HUAWEI

Striding Towards the
Intelligent World White Paper

Data Storage

New Data Paradigm Unleashes the Power of AI

Building a Fully Connected, Intelligent World

FOREWORD

AI foundation models are currently a topic of hot debate. In fact, the growth of machine intelligence looks quite similar to the growth of human civilization. Humans have been walking the earth for hundreds of thousands of years, but civilization has only existed for a fraction of that time. The key to civilization is the emergence of words. The words have enabled us to record and pass down experience and knowledge for group learning, replication, evolution, and development. This has contributed to the rapid development of human civilization over the past thousands of years.

AI is similarly driving humanity into a new era of civilization – one of machine-powered intelligence. Today, machines already have good algorithms that enable them to learn. But what are they learning from? What materials are they using? There is no AI without sufficient data.

If we have only a methodology but lack knowledge bases and corpuses, AI foundation models are simply not practical. We must feed AI foundation models with knowledge bases and corpuses so that they can develop a self-learning brain and generate bots serving as consultants, programmers, and customer support for different service scenarios.

Although computing power is critically important in AI systems, it is also important to store information on digital storage and turn it into knowledge bases for better utilization.

Any enterprise that wants to be a leader in the coming AI era must have advanced data infrastructure underpinned by data storage.

For over ten years, Huawei has invested heavily in data storage to produce a portfolio of cutting-edge offerings

for 25,000 customers in over 150 countries and regions worldwide, for sectors like carrier, finance, government, energy, healthcare, manufacturing, and transportation. The Striding Towards the Intelligent World – Data Storage white paper was made possible through extensive communication with industry experts, customers, and partners. This paper details today's hot topic of AI foundation models and focuses on five perspectives: new apps, new data, new resilience, new technologies, and energy saving. It provides viewpoints about data storage trends and suggestions for action. I believe this very meaningful research will bring together more industry forces to drive the data storage industry forward.

Over the past three decades, data storage has been a premium foundation for high-value data. As we usher in an era when data generated by new technologies and applications is continuously pouring into the data ocean, Huawei Data Storage will continue working closely with all parties in the industry to create a better future for data storage.

Dr. Peter Zhou
President, Huawei IT Product Line

CONTENTS

Executive Summary

AI foundation models have surpassed our wildest imaginations, propelling us into a world of unparalleled intelligence. The three elements of AI are computing power, algorithms, and data. Computing power and algorithms are tools used to serve AI foundation models, while the scale and quality of data truly determine the height of AI. Data storage enables disparate information to be collated into corpuses and knowledge bases. Together with computing, data storage is becoming the most important part of infrastructure for AI foundation models.

Intelligent enterprise applications, especially AI foundation models, have become as crucial and widely used as mainstream database applications and are poised to surpass them in the near future. Typically, with every application transformation, comes a corresponding evolution in the architecture of data infrastructure. Reliable, performant, and shared data storage is the optimal data infrastructure for databases such as Oracle Database. However, as new intelligent enterprise applications continue to push their limits, a new data paradigm is forming.

This report provides the following outlook on enterprise data storage:

---

1. To take AI to the next level, AI foundation models require more efficient collection and preprocessing of massive amounts of raw data, higher-performance training data loading and model data storage, and more responsive and accurate industry inference knowledge bases. A new AI data paradigm, represented by near-memory computing and vector storage, is rapidly gaining momentum.

2. Big data applications have undergone the stages of statistics collection and trend prediction, and are now moving towards the stage of supporting precise real-time decision-making and intelligent decision-making. The new data paradigm represented by near-memory computing will greatly improve the analytical efficiency of the lakehouse big data platform.

3. Open-source distributed databases are serving more mission-critical enterprise applications. A new architecture that features high performance and reliability is being created based on distributed databases and shared storage.

4. Multi-cloud has become the new normal for enterprise data centers. Enterprises' on-premises data centers and public clouds complement each other. Cloud computing is shifting from a closed and full-stack construction model to an open and decoupled one. This new model facilitates multi-cloud application deployment and sharing and centralized management of data and resources. Container-based cloud-native applications are shifting from stateless to stateful, so data storage needs to be able to support new cloud-native applications in addition to enhancing resource provisioning efficiency. The storage as a service (STaaS) business model is expanding its reach beyond public clouds and into enterprise data centers.

5. 80% of new enterprise data will be unstructured. AI foundation models are accelerating the use of unstructured data in production and decision-making systems. All-flash scale-out storage is the best choice of data infrastructure for mass unstructured data.

6. AI foundation model applications aggregate massive amounts of private domain data, exposing enterprises to increasing data resilience risk. Therefore, there is an urgent need for a comprehensive data resilience system with intrinsic resilience of storage.

7. All-flash storage features high performance, superb reliability, and low TCO. It can be used to build all-flash data centers to replace both high-performance and large-capacity HDDs.

8. AI foundation models are transforming the compute-storage architecture of data centers from CPU-centric to data-centric. A new system architecture and ecosystem are being built.

9. AI technologies are becoming increasingly integrated into data storage products and their management, greatly improving the service level of data infrastructure.

10. Energy-saving initiatives for data storage are seeing wider implementation. Data storage typically accounts for over 30% of a data center's energy consumption. Energy consumption indicators are now being integrated into construction standards.

A new architecture for data infrastructure is emerging to adapt to new intelligent enterprise applications like AI foundation models. We provide the following advice on how to build optimal data infrastructure for foundation models:

1. Shift your digital transformation focus from application innovation to collaborative innovation of both applications and data infrastructure. This will enable you to fully unleash the potential of data.

2. Establish a collaborative design team for new applications and data storage to create the best data infrastructure together, so that you can better harness the power of AI, big data, distributed databases, and cloud-native applications.

3. Consistently promote the decoupled storage-compute architecture for new applications to take full advantage of both applications and storage.

4. Explore a new data-centric storage system architecture, build a new data paradigm, and upgrade data storage to support near-memory computing, new data formats, new data access protocols, and high-performance application data caching. A combination of the new architecture and paradigm can greatly improve the efficiency of new digital applications, enabling them to achieve a higher service level while reducing the costs of rebuilding traditional applications.

5. Deploy mission-critical applications in your on-premises data centers. If needed, use public clouds to deploy innovative applications that may have uncertainties. For cloud construction, use the hierarchical decoupling model that features multi-cloud application deployment and sharing and centralized management of data

and resources. Enable application development teams and data storage teams to jointly develop best practices for container-based cloud-native applications. Select appropriate business models, managed service providers (MSPs), and storage vendors based on your business strategy, operational status, and forecasts.

6. Accelerate the application of all-flash storage. Use technologies such as new data-centric storage architecture, high-density hardware, data reduction, system convergence, and governance of mass unstructured data to reduce the TCO of mass data storage and build green data centers.

7. Incorporate a data storage team into a joint data resilience team and develop intrinsic resilience standards for data storage, thereby building the last line of defense for data resilience.

8. Actively try AI-enabled data storage products and management, and improve team members' AI skills to improve the service level of data infrastructure.

New Apps

Outlook 1

AI Foundation Models

No AI Without Sufficient Data

With the rapid development of the GPU computing power and AI algorithms, the era of AI foundation models represented by generative AI has arrived. AI foundation models are now able to demonstrate higher levels of intelligence in conversations and knowledge feedback than ordinary human beings, and they will bring fundamental changes to the Internet, industrial manufacturing, finance, media, and other industries. Currently, we are entering the first boom in AI foundation models. Decision-makers in enterprise IT construction need to proactively embrace changes and explore approaches to enable efficient production by taking advantage of foundation models.

