ThinkSystem Qualcomm Cloud AI 100 Accelerator
Product Guide

The Qualcomm Cloud AI 100 is designed for AI inference acceleration, and addresses the unique requirements in the cloud, including power efficiency, scale, process node advancements, and signal processing. The AI 100 enables data centers to run inference on the edge cloud faster and more efficiently. Qualcomm Cloud AI 100 is designed to be a leading solution for datacenters who increasingly rely on infrastructure at the edge-cloud.

The following figure shows the Qualcomm Cloud AI 100.

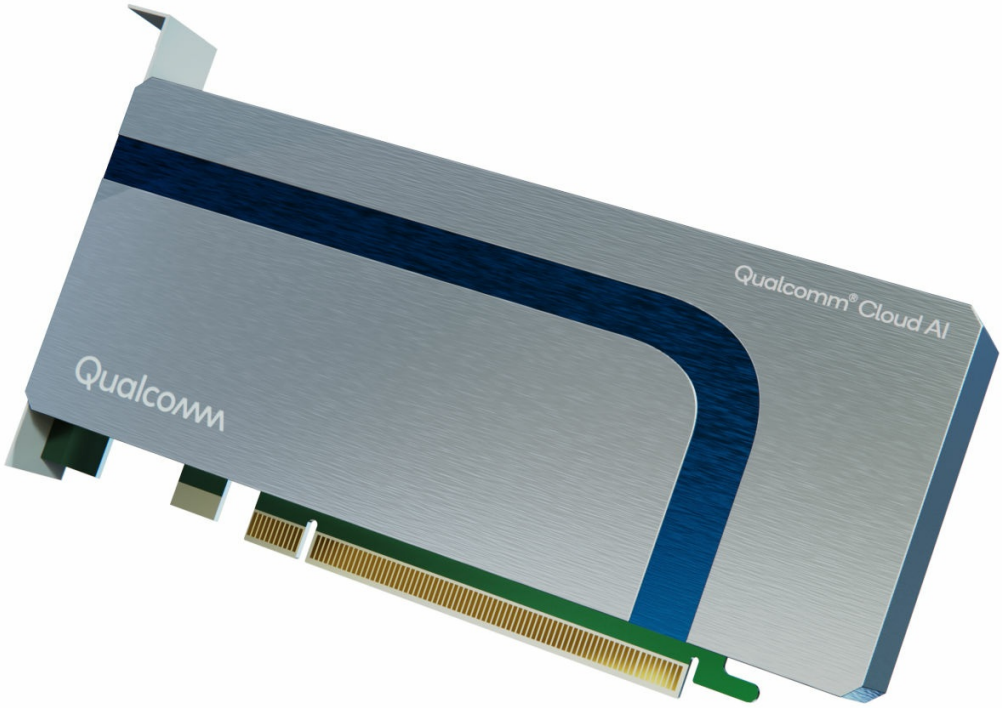

Figure 1. ThinkSystem Qualcomm Cloud AI 100

Did you know?

The ThinkSystem Qualcomm Cloud AI 100 accelerator is offered on ThinkEdge servers to enable customers to deploy AI workloads at the edge of their network. The AI 100 supports over 150 neural networks across multiple categories, including image classification, object detection, semantic segmentation, and natural language processing.

Part number information

The following table shows the part numbers for the GPU.

Table 1. Ordering information

Part number	Feature code	Description	Vendor part number
4X67A84009	BS49	ThinkSystem Qualcomm Cloud AI 100	QAIC-100P-0-MPA001-MT-01-0-BE

The option part number includes the following:

- One Qualcomm Cloud AI 100 (PCIe HHHL-Standard)
- Full height (3U) and Low Profile (2U) adapter brackets
- Documentation

Features

The Qualcomm Cloud AI 100 accelerator supports more than 150 deep learning networks, with strong emphasis on computer vision use cases and natural language processing.

Target applications include:

- Image classification
- Object detection and monitoring
- Semantic segmentation
- Face detection
- Point cloud
- Pose estimation
- Natural language processing (NLP)
- Recommendation systems

The Qualcomm Cloud AI 100 accelerator and accompanying software development kits (SDKs) offer superior power and performance capabilities to meet the growing inference needs of Cloud Data Centers, Edge, and other machine learning (ML) applications. The Cloud AI 100 card is powered by the AIC100 system-on-chip (SoC), which is designed for ML inference workloads.

