AMD Instinct MI455X Specs, Price, and Azure Deal Explained

Published July 26, 2026 38 min read

I'll add the internal links directly based on the article content provided.

Executive Summary

AMD unveiled the **Instinct MI455X** GPU and the accompanying **Helios** rack-scale platform at its Advancing AI 2026 event in San Francisco on July 23, 2026, positioning the chip as the flagship of its new **MI400 series** and its most direct challenge yet to Nvidia's data center dominance (Source: phoronix.com) (Source: heise.de). Built on 5th generation **CDNA 5** architecture and manufactured with TSMC 2 nanometer compute dies alongside 3 nanometer cache and I/O dies, the MI455X packs roughly **320 billion transistors** across 24 chiplets, delivers up to **40 PFLOPs of OCP MXFP4 compute** and over **20 PFLOPs of FP8 compute**, and carries **432 GB of HBM4 memory** with peak theoretical bandwidth AMD's own datasheet lists as **23.3 TB/s** (Source: amd.com) (Source: heise.de). Elsewhere on AMD's own Helios product page, however, the identical GPU is described as reaching only **19.6 TB/s** of bandwidth, a discrepancy this report addresses directly rather than picking one number silently (Source: amd.com).

AMD has not published a per-unit list price for the MI455X as of July 2026; early cloud partner materials list pricing simply as pending, consistent with AMD's practice of selling MI-series accelerators through negotiated hyperscaler and OEM contracts rather than retail channels (Source: tensorwave.com). The closest proxy is the prior-generation **MI355X**, which cloud providers rent for between **$2.95 and $8.60 per GPU-hour** depending on provider and commitment level (Source: gpus.io) (Source: getdeploying.com). The full **Helios rack** houses **72 MI455X GPUs** across 18 liquid-cooled compute trays plus AMD "Venice" EPYC CPUs and Pensando networking, delivering up to **2.9 exaFLOPS of FP4 compute**, **1.4 exaFLOPS of FP8 compute**, **31 TB of aggregate HBM4 memory**, and **260 TB/s of scale-up bandwidth** according to independent teardown analysis published alongside the launch (Source: chipsandcheese.com). AMD claims this beats Nvidia's competing **Vera Rubin NVL72** rack on compute, memory, and scale-out bandwidth, though independent comparison of each vendor's own published rack specifications (detailed in Table 2 below) shows a more mixed picture: AMD leads on memory capacity, 31 TB versus 20.7 TB per rack, while Nvidia's own datasheet claims higher aggregate FP4 throughput, 3.6 exaFLOPS versus AMD's 2.9 exaFLOPS (Source: nvidia.com).

On the commercial side, **Microsoft** announced on July 20, 2026 that Azure will deploy Helios at scale to power frontier-model inference, including new **ND MI455X v7** virtual machines, with shipments beginning in the second half of 2026 (Source: ir.amd.com) (Source: techwireasia.com). **Oracle Cloud Infrastructure** committed to an initial 50,000-GPU MI450 Series supercluster starting in Q3 2026 (Source: amd.com. **OpenAI** and **Meta** each signed 6-gigawatt, multi-year Instinct GPU agreements (October 2025 and February 2026, respectively), each paired with a warrant for up to 160 million AMD shares (Source: morethanmoore.substack.com (Source: fortune.com), with trade press estimating each deal's potential value near $90 billion and $100 billion respectively (Source: morethanmoore.substack.com) (Source: fortune.com). Most recently, **Anthropic** agreed on July 22, 2026 to deploy up to 2 gigawatts of MI455X-based Helios systems alongside a possible $5 billion AMD equity investment (Source: eweek.com). These commitments have coincided with rapid AMD Data Center segment growth, which reached **$5.775 billion in Q1 2026**, up 57 percent year over year (Source: ir.amd.com) (Source: techi.com). The remainder of this report details MI455X specifications, Helios rack architecture, pricing dynamics, the MI400 series lineup (including the HPC-focused **MI430X**), a full Nvidia Vera Rubin comparison, and named deployment case studies.

Introduction and Background

For more than a decade, AMD's Instinct line of data-center GPUs has chased Nvidia's dominant position in AI accelerator silicon, moving from the MI100 through MI250X, MI300X, MI325X, and the MI350 series before arriving at the MI400 series in 2026. The **AMD Instinct MI455X**, launched at AMD's Advancing AI 2026 keynote in San Francisco on July 23, 2026, represents the company's first GPU purpose-built for full rack-scale AI deployment rather than as a standalone accelerator retrofitted into someone else's server design (Source: chipsandcheese.com). AMD Chair and Chief Executive Officer **Dr. Lisa Su** framed the launch as a system-level bet: "AMD Instinct™ MI455X GPU is the Engine. AMD Helios is the System," according to AMD's own launch materials (Source: amd.com).The MI455X sits inside a broader **MI400 series** portfolio that also includes the **MI430X**, a variant tuned for sovereign AI and scientific high-performance computing (HPC) with native double-precision floating point (FP64) acceleration (Source: wccftech.com). Both chips are manufactured using TSMC's advanced packaging and sit inside AMD's new **Helios** rack-scale reference design, a full rack of 72 GPUs, EPYC "Venice" server CPUs, and Pensando networking hardware that Phoronix reported as already "in full production" on the July 23, 2026 launch date (Source: phoronix.com). Reviewers who saw the hardware in person described it as directly targeting Nvidia's **Vera Rubin NVL72**, the rack-scale system Nvidia is shipping through 2026 as the successor to its GB200/GB300 NVL72 platforms (Source: heise.de).

The stakes behind this launch are large. AMD's Data Center segment, which houses both EPYC server processors and Instinct GPUs, has become the company's largest and fastest-growing business, posting **$5.775 billion** in revenue for the first quarter of 2026 alone, a 57 percent year-over-year increase (Source: ir.amd.com) (Source: techi.com). AMD has signed multi-gigawatt supply agreements with OpenAI, Meta, Oracle, Microsoft, and Anthropic, each explicitly built around MI450-series silicon and the Helios rack, and each announced within the twelve months preceding this report's July 26, 2026 publication date. This report answers, with citations to primary AMD documentation, independent hardware analysis, and named commercial deployments verified during research in July 2026, what the MI455X actually is, what it costs to access, how the Helios rack compares to Nvidia's competing system, and where the chip is already being deployed.

