www.qct.io

5G AI

Accelerate your AI journey with QCT: From system-level to AI Factory

Jack Luoh
Head of Product Marketing, QCT/Quanta

QCT CONFIDENTIAL

TOKENS

The fundamental unit of AI processing

The units AI models process and generate — from understanding inputs to reasoning and producing outputs.

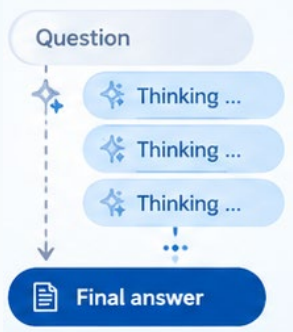


Why AI Is Processing More Tokens Per Task?

1 Reason More
Multi-Step Thinking

AI thinks through multiple steps before producing the final answer, especially for complex tasks.

More reasoning tokens per task



2 Act More
Agentic AI

AI agents repeatedly reason, call tools, receive results, and continue until a task is completed.

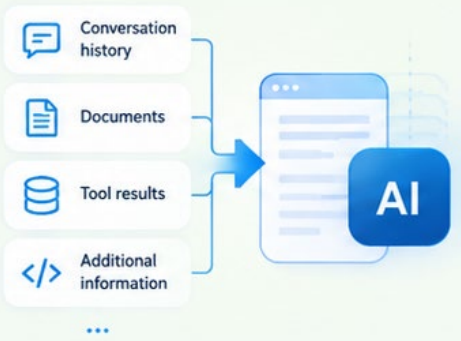
More model calls & tokens per task



3 Remember More
Long Context

AI needs to process longer conversations, documents, and historical information across interactions.

More input tokens to process



From More Tokens to Massive AI Compute Infrastructure

1 Reason More
Reasoning

AI thinks through multiple steps before producing the final answer, especially for complex tasks.

More reasoning tokens per task

2 Act More
Agentic AI

AI agents repeatedly reason, call tools, receive results, and continue until a task is completed.

