ThinkSystem AMD MI300X 192GB 750W 8-GPU Board
Product Guide

The AMD Instinct MI300X 192GB 750W Accelerator is a GPU based on next-generation AMD CDNA 3 architecture, delivering leadership efficiency and performance for the most demanding AI and HPC applications. Eight MI300X accelerators are integrated into servers such as the ThinkSystem SR685a V3.

It is designed with 304 high throughput compute units, AI-specific functions including new data-type support, photo and video decoding, plus an unprecedented 192 GB of HBM3 memory on a GPU accelerator. Using state-of-the-art die stacking and chiplet technology in a multi-chip package propels generative AI, machine learning, and inferencing, while extending AMD leadership in HPC acceleration.

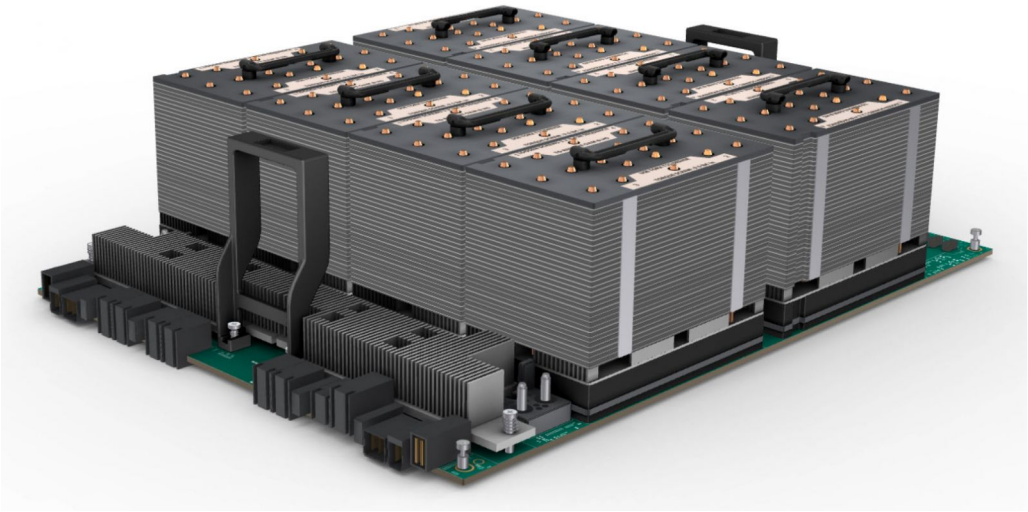

Figure 1. ThinkSystem AMD MI300X 192GB 750W 8-GPU Board

Did you know?

The ThinkSystem SR685a V3 includes 8x MI300X GPUs that are fully interconnected using AMD Infinity Fabric which provides 128 GB/s bandwidth between each of the 8 GPUs, for a total of 896 GB/s.

Part number information

The following table shows the part numbers for the 8-GPU board.

Table 1. Ordering information

Part number	Feature code	Description
CTO only	C1HK	ThinkSystem AMD MI300X 192GB 750W 8-GPU Board

Feature code C1HK contains 8x MI300X GPUs plus the Infinity Fabric high-speed interconnections.

Features

The AMD Instinct MI300X accelerator offers the following features:

- Designed to Accelerate Modern Workloads
  The increasing demands of generative AI, large-language models, machine learning training, and inferencing puts next-level demands on GPU accelerators. The discrete AMD Instinct MI300X GPU delivers leadership performance with efficiency that can help organizations get more computation done within a similar power envelope compared to MI250X accelerators from AMD. For HPC workloads, efficiency is essential, and AMD Instinct GPUs have been deployed in some of the most efficient supercomputers on the Green500 supercomputer list2, these types of systems— and yours —can take now take advantage of a broad range of math precisions to push highperformance computing (HPC) applications to new heights.

- Based on 4th Gen Infinity Architecture
  The AMD Instinct MI300X is one of the first AMD CDNA 3 architecturebased accelerators with high throughput based on improved AMD Matrix Core technology and highly streamlined compute units. AMD Infinity Fabric™ technology delivers excellent I/O efficiency, scaling, and communication within and between industry-standard accelerator module (OAM) device packages. Each discrete MI300X offers a 16-lane PCIe® Gen 5 host interface and seven AMD Infinity Fabric links for full connectivity between eight GPUs in a ring. The discrete MI300X is sold as an AMD Instinct Platform with eight accelerators interconnected on an AMD Universal Base Board (UBB 2.0) with industry-standard HGX host connectors.

- Multi-Chip Architecture
  The MI300X uses state-of-the-art die stacking and chiplet technology in a multi-chip architecture that enables dense compute and highbandwidth memory integration. This helps reduce data-movement overhead while enhancing power efficiency.