Trends

AI has developed rapidly and far beyond expectations

No one could have imagined that AI foundation models would bring such fundamental changes to human society at the end of 2022 with the release of ChatGPT by OpenAI. Before 2022, AI was used as a niche tool in specialized fields such as computer vision and Internet recommendation to help perceive and understand the world. Today, however, AI is considered an all-round expert in all domains that can generate and create new content for the world. It is able to learn, understand, and think, and it can write code, make important decisions, and generate new ideas, greatly improving our productivity in everyday life and at work.

Take Midjourney, a generative AI drawing software also released in 2022, as an example. It can produce incredible works rivaling the best artists based on simple, text-based descriptions in a minute. Based on the GPT model, Midjourney is now in use in 33 design fields, including wool weaving, mobile phone cases, blind boxes, refrigerator stickers, greeting cards, toys, cartoon profile images, company logos, other logos, movie posters, carpet grains, tile patterns, furniture modeling, and so on.

Foundation models are evolving into industry-specific models

AI foundation models have gained more popularity and are applied to various industries. In the past, different AI models needed to be developed and trained in different scenarios, resulting in huge investments and low efficiency. In addition, development had to start from the most basic model, meaning that technical requirements were high. Today, foundation models have removed the bottlenecks in AI generalization and industrialization, and provide more general and basic capabilities for upper-layer applications. Enterprises no longer need to develop and train independent basic models from scratch based on various usage scenarios, but can instead integrate private domain data accumulated from production services into mature foundation models to implement professional model training, while at the same time ensuring accuracy and resilience in specific fields. Huawei estimates that 95% of medium- and large-sized enterprises will build their own industry models based on domain-specific data, such as enterprise accounts and personal financial information of banks, self-driving video records of automobile enterprises, and health data of healthcare institutions.

It is also found that enterprises are very cautious about using public foundation models. Enterprises cannot tolerate the disclosure of their confidential data due to the use of foundation models, because data is considered as high-value assets. According to Cyberhaven's survey of 1.6 million employees from various industries, 4.2% of employees have

copied their companies' data to ChatGPT, and confidential data accounts for 11% of the content pasted to ChatGPT. 100,000 employees have uploaded 199 confidential documents, 173 pieces of customer data, and 159 pieces of data source code to ChatGPT within a week. "Generative AI was the second most-frequently named risk in our second quarter survey, appearing in the top 10 for the first time," said Ran Xu director, research in the Gartner Risk & Audit Practice. "Information entered into a generative AI tool can become part of its training set, meaning that sensitive or confidential information could end up in outputs for other users. Moreover, using outputs from these tools could well end up inadvertently infringing the intellectual property rights of others who have used it."

Data determines the development of AI

Data, computing power, and algorithms are the three elements that make up AI foundation models. With the rapid development of AI, enterprises are utilizing the same computing power, which is dominated by NVIDIA's graphics processing units (GPUs) and Huawei's Ascend AI processors. In addition, enterprises also tend to use the same algorithms, with the Transformer model infrastructure and development frameworks like PyTorch, TensorFlow, and MindSpore now dominating the industry. For these reasons, what determines the future development of AI is data, and therefore enterprises need to think about how to unlock the value of data.

First, the volume of the training data is critical. As large language models (LLMs), Meta's LLaMA has 65 billion parameters and 4.5 TB of training data, while OpenAI's GPT-3.5 has 175 billion parameters and 570 GB of training data. Although LLaMA has less than half the parameters of GPT-3.5, it outperforms the latter on most benchmarks. Moreover, LLaMA is on a par with Chinchilla, a model with 70 billion parameters from DeepMind, and PaLM, a model with 540 billion parameters from Google. This shows that the volume of training

data is more crucial in improving AI precision than the model's parameter scale.

Second, the quality of the data is of equal importance. The reason why AI foundation models may generate incorrect, ambiguous, meaningless, or inauthentic results is that they lack standard, complete, timely, and high-quality data sources. For foundation models, the key to solving this problem is to improve the quality of the data that foundation model vendors can obtain from the public. For industry-specific foundation model training and scenario-specific inference applications, the effectiveness of the model depends on the quality of industry-specific private domain data, including original enterprise data and incremental data updated in real time, that is, the industry knowledge base.

Data storage is becoming a critical infrastructure for AI foundation models

Data storage serves as the carrier of data and has become a form of critical infrastructure for AI foundation models. Data storage is essential for data collection, preprocessing, training, and inference of AI foundation models, because it determines data capacity, data read efficiency in the training and inference processes, data reliability, and data resilience.

First, the collection efficiency of mass raw data. In this phase, data is collected and summarized across regions as well as online and offline. To be specific, data is transmitted between data centers, edges, and clouds in various protocol formats. It is estimated that PB-level data collection usually takes three to five weeks, accounting for 30% of the entire AI foundation model process. Storage systems need to be able to provide efficient aggregation, multi-protocol interworking, and on-demand capacity expansion to accelerate data collection and reduce the idle time for subsequent analytics.

Second, data preprocessing efficiency. PB-level collected and crawled raw data is read by CPUs and GPUs for parsing, cleaning, and deduplication before it can be used for model

training. The data preprocessing phase requires at least three full data reads and migrations, which consume more than 30% of all CPU, GPU, network, and memory resources. Huawei estimates that the preprocessing phase takes more than 50 days, which is more than 40% of the full pipeline of AI foundation models. Storage systems need to be able to implement near-data processing to enhance data processing efficiency and reduce resource waste.

Third, data access efficiency in the model training phase. In the startup phase, the training can proceed only after the GPU server has randomly read tens of thousands of small files. Storage systems need to provide tens of millions of IOPS to shorten the idle time for loading training data on GPUs. In addition, the high failure rate of GPU server hardware hinders model training. On average, model trainings encounter hardware failures of GPU servers every 2.8 days. If a model is retrained from the beginning each time, the training will be extended indefinitely. To alleviate the problem, dozens or even hundreds of periodic checkpoints are set during the process to save intermediate process data so that resumable training can be performed upon the occurrence of a fault. The GPU is suspended during the checkpoint and continues to run only after the data is completely saved. Therefore, storage systems need to provide hundreds of GB/s of write bandwidth to shorten the GPU idle time.

Furthermore, the timeliness and accuracy of the inference phase. When a foundation model is used for inference, private data continuously generated by enterprises is transmitted to the model to prevent issues such as irrelevant answers and fabricated information. Retraining or fine-tuning these data is time-consuming and costly. There are ongoing efforts in the industry to create an industry knowledge base that can import incremental and real-time data updates to foundation models, requiring a new type of storage from which key information can be efficiently retrieved.

Last but not least, in the long data chain of AI foundation models,

attacks have evolved from traditional ransom-requesting malware to more modern types. Huawei estimates that data ransomware attacks occurred every 11 seconds in 2022. These attacks brought enterprises not only ransom losses, but also damage to their reputations and business opportunities, as well as lawsuits, and additional labor and time costs. It is estimated that the collateral damage costs enterprises more than 23 times the actual losses in ransom fees.

Modern data attacks, different from traditional ones, typically add noisy data, for example, adding violent and ideologically distorted contents to training data. As a result, the model quality deteriorates, inference accuracy is affected, and model hallucinations occur, ultimately interfering in enterprise decision-making. This is where storage systems are needed for data resilience.

