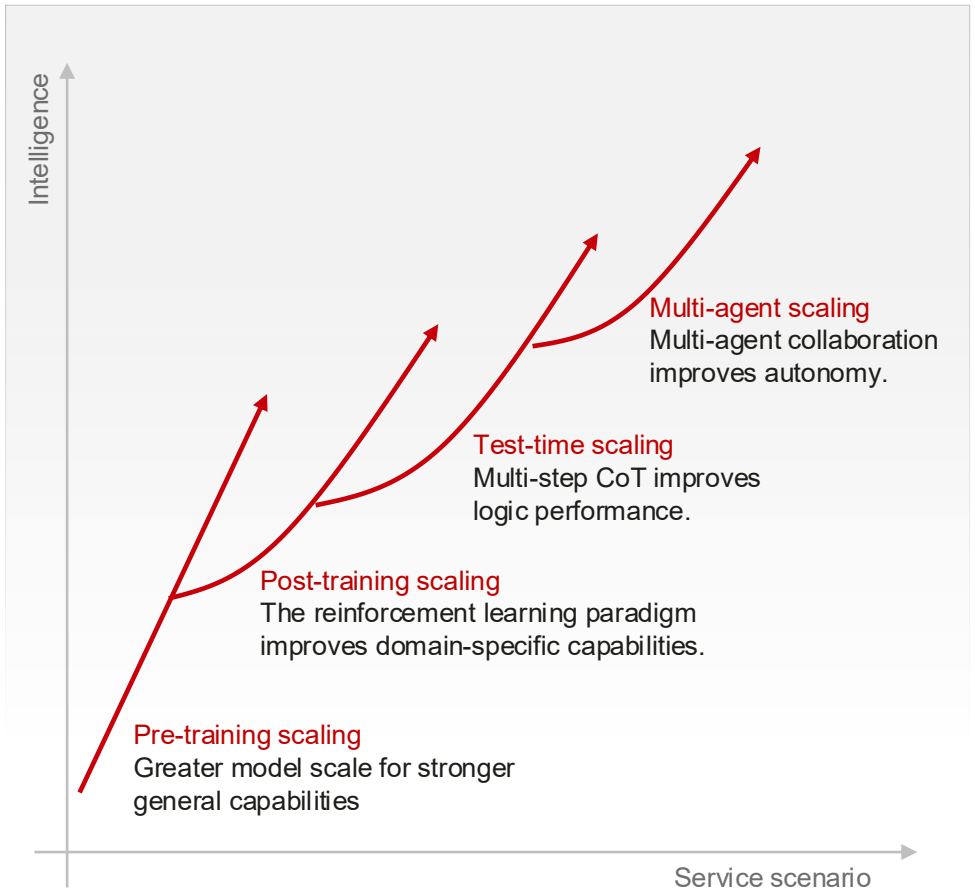
AI Full Stack Redefined: Compute, Software & Storage

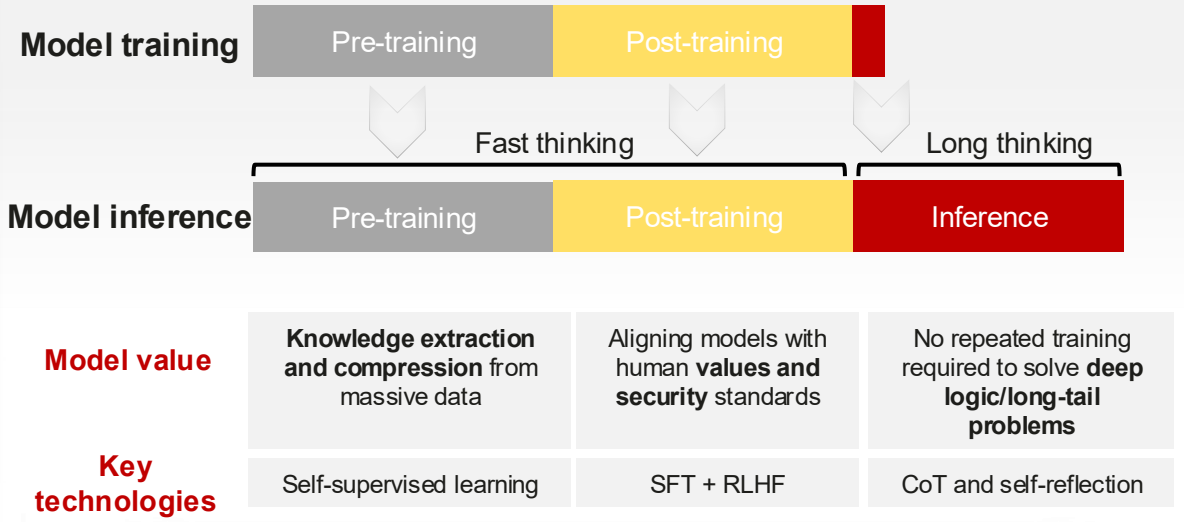
01

Huawei Next-Gen AI Computing

Scaling Law continues: Computing Power remains the Core Driver for AI Innovation

As models grow more intelligent, **the availability of computing power** will remain crucial for breaking their performance limits





2025 → 2030

Model parameters
(100 bn → 10 tn)

Computing power demand
(↑ 1,000x+)

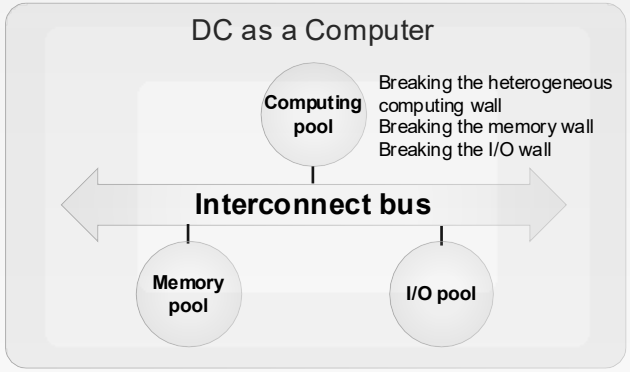
HUAWEI

Evolution of Computing, Network, and Energy requirements for Intelligent Infrastructure

Computing infrastructure will move towards all-domain coupling

Single-domain coupling → All-domain coupling

(Compute + Memory + I/O)

