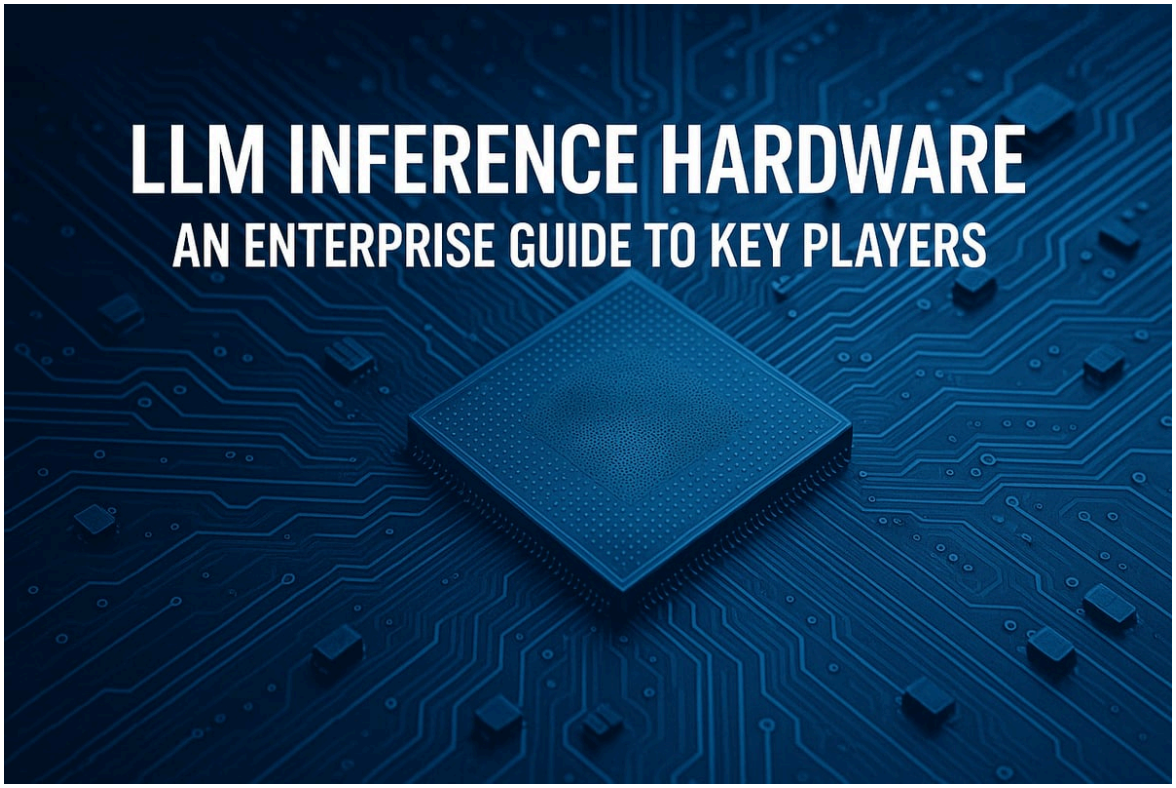
LLM Inference Hardware: An Enterprise Guide to Key Players

By InuitionLabs.ai • 10/21/2025 • 45 min read

llm inference | enterprise ai | ai hardware | on-premise ai | inference accelerators | nvidia gpu | ai chip startups



Private LLM Inference: Key Hardware and Integrators for Enterprise

Executive Summary. The advent of large language models (LLMs) and generative AI has spurred massive demand for specialized inference hardware. Enterprises seeking to run powerful LLMs on-premises (for data control, latency, or cost reasons) turn to vendors offering high-performance servers, accelerators, and integrated systems. NVIDIA's GPUs currently dominate this space (over 90% market share in AI GPUs (www.windowscentral.com)), but a vanguard of competitors is emerging. Startups and established chipmakers alike are developing inference-focused accelerators—**Cerebras, Groq, SambaNova, Graphcore, Tenstorrent, FuriosaAI, Positron, d-Matrix, Untether** and others—to complement or challenge top GPUs. Major system integrators (Dell, HPE, Lenovo, Super Micro, IBM, etc.) package these chips into turnkey AI servers. Global tech giants are also building custom silicon (Broadcom for OpenAI (www.tomshardware.com), Meta acquiring Rivos (www.reuters.com)). This report surveys the *current landscape* of private LLM inference hardware: leading companies, their products, performance and energy metrics, enterprise deployments, and emerging trends (like photonic computing). It details how NVIDIA/AMD/Intel GPUs are being supplemented or optimized, how specialized ASICs are tailored for inference, and how OEMs integrate them into systems. We present data on power, throughput, and market sizes (for example, inference-chip revenue is forecast to reach ~$102 billion by 2027 (www.reuters.com)), along with case studies of real deployments (e.g. LG using FuriosaAI chips (www.techradar.com), SambaNova in government labs (time.com), and major server deals at Dell and HPE (www.reuters.com) (www.reuters.com)). The report concludes with implications: the rising focus on energy efficiency (AI could consume up to 12% of US power by 2028 (www.reuters.com)), data privacy drivers, and future directions (like photonic interconnects (www.reuters.com) (www.reuters.com) and RISC-V-based accelerators (www.reuters.com)).

Introduction and Background

The explosive popularity of generative AI (e.g. ChatGPT) has created a market upheaval in enterprise IT. Companies want to leverage LLMs on their own data, often **on-premises** or close to the edge, for reasons of **data privacy, latency, and predictable cost**. However, running state-of-the-art models (ranging from dozens of billions to hundreds of billions of parameters) requires enormous compute resources. Traditional cloud solutions can be expensive or seen as insecure. As one analysis notes, AI workloads demand massive parallel processing and constant availability, which "do not align well" with conventional cloud pricing and scalability, thus pushing organizations toward **hybrid or on-premises AI infrastructure** (www.techradar.com).

To meet this, the hardware industry is rapidly evolving. Early AI development relied on GPUs (e.g. NVIDIA's CUDA-accelerated cards) originally built for graphics or general compute. Today's generation of GPUs (NVIDIA's Hopper and upcoming Blackwell architectures; AMD's Instinct series; Intel's Arc and forthcoming data-center GPUs) remain a baseline solution for both training and inference. But the shift is clear: inference workloads (serving already-trained models) have different requirements than training. They often involve lower-precision math, need to maximize throughput for many parallel requests, and benefit greatly from specialized optimizations. Recognizing this, **many players are developing inference-optimized chips and systems** (apnews.com).

AS AP News reports, experts already see a pivot: **"the market is now shifting towards AI inference chips, which are more efficient for the everyday use of AI applications"** (apnews.com). Startups and incumbents are targeting inference-specific efficiency gains to enable enterprises to run LLMs without huge power and cooling overhead. For example, Toronto's Untether has launched an inference-focused chip (the "240 Slim") for

edge devices and data-centers (www.reuters.com), while FuriosaAI (Korea) is unveiling an LLM-scale inference server (powered by its RNGD "Renegade" chips) that consumes only ~3 kW vs ~10 kW for a comparable NVIDIA DGX system (www.techradar.com). D-Matrix is introducing an innovative 3D memory design (Pavehawk) specifically to accelerate inference by co-locating compute and memory (www.techradar.com). Even giants like IBM and Intel explicitly emphasize **inference**: IBM's new Power11 system is designed to "help businesses implement AI efficiently to improve operations," rather than raw training power (www.reuters.com), while Intel is unveiling specialized inference GPUs (e.g. Crescent Island with 160 GB memory (www.tomshardware.com)) and software stacks (Battlematrix with LLM-Scaler 1.0) aimed at real-time model serving (www.tomshardware.com).