What Is the AMD Instinct MI455X? Definition and Product Taxonomy

The **AMD Instinct MI455X** is a data-center graphics processing unit (GPU) built for artificial intelligence (AI) training, fine-tuning, and inference, along with select high-performance computing (HPC) workloads. It is the flagship accelerator of AMD's **MI400 series**, built on the company's **5th generation CDNA (Compute DNA) architecture** and packaged in a new liquid-cooled **Enhanced Accelerator Module (EAM)** form factor that replaces the OCP Accelerator Module (OAM) form used on the MI300X, MI325X, and MI350 series (Source: amd.com) (Source: chipsandcheese.com). CDNA 5 marks a structural break from AMD's prior compute architectures: independent hardware analysis found the underlying compute unit design borrows heavily from AMD's consumer RDNA4 graphics architecture rather than extending the GCN-derived CDNA lineage that powered every AMD data-center GPU since 2012, from the original MI6/MI8 parts through the MI300X and MI350 series (Source: chipsandcheese.com).

The MI400 series has three members disclosed as of July 2026:

- **MI455X**: the frontier AI accelerator that powers the 72-GPU AMD Helios rack, targeted at hyperscale AI training and inference (Source: phoronix.com).
- **MI450 Series**: the broader commercial designation AMD and its cloud partners use across most customer announcements (Oracle, OpenAI, Meta, Anthropic), of which the MI455X is one specific configuration; AMD's own Anthropic press release explicitly states the deployment will feature "AMD Instinct™ MI455X GPUs, part of the AMD Instinct MI450 Series" (Source: ir.amd.com).
- **MI430X**: a sovereign AI and HPC-oriented variant carrying native FP64 acceleration, aimed at national laboratories, research institutions, and governments rather than hyperscale commercial inference (Source: wccftech.com).

Both the MI455X and MI430X share the same 432 GB HBM4 (High Bandwidth Memory, generation 4) memory capacity. AMD's product page for the MI400 series lists the MI430X's memory bandwidth at a much lower **2.3 TB/s**, reflecting a memory controller configuration optimized for FP64 accuracy and capacity over raw throughput (Source: amd.com). Yet AMD's own investor-relations announcement of the Alice Recoque supercomputer, discussed further in the Case Studies section, describes MI430X GPUs delivering "432 GB of HBM4 memory and 19.6 TB/s of bandwidth," an eight-times-higher figure for the same chip that this report flags as a second unresolved internal AMD bandwidth discrepancy, alongside the MI455X bandwidth question addressed later in this report (Source: ir.amd.com). The MI430X claims up to **288 TFLOPS** of hardware-based FP64 performance, which independent trade press reported AMD describing as the "highest performance" FP64 capability among the MI400 series (Source: wccftech.com).

Four MI455X GPUs, together with a single AMD EPYC "Venice" server CPU, form a single **compute tray**, the fundamental building block of the Helios rack described in the following sections (Source: chipsandcheese.com).

AMD Instinct MI455X Technical Specifications

Compute Performance

AMD's official datasheet lists the MI455X's peak theoretical AI compute figures across multiple numeric precisions used in modern machine learning. At **4-bit precision** using the Open Compute Project's Micro-scaling FP4 (OCP MXFP4) format, the GPU reaches **40,265 TFLOPS**, commonly rounded to "up to 40 PFLOPs" in AMD marketing materials (Source: amd.com). At **8-bit precision** (OCP MXFP8 or standard FP8), independent trade press corroborates a peak throughput of roughly **20 PFLOPs** (Source: wccftech.com). Independent technical analysis published alongside the launch found the GPU delivers peak compute "ranging from 315 TFLOP for matrix/vector FP32 and vector FP16, and up to 40.26 PFLOP for OCP MXFP4," while double-precision FP64 tops out at just a few TFLOPS, underscoring that the MI455X (unlike the sibling MI430X) is optimized overwhelmingly for low-precision AI workloads rather than scientific simulation (Source: chipsandcheese.com).

The same independent technical analysis found the GPU contains **256 Work Group Processors (WGPs)** spread across **8 Accelerator Complex Dies (XCDs)**, running at a peak engine clock of **2.4 GHz** (Source: chipsandcheese.com. Each WGP contains four dual-issue Wave32 SIMD32 (single instruction, multiple data) units, a substantial redesign from the single-issue Wave64 units used in the prior CDNA4-based MI350 series, and AMD doubled the register file addressable per wavefront to 1,024 vector general-purpose registers (VGPRs) to support the new execution model (Source: chipsandcheese.com). A new memory multicasting capability lets a single load instruction broadcast shared operand data (such as matrix weights reused across a general matrix multiplication) to every relevant WGP simultaneously, which AMD's hardware documentation describes as delivering "up to 4x" effective bandwidth amplification for data already resident in the on-package L2 cache, though this multiplies locally delivered data rather than the physical bandwidth to HBM4 itself (Source: chipsandcheese.com).

Memory Subsystem: HBM4 Capacity and Bandwidth

The MI455X integrates **432 GB of HBM4** memory across **12 stacks**, each with a capacity of 36 GB running over a 2,048-bit bus, giving a combined 24,576-bit total memory interface (Source: wccftech.com). AMD's formal datasheet describes this as "up to 2.9x higher peak memory bandwidth than the previous-generation AMD Instinct MI355X GPU," which carried 8 TB/s of HBM3E bandwidth (Source: gpus.io). Independent teardown analysis corroborates a peak figure of 23.3 TB/s, calculating roughly 7.6 giga-transfers per second per pin across the 2,048-bit bus per stack (Source: chipsandcheese.com).

Yet on the very same AMD Helios product page, elsewhere in a components description and again in an FAQ section, AMD describes the MI455X as delivering "432 GB HBM4 memory and up to 19.6 TB/s bandwidth per GPU," a figure that also appears in independent trade coverage of the Microsoft deployment (Source: amd.com) (Source: techwireasia.com). This report treats **23.3 TB/s** as the more authoritative figure, since it is the number stated in AMD's formal datasheet PDF and corroborated independently by chip analysts examining the physical die, and is consistent with the rack-level total of 1.7 PB/s across 72 GPUs; the 19.6 TB/s figure likely reflects either a sustained (rather than peak theoretical) bandwidth rating or an earlier engineering estimate not fully updated across every AMD and partner web property before launch. Readers evaluating vendor quotes or system configurations referencing the MI455X should expect to see both figures in circulation and should request the specific bandwidth basis (peak theoretical versus sustained) from AMD or an OEM before making a procurement decision.

The GPU's on-package **Level 2 (L2) cache** grew to **192 MB total**, split across two Fabric and Cache Dies (FCDs) of 96 MB each, up from AMD's prior "Infinity Cache" design that was allocated per-compute-unit; critically, CDNA 5 removed the ability for a compute unit attached to one FCD to directly access the L2 cache on the other FCD within the same package, a deliberate architectural tradeoff AMD made to simplify atomic-operation coherency and eliminate a kernel-flush boundary required by the MI300-series chiplet design (Source: heise.de) (Source: chipsandcheese.com).