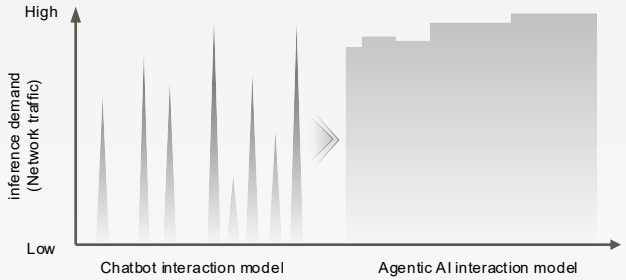


- **Full-scale data center pooling:** DC as a Computer and flexible resource configuration to meet diverse service requirements
- **High-concurrency and low-latency computing:** Agentic memory retrieval < 50 ms, single-token inference in milliseconds, and million-level long-sequence processing
- **Multimodal hierarchical memory storage:** Agents serving as real-time, short-term, and long-term hierarchical memory systems like humans

Networks will be vital to eliminating information silos for AI

Single-round AI conversation → Agentic AI

Continual interactions and heavy network traffic



- Agents reshape traffic models: **AI-driven traffic will account for 64% of total traffic by 2030**
- On-device AI application boosts uplink traffic: **Increase from 10% to over 40% in the next 3 years**
- Cross-DC agentic execution: **Inter-DC traffic changes from point-to-point to mesh, doubling every 18 months**
- AI optimizes communication networks and **comprehensively improves network capabilities and user experience**

Energy network management will become token-based

Data center construction has entered the GW era

Single-point load exceeds 1 GW, with high density and liquid cooling as the norm

Energy mix transformation

Wind and solar will become primary energy sources. Energy storage systems (ESS) and hydrogen energy will become widespread.

By 2035

> 50%
Proportion of wind and solar power

Levelized cost of electricity (LCOE)
< US$0.01

Token-based "energy nervous system"

AI will become the "nervous system" of energy, with each joule of energy programmable and autonomously manageable.

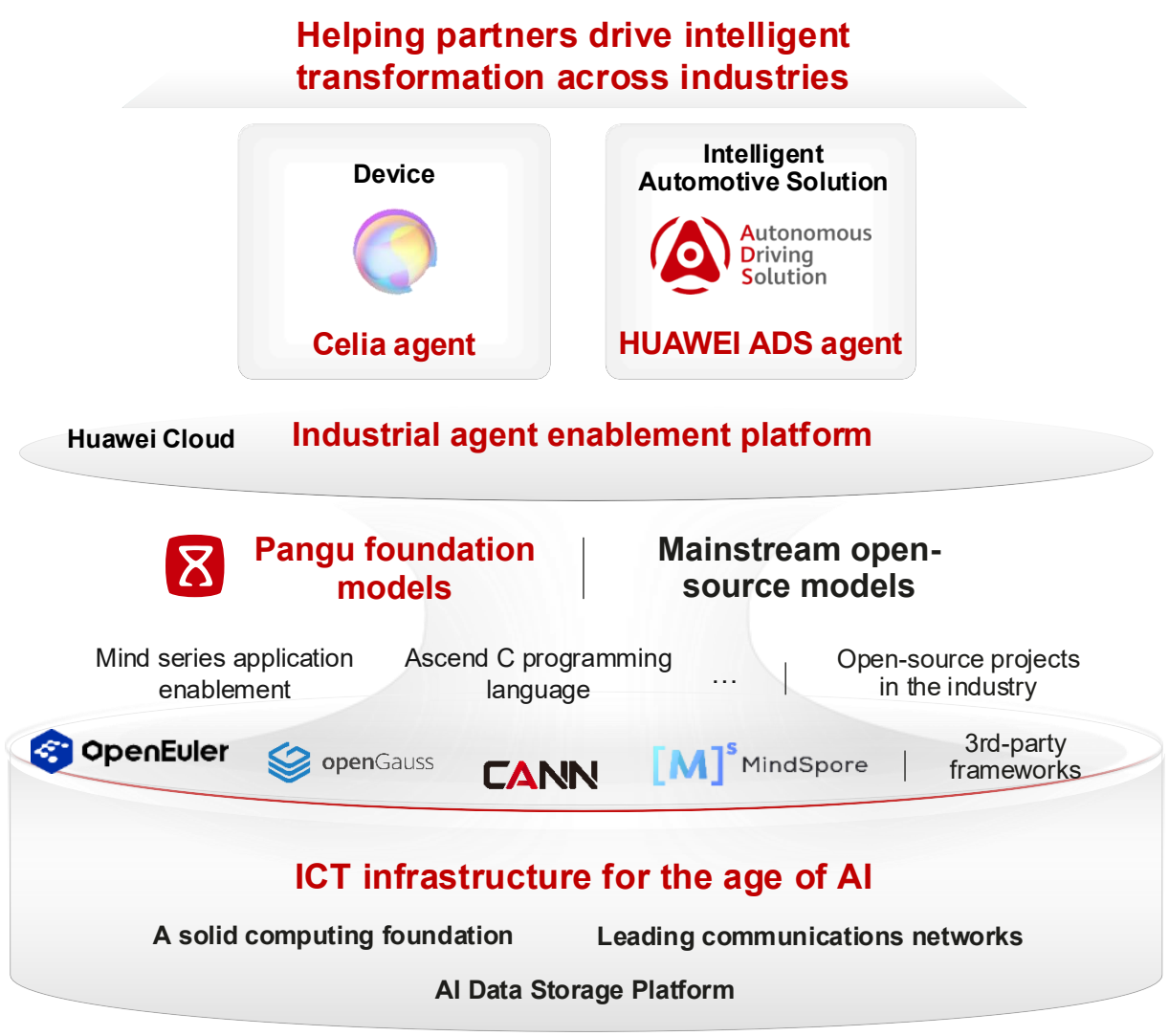
By 2035

> 99.9%
Grid self-healing rate

100%
AI-based electricity transactions

HUAWEI

Advancing the All Intelligence Strategy to Support a vast range of Models and Applications across Industries



1. A solid computing foundation that powers a vast range of models and applications

- SuperPoDs and SuperClusters: Massive computing and superior competitiveness
- Open access to the UnifiedBus protocol and the SuperPoD reference architecture, and open source for the UB OS Component

2. An open and vibrant AI ecosystem

- CANN, MindSpore, and Mind series application enablement kits and toolchains: Open source and open access to help partners and developers innovate faster
- Embracing mainstream open-source frameworks and programming languages, and providing equal support for mainstream model communities

3. Continuous investment in foundation models

- Pangu foundation models with enhanced competitiveness

4. Agents and agent enablement platform

- The device agent Celia, and HUAWEI ADS agent
- B2B: An industrial agent enablement platform that helps partners drive intelligent transformation across industries