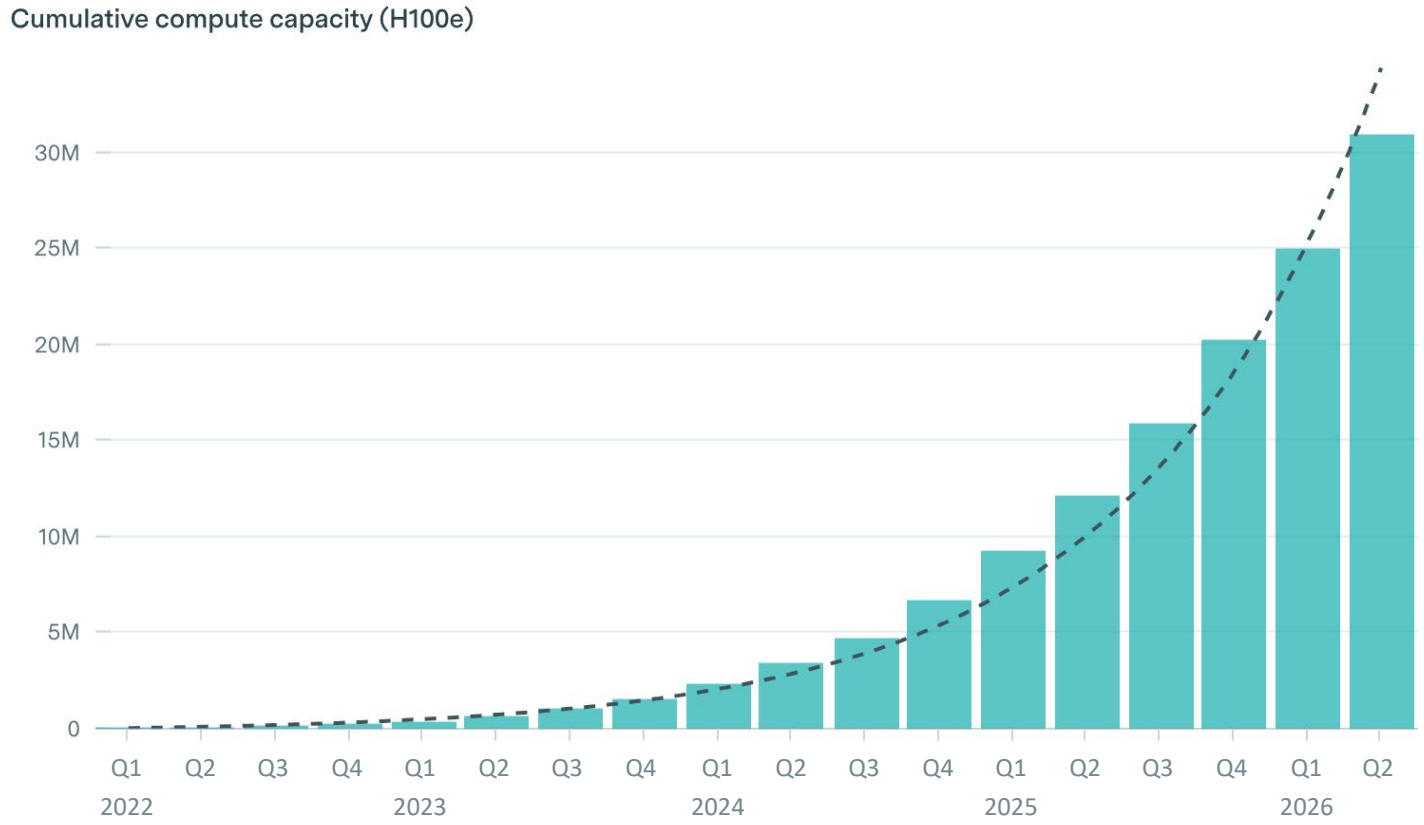
More model calls & tokens per task

3 Remember More
Long Context

AI needs to process longer conversations, documents, and historical information across interactions.

More input tokens to process

Global AI chip sales translate to rapidly expanding compute capacity



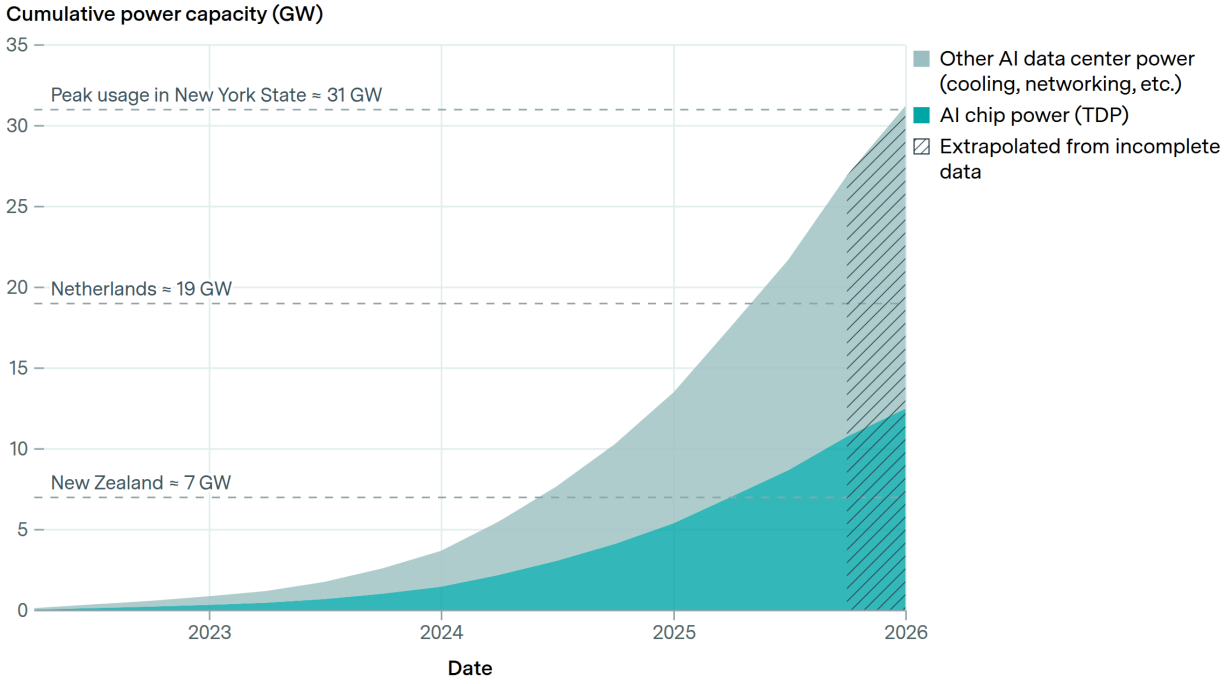
Source: EPOCH AI Dataset

QCT CONFIDENTIAL

AI Growth Is Outpacing the Power Infrastructure Behind It

Frontier AI power requirements are rising rapidly, while grid expansion struggles to keep pace.

