East-west rack-to-rack networking; up to 5x greater optical power efficiency and 10x higher resiliency than pluggable transceivers (Source: nvidianews.nvidia.com)

These five racks combine into what NVIDIA calls a POD-scale system: a single Vera Rubin POD spans 40 racks, roughly 1.2 quadrillion transistors, nearly 20,000 NVIDIA dies, 1,152 Rubin GPUs, 60 exaFLOPS of compute, and 10 PB/s of bandwidth (Source: developer.nvidia.com). The centerpiece Vera Rubin NVL72 rack itself carries 20.7 terabytes (TB) of HBM4 GPU memory at 1,580 TB/s of bandwidth, delivers 3,600 petaFLOPS of NVFP4 inference compute, and connects its 72 Rubin GPUs with a sixth-generation NVLink switch offering 260 TB/s of scale-up bandwidth (Source: nvidia.com). A single Rubin GPU, meanwhile, packs 288 GB of HBM4 memory paired with a 50-petaFLOP NVFP4 Transformer Engine, more than 570 times the SRAM capacity of a single Groq 3 LPU, which underscores why the two chip types are designed to be complementary rather than substitutable (Source: au.pcmag.com) (Source: nvidia.com).

NVIDIA's stated case for the seven-chip approach is anchored in shifting AI usage patterns. Agentic systems, in which AI agents call tools, retrieve data, and communicate with other agents in extended loops, "consume up to 15x more tokens than traditional AI applications," and the Vera Rubin platform is designed to meet that demand across "every phase of AI, from massive-scale pretraining, post-training and test-time scaling to real-time agentic inference" (Source: nvidia.com) (Source: nvidianews.nvidia.com). Frontier AI labs Anthropic, Meta, Mistral AI, and OpenAI were named in NVIDIA's press release as looking to use the platform "to train larger, more capable models and to serve long-context, multimodal systems at lower latency and cost than with prior GPU generations" (Source: nvidianews.nvidia.com).

NVIDIA Groq 3 LPU and LPX Architecture: Specs and Performance

The chip NVIDIA calls the Groq 3 LPU is internally designated LP30 (short for "LP3.0"), reflecting Groq's decision to skip a formal second generation entirely. Groq's product and commercial marketing head Stuart Pitts, now working within NVIDIA's accelerated computing group, explained: "We skipped V2. We had been working with Nvidia for a little bit before the licensing agreement was signed, and then once it was signed, Jensen said, let's go, I'll take V3 please, and I'll take it tomorrow" (Source: eetimes.com). The chip is manufactured by Samsung, and Huang confirmed at GTC that "We're in production with the Groq chip," adding it would "likely ship in Q3" of 2026, consistent with NVIDIA's public statement that Vera Rubin-based products, including LPX, will be available from partners in the second half of 2026 (Source: au.pcmag.com) (Source: nvidianews.nvidia.com).

At the chip level, each NVIDIA Groq 3 LPU carries 500 MB of on-chip SRAM as its primary working memory, not a cache, holding the "active working set, including weights, activations, and KV state," 150 TB/s of SRAM bandwidth, and 2.5 TB/s of scale-up chip-to-chip (C2C) bandwidth delivered through 96 direct links running at 112 gigabits per second (Gbps) each (Source: developer.nvidia.com) (Source: developer.nvidia.com). Internally, the chip organizes compute and communication around fixed 320-byte vectors, using three specialized execution modules: matrix execution modules (MXM) for dense tensor multiply-accumulate operations, vector execution modules (VXM) for pointwise arithmetic and activation functions, and switch execution modules (SXM) for structured data movement such as permutation and transposition (Source: developer.nvidia.com).

Table 2 below compares the rack-level specifications of a Groq 3 LPX rack against the Vera Rubin NVL72 GPU rack, using NVIDIA's own published datasheets.