In a word, the emergence of AI foundation

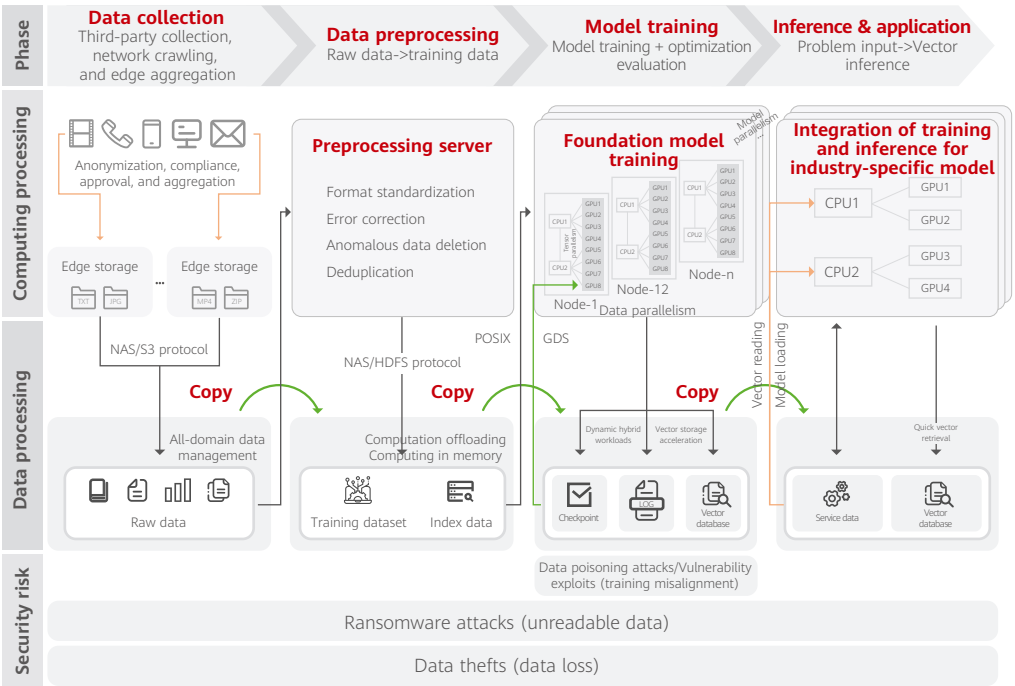


Figure 1: AI foundation model lifecycle

models presents three goals for data storage, including mass unstructured data management, tenfold higher performance, and intrinsic resilience of storage. To be specific, not only does the storage system need to provide EB-level scalability, bus it also has to deliver tenfold higher performance with hundreds of GB/s of bandwidth and tens of millions of IOPS.

Data fabric helps collect and manage cross-region mass data

Data fabric leverages the global data view for data visualization and management and on-demand cross-region, cross-system data scheduling, to achieve optimal data layout without affecting the services and performance, and enable seamless collection and mobility of valuable data from multiple sources, supercharging management efficiency of mass and complex data and shortening the end-to-end(E2E) AI training period.

In addition, data fabric can be used for on-demand dataset filtering, in which data profiles are identified by location, creation time, and labels to help simplify data tiering and classification, improve data governance, and meet scenario-specific requirements of AI foundation models. Intelligent tiering of hot, warm, and cold data can be implemented to deliver optimal TCO

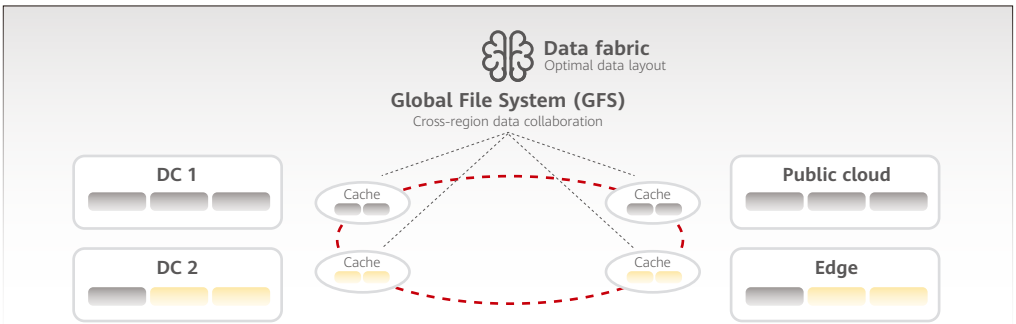


Figure 2: Global data view and scheduling

by identifying access date, format and type, and access frequency.

High-performance access of AI foundation model data demands all-flash data storage

High-performance data read/write is key to improving GPU utilization and streamlining the training pipeline.

Conventional HDD storage cannot meet the needs for fast access and large-scale data processing. Flash storage, however, features high-speed read/write and low latency and takes advantage of breakthroughs in stacking layers and chip types to reduce costs significantly, making it the ideal choice for processing AI foundation models. When the read/write ratio is 6:4, HDD storage provides between 50,000 to 100,000 IOPS, whereas all-flash storage delivers over 1 million IOPS. The tenfold boosts in data read/write performance reduces the idle time for computing, network, and other resources, accelerating the rollout and application of foundation models. Huawei concludes that in the scenario of GPT-3 using 100 PFLOPS of computing power, if the storage read/write performance is enhanced by 30%, the compute utilization will be improved by 30%. As a result, the training period will be reduced from 48 days to 36 days, shortening the overall training by 32%.

The high-performance compute and storage architectures evolve from CPU-centric to data-centric

The emergence of AI foundation models has displaced the convention of centering computing power on a CPU and popularized a heterogeneous computing method across CPUs, GPUs, NPUs, and DPUs. Currently, the model training still accesses memory through CPUs, but the memory bandwidth and capacity have hit a bottleneck due to the slowdown in CPU development.

The IT industry adopts a solution using high-speed interconnect buses such as the Compute Express Link (CXL) to decouple compute, storage,

and memory resources and make them form separate sharing resource pools, so that GPUs can directly access memory and storage resources at a higher speed, greatly improving data loading and forwarding efficiency of AI foundation models and driving the architecture evolution from CPU-centric to data-centric.

Currently, storage systems still center on CPUs, but to provide higher-performance data services for the efficient training and reference of AI foundation models in the future, storage will use high-speed interconnect buses for data interaction and shift towards data-centric.

New data paradigms accelerate the training and reference of AI foundation models with new data architectures

The rise of AI foundation models triggers the convergence of high computing power + big data + foundation models, driving the innovation in new storage paradigms like vector storage and near-memory computing.

【 Vector storage 】

External knowledge bases are becoming a necessary part of foundation model applications. This shift is an example of a new data paradigm, in which knowledge bases work as a new type of external storage that is also known as vector storage. Adhering to the principle of "everything can be vectorized", the vector storage converts all knowledge content and all question input into vectors. It then extracts features from multi-modal and high-dimensional unstructured data, efficiently retrieves the most related information to the user query (the shortest distance between vectors means the most relevant information), and input questions and related information into the foundation model to generate accurate answers. As an external component of AI foundation models, vector storage can store data for long periods for anytime access, and also allow convenient updates.

It is predicted that unstructured data stored in vector knowledge bases will account for 30% of global data by 2025. The vector storage is very likely to become the base for data of all AI foundation models. Despite the clear advantages, vector storage needs to, first, provide a retrieval speed of 10,000 operations per second to support fuzzy search and exact matching from tens of billions of vectors, and second, support cross-region and cross-modal data index search to enable efficient aggregation and association of data about the same subject from pictures, voices, and text sources.