Chiplet Packaging and Process Node

A complete MI455X package integrates **24 individual chips**: eight XCDs (the compute dies) manufactured on TSMC's 2 nanometer (N2) process, two FCDs and I/O dies manufactured on TSMC's 3 nanometer (N3P) process, and twelve HBM4 memory stacks, all assembled using TSMC's CoWoS-L advanced packaging technology (Source: heise.de) (Source: chipsandcheese.com). Trade press covering the launch reported that AMD's compute dies use "3D Hybrid Bonded XCDs," providing higher interconnect density than the 2.5D packaging used on the MI300 series (Source: wccftech.com). The two FCDs are linked by a die-to-die interconnect independent analysis estimates at roughly **14 TB/s** of bidirectional bandwidth, while the overall transistor count across the full package reaches approximately **320 billion**, about 70 percent more than the MI355X and roughly 16 billion fewer than the 336 billion transistors independent teardown estimates attribute to Nvidia's competing Rubin GPU (Source: chipsandcheese.com) (Source: tensorwave.com) (Source: wccftech.com).

Notably, AMD has not publicly disclosed a thermal design power (TDP) figure for the MI455X as of the July 2026 launch. German technology outlet Heise observed that "one piece of information is conspicuously missing from AMD's documentation: how much electrical power an Instinct MI455X or an entire Helios system consumes," and estimated that a fully populated Helios rack could plausibly draw "more than 200 kilowatts," while noting the predecessor MI355X already consumed up to 1.4 kW per GPU (Source: heise.de) (Source: heise.de). AMD's own footnotes confirm that "actual power consumption, thermal design, and system configuration may vary by deployment and are available to qualified customers and partners under a mutual non-disclosure agreement," meaning prospective buyers cannot obtain firm power figures from public documentation alone (Source: amd.com).

Networking and Interconnect

The MI455X connects to its host EPYC "Venice" CPU using a dedicated 16-lane AMD Infinity Fabric link rather than PCI Express, delivering **256 GB/s** of bidirectional bandwidth for coherent CPU-to-GPU memory access, a change from the PCIe-based host link used on the MI300 series (Source: chipsandcheese.com). I/O dies retain PCI Express 6.0 links used to connect network controllers via the **Ultra Accelerator Link (UALink)** open standard (Source: heise.de. For scale-up connectivity within a rack, each GPU exposes **36 UALink-over-Ethernet (UALoE) interfaces** at 400 Gbit/s each, aggregating to **3.6 TB/s of peak bidirectional bandwidth per GPU** (Source: amd.com) (Source: chipsandcheese.com). For scale-out connectivity between racks, independent teardown analysis lists peak bandwidth of **43 TB/s** aggregated at the rack level via up to three 800 Gbit/s AMD Pensando "Vulcano" AI network interface cards (NICs) per GPU (Source: chipsandcheese.com). The GPU also supports advanced partitioning: up to eight compute partitions and four spatial memory partitions per device, letting operators subdivide a single MI455X into smaller isolated instances for multi-tenant cloud use, according to independent reporting on AMD's technical pre-briefing materials (Source: wccftech.com).

Table 1 below summarizes the MI455X's core specifications alongside its immediate predecessor and its direct HPC-focused sibling.

SPECIFICATION	AMD INSTINCT MI355X (PRIOR GEN)	AMD INSTINCT MI455X	AMD INSTINCT MI430X
Architecture	CDNA 4	CDNA 5	CDNA 5
Process node	TSMC N3	TSMC N2 (compute) / N3P (I/O, cache)	TSMC N2 (compute) / N3P (I/O, cache)
Transistors	Not disclosed	Approximately 320 billion (Source: tensorwave.com)	Not separately disclosed
Memory capacity	288 GB HBM3E (Source: gpus.io)	432 GB HBM4 (Source: wccftech.com)	432 GB HBM4 (Source: amd.com)
Peak memory bandwidth	8 TB/s (Source: gpus.io)	23.3 TB/s per AMD datasheet; 19.6 TB/s per some AMD/partner pages (Source: heise.de)	2.3 TB/s (Source: amd.com)
Peak MXFP4 compute	Approximately 20 PFLOPs (Source: wccftech.com)	40,265 TFLOPS (~40 PFLOPs) (Source: chipsandcheese.com)	Not primary use case
Peak FP64 compute	Not primary use case	Approximately 5 TFLOPS (Source: chipsandcheese.com)	288 TFLOPS (Source: amd.com)
TDP (max)	1,400 W (Source: gpus.io)	Not disclosed publicly as of July 2026 (Source: heise.de)	Not disclosed publicly
Cloud rental price (July 2026)	$2.95 to $8.60 per GPU-hour (Source: gpus.io) (Source: getdeploying.com)	Not yet publicly listed (Source: tensorwave.com)	Not applicable (2027 availability)

The table illustrates that AMD's central engineering tradeoff between the MI455X and MI430X is memory bandwidth versus double-precision compute: the MI455X sacrifices FP64 throughput almost entirely in exchange for triple-digit-terabyte-per-second HBM4 bandwidth suited to serving large language models, while the MI430X inverts that tradeoff to serve scientific simulation workloads that depend on numerical precision more than raw memory throughput. Both chips nonetheless share the same 432 GB memory ceiling, reflecting a design philosophy of standardizing memory capacity across the MI400 family while varying bandwidth and compute mix by target workload.

The AMD Helios Rack-Scale Platform

AMD Helios is the company's first rack-scale reference architecture designed jointly with the hardware itself rather than assembled afterward by system integrators. The design uses the **Open Rack Wide (ORW)** specification, a double-wide rack form factor Meta submitted to the Open Compute Project (OCP) in 2025, with AMD and Meta co-developing the Helios architecture around it (Source: ir.amd.com). Independent reporting on the platform notes it "uses OCP's new Open Rack Wide form factor, comprising a cabinet 1.2m wide and 1.3m deep," a physically larger footprint than a conventional single-wide data-center rack (Source: chipsandcheese.com).

A single Helios rack contains **72 MI455X GPUs** organized into **18 compute trays**, each housing four liquid-cooled GPUs paired with one 96-core AMD EPYC "Venice" server CPU, for a rack total of roughly **4,600 CPU cores and 18,000 GPU compute units**, according to hands-on reporting from the launch event (Source: phoronix.com). Independent teardown analysis further details that the compute trays are organized into two groups of nine, and that six additional **Helios Switch Trays** house twelve total network switches, each delivering 21.6 TB/s of bidirectional bandwidth for an aggregate scale-up fabric of 260 TB/s (Source: chipsandcheese.com) (Source: heise.de).