SPECIFICATION	NVIDIA GROQ 3 LPX RACK	NVIDIA VERA RUBIN NVL72 RACK
Chip count	256 Groq 3 LPUs (LP30) (Source: nvidia.com)	72 Rubin GPUs, 36 Vera CPUs (Source: nvidia.com)
AI inference compute	315 PFLOPS (FP8) (Source: developer.nvidia.com)	3,600 PFLOPS (NVFP4) (Source: nvidia.com)
On-chip / GPU memory	128 GB total SRAM (Source: nvidia.com)	20.7 TB HBM4 (Source: nvidia.com)
Memory bandwidth	40 PB/s on-chip SRAM bandwidth (Source: developer.nvidia.com)	1,580 TB/s HBM4 bandwidth (Source: nvidia.com)
Scale-up bandwidth	640 TB/s rack-scale C2C (Source: developer.nvidia.com)	260 TB/s NVLink 6 switch bandwidth (Source: nvidia.com)
Cooling	Fully liquid cooled, MGX ETL rack (Source: nvidianews.nvidia.com)	Liquid cooled, MGX NVL72 rack (Source: nvidia.com)
Primary role	Latency-sensitive decode (FFN/MoE)	Prefill, decode attention, training

Table 2 illustrates the deliberate asymmetry between the two rack types: the LPX rack trades roughly an order of magnitude less raw compute and memory capacity for far higher relative bandwidth density per unit of stored data, which is precisely the profile needed for fast, deterministic per-token generation rather than bulk parallel throughput. NVIDIA's own analysis frames the two chips as "processors of extreme differences," uniting "extreme FLOPS" (Rubin GPU) with "extreme bandwidth" (Groq 3 LPU) rather than treating LPX as a faster GPU substitute (Source: developer.nvidia.com).

The mechanism NVIDIA uses to combine the two chip types is called **attention-FFN disaggregation (AFD)**. In an LLM's decode phase, where tokens are generated one at a time, NVIDIA splits the workload so Rubin GPUs handle self-attention over the accumulated key-value (KV) cache, a task that benefits from large memory capacity, while Groq LPUs execute the feed-forward network (FFN) and mixture-of-experts (MoE) portions of each layer, exchanging intermediate activations between chips for every generated token (Source: developer.nvidia.com). Nvidia's Ian Buck described the intended outcome plainly: engineers "united two processors of extreme differences. One for high throughput, one for low latency" (Source: eetimes.com). This approach effectively shelved NVIDIA's previously previewed "Rubin CPX" large-context compute engine, a GDDR7-based chip aimed at faster prefill, because "we decided to focus [on decode] to improve the dollars per token, and the token rate," according to Buck, though he added Rubin CPX "is still a good idea; I think we'll revisit it in the Feynman generation," referring to NVIDIA's subsequent architecture (Source: eetimes.com).

The rack-level results NVIDIA advertises are substantial. When paired with Vera Rubin NVL72, the Groq 3 LPX system delivers "up to 35x higher throughput per megawatt (MW) for trillion-parameter models" and "up to 10x more revenue per watt," predicated on the idea that faster, "premium" tokens command higher prices than slower, bulk-tier tokens (Source: nvidia.com) (Source: nvidia.com). SiliconANGLE reported that NVIDIA's Ian Buck framed this in concrete terms: "Though 100 tokens per second might seem reasonable for humans, such speeds would seem glacial for agentic systems," and NVIDIA is aiming to support throughputs "of up to 1,500 tokens per second for agentic communications" (Source: siliconangle.com). Huang himself pegged the business opportunity at close to $300 billion per gigawatt of deployed compute for NVIDIA customers, driven mainly by the ability to sell higher-value, faster tokens (Source: eetimes.com). Independent analyst estimates on volume are more modest but still large: semiconductor analyst Ming-Chi Kuo projected NVIDIA would ship 4 to 5 million LPUs through 2026 and 2027, per PCMag's reporting (Source: au.pcmag.com).

Huang was explicit that LPX is a complement, not a full replacement, for GPU-based inference: "If most of your workload is high-throughput, I would stick with just 100% Vera Rubin. If a lot of your workload is coding and very high-value engineering token generation, I would add Groq to it. I would add Groq to maybe 25% of my total data center. The rest of my data center is 100% Vera Rubin" (Source: eetimes.com). The Next Platform's analysis put a number on that mix, reporting Huang's estimate that "low latency, premium priced token generation should represent somewhere on the order of 25 percent of the compute in an AI cluster," and cited NVIDIA guidance recommending "one Vera Rubin rack for every one to four Groq LPX racks" as the practical deployment ratio (Source: nextplatform.com) (Source: eetimes.com).