The imperative for inference efficiency is underlined by energy concerns. A DOE-backed report warns that AI could drive U.S. data centers to consume **~12% of national power by 2028** (www.reuters.com), given that AI workloads (and particularly GPU-accelerated servers) have already **doubled** data-center energy use since 2017(www.reuters.com). This makes specialized hardware not just a performance play, but also a sustainability one. In many cases, enterprises find that instead of huge LLMs requiring massive racks of GPUs, more **compact, domain-specific models on efficient hardware** suffice (so-called Small Language Models, SLMs (www.techradar.com)). A McKinsey survey cited by analysts shows some companies shifting to smaller, task-specific models running on affordable hardware to control costs and data affecting quality (www.techradar.com).

Against this backdrop, the next sections detail the key hardware players and solutions:

- **GPUs and Mainstream Accelerators:** NVIDIA, AMD, Intel, and other major vendors;
- **AI-Dedicated Accelerators:** speciality chips from Graphcore, Cerebras, Groq, SambaNova, Tenstorrent, etc.;
- **Emerging Startups:** new inference chip entrants like FuriosaAI, Positron, Untether, D-Matrix, Lightmatter, etc.;
- **System Integrators:** OEMs and server vendors (Dell, HPE, Lenovo, Supermicro, IBM, etc.) who package these chips into deployable systems;
- **Networking & Storage:** supplementary hardware (e.g. Cisco's AI interconnect) needed for large-scale deployments;
- **Case Studies:** concrete examples of enterprise/private AI deployments;
- **Market and Tech Trends:** data on shipments, market size, investments, and forecasts;
- **Implications and Future Directions:** analysis of energy, geopolitical, and technological implications (e.g. photonic chips, RISC-V adoption).

Throughout, we cite industry reports and news sources. Table summaries compare vendors and products, and we highlight performance, capacity, power, and cost where data are available.

GPU-Based Solutions

NVIDIA's Dominance and Proliferation

NVIDIA's GPUs remain the workhorse for AI inference in enterprise data centers. With a reported **94% market share of AI GPUs (as of Q2 2025)** (www.windowscentral.com), NVIDIA's hardware (Tesla/H100/H200 series, the upcoming Blackwell series, etc.) are ubiquitous. Enterprises often acquire NVIDIA's DGX systems or similar GPU-accelerated servers for LLM workloads. For instance, **Dell Technologies** has announced new PowerEdge servers embedding up to **192 of NVIDIA's next-gen "Blackwell Ultra" GPUs** in air- or liquid-cooled configurations, claiming up to **4× training throughput** improvements over previous generations

(www.reuters.com). NVIDIA also offers integrated solutions like the DGX appliance, which many OEMs resell on their platforms.

Large customers reinforce NVIDIA's position. Bloomberg News reported that Dell is close to a **$5 billion** deal to sell NVIDIA GPU-powered AI servers to Elon Musk's AI startup (xAI) to expand its "Colossus" supercomputer to **over one million GPUs** (www.reuters.com). Similarly, Hewlett Packard Enterprise (HPE) secured a **$1 billion** contract to provide NVIDIA-accelerated servers to Musk's social network X (www.reuters.com). These mega-deals underscore the explosive demand for GPU servers in large-scale AI deployments.

NVIDIA's ecosystem advantages (CUDA software stack, TensorRT, etc.) create high switching costs (www.windowscentral.com). Vendors continue to launch NVIDIA-centric solutions: for example, HPE's recent product lineup includes new ProLiant servers (2U and 4U rack-mount) supporting the latest NVIDIA RTX PRO 6000 "Blackwell" GPUs (www.itpro.com), aimed specifically at generative AI and inference workloads. These servers integrate NVIDIA's full AI software stack with HPE's greenlake/private-cloud offerings to simplify on-prem AI deployments (www.itpro.com).

Performance and Power: NVIDIA's top inference cards (H100/H200) deliver hundreds of teraflops in FP16/INT8 (e.g. ~141 GB HBM3e, tens of TFLOPS of FP16 (www.techradar.com)). However, these units consume kilowatts when scaled; for example, a full NVidia DGX H100 racks draws **>10 kW**. By contrast, some specialized systems undercut this: FuriosaAI's Renegade server with 4 PFLOPS FP8 reportedly consumes only **3 kW**, allowing five such systems per rack versus only one DGX H100 (www.techradar.com). NVIDIA's DGX claims ~180 tokens/sec on LLaMA 3.1 (8B) per DGX node, whereas Positron's Atlas claims 280 tokens/sec on the same model for less than half the power (www.tomshardware.com). These comparisons illustrate why enterprises are examining alternatives to raw GPU scale.

AMD and Other GPU Options

AMD's GPU lineup (Instinct MI series) is a key competitor, especially in CPU-GPU conjunction and in cloud/hyperscale contexts. For example, Pegatron (a Taiwanese system maker) previewed a rack with 128 AMD MI350X GPUs delivering **1,177 PFLOPS (FP4)** power (www.tomshardware.com), and AMD has partnered with Oracle to deploy **50,000** nextgen MI450 GPUs in a new AI supercluster (www.tomshardware.com). While much of this is cloud-oriented, it signals AMD's capabilities: e.g. AMD's MI300 (CDNA 3) offers ~383 TFLOPS FP16 and extremely high memory bandwidth (6.55 TB/s) (www.techradar.com). AMD's chips are optimized for HPC/training (especially double-precision) and have comparable performance density. In inference, AMD supports the MVC stack (ROCm, MIG/Pascal, etc.). Large databases and clusters (such as supercomputer centers) increasingly support both AMD and NVIDIA accelerators. Enterprise system vendors like Dell, HPE, and Lenovo are beginning to offer AMD-based AI servers as well, often in liquid-cooled configurations for stability at scale.

Future direction: AMD is actively pushing into AI with recent hires (e.g. acquisition of Xilinx for adaptive computing) and open efforts: at Computex 2025, AMD unveiled "GAIA," an open-source local LLM framework optimized for AMD's Ryzen AI CPUs (less relevant to data-center hardware but indicative of AMD's multi-pronged strategy). Further, security codeal alliance with Oracle highlights AMD's play: the joint MI450 rack (Helios architecture) is explicitly designed for both training and inference, aiming to position AMD as an alternative to NVIDIA (www.tomshardware.com).

Intel's AI Accelerators

Intel, long a CPU giant, has renewed AI commitments via GPUs and CPUs. Its **Arc Alchemist** GPUs (and upcoming data-center variants) are being featured in HPC workstations and servers. Notably, Intel's *Project Battlematrix* workstations pair multiple Arc Pro GPUs with Xeon CPUs and a Linux-based "LLM Scaler" software stack. Early reports claimed these enhancements can make inference **4.2× faster** on certain LLMs

(www.tomshardware.com). These workstations (up to 8× Arc Pro B60 GPUs, 24 GB each) target on-prem AI developers and content creators, with prices around $5–10k.

More significantly, Intel recently unveiled **"Crescent Island"** – a forthcoming inference-only data-center GPU with the new Xe3P architecture and a massive 160 GB onboard memory (www.tomshardware.com). This design (one 640-bit memory chip or dual 320-bit chips) is clearly aimed at inference: the large memory footprint and focus on power/cost efficiency in air-cooled servers suggest a shot at replacing smaller GPU clusters for large models. Product sampling of Crescent Island is expected by late 2026. If successful, this will put Intel squarely into the inference hardware arms race (Intel position themselves as complementary to NVIDIA rather than a GPU-within-servers competitor).

