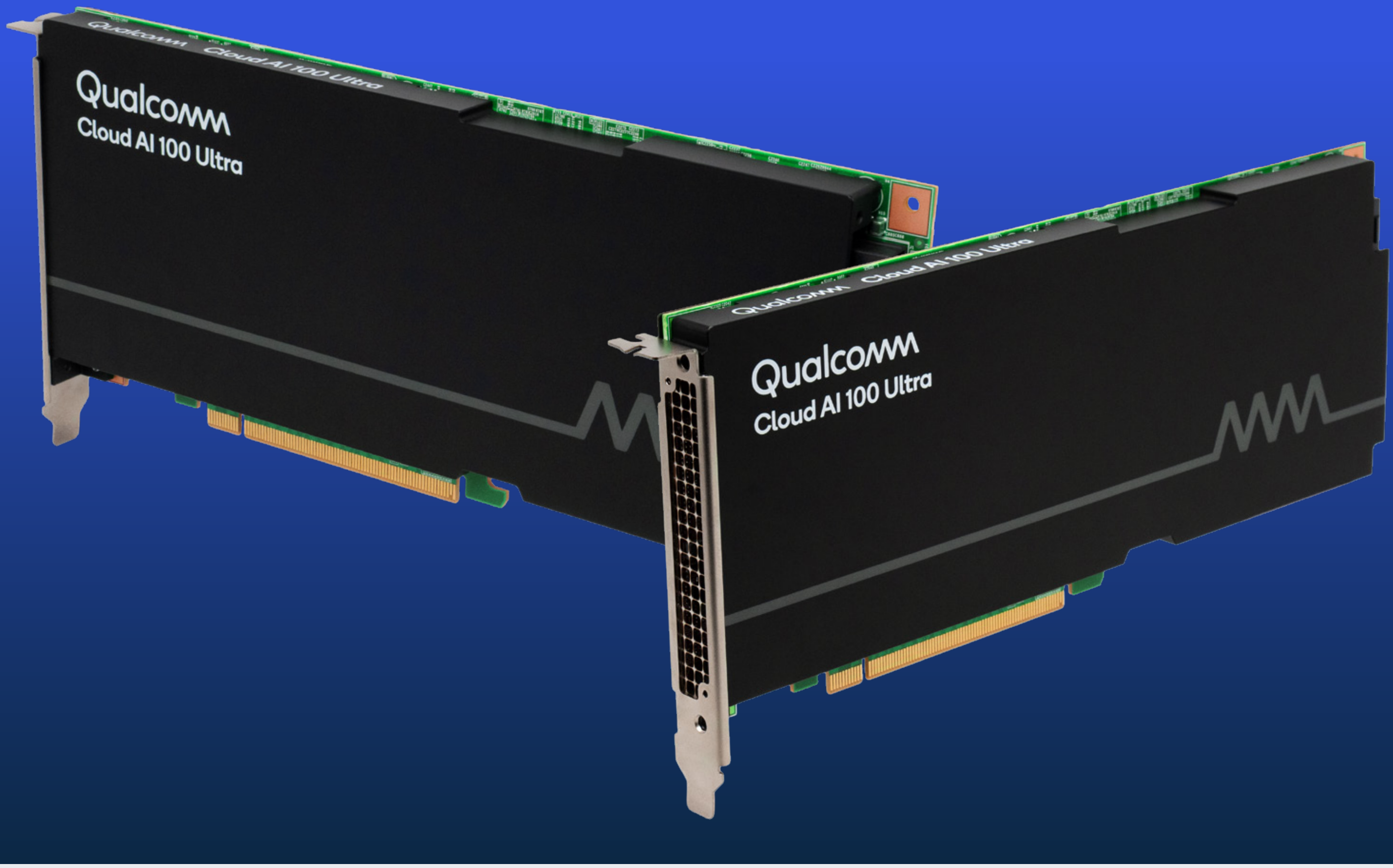
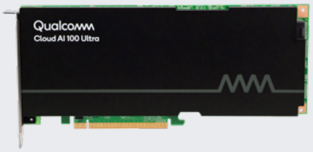


Qualcomm

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).

Form factor:	**PCIe FH3/4L**
TDP:	**150W**
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

*Compared to competitive solutions.

Qualcomm Cloud AI 100 Ultra is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.