At full capacity, independent chip analysts who examined the reference design at launch corroborate the following aggregate rack-level specifications for a Helios deployment: **2.9 exaFLOPS** of dense FP4 compute and **22.6 PFLOPS** of dense FP32 compute; **31 TB** of combined HBM4 memory across all 72 GPUs, with an aggregate memory bandwidth of **1.7 PB/s**; and **260 TB/s** of scale-up bandwidth linking all 72 GPUs into a single load/store domain, with **43 TB/s** of scale-out bandwidth for connecting multiple racks into larger clusters (Source: chipsandcheese.com) (Source: chipsandcheese.com).

AMD's own comparison against Nvidia's rival Vera Rubin NVL72 rack, based on the company's internal performance-lab modeling, claims Helios delivers "up to 15% more AI compute, 50% more HBM capacity and 50% more scale-out bandwidth" than the Nvidia system; on a modeled inference workload using the Kimi K2 Thinking model with a 32,000-token input and 8,000-token output sequence, AMD's performance labs further estimate Helios delivers 15% higher throughput per GPU at low interactivity, 12% higher at medium interactivity, and 10% higher at high interactivity versus the equivalent Nvidia rack configuration, translating to an estimated "up to 30% more tokens per dollar" compared to Vera Rubin NVL72 (Source: amd.com).

Phoronix reported from the launch event that the platform is already "in full production" with shipments to customers "happening later in the quarter and ramping into Q4 and H1'2027" (Source: phoronix.com). AMD's launch materials list Dell, HPE, IBM, and Cisco as infrastructure partners for building and servicing Helios systems, alongside OpenAI, Meta, Microsoft, and Oracle as its principal "at scale" AI partners for the platform (Source: techwireasia.com).

AMD Instinct MI455X Price and Availability

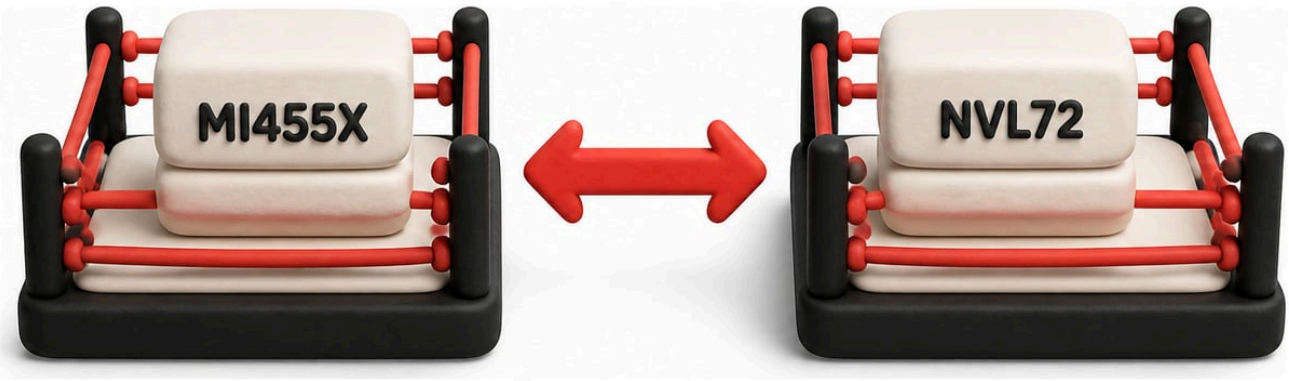
AMD does not publish a per-unit list price for the MI455X, consistent with its longstanding practice of selling Instinct-series accelerators through direct hyperscaler agreements, OEM system partners, and cloud rental partners rather than a public price list. Early cloud partner materials confirm this: one prominent Instinct-focused GPU cloud provider lists MI455X pricing simply as **"TBD"** as of its product page publication in mid-2026 (Source: tensorwave.com). This makes the closest available pricing proxy the previous-generation **MI355X**, which AMD's own product page lists with **288 GB of HBM3E memory and 8 TB/s of bandwidth**, and which cloud providers currently rent for between **$2.95 per GPU-hour** (TensorWave, on-demand) and **$8.60 per GPU-hour** (Oracle Cloud Infrastructure, on-demand, no term commitment) (Source: amd.com) (Source: gpus.io) (Source: getdeploying.com).

As a genuinely full 72-GPU Helios rack, one German outlet's technical analysis estimated the purchase price of a complete system "is likely to be several million US dollars," though AMD has not confirmed a system-level list price and the estimate should be read as an informed approximation rather than a disclosed figure (Source: heise.de). In practice, the overwhelming majority of MI455X capacity in 2026 and 2027 will be sold not as discrete hardware but as bundled multi-gigawatt supply agreements with hyperscalers and frontier AI labs, discussed in the Case Studies section below, or accessed indirectly through cloud rental pricing once general-availability cloud instances launch. AMD's Microsoft announcement specifically references upcoming **Azure ND MI455X v7** virtual machines, though as of the July 2026 announcement, "Microsoft has not announced pricing, launch dates, or the Azure regions where [new instances] will initially be available" (Source: techwireasia.com).

Regarding release timing, Phoronix reported directly from AMD's launch event that shipments would be "happening later in the quarter and ramping into Q4 and H1'2027," while Oracle's press materials specify its own initial 50,000-GPU deployment beginning "in calendar Q3 2026" (Source: phoronix.com) (Source: amd.com), while the MI430X HPC variant is not expected to be broadly available until 2027 (Source: amd.com). Buyers should therefore treat "MI455X availability" as a staggered rollout across the second half of 2026 into 2027, gated by individual partner timelines rather than a single universal launch date, notwithstanding the July 23, 2026 product unveiling itself.