Global AI power capacity is now comparable to peak power usage of New York State



Source: EPOCH AI Dataset

~31 GW

Global AI data center power capacity by Q4 2025

Comparable to peak power usage in New York State, and outstripping many developed countries.



Source: EPOCH AI Dataset

~40%

of facility power goes directly to GPUs

- Total server power is 1.53× GPU power alone.
- IT equipment uses 1.14× as much power as the servers.

Source: EPOCH AI Dataset

Grid expansion struggles to keep pace

Source: EPOCH AI Dataset

~268 GW Power supply gap by 2030



As AI Demand Scales, Power Becomes the Constraint

AI Demand
More sophisticated workloads



More AI Workloads → More Tokens

300T+ / DAY
AI tokens already disclosed

Source: Tokens Per Day

Drives demand for

Resource Demand
Expanding infrastructure requirements



COMPUTE
Processing Resource

More Tokens → More Compute

~30M H100e
Cumulative AI compute capacity by 2026

Source: EPOCH AI Dataset

Requires power to operate

Physical Limit
Constrained by grid and site infrastructure



POWER
Physical Infrastructure & Constraint

More Compute → More Power

~31 GW
Global AI data center power capacity by Q4 2025

Source: EPOCH AI Dataset

The focus is no longer just
More Compute
but how to generate

More Tokens Per Watt

QCT CONFIDENTIAL

Tokens-per-Watt Optimization with NVIDIA DSX

QCT collaborates with the DSX ecosystem to optimize tokens per watt from AI Factory perspective

Industrial Workloads

Manufacturing *Healthcare* *Education & Research* *Logistics* *Entertainment* *Transportation*

QCT AI Cluster Infrastructure

Compute **Storage** **Network** **Software**

DSX AIDC

DSX Sim MaxLPS DSX Exchange DSX Flex

Cooling system *Power system*

1 POD-Level Efficiency



Deliver software-defined systems tailored to workload requirements.

2 Infrastructure-Level Optimization



Integrate cooling and power systems to optimize data center design and operations.

QCT CONFIDENTIAL

POD-Level Efficiency: Optimized for Every AI Workload

Purpose-built architectures to deliver higher performance and efficiency across diverse AI workloads.

1 POD-Level Efficiency

Match each **AI workload** with the right **compute architecture**

Right Workload X Right Architecture → Optimize performance

ACCELERATE



Large-scale AI Compute

Vera Rubin NVL72

- Highest AI compute density
- Multi-GPU at scale
- Optimized for training and advanced reasoning

REASON



Agentic AI & Reinforcement Learning

Vera CPU Rack

- High-performance CPUs
- Optimized for agentic AI and RL
- Scalable and flexible architecture

REMEMBER



Context & Memory

Vera BlueField-4 STX Storage

- High-capacity memory
- Optimized for long context
- Efficient data access and retrieval

REAL-TIME RESPONCE



High-throughput Inference

Groq 3 LPX

- Low-latency inference
- High throughput
- Optimized for real-time and large-scale serving

CONNECT



High-speed AI Fabric

Spectrum-6 SPX Networking

- Scalable rack-to-rack connectivity
- High bandwidth, low latency
- Optimized for AI clusters

QCT CONFIDENTIAL

POD-Level Efficiency: Optimized for Every AI Workload

Purpose-built architectures to deliver higher performance and efficiency across diverse AI workloads.

Beyond accelerating AI, new architectures are emerging to optimize how AI reasons and remembers.

Right Workload X Right Architecture → Optimize performance

ACCELERATE

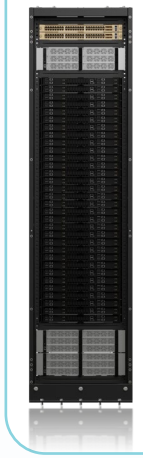
Large-scale AI Compute

Vera Rubin NVL72

- Highest AI compute density
- Multi-GPU at scale
- Optimized for training and advanced reasoning

REASON

Agentic AI & Reinforcement Learning



Vera CPU Rack

- High-performance CPUs
- Optimized for agentic AI and RL
- Scalable and flexible architecture

REMEMBER

Context & Memory



Vera BlueField-4 STX Storage

- High-capacity memory
- Optimized for long context
- Efficient data access and retrieval

REAL-TIME RESPONCE

High-throughput Inference

Groq 3 LPX

- Low-latency inference
- High throughput
- Optimized for real-time and large-scale serving

