QNAP On-Premise AI Solutions

Enterprise-Grade On-Premise AI Infrastructure

INVESTOR BRIEFING · 2026

Local AI · Self-Deployment

Data Stays In-house

AI Stays Secure

Secure & Compliant · Rapid Deployment · Agent-Driven ·

Agenda

01 **QNAP Company & Brand Overview**
22 Years of Storage Technology · Global Market Position

02 **The AI Inflection Point**
ChatGPT → RAG → Agent Evolution

03 **Why Enterprises Choose On-Premise AI**
Security & Compliance · Cost Control · System Integration

04 **QAI Full Product Line**
Desktop to Data Center · h1290FX Deep Dive

05 **Enterprise Use Cases & Business Value**
Validated Cases · TCO · Market Opportunity

06 **Closing · Q&A**

ABOUT QNAP

QNAP

22 Years of Storage Expertise

Taiwan-listed company (Stock Code: 7805)
Taipei, Taiwan · Global Market Reach
Market Focus: AI Storage & Security

STOCK CODE: 7805

01

Comprehensive Smart Storage

Personal to enterprise-grade private cloud. All-Flash and hybrid architectures for massive-scale expansion and high-performance backup.

02

High-Speed Networking

Leading adoption of 25GbE / 100GbE high-speed specs. QSW switches build an enterprise-grade AI transmission environment.

03

Cybersecurity & Hybrid Cloud

Multi-layer anti-ransomware, Zero Trust architecture, immutable backup — seamlessly bridging public and private cloud.

04

Server Application Ecosystem

Container Station supports VMs and containers. OpenClaw / Hermes AI Agent works right out of the box.

COMPETITIVE MOAT · Core Competitive Advantages

Global Leadership · NAS & AI Storage

TIER 1 Leading Brand

22 years

Tier 1 global NAS market share; 22 years of enterprise storage expertise

Gross Margin Moat

56%

Far above industry average — sustains intensive R&D investment

R&D Capability

500+

Elite R&D engineering team focused on AI storage and cybersecurity

Scale · Expansion

Supports AI-scale data; MEGA Scale-out storage expands horizontally without limit

Secure · Security

Respects data sovereignty and regulations; private / hybrid cloud / Zero Trust full coverage

Sovereignty · Data Sovereignty

QNAP is the indispensable on-premise infrastructure enabler for the AI era

PART 01

The AI Inflection Point

01

From “Chat Tool” to “Enterprise Intelligence”
Three eras of evolution reshaping competitive advantage

01 ChatGPT · Conversational Q&A

02 RAG · Document Knowledge Enhancement

03 Agent · Active Task Execution

EVOLUTION · Timeline

Three Eras of AI Applications

STAGE 01

01

Chat Model

Represented by ChatGPT — general-purpose and easy to use, but answers come from the model's fixed knowledge and cannot access internal enterprise data.

// GPT // Claude // Gemini

STAGE 02

02

RAG Knowledge Base

Retrieves enterprise documents before generation, letting AI answer using "your data" with every answer traceable to its source.

// AnythingLLM // Open WebUI // Dify

STAGE 03

03

AI Agent

AI proactively breaks down tasks, calls tools, and executes across systems — capable of replacing portions of repetitive white-collar work.

// Function Calling // Tools

THE PROBLEM

The Core Dilemma of Enterprise AI

Cloud AI may seem fast, but is often incomplete for enterprises

01

Cloud AI Security Risks

Uploading enterprise data to the cloud creates exposure and regulatory risks (GDPR / finance / healthcare).

02

Runaway Cloud Costs

Per-API-call billing means total cost of ownership far exceeds projections at scale, hurting AI adoption ROI.

03

Lack of Enterprise Context

Without internal documents, SOPs, and terminology, AI cannot integrate into enterprise processes.

04

Insufficient Reliability & Continuity

Cloud outages or network instability directly impact AI business continuity and stable operation.

PART 02

What Modern Enterprises Need Is On-Prem AI

02

When AI enters core business operations, the cloud is no longer the only answer. Data compliance, long-term TCO, and system integration all push enterprises toward on-premise deployment.

01 Data Sovereignty · Data never leaves the server room

02 Controllable Cost · One-time investment, long-term use

03 System Integration · Fits into existing architecture

04 Compliance · Meets GDPR / financial / healthcare regulations

USE CASES · Real-World Deployments

AI Can Already Replace a Lot

01

Private RAG Knowledge Base

Employee manuals, SOPs, technical documents — instant Q&A with source citations.