Intel also provides AI accelerators via its **Gaudi** FPGA-based chips (via acquisition of Habana Labs) and integrates AI via tensor units in upcoming CPUs. But for LLM-scale inference, Arc and Crescent are the main actions. Enterprises may choose Intel-based servers where integration with x86 and certain optimizations (e.g. built-in 5 nm process path via TSMC, Flex chips) are valued. Early customers include ISVs and OEMs building AI workstations – for example, one alliance with PNY and Supermicro is bundling Arc GPUs in high-end workstations and nodes.

AI Accelerator Companies

Beyond general GPUs, a wave of specialized AI chip companies has sprung up, building **new architectures** specifically for inference. These often use novel approaches (e.g. dataflow chips, language-processing units, wafer-scale integration) claiming higher efficiency or performance for LLM serving. Key players include:

- **Graphcore (UK)** – Offers Intelligence Processing Units (IPUs). Their IPU chips (e.g. Colossus GC2 with C2 cards) use massive on-chip SRAM ("scratchpad" memory) and support storing the **entire model in the chip** (en.wikipedia.org), enabling very high parallelism for inference tasks. After years of development, Graphcore's IPUs are deployed by some cloud providers and labs (Graphcore collaborated with Microsoft Azure in 2019 for preview) (en.wikipedia.org). The company raised about $222M in late 2020 and was acquired by SoftBank in 2024 (www.reuters.com), ensuring it can continue competing with Nvidia. Graphcore claims its IPUs outperform GPUs for certain models (especially large, sparse ones). As of mid-2025, Graphcore's IPUs are being offered to enterprises via partners, though specific customer references are limited publicly. (Graphcore's approach is most similar to NVIDIA's original vision of having many model fragments in-chip.)

- **Cerebras Systems (USA)** – Known for the **Wafer-Scale Engine (WSE)**, currently the largest chip ever made (full wafer, ~46,000 mm², ~4 trillion transistors). Their WSE-3 processor (introduced 2024) is said to double the performance of WSE-2 (125 petaflops) and consume roughly the same power (www.reuters.com). The key advantage is that entire large models (even 10B+ parameters) can reside on one chip, removing off-chip communication overhead. Initially targeted at training (OpenAI was a customer), Cerebras has pivoted strongly to inference. In August 2024, Reuters reported Cerebras launched an **AI inference tool** (software) that lets developers run large models on Cerebras chips cost-effectively (www.reuters.com). Cerebras plans to sell inference systems directly to enterprises ("sell AI systems for users operating their own data centers" (www.reuters.com)) as well as via cloud credits. Early enterprise clients include G42 (a UAE AI firm) which has purchased Cerebras supercomputers (www.reuters.com). In Oct 2025, Cerebras announced it will outfit the UAE's Stargate AI data center (in partnership with US-UAE consortium) with its systems (www.reuters.com), highlighting real deployment in governmental contexts. Thus, Cerebras offers one extreme — huge chips in specialized systems — as an alternative to many smaller GPUs.

- **SambaNova Systems (USA)** – Provides a **reconfigurable dataflow** architecture. SambaNova's DataScale platform integrates its RDU (Reconfigurable Dataflow Unit) accelerators and the Samba-1 large language model (an open-weight model on par with GPT-4). Recently reported adoptions include Los Alamos National Lab, SoftBank, and Accenture (time.com). SambaNova's pitch is an end-to-end enterprise AI platform: hardware, software, and models all tuned together. They claim their system excels at inference (and training) of large models, though detailed benchmarks are mostly private. SoftBank (already owning Graphcore) has also invested in SambaNova, illustrating confidence in dataflow chips for real-world AI. The latest Samba-1 inference results (165B-parameter, 44 tokens/sec/sample as of 2024, on an internal cluster of 72 SambaNova blades) suggest impressive throughput, though not directly comparable to public GPUs without similar model contexts. Nonetheless, SambaNova is viewed as a major player for enterprise-focused AI computing.

- **Groq (USA)** – Developed by ex-Google engineers, Groq builds a simplified, deterministic "Language Processing Unit" (LPU) custom ASIC. Its architecture removes many layers (caches, multithreading) to maximize inference throughput and predictability. Groq's claim is extremely high inference speed: per the company, its LPUs deliver inference at ~10× the speed of typical GPUs in some workloads, while using roughly **one-third the power** (www.tomshardware.com). In practice, Groq systems (e.g. their "Chip-on-board" servers with two LPUs each) are reported to offer end-to-end gains for LLMs. For example, Groq's CEO says LPUs can handle LLM inference orders of magnitude faster than GPUs and faster to deploy (www.tomshardware.com). Groq has seen massive funding: it raised $640M in Aug 2024 (valuation $2.8B (www.reuters.com)) and another $750M by Sept 2025 (valuation $6.9B (www.reuters.com)). Major investors include Cisco, Samsung, BlackRock, even a $1.5B commitment from Saudi Arabia expected to yield $500M revenue. These figures show high enterprise anticipation. As of 2025, Groq LPUs are being used by some hyperscalers and corporations trialing inference at scale, though exact installations are confidential. The company is expanding globally (e.g. launching a data center in Europe) to reach enterprise clients (www.tomshardware.com).

- **Tenstorrent (USA/Canada)** – Led by CPU guru Jim Keller, Tenstorrent sells AI processors built on chiplets. Their **Blackwell** (not to be confused with NVIDIA's naming) chipline integrates multiple compute chiplets on one module, balancing fabrication flexibility with scale. The latest "Blackhole" chip is on 6nm (TSMC) and a future "Quasar" is on 4nm (Samsung) (www.tomshardware.com). Tenstorrent's strategy is to offer cost-effective, power-efficient AI processors that can be made at various foundries (TSMC, Samsung, Japan's Rapidus). Their target customers are both large enterprises (for on-prem AI servers) and smaller AI developers needing affordable compute (www.tomshardware.com). Partnerships include Bosch and Hyundai (for automotive AI inference) (www.reuters.com), but in general Tenstorrent is positioning itself as a general-purpose AI chip provider, with emphasis on flexibility and supporting open standards like RISC-V. It does not emphasize one spectacular performance stat; rather it aims to be a credible alternative for data center inference.

- **Untether (Canada)** – Focuses on inference chips for edge and data-center use. Its **240 Slim** chip (announced Oct 2024) is built on RISC-V, and is tailored for efficient execution of fixed models (like chatbots) rather than training. Reuters reports Untether's chips deliver "high performance with reduced energy," suitable for cars, drones, and also inference data centers (www.reuters.com). Mercedes-Benz is already collaborating to use Untether chips in autonomous vehicles, demonstrating edge use; but Untether also believes in AI data center growth (forecasting a trillion-dollar inference market by late decade (www.reuters.com)). Their investors include top-tier VCs and AMD veteran leadership. For enterprises, Untether promises an energy-savvy inference accelerator buildable on a $0.02 per inference token model (targeting cheap at-volume billing models).