LPU vs. GPU vs. TPU: Comparing AI Accelerator Architectures

Answering "LPU vs TPU vs GPU" requires separating three distinct engineering tradeoffs: flexibility, memory locality, and scheduling philosophy. GPUs, exemplified by NVIDIA's own Rubin and Hopper-generation chips as well as AMD's MI300X and MI325X, are general-purpose parallel processors that can hot-swap models, precision formats (FP16, INT8, FP8), and batch sizes at runtime without recompilation, which is why they remain the default choice for most inference workloads despite not being purpose-built for the task (Source: digitalocean.com). TPUs, Google's custom accelerators, and similar fixed-function chips like AWS's Inferentia2, take the opposite approach: models are compiled once for a specific configuration using toolchains like Google's XLA or AWS's Neuron SDK, then served cheaply at scale, but any change to the model, precision, or input shape typically requires a full recompile (Source: digitalocean.com).

Groq's LPU occupies a third position on this spectrum, prioritizing **deterministic, predictable latency** above both flexibility and raw compute density. Because the compiler fixes the entire execution schedule, including inter-chip data movement, before any request arrives, LPU-based systems keep their 50th-percentile (p50, or median) and 99th-percentile (p99, the slowest 1% of requests) latencies close together, a property DigitalOcean's guide calls "important for real-time, single-request workloads" (Source: digitalocean.com). By contrast, GPUs serving mixed, concurrent traffic exhibit "statistical latency," where scheduling decisions happen at runtime and resource contention among users can cause the tail of the latency distribution (p99) to spike well above the median (Source: digitalocean.com).

An architectural comparison from analyst Timothy Prickett Morgan at The Next Platform put the raw compute gap between NVIDIA's incoming R200 (Rubin) GPU and the LP30 (Groq 3) accelerator at roughly 21 times more theoretical performance for the GPU at equivalent (FP8) precision, and up to 42 times if the workload can exploit the GPU's FP4 numeric format, underscoring that the LPU's advantage is architectural and latency-based rather than raw-throughput-based (Source: nextplatform.com). Groq's own documentation similarly frames the LPU's relative advantage in terms of memory bandwidth rather than raw FLOPS, contrasting on-chip SRAM bandwidth "upwards of 80 terabytes/second" with the roughly 8 TB/s of contemporary off-chip GPU HBM at the time that comparison was published (Source: groq.com).

Table 3 below summarizes the practical tradeoffs across the three architectures for teams evaluating LLM inference hardware.

DIMENSION	GPU	TPU (AND SIMILAR FIXED-FUNCTION ASICS)	LPU (GROQ / NVIDIA GROQ 3)
Best for	Training and mixed, flexible inference workloads (Source: digitalocean.com)	High-volume serving of one fixed, stable model (Source: digitalocean.com)	Low-latency, single-request, high-concurrency decode (Source: digitalocean.com)
Memory design	Off-chip HBM (large capacity, e.g. 288 GB per Rubin GPU) (Source: nvidia.com)	Off-chip HBM/DRAM, compiled per shape	On-chip SRAM (small capacity, e.g. 500 MB per LPU) (Source: nvidia.com)
Scheduling	Dynamic, runtime hardware scheduling	Compiled per model/shape (XLA, Neuron) (Source: digitalocean.com)	Fully static, compiler-scheduled at compile time (Source: nextplatform.com)
Model swap cost	Seconds (hot-swap weights) (Source: digitalocean.com)	Minutes to hours (recompile) (Source: digitalocean.com)	Requires model sharding across many chips
Latency profile	Statistical (variable under load) (Source: digitalocean.com)	Statistical, generally predictable per fixed shape	Deterministic (p50 and p99 close) (Source: digitalocean.com)