CONNECT

High-speed AI Fabric

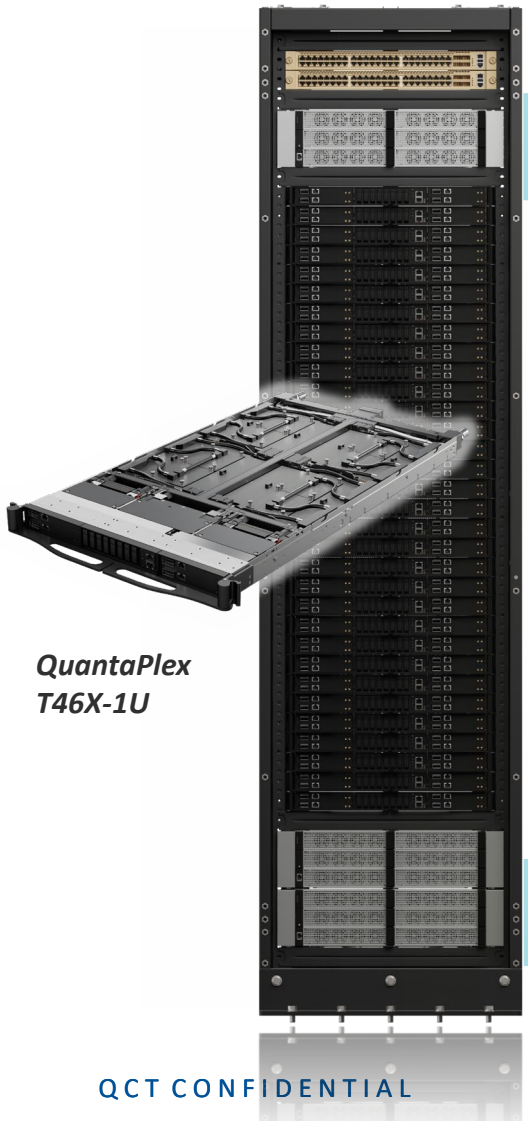
Spectrum-6 SPX Networking

- Scalable rack-to-rack connectivity
- High bandwidth, low latency
- Optimized for AI clusters

QCT CONFIDENTIAL

Vera CPU Rack Accelerate Execution & Generate More Tokens

High-performance, energy-efficient CPU platform for agentic AI.



QuantaPlex T46X-1U

Accelerate Agentic AI Workflows

Independent CPU Scaling

Scale CPU capacity independently from GPU compute, **adding execution resources** where agentic workloads need them.

Massive Parallel Execution

Support **22,500+ concurrent CPU environments** for sandbox execution, tool use, and RL evaluation.

High-Density CPU Deployment

Pack massive CPU capacity into a dense rack-scale architecture, **maximizing compute** within the data center footprint.

QCT CONFIDENTIAL

Vera Bluefield-4 STX Storage Reuse Context & Generate More Tokens.

AI-native context memory infrastructure for efficient inference.

QCT collaborates with leading storage solution providers to deliver AI-native storage solutions

Efficient Context Reuse

Reuse KV Cache
Shared context memory enables KV cache reuse across AI workloads.

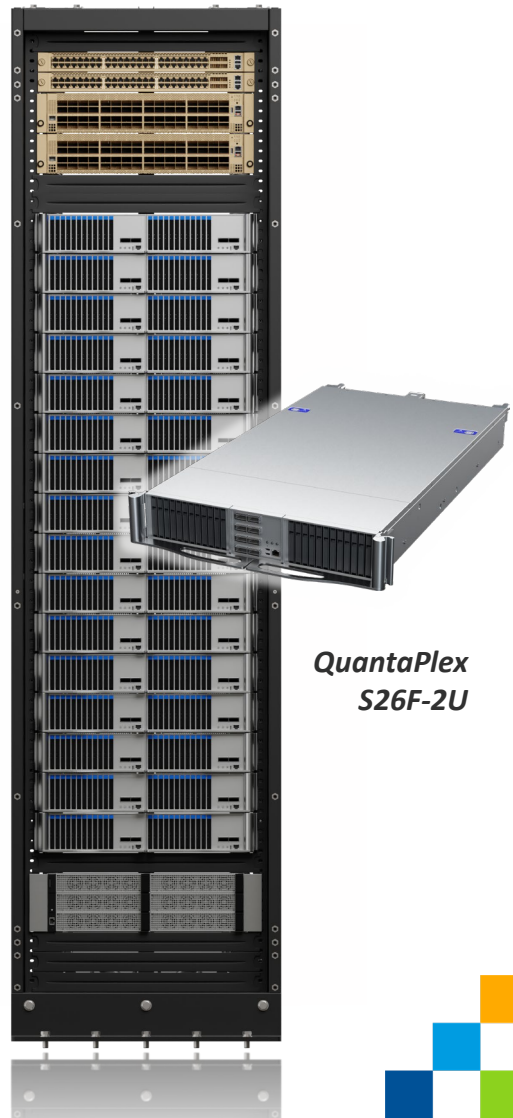
Reduce Recomputation
Avoid redundant GPU computation to improve compute efficiency.

