GPU SMITH
COMPUTE ENGINEERING

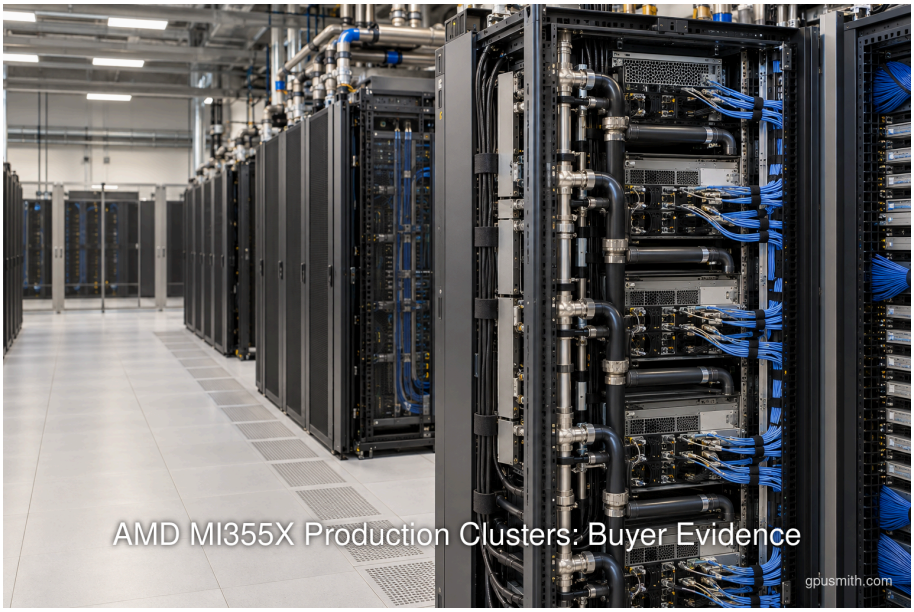
GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

AMD MI355X Production Clusters: Buyer Evidence

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-09-19 · amd mi355x production clusters · mi355x deployment · mi355x cluster · amd instinct · rocm · gpu infrastructure · cisco 800g · nvidia b200 · private ai

Original article: https://gpusmith.com/articles/en/amd-mi355x-production-cluster-analysis



In this article

Executive Summary

As of **September 19, 2026**, enterprise buyers can conclude one important thing about **AMD Instinct MI355X production clusters**: the architecture has crossed from reference designs into a named, customer-serving production deployment. AMD, Cisco, and HUMAIN said on August 31, 2026 that the installation was live in Saudi Arabia and serving customers (newsroom.cisco.com). The disclosed stack combines MI355X graphics processing units (GPUs), AMD EPYC central processing units (CPUs), Cisco Silicon One networking, Nexus 9000 systems, and **800G optics** (newsroom.cisco.com). This is credible integration evidence.

It is not scale, performance, availability, or economic evidence. The release does not publish the installed GPU, node, rack, or megawatt count, nor measured throughput, utilization, uptime, pricing, or cost per token. Its **250 MW beginning in 2027** and **1 GW by 2030** figures describe a future program, not the live MI355X installation (newsroom.amd.com) (newsroom.amd.com). Buyers should therefore treat HUMAIN as proof that the component stack can be integrated and offered as a service, not as proof that a buyer's own workload will scale.

The hardware envelope is concrete. AMD specifies **288 GB HBM3E**, **8 TB/s** memory bandwidth, and **1,400 W maximum board power** per MI355X module (amd.com) (amd.com). An eight-accelerator Universal Baseboard (UBB) therefore represents **11.2 kW of accelerator nameplate power alone** before CPUs, memory, network interface cards, storage, pumps, fans, and conversion losses. Facility input is IT load multiplied by power usage effectiveness (PUE), because the US Environmental Protection Agency defines PUE as total annual source energy divided by IT source energy (portfoliomanager.energystar.gov).