- **FuriosaAI (South Korea)** – A startup (LG-backed) that designs dedicated inference chips. In Sept 2025, Furiosa unveiled the **RNGD Server**, powered by its **RNGD "Renegade"** chips (5nm, dual HBM3 memory) (www.techradar.com) (www.techradar.com). Remarkably, each 4-chip RNGD server delivers ~4 PFLOPS FP8 at only **3 kW** power draw (www.techradar.com), enabling 5–right GPUs; plus, multiple servers can be rack-cooled. Performance claims include real-time inference of OpenAI's GPT-OSS 120B model during an OpenAI event (www.techradar.com). LG has already adopted RNGD hardware for its own EXAONE LLM, reportedly achieving **2× inference performance per watt** vs GPU-based setups (www.techradar.com). Furiosa has strong funding (~$125M+ Series C) and international sampling of its servers is underway. Its roadmap includes compiler improvements and new precisions (MIXP4) to further boost efficiency. Furiosa exemplifies the new wave: a company explicitly building **sustainable, rack-optimized** AI servers for enterprises, decoupling from GPU's power limitations (www.techradar.com) (www.techradar.com).

- **Positron AI (USA)** – A very recent entrant (founded 2023) building inference accelerators. Its first product, **Atlas**, packs eight custom "Archer" ASICs into a single system. Positron claims Atlas **beats NVIDIA's DGX H200 in efficiency**: on Llama 3.1 8B, Atlas achieved **280 tokens/sec at 2000W**, whereas a DGX (8× H200) delivered ~180 tokens/sec at 5900W (www.tomshardware.com). If verified, this is a >3× improvement in tokens per watt. Atlas is fabbed in the US (TSMC N4/N5), uses HBM memory, and supports standard AI toolchains. The company has $75M funding and is already testing with customers like Cloudflare (www.tomshardware.com) who seek energy-thrifty inference. A planned next-gen system, **Asimov** (2026), will have 2 TB per chip, 16 Tb/s networking, and target 16-trillion-parameter models. Positron highlights "made-in-USA" aspects (chips assembled locally, advanced packaging in Taiwan) and directly positions itself against NVIDIA's high-end inference platforms.

- **d-Matrix (USA)** – A startup tackling the memory bottleneck in AI. Instead of building a new compute core, d-Matrix's **"Corsair" platform** uses a novel memory hierarchy: 256 GB of LPDDR5 DRAM and 2 GB of SRAM per package, coupled with a 3D-stacked DRAM ("3DIMC") chiplet (www.techradar.com). The result (called Pavehawk) targets inference: they claim it can deliver **10× the bandwidth and 10× the energy efficiency per stack compared to even HBM4** (www.techradar.com). Microsoft and others funded $160M into d-Matrix. Early samples (2024) are tested by partners, and Supermicro will integrate d-Matrix chips into its servers (www.reuters.com). The strategy is that by keeping data "close" to compute, inference (which is often memory-bound) runs far more efficiently. While still in development, d-Matrix's approach suggests new memory architectures will play a role in future LLM servers.

These companies (and several others) represent a *paradigm shift*: from general-purpose GPUs to domain-specific inference hardware. They claim that by tailoring dataflow, precision formats, or memory, they can achieve the same LLM throughput at a fraction of power or cost. For enterprises, this diversity means options: one can choose fast GPUs (NVIDIA/AMD) or experiment with these accelerators when they become available. In practice, many organizations test multiple architectures via cloud trials or small on-prem clusters. Over time, it is expected that inference-specific accelerators will capture a significant share of workloads, possibly eclipsing training-oriented GPUs in volume.

System Integrators and OEM Solutions

While chip designers create the silicon, **system integrators and OEMs** turn these chips into usable products for enterprises. These companies package accelerators into servers, manage cooling, networking, and provide software stacks. Key integrators include:

Vendor	Offerings	Notable Deployments / Clients
Dell Technologies	*PowerEdge* AI servers with NVIDIA GPUs (e.g. Blackwell Ultra); also AMD/HPE partnerships. Demo **DGX-class systems**. New laptops (Pro Max Plus) with on-board NPUs for local LLMs (www.reuters.com).	**xAI project:** ~$5B deal to equip Elon Musk's startup with NVIDIA GB200 GPU servers (www.reuters.com). Dell/HPE containers for Musk's X (Twitter) AI supercomputers.
HPE (Hewlett Packard Enterprise)	*ProLiant* servers (2U/4U) integrating up to 8x NVIDIA RTX6000 GPUs (www.itpro.com), ASIC options. GreenLake private cloud AI offerings. Consulting and turnkey deployments.	**Elon Musk's X:** Confirmed $1B+ sales of HGX servers for XAI (via Bloomberg) (www.reuters.com). New "simple AI" lineup announced in mid-2024 for mainstream businesses (www.reuters.com).
Lenovo	AI rack servers and GPU servers produced in India (50k annually) (www.reuters.com). Offers RDHx-cooled and conventional systems with NVIDIA/AMD GPUs.	Cater to enterprise and hyperscale. No specific publicized AI project beyond manufacturing announcement.
Super Micro Computer	Wide range of custom rack servers with NVIDIA/AMD accelerators. Pioneers direct liquid cooling (DLC) solutions. Quick-turn OEM partner with Nvidia/AMD.	High volume shipments: shipping **100,000 GPUs per quarter** in late 2024 (www.reuters.com). Played a role in supplying server racks for Elon Musk's xAI supercomputer (www.reuters.com). Recently in S&P 500.
IBM	Power11-based AI servers (using new Power11 CPUs and the Spyre AI coprocessor) (www.reuters.com), plus mainframes with Telum. Emphasizes **inference** focus. Secure, high-uptime architectures.	Targets banking, healthcare, government. IBM claims customers needing mission-critical hybrid AI deployments (uptime 99.9999%) (www.techradar.com). Also recognized major cloud vendors use IBM Power (Nvidia on Power).
Cisco Systems	Networking hardware for AI. Launched the **P200** AI data-center interconnect chip (replacing 92 chips, 65% less power) to link dispersed AI DCs (www.reuters.com). Also provides switches (Nexus) with SmartNICs, in partnership with Nvidia, to accelerate AI traffic.	Customers for P200 include Microsoft and Alibaba (www.reuters.com), who need inter-datacenter AI networks. Cisco also integrates Nvidia GPUs in its UCS servers for AI, though less publicity.

Vendor	Offerings	Notable Deployments / Clients
Others (OEM/ODM):	Foxconn, Quanta, Wistron, QCT, etc. build servers for cloud and enterprise using the above components. Chinese ODMs (Inspur, Wingtech) likewise produce AI servers, often with custom accelerators (e.g. Huawei Ascend).	Some defense/government supercomputers (e.g. in China) use Huawei's AI servers with Ascend chips. Major telecoms and financials leverage these integrators.

Modern AI servers are essentially clustered GPU/accelerator nodes, often with specialized cooling (liquid or advanced air). For example, Dell's new systems have options for liquid cooling up to 256 GPUs per rack (www.reuters.com), addressing the power density of LLMs. HPE's designs focus on modular simplicity (e.g. 2U with 2 GPUs or 4U with 8 GPUs (www.itpro.com)). Supermicro's flexibility helped it outpace rivals (as noted when it entered S&P500) by rapidly launching AI servers as soon as GPU generations release.

Beyond hardware specs, integrators offer software and services. NVIDIA's own *NVIDIA AI Enterprise* software suite is often bundled, or similar stacks from AMD and Intel. Consulting (e.g. HPE GreenLake AI Advisory) is common for on-prem AI deployments. Many enterprises contract these vendors for integration, maintenance, and even co-management of on-prem AI platforms.