Efficient Context Access

Faster Context Access
Prestage KV cache before decoding to reduce GPU stalls and idle time.

Free GPU Resources
Offload context from GPU memory to free valuable HBM for AI compute.

Efficient Context Memory for AI



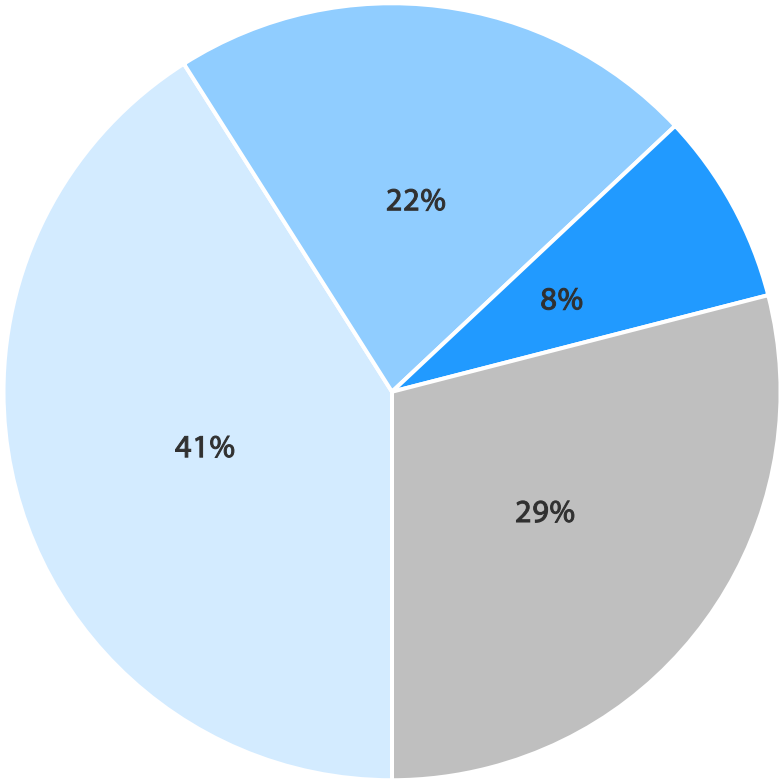
QuantaPlex S26F-2U

QCT CONFIDENTIAL

Beyond System-Level Efficiency: Addressing the Remaining 29%

1 **POD-Level Efficiency**

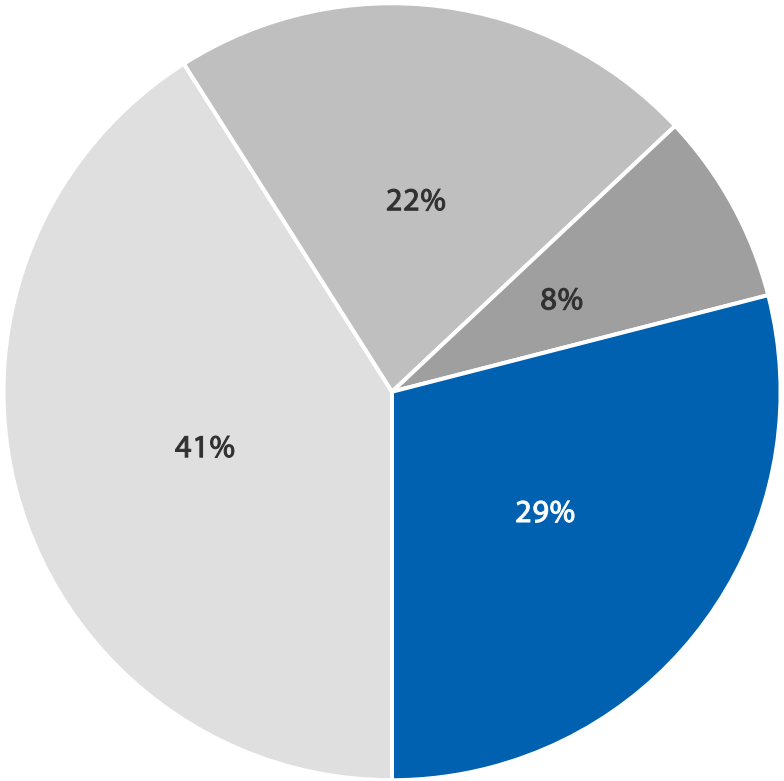




Estimates are for a data center using GB200 NVL72 chips, at peak operation
Source: EPOCH AI