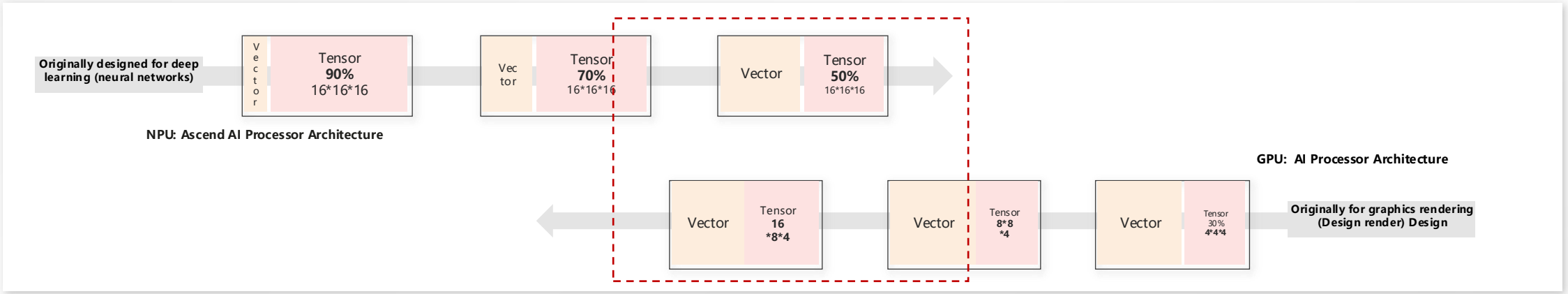
5. AI for more competitive Huawei products

6. AI for Huawei's internal operations

HUAWEI

Ascend Huawei's Self-developed NPU Roadmap

Both NPU and GPU continue to evolve along different technological paths, complementing each other.



First-generation Ascend chips for international markets

Ascend 910

Q1 — Ascend 910C

2025

Ascend 950PR

Q1 — Ascend 950PR

Ascend 950DT

Q4 — Ascend 950DT

2026

- Supports mxFP8, FP8, mxFP4, and HiF8 low-precision data formats
- Improves vector computing power and supports SIMD+SIMT
- Improves interconnect bandwidth.

Ascend 960

Q4 — Ascend 960

2027

- Improves computing power, memory capacity, and bandwidth.
- Continuously improves training/inference performance.

Ascend 970

Q4 — Ascend 970

2028

HUAWEI

SuperPoD Architecture: Breaking the Communication Wall with Unified Bus

Ultra-high bandwidth

10x↑ bandwidth

100 GB/s → TB-level

Optical-electrical integration, lossless interconnect

Ultra-low latency

50%↓ RTT latency

7 μs → 3 μs

Unified protocol, high-density design

Unified memory addressing

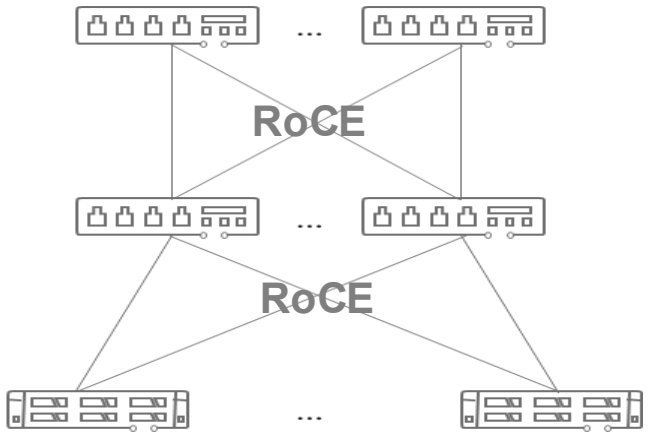
Unified addressing of global memory

256 TB on-chip memory

Peer-to-peer interconnect, resource pooling

Conventional cluster

Interconnect via network devices



Stacked AI servers, communication of idle computing resources

10x↑ bandwidth

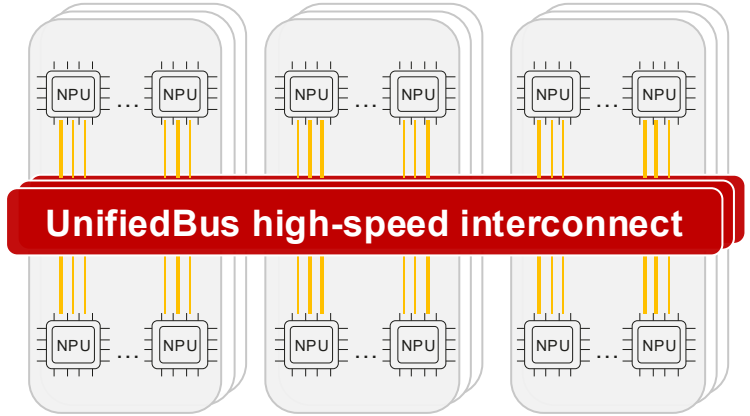
50%+↓ latency

100x↑ memory capacity

TB-level → 100 TB-level

Huawei SuperPoD

Interconnect via UnifiedBus



High-speed NPU interconnect and unified global memory addressing

HUAWEI

Focus on Three Major Scenarios

Three types of Ascend solutions

AI Computing Center (AICC) Solution | Ascend open source solution reference design | Partner scenario-based solutions

Internet recommendation

Atlas 350 accelerator card

Integrated into Partner server

KunLun

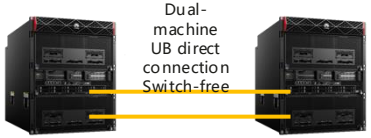
Industry intelligence

Inference server (A2 single server with 8 cards)

INT8/4 quantization, support next-generation DSV4 300B

Upgrading air-cooled servers

Atlas 850E Super Node Server (single/dual-node)
Enterprise-level air-cooled data center

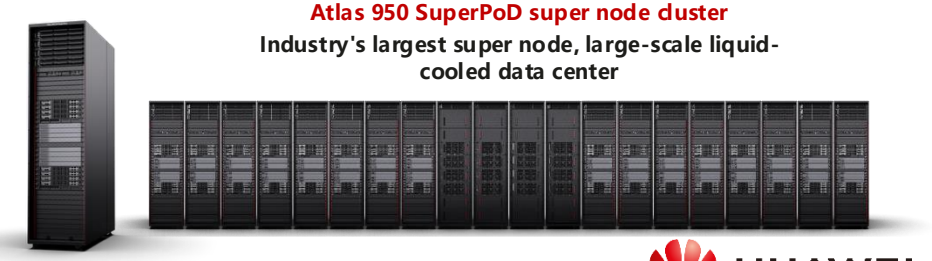


Dual-node, best affinity for next-generation Deepseek, supporting 10 ms low latency

