PRIVATE AI INFRASTRUCTURE

xSONiC AI Inference Server

AMD Instinct MI355X and MI300X platform options for private LLM, RAG, multimodal and enterprise AI services.

PRODUCT OVERVIEW

AMD Instinct-class inference capacity, integrated by xSONiC.

xSONiC AI Inference Server integrates AMD Instinct MI355X and MI300X GPU platform options with high-density HBM memory, PCIe Gen5 host connectivity, GPU-to-GPU fabric and deployment services. It is built for organizations running private assistants, enterprise search, document intelligence, coding support and multimodal workflows where data locality, predictable throughput and operational ownership matter.

8 GPUs
OAM high-density accelerator platform

2.3 TB
HBM3E memory pool with MI355X option

Fabric
Infinity Fabric for scale-up inference

PCIe 5
Gen5 host I/O with ROCm ecosystem support

Accelerated inference node

Designed as the compute building block for AMD Instinct MI355X or MI300X based private model service, AI fabric integration, storage access and observability.

Compute Node
Dense GPU acceleration and large model memory residency.

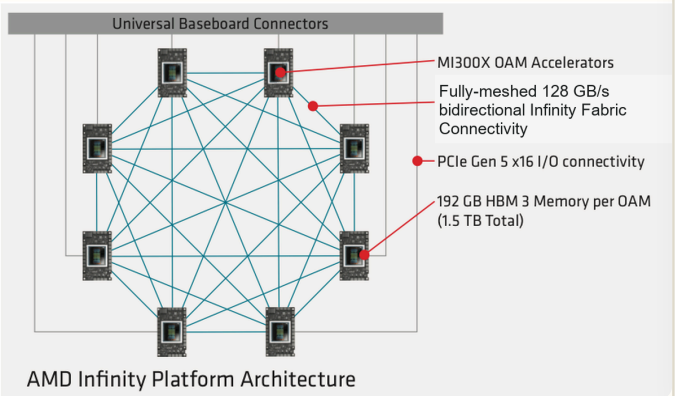
AI Fabric
Low-latency switching, optics and rack-scale connectivity.

Data Layer
Model, vector and enterprise content storage paths.

Service API
Routing, batching, monitoring and application access.

MI300X scale-up GPU fabric

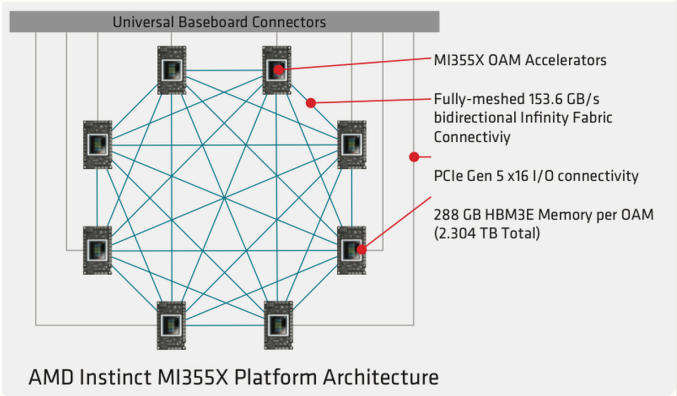
Universal baseboard architecture with fully meshed 128 GB/s bidirectional Infinity Fabric links.



AMD Infinity Platform Architecture

MI355X next-generation fabric

Updated OAM architecture with 153.6 GB/s links and a 288 GB HBM3E memory pool per accelerator.



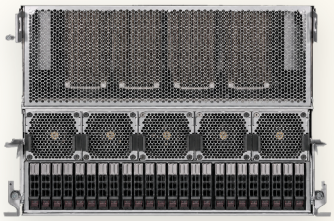
AMD Instinct MI355X Platform Architecture

PLATFORM OPTIONS + WORKLOADS

GPU platform specification and model ecosystem.

AMD remains the accelerator platform foundation while xSONiC owns the integrated server, deployment, support, network integration and solution packaging story.