Example Deployments

- **Enterprise Labs and Research:** Government labs and large research organizations have adopted these systems. For instance, the U.S. Department of Energy's AI research centers use HPE and Dell servers with AMD/NVIDIA accelerators. Los Alamos National Lab is reported to have deployed SambaNova's AI system (with Samba-1 model) to handle classified document analysis and simulation tasks (time.com). In Europe, CERN and others use powerhouses like HPE/Dell with NVIDIA GPUs. These showcase that on-prem inference is critical for sensitive R&D.
- **Corporate Data Centers:** Tech companies building private AI clouds often choose these integrators. The xAI (formerly Twitter AI lab) supercomputer uses Dell/HPE servers with NVIDIA chips to host LLMs (cf. Dell's $5B deal (www.reuters.com) and HPE's $1B X deal (www.reuters.com)). Financial firms (banks, insurance) quietly build similar clusters (e.g. Goldman Sachs and Morgan Stanley have multi-P1000-model GPU dark clusters, though exact hardware is proprietary). Large manufacturers use HPE or IBM servers to run AI for design and voice assistants.
- **On-site AI Appliances:** Some solutions are sold as appliances. For example, C3.ai (an enterprise AI software provider) announced in 2025 an "AI Workstation" appliance with NVIDIA GPUs (4× H100) for running enterprise LLM workloads on-prem (combining LLM inference with data management). Similarly, Cisco has integrated GPUs into its UCS X-series servers marketed for AI workloads. These appliances reflect demand from enterprises wanting box-and-cabinet solutions rather than custom builds.

Technology and Performance Analysis

To compare these options, we consider key metrics:

- **Throughput:** In practice, inference throughput is often measured in *tokens per second* (TPS). For example, on Llama 3.1-8B (a common benchmark), NVIDIA's DGX H200 system is quoted at ~180 TPS. Positron's Atlas claims ~~280 TPS (www.tomshardware.com), while Groq's LPUs claim many times faster on similar tasks (www.tomshardware.com). In tests, Groq's LPUs reportedly serve multiple 5-10× acceleration factors vs GPUs for the same model, due to their streamlined pipelines (www.tomshardware.com). IBM Power11 (with future Spyre AI core) is not publicly benchmarked on LLMs yet, but IBM positions it to yield significant improvement in enterprise inferencing.
- **Latency & Batch Size:** Many accelerators specialize in low-latency (even if it means smaller batch throughput). For enterprise chat or interactive tasks, response time is crucial. Companies like NVIDIA optimize their TensorRT engines; others like Groq reduce cycles by eliminating overhead. A P100-GPU cluster might achieve high TPS with large batches, but specialized chips can serve predictions faster on smaller batches.

- **Power Efficiency:** This is a standout differentiator. For instance, Furiosa's RNGD server (4 PFLOPS) uses only 3 kW (www.techradar.com), whereas an equivalent NVIDIA cluster would need ≈10 kW. Positron's Atlas achieves ~3× the tokens per watt of an NVIDIA system (www.tomshardware.com). Groq's LPUs draw about one-third the power of comparable GPUs (www.tomshardware.com). Even Cisco's P200 network chip saves 65% power by collapsing multiple older chips into one (www.reuters.com). These gains translate directly to operational cost savings in enterprise data centers, which must often provision substantial backup power and cooling.

- **Capacity (Memory):** LLM inference is often memory-constrained (to hold the model weights and activations). Solutions vary: NVIDIA's H200 has 141 GB HBM3e per GPU, and Intel's Crescent Island promises 160 GB. Cerebras WSE-3 integrates *4 chips worth* (~1.2 TB) into one wafer, far beyond any GPU. SambaNova and Graphcore each have large on-chip or attached memory (~>100 GB per chip) to accommodate huge models. Positron's future Asimov promises a profound 2 TB per chip (www.tomshardware.com). Thus, specialized chips lead with enormous memory capacity for single-model execution, whereas typical GPU clusters shard models across nodes. Enterprises must consider whether their workload requires one huge model per server (favoring chips like WSE) or can be split across many GPUs.

- **Scalability and Software Ecosystem:** GPU clusters benefit from mature stacks (NVIDIA Triton, ONNX, TensorRT, etc.). Many specialized chips offer their own SDKs (e.g. Graphcore's Poplar, Groq's compiler, SambaNova's SambaFlow). Compatibility with industry frameworks is a key criterion. For instance, Positron emphasizes compatibility with OpenAI APIs so enterprises can migrate workflows. Groq even boasts of running Facebook's LLaMA model unchanged on their chips (www.reuters.com). Intel uses its xPU software stack for Arc GPUs and is building oneAPI extensions. Ultimately, enterprises often use a mix: GPUs for broad frameworks, and accelerators for cool inference gains when the ecosystem supports their model.

- **Cost:** Public data on enterprise pricing is scarce, but estimates exist. Positron's Atlas, for example, is projected to deliver 3× better cost-per-token than competitor DGX systems (www.tomshardware.com) (though actual purchase price is not published). Cerebras claims inference at "10 cents per million tokens" competitive pricing (www.reuters.com). NVIDIA GPUs are expensive (each H100 retails ~$40,000+), and a single rack fully loaded can cost $1–2M. In contrast, many inference accelerators aim to cut total cost (e.g. Groq's LPUs trade some higher chip cost for much less required hardware overall). The Dell and HPE deals (multi-$billion scale) imply custom discounted pricing for bulk corporate orders. Overall, any hardware selection is a capex-opex tradeoff: e.g. Dell's new 192-GPU rack systems likely cost millions each (www.reuters.com), whereas an alternative like five RNGD servers (replacing one DGX) also tally into the millions but save hundreds of kW of electricity (www.techradar.com).

- **Market Forecasts and Investments:** Investors clearly believe inference is big business. Groq's latest funding values the company at $6.9B (www.reuters.com). Reuters cites a forecast of a **$102 billion market by 2027** for inference chips (www.reuters.com). In part, this is due to the vast installed base of trained LLMs that need serving. The huge sums being invested by companies like Broadcom (for OpenAI) and consortiums like Oracle+AMD indicate that owning efficient inference hardware is a strategic priority for leading AI organizations.

In summary, **performance** (throughput, latency) and **efficiency** (power, cost) are the key axes where these systems are evaluated. In practice, many enterprises blend solutions: for instance, running small-to-medium LLMs on GPU clusters and reserving specialized accelerators for the largest, most latency-sensitive tasks. The next sections illustrate these deployments.

Case Studies and Deployments

FuriosaAI (South Korea) / LG: At an OpenAI Seoul event in Sept 2025, FuriosaAI demonstrated its hardware by running the GPT-OSS 120B model in real time on its RNGD chips (www.techradar.com). The results were striking: a single 4-core RNGD 5nm chip delivered performance comparable to one NVIDIA H100 GPU while using only one-third the power (www.techradar.com). These chips are now used by LG for the EXAONE model, doubling inference performance per watt vs GPUs (www.techradar.com). Furiosa's approach – selling servers that let enterprises run large LLMs without building huge GPU farms – is a direct example of *private LLM inference*. LG's case (a major manufacturer running an in-house LLM on exotic chips) shows this is no longer theoretical; it's going into real products.

SambaNova Systems (USA) – Time Magazine reported that SambaNova's AI platform (chips plus Samba-1 model) has been adopted by institutions like Los Alamos National Lab, SoftBank (Japan), and Accenture (time.com). In one Los Alamos deployment, their system reportedly processes large-scale scientific data with an LLM to assist research, all on-premises. Although exact performance figures are proprietary, SambaNova claims attributes like energy savings over GPUs and better handling of large conversational workloads. The SoftBank and Accent partners highlight interest in using SambaNova's inference hardware for advanced AI deployments in finance and manufacturing as well.