Barron's has flagged the historical lineage connecting the LPU and TPU categories: because Groq's founding team, led by Jonathan Ross, previously worked on Google's TPU project, "we believe Groq's LPU architecture is likely similar to that of the TPU, and designed for better latency and energy performance in large scale inference," a link that also helps explain why NVIDIA viewed acquiring Groq's technology as a defensive move against Google's chip ambitions (Source: barrons.com). Reuters Breakingviews similarly noted that Google's TPUs "are now nearly half as expensive for certain tasks as Nvidia's," according to Bank of America analysts, a cost pressure that shaped NVIDIA's calculus in licensing Groq's architecture (Source: breakingviews.com).

Implementation Considerations: When to Deploy LPUs, GPUs, or Hybrid Inference

For organizations evaluating dedicated inference hardware, the practical decision rarely reduces to a single "best chip." Several factors should guide the choice:

- **Workload latency sensitivity**: Real-time, single-stream applications (voice assistants, coding copilots, agentic multi-step reasoning) benefit most from LPU-class determinism, where p50 and p99 latency stay close together even under variable load (Source: digitalocean.com).
- **Model size and change frequency**: Teams that ship new models weekly or need flexible precision and batching should default to GPU serving, since specialized chips "make you pay again for each of those changes" through recompilation (Source: digitalocean.com).
- **Throughput versus interactivity mix**: NVIDIA's own guidance suggests reserving roughly 25% of a data center's compute for LPU-accelerated, high-value, low-latency workloads such as coding and agentic token generation, with the remainder on GPU-only infrastructure for bulk or offline processing (Source: eetimes.com).
- **Deployment ratio**: If adopting the Vera Rubin plus LPX approach, NVIDIA recommends provisioning "one Vera Rubin rack for every one to four Groq LPX racks," depending on workload composition (Source: eetimes.com).
- **Data sovereignty and compliance**: Regulated industries such as banking, government, and healthcare should evaluate managed inference providers offering in-region hosting, as illustrated by Unifonic's partnership with Groq and Saudi Arabia's HUMAIN for compliant, in-kingdom deployment (Source: groq.com).
- **Software stack maturity**: Nvidia's Ian Buck confirmed the company licensed all of Groq's compiler and disaggregation software, noting "It's not about who has the faster chip, it's who has the integration and the software to actually execute and run," a reminder that raw chip specifications alone rarely determine real-world serving performance (Source: eetimes.com).
- **Access model**: Early access to the NVIDIA-integrated LPX programming environment is limited. Buck stated that the initial rollout would work "with our biggest customers" before broader availability, similar to how Groq itself historically served tokens primarily through a managed API rather than exposing low-level hardware access (Source: eetimes.com).

Teams that want LPU-class inference today, ahead of the Vera Rubin LPX rollout expected in the second half of 2026, can access Groq's existing LPU-based GroqCloud service directly, which as of July 2026 prices models on a linear, per-token basis with no idle infrastructure charges, for example $0.075 per million input tokens and $0.30 per million output tokens for the GPT OSS 20B model running at up to 1,000 tokens per second, or $0.05 input and $0.08 output per million tokens for Llama 3.1 8B Instant running at 840 tokens per second (Source: groq.com) (Source: groq.com). Groq describes its pricing philosophy as intentionally predictable: "Other inference providers spike costs without warning. Some hide behind elastic pricing. Groq pricing is linear and predictable, with no hidden costs or idle infrastructure" (Source: groq.com).

Data Analysis and Evidence

Several categories of quantitative evidence anchor the claims made about NVIDIA's Groq 3 LPU and the broader inference-chip market as of mid-2026. On the deal economics, CNBC and Reuters both reported the license-and-hire transaction at approximately $20 billion, against Groq's most recent private valuation of $6.9 billion set in a $750 million funding round roughly three months prior (Source: cnbc.com) (Source: reuters.com). Reuters Breakingviews computed that price as roughly 40 times Groq's revenue and noted the deal was worth "only a third of Nvidia's cash and short-term investments," a scale that helps contextualize why NVIDIA's stock declined only modestly (1.6%) on the news rather than sharply (Source: breakingviews.com (Source: barrons.com).