National Intelligent Computing Center

- High-density design
- Supports low-precision formats
- Flexible capacity expansion

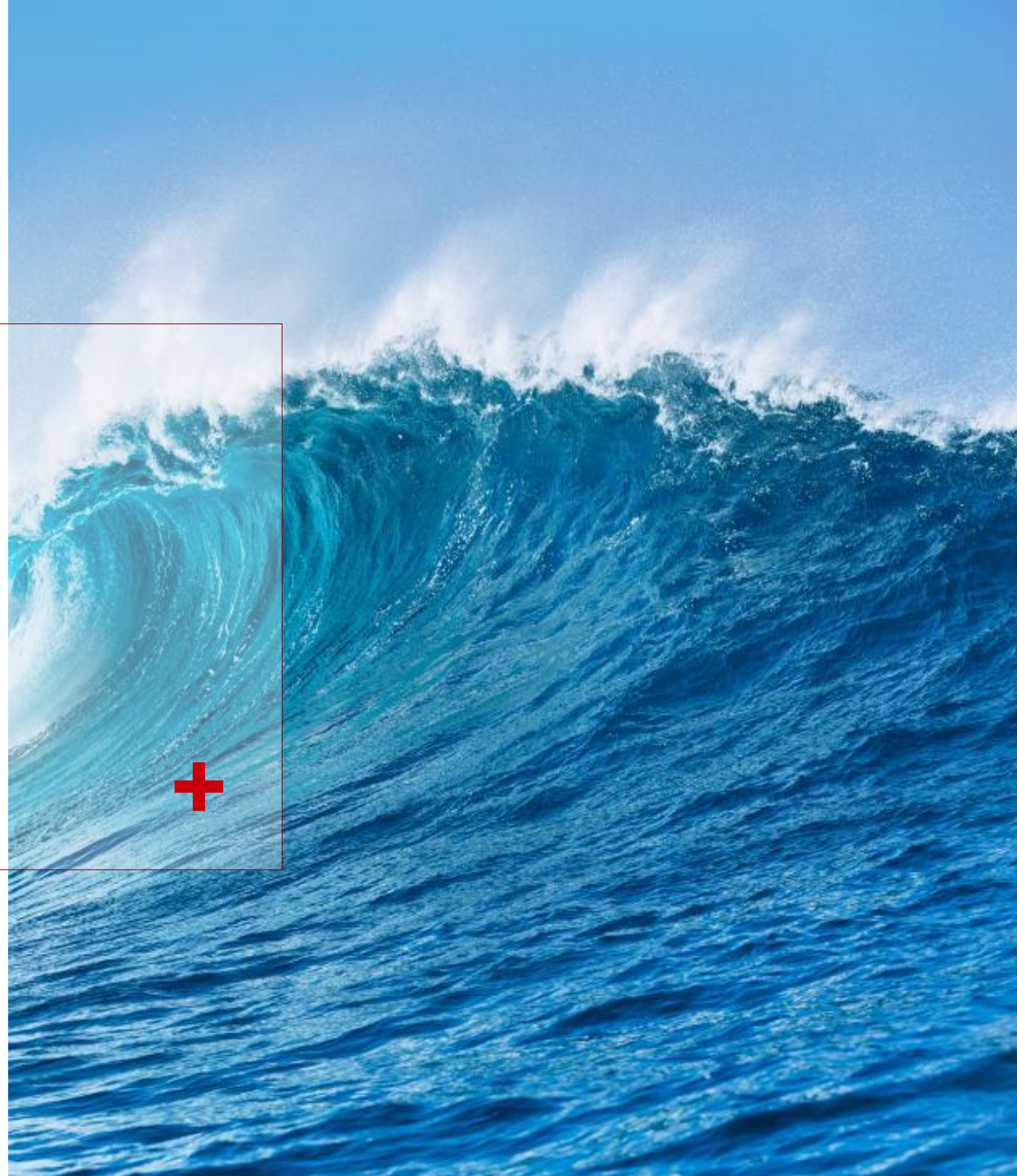
Atlas 950 SuperPoD super node cluster
Industry's largest super node, large-scale liquid-cooled data center