The appropriate decision is **production shortlist, conditional on a controlled pilot**. AMD publishes a node acceptance target of at least **304 GB/s in-place RCCL all-reduce bus bandwidth** and a ten-hour multi-node collective test procedure (instinct.docs.amd.com) (instinct.docs.amd.com). An RFP should turn those documents into contractual, workload-specific gates for burn-in, collectives, error telemetry, link failure, scheduler recovery, security, firmware ownership, and escalation. No public MI355X server price, lead time, inventory level, or general availability commitment was found, so commercial readiness must be established by dated, configuration-specific quotes.

Introduction and Background

The central procurement question is not whether MI355X exists, or whether its peak specifications are competitive. It is whether an enterprise can place MI355X into a request for proposal (RFP) with enough evidence to bound implementation, facility, software, and support risk. The first named live deployment changes that assessment, but only within a precise evidentiary boundary.

The August 31 statement says the infrastructure is live and serving HUMAIN customers, and a September follow-up repeats that status (newsroom.amd.com). This establishes operation beyond a laboratory announcement. It does not establish the installed size, a service-level agreement (SLA), or measured application results.

That distinction matters because private-cluster procurement joins several systems that can each invalidate the whole: accelerator modules, host servers, scale-up links, scale-out fabric, cooling, power distribution, firmware, operating system, ROCm, containers, orchestration, telemetry, and support. Product readiness is therefore not a single binary state. It is an evidence chain from an orderable bill of materials through repeatable workload acceptance.

GPU Smith is an adjacent engineering advisor, not an accelerator or cloud provider. Its stated method spans specification, procurement, integration, and validation (gpusmith.com), and it describes acceptance against written criteria and an as-built documentation set (gpusmith.com). That procurement lens is used here without treating the consultancy as a competing hardware option.

Key Changes

A named production deployment now exists

Before the HUMAIN disclosure, a buyer could point to product pages, reference architectures, OEM listings, and benchmark submissions. The new evidence adds a named operator and a customer-serving service. Cisco says the Nexus 9000 platform interconnects the MI355X GPUs (newsroom.cisco.com), while AMD names EPYC CPUs and Cisco infrastructure in the production stack (newsroom.amd.com). HUMAIN offers that capacity as GPU as a service, with training and inference named as use-case categories (newsroom.cisco.com).

The statement does not identify a model, customer, job size, topology, network interface card (NIC), ROCm version, or scheduler. It also does not disclose whether the training and inference language describes current workload mix or the offered service envelope. The prudent conclusion is that integration is proven at an undisclosed scale, while workload suitability remains a buyer validation question.

Planned capacity is separate from installed capacity

The parties say **up to 250 MW** of additional infrastructure is planned from 2027 (newsroom.amd.com). The source says this expansion uses MI400 Series GPUs, and capacity is expected to start coming online in the second half of 2027 (newsroom.amd.com). A prior 2025 announcement similarly described a **100 MW** first phase as planned and named MI450, not MI355X (newsroom.cisco.com).

Those plans may indicate ecosystem commitment, but they cannot be added to the live deployment or converted into an MI355X GPU count. Any article or RFP that labels 250 MW or 1 GW as today's installed MI355X capacity would mix different generations and time horizons (newsroom.amd.com) (newsroom.amd.com).

The evidence ledger changes the buying posture

Table 1 separates current proof, announced product evidence, and forward-looking statements. Confidence refers to what the cited source actually supports, not to a prediction about execution.

Claim class	What the record supports	What remains unproven	Buyer confidence
Proven in a named deployment	MI355X, EPYC, Cisco Silicon One, N9000, and 800G optics form a live customer-serving stack (newsroom.cisco.com).	Installed count, topology, performance, uptime, power, and economics.	**High for integration, low for extrapolation**
Published product evidence	MI355X provides 288 GB HBM3E, 8 TB/s, and an eight-OAM UBB design (amd.com).	Sustained workload throughput and multi-node scaling for a buyer's stack.	**High for specification, conditional for workload**
Published acceptance evidence	AMD defines a repeatable method for configuration, validation, benchmarking, and baselining (instinct.docs.amd.com).	Contracted thresholds, version freeze, failure budget, and joint sign-off.	**High as a starting template**
Forward-looking plan	Up to 1 GW is targeted by 2030 (newsroom.amd.com).	Timing, delivered capacity, and relationship to the live MI355X system.	**Not current production evidence**

The ledger supports placing MI355X on a shortlist, but not waiving a proof of concept. The strongest evidence is architectural and procedural. The weakest evidence is commercial and operational at fleet scale.