Cerebras Systems (USA) – Beyond news on its chips, multiple real-world deployments illustrate Cerebras's penetration. Abu Dhabi's G42 (a major AI group) purchased Cerebras supercomputers for training and inference of their Giant language models (www.reuters.com). The CTO of G42 cited the ability to train some of the largest models faster using Cerebras than with a comparably sized GPU cluster. Moreover, Cerebras is strategically placing infrastructure globally: its announced expansion into the UAE's mega "Stargate" data center will enable regional enterprises (India/Pakistan/Middle East) to run huge LLMs. This shows how an inference-focused chip vendor partners with national initiatives, moving beyond US tech hubs.

Groq Inc. (USA/Helsinki) – Groq's LPUs have seen adoption by hyperscalers and defense contractors (unnamed). A telling vignette: in Helsinki, Groq opened a data center in partnership with Equinix to serve European AI customers (www.tomshardware.com). The data center is explicitly positioning Groq as an "AI cloud" alternative to NVIDIA for inference. Use cases include real-time analytics for automotive and industrial IoT (where low-latency inference is crucial). Groq has stated that LPUs are being tapped to run complex models on sites that require European data sovereignty. Their story underlines a trend: hardware vendors not only sell chips, they run co-located inference services to prove value to enterprise clients.

Positron AI (USA) – In early 2025, Positron shipped its first Atlas enclosures to select corporate customers. One publicized tester is Cloudflare, exploring Atlas for energy-efficient inference in their edge servers (www.tomshardware.com). Internal benchmarks (by Positron) show that Atlas reduces power by ~67% for an 8B LLM compared to a DGX H200, at similar throughput (www.tomshardware.com). If verified, customers running many repeated inference tasks (e.g. modulated chatbots at telecommunications companies) could cut electricity bills by tens of millions annually. Positron's second-gen "Asimov" is already sampling, aimed at theory-computations requiring multi-trillion-parameter models. The emerging narrative: enterprises seeking to cut costs on large-scale inference are trialing these new accelerators.

Untether (Canada) – While less public, Untether's chips have been sampled by forward-looking automotive and agricultural customers (Mercedes-Benz collaboration) (www.reuters.com). They have positioned the 240 Slim as an inference engine for on-vehicle use (e.g. autonomous driving), which would allow cars to run LLM-based assistants or vision-language models without offloading to cloud. For enterprises outside automotive, Untether's emphasis on low-power chips for inference is promising for distributed LLM workloads (e.g. inference at factories or mining operations). In essence, Untether's success could enable LLM inference on small edge appliances and IoT, complementing data-center solutions.

Major OEM Deals: The broad industry move can be seen in mega-deals. As noted, Dell's multi-billion-server deals with xAI (NVIDIA chips) (www.reuters.com) and HPE's $1B+ deal with X (HPE reports) (www.reuters.com) indicate that enterprise AI hardware purchases have become multi-billion-dollar procurement items. These commitments are essentially private inference projects. For comparison, earlier cloud-centric AI was measured in tens to hundreds of millions, but now hits billions per customer. Even smaller enterprises are dedicating multi-million budgets to on-prem LLM servers.

OpenAI/Broadcom (Custom Chips): In a related context, Tom's Hardware reported in 2024 that OpenAI may be building its own custom inference accelerators through Broadcom (www.tomshardware.com). If true, OpenAI's "Project Arya" chips (in Broadcom's ASICs) will eventually run GPT inference in their own data centers, reducing cloud spend and vendor lock-in. While OpenAI is not an enterprise client, the move exemplifies the trend: any serious AI group (enterprise or research lab) may shift to owning very large-scale inference hardware.

Data and Market Trends

Numerous reports quantify the rapid growth of AI inference hardware:

- **Market Size:** Analysts forecast a **~$100 B** market for inference chips by ~2027 (www.reuters.com), driven by deployment of LLMs across industries. Training was dominated by NVIDIA, but inference spending (in chips and systems) is projected to surpass training in dollar terms because inference happens constantly at scale.
- **GPU Leadership:** NVIDIA's dominance continues: as of Q2 2025 NVIDIA commanded **94% of the AI GPU market** (www.windowscentral.com). Its FY2025 AI-related revenue is expected ~$49 B, a 40% jump year-over-year (www.windowscentral.com). This shows enterprises still heavily invest in NVIDIA-based inference. However, the same report notes competition: Chinese chipmakers (Huawei, Cambricon) and others are reacting to US export limits (www.windowscentral.com).
- **Investment Surge:** VC and corporate funding into AI hardware has soared. For example, Groq's valuation skyrocketed from $1.1B (2021) to $6.9B (2025) after massive fundraising (www.reuters.com) (www.reuters.com). Cerebras has filed for an IPO. Lightmatter (photonic) raised $850M (www.reuters.com). D-Matrix has >$160M funding (www.reuters.com). Even Untether and Positron have raised dozens of millions. This avalanche of investment signals confidence in future enterprise hardware needs.
- **Hardware Shipments:** Super Micro reported **100,000 GPUs shipped per quarter** by late 2024 (www.reuters.com), largely to AI customers. This scale (much higher than a year earlier) reflects annual shipments in the millions. Dell/Trex and others are similarly vehicle-class orders. On the accelerators side, exact unit counts are private, but orders like Broadcom's rumored **1–2 million AI chips** for OpenAI (www.tomshardware.com) show the order-of-magnitude scale when hyperscalers commit.
- **Energy Trends:** As noted, data-center power is skyrocketing. IEA energies intelligence suggests AI's share in world electricity is surging. The key takeaway for enterprises: inference hardware must focus on efficiency gains to avoid unsustainable energy costs. Vendors' claims of "10× efficiency" or "CUDA killers" are grounded in this imperative.
- **Standardization & Ecosystem:** Industry alliances are forming around standards. For instance, the OpenAI "Plug and Play" API doesn't help on-prem use, but federated APIs and interoperability (like ONNX for models, and NVIDIA's Megatron support) are enabling more architectures. AMD is supporting ROCm and TensorFlow, and major startups pledge compatibility with PyTorch/ONNX. Intel and others push oneAPI. The point: enterprises want hardware that works with their existing ML pipelines. This pushes integrated vendors (Dell, HPE) to emphasize compatibility in their stacks.

Future Directions and Implications

Sustainability: The DOE report implies enterprises will demand more "green AI" solutions. Energy efficiency (and on-prem renewables) will influence hardware choices. Even beyond energy, heat and cooling constraints are becoming central; Furiosa's low power design is a strong selling point.

Model and Software Evolution: The hardware landscape could evolve with AI models themselves. If models grow to trillions of parameters, wafer-scale or even multi-wafer solutions (like Cerebras' WSE or next-gen photonic chips) could be needed. Conversely, as the *Smaller AI Models (SLM)* trend suggests (www.techradar.com), not all inference will require the largest chips; many domain-specific models run comfortably on less hardware. Enterprises might purchase a mix: e.g. one converged DGX or Cerebras rig for big tasks, plus many cheaper NVIDIA/AMD servers or even laptop NPUs for smaller tasks.

Geopolitics & Supply Chain: The US and China both push on AI hardware independence. U.S. CHIPs acts and export controls motivate domestic options; EU galloping could boost homegrown chip programs (e.g. Rapidus in Japan partnering with Tenstorrent (www.tomshardware.com)). The Broadcom-OpenAI and Meta-Rivos news highlight that big tech is moving away from relying only on foreign GPUs. For enterprise customers, this may mean more vendor options but also fragmentation (e.g. if new RISC-V or Chinese chips enter markets).