Beyond System-Level Efficiency: Addressing the Remaining 29%

AI Data Center Power Distribution



Estimates are for a data center using GB200 NVL72 chips, at peak operation
Source: EPOCH AI

2 **Infrastructure-Level Optimization**



QCT Supports Infrastructure-level optimization

2 **Infrastructure-Level Optimization**

QCT collaborates with ecosystem partners to support the end-to-end AI Factory journey and optimize tokens per watt.



Design | **Construction** | **Operation**

Collection → Simulation → Report → Delivery → Deployment → DC Bring-up → Optimization & Maintenance

DSX SIM, Omniverse, MaxLPS, DSX Exchange, DSX Flex

- Collect AI factory capacity, power & density requirements and assemble required SimReady assets.
- Build AI factory digital twins and validate design with simulation tools.

- Utilizes simulation result to streamline construction processes.
- Ensure system operation to meet with digital twins for customer acceptance.

- Enables continuous operational feedback and traceability for AI Factory management.

Streamline AI Factory Innovations with Digital Twins

By leveraging digital twins for simulation and validation,
QCT optimize AI Factory performance and accelerate data center deployment

Real World

Green Field / Brown Field Data Center

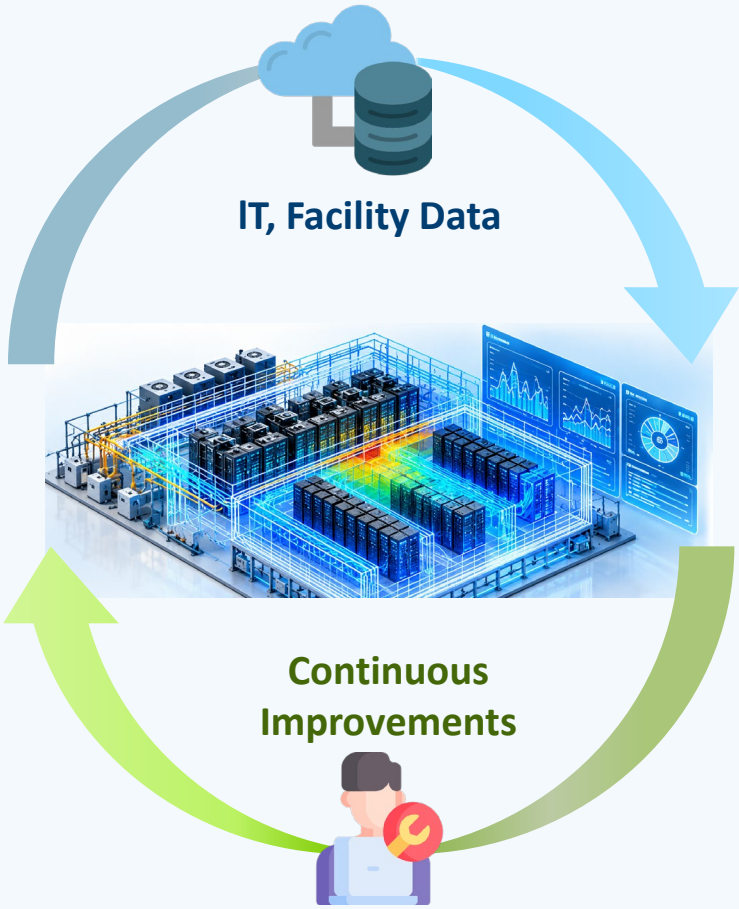


Physical Systems & Equipment

Racks, servers, cooling and power systems

Ecosystem Partners

Provide equipment models, parameters and real-world data



Virtual Environment

Simulate, analyze and validate in the virtual environment



Thermal & Airflow Simulation

Analyze temperature, detect hotspots and optimize cooling

Power & Energy Analysis

Evaluate power capacity, PUE and energy efficiency

Configuration & Layout

Adjust rack and equipment configurations

QCT AI Factory Based on NVIDIA DSX

Together with ecosystem partners, QCT enables a Neocloud customer to reduce execution risks, accelerate AI Factory deployment, scale out faster, and optimize performance

Power Digital Twins with IT & Facility Data

Provide data center inputs for digital twin simulations, reducing AI Factory design iterations and accelerating deployment.