On hardware specifications, the comparative gap between the Groq 3 LPU and NVIDIA's own Rubin GPU is stark on memory capacity but favors the LPU on bandwidth-per-byte: a single Groq 3 LPU holds 500 MB of SRAM against 288 GB of HBM4 on a single Rubin GPU, a 576-times capacity difference, while a full Groq 3 LPX rack delivers 40 PB/s of on-chip bandwidth compared with 1,580 TB/s (1.58 PB/s) for a Vera Rubin NVL72 rack's HBM4 pool, meaning the LPX rack moves data roughly 25 times faster per rack despite storing far less of it (Source: au.pcmag.com) (Source: developer.nvidia.com) (Source: nvidia.com).

On projected system-level performance, NVIDIA's own marketing materials, explicitly labeled "projected performance subject to change," claim up to 35 times higher throughput per megawatt and up to 10 times more revenue opportunity when LPX is paired with Vera Rubin NVL72 for trillion-parameter models with large key-value (KV) cache contexts (Source: nvidia.com). Independent analyst Timothy Prickett Morgan offers a more conservative, precision-normalized figure of 21 times higher theoretical FP8 compute for the Rubin GPU alone relative to the LP30 chip, rising to 42 times if FP4 precision applies, a reminder that NVIDIA's headline multiples describe a heterogeneous system's economics rather than a like-for-like chip comparison (Source: nextplatform.com). NVIDIA also states that Vera Rubin NVL72 alone delivers training with one-fourth the GPU count and inference at one-tenth the cost per million tokens compared with the prior Blackwell generation, and separately reports that in an earlier 1-gigawatt Hopper-to-Rubin comparison, halving the GPU count still yielded 13.3 times more AI processing performance, with roughly half of that gain attributable to a shift from FP8 to FP4 numeric precision (Source: nvidianews.nvidia.com) (Source: nextplatform.com).

On market context, NVIDIA's data center segment revenue rose to $193.5 billion in fiscal 2026, up from $116.2 billion the prior year, a 66% increase that underscores the scale of the inference and training market NVIDIA is defending with the Groq licensing deal (Source: siliconangle.com). Hyperscale cloud providers, including Amazon Web Services, Google, Microsoft, and Meta Platforms, were collectively set to spend $650 billion on data center buildouts in 2026, a figure that frames the addressable market NVIDIA and its LPU-equipped Vera Rubin platform are targeting (Source: siliconangle.com). Within the competing inference-chip startup landscape, D-Matrix raised $275 million at a $2 billion valuation and Fireworks raised $250 million at a $4 billion valuation, both in the months surrounding the Groq deal, illustrating how quickly capital flowed toward inference-specialized chip and software startups once NVIDIA validated the category (Source: fortune.com) (Source: fortune.com). Cerebras Systems' own most recent flagship chip, the WSE-3, offers a useful reference point for a different SRAM-centric architecture: 4 trillion transistors, 900,000 AI-optimized cores, 44 GB of on-chip SRAM, and 125 petaFLOPS of peak AI performance built on a 5-nanometer TSMC process (Source: cerebras.ai).

On real-world adoption, GroqCloud pricing pages document sub-second inference speeds across its current model lineup, ranging from 394 tokens per second for Llama 3.3 70B Versatile to 1,000 tokens per second for GPT OSS 20B, at prices as low as $0.05 per million input tokens (Source: groq.com). One customer testimonial published on Groq's own site describes a 7.41 times increase in chat speed and an 89% reduction in costs after switching to GroqCloud, prompting the customer to triple its token consumption (Source: groq.com). These figures, while vendor-reported rather than independently audited, are broadly consistent with the third-party customer case evidence detailed in the next section.