Cluster Architecture and Availability

Accelerator and node layer

MI355X is an Open Compute Project Accelerator Module (OAM), not a standalone PCI Express add-in card. AMD's UBB 2.0 platform integrates **eight OAM accelerators** (amd.com). Each GPU has seven bidirectional Infinity Fabric links rated at **153.6 GB/s each**, forming a fully connected eight-GPU node (rocmdocs.amd.com). AMD describes aggregate intra-node communication as more than **1 TB/s per GPU** (rocmdocs.amd.com).

The reference node configuration calls for dual EPYC 9004 or 9005 processors and high-speed back-end networking (asrockrack.com). These are reference choices, not a disclosure of HUMAIN's exact host or NIC models.

Scale-out network layer

The live deployment disclosure names Cisco's N9000 family, Silicon One, and 800G optics, but no switch SKU, ASIC, port count, topology, rail mapping, or oversubscription. Cisco's wider portfolio supports 400G and 800G optics in QSFP-DD and OSFP form factors (cisco.com). One available Nexus 9364E-SG2 design has 64

800GbE ports, but the HUMAIN release never says this is the deployed model (cisco.com).

AMD's separate cluster reference says the choice between rail-optimized and fat-tree designs depends on workload communication patterns (instinct.docs.amd.com). RFPs should therefore demand an explicit topology diagram and measured collective curves, not a generic claim of 800G capability.

OEM and software availability

There is meaningful ecosystem evidence. Dell lists eight 288 GB, 1,400 W MI355X OAMs in PowerEdge XE9785 systems (dell.com). Supermicro lists both a 4U liquid-cooled eight-GPU system and a 10U air-cooled system (supermicro.com) (supermicro.com). GIGABYTE publishes a 64-GPU liquid-cooled rack configuration with a stated maximum rack consumption of **120 kW** (gigabyte.com) (gigabyte.com).

Canonical lists a MiTAC MI355X server in its certified catalog, and Red Hat lists Supermicro's AS-A126GS-TNMR as working with Red Hat Enterprise Linux (ubuntu.com) (catalog.redhat.com). These listings show a maturing supply and certification ecosystem, not public inventory. None of the reviewed official pages supplied a public price, guaranteed lead time, committed delivery date, or inventory level.

Implementation Considerations and Process Changes

Freeze a reproducible software image

AMD's current acceptance prerequisite is **ROCm 7.0.1 or later** (instinct.docs.amd.com). A minimum version is not a production image. The contract should freeze the server BIOS, baseboard management controller, GPU firmware, NIC firmware, operating system kernel, AMDGPU driver, ROCm user space, collective library, container digest, framework build, scheduler, and orchestration manifests.

Current ecosystem artifacts make this feasible. AMD's GPU Operator hardware matrix lists MI355X (instinct.docs.amd.com). AMD also documents a `rocm/primus:v26.7` PyTorch training image for MI355X (rocm.docs.amd.com), while its JAX guide specifies a `maxtext-v26.2` image (rocm.docs.amd.com). Image names should be resolved to immutable digests at acceptance.

Make observability and remediation testable

AMD System Management Interface exposes error-correcting code (ECC) counts and Common Platform Error Record data (rocm.docs.amd.com). Its cluster checker collects logs and metrics into a health report and checks RAS, PCIe, XGMI, and network counters (rocm.docs.amd.com) (rocm.docs.amd.com). These capabilities become operational evidence only when alert thresholds, retention, dashboards, and escalation owners are specified.

The GPU Operator remediation flow can taint affected nodes and validate health after reboot (instinct.docs.amd.com). It does not by itself guarantee application checkpointing or distributed-job restart. Scheduler, framework, and application responsibilities must be separately accepted.