Reduce Deployment Risk through Simulation

Minimize deployment risk through pre-deployment simulation and validation.

Scale Faster with Modular Architecture

Validate modular designs with ecosystem partners to ensure standardized infrastructure and future scalability.

Optimize AI Factory Performance

Continuously synchronize data with the digital twin to optimize AI factory performance.

.

QCT CONFIDENTIAL

Different Partners. Multiple Paths to AI Infrastructure.

Choose the path that best accelerates your AI journey.

AI Factory

Full-stack NVIDIA AI infrastructure platform for large-scale AI training and accelerated workloads.



End-to-end NVIDIA platform

Large-scale AI

Building a More Open AI Future

Heterogeneous Systems

Flexible system integration with different platforms to support diverse AI workloads.



Flexible

Workload-Driven

QCT CONFIDENTIAL

QCT AI POD: Cluster-Level Infrastructure Optimized for Diverse AI Workloads

Choose The Right Solution Package Tailored to Your Workloads Needs

- Computer Vision
- LLM Inference
- LLM Fine-tune Training
- Trillion-parameter LLM Training
- Generative AI Inference with RAG
- Deep Learning Training
- Agentic AI
- AI Reasoning

NVAIE

For Enterprise

- AI Application
- Workload Orchestration
- System Provisioning

QCT SW Stack

For Enterprise

- Rapid Deployment
- System Management
- Realtime Monitoring
- Development Environment

DSX OS

For NCP

Platform Software
Orchestration and Scheduling

Infrastructure Software
Infrastructure Provisioning & System Management

QCT AI Cluster Infrastructure

NVIDIA



4/8 GPU AI server
D75E-4U



8 GPU server
D75H-10U



8 GPU server
D76T-2U

Intel



General server
D55Q-2U



Storage
S55J-2U

AMD



8 GPU server
D75T-7U



General server
S44N-1U

Build the Right AI POD for Your Workloads

Intel - Lighter AI Workloads: Right-sized infrastructure for efficient inference and everyday AI workloads.



Lighter AI Workloads

- Computer Vision
- LLM Inference
- Generative AI Inference with RAG

Built on intel Platform

Workload

Compute

GPU-accelerated compute for inference, fine-tuning, perception & visual computing

4-GPU / 8-GPU



D75E-4U

Intel GNR-2P

- Unlock Infinite AI Possibilities

Management

Manage, monitor and orchestrate AI resources for reliable and efficient operations

Management node



D55Q-2U

Intel SRF/GNR-2P

-High Flexibility for Expansion

Storage

Keep datasets, models and application data continuously available to AI compute

Capacity Node



S55J-2U

Intel SRF/GNR-1P

- Flexible, Ultra-Dense Architecture

Build the Right AI POD for Your Workloads

Intel - Advanced AI Workloads: High-performance infrastructure for compute-intensive AI training and reasoning.



Advanced AI Workloads

- LLM Fine-tune Training
- Agentic AI
- AI Reasoning
- Deep Learning Training

Built on intel Platform

Workload

Compute

GPU-accelerated compute for inference, fine-tuning, perception & visual computing

HGX Platform



D75E-4U

Intel GNR-2P

-Enabling Diverse AI Workloads

-Unlocking Infinite AI Possibilities

Management

Manage, monitor and orchestrate AI resources for reliable and efficient operations

Management node



D55Q-2U

Intel SRF/GNR-2P

-High Flexibility for Expansion

Storage

Keep datasets, models and application data continuously available to AI compute

Capacity Node



S55J-2U

Intel SRF/GNR-1P

- Flexible, Ultra-Dense Architecture

Build the Right AI POD for Your Workloads

AMD - Advanced AI Workloads: Flexible GPU infrastructure for scalable AI training and HPC workloads.



Advanced AI Workloads

- LLM Fine-tune Training
- Agentic AI
- AI Reasoning
- Deep Learning Training

Built on AMD Platform

Workload

Compute

GPU-accelerated compute for inference, fine-tuning, perception & visual computing

8-GPU



D75H-7U

AMD MI325X/MI300X-2P

- Enabling Diverse AI HPC Workloads

Management

Manage, monitor and orchestrate AI resources for reliable and efficient operations