【 Near-memory computing 】

The pre-processing of AI foundation models involves data movement between storage, memory, and CPUs, and consumes 30% of compute and network resources. To reduce extra system overheads caused by data movement, near-memory computing and supplementary computing enhanced by storage implement computation offloading to the storage for in-line computing, and perform part of the data filtering, aggregation, and transcoding on the storage side. This will help alleviate the pressure of CPU, GPU, network, and memory resources by 20%, reducing the dependency on GPUs to some extent.

The intrinsic resilience of storage is the last line of defense

Foundation models are trained based on mass data containing sensitive information such as personal data and core business secrets. Data storage, as the final carrier of data, must be absolutely secure, highlighting the importance of intrinsic resilience. The data protection capabilities of foundation models must be enhanced to function as the ultimate line of defense for data resilience. The intrinsic resilience of storage includes software and hardware resilience, data resilience, and resilience management.

AI foundation models are adopting data lake construction mode, sharing the same data sources as HPC and big data

Enterprises need large amounts of raw data when using AI foundation models, HPC, and big data. These applications share the same raw data source accumulated by enterprises, including production and transaction data, data from scientific experiments, and data on user behavior. It is cost-effective and highly efficient to have foundation models use the same data source as HPC and big data because one copy of data can be utilized in different environments. If they could not share data, independent clusters would be constructed repeatedly, and this uses a large number of storage devices and equipment rooms and generates more data silos, which increases construction and O&M costs and reduces the efficiency of data transfer. Huawei's customers, which include the Peng Cheng Laboratory, Wuhan AI Computing Center, China Mobile, and China Telecom, have now started to use the data lake construction mode.

HPC, big data, and AI foundation models are shifting towards the data lake construction mode. However, the larger data scale and workloads of AI foundation models means that a 10x more performant storage system with a much larger capacity is needed. This is driving enterprises to upgrade the performance and expand the capacity of their existing data lake storage systems. In addition, data lifecycle management is also an important factor in the construction.

The one-stop training/inference HCI appliance is the mainstream deployment mode for enterprise segments

Due to the limitations on technologies, talent, and funds, enterprises face a series of challenges, such as device integration, model deployment, resource utilization, and O&M.

To address these challenges, Huawei released the one-stop training/inference HCI appliance featuring out-of-the-box service, elastic compute and storage scaling, and one-click model deployment. This is the optimal solution for enterprises looking to

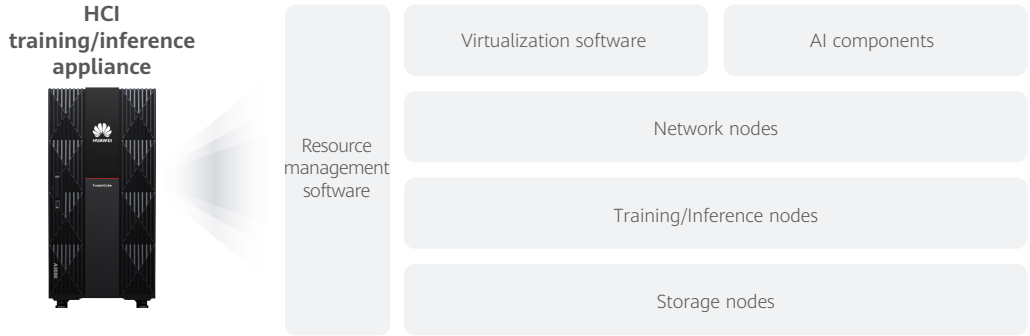


Figure 3: HCI training/inference appliance

embrace industry-specific models. The all-in-one delivery mode integrates storage, network, and diversified computing, meaning deployment can be completed within 2 hours, and eliminating the need for adaptation, optimization, and system construction from scratch. In addition, compute and storage nodes can be flexibly expanded, and efficient resource scheduling and virtualization technologies achieve the full utilization of all resources. Moreover, diversified and pre-trained foundation models can perform fine-tuning and inference based on the enterprise's private knowledge base. Therefore, an environment that focuses on segmented applications, including customer service robots, office assistant robots, and programmer robots, can be built on the device side. This lowers the threshold for enterprises to deploy AI foundation models, making them inclusive for all.

Suggestions

Build a reliable foundation model infrastructure that attaches equal importance to compute and storage

AI foundation models are widely used in various industries. Data quality and scale are critical factors determining the potential of AI. In addition to computing power stacking, enterprises should consider employing a storage-centered data infrastructure that provides governance of mass

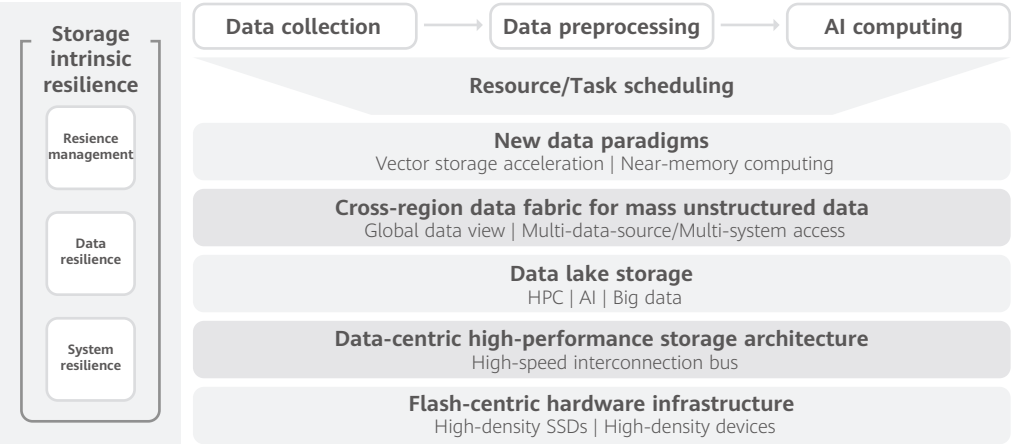


Figure 4: AI data infrastructure

unstructured data, optimal throughput performance, and robust data resilience.

Adopt data lake construction mode for foundation models that share the same data sources as HPC and big data, and upgrade the performance of the current data lake storage

The data lake helps break down data silos and enables elastic capacity expansion for mass data, reducing TCO. In addition, the performance of the existing data lake storage should be upgraded on demand in order to meet the increasing real-time performance demands of AI foundation models.

Build forward-looking data infrastructures that include all-flash storage, data-centric architecture, data fabric, new data paradigms (vector storage and near-memory computing), and intrinsic resilience of storage

All-flash storage greatly improves performance and accelerates the development and implementation of AI foundation models. Meanwhile, the data-centric architecture decouples and interconnects hardware resources, accelerating on-demand data mobility. Emerging data processing technologies, such as data fabric, vector storage, and near-memory computing, minimize the threshold for enterprises to integrate

and use data and meet efficient resource utilization requirements, while make it easier for industries to access AI foundation models. The intrinsic resilience system will protect enterprises' core private data assets and enable enterprises to easily use AI foundation models.

The one-stop training/inference HCI appliance is recommended for enterprise segments

The hyper-converged infrastructure (HCI) is highly recommended for segmented industries due to its compact design. It integrates data storage nodes, compute (training/inference) nodes, switches, AI platform software, and management and O&M software, delivering one-stop services. HCI also reduces the costs of a large amount of adaptation, optimization, and system construction.