  Each OAM module includes:

  - Eight accelerated compute dies (XCDs) with 38 compute units (CUs), 32 KB of L1 cache per CU, 4 MB shared L2 cache shared across CUs, and 256 MB of AMD Infinity Cache™ shared across 8 XCDs. The compute units support a broad range of precisions for both AI/ML and HPC acceleration, native hardware support for sparsity, and enhanced computational throughput.
  - Four supported decoders for HEVC/H.265, AVC/H.264, V1, or AV1, each with an additional 8-core JPEG/MPEG CODEC
  - 192 GB of HBM3 memory shared coherently between CPUs and GPUs with 5.3 TB/s on-package peak throughput
  - SR-IOV for up to 8 partitions

- Coherent Shared Memory and Caches
  Machine-learning and large-language models have become highly data intensive, and they need to split jobs across multiple GPUs. AMD Instinct accelerators facilitate large models with shared memory and caches. The large amount of HBM3 memory is supported with 5.3 TB/s of local bandwidth, and direct connectivity of 128 GB/s bidirectional bandwidth between each GPU, accelerating memory-intensive AI, ML, and HPC models.

- AMD ROCm 6 Open Software Platform for HPC, AI, and ML Workloads
  Whatever your workload, AMD ROCm software opens doors to new levels of freedom and accessibility. Proven to scale in some of the world's largest supercomputers, ROCm software provides support for leading programing languages and frameworks for HPC and AI. With mature drivers, compilers and optimized libraries supporting AMD Instinct accelerators, ROCm provides an open environment that is ready to deploy when you are.

- Propel Your Generative AI and Machine Learning Applications
  Support for the most popular AI & ML frameworks—PyTorch, TensorFlow, ONYX-RT, Triton and JAX—make it easy to adopt ROCm software for AI deployments on AMD Instinct accelerators. The ROCm software environment also enables a broad range of AI support for leading compilers, libraries and models making it fast and easy to deploy AMD based accelerated servers. The AMD ROCm Developer Hub provides easy access point to the latest ROCm drivers and compilers, ROCm documentation, and getting started training webinars, along with access to deployment guides and GPU software containers for AI, Machine Learning and HPC applications and frameworks.

- Accelerate Your High Performance Computing Workloads
  Some of the most popular HPC programing languages and frameworks are part of the ROCm software platform, including those to help parallelize operations across multiple GPUs and servers, handle memory hierarchies, and solve linear systems. Our GPU Accelerated Applications Catalog includes a vast set of platform-compatible HPC applications, including those in astrophysics, climate & weather, computational chemistry, computational fluid dynamics, earth science, genomics, geophysics, molecular dynamics, and physics. Many of these are available through the AMD Infinity Hub, ready to download and run on servers with AMD Instinct accelerators.

Technical specifications

The following table lists the MI300X accelerator specifications.

Table 2. GPU specifications

Feature	Specification
Form Factor	OAM module
FP64 Performance	81.7 teraFLOPS
FP64 Matrix Performance	163.4 teraFLOPS
FP32 Performance	163.4 teraFLOPS
FP32 Matrix Performance	163.4 teraFLOPS
TF32 Matrix Performance	653.7 / 1,305 teraFLOPS*
BFLOAT16 Performance	1,305 / 2,610 teraFLOPS*
FP16 Performance	1,305 / 2,610 teraFLOPS*
FP8 Performance	2,610 / 5,220 teraFLOPS*
INT8 Performance	2,610 / 5,220 TOPS*
GPU Memory	192 GB HBM3
GPU Memory Bandwidth	5.3 TB/s
Total Graphics Power (TGP) or Continuous Electrical Design Point (EDPc)	750W
Partitions	8 partitions**
Interconnect	Infinity Fabric: 128 GB/s between each of the 8 GPUs, 896 GB/s total (fully interconnected)
Thermal solution	Passive

* Without / with structural sparsity enabled
** Partitions is a planned feature. Consult the latest ROCm release notes for availability: https://rocm.docs.amd.com/en/latest/about/release-notes.html

Server support

The following tables list the ThinkSystem servers that are compatible.

Table 3. Server support (Part 1 of 4)

		AMD V3				2S Intel V3			4S 8S Intel V3			Multi Node			GPU Rich				1S V3		
Part Number	Description	SR635 V3 (7D9H / 7D9G)	SR655 V3 (7D9F / 7D9E)	SR645 V3 (7D9D / 7D9C)	SR665 V3 (7D9B / 7D9A)	ST650 V3 (7D7B / 7D7A)	SR630 V3 (7D72 / 7D73)	SR650 V3 (7D75 / 7D76)	SR850 V3 (7D97 / 7D96)	SR860 V3 (7D94 / 7D93)	SR950 V3 (7DC5 / 7DC4)	SD535 V3 (7DD8 / 7DD1)	SD530 V3 (7DDA / 7DD3)	SD550 V3 (7DD9 / 7DD2)	SR670 V2 (7Z22 / 7Z23)	SR675 V3 (7D9Q / 7D9R)	SR680a V3 (7DHE)	SR685a V3 (7DHC)	ST50 V3 (7DF4 / 7DF3)	ST250 V3 (7DCF / 7DCE)	SR250 V3 (7DCM / 7DCL)
C1HK	ThinkSystem AMD MI300X 192GB 750W 8-GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	1[1]	N	N	N