Management node



S45Q-2U

AMD GENOA-2P

-High Flexibility for Expansion

Storage

Keep datasets, models and application data continuously available to AI compute

Capacity Node



S81N-2U

AMD GENOA-1P

- Flexible, Ultra-Dense & Cost-Effective

From More Tokens to a Smarter AI Factory

OPTIMIZE EVERY WATT

- POD-level efficiency
- More AI output from every watt

OPTIMIZE THE AI FACTORY

- Infrastructure-level optimization
- Power, cooling, networking & operations

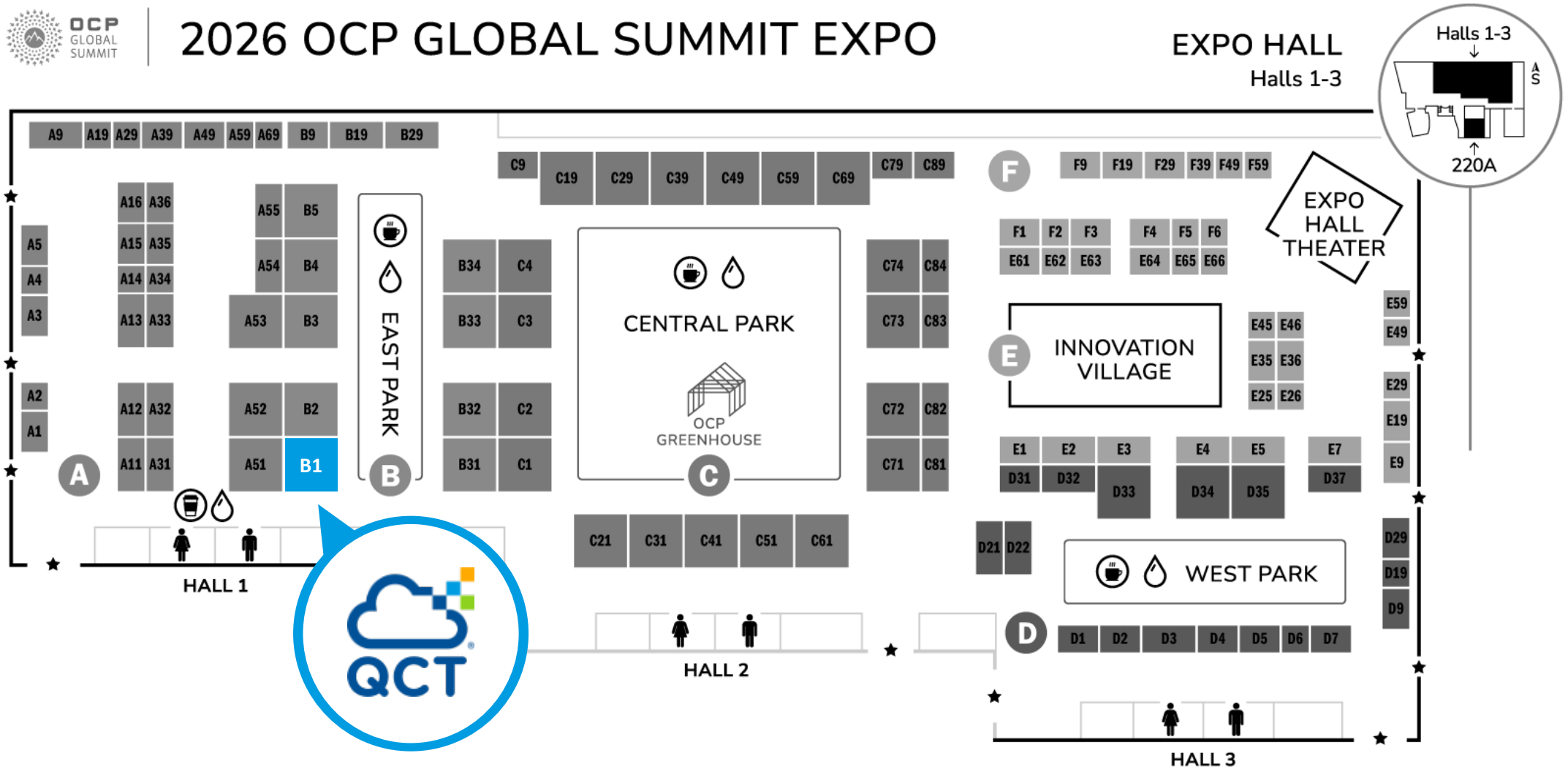
BUILD AROUND REAL WORKLOAD

- Flexible AI POD solutions
- Right platform. Right scale. Right workload.

Accelerate your AI journey with QCT

from optimized systems and AI Factory infrastructure to workload-ready AI solutions.

2026 OCP GLOBAL SUMMIT EXPO



Visit QCT Booth #B1 to Learn More !



www.qct.io

Thank You

5G AI