Create a professional technical team with enhanced professional skills of AI foundation models, particularly in storage aspects

Enterprises should cultivate more professionals who have in-depth understanding and practical experience in AI foundation models, especially in related storage services.

New Apps

Outlook 2

Big Data

Big data applications are shifting from describing the past to making decisions for the future, and new data paradigms are driving increases in data application efficiency

Big data applications have been in development for more than 10 years. They no longer simply collect statistics on historical data, but stride towards proactive and intelligent decision-making. By optimizing the big data platform and infrastructure, enterprises can build leading data value mining capabilities and application efficiency to gain competitive advantages.

Trends

Data lake storage is key to facilitating big data applications to assist real-time, precise, and intelligent decision-making and driving big data platforms to use lakehouse architecture

The development of big data applications can be described as having three phases: traditional data application, predictive analytics, and proactive decision-making.

The first phase was the traditional data era from 2000 to 2012 when data technology was mainly used to describe historical phenomena more accurately. For example, it could be used to query historical bank details, carrier CDRs, customer churn rate statistics, or statistics on a city's electricity, gas, and water usage.

The second phase was the predictive

analytics era from 2012 to 2022 when big data applications predicted what would happen in the future based on historical statistics to assist managers with judgments and decision-making. For example, applications could be used to target customer profiles and make recommendations for credit cards or mobile services, monitor public opinion, and assess disasters.

In the third phase that starts in 2023 and will continue for the foreseeable future, big data will enter the era of proactive decision-making. By instantly analyzing what happened and what is happening, big data applications can make precise decisions in real time. For example, in urban traffic management, big data technologies can collect and analyze vehicle location data and traffic volume data in real time, in order to implement the automatic optimization of suggested traffic routes and reduce congestion.

During this process, the evolution of the big data analytics platform also went through three phases:

Traditional data warehouse era: Enterprises used data warehouses to build a subject-oriented, time-variant set of data which would accurately describe and collect historical data, helping make analysis and informed decisions. However, only TB-level structured data could be processed.

Traditional data lake era: Enterprises used the Hadoop technology to build data lakes and process structured and semi-structured data in order to use historical data to predict what would happen in the future. However, this caused data silos as data lakes and warehouses would coexist and data had to be transferred between them, which hindered real-time and proactive decision-making.

New lakehouse era: Enterprises start to optimize IT stacks to achieve real-time and proactive decision-making, and this attempt has greatly accelerated the evolution of the big data platform into a new lakehouse architecture. To be specific, enterprises work with storage vendors to decouple storage and compute for the big data IT stacks. With the use of data lake storage, the data lakes and warehouses share the same copy of data without

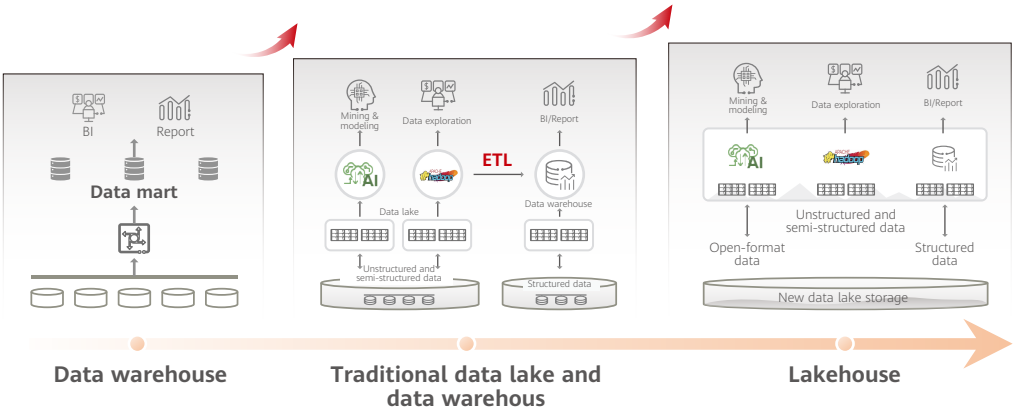


Figure 5: Three-phase evolution of the big data analytics platform

the need for data transfer between them, thereby facilitating real-time and proactive decision-making.

China Mobile cooperates with Huawei data storage team to research on decoupled storage and compute for big data, and focuses on applying the lakehouse architecture to make big data services more convenient. The data lake storage of over 180 PB is constructed in nine regional data centers for large-scale pilot programs. It provides analysis and processing capabilities of more than 200,000 jobs per day and more than 200 million data records per second, leading the world in terms of scale.

Access for diverse workloads is the basic feature of new data lake storage

New data lake storage provides a unified storage pool by integrating data sources of different applications in multiple fields including data science, AI, and knowledge mining. Therefore, it must be able to handle diverse data access requests from various application tool sets, including diverse data access protocols and different I/O loads.

Data lake storage supports near-data computing, and the new data paradigm allows big data to support applications more efficiently

When the number of computing clients reaches tens of thousands or even hundreds of thousands and data volume reaches tens of PBs, the key to accelerating data query and analytics will be optimizing metadata query performance. To do this, a high-speed cache is added between the big data platform and persistent data storage to function as a data acceleration engine. With near-data computing, queries involving hundreds of PBs of data can be shortened from 10 minutes to just 10 seconds, making real-time data analytics (T+0) possible.

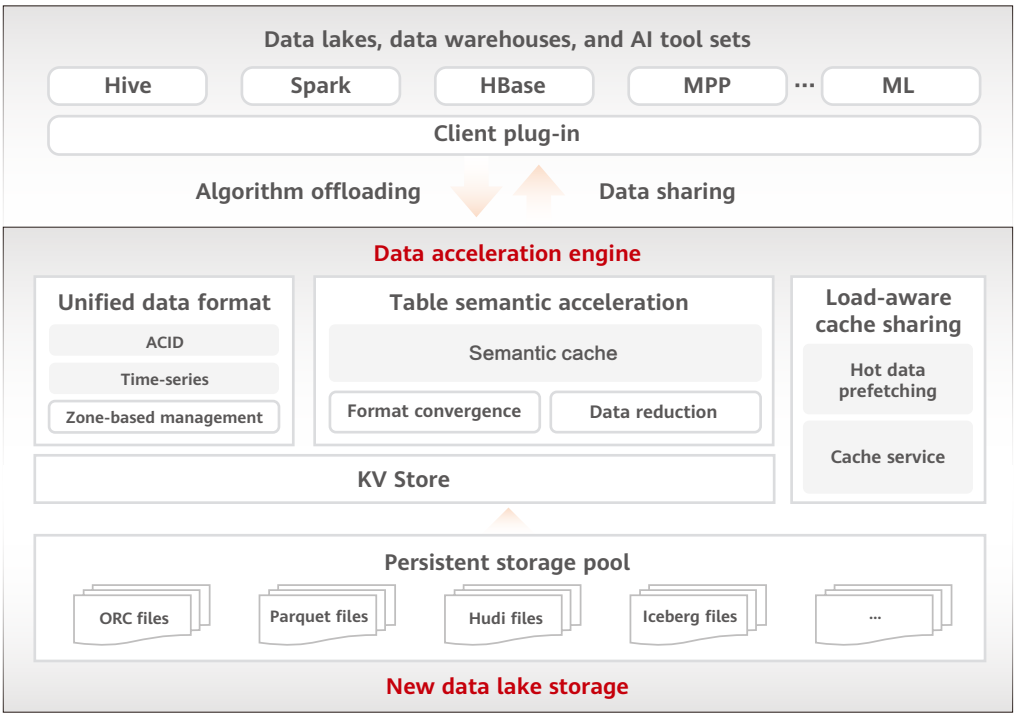


Figure 6: Real-time big data analytics achieved thanks to near-memory computing

Suggestions

Enterprises should focus on innovative collaboration between big data platforms and storage to promote real-time data analytics

Enterprises should shift their focus from the construction of big data platforms towards implementing innovative collaboration between big data platforms and storage. This will enable sharing and converged analytics of real-time data and offline data on existing big data platforms, realize real-time data updates, analytics, and supply, as well as allow data of different types, sources, and formats to be managed and processed in a unified manner.