02

Automated Report Generation

From raw data to weekly reports, quarterly reviews, contract drafts, and meeting minutes — one click.

03

Agentic AI Automation

Anomaly detection, compliance review, intelligent reporting — all executed in multiple steps on the internal network.

04

Text-to-SQL Business Intelligence

Executives query ERP/CRM in natural language for instant revenue and inventory analysis.

05

Customer Service & IT Self-Service

AI handles repetitive questions; staff focuses on exceptions and escalations.

06

Multimodal Image Generation

Marketing assets and product visuals generated with on-premise ComfyUI.

COMPARISON

Cloud AI vs. On-Premise AI

CLOUD API

Cloud API Service

- Conversation data uploaded to third parties — difficult to pass security and privacy audits
- Dependent on external networks — outages halt operations; response speed varies with quality
- Usage-based billing — 3-year cumulative TCO typically exceeds on-premise by 2× or more
- Difficult to integrate internal databases, SOP documents, and existing systems

ON-PREMISE

Self-Deployed On-Premise

- All conversations and vector data remain entirely on the internal network — zero exfiltration
- Zero latency, always-on — network anomalies have no effect on inference performance
- One-time hardware investment for long-term use — predictable and controllable costs
- Direct integration with ERP, CRM, file servers, and internal databases

PART 03

QNAP QAI
Full Product Line

03

From desktop personal knowledge bases to data-center-grade enterprise AI platforms, QAI offers a complete hardware portfolio for on-premise AI at every scale.

01 Desktop · QAI-6300 / 8300 Pro

02 Departmental · QAI-16700

03 Enterprise Flagship · QAI-h1290FX

QAI PRODUCT LINE · Full Portfolio

Desktop to Data Center — The Complete AI Ecosystem

01

Desktop
Desktop

QAI-6300 / 8300 Pro

✓ Built-in NPU acceleration

✓ Compact tower design

✓ Silent, low-power operation

Personal knowledge base, small-scale RAG applications

02

Departmental
Departmental

QAI-16700

✓ Hybrid storage architecture

✓ Cross-department sharing

✓ Flexible capacity expansion

Cross-department AI workflows, mid-scale RAG

03

Enterprise Flagship
Enterprise

QAI-h1290FX

✓ AMD EPYC processor

✓ All-Flash storage

✓ Enterprise-grade HA design

Large-scale LLM inference / Enterprise AI platform

PRODUCT OVERVIEW · QAI-H1290FX

Built for On-Prem AI

Integrates enterprise storage, multi-GPU inference capability, and 25GbE high-speed networking — eliminating all the hassle of assembling your own server.

FORM	Tower workstation — storage, compute, and inference in one unit
CPU	AMD EPYC 7302P · 16 cores / 32 threads · 3.3 GHz
GPU	4 PCIe slots · Supports NVIDIA RTX PRO Blackwell
MEMORY	DDR4 ECC · 128 GB base, expandable to 1 TB
STORAGE	12× U.2 NVMe PCIe Gen4 All-Flash hot-swap
NETWORK	2× 25GbE + 2× 2.5GbE built-in

CORE SPECS · Hardware at a Glance

The Robust Specs to Support Large Model Inference

Every specification is designed for large-scale LLM inference and multi-user concurrent scenarios.

CPU CORES

16 C / 32T

AMD EPYC 7302P · 3.3 GHz Base

MEMORY

1 TB

DDR4 ECC · 128 GB base, expandable

NVME BAY

12 U.2

PCIe Gen4 All-Flash hot-swap

NETWORK

25 GbE × 2

+ 2.5GbE × 2 built-in

AI COMPUTE POWER · Enterprise-Grade AI Computing

Enterprise-Grade Edge AI High-Performance Computing

3511

AI TOPS (FP4)

333

TFLOPS (RT Core) · NVIDIA RTX PRO Blackwell

DUAL GPU LAYOUT · Usage Example

GPU 1 · LLM Primary	GPU 2 · Multimodal Assist
RTX PRO 6000 Blackwell Max-Q Workstation · 96 GB · 300W	**RTX PRO 4000 Blackwell SFF** 24 GB · 70W LP
70B FP16 Inference	Embedding Model
Agent Primary Model	Stable Diffusion
RAG Generation Stage	Whisper Voice

Combined 120 GB VRAM · Multi-workload parallel computing

DUAL GPU CONFIGURATION

One System. Dual GPUs. Shared Power.