Buyer Verification and RFP Gates

Assign support ownership

A cluster crosses organizational boundaries. AMD says the server or infrastructure provider publishes GPU and baseboard firmware bundles (rocm.docs.amd.com). It advises using firmware qualified for the exact GPU model and escalating missing platform firmware to the OEM or AMD support (instinct.docs.amd.com). Cisco owns the fabric and optics in the HUMAIN disclosure, but the end-to-end service boundary is not public.

Table 2 is a minimum support ownership matrix. The final RFP should name legal entities, response times, evidence handoffs, and one incident coordinator.

Layer	Primary acceptance artifact	Owner to name in RFP	Handoff test
GPU and UBB	Serial inventory, firmware manifest, field-health JSON	Server OEM plus AMD	Reproduce health pass and one injected error alert
Host	BIOS, BMC, CPU, memory, PCIe topology	Server OEM	Verify every GPU at its contracted PCIe speed and width, with no fatal flag
Scale-up fabric	Infinity Fabric topology and collective baseline	OEM plus AMD	RCCL curve across message sizes, not one peak point
Scale-out fabric	Port map, optics, NIC firmware, routing, congestion policy	Network integrator plus Cisco	Link loss, reroute, congestion, and oversubscription evidence
Cooling and power	CDU, flow, temperature, PDU, and breaker schedules	Facility engineer plus OEM	Worst-case thermal soak with alarms and safe shutdown
ROCm and containers	Versioned software bill and immutable image digests	Platform team plus AMD	Rebuild a node from approved artifacts
Scheduler and applications	Checkpoint, retry, queue, and data-integrity policy	Buyer application owner	Kill a node during a representative distributed job
Operations	Dashboards, retention, runbooks, and escalation tree	Managed service or buyer operations	Timed detect, isolate, remediate, and return-to-service exercise

The matrix prevents a common ambiguity: component warranties do not create an end-to-end recovery objective. A single escalation coordinator and a clock that continues across vendor handoffs should be contractual.

Turn the pilot into measured acceptance

Table 3 defines a practical acceptance matrix. Thresholds should be set from the buyer's workload and bid configuration. AMD's published values are useful baselines, not universal promises.

Gate	Method	Evidence required	Pass logic
Inventory and image	Reconcile serials, firmware, packages, containers, and configuration	Signed machine-readable manifest	Exact match to released build
Burn-in	Exercise PCIe, HBM, compute, power, thermals, and fabric	Per-GPU logs and JSON; AMD says each run generates both (instinct.docs.amd.com)	Zero unresolved failures before sign-off
Intra-node collectives	RCCL all-reduce over multiple sizes and iterations	Median, tail, variance, and topology	Meet the contracted curve across sizes
Multi-node collectives	All-reduce, all-gather, reduce-scatter, and all-to-all	Per-node and per-rail endurance results	No outlier rail; stable performance over time
ECC and error telemetry	Observe correctable, uncorrectable, replay, XGMI, and NIC counters	Dashboard, alert, ticket, and retained raw event	Alert reaches named owner within contracted time
Link failure	Disable one link or port under load	Reroute trace, job impact, recovery, and counter evidence	Behavior matches topology and failure budget
Job recovery	Remove one node during a representative job	Checkpoint, restart time, completed output, and scheduler record	Recovery point and time objective met
Security	Test role-based access control, TLS, image trust, and Secure Boot	Access logs and signed modules	Unauthorized metrics access denied; approved image runs
Support escalation	Open a realistic multi-vendor case	Timeline and evidence-transfer record	One owner maintains the clock through closure

AMD's broader guide specifies a **ten-hour** multi-node RCCL test and validates GPU-to-NIC, NIC-to-NIC through the switch, and GPU-to-GPU through the switch (instinct.docs.amd.com). The buyer should retain raw outputs, not only a vendor certificate. GPU Smith's own adjacent-advisor description emphasizes architecture assessment and vendor-quote verification (gpusmith.com), which is the appropriate role for independent review here.