MI455X vs Nvidia Vera Rubin NVL72: Performance Comparison

Nvidia's directly competing rack-scale system is the **Vera Rubin NVL72**, which Nvidia's own datasheet describes as combining **72 Rubin GPUs and 36 Vera CPUs** interconnected by sixth-generation NVLink (Source: nvidia.com). Each individual Rubin GPU carries **288 GB of HBM4** memory at up to **22 TB/s** of bandwidth, delivers **50 PFLOPS of NVFP4 inference** compute (Nvidia's own proprietary 4-bit floating point format) and **35 PFLOPS of dense NVFP4 training** compute, and is built with an estimated **336 billion transistors** using TSMC's 3 nanometer process (Source: nvidia.com) (Source: hashrateindex.com). At the full rack level, Nvidia's own specification sheet lists **3,600 PFLOPS of NVFP4 inference**, **2,520 PFLOPS of NVFP4 training**, **20.7 TB of total HBM4 memory**, and **260 TB/s** of aggregate NVLink 6 scale-up bandwidth, the same headline scale-up figure AMD claims for Helios (Source: nvidia.com) (Source: nvidia.com. Scale-out networking on Vera Rubin NVL72 reaches **28.8 TB/s**, and independent trade coverage of Nvidia's own datasheet places rack power draw at roughly **190 kW in Max Q mode and 230 kW in Max P mode**, with 100 percent liquid cooling using 45 degrees Celsius inlet water (Source: nvidia.com (Source: hashrateindex.com.



At the individual chip level, per-chip figures reported by independent hardware press credit the MI455X with a slight edge in FP8 throughput (roughly 20 PFLOPs versus Nvidia's reported 17.5 PFLOPs dense FP8), while Nvidia's Rubin GPU claims a higher headline 4-bit figure (50 PFLOPS NVFP4 versus AMD's 40 PFLOPS OCP MXFP4), though German outlet Heise cautioned that "Nvidia reports 35 to 50 petaflops for its current Rubin GPU, depending on the workload, including sparsity, which means AMD's accelerator is faster at least on paper," before adding the standard caveat that "manufacturer benchmarks should therefore be viewed with caution" since both vendors measure peak theoretical performance under favorable, and not always directly comparable, precision-format assumptions (Source: wccftech.com) (Source: heise.de) (Source: heise.de). On memory, AMD's own competitive comparison slides summarized the matchup at the chip level as "1.5x Memory Capacity vs Competition," "Same Memory Bandwidth vs Competition," "Same FP4/FP8 FLOPs vs Competition," "Same Scale-Up Bandwidth vs Competition," and "1.5x Scale-Out Bandwidth vs Competition," a more conservative self-assessment than some of the rack-level "15% more compute" marketing claims elsewhere in AMD's materials (Source: wccftech.com).

Table 2 below compares the two rack-scale platforms directly using each vendor's own published specifications.

METRIC	AMD HELIOS (72X MI455X)	NVIDIA VERA RUBIN NVL72 (72X RUBIN GPU)
GPU count per rack	72 (Source: heise.de)	72 (Source: nvidia.com)
Host CPU	AMD EPYC "Venice" (up to 256 cores) (Source: techwireasia.com)	Nvidia Vera (88 custom Olympus cores per CPU) (Source: hashrateindex.com)
Per-GPU memory	432 GB HBM4 (Source: wccftech.com)	288 GB HBM4 (Source: nvidia.com)
Per-GPU FP4 compute	~40 PFLOPS OCP MXFP4 (Source: chipsandcheese.com)	50 PFLOPS NVFP4 (Source: nvidia.com)
Rack HBM4 capacity	31 TB (Source: chipsandcheese.com)	20.7 TB (Source: nvidia.com)
Rack scale-up bandwidth	260 TB/s (Source: chipsandcheese.com)	260 TB/s (Source: nvidia.com
Rack scale-out bandwidth	43 TB/s (Source: chipsandcheese.com)	28.8 TB/s (Source: nvidia.com)
Rack-level FP4 compute	2.9 exaFLOPS (Source: chipsandcheese.com)	3.6 exaFLOPS (Source: nvidia.com)
Rack form factor / cooling	OCP Open Rack Wide, double-wide, fully liquid-cooled (Source: chipsandcheese.com)	Standard NVL72, fully liquid-cooled, 45C inlet (Source: hashrateindex.com)
Estimated rack power draw	Not disclosed publicly; predecessor MI355X GPU alone drew up to 1.4 kW (Source: heise.de)	Approximately 190 to 230 kW (Source: hashrateindex.com

The comparison in Table 2 shows a system-level split rather than a clean AMD win: AMD leads decisively on raw memory capacity (432 GB versus 288 GB per GPU, and 31 TB versus 20.7 TB per rack) and on scale-out networking (43 TB/s versus 28.8 TB/s), while Nvidia leads on raw per-GPU FP4 throughput and, by extension, aggregate rack-level FP4 exaFLOPS. Because AI inference throughput depends heavily on memory capacity and bandwidth (which determine how large a model and its key-value cache can remain resident without costly offloading) as much as on raw compute FLOPs, AMD's memory advantage is a meaningful practical differentiator even where its peak compute trails Nvidia's on paper, and it is the basis for AMD's own modeled claim of higher tokens-per-dollar economics discussed above.

Implementation Guidance: Deploying MI455X and Helios

Organizations evaluating MI455X or Helios adoption in 2026 and 2027 face a materially different procurement path than for a conventional GPU purchase. Because AMD is not selling the MI455X through retail or small-volume distribution channels at launch, prospective adopters fall into three practical tiers:

- **Hyperscale cloud and frontier AI labs**: entities such as Microsoft, Oracle, Meta, OpenAI, and Anthropic negotiate multi-gigawatt, multi-year supply agreements directly with AMD, often including equity warrants or direct AMD investment as part of the commercial structure, as detailed in the Case Studies section below.
- **Enterprises seeking managed cloud access**: customers can access AMD-based capacity indirectly once cloud providers stand up MI455X or Helios-based instances; Microsoft's Azure Foundry Managed Compute service, for example, lets customers select a model, accelerator family, and deployment template while Microsoft manages the underlying GPU topology and container runtime, though as of the July 2026 announcement this service "remains in public preview, has no service-level agreement, and is not currently recommended by Microsoft for production workloads" (Source: techwireasia.com).
- **Research institutions and sovereign buyers**: government-backed programs such as France's Alice Recoque supercomputer procure MI430X-based systems through system integrators like Eviden rather than direct AMD contracts, discussed further below.

For teams planning software migration onto MI455X or Helios, AMD's **ROCm (Radeon Open Compute)** software stack remains the primary programming interface, with native support for **PyTorch, TensorFlow, JAX, ONNX Runtime, vLLM, SGLang, and DeepSeek** frameworks according to AMD's own datasheet (Source: amd.com). AMD's Primus tool drives distributed training and fine-tuning optimized for rack- and cluster-scale deployments, and the company has publicly committed to a **six-week ROCm release cadence** intended to accelerate feature parity with rival software ecosystems (Source: phoronix.com). Teams should also plan for the platform's **advanced partitioning model**: independent teardown analysis describes how a full 72-GPU Helios rack can be subdivided into virtual pods ranging from four GPUs sharing a common CPU within a single compute tray up to the full complement of compute nodes in the rack, a useful property for multi-tenant scheduling (Source: chipsandcheese.com).