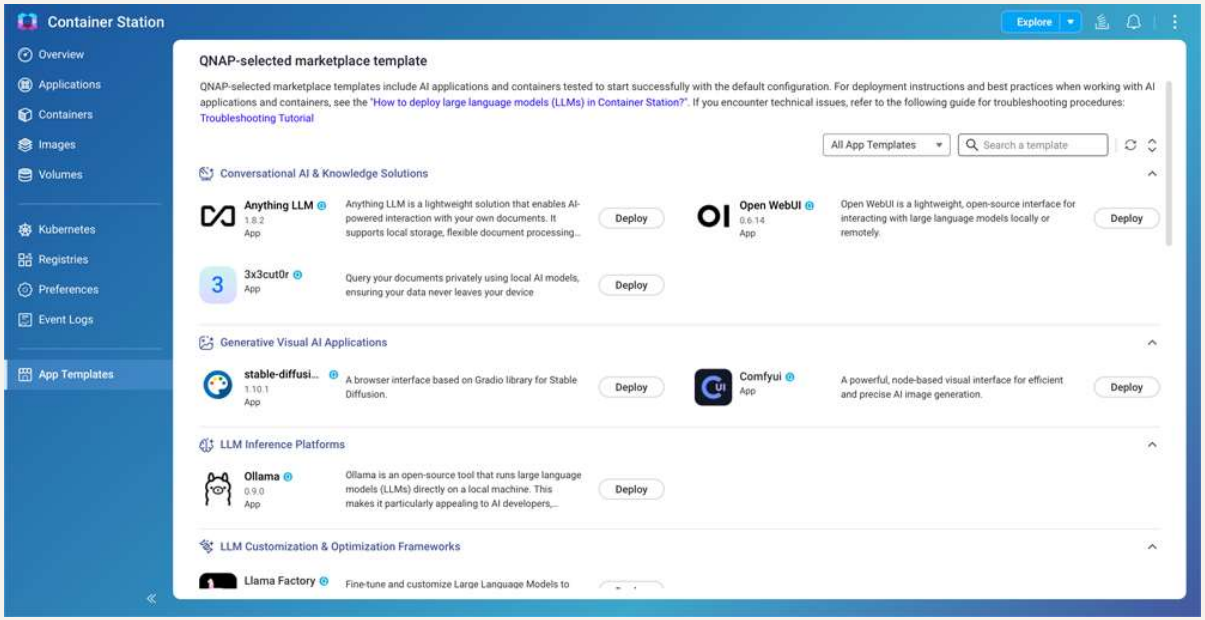
GPU 1	RTX PRO 6000 Blackwell Max-Q · 96GB Primary LLM inference — supports 70B FP16
GPU 2	RTX PRO 4000 Blackwell SFF · 24GB Handles Embedding, Stable Diffusion
Total	120GB VRAM · Two GPUs operating independently without interference

SOFTWARE PLATFORM

Seamless MLOps Platform Deployment

OpenClaw & Hermes · Zero MLOps knowledge required — IT completes the AI environment setup in a single day. Supports the latest open-source LLMs including Gemma 4 and Qwen 3.6.

AI Deployment in 1 Day
Zero MLOps Skills Required
100% Local Inference



PART 04

Enterprise Use Cases & Business Value

04

When on-premise AI is deployed, these scenarios become everyday operations — no longer constrained by data compliance or per-token costs.