HUAWEI

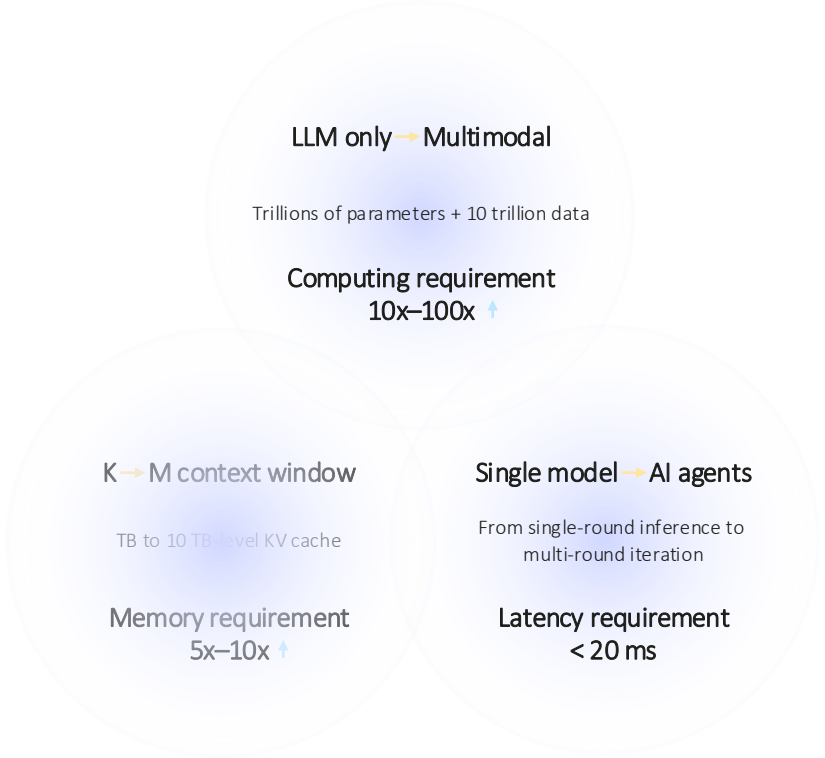
02

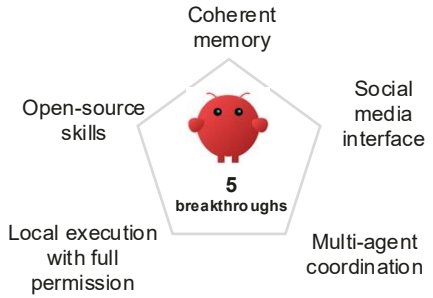
Huawei AI Data Platform

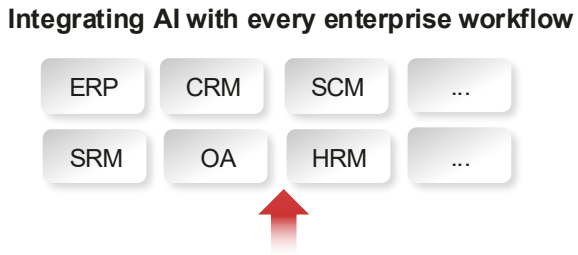


Strategic Hybrid for Sovereign AI

- ***Unpredictable Workload, Non-sustainable Advantage, Unmanageable Cost, Inconsistent Work Quality, Privacy and Security Challenging, …***
- Building IT strategy for Agentic AI and Inference economics
  - ➢ Token cost have dropped 280-fold in two years, yet some enterprises are seeing monthly bills in the tens of millions
  - ➢ The existing infrastructure strategies are not designed **to scale AI to production-scale deployment**
  - ➢ Shifting to strategic hybrid: on-premises for consistency & competence & sovereignty … and edge for immediacy, cloud for elasticity and scale with **a single unified data platform**
  - ➢ Multi-Agent is **the new composable platform**







HUAWEI

Data Explosion in the Agent Era

Storage opportunity

- Domain-specific language models
- AI training datasets – (PB-scale)
- Feature stores
- Trajectories/Decisions
- Vector knowledge base and RAG
- Model checkpoints

Data value continues to rise

Conventional pathology

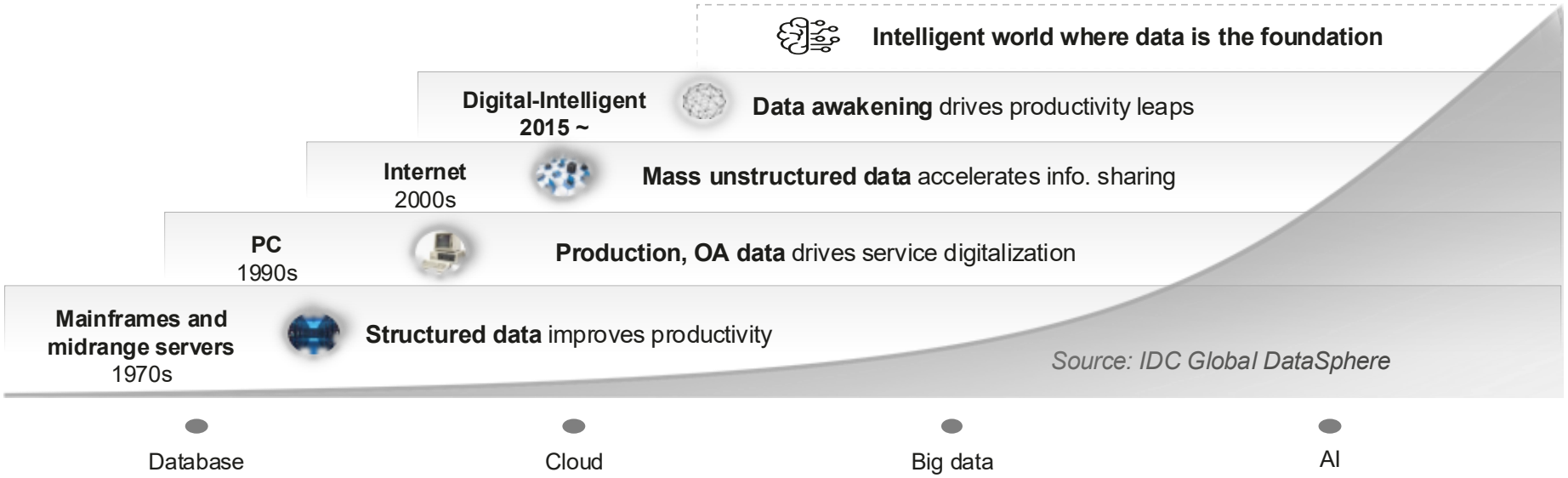
Pathologist shortage in China
120,000

Digital pathology

Slice-examining speed per pathologist
100 → 300 slices/day

Intelligent pathology

AI for personal health
Real-time slice examining



HUAWEI

Physical Infrastructure Impact – Token Economy : Data-Centric

AI infrastructure's success is measured by how fast STORAGE can feed GPUs with relevant DATA.

The most important metrics in the AI era are shifting from traditional KPIs to **AI pipeline metrics**:

Traditional Storage Metrics

- Capacity
- IOPS
- Reliability

AI-Pipeline Metrics

- **GB/s per GPU**
- **TTFT (Time to First Token)**
- **TPS (Tokens Per Second)**
- **PB-scale datasets**
- **TB/s cluster bandwidth**
- **Billions of vectors**
- **Data ingestion rate**
- **Global context and memory**

Infrastructure impact

Network upgrades:

- Ultra high-speed
- Fiber optical networks
- RDMA

Storage upgrades:

- Inference Context Memory Storage
- DPU Capability
- Global Namespace
- KV Cache Acceleration
- High-speed Data Path
- Massive Scalability