The final RFP evidence package should be explicit:

- **Deployment manifest:** serialized hardware and installed location.
- **Bill of materials:** orderable parts, substitutions, and spare quantities.
- **Topology:** physical and logical port maps with rail assignments.
- **Firmware baseline:** signed versions for every managed component.
- **Software bill:** operating system, driver, ROCm, libraries, and frameworks.
- **Container record:** immutable digests, build inputs, and registry provenance.
- **Facility schedule:** power, cooling, weight, floor, and piping requirements.
- **Collective curves:** raw results by operation and message size.

- **Workload baseline:** commands, data, quality target, and completion time.
- **Power boundary:** meters, sampling interval, and included equipment.
- **Health evidence:** per-GPU logs, machine-readable summaries, and counters.
- **Failure evidence:** link, port, GPU, and node interruption outcomes.
- **Recovery evidence:** checkpoint age, restart time, and output verification.
- **Security evidence:** roles, certificates, image trust, and access logs.
- **Support map:** named owner, response clock, and evidence handoff.
- **Change control:** revalidation triggers and approved rollback procedure.
- **Acceptance record:** exceptions, remedies, signatures, and as-built package.

Data Analysis and Evidence

Power and facility scenarios

At **1.4 kW per GPU**, accelerator-only theoretical power is based on AMD's maximum module rating (rocmdocs.amd.com):

```
accelerator nameplate kW = GPU count × 1.4
```

An eight-GPU node is therefore **11.2 kW**, 64 GPUs are **89.6 kW**, 256 GPUs are **358.4 kW**, and 1,024 GPUs are **1,433.6 kW** before the rest of the IT equipment. These are arithmetic scenarios based on AMD's module maximum, not measured cluster draw. AMD describes MI355X as a 1,400 W direct-liquid-cooled OAM (rocmdocs.amd.com). Dell separately offers a factory 1 kW cap, showing that bid configurations can alter the envelope (dell.com). Performance under that cap must be measured.

For total facility input, use:

```
facility input kW = total IT load kW × PUE
```

If a complete rack measures 120 kW at IT input, PUE scenarios of 1.2, 1.4, and 1.6 imply **144 kW**, **168 kW**, and **192 kW** of facility source power. The Lawrence Berkeley National Laboratory guide uses **1.6** as an average-data-center reference and cautions that PUE does not measure whole-data-center efficiency (datacenters.lbl.gov). The RFP should model the facility's measured seasonal PUE, coolant-distribution-unit overhead, redundancy state, and utility limits rather than adopt 1.6 as a forecast.

Server and rack specifications illustrate why GPU multiplication is only a lower bound. Dell lists a **33 kW power shelf** for an eight-GPU XE9785L, while Supermicro provisions four 6,600 W supplies in a redundant 2+2 arrangement (dell.com) (supermicro.com). Provisioned supply capacity is not measured consumption, but it is relevant to breakers, busways, and redundancy design.

Network bandwidth worksheet

The reference eight-NIC node has **3.2 Tb/s nominal aggregate line rate**, or 400 Gb/s per GPU under a one-to-one mapping (instinct.docs.amd.com). That number is before Ethernet framing, protocol overhead, congestion, and oversubscription. A useful worksheet contains:

- **GPU count:** total accelerators participating in the job.
- **NIC count and line rate:** physical adapters per node and configured speed.
- **Rail mapping:** which GPU, CPU socket, PCIe root, and NIC share a path.
- **Downlink capacity:** active NIC ports multiplied by negotiated line rate.
- **Uplink capacity:** leaf-to-spine links multiplied by negotiated line rate.
- **Subscription ratio:** uplink capacity divided by downlink capacity.
- **Useful collective bandwidth:** measured bus bandwidth across message sizes.
- **Tail behavior:** slowest node, rail, and iteration, not only the median.

Cisco's reference example uses equal aggregate downlink and uplink capacity for a **1:1 leaf ratio** (cisco.com). An RFP should define the numerator, denominator, failure state, and traffic pattern so terminology cannot hide a different calculation.