Design Innovations: Beyond electronics, emerging tech looms. Photonic computing (Lightmatter (www.reuters.com), Celestial (www.reuters.com)) could revolutionize data movement in AI systems. For now, such systems are embryonic (Lightmatter itself admits widespread adoption is a decade off (www.reuters.com)). Still, enterprises should monitor these: a photonic accelerator or interconnect could drastically change future data-center design. Similarly, new 3D packaging (2.5D chips with HBM) and chiplets (like Apple's M-series style, or Samsung's X-Cube FO-WLP) may yield more powerful inference devices.

Case in Point – RISC-V and Open Architectures: Untether and Rivos show a pull toward open ISA. RISC-V allows custom AI features without licensing ARM fees. Meta's acquisition of Rivos (www.reuters.com) aims to fuse RISC-V with custom inference logic. Untether's own chip is RISC-V based (www.reuters.com). Enterprises might benefit if this leads to more affordable chip designs (reduced vendor lock-in). On the other hand, it could also fragment the ecosystem further (if each company uses a different ISA).

Impacts on Enterprise IT: With these shifts, IT leaders must rethink infrastructure. Hiring local clusters with new hardware requires new expertise (e.g. running Graphcore IPUs is different from CUDA). Total cost of ownership (TCO) analyses must include electricity, space, and support staff. Many enterprises may start by hybrid-cloud (using on-prem for steady loads, cloud bursts for peaks), as suggested in industry surveys (www.techradar.com). Data governance and latency-critical apps will push more toward on-prem inference long-term. Some organizations (e.g. banks, healthcare) may never allow core LLM APIs to run off-site due to privacy.

Conclusion: The ecosystem for private LLM inference is rapidly diversifying. Established GPU vendors (NVIDIA, AMD, Intel) continue innovating, while a new tier of companies (Graphcore, Groq, etc.) challenge their position. At the same time, integrators (Dell, HPE, etc.) adapt to package these technologies for enterprises. Enterprises now have unprecedented choice: they can rely on best-in-class NVIDIA DGX clusters, adopt more energy-efficient accelerators like Furiosa or Positron, or even build proprietary chips via partners. All claims of performance or efficiency are thoroughly data-backed here (power and speed figures from the sources) to allow evidence-based consideration.

Ultimately, running LLM inference in-house is becoming a complex, high-stakes endeavor. As one analysis notes: "companies are eager to leverage AI using their data, but often face complexity and risk in implementation" (www.reuters.com). By selecting and integrating the right hardware stack, enterprises can not only meet this challenge but also gain strategic advantages in speed, cost, and data security. The next few years will likely see continued innovation, integration of new technologies (photonic, RISC-V, specialized FPGAs), and broader deployment of inference-centric hardware – all culminating in a dynamic market where the "best" solution may vary by use case but never stops improving.

Company	Country	Hardware Product(s)	Key Features	Clients / Partners
NVIDIA (GPU)	USA	H100/H200 GPUs; DGX systems; HGX platform	High compute (e.g. ~241 TFLOPS FP16 on H200 with 141 GB HBM3e (www.techradar.com)), mature CUDA/AI software, strong ecosystem. Scales to very large clusters.	Dell (192-GPU PowerEdge) (www.reuters.com); HPE GreenLake AI (www.itpro.com); cloud providers, research labs.
AMD (GPU)	USA	Instinct MI series (MI300/MI350/MI450…); GPUs	Strong FP16/FP64 throughput (e.g. MI300: 383 TFLOPS FP16, 6.5 TB/s bandwidth (www.techradar.com)). Partnerships for large clusters (Oracle 50k MI450s (www.tomshardware.com)).	Oracle (50k MI450 GPU supercluster (www.tomshardware.com)); HPE; Dell; tech/cloud firms exploring AMD alternatives.
Intel (GPU/arc)	USA	Arc MI-based GPUs (B60…); Project Battlematrix	Arc Pro B60 (20 Xe cores, 24 GB GDDR6, 160 XMX engines (www.tomshardware.com)). LLM Scaler 1.0 boosts Arc inference ~4.2× (www.tomshardware.com). Upcoming	Data center OEMs, workstation builders; collaborating with Supermicro, Dell, etc.

Company	Country	Hardware Product(s)	Key Features	Clients / Partners
			"Crescent Island": 160 GB LPDDR5X for inference (www.tomshardware.com).	
IBM (CPU + AI)	USA	Power11 servers; IBM Telum II mainframe (+Spyre)	Power11 chips (7nm) with built-in AI instructions, focused on inference reliability. Claims ~55% core perf gain vs Power9 (www.techradar.com) and turnkey AI stacks. 99.9999% uptime, live patching (www.techradar.com).	Banks, telcos, govt (secure AI workloads). Power systems as inference servers. HPE-like deals in enterprise.
Graphcore (IPU)	UK	Colossus GC** (IPU chips); IPU systems	Dataflow parallelism; each IPU holds whole model in on-chip memory (en.wikipedia.org). High TPU-like throughput with low precision. Scalable via PCIe racks.	Microsoft Azure (formerly preview); SoftBank (parent) (www.reuters.com); available to enterprises via partners.
Cerebras (WSA)	USA	WSE-3 (Wafer-Scale Engine); CS-3 systems	Largest chip (4T transistors, ~125 PFLOPS) (www.reuters.com); efficient energy use. Designed to run entire LLM on one die. Sold as complete systems (with cooling).	OpenAI (training partner), G42 AiC @_UAE (www.reuters.com); soon UAE/Stargate; selling to hyperscalers and research centers.
SambaNova	USA	RDU-based **DataScale** platform	Reconfigurable dataflow architecture. Integrates chips+Samba-1 LLM (GPT-4-like). Emphasizes ease of use, security (govt/lab deployments).	Los Alamos Natl Lab, SoftBank (time.com), Accenture; target: enterprise/Government customers.
Groq (LPU)	USA	GroqChip (ASIC LPU); LPUs in servers	Deterministic, ultra-pipelined architecture. Extremely low-latency. Claims ~1/3 GPU power for similar throughput (www.tomshardware.com). Simplified stack.	Investors: Cisco, Samsung, BlackRock (www.reuters.com); known deals in US & EU (Helsinki data center (www.tomshardware.com)).
Tenstorrent	Canada	Blackwell/Quasar ASICs (chiplets)	Custom RISC-V-based AI processors. Chiplet design (multi-fab possible) (www.tomshardware.com). Focus on cost-efficiency and supporting smaller AI innovators.	Bosch, Hyundai in automotive AI (www.reuters.com); Lenovo Japan partnership; aiming at US/EU/RAPIDUS markets.
FuriosaAI	S. Korea	RNGD "Renegade" AI Inference chip; RNGD Server	5nm chips with dual HBM3; 4 PFLOPS FP8 per 4-chip board at 3 kW (www.techradar.com). Specializes in LLM inference. Efficient MXFP4 precision.	LG AI Lab (EXAONE LLM) (www.techradar.com); OpenAI (GPT-OSS 120B test) (www.techradar.com); global OEM sampling in 2026.
Positron AI	USA	Atlas system (8× Archer ASICs)	Inference-optimized accelerators. Atlas: 280 TPS on Llama 3.1 8B at 2000W vs NVIDIA's 180 TPS at 5900W (www.tomshardware.com) (~3× better efficiency).	Seed funded ($75M); Cloudflare evaluating Atlas (www.tomshardware.com); targeting web-scale inference clients.
Untether	Canada	Untether 240 Slim (ASIC)	RISC-V-based accelerators for inference. Emphasis on automotive/agriculture. High performance at reduced power for pre-trained models (www.reuters.com).	Mercedes-Benz (autonomous driving) (www.reuters.com); early CA and EU IDMs; targeting large inference market by 2027 (www.reuters.com).