Area	MI355X Platform Option	MI300X Platform Option
GPU configuration	8 AMD Instinct MI355X OAM GPUs on UBB 2.0 module.	8 AMD Instinct MI300X OAM GPUs on UBB 2.0 module.
Architecture	4th Gen AMD CDNA, TSMC 3nm/6nm FinFET.	3rd Gen AMD CDNA, 5nm/6nm FinFET.
GPU memory	288 GB HBM3E per GPU / approx. 2.3 TB total.	192 GB HBM3 per GPU / 1.5 TB total.
Memory bandwidth	Up to 8 TB/s per GPU.	Up to 5.3 TB/s per GPU.
GPU-to-GPU links	7 x 153.6 GB/s bidirectional Infinity Fabric links per GPU.	7 x 128 GB/s bidirectional Infinity Fabric links per GPU.
Host I/O	8 PCIe Gen5 x16 connections.	8 PCIe Gen5 x16 connections.
Precision support	FP16, BF16, FP8, MXFP6, MXFP4 for efficient inference and training.	FP32/FP64 for HPC plus FP16/BF16/FP8/INT8 for AI.
Software ecosystem	ROCm 7.0, PyTorch, TensorFlow, JAX, ONNX Runtime.	ROCm 6, PyTorch, TensorFlow, ONNX Runtime, Triton, JAX.



High-density chassis for MI355X or MI300X platform integration.

The server form factor supports dense accelerator deployments with large intake area, serviceable fan modules and high drive density for model and data paths.

Workload	Primary compute requirement	Deployment consideration
LLM serving	High token throughput, batching, FP8/BF16 inference and multi-user concurrency.	Capacity plan around latency target, context length, concurrency and model size.
RAG and search	Embedding, reranking and generation pipelines with model and vector data paths.	Coordinate GPU nodes with storage, retrieval services and private knowledge access.
Document intelligence	Long-context inference, OCR/vision-language model support and batch processing.	Keep documents, prompts and extracted knowledge inside controlled infrastructure.
Multimodal AI	Vision-language inference and mixed workload scheduling across accelerator memory.	Validate image/text data flow, API integration and monitoring before production.

Model ecosystem

Llama, Qwen, DeepSeek, Mistral, GLM and other mainstream model families for assistants, search and enterprise workflows.

Service readiness

- Token throughput planning
- Long-context scheduling
- Multi-user concurrency

Private deployment

Keep models, prompts and enterprise knowledge inside a controlled infrastructure domain.

Scale-out path

Pair compute nodes with xSONiC switching, optics, packet visibility and storage design services.

ALL-IN-ONE INFERENCE SERVICE STACK

From large-model compute to deployable private AI services.

The platform is positioned as a hardware-to-service-stack foundation: high-throughput token generation, large-context memory residency, standard APIs, observability, cost planning inputs and tuning for enterprise AI applications.

Capability layer	What the platform provides	Production requirement
Accelerator compute	8-GPU AMD Instinct platform options with large HBM memory pools and high memory bandwidth.	Size GPU count, memory capacity and precision mode against model family, batch size and latency target.
Model serving	Inference runtime path for mainstream LLM, code, search, RAG and multimodal model services.	Validate model compatibility, tokenizer behavior, concurrency profile and expected service API.
Data and retrieval	Support for model files, vector data, enterprise knowledge sources and retrieval service integration.	Plan storage, access control, retrieval latency and data locality before production deployment.
Operations	Deployment planning, telemetry, health monitoring, tuning, cost-per-inference inputs and lifecycle support around the GPU node.	Define monitoring metrics, failure response, target service objectives, update process and capacity expansion path.

Supported mainstream models

Llama, Qwen, DeepSeek, Mistral, GLM and other mainstream model families for assistants, search and enterprise workflows.

RAG retrieval stack

Embedding and reranking models support retrieval-augmented generation, enterprise search and document intelligence.

Multimodal inference

Vision-language inference supports image-text understanding, review workflows and multimodal generation tasks.

High-throughput tokens

Multi-user concurrency, batching and accelerator memory bandwidth improve useful throughput for production service loads.

Long-context support

Large HBM capacity supports longer context windows, larger batches and memory-intensive AI workloads.

API-ready services

Standard interfaces make inference service integration easier for existing applications and workflow systems.

TCO planning

Use large HBM capacity, low-bit precision paths, utilization tuning and private deployment assumptions as inputs to a TCO model; actual results depend on workload mix, power, cooling and operations.

Cost per inference

Plan cost per request or token by validating batching, concurrency, precision mode and model size on target workloads before committing to a fixed production baseline.

SLA readiness

Support production service objectives with endpoint health checks, monitoring, runbooks and defined support boundaries; final SLA targets should be agreed and tested during deployment.

Deployment item	Planning detail	Acceptance check
Network fabric	Define GPU node uplinks, east-west traffic path, RoCE/Ethernet policy and service ingress.	Latency, packet loss and bandwidth profile validated under model-serving load.
Model storage	Plan model file storage, vector data path, knowledge-source access and backup policy.	Model load time, retrieval latency and access-control behavior verified.
Runtime stack	Select ROCm runtime, serving framework, API gateway, batching and health-check method.	Service endpoint, failover behavior and monitoring metrics confirmed.
Operations	Document update process, capacity expansion trigger, cost baseline, incident response and support boundary.	Runbook, ownership model and target service objectives accepted before production handoff.