Set up a team to design joint solutions for big data platforms and storage and develop a mechanism for regular teamwork

Currently, enterprises' big data platform teams mainly focus on building stable and reliable big data computing platforms, as well as exploring application scenarios by computational modeling, analysis, and mining of mass data. Setting up a team to design joint solutions for big data platforms and storage, with a mechanism for regular teamwork, helps cover more steps of big data analytics than just data computing. Such a team can explore and optimize the entire process of big data analytics, from data generation to data computing, storage, and application, and develop a powerful and innovative engine to drive exploration of data applications.

Explore the new data paradigm to achieve real-time (T+0) decision-making as big data platforms evolve towards lakehouses on the basis of decoupled storage-compute architecture

The construction of new data lake storage and new data paradigms (represented by near-memory computing) will help enterprises transform their big data platforms into lakehouses, and achieve real-time and proactive decision making.

New Apps

Outlook 3

Distributed Databases

The increasing popularity of the Internet and rising costs are driving core systems to embrace distributed databases, which are shifting from a coupled to a decoupled storage-compute architecture

Open-source databases MySQL and PostgreSQL are the top 2 in the global database market. Open-source databases are reshaping enterprise core systems. Furthermore, to ensure smooth service operation, the decoupled storage-compute architecture has become the de facto standard for distributed databases.

Trends

Distributed databases built based on an open-source ecosystem are replacing traditional core systems in order to better suit service changes, achieve higher efficiency at a lower cost, and facilitate long-term technology evolution

The development of digital and mobile technologies has greatly changed the interaction channels between enterprises and their customers. Internet applications such as mobile apps have become the best medium for triggering customers' purchase behavior. This leads to rapid business growth, but also brings unpredictable and fluctuating workload surges to core systems. The core system must be sufficiently elastic to ensure that resources are quickly expanded during peak hours to keep services running smoothly while

idle resources are released during off-peak hours to prevent the waste of resources.

High O&M cost is another factor driving enterprises to reconstruct traditional core systems. According to an Oracle user survey conducted by Rimini Street, a company providing third-party support for Oracle databases, 97% believe that cost is the biggest challenge associated with using Oracle databases, and 35% are turning to open-source or other non-Oracle cloud databases.

According to 6sense, MySQL tops the database ranking with a 42.95% market share. Another open-source database PostgreSQL ranks second, and Oracle Database only takes the third place.

The decoupled storage-compute architecture has become the de facto standard for distributed databases because it keeps services running smoothly

Stability is the top consideration for core databases. Performance, functionality, and energy efficiency are also important appraisal criteria.

In the early stages of using distributed databases, both the pilot service scale and the data volume are small. To minimize the initial investment, many enterprises deploy both database applications and data on the same server. This is what we call a coupled storage-compute architecture. However, it is vulnerable to the risks that come with putting all of your eggs in one basket. Therefore, some enterprises choose to use multiple servers and redundant data copies to temporarily solve service stability issues. However, as the scale of a distributed database expands, the volume of data and the quantity of servers also increase exponentially. Data redundancy can result in a significant waste of investment. As the volume of data continues to grow, the synchronization of redundant data will consume more and more network bandwidth. Especially in the multi-site disaster recovery (DR) architecture, network bottlenecks may cause data loss if a disaster occurs.

As these problems become increasingly prominent, distributed database construction has gradually evolved from a coupled to a decoupled

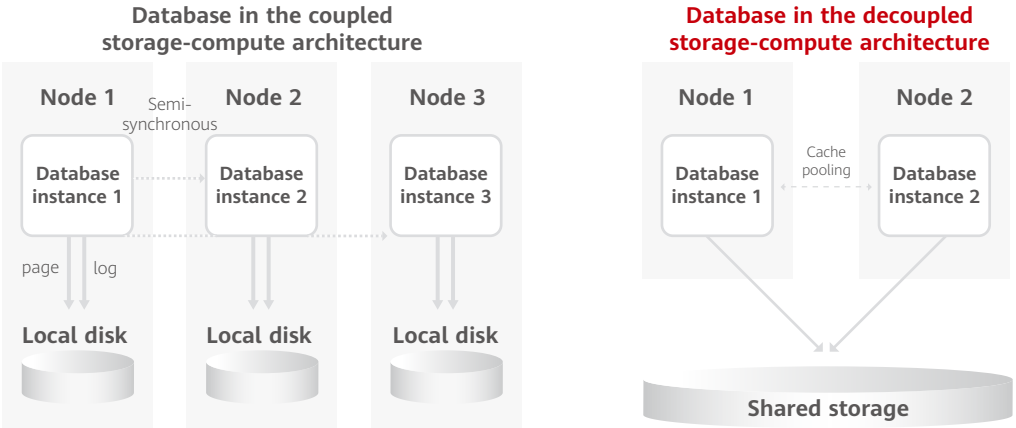


Figure 7: Database architecture evolution from coupled to decoupled storage-compute

storage-compute architecture. In the decoupled storage-compute architecture, enterprises can use performant, reliable, and shared enterprise-level all-flash storage pools to ensure high data availability. The decoupled storage-compute architecture isolates applications from data, eliminating the need to use multiple redundant data copies for high availability. In addition, the powerful and mature DR capabilities of storage systems can compensate for the insufficient DR capabilities of open-source databases. Most importantly, the decoupled storage-compute architecture has been comprehensively tested through its use in traditional core systems, resulting in the development of mature product systems and O&M expertise. Enterprises can focus on how distributed databases can help them drive business growth without concerns about frequent O&M issues.

Currently, major banks around the world have built new core systems using distributed databases that adopt the decoupled storage-compute architecture. New mainstream database solutions, such as Amazon Aurora, Alibaba PolarDB, Huawei GaussDB, and Tencent TDSQL, have also shifted to the decoupled storage-compute architecture. It has become the de facto standard for distributed database construction.

Distributed databases are driving the development of a new data paradigm

Open-source databases, such as MySQL and PostgreSQL, are deployed on standalone servers. Unlike Oracle RAC, open-source databases cannot coordinate multiple database nodes to simultaneously read from and write to the same database. As a result, distributed databases have significant bottlenecks in performance expansion. However, by using professional storage devices to provide shared data access across nodes and implement a consistent cache layer between database nodes, distributed databases can also provide the same multi-primary capabilities as Oracle RAC. For example, GreatDB and TeleDB have worked with Huawei storage to implement multi-primary capabilities through Huawei's Cantian database storage engine, and this improved database performance by up to a factor of 10.

