AUK Computing®

AUK ZEUS

Qualcomm

Features

- Supports X86-64 Scalable Processors for highest performance and improved power efficiency.
- Supports 8 DDR5 DIMM slots up to 6400GT/s
- Supports Compute Express Link (CXL2.0)
- Supports MCR DIMMs (8000MTs)
- Supports three PCIe5.0 x16 slots and four MCIO PCIe5.0 x8, total 80 PCIe 5.0 lanes
- Supports two RJ45 10GBase -T connectors
- Supports 2 Qualcomm Cloud AI 100 Ultra cards

Specifications

Dimensions	401x 215 x 347 mm [LxWxH]
Mother Board	Micro-ATX
CPU	Supports X86-64 Processors
Memory	8 Channel Per CPU, 8 x DIMMs
Expansion Slots	3 PCI-E
Internal Drive Bays	2 x 3.5" / 2.5" 1 x 2.5"
Accelerator Card	Qualcomm® Cloud AI 100 Ultra Up to 870 TOPS INT8 128GB on-card DRAM 64 AI Cores
TDP	150W Power
Power Supply	Supports: • 750W , 80 +Gold
Performance & AI Inference	The Qualcomm Cloud AI 100 Ultra, the newest member of our portfolio of cloud artificial intelligence (AI) inference cards, is a performance and cost-optimized AI inference solution, designed for Generative AI and large language models (LLMs). With up to 576 MB of on-die SRAM and 64 AI cores per card - and programmability for a wide range of workloads and acceleration techniques - Qualcomm Cloud AI 100 solutions address the unique requirements for scaling classic and generative AI workloads, ranging from computer vision and natural language processing to transformer-based LLMs.
Highest compatibility	Supports upto 2 graphics cards scalable, flexible & energy efficient. Supports CPU cooler height: 162 mm
Cooling Fans	Top: 2 x 120 mm Rear: 1 x 120 mm Bottom: 2 x 120 mm Water Cooling Radiator: Top: 1 x 240 mm

- The Above picture is for just illustration only actual product may be different
- All Specifications are subject to change without notice. Please visit our website for the latest information.
- Qualcomm are trademark and or registered trademarks of Qualcomm in the U.S. and other countries.
- All other brands, logos and names are property of their respective owners.

www.aukcomputing.com
sales@aukcomputing.com

AUK Computing®

Cloud AI 100 Ultra

Qualcomm® Cloud AI 100 Ultra is a performance - and cost-optimized AI inference solution, purpose-designed for Generative AI and large language models (LLMs).

Cloud AI Features

Form factor:	PCIe FH3/4L
ML capacity (INT8):	870 TOPs
On-die SRAM:	576 MB
On-card DRAM:	128 GB LPR4x 548 GB/s
Host interface:	PCIe Gen 4, 16 Lanes
Number of cores:	64 AI cores on single card

- Best Perf/TCO$
- 100B Gen AI models on a single card
- Software tools for frictionless porting of pre-trained models
- 8x larger models within a single server
- Fully programmable and with support for recent AI techniques and data formats

AI Inference Suite

for cloud and on-prem deployments

The Qualcomm® AI Inference Suite for Cloud and for On-Prem enables the deployment of AI models and applications with a single click, supporting efficient and scalable AI, while eliminating the need for complex infrastructure management.

Unlock AI with ease

Seamless one-click deployment. Easily swap or add your own models as needed, including gen AI, computer vision, and natural language processing. Build custom apps with common frameworks.

Deploy your way

Choose your preferred deployment: On-prem or cloud, powered by Qualcomm® Cloud AI roadmap of accelerators.

Top performance, future-proofed

Maximize performance and cost efficiency with powerful inference accelerators, embedded optimization techniques, and state-of-the-art models.

Run with confidence

High-availability and strict data privacy; no storage of model inputs or outputs. Designed and pressure tested by enterprise, for enterprise.

Highlights

- Powered by Qualcomm® Cloud AI roadmap of accelerators
- Robust APIs, OpenAI compatible
- Ready to use gen AI applications
- Configurable for any use case
- Supports 1000s of models
- Supports multi-tenancy

Accelerate gen AI development with ready-to-use applications and agents

- Chatbot
- AI agents
- Retrieval augmented generation (RAG)
- Summarization
- Image generation
- Real-time translation
- Code development
- Real-time transcription
- Your next use case

- The Above picture is for just illustration only actual product may be different
- All Specifications are subject to change without notice. Please visit our website for the latest information.
- Qualcomm are trademark and or registered trademarks of Qualcomm in the U.S. and other countries.
- All other brands, logos and names are property of their respective owners.