Benchmark evidence and its limit

MLPerf Training v6.0 includes single-node results from systems with eight MI355X accelerators and three large-language-model workloads (mlcommons.org). An official result log records a successful Llama 3.1 8B training run of **94.99 minutes** (github.com). MLCommons explicitly says RoCE was not used for communication in that single-node submission (mlcommons.org).

This is useful node-level software and compute evidence. It does not validate Cisco scale-out networking, multi-node collective efficiency, HUMAIN's service, or a buyer's model. No public benchmark found in this research closes those gaps. A buyer should run the same container and dataset on one node, then extend the identical workload across nodes and report scaling efficiency relative to that baseline.

Comparative Context: MI355X and B200 Clusters

A responsible **MI355X versus NVIDIA B200 cluster** comparison begins with system boundaries. NVIDIA documents **180 GB HBM3E per B200 GPU**, while AMD specifies 288 GB per MI355X (docs.nvidia.com). NVIDIA says B200 is configurable up to **1 kW per GPU**, compared with MI355X's 1.4 kW maximum board power (docs.nvidia.com). Those are component limits, not energy per completed job.

At the eight-GPU system level, NVIDIA reports **14.4 TB/s aggregate NVLink bandwidth** and **14.3 kW maximum system input** for DGX B200 (docs.nvidia.com) (docs.nvidia.com). AMD reports **64 TB/s peak aggregate HBM bandwidth** for an eight-GPU MI350-series platform, a different metric that must not be compared to NVLink bandwidth (amd.com).

The buying comparison should use matched workloads, precision, software versions, model quality, batch and sequence lengths, node counts, networking, power measurement boundaries, and availability terms. Without these controls, memory capacity and nameplate power can frame the experiment but cannot select a production cluster.

Implications and Future Directions

MI355X has enough evidence for an enterprise RFP in 2026. The live HUMAIN service, multiple OEM systems, operating-system certifications, ROCm support, containers, health tooling, and an AMD acceptance template collectively move it beyond a paper product. HPE's XD685 data sheet, for example, names MI355X in a 5U direct-liquid-cooled chassis and offers factory integration and validation services (hpe.com). ASRock Rack also publishes a 4U MI355X system supporting EPYC 9005 and 9004 processors (asrockrack.com).

The evidence still favors staged commitment:

- **RFP now:** solicit configuration-specific bids, topology, software bill, support matrix, facility schedule, and acceptance plan.
- **Pilot before volume:** test the buyer's largest model, representative data path, collective mix, checkpoint behavior, and failure recovery.
- **Option volume, do not assume it:** link production expansion to measured gates and dated commercial commitments.
- **Preserve benchmark artifacts:** retain containers, commands, raw logs, telemetry, and power-meter boundaries.
- **Revalidate after changes:** firmware, kernel, ROCm, collective libraries, NIC code, and switch software can change the accepted system.

Commercial evidence should mature next. Public product and certification pages do not answer price, lead time, allocation, spare policy, or service response. The planned MI400 expansion may broaden the ecosystem, but it should not be used to underwrite MI355X residual value or current delivery.

Operational evidence should also become more specific. AMD's public test runner supports ROCm Validation Suite recipes, while the AMD GPU Field Health Check image is separately licensed (instinct.docs.amd.com). Buyers should make access, license, version, delivery, and support for restricted acceptance artifacts explicit before sign-off.

Frequently Asked Questions (FAQs)

What does the AMD MI355X production deployment prove?

Yes, in the bounded sense that AMD, Cisco, and HUMAIN named a live Saudi deployment serving customers on August 31, 2026 (newsroom.amd.com). Public evidence does not disclose its size, measured performance, availability, or economics. That makes it production integration evidence, not a substitute for a buyer-specific pilot.

What is AMD MI355X availability from OEMs?

Multiple official OEM pages list MI355X systems, including Dell, HPE, Supermicro, GIGABYTE, and ASRock Rack. GIGABYTE publishes an ordering number for its 4U liquid-cooled server (gigabyte.com). A product page or configurator does not prove inventory or lead time. Require a dated quote with exact parts, delivery milestones, substitutions, spares, and acceptance remedies.

What AMD MI355X cluster performance and benchmark results exist?