Before committing capital or long-term contracts, procurement teams should specifically request from AMD or an OEM partner: (1) the **basis for any quoted HBM4 bandwidth figure** (peak theoretical versus sustained, given the 23.3 TB/s versus 19.6 TB/s discrepancy documented above), (2) **actual power and cooling requirements** under a non-disclosure agreement, since public specifications omit TDP entirely, and (3) **realistic delivery timelines** given that AMD's own materials describe 2026 as a ramp period extending into 2027 rather than a single general-availability date.

Data Analysis and Evidence

AMD's Data Center segment, which combines EPYC server CPU and Instinct GPU revenue, posted **$5.775 billion** in net revenue for the first quarter of fiscal 2026, an increase of 57 percent year over year from $3.674 billion in the equivalent quarter of 2025, according to AMD's own quarterly financial results (Source: ir.amd.com). The segment generated **$1.599 billion** in operating income for the quarter, and total AMD net revenue across all segments reached **$10.253 billion**, up from $7.438 billion in the prior-year quarter (Source: ir.amd.com). Trade press coverage of the same results reported that AMD guided second-quarter 2026 revenue to approximately **$11.2 billion**, "well above the roughly $10.5 billion Street setup," and specifically cited management commentary that "customer engagement around MI450 and Helios is strengthening" (Source: techi.com) (Source: techi.com).

Turning to the deals underpinning that growth, AMD and its partners disclosed unit and capacity figures across five major gigawatt-scale supply agreements, summarized in *Table 3* below. Note that only capacity figures (gigawatts, GPU counts, dates) come from AMD's own or the counterparty's official press releases; specific dollar valuations are drawn from trade press estimates since neither AMD nor its partners have disclosed exact contract values.

CUSTOMER	ANNOUNCEMENT DATE	COMMITTED CAPACITY	FIRST DEPLOYMENT	ESTIMATED DEAL VALUE (TRADE PRESS)
OpenAI	October 6, 2025 (Source: morethanmoore.substack.com)	Up to 6 gigawatts, multi-generation (Source: morethanmoore.substack.com)	1 GW of MI450 Series GPUs, 2H 2026 (Source: ir.amd.com)	~$90 billion potential sales value (Source: morethanmoore.substack.com)
Meta	February 24, 2026 (Source: fortune.com)	Up to 6 gigawatts, multi-generation (Source: fortune.com)	Custom MI450-based GPU, 1 GW, 2H 2026 (Source: fortune.com)	Potentially exceeding $100 billion (Source: fortune.com)
Oracle (OCI)	October 14, 2025 (Source: amd.com)	50,000 GPUs, expanding 2027+	MI450 Series, Q3 2026	Not publicly estimated
Microsoft (Azure)	July 20, 2026 (Source: ir.amd.com)	Not disclosed in gigawatts	Helios at scale, 2H 2026 (Source: ir.amd.com)	Not publicly estimated
Anthropic	July 22, 2026 (Source: eweek.com)	Up to 2 gigawatts (Source: eweek.com)	First GW, H1 2027 (Source: eweek.com)	Tens of billions of dollars (Reuters, via eWeek) (Source: eweek.com)

The pattern visible in Table 3 is a shift in AMD's revenue model itself: rather than selling individual GPUs at list price, AMD is increasingly structuring its largest AI deals around performance-based equity warrants tied to shipment milestones. Both the OpenAI and Meta agreements include a warrant for **up to 160 million shares of AMD common stock**, vesting in tranches as gigawatt deployment milestones are reached and as AMD's stock price crosses specified thresholds (Source: ir.amd.com) (Source: fortune.com). AMD's own chief financial officer, Jean Hu, stated the OpenAI partnership "is expected to deliver tens of billions of dollars in revenue for AMD" and would be "highly accretive to AMD's non-GAAP earnings-per-share" (Source: ir.amd.com). Rather than a straightforward chip sale, the Anthropic deal folds in a reciprocal element: AMD committed to make an equity investment of up to $5 billion in Anthropic, while Anthropic agreed to use its Claude models to optimize workloads for AMD Instinct GPUs and accelerate ROCm software development (Source: eweek.com) (Source: eweek.com).

On raw inference performance, AMD's own benchmarking on the open-weight **DeepSeek-V4-Flash** model, a model that independent analysis firm Artificial Analysis "places among the leading open-weight models in intelligence when compared with models of similar size," reports that MI455X GPUs deliver "up to 34X higher token throughput at high interactivity and up to 18X lower token cost compared with AMD Instinct MI355X GPUs," figures AMD's footnotes attribute to internal performance-lab measurement rather than independent third-party benchmarking (Source: amd.com).

Case Studies and Real-World Examples

Microsoft Azure and the ND MI455X v7 Virtual Machines

On July 20, 2026, AMD and Microsoft announced an expanded strategic partnership under which "Microsoft will deploy the AMD Helios Rackscale Solution, to power frontier model AI inference for Microsoft, its AI customers and Azure AI services" (Source: ir.amd.com). Microsoft plans to offer this capacity to customers through new **Azure ND MI455X v7** virtual machines "designed for production-scale reasoning, search, and agentic inference workloads," though it "has not provided a timetable for customer access to the instances" as of the announcement (Source: techwireasia.com) (Source: techwireasia.com). The partnership also introduced two new AMD EPYC "Venice"-powered VM series, **Azure HDv2** for agentic AI and data pipelines, and **Azure HXv2** for semiconductor design workloads with up to 800 Gigabit InfiniBand connectivity (Source: techwireasia.com). Satya Nadella, Microsoft's Chairman and CEO, said the collaboration is "expanding the Azure infrastructure portfolio with AMD Helios to give customers the performance, scale and choice they need to build and run the next generation of AI applications" (Source: ir.amd.com).

Oracle Cloud Infrastructure's 50,000-GPU Supercluster

Announced October 14, 2025 at Oracle AI World, Oracle Cloud Infrastructure (OCI) committed to being "a launch partner for the first publicly available AI supercluster powered by AMD Instinct™ MI450 Series GPUs, with an initial deployment of 50,000 GPUs starting in calendar Q3 2026 and expanding in 2027 and beyond" (Source: amd.com. Oracle's own executive vice president, Mahesh Thiagarajan, said the deployment would let "Oracle customers gain powerful new capabilities for training, fine-tuning, and deploying the next generation of AI," and the deal builds on Oracle's existing general availability of MI355X-powered compute within a zettascale OCI Supercluster capable of scaling to **131,072 GPUs** (Source: amd.com). Independent trade press separately corroborated that this initial supercluster commitment specifically involves MI455X-class hardware within the broader MI450 Series designation (Source: wccftech.com).