OceanStor A series storage

HUAWEI

Enterprise Data Platforms – the Foundation for Intelligence

Application Impact

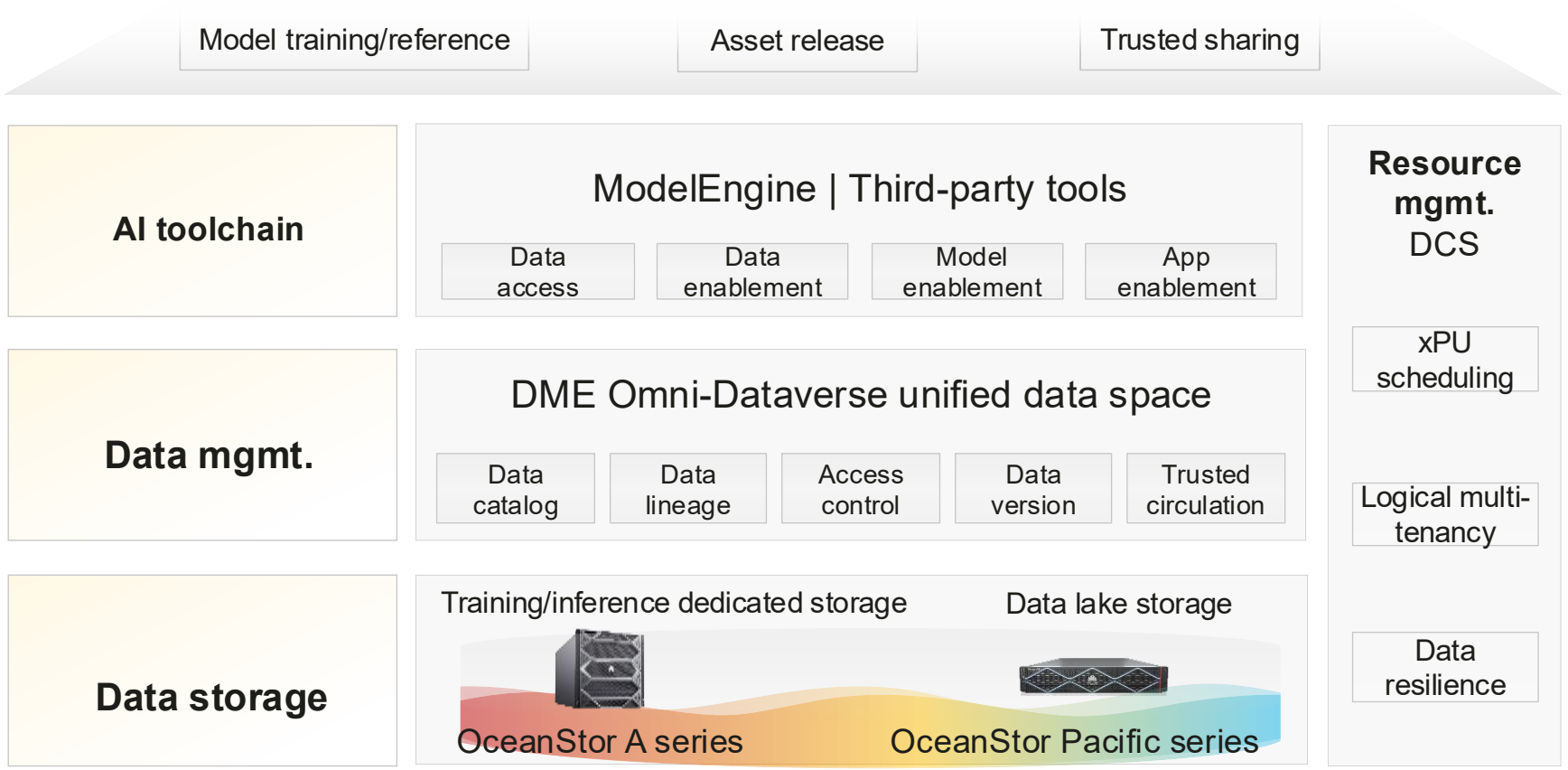
All applications require data to be:

- Discoverable
- Persistent
- Searchable

Infrastructure impact

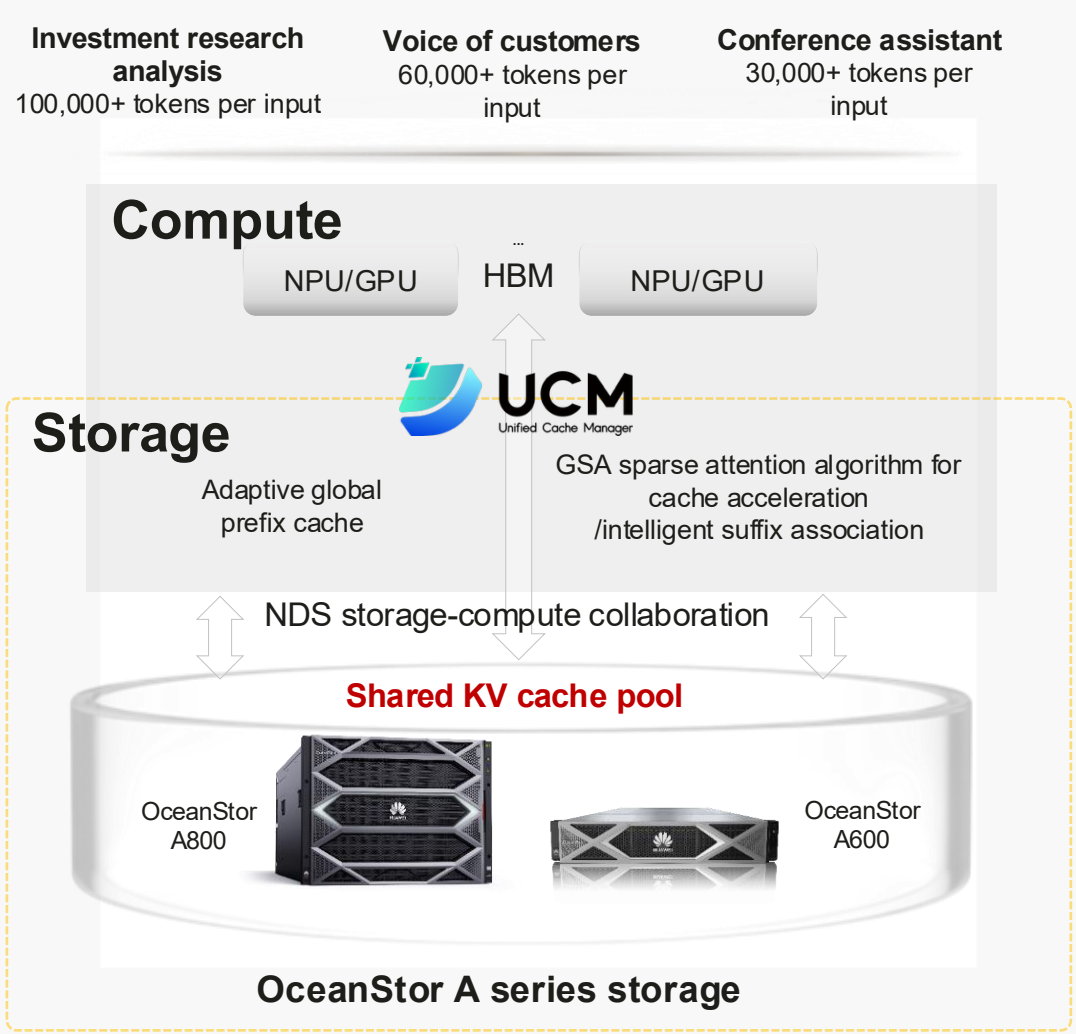
Enterprises must store:

- context memory
- training datasets
- embeddings
- real-time telemetry
- prompts
- Logs
- … …



HUAWEI

UCM: Unified Cache Manager for KV cache Management



Sufficient storage for long-sequence KVs

Up to 90% lower TTFT

For long sequences, Prefix Cache is used to eliminate repeated computations via queries.