1. Contains 8 separate GPUs connected via high-speed interconnects

Table 4. Server support (Part 2 of 4)

Part Number	Description	Edge					Super Computing					1S Intel V2			2S Intel V2		
		SE350 (7Z46 / 7D1X)	SE350 V2 (7DA9)	SE360 V2 (7DAM)	SE450 (7D8T)	SE455 V3 (7DBY)	SD665 V3 (7D9P)	SD665-N V3 (7DAZ)	SD650 V3 (7D7M)	SD650-I V3 (7D7L)	SD650-N V3 (7D7N)	ST50 V2 (7D8K / 7D8J)	ST250 V2 (7D8G / 7D8F)	SR250 V2 (7D7R / 7D7Q)	ST650 V2 (7Z75 / 7Z74)	SR630 V2 (7Z70 / 7Z71)	SR650 V2 (7Z72 / 7Z73)
C1HK	ThinkSystem AMD MI300X 192GB 750W 8-GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Table 5. Server support (Part 3 of 4)

Part Number	Description	AMD V1					Dense V2				4S V2		8S	4S V1			1S Intel V1			
		SR635 (7Y98 / 7Y99)	SR655 (7Y00 / 7Z01)	SR655 Client OS	SR645 (7D2Y / 7D2X)	SR665 (7D2W / 7D2V)	SD630 V2 (7D1K)	SD650 V2 (7D1M)	SD650-N V2 (7D1N)	SN550 V2 (7Z69)	SR850 V2 (7D31 / 7D32)	SR860 V2 (7Z59 / 7Z60)	SR950 (7X11 / 7X12)	SR850 (7X18 / 7X19)	SR850P (7D2F / 2D2G)	SR860 (7X69 / 7X70)	ST50 (7Y48 / 7Y50)	ST250 (7Y45 / 7Y46)	SR150 (7Y54)	SR250 (7Y52 / 7Y51)
C1HK	ThinkSystem AMD MI300X 192GB 750W 8-GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Table 6. Server support (Part 4 of 4)

Part Number	Description	2S Intel V1								Dense V1			
		ST550 (7X09 / 7X10)	SR530 (7X07 / 7X08)	SR550 (7X03 / 7X04)	SR570 (7Y02 / 7Y03)	SR590 (7X98 / 7X99)	SR630 (7X01 / 7X02)	SR650 (7X05 / 7X06)	SR670 (7Y36 / 7Y37)	SD530 (7X21)	SD650 (7X58)	SN550 (7X16)	SN850 (7X15)
C1HK	ThinkSystem AMD MI300X 192GB 750W 8-GPU Board	N	N	N	N	N	N	N	N	N	N	N	N

Operating system support

Operating system support is based on that of the supported servers. See the SR685a V3 server product guide for details: https://lenovopress.lenovo.com/lp1910-thinksystem-sr685a-v3-server

Regulatory approvals

The MI300X accelerator has the following regulatory approvals:

- Electromagnetic Compliance
  - Australia and New Zealand: CISPR 32: 2015 +COR1: 2016, Class A
  - Canada ICES-003, Issue 7, Class A
  - European Countries: EN 55032: 2015 + A11: 2020 Class B, EN 55024: 2010, EN 55035: 2017
  - Japan VCCI-CISPR32:2016, VCCI 32-1: 2016 Class A
  - Korea KN32, Class A, KN35, RRA Public Notification 2019-32
  - Taiwan CNS 13438: 2016, C6357, Class A
  - USA FCC 47 CFR Part 15, Subpart B, Class A
- Product Safety Compliance
  - UL 62368-1, 2nd Edition, 2014-12
  - CSA-C22.2 No. 62368-1, 2nd Edition, 2014-12
  - EN 62368-1, 2nd Edition, 2014 + A1: 2017
  - IEC 62368-1, 2nd Edition, 2014
  - RoHS Compliance: EU RoHS Directive (EU) 2015/863 Amendment to EU RoHS 2 (Directive 2011/65/EU)
  - REACH Compliance
  - Halogen Free: IEC 61249-2-21:2003 standard

Warranty

The MI300X accelerator assumes the server's base warranty and any warranty upgrades.

Related publications

For more information, refer to these documents:

- ThinkSystem and ThinkAgile GPU Summary:
  https://lenovopress.lenovo.com/lp0768-thinksystem-thinkagile-gpu-summary
- ServerProven compatibility:
  https://serverproven.lenovo.com/
- AMD MI300X product page:
  https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html

Related product families

Product families related to this document are the following:

- GPU adapters

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1943, was created or updated on June 4, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1943
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1943.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ServerProven®
ThinkAgile®
ThinkSystem®

The following terms are trademarks of other companies:

AMD, AMD CDNA™, AMD Infinity Cache™, AMD Instinct™, AMD ROCm™, and Infinity Fabric™ are trademarks of Advanced Micro Devices, Inc.

Intel® is a trademark of Intel Corporation or its subsidiaries.

Other company, product, or service names may be trademarks or service marks of others.