OpenAI's 6-Gigawatt Commitment

OpenAI's agreement with AMD, announced October 6, 2025, calls for OpenAI to "purchase and deploy (through affiliates) up to six gigawatts of AMD Instinct GPUs over multiple hardware generations, beginning with the MI450 series in the second half of 2026" (Source: morethanmoore.substack.com. OpenAI CEO Sam Altman called the agreement "a major step in building the compute capacity needed to realize AI's full potential" (Source: ir.amd.com), while AMD structured the deal with a stock warrant "for up to 160 million shares of AMD common stock, structured to vest as specific milestones are achieved," with the first tranche vesting alongside the initial 1-gigawatt deployment (Source: ir.amd.com). Independent semiconductor analyst Dr. Ian Cutress noted AMD's stock "was up 35% in early trading on this news" the morning the deal was announced (Source: morethanmoore.substack.com).

Meta's Custom MI450 Deployment

Meta and AMD announced an expanded 6-gigawatt agreement on February 24, 2026. Fortune's coverage of the announcement reported that "Meta will buy AMD's latest chips, the MI450, to help power data centers," and that "the 6-gigawatt agreement will see shipments supporting the first gigawatt deployment set to start during the second half of this year," with the deal "potentially worth more than $100 billion" (Source: fortune.com). Meta

founder and CEO Mark Zuckerberg said the company was "excited to form a long-term partnership with AMD to deploy efficient inference compute and deliver personal superintelligence," calling it "an important step for Meta as we diversify our compute" (Source: ir.amd.com). Shares of AMD "finished Tuesday's regular trading session up nearly 9%" on news of the deal, according to Fortune's reporting citing the Associated Press (Source: fortune.com), and analyst commentary characterized the agreement as validating "the Helios rack-scale platform as a credible alternative to NVIDIA's NVL72" (Source: introl.com).

Anthropic's Diversification Beyond Nvidia

On July 22, 2026, just one day before the MI455X's formal unveiling, AMD and Anthropic announced a partnership under which "Anthropic will deploy AMD Helios rack-scale solutions featuring AMD Instinct™ MI455X GPUs, part of the AMD Instinct MI450 Series, together with AMD EPYC™ 'Venice' CPUs, AMD Pensando™ networking and ROCm™ software" for up to 2 gigawatts of capacity (Source: ir.amd.com), with deployment of the first gigawatt beginning in the first half of 2027 (Source: eweek.com). Notably, this deployment builds on Anthropic's existing use of AMD Instinct MI355X GPUs, meaning Anthropic was already an AMD customer prior to this expanded commitment (Source: ir.amd.com). Anthropic's co-founder and chief compute officer, Tom Brown, framed the diversification rationale directly: "Running across a diversified range of hardware lets us map the right workloads to the right hardware" (Source: ir.amd.com). Trade coverage noted that "Anthropic already runs workloads on Nvidia GPUs, Google's TPUs, and Amazon's Trainium chips," positioning AMD as a fourth compute vendor in Anthropic's hardware portfolio rather than a replacement for any existing supplier (Source: eweek.com). Emarketer analyst Jacob Bourne, cited by Reuters, said the win "reinforces AMD's position as Nvidia's closest competitor in AI accelerators" (Source: eweek.com).

Alice Recoque: Sovereign AI in Europe Using the MI430X

Beyond hyperscale commercial deployments, AMD's MI400 series also anchors a sovereign HPC project in Europe. Announced November 18, 2025, **Alice Recoque** will be "France's first and Europe's second Exascale supercomputer," a project GENCI leads and CEA operates, built on Eviden's new BullSequana XH3500 platform and powered by "next-gen AMD EPYC CPUs, codenamed 'Venice,' AMD Instinct MI430X GPUs, a new MI400 Series accelerator engineered for sovereign AI and scientific computing, and AMD FPGAs" (Source: ir.amd.com). Eviden's own announcement of the project put its overall cost at **554 million euros over five years** of operation, funded jointly by the EuroHPC Joint Undertaking's Digital Europe Programme and the Jules Verne consortium of France, the Netherlands, and Greece (Source: bull.com). The system is expected to deliver "more than one exaflop of HPL [High Performance Linpack] performance" and, according to Eviden, will use "25% less racks and components than other Exascale systems and up to 50% better energy efficiency per GPU" compared with prior-generation exascale designs (Source: ir.amd.com) (Source: bull.com). Composed of **94 racks**, Alice Recoque is expected to be "one of the top supercomputers in Europe for double-precision HPC workloads" (Source: ir.amd.com). The European supercomputing agency EuroHPC noted the system "will employ warm-water direct-liquid cooling for the unified racks and chilled-door technologies for the scalar racks," managed through "CEA's Ocean suite completed by Eviden's management suite with widely used open-source components such as SLURM, Kubernetes, LUSTRE, Grafana or Prometheus" (Source: eurohpc-ju.europa.eu).

Implications and Future Directions

AMD has committed to an **annual cadence** of Instinct GPU releases going forward, with the **MI500 series (CDNA 6)** already confirmed for 2027 and the **MI600 series (CDNA-Next)** targeted for 2028, according to reporting from the Advancing AI 2026 launch event (Source: phoronix.com). AMD itself frames this shift as strategically significant beyond any single product generation, stating that "Helios also marks a broader shift in how AMD advances AI infrastructure. Annual execution now connects successive AMD Instinct GPU generations with progress in memory, interconnect and GPU scale-up" (Source: amd.com).

This pace mirrors, and is explicitly framed by outside observers as a direct response to, Nvidia's own accelerated release cadence, which similarly compresses what was once a roughly two-year generational gap into an annual "standard and Ultra" pattern (Source: wccftech.com). Nvidia itself is already previewing its next rack, "Rubin Ultra," featuring a new "Kyber" rack design housing 144 GPUs at roughly 600 kW and 15 exaFLOPS of FP4 compute, expected in the second half of 2027, underscoring that AMD's MI455X launch does not represent a static target but a moving one (Source: hashrateindex.com.