The Qualcomm Apps and Platform SDKs provide the ability to compile, optimize, and run deep learning models from popular frameworks including:

- PyTorch
- TensorFlow
- ONNX
- Caffe
- Caffe2

Technical specifications

The Qualcomm Cloud AI 100 has the following specifications:

- Low profile form factor
- PCIe 4.0 x8 host interface
- Supports data types: FP32, FP16, INT16, INT8
- Security features include Hardware Root of Trust, Secure boot, Firmware rollback protection

The following table lists the processing specifications and performance of the Qualcomm Cloud AI 100.

Table 2. Specifications

Feature	Specification
Qualcomm AI Cores	16
Peak FP16 Floating Point performance	175 TFLOPS
Peak INT8 Integer Performance	350 TOPS
GPU Memory	16 GB LPDDR4x @ 2133 MHz
Memory Bandwidth	136.5 GB/s
ECC	Yes
Host Interface	PCIe Gen 4, x8 lanes
Form Factor	PCIe low profile (168mm x 69mm), single width
Max Power Consumption	75 W
Thermal Solution	Passive
Display connectors	None

Server support

The following tables list the ThinkSystem servers that are compatible.

Table 3. Server support (Part 1 of 4)

Part Number	Description	2S AMD V3				2S Intel V3			4S 8S Intel V3			Multi Node			GPU Rich				1S V3	
		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)	ST250 V3 (7DCF / 7DCE)	SR250 V3 (7DCM / 7DCL)
4X67A84009	ThinkSystem Qualcomm Cloud AI 100	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

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)
4X67A84009	ThinkSystem Qualcomm Cloud AI 100	1	N	2	4	3	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)
4X67A84009	ThinkSystem Qualcomm Cloud AI 100	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)
4X67A84009	ThinkSystem Qualcomm Cloud AI 100	N	N	N	N	N	N	N	N	N	N	N	N

Operating system support

The following table lists the supported operating systems.

Tip: These tables are automatically generated based on data from Lenovo ServerProven.

Table 7. Operating system support for ThinkSystem Qualcomm Cloud AI 100, 4X67A84009

Operating systems	SE360 V2	SE450	SE455 V3	SE350
Red Hat Enterprise Linux 8.4	N	Y	N	Y
Red Hat Enterprise Linux 8.6	Y	Y	Y	Y
Red Hat Enterprise Linux 8.7	N	Y	N	Y
Red Hat Enterprise Linux 9.0	Y	Y	Y	Y
Red Hat Enterprise Linux 9.1	N	Y	N	Y
Ubuntu 18.04.6 LTS	N	Y	N	Y
Ubuntu 20.04.5 LTS	Y	Y	Y	Y
Ubuntu 22.04 LTS	Y	Y	N	Y
Ubuntu 22.04.2 LTS	N	N	Y	N
VMware vSphere Hypervisor (ESXi) 7.0	N	N	N	Y

Auxiliary power cables

The Qualcomm Cloud AI 100 does not require an auxiliary power cable.

Regulatory approvals

The Qualcomm Cloud AI 100 has the following regulatory approvals:

- International: IEC 62368-1, EN62368-1 2nd, and 3rd Ed.
- United States of America: FCC
- Canada: ICES-003
- EU/UK: EN 55032, EN55024, EN55035, EN 61000-3-2, EN 61000-3-3, EN62368-1 2nd, and 3rd Ed.
- Taiwan: BSMI
- Korea: KN32 / KN35
- Japan: VCCI
- China: CNS 15663, RoHS
- Australia / New Zealand: AS/NZS CISPR 32
- Logos: cUL, FCC, ICES, RCM, VCCI

Operating environment

The Qualcomm Cloud AI 100 has the following operating characteristics:

- Ambient temperature
  - Operational: 0°C to 50°C (-5°C to 55°C for short term*)
  - Storage: -40°C to 85°C
- Relative humidity:
  - Operational: 5-90%
  - Storage: 5-93%

Warranty

One year limited warranty. When installed in a Lenovo server, the adapter 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/
- Qualcomm Cloud AI 100 product page:
  https://www.qualcomm.com/products/technology/processors/cloud-artificial-intelligence/cloud-ai-100

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, LP1772, was created or updated on July 18, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1772
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1772.

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®
ThinkEdge®
ThinkSystem®

The following terms are trademarks of other companies:

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

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

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