The TTFT is reduced by up to 90%.

Multi-round inference support

Concurrent multi-round inference support

2x higher throughput (TPS)

The intelligent suffix association for acceleration based on user habits improves the throughput by more than twice.

Sharing support in multi-node inference

Avoiding repeated computations of 50% + KV cache

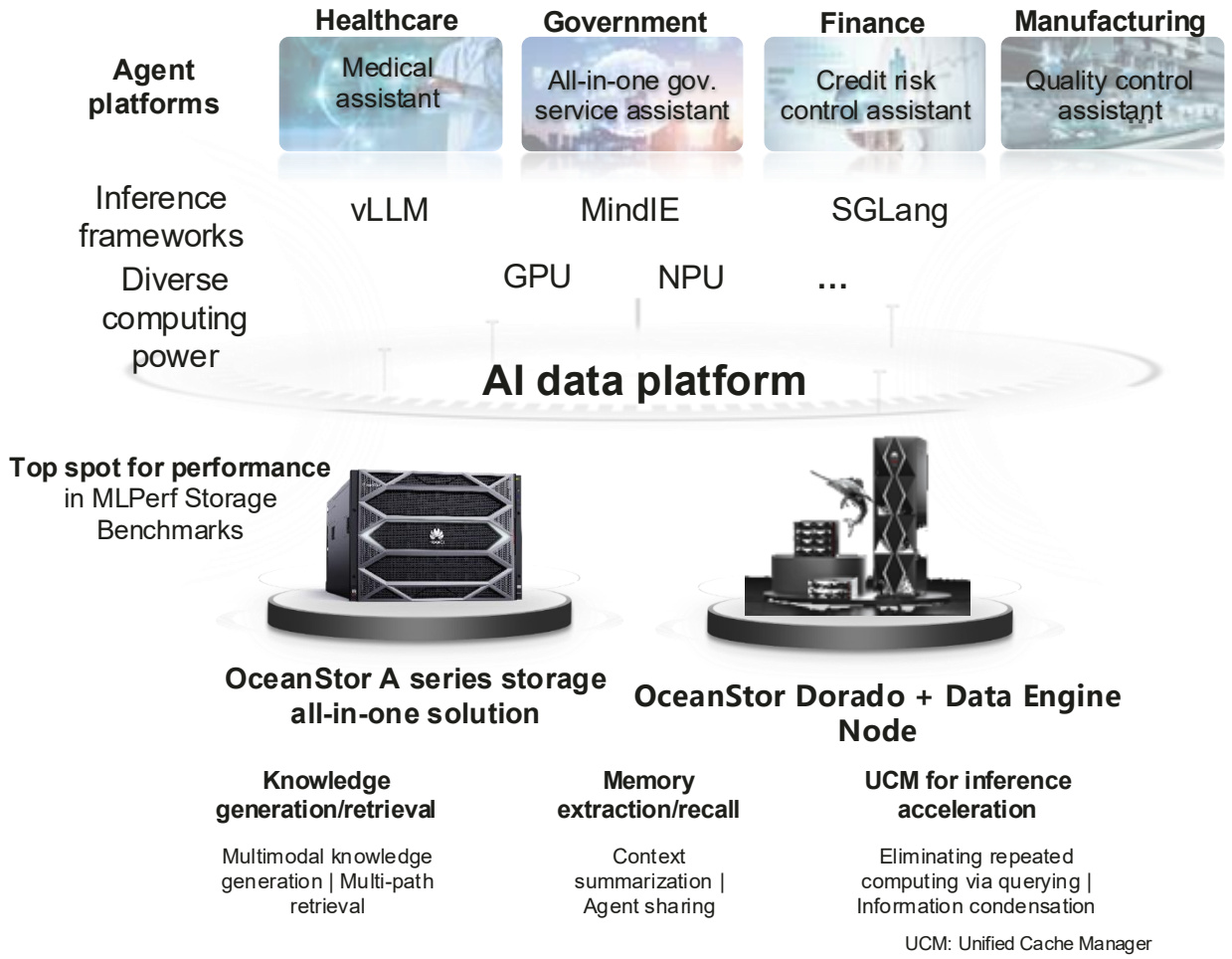
Persisted KV cache avoids repeated computations, enables multi-node sharing, improves KV cache hit rates, and boosts inference efficiency.