Case Studies and Real-World Examples

Stats Perform: Sports Data at Real-Time Speed

Stats Perform, the sports-data authority behind the Opta brand, manages 7.2 petabytes of proprietary sports data and AI models embedded in more than 200 software modules (Source: groq.com). Chief Innovation Officer Christian Marko said the company's prior open-source inference providers could not meet real-time delivery requirements for live broadcasts: "For us, one second is too much. I need inference in real-time, and our old

approach simply wasn't going to work" (Source: groq.com). After migrating standard open-source model inference to GroqCloud, Stats Perform measured "the average inference speed with Groq is 7 to 10x faster than anything else we tested," with Marko noting the company now runs "over 60 generative AI initiatives" on the platform, describing that figure as "less than 1% of what's coming" (Source: groq.com) (Source: groq.com).

Unifonic and HUMAIN: Sovereign AI in the Middle East

Unifonic, a Middle East customer engagement platform serving banks, government entities, and healthcare providers, needed real-time AI with in-region hosting to satisfy strict data sovereignty rules, a requirement local GPU-only cloud capacity could not reliably meet (Source: groq.com). Partnering with Groq and HUMAIN, the Saudi AI company established under the Public Investment Fund, Unifonic's AI Engineering Director Mohamed Sakher Sawan said the combination let engineering teams "focus on product, not infrastructure," delivering "consistently high performance and the flexibility to adopt new models without building our own hosting layer" (Source: groq.com). The deployment supports Unifonic's use cases across agent co-pilots, agentic task automation, and retrieval-augmented generation for regulated banking and government customers (Source: groq.com).

Stash: Fintech Compliance Under Latency Pressure

Stash, a U.S. financial-technology company serving 1.4 million users through fractional investing and its Stock-Back Card product, built an AI-powered "Money Coach" that ran into a fundamental latency problem on OpenAI's API: "With OpenAI, it was roughly 300 milliseconds just to start a request," said Joel Parrish, Stash's Head of AI, a delay that forced the company to cut agent reasoning steps and run compliance checks in overnight batches rather than in real time (Source: groq.com). After switching its agent network to GroqCloud running OpenAI's open-weight GPT OSS 120B and 20B models, Stash reported it could "do 12 to 15 times the tasks we were doing before" while maintaining full compliance guardrails, and measured "an approximately 37% decrease in average request time and a approximately 10% increase in total requests processed" (Source: groq.com) (Source: groq.com). The company also said its monthly OpenAI inference spend of approximately $40,000 fell meaningfully after the switch, freeing budget to expand additional internal use cases (Source: groq.com).

Frontier AI Labs: Anthropic, OpenAI, Meta, and Mistral Commit to Vera Rubin

Beyond individual application deployments, the launch of the Vera Rubin platform itself drew public endorsements from the leading frontier AI labs whose training and inference budgets shape overall demand for both GPUs and LPUs. Anthropic CEO Dario Amodei said "Enterprises and developers are using Claude for increasingly complex reasoning, agentic workflows and mission-critical decisions. That demands infrastructure that can keep pace," crediting NVIDIA's platform with providing "the compute, networking and system design to keep delivering while advancing the safety and reliability our customers depend on" (Source: nvidianews.nvidia.com). OpenAI CEO Sam Altman said "NVIDIA infrastructure is the foundation that lets us keep pushing the frontier of AI," adding that with Vera Rubin, OpenAI would "run more powerful models and agents at massive scale and deliver faster, more reliable systems to hundreds of millions of people" (Source: nvidianews.nvidia.com). Mistral AI cofounder and chief technology officer Timothée Lacroix highlighted the platform's BlueField-4 STX storage tier specifically, saying it "will enable a critical performance boost needed to exponentially scale our agentic AI efforts" by ensuring "our models can maintain coherence and speed when reasoning across massive datasets" (Source: nvidianews.nvidia.com). Meta was also named alongside these labs as a frontier developer evaluating the platform to serve long-context, multimodal systems at lower latency and cost than prior GPU generations allowed (Source: nvidianews.nvidia.com).

McLaren Formula 1: Real-Time Decision-Making at the Track

Groq's homepage highlights its partnership with the McLaren Formula 1 racing team, noting that "the McLaren F1 Team is fueled by decision-making, analysis, development and real-time insights. So the McLaren F1 Team chose Groq" (Source: groq.com). The partnership illustrates a use case common to motorsport and other latency-critical operational settings, where inference speed translates directly into decision quality within a fixed time budget, the same underlying property (deterministic, low-latency token generation) that NVIDIA is now productizing at data-center scale through the Groq 3 LPU.