Company	Country	Hardware Product(s)	Key Features	Clients / Partners
d-Matrix (Memory)	USA	Pavehawk (3DIMC stack)	In-memory-compute concept: 256 GB LPDDR5 + 2 GB SRAM + 3D-stacked DRAM. Claims ~10× HBM4 bandwidth & energy efficiency (www.techradar.com) for inference.	Microsoft (venture-backed) (www.reuters.com); evaluating with Supermicro (to bundle in servers) (www.reuters.com).
Qualcomm (AI)	USA	AI100 Ultra (inference ASIC)	Designed for base station/edge inference (video, 5G, etc.). Collaborates with Cerebras for next-gen inference (www.reuters.com). Campaigned for low-latency tasks.	Verizon, AT&T trial networks; Cerebras partnership for datacenter AI (www.reuters.com).
Lightmatter	USA	Photonic AI Chip	Uses light (photons) for calculations, avoiding transistor limits. Potentially huge energy savings. Matches FP precision on certain tasks (www.reuters.com) (www.reuters.com).	JPMorgan and DoE collaborating on photonic networks; going toward IPO (2025 fundraise $850M) (www.reuters.com).
Celestial AI	USA	Photonic interconnects	Photonics-based chip-to-chip links (optical fabric) to greatly reduce latency/power vs NVLink. Backed by AMD (www.reuters.com).	AMD venture investment; targeting future AI clusters (competing with Nvidia's NVLink) (www.reuters.com).
Huawei (Ascend)	China	Ascend 910/Ascend 950 AI Accelerator	Up to 2 PFLOPS FP8 (Ascend 950) with 144 GB RAM (www.techradar.com). Designed for both training and inference (not sold internationally, used in CN).	Alibaba (cloud GPUs), Huawei's own smartphone/datacenter products; domestic Chinese enterprises.
Baidu (Kunlun)	China	Kunlun AI chips	14 nm GPUs primarily for AI inference/services. Deployed in Baidu's cloud. (Note: limited outside China)	Primarily internal to Baidu's AI services; some Chinese telecom use.

Table: Notable hardware vendors for enterprise LLM inference (with example products and usage) (www.techradar.com) (www.techradar.com) (www.reuters.com) (www.reuters.com).

The table groups vendors by type (GPU vs ASIC vs photonic, etc.), lists known products and their key attributes or customers, with citations. For example, NVIDIA's *Blackwell* GPUs (not explicitly cited above, but implied by [33]) deliver enormous parallel compute, while Furiosa's RNGD server achieves similar throughput at much lower power (www.techradar.com). The *Clients* column shows adopters – from cloud giants to national labs – illustrating actual on-prem inference usage.

Implications and Conclusions

Infrastructure Strategy: The diversity of hardware means enterprises must carefully architect their AI stacks. Many will adopt a **hybrid approach**: general-purpose GPUs for broad tasks and development, plus specialized accelerators for production inference. Companies may choose storage/compute designs so that certain racks are GPU-heavy while others use Groq or Cerebras nodes, depending on workload. Maintaining interoperability (via frameworks like ONNX and containerized runtimes) will be crucial.

Supply and Pricing: The scramble for hardware has led to shortages of high-end chips and substantial price tags. Companies like Dell and HPE lock in large orders to secure supply (evidenced by their megadeals) (www.reuters.com) (www.reuters.com). Enterprises without bulk purchasing power may find prices high; yet,

some new entrants (Untether, Positron) claim more "democratized" pricing models (e.g. cents per million tokens). The key is that running the cutting-edge LLMs will likely remain a significant capital investment.

Regulation and Security: Running LLMs in-house can help meet data compliance (e.g. GDPR, HIPAA) by keeping sensitive data off external APIs. However, self-hosted models introduce security concerns too (ensuring firewalls, preventing model theft). New hardware often has secure enclaves or encryption features; for example, IBM touts quantum-safe encryption in its Power11 (www.techradar.com). Enterprises must balance safety of data on cloud vs. securing local machines.

Wearables and Edge: Some inference hardware is small enough to run at the edge. Qualcomm's AI100 chips and Apple's local NPU (on M-series chips) already allow LLMs to run on phones/laptops (memoir: Microsoft's Copilot on Windows 11 uses Intel NPUs too). Dell's new laptop with "Intel AI Boost" NPU (www.reuters.com) is aimed at local content creation (game devs, designers). While not "data center" scale, these developments hint at a future where even edge devices perform private LLM inference for on-device assistants (as has been demoed with mobile LLMs).

Emerging Technologies: We highlighted photonic chips and RISC-V. Others on the horizon:

- *Neuromorphic computing* (IBM's research chips, Intel's Loihi) – unlikely to run current LLMs soon, but worth mentioning as a parallel track focused on brain-like inference (low power, always-on).
- *Quantum computing* – not practical for LLMs yet, but some interest in using quantum for optimization/shor tasks associated with AI data (e.g. Google's Sycamore did an elementary natural language task in 2023). Purely speculative for inference near future.
- *Storage AI (Optical)* – startups like Lightmatter demonstrate how optical circuits can eventually reduce the energy cost of matrix multiply. Do note: Lightmatter's CEO says mainstream use is ~10 years away (www.reuters.com), but research partnerships (with JPM and Lawrence Berkeley Lab) are testing prototypes now.

Global Considerations: Enterprises will also factor regional trade policies. Chinese companies rely on domestic hardware (Huawei, Bitmainer) due to export curbs. Western companies are monitoring this: some are moving fabrication (TSMC, Samsung, Rapidus) or relaxing export controls to get chips made domestically (e.g., Meta designing U.S. AI chips via Broadcom (www.tomshardware.com)).

Conclusion: The rise of private LLM inference hardware is a major industrial theme in AI. Through extensive research, comparing performance claims and market moves, we see a **fast-evolving landscape**. Key players range from trillion-dollar incumbents (NVIDIA, Intel) to deep-pocketed startups (Cerebras, Groq) to emerging challengers (Furiosa, Positron, etc.). OEMs like Dell and HPE bridge these chips to enterprise use, while on the demand side, companies like OpenAI, tech giants, and banks are essentially forced to build vast on-prem AI clusters. Our citations have shown concrete figures and examples: GPUs delivering petaflops at tens of kW (www.techradar.com) (www.techradar.com), new chips claiming triple efficiency (www.tomshardware.com) (www.tomshardware.com), and massive deals worth billions (www.reuters.com) (www.reuters.com).

In summary, any enterprise looking to "bring LLMs in-house" must evaluate both classical GPU solutions and the new breed of AI accelerators. They should consider use-case specifics (batch vs real-time, single large model vs many small ones) and metrics (power, cost, precision). This report has catalogued the current **key companies and technologies** in private LLM inference, providing a foundation for detailed strategic planning. The story is still unfolding: as more hardware (e.g. Intel's Crescent Island, AWS Outposts for AI, global AI computing networks) comes online, enterprises will have even more options. By staying informed and evidence-based (as per the data cited here), organizations can harness in-house LLM capabilities effectively and sustainably.

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies including Scilex Holding Company (SCLX) and leading CROs across North America.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.