01 Private RAG Knowledge Base · Document Q&A

02 Agentic AI Workflow Automation

03 Natural Language Database Query (Text-to-SQL)

04 Multimodal Generation · ComfyUI Image Output

DATA RESILIENCE & SECURITY · Defensible Moat

Data Resilience & Zero Trust Security

RTO < 90 sec — Automatic Failover

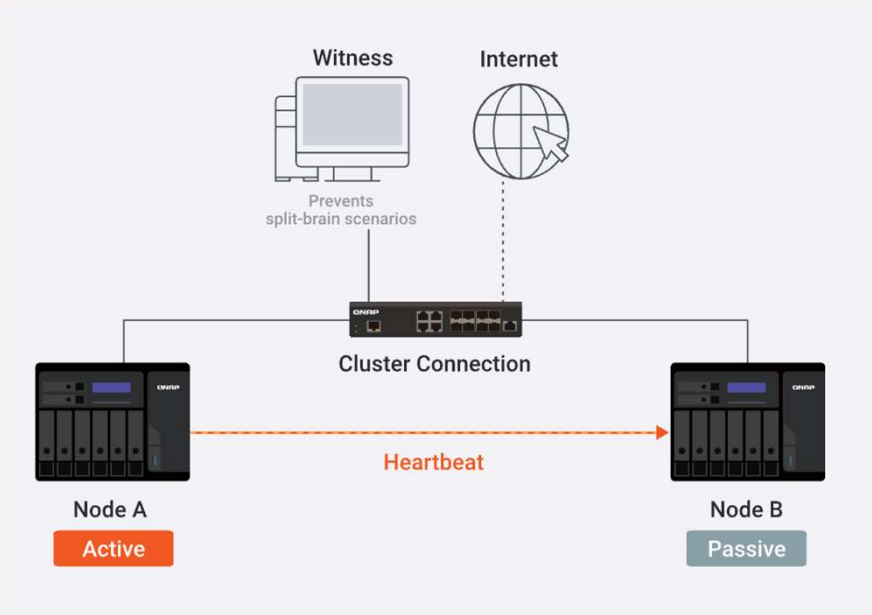
Dual-node Active+Standby HA eliminates single points of failure — AI service continuity with zero interruption.

3-2-1-1-0 Immutable Backup

Immutable Backup guards against ransomware; Video Verification confirms recoverability.

Zero Trust Architecture

Multi-layer anti-ransomware, hybrid cloud security integration, and AI threat detection — a comprehensive data protection matrix.



BUSINESS VALUE · Business Value

Infrastructure Becomes Your AI Brain

Significant TCO Reduction

60%

3-year deployment TCO reduced by over 60%
vs. continuous cloud API billing

Agentic AI Workflows

Private RAG knowledge base drives multi-step automation tasks — all executed on the internal network.

Qsirch AI Semantic Search

Full AI semantic search across documents, images, and video — auto-transcription and summary generation.

QuAgent Natural Language Control

Execute cross-system tasks with natural language — instant intelligent cross-folder search.

MARKET OPPORTUNITY · Market Opportunity

A Vast Market — QNAP at the Forefront

TAM

$420B

Global AI Infrastructure Market

SAM

$85B

Enterprise On-Premise AI Serviceable Market

SOM

$12B

QNAP Obtainable Market

Annual Growth Rate (Enterprise AI)

35%+

IDC 2025 · Gartner AI

82% Enterprises Concerned About AI Security & Compliance

78% Enterprise GenAI Adoption Rate (2025)

WHY QAI-H1290FX · Why Choose QNAP

Five Key Advantages — All in One Machine

01

Storage as Foundation

ZFS All-Flash + snapshot replication — vector database, models, and enterprise documents all on one unit.

STORAGE

02

Compute is Inference

Dual GPU · 120GB VRAM, primary + auxiliary roles — even 70B models run smoothly.

COMPUTE

03

Container is Deployment

Container Station deploys vLLM, AnythingLLM, and n8n with one click.

DEPLOY

04

Network is Integration

25GbE internal direct access to ERP, CRM, and file servers — zero-latency access.

NETWORK

05

Data is Sovereignty

Conversations, vectors, and inference results all remain in the corporate server room — zero exfiltration.

PRIVACY

COMPETITIVE COMPARISON · Competitive Comparison

QAI-h1290FX vs. Other AI NAS · AI Workstations

Comparison Item	QAI-h1290FX	Traditional QNAP NAS	Competitor AI NAS	Cloud AI Services
AI Computing Power	✓ 3,511 AI TOPS	✗ No GPU	△ Limited GPU Power	✓ Elastic Compute
Data Sovereignty	✓ Fully On-Premises	✓ On-Premises	✓ On-Premises	✗ Data Leaves Your Premises
Storage Integration	✓ All-in-One AI Storage	✓ High-Capacity Storage	△ Basic Integration	✗ Additional Storage Costs
AI Agent / RAG	✓ Out-of-the-Box	✗ Requires External Hardware	△ Partial Support	✓ Cloud LLM Support
5-Year TCO	✓ Lowest	✓ Low (Without AI Compute)	△ Medium	✗ High (Accumulating Subscription Costs)

“

QAI is not just an AI computing machine,
We are delivering an enterprise intelligent infrastructure that is “Secure & Compliant · Ready Out of the Box · Agent-Driven · Protected with Backup & DR”