Suggestions

Consistently promote the decoupled storage-compute architecture for distributed databases

Although the industry has seen numerous examples of distributed databases built on the coupled storage-compute architecture, the decoupled storage-compute architecture is now a necessary choice from technology, O&M, and evolution perspectives. If enterprises plan to build a distributed database, they should adopt the decoupled storage-compute architecture to prevent repeated construction and resource wastage. Enterprises that have already used the coupled storage-compute architecture should gradually shift it to the decoupled architecture to prepare for future cost reduction, efficiency improvement, and continuous expansion.

Encourage the database team and storage team to jointly incubate a new data paradigm

The new data paradigm built with

distributed databases centers on the idea that database software is no longer a one-size-fits-all solution. To meet enterprise requirements, a collaborative effort between the database and hardware infrastructure is essential. Therefore, the database team should not work alone when building enterprise core systems. Instead, the database and storage teams should work together to build databases and core systems so that they can fully leverage both the software and hardware advantages and establish a new data paradigm.

New Apps

Outlook 4

Cloud Native

Innovation and cost efficiency requirements are propelling the transformation of cloud-native, driving cloud-native infrastructures towards an open and decoupled multi-cloud architecture

As 89% of enterprises are constructing multi-cloud IT architecture, cloud-native containers have become the ideal technology foundation and support infrastructures such as storage to help enterprises construct multi-cloud architecture extensively. Through open and decoupled architecture, cloud-native infrastructures pave the way for enterprises to integrate different optimal services and realize cross-cloud resource sharing.

Trends

Driven by the need to accelerate innovation in applications, reduce costs, and increase efficiency, multi-cloud architecture has become the new normal for enterprise IT

Enterprise cloud computing infrastructures have shifted from a single-cloud to a multi-cloud architecture. Enterprise application and cost requirements cannot be met by any single cloud. Consequently, 89% of enterprises have turned to a multi-cloud IT architecture that incorporates both public and private clouds.

Enterprises are keeping their stable, key services locally and building new services or services with uncertain traffic on clouds.

The IT infrastructure provides two key capabilities for multi-cloud architecture:

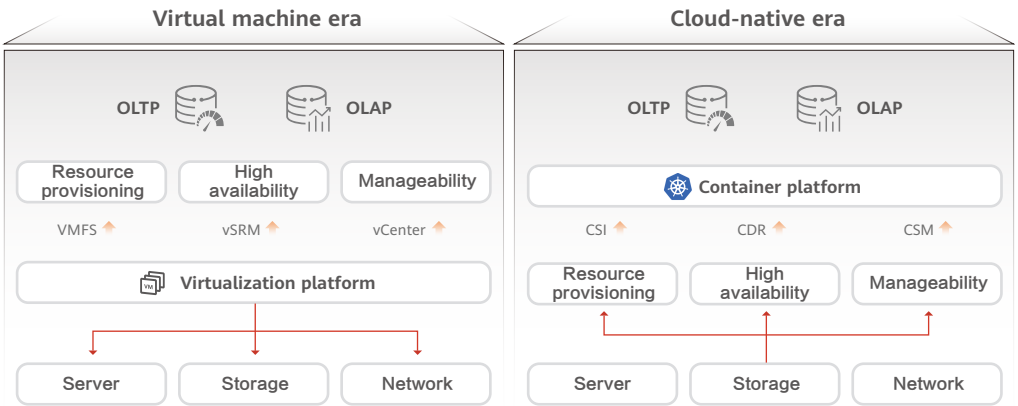


Figure 8: Comparison between the virtual machine era and the cloud-native era

cross-cloud data mobility and cross-cloud data management. For example, Huawei and NetApp storage devices support cross-cloud tiering and backup, enabling data to always use the most cost-effective storage service. Cross-cloud data management allows users to see overall data status through a global data view and schedule data to applications that generate the most value.

As container-based cloud-native applications are increasingly used in key services, storage support for containers will become a necessity

According to a CNCF survey, 96% of customers are building container platforms, and 95% of new applications are being deployed in containers.

As the migration of critical enterprise applications to containers is well underway, there are currently 61% of containerized applications that are stateful applications, which require their interaction data to be stored persistently. Therefore, more highly reliable enterprise-level storage will be needed to offer support. For one thing, storage must support the configuration and extension of container storage interfaces so as to facilitate fast resource provisioning and the DR of containerized applications. For another, storage needs to collaborate with new cloud-native applications running on the containers to build the best practices.

Cloud-native infrastructures are becoming more open and decoupled

Globally, cloud infrastructure construction can be broken down into two categories: full-stack closed construction and open and decoupled construction. As enterprises continue pursuing multi-cloud construction and increasingly demand optimal services with lower cost and higher efficiency, openness and decoupling is establishing itself as a mainstream mode.

This structure makes sharing hardware resources over multiple clouds possible and enables data to flow between them freely, which is a major way of showcasing the strengths of a multi-cloud architecture.

Optimal services, from hardware and platforms to applications, are often provided by different vendors. Therefore, open and decoupled construction can help enterprises build optimal IT stacks. For example, the most common AI foundation model suppliers in the market, such as OpenAI and Meta, cannot compete with IT giants such as NVIDIA, DDN, and Huawei in terms of hardware infrastructure capabilities. No vendor can provide the optimal end-to-end AI training/inference solution on its own. So, when deploying AI

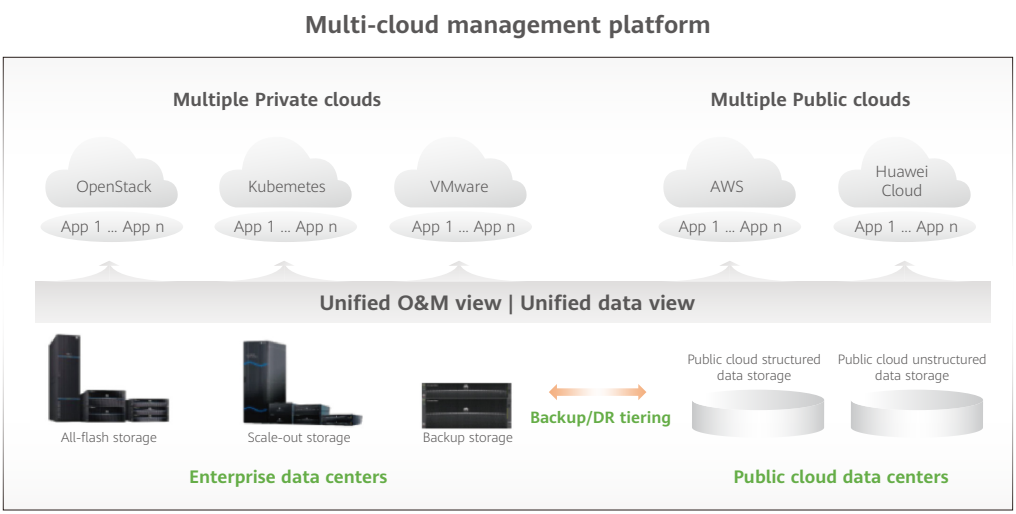


Figure 9: Enterprise multi-cloud IT architecture

training/inference clusters, enterprises often choose open and decoupled architectures so that they can select the optimal hardware as well as training/inference models.

Data storage will strike a balance between CAPEX and OPEX business models

The business model of clouds is shifting from CAPEX-based to OPEX-based, which is simultaneously reshaping the business model of data storage for enterprises. Enterprises are shifting their attention away from assets, features, and functions, and placing a greater emphasis on the business outcomes delivered by procurement services. As a result, the SLA- and result-based OPEX business model is gaining popularity. Furthermore, due to mounting global economic pressures, enterprises are becoming more sensitive to the cost of trial and error. This makes the OPEX business model, which offers high elasticity and low initial costs, a less risky and more preferable choice for enterprises.

