Qualcomm Dragonwing

On-Prem Enterprise-grade AI That You Own, Control, and Scale

The Qualcomm Dragonwing™ AI on-prem appliance: High-performance AI without cloud dependence

AI is no longer optional, it is essential for staying competitive. Yet for enterprise leaders, adoption can be challenging. Rising cloud costs, data privacy concerns, low-latency applications, and strict compliance requirements often make public cloud services a poor fit for sensitive workloads. The Dragonwing AI on-prem appliance can help to address these challenges directly. This high-performance solution enables enterprises to run generative AI and computer vision workloads entirely within their own infrastructure. It provides data controls, ensures consistent performance, and lowers long-term operational costs, all without dependence on cloud providers.

Dragonwing AI on-prem appliance solution for enterprises

At the heart of the appliance is the Qualcomm® Cloud AI cards, part of the Qualcomm Cloud AI family. Engineered for enterprise-scale inferencing, the appliance delivers efficient performance for generative AI, computer vision, and high-performance computing (HPC) applications. With up to 37% lower power and nearly three times better performance/dollar than competitors*, the solution enables energy-efficient, scalable AI deployments.

Designed for flexibility, the plug-in hardware supports a range of form factors, from standalone desktop setups to compact, wall-mounted appliances. It requires no dedicated infrastructure, making it easy to deploy AI across diverse enterprise environments.

Qualcomm Cloud AI 100 ultra specifications

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**

*Source: Qualcomm internal benchmarks. Study available on request.

Helping to address key enterprise challenges

- **Run AI locally:** Process large AI workloads in real time and on-prem with 870 TOPS performance with no cloud delays, risks, or surprise costs.
- **Security features:** Keep data and IP on the device, mitigating security risk and leaks associated with using the cloud.
- **Achieve predictable AI costs:** Avoid ongoing cloud fees. Works on a fixed-capex model, pay once for long-term AI processing across your enterprise.
- **Easily integrate with enterprise infrastructure:** No special infrastructure needed. The appliance fits into existing enterprise stacks with Kubernetes, containers, and standard power.
- **Scale across the enterprise with confidence:** Compact and efficient, it scales easily across locations and departments, fitting easily in from data centers and edge sites.

To learn more visit: Qualcomm Cloud AI 100 Page

Enterprise-ready AI tools for on-prem deployment

The Qualcomm® AI Inference suite powers the Dragonwing AI on-prem appliance, enabling enterprises to build, deploy, and manage generative AI applications within their own infrastructure. Designed for scalability and ease of use, the suite includes a comprehensive SDK and OpenAI-compatible APIs, making it simple to integrate with your existing systems and workflows. With this platform, enterprises can run a wide range of familiar AI use cases locally including:

- Voice-enabled virtual agents
- Security-focused, domain-specific chatbots using SLMs, LLMs, and LMMs
- Retrieval-augmented generation (RAG) for intelligent search and summarization
- Custom internal AI assistants
- Multilingual search across enterprise data
- Automated document drafting and note capture
- Image and code generation
- AI-enabled video and image analysis for security, worker safety, and site monitoring

The suite provides easy-to-use API endpoints for core functions such as user management, chat, image generation, search, and audio/video-based AI. It supports widely used development frameworks like LangChain, CrewAI, and AutoGen, helping your teams build and scale AI agents faster. Plus, here's something powerful: The Qualcomm Cloud AI 100 Ultra supports most Hugging Face models, enabling low-latency inference at the edge.

All components can be deployed as containers, either bare-metal or within Kubernetes, with support for auto-scaling in enterprise environments.

The Qualcomm AI Inference Suite includes **Full API documentation and step-by-step tutorials**, helping your IT teams get applications up and running quickly.

Ready to see what on-prem ai can do for your enterprise?

In just minutes, your team can explore the Qualcomm AI Inference Suite on the Dragonwing AI OnPrem appliance—with no complex setup required.

Test drive key enterprise use cases using familiar, OpenAI-compatible endpoints. Everything runs locally, giving you a true sense of performance.

Get started with a simple Google sign-in. Experience the platform firsthand and see how quickly you can bring GenAI into your enterprise environment.

Get started now

Qualcomm Dragonwing