Implications and Future Directions

The NVIDIA-Groq alliance signals a broader shift in how the AI infrastructure industry values speed as a distinct, monetizable product attribute rather than a secondary consideration behind raw model capability. NVIDIA's own framing, that faster tokens can be sold at "premium" prices while slower tokens remain free or low-cost, effectively creates a tiered token economy in which chip architecture choice becomes a direct revenue lever rather

than only a cost-efficiency one (Source: nvidia.com). Reuters Breakingviews cautioned that this remains "largely a theoretical opportunity" for now, since Groq itself was targeting only $500 million in 2026 revenue and Cerebras, its closest architectural peer, called off its own IPO after disclosing that most of its sales came from a single customer (Source: breakingviews.com) (Source: breakingviews.com).

Skeptics of specialized inference silicon argue the entire category, LPUs included, may itself be an intermediate step rather than an endpoint. Naveen Rao, former SVP of AI at Databricks and founder of MosaicML, who left to start Unconventional AI with a $475 million seed round, contends that Groq, D-Matrix, and Cerebras "are still optimizing within the same digital computing paradigm," and that a more radical hardware redesign will eventually be needed as AI workloads come to dominate total global compute demand: "We've been building the same fundamental machine for 80 years, a numeric digital machine... But there was never one workload that dominated even more than 2% of all compute cycles," a share he expects AI to grow to roughly 95% within a few years (Source: fortune.com).

For competitors, the deal simultaneously validated the inference-specialized chip category and consolidated its most credible early player under NVIDIA's umbrella, leaving Cerebras, SambaNova (reportedly under a term sheet for acquisition by Intel), D-Matrix, and newer entrants like UK-based Fractile to compete for the remaining independent inference-hardware market, alongside inference-software platforms such as Fireworks and Baseten that gained perceived strategic value from the same announcement (Source: fortune.com). Menlo Ventures partner Matt Murphy struck a more cautious note about the broader inference-startup enthusiasm, observing that "it's hard to tell who's really got something significant versus the tide is [raising] all boats," and predicting further consolidation across the sector (Source: fortune.com).

Looking ahead, EE Times reported that NVIDIA's next-generation Groq chip, referred to as "Groq 4," is expected to add NVLink-C2C connectivity, which today already enables 72 GPUs to share a single coherent NVLink domain and is projected to scale to 576 GPUs with the co-packaged optics planned for NVIDIA's subsequent "Rubin Ultra" generation, suggesting the LPU roadmap will increasingly converge with NVIDIA's own GPU interconnect fabric rather than remaining an isolated architecture (Source: eetimes.com). NVIDIA also indicated it plans to gradually open the LPX programming environment beyond its largest initial customers "to the rest of the world" over time, a step that would let independent developers write lower-level code for the chip in a manner more analogous to CUDA on GPUs (Source: eetimes.com).

Frequently Asked Questions (FAQs)

What is a Groq LPU? A Language Processing Unit (LPU) is an AI accelerator chip category, pioneered by Groq in 2016, purpose-built for inference (running trained AI models) rather than training, using large pools of on-chip SRAM and a compiler-scheduled, deterministic execution model instead of the off-chip HBM and dynamic hardware scheduling used by GPUs (Source: groq.com) (Source: groq.com).

Is the NVIDIA Groq 3 LPU the same as NVIDIA acquiring Groq? No. NVIDIA and Groq describe the arrangement as a non-exclusive technology licensing agreement, reported at roughly $20 billion, under which Groq founder Jonathan Ross and other key staff joined NVIDIA, but Groq continues to operate independently under CEO Simon Edwards and keeps its own GroqCloud service running (Source: groq.com) (Source: cnbc.com).