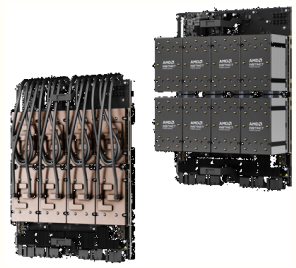
PLATFORM EVIDENCE

AMD Instinct platform renders and expanded technical indicators.

Additional platform images and technical rows carry deeper MI355X and MI300X evidence instead of only a short summary table.

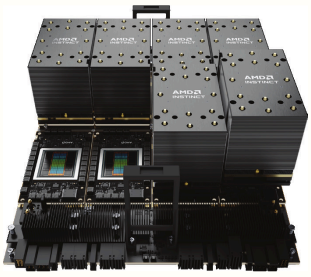
MI355X liquid-cooled platform option

Eight MI355X OAM accelerators, 288 GB HBM3E memory per OAM, 153.6 GB/s fabric links and direct liquid cooling support.



MI300X UBB 2.0 accelerator platform

Eight OAM accelerators, 192 GB HBM3 memory per OAM, 128 GB/s bidirectional fabric links and PCIe Gen5 host I/O.



Detail	MI355X Platform Option	MI300X Platform Option
Form factor	UBB 2.0 module with 8 AMD Instinct MI355X OAM GPUs.	Universal Baseboard UBB 2.0 module with 8 AMD Instinct MI300X OAM GPUs.
Compute units / cores	2,048 GPU compute units, 8,192 matrix cores and 131,072 stream processors.	2,432 GPU compute units, 9,728 matrix cores and 155,648 stream processors.
Engine / cache	Peak engine clock 2,400 MHz; 256 MB AMD Infinity Cache per GPU.	Peak engine clock 2,100 MHz; 256 MB AMD Infinity Cache per GPU.
Memory system	288 GB HBM3E per OAM, approx. 2.304 TB total; 8 TB/s per GPU maximum theoretical bandwidth.	192 GB HBM3 per OAM, 1.5 TB total; 5.3 TB/s per GPU maximum theoretical bandwidth.
Scale-up fabric	7 bidirectional AMD Infinity Fabric links per GPU at 153.6 GB/s.	7 bidirectional AMD Infinity Fabric links per GPU at 128 GB/s; ring of 8 aggregate bandwidth 896 GB/s.
Host I/O	8 PCIe Gen5 x16 connections to host CPU.	8 PCIe Gen5 x16 connections, 128 GB/s per GPU scale-out network bandwidth.
Virtualization / RAS	SR-IOV up to 64 GPU partitions, 1 or 4 memory partitions per module, ECC memory, page retirement and page avoidance.	SR-IOV up to 64 partitions, full-chip ECC memory, page retirement and page avoidance.
Video decode	32 groups for HEVC/H.265, AVC/H.264, VP9 or AV1; JPEG/MJPEG codec support.	32 groups for HEVC/H.265, AVC/H.264, VP9 or AV1; JPEG/MJPEG codec support.
Power / cooling	Direct liquid cooled platform option; PG-25 coolant, 43 C maximum liquid inlet temperature, 2.1 l/min recommended flow per OAM, up to 1400W module TBP.	750W maximum TBP per GPU in the platform specification.

ROCm software environment

ROCm provides an open software platform for AI and HPC deployment paths.

- PyTorch, TensorFlow and JAX support
- ONNX Runtime and Triton service paths

Private AI infrastructure fit

The GPU node pairs with xSONiC infrastructure services for private AI production rollout.

- Switching, optics and packet visibility
- Storage design and deployment planning