Public **AMD MI355X benchmark results** include MLPerf single-node workload evidence. AMD also publishes theoretical accelerator and platform specifications, an intra-node RCCL acceptance floor, and reference network designs. The HUMAIN announcement supplies no measured result, and the located MLPerf evidence does not test RoCE scale-out. No public evidence reviewed here establishes HUMAIN cluster throughput or multi-node scaling efficiency.

What does AMD MI355X production readiness require?

It should include the final server and NIC configuration, frozen production image, representative model and dataset, one-node baseline, multi-node scaling, at least ten hours of collectives, thermal soak, telemetry, link failure, node loss, checkpoint recovery, security controls, and a multi-vendor support exercise. Every result should be machine readable and reproducible.

How should an AMD MI355X cluster buying guide estimate power?

Multiply GPU count by 1.4 kW for an accelerator-only nameplate scenario (rocmdocs.amd.com). Add measured or vendor-qualified consumption for hosts, NICs, storage, fans, pumps, and conversion. Then multiply total IT load by the site's PUE for facility input. Do not divide by PUE, and do not present nameplate arithmetic as workload energy.

How do AMD MI355X vs Nvidia B200 clusters compare?

The public specifications alone cannot answer that question. MI355X provides more HBM capacity per GPU in the cited documents, while B200 has a lower configurable per-GPU maximum. Production selection depends on matched application throughput, model quality, scaling, power per completed job, software effort, availability, support, and total cost. A controlled bake-off is the defensible method.

What defines AMD MI355X cluster architecture?

The baseline is an eight-OAM UBB node with Infinity Fabric scale-up links, EPYC hosts, and a scale-out network built from validated 400G or 800G components (amd.com) (cisco.com). The procurement architecture must go further by naming PCIe locality, NIC rail mapping, leaf and spine models, optics, subscription ratio, storage paths, management network, cooling, power, and the exact software image.

Conclusion

The first named live HUMAIN deployment changes MI355X procurement from a speculative architecture review to an evidence-backed production shortlist. It proves that MI355X, EPYC, and the disclosed Cisco fabric can be assembled into a customer-serving service. It does not disclose the scale, service level, workload performance, availability, price, or economics needed to underwrite a large private cluster.

Enterprise buyers should therefore issue an RFP, but make award and volume conditional. The bid should freeze the bill of materials and software image, state topology and oversubscription, assign end-to-end support, price the facility envelope, and bind acceptance to reproducible workload and failure tests. AMD's published collective thresholds, endurance procedures, telemetry interfaces, and remediation tooling provide a useful starting point. They are inputs to a contract, not the contract itself.

The appropriate conclusion as of September 2026 is neither blanket readiness nor immaturity. **MI355X is mature enough for an RFP and controlled production pilot, but public evidence is not yet sufficient to skip buyer-specific proof or commit fleet-scale capital without conditional acceptance.**

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m08/amd-cisco-and-humain-expand-saudi-arabia-ai-infrastructure.html
2. https://newsroom.amd.com/news/leap-2026-advancing-ai-across-the-middle-east/
3. https://newsroom.amd.com/news/amd-cisco-humain-expand-saudi-arabia-ai-infrastructure/
4. https://www.amd.com/en/products/accelerators/instinct/mi350/mi355x.html
5. https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/amd-instinct-mil355x-platform-brochure.pdf
6. https://gpusmith.com/articles/en/gpu-rack-power-requirements-planning-guide
7. https://portfoliomanager.energystar.gov/pm/glossary
8. https://gpusmith.com/articles/en/gpu-cluster-acceptance-testing-runbook
9. https://instinct.docs.amd.com/projects/system-acceptance/en/latest/gpus/mi355x.html
10. https://instinct.docs.amd.com/projects/system-acceptance/en/latest/index.html
11. https://gpusmith.com/
12. https://gpusmith.com/articles/en/amd-instinct-mi455x-specs-price
13. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2025/m11/amd-cisco-and-humain-to-form-joint-venture-to-deliver-world-leading-ai-infrastructure.html
14. https://www.amd.com/en/products/accelerators/instinct/mi350.html
15. https://gpusmith.com/articles/en/nvlink-vs-infinity-fabric-gpu-interconnects
16. https://rocmdocs.amd.com/en/latest/reference/gpu-arch/mi350.html
17. https://www.asrockrack.com/general/productdetail.de.asp?Model=4U8M-TURIN2%2FDLC+SYN+MI355X
18. https://www.cisco.com/site/us/en/products/networking/cloud-networking-switches/nexus-9000-switches/index.html