What is the NVIDIA Vera Rubin platform? Vera Rubin is NVIDIA's next-generation AI data center architecture, comprising seven chip types across five rack-scale systems (Vera Rubin NVL72, Vera CPU, Groq 3 LPX, BlueField-4 STX, and Spectrum-6 SPX) designed to handle every phase of AI, from pretraining to real-time agentic inference (Source: nvidianews.nvidia.com).

When does NVIDIA Vera Rubin, including the Groq 3 LPU, release? NVIDIA stated Vera Rubin-based products, including LPX, will be available from partners starting in the second half of 2026, and Jensen Huang indicated the Groq chip specifically would likely ship in the third quarter of 2026 (Source: nvidianews.nvidia.com) (Source: au.pcmag.com).

How does an LPU differ from a GPU for inference? GPUs use large off-chip HBM memory and flexible, dynamically scheduled execution, making them well suited to varied and rapidly changing workloads, while LPUs keep model data in smaller but far faster on-chip SRAM and use a fully compiler-scheduled, deterministic execution model, trading memory capacity and flexibility for predictable, low-jitter latency (Source: digitalocean.com).

How does an LPU compare to a TPU? Both LPUs and TPUs are custom, non-GPU AI accelerators, and Groq's founding team includes engineers who previously built Google's TPU, but TPUs were originally optimized primarily for large-scale training with a systolic-array architecture, while LPUs were designed specifically for low-latency inference from the outset (Source: barrons.com).

What are the Groq 3 LPU specs? Each chip carries 500 MB of on-chip SRAM, 150 TB/s of SRAM bandwidth, and 2.5 TB/s of scale-up bandwidth; a full Groq 3 LPX rack of 256 chips delivers 128 GB of total SRAM, 40 PB/s of on-chip bandwidth, 640 TB/s of scale-up bandwidth, and 315 petaFLOPS of FP8 inference compute (Source: nvidia.com) (Source: developer.nvidia.com).

Can I use LPU-class inference today, before Vera Rubin LPX ships? Yes. Groq's existing GroqCloud service already offers LPU-based inference commercially, with per-token pricing published on Groq's pricing page and speeds up to 1,000 tokens per second on certain models as of July 2026 (Source: groq.com).

Conclusion

The NVIDIA Groq 3 LPU is not a marketing rebrand but the tangible output of a genuine, closely watched industry event: NVIDIA's roughly $20 billion, non-exclusive licensing deal for Groq's inference technology, announced on December 24, 2025, and productized as the seventh chip of the Vera Rubin platform at GTC 2026 on March 16, 2026. It represents the convergence of two previously competing philosophies of AI compute, NVIDIA's flexible, HBM-backed GPU model and Groq's deterministic, SRAM-backed LPU model, into a single heterogeneous inference architecture built around attention-FFN disaggregation, where Rubin GPUs handle prefill and attention while Groq 3 LPUs accelerate the latency-sensitive decode steps that determine how fast an agentic AI system can think and respond.

The evidence gathered here, drawn from NVIDIA's and Groq's own technical documentation, wire-service and financial press coverage of the deal's economics, independent analyst commentary questioning both the price and the strategy, and named customer deployments already running on Groq's existing LPU technology, supports a consistent picture: LPUs are not a GPU replacement but a complementary tool best suited to roughly a quarter of a modern AI data center's workload, specifically the latency-critical, high-value token generation that agentic AI systems increasingly demand. Buyers evaluating "LPU vs GPU vs TPU" decisions should weigh workload latency sensitivity, model-change frequency, and compliance or sovereignty requirements rather than chasing a single "fastest chip" answer. With Vera Rubin and Groq 3 LPX products expected from cloud partners in the second half of 2026, and with GroqCloud already delivering measurable, third-party-documented speed and cost improvements for customers like Stats Perform, Stash, and Unifonic, the practical case for LPU-class inference no longer rests on vendor promises alone, even as the ultimate scale and profitability of the "premium token" business model NVIDIA is betting on remains, as several analysts have cautioned, still largely unproven.

Tags: nvidia groq 3 lpu, groq lpu, vera rubin platform, nvidia vera rubin, lpu vs gpu vs tpu, ai inference chips, language processing unit, nvidia gtc 2026, groq nvidia deal, ai accelerator chips

---