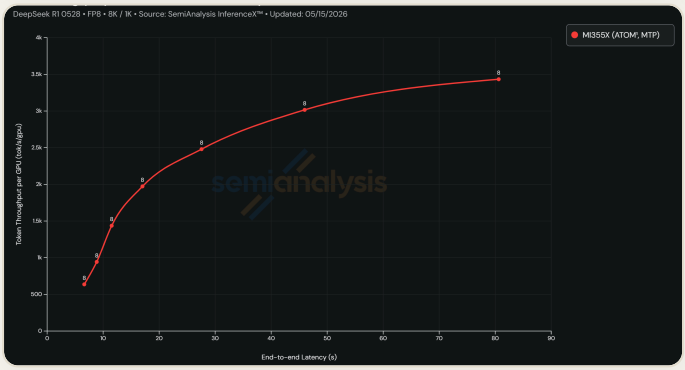
FROM GPU THROUGHPUT TO PRODUCTION AI INFRASTRUCTURE

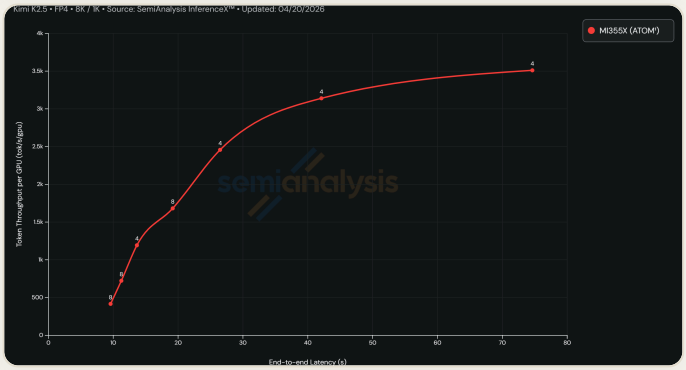
AMD Instinct performance options, delivered as a private AI platform.

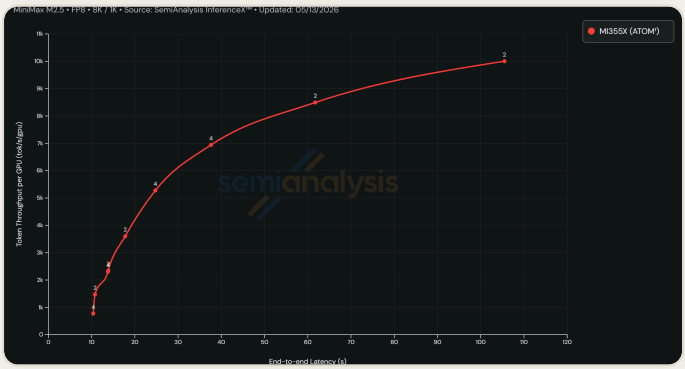
The server supplies AMD Instinct MI355X or MI300X GPU compute and inference services; xSONiC switching, optics, packet visibility, storage and deployment planning complete the path from rack design to production operations.

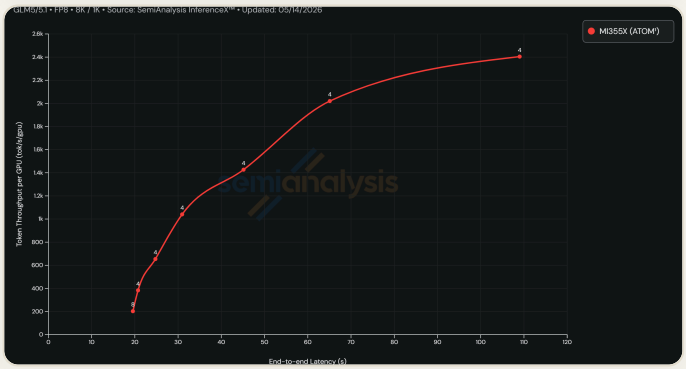
MI355X Peak Metric	Peak	w/ sparsity
FP16 vector	1,258.4 TFLOPS	N/A
FP16 matrix	20.1328 PFLOPS	40.2656
BFLOAT16 matrix	20.1328 PFLOPS	40.2656
INT8 matrix	40.2656 POPS	80.5312
MXFP8	40.2656 PFLOPS	N/A
OCP-FP8	40.5312 PFLOPS	80.5312
MXFP6	80.5304 PFLOPS	N/A
MXFP4	80.5304 PFLOPS	N/A

MI300X Peak Metric	Peak	w/ sparsity
TF32	5.2 PFLOPS	10.5 PFLOPS
FP16	10.5 PFLOPS	20.9 PFLOPS
BFLOAT16	10.5 PFLOPS	20.9 PFLOPS
INT8	20.9 POPS	41.8 POPS
FP8	20.9 PFLOPS	41.8 PFLOPS
FP64 vector	653.7 TFLOPS	N/A
FP32 vector	1,307.4 TFLOPS	N/A
FP64 / FP32 matrix	1,307.4 TFLOPS	N/A









Source: SemiAnalysis InferenceX, github.com/SemiAnalysisAI/InferenceX. Original benchmark source markings are retained.

Compute
Accelerator nodes for model serving.

Fabric
Switching, optics and packet visibility.

Storage
Model, vector and knowledge data paths.

Private AI
Enterprise model service foundation.