19. https://www.cisco.com/c/en/us/products/collateral/networking/cloud-networking-switches/nexus-9000-switches/nexus-9000-ai-networking-wp.html
20. https://instinct.docs.amd.com/projects/MI3XX-reference/latest/index.html
21. https://www.dell.com/en-ca/shop/ipovw/poweredge-xe9785
22. https://www.supermicro.com/en/products/system/datasheet/as-4126gs-nmr-lcc
23. https://www.supermicro.com/en/products/system/gpu/10u/as-a126gs-tnmr
24. https://www.gigabyte.com/Enterprise/GIGAPOD-Rack-Scale/AI-DLC-Rack_AMD-Instinct-MI355X
25. https://ubuntu.com/certified/servers?offset=900
26. https://catalog.redhat.com/en/hardware/system/detail/310147
27. https://instinct.docs.amd.com/projects/gpu-operator/en/main/
28. https://rocm.docs.amd.com/projects/primus/en/latest/02-user-guide/megatron-lm-training.html
29. https://rocm.docs.amd.com/en/docs-7.2.0/how-to/rocm-for-ai/training/benchmark-docker/jax-maxtext.html
30. https://rocm.docs.amd.com/projects/amdsmi/en/develop/conceptual/ras.html
31. https://rocm.docs.amd.com/projects/cvs/en/docs/how-to/run-cluster.html
32. https://instinct.docs.amd.com/projects/gpu-operator/en/main/autoremediation/auto-remediation.html
33. https://rocm.docs.amd.com/en/docs-7.2.4/about/release-notes.html
34. https://instinct.docs.amd.com/projects/system-acceptance/en/latest/common/firmware-updates.html
35. https://instinct.docs.amd.com/projects/system-acceptance/en/latest/common/system-validation.html
36. https://gpusmith.com/articles/en/gpu-data-center-cooling-design
37. https://www.dell.com/support/manuals/en-in/poweredge-xe9785/xe9785_ism/power-capping-configuration-with-mi355x-gpu?guid=guid-73e7b829-65b2-47bb-9c88-9b73659eb35e&lang=en-us
38. https://datacenters.lbl.gov/sites/default/files/2025-07/best-practice-guide-data-center-design.pdf
39. https://www.dell.com/en-us/shop/ipovw/poweredge-xe9785l
40. https://gpusmith.com/articles/en/data-center-electrical-design-gpu-clusters
41. https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-743245.pdf
42. https://mlcommons.org/wp-content/uploads/2026/06/Final-MLPerf-Training-v6.0-Supplemental-Discussion-UNDER-EMBARGO-UNTIL-6_16_26-8_00-AM-PT.pdf
43. https://github.com/mlcommons/training_results_v6.0/blob/main/HPE/results/HPE-Compute-XD685-MI355X-288GBx8/llama31_8b/result_1.txt
44. https://docs.nvidia.com/enterprise-reference-architectures/hgx-ai-factory-h100-h200-b200/latest/components.html
45. https://docs.nvidia.com/dgx/dgxb200-user-guide/introduction-to-dgxb200.html
46. https://www.hpe.com/psnow/generateDDS/HPE%20ProLiant%20Compute%20XD685%20data%20sheet-PSN1014862906WWEN.pdf?cc=WW&contentDisposition=attachment&isLinearized=false&lc=EN&oid=1014862906&prelaunch=false&softroll=0
47. https://instinct.docs.amd.com/projects/gpu-operator/en/release-v1.4.0/releasenotes.html
48. https://www.gigabyte.com/Enterprise/GPU-Server/G4L3-ZX1-LAT4

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com