HUAWEI

AI Data Platform

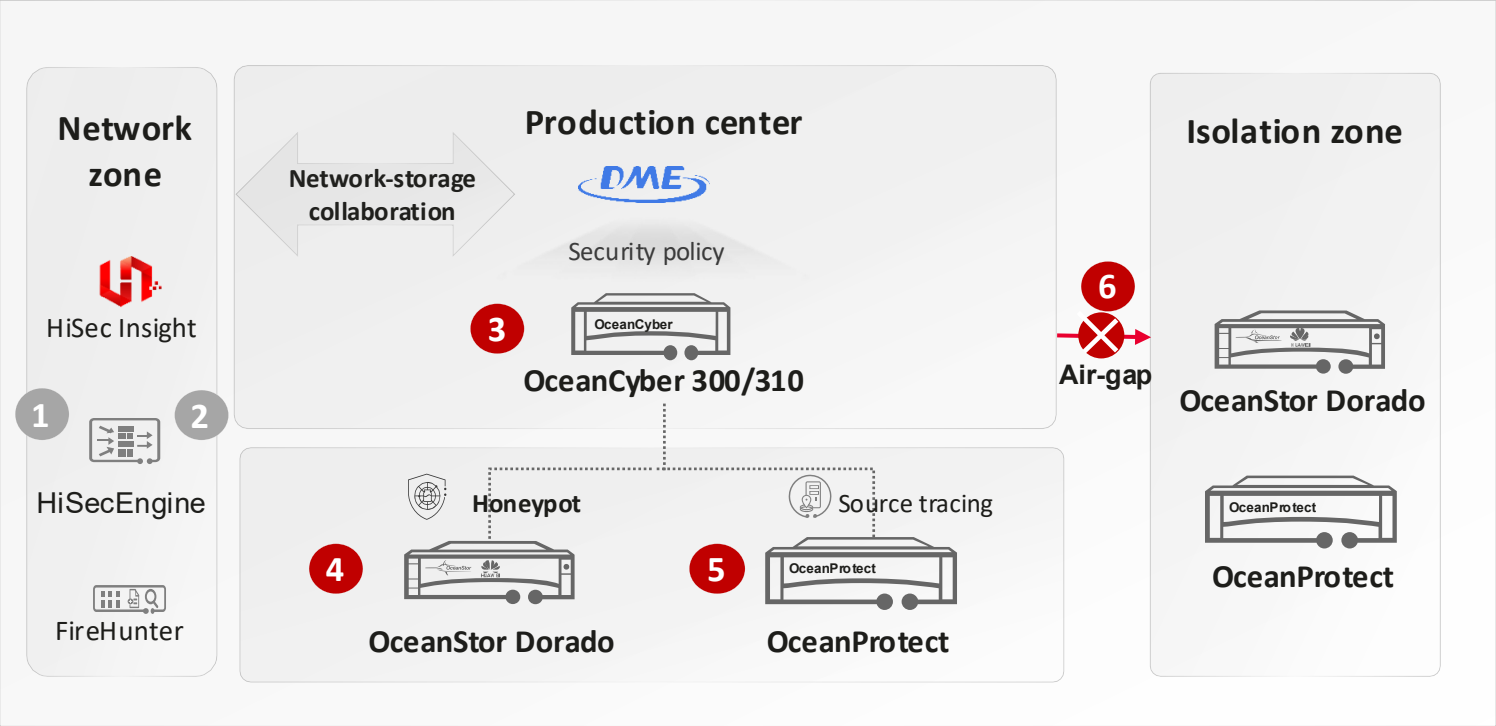
Storage transforms from a **passive repository** into an **active intelligence layer** that continuously feeds, organizes, and enriches data for AI systems.

- All-in-one Data, Insight, and Agent Platform
  - Insight-to-action workflows
- Unified Data Lakes for AI
  - Petabyte-exabyte scale storage
  - Data lake-house architecture supporting both analytics and AI
- Vector and Embedding Storage
  - Storage for billions of embeddings
  - High performance indexing and metadata storage
- Storage for AI Training & Model Pipelines
  - High-throughput NVMe storage feeding GPU cluster
  - Parallel file systems and distributed storage
- Intelligent Tiered Data Lifecycle Storage
  - Auto-tiering for hot, warm and cold data



HUAWEI

Cybersecure Storage



Key Technologies

Network

1. Anti-intrusion
2. Anti-proliferation

Storage

3. Detection & Analysis
4. Immutable Storage
5. Fast & Reliable Backup
6. Air-gap Isolation

- **Cyber-resilient storage platform**
  - AI-driven ransomware detection
  - Anomaly detection in file access
  - Automated isolation
  - Immutable storage and ransomware-proof backups
  - Air-gapped fast recovery
- **AI-driven data integrity monitoring**
  - Abnormal data access detection
  - Early ransomware pattern detection
- **Data protection integrated with storage**
  - Storage platforms include built-in backup, disaster recovery and cyber-resilience capabilities

HUAWEI

AI Storage: Leading Performance to Accelerate the Entire AI Training Process, and 30% Higher Cluster Utilization

AI Training Cluster Availability

30% » 60%+

Data Loading	CKPT Saving		Loading CKPT	
30 min → 1min	Minute-Level	Second-Level	Hour-Level	Second-Level

OceanStor A800

500 GB/s per enclosure

10 million IOPS per enclosure

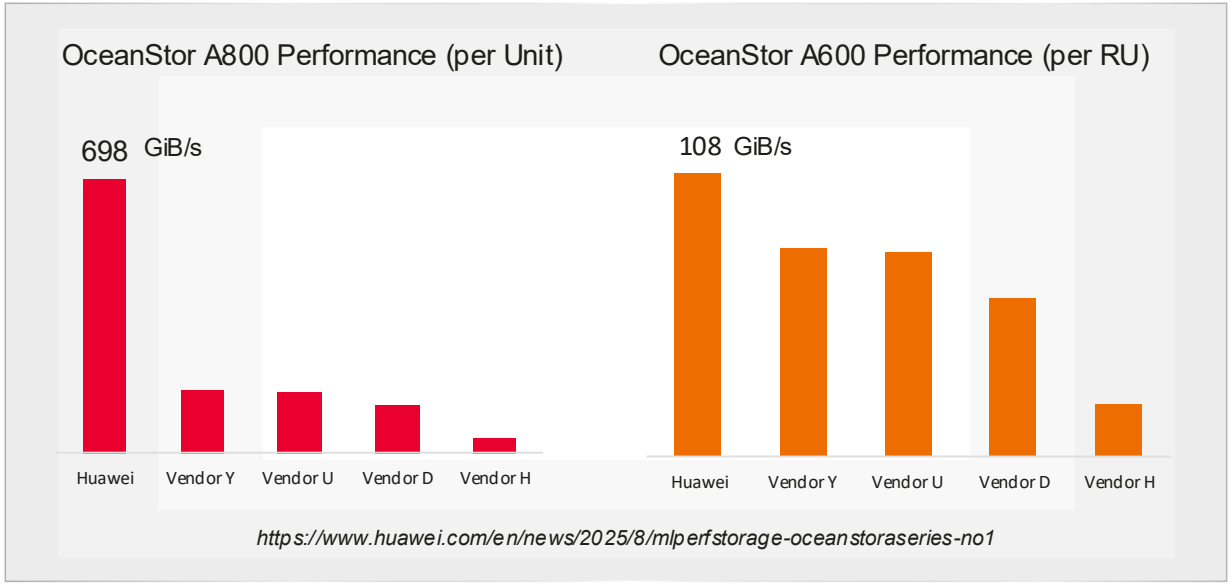
OceanStor A600

180 GB/s per enclosure

4.8 million IOPS per enclosure

No.1

MLPERF™ 2025 Storage Benchmark Test



HUAWEI

众行远

Go Far Go Together