The industry-wide adoption of open standards is a recurring theme across AMD's MI400-series messaging: the Helios rack is built on the OCP-submitted Open Rack Wide specification, UALink for scale-up interconnect, and the Ultra Ethernet Consortium's standards for scale-out networking, all explicitly positioned by AMD as an alternative to Nvidia's more proprietary NVLink-centric approach (Source: amd.com. Analysts covering the Meta deal have observed that "AMD has historically priced its data center GPUs 20-30% below NVIDIA's equivalent products to compensate for the CUDA

ecosystem advantage," and argue that AMD's entry at multi-gigawatt scale with both Meta and OpenAI "introduces genuine price competition for the first time," even while cautioning that "NVIDIA will retain training workload dominance for at least 2-3 more years, supported by CUDA's entrenched software ecosystem and NVLink's superior GPU-to-GPU bandwidth" (Source: introl.com) (Source: introl.com).

For the broader AI infrastructure market, the scale of the commitments detailed in this report, 6 gigawatts each from OpenAI and Meta, up to 2 gigawatts from Anthropic, and 50,000 GPUs from Oracle, signal that AI accelerator procurement has moved decisively from single-generation purchase orders toward multi-year, multi-generation infrastructure pacts with equity components attached. This shift raises durable questions about supply-chain concentration risk (given that both AMD's and Nvidia's most advanced chips still depend on TSMC's most advanced process nodes and CoWoS packaging capacity), about the accuracy of vendor-supplied performance claims that mix precision formats (as the MI455X-versus-Rubin FP4 comparison in this report illustrates), and about power availability, since even a single 72-GPU Helios rack likely draws well over 100 kilowatts continuously, a fact AMD itself has not yet quantified publicly.

Frequently Asked Questions (FAQs)

What is the AMD Instinct MI455X? It is AMD's flagship data-center GPU in its new MI400 series, built on 5th generation CDNA architecture, featuring 432 GB of HBM4 memory and up to 40 PFLOPs of 4-bit AI compute, designed specifically to power the 72-GPU AMD Helios rack-scale platform (Source: chipsandcheese.com).

When was the AMD Instinct MI455X released? AMD unveiled the MI455X and Helios at its Advancing AI 2026 event on July 23, 2026, describing the platform as already "in full production," with customer shipments beginning later in 2026 and ramping into the first half of 2027 (Source: phoronix.com) (Source: phoronix.com).

How much does the AMD Instinct MI455X cost? AMD has not published a public per-unit list price; one early cloud partner lists pricing simply as "TBD" as of mid-2026 (Source: tensorwave.com). The predecessor MI355X currently rents for $2.95 to $8.60 per GPU-hour across cloud providers, offering the closest available pricing proxy until MI455X cloud instances launch (Source: gpus.io) (Source: getdeploying.com).

What is the AMD Helios rack? Helios is AMD's first fully co-designed rack-scale AI infrastructure solution, combining 72 MI455X GPUs, AMD EPYC "Venice" CPUs, and AMD Pensando networking in an Open Compute Project Open Rack Wide form factor, delivering up to 2.9 exaFLOPS of FP4 compute and 31 TB of HBM4 memory (Source: heise.de) (Source: chipsandcheese.com).

How does the AMD Instinct MI455X compare to Nvidia? Independent comparison of each vendor's own published rack specifications shows a mixed picture: AMD's Helios leads on HBM4 capacity (31 TB versus 20.7 TB per rack) and scale-out bandwidth (43 TB/s versus 28.8 TB/s), while Nvidia's Vera Rubin NVL72 claims higher aggregate FP4 throughput (3.6 exaFLOPS versus 2.9 exaFLOPS), making the comparison workload- and precision-format-dependent rather than a uniform advantage for either vendor (Source: nvidia.com) (Source: chipsandcheese.com).

What is the AMD Instinct MI400 series, and how does it differ from the MI450 series naming? The MI400 series is AMD's overall product generation encompassing the MI455X (frontier AI), MI450 Series (the commercial designation used in most customer deal announcements, of which MI455X is a member), and MI430X (sovereign AI and HPC, with FP64 acceleration) (Source: wccftech.com) (Source: ir.amd.com).

Can enterprises buy MI455X directly, or only through cloud providers? As of July 2026, AMD's MI455X capacity is overwhelmingly allocated through multi-gigawatt hyperscaler and frontier-lab agreements (Microsoft, Oracle, OpenAI, Meta, Anthropic); enterprises seeking access will most likely do so indirectly through cloud rental instances such as Azure's upcoming ND MI455X v7 VMs once they reach general availability, or through Microsoft's Azure Foundry Managed Compute preview service, which is explicitly "not currently recommended by Microsoft for production workloads" as of the announcement (Source: techwireasia.com).

Conclusion

The AMD Instinct MI455X, launched alongside the Helios rack-scale platform on July 23, 2026, represents AMD's most architecturally ambitious data-center GPU to date: a 432 GB HBM4, roughly 320-billion-transistor CDNA 5 accelerator built from the outset to operate as one component of a 72-GPU rack rather than a standalone card. Its headline specifications, up to 40 PFLOPs of 4-bit compute, 23.3 TB/s of memory bandwidth, and a rack that AMD claims outperforms Nvidia's Vera Rubin NVL72 by double-digit percentages on compute, memory, and networking, place it as a credible, if not uniformly superior, rival to Nvidia's current flagship rack-scale system. The comparison is genuinely mixed rather than one-sided: AMD leads on memory capacity and scale-out bandwidth while Nvidia's individual GPU claims a higher peak FP4 compute figure, and even AMD's own documentation contains unresolved internal discrepancies, over both the MI455X's exact memory bandwidth rating and, separately, the MI430X's.

Commercially, the MI455X's success will be decided less by any single benchmark than by the execution of the five major deployment commitments detailed in this report, from Microsoft's Azure integration and Oracle's 50,000-GPU supercluster to the OpenAI, Meta, and Anthropic gigawatt-scale agreements that collectively represent tens of billions of dollars in disclosed or estimated contract value. AMD has not published a public MI455X price, has not disclosed the chip's power consumption, and continues to sell the vast majority of this capacity through negotiated hyperscaler contracts rather than open market channels, meaning most enterprise buyers will encounter the MI455X indirectly through cloud instances over the next twelve to eighteen months rather than as a direct purchase. With the MI500 series already confirmed for 2027 and MI600 for 2028, the MI455X marks the opening chapter of what AMD describes as a new annual product cadence, one whose ultimate market impact will depend on ROCm software maturity, actual delivered performance under production workloads, and whether AMD can sustain supply at the scale its 2026 contracts already promise.

Tags: amd instinct mi455x specs, amd instinct mi455x price, amd helios rack specs, amd mi455x vs nvidia, amd instinct mi400 series, amd helios, cdna 5, hbm4, vera rubin nvl72, ai accelerator

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 GPUSmith 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.

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