However, as data storage scales and contract duration extends, subscription-based services may not always be the most cost-efficient choice for construction. A combination of CAPEX and OPEX business models may therefore be the optimal solution. However, for large-scale enterprises with stable service revenues, the CAPEX model tends to be the preferred choice.

Suggestions

Migrate innovative services that have uncertainties, along with emerging services like OA to public clouds, while retaining core services in their on-premises data centers

Innovative and emerging services require IT systems to deliver elastic scaling and on-demand resource application and release. Public clouds provide benefits that are less costly, more reliable, and offer good elasticity. To preserve the competitiveness of core services, enterprises must develop strong

capabilities in IT platform R&D and prioritize service confidentiality. Therefore, retaining core services in enterprises' on-premises data centers can truly stimulate IT R&D and innovation, implement independent data control and operation, and prevent cloud vendor lock-in.

Container platform teams should collaborate with storage teams to build agile and highly reliable container platforms and develop best practices for containerized applications

Container teams should work with storage teams to build highly reliable container platforms and develop interface standards for containers and storage devices, so that storage resources and containers can be provisioned simultaneously in an agile manner. Furthermore, enterprises can gradually develop their own best practices for containerized application construction during the cloud-native transformation. Through continuous optimization, enterprises can accumulate valuable experience and knowledge that will prove invaluable in navigating the multi-cloud era.

Develop an open and decoupled architecture for cloud construction

Thanks to its open and decoupled architecture, cloud-based IT infrastructure can optimize services, costs, and flexibility, and is becoming the mainstream choice for enterprises. It is crucial that enterprises open up procurement models to select the optimal component vendors. In addition, cloud platform vendors must open up their interfaces and take the lead in establishing interconnection standards with infrastructure providers.

Select the suitable business models based on enterprise and service requirements

When it comes to selecting a business model, enterprises should carefully take their business development status into consideration, and make flexible choices based on the overall benefits and risks involved.

New Data

Outlook 5

Unstructured Data

Unstructured data accounts for more than 80% of new enterprise data and is increasingly important to production and decision-making

According to Huawei's Global Industry Vision (GIV) report, the global data volume will reach 180 ZB by 2025, of which over 80% will be unstructured data. 25% of unstructured data at that time is predicted to be used for production and decision-making, with that number soaring to 80% by 2030.

Trends

New applications will give rise to mass unstructured data, and AI foundation models will accelerate the use of unstructured data in production and decision-making systems

Unstructured enterprise data is rapidly growing from PBs to EBs with the development of new technologies and applications such as 5G, cloud computing, big data, AI, and High Performance Data Analytics (HPDA). This data includes a mix of video, image, and file types. A major carrier can process up to 15 PB of data on average every day. In HPDA scenarios, a single DNA sequencer, remote sensing satellite, or autonomous-driving training car can generate 8.5 PB, 18 PB, or 180 PB of data every year, respectively.

Production and decision-making systems have also started using unstructured data, and the adoption of AI foundation models across industries is set to expedite this evolution.

In the financial industry, one major bank is using its financial big data platform and AI analysis platform to facilitate online real-time credit extensions. This helps the bank shorten the loan application process from 15 minutes to just 1 minute and improve risky applicant identification accuracy by 80%. In the healthcare industry, the Pangu drug molecule model has learned the chemical structure of 1.7 billion drug-like molecules to help reduce both drug R&D periods and costs.

A growing number of industries are looking for professional-grade scale-out storage solutions for enterprise data centers to efficiently and securely store unstructured data

First, storage must provide sufficient capacity to store more data with the minimum cost, footprint, and power consumption.

- Enterprises need to use mass unstructured data. So, storage scale and scalability become top considerations. A single cluster must support thousands of nodes to simplify storage resource allocation and management. In addition, capacity and performance must increase linearly as the number of nodes increases.

- The traditional multi-copy technique is also a capacity barrier to unstructured data storage. To optimize storage space utilization, the data reduction techniques provided by professional scale-out storage are needed, such as erasure coding (EC), deduplication, and compression. Replacing general-purpose servers with high-density storage hardware also helps reduce footprints, power consumption, and O&M complexity to achieve optimal TCO.

- Industry players can use professional scale-out storage that integrates software and hardware to provide enterprise customers with end-to-end solutions that deliver high reliability, performance, and scalability. This simplifies deployment, management, and services.

Second, storage must deliver efficient, on-demand, and policy-driven data

mobility both within and between data centers.

- Multi-region and multi-form data center deployment needs the data fabric function to enable the sharing of data resources across different regions, clusters, vendors, and forms, with efficient and on-demand data scheduling through a graphical topology view.

- In a data center, professional scale-out storage can be used to implement hot, warm, and cold data tiering that automatically relocates data to different tiers for optimal ROI.

Third, storage must ensure premium data usability to easily handle hybrid workloads involving video, audio, image, and text data.

- Mass unstructured data serves a wide variety of applications. Therefore, all-flash scale-out storage specially designed for hybrid workloads is the best choice to prevent data silos and provide both high bandwidth for video, audio, and file scenarios and high IOPS for image, retrieval, and query scenarios. All-flash scale-out storage delivers significantly lower read and write latency than traditional HDD storage for faster data processing.

- In scenarios that involve mass data and hybrid workloads, technologies powered by unstructured data often involve multiple access protocols (such as those for file, object, and HDFS access) in a single data processing flow. To ensure premium usability, professional scale-out storage should reduce data redundancy by implementing multi-protocol interworking without data copying.

- In addition to storing mass unstructured data, storage systems also need to manage it effectively, for example, by accelerating queries and retrieval based on metadata as well as identifying hot and cold data for proper data lifecycle management.

- Storage also acts as the last line of defense and so must ensure

high intrinsic resilience and reliability, through ransomware protection, DR, and backup functions.

Suggestions

Enterprise IT teams should strengthen their mass unstructured data processing capabilities

As enterprises use unstructured data more widely, especially in their production and decision-making systems, the ability to efficiently store mass unstructured data and extract its huge value to help make informed decisions becomes a key competitive edge. Enterprise IT teams therefore need to strengthen their mass unstructured data processing capabilities, and transform their structured data-centric capabilities to design, planning, and management of mass unstructured data.

Choose professional scale-out storage to build a foundation for mass unstructured data

To improve the efficiency of using mass unstructured data for production, use a professional scale-out storage system to build a global unified data storage foundation centered on unstructured data. It is best to choose a scale-out storage system that supports hybrid workloads, multi-protocol interworking (file, object, and HDFS), data reduction, high-density hardware, and all-flash configurations to ensure sufficient capacity, superb data mobility, and premium usability.

New Resilience

Outlook 6

Intrinsic Resilience of Storage

Trends

The risks associated with a lack of resilience continue to grow as we move to the AI era and begin to aggregate massive amounts of data. Consequently, enterprises are embracing both data resilience and network resilience as key parts of their protection systems

In 2023, AI foundation models, and ChatGPT in particular, gave rise to a new wave of AI technology development around the world. After being trained by foundation models, AI can make valuable inferences from the massive amounts of aggregated data. Data is the foundation of AI and also the core asset of enterprises. Data resilience is of paramount importance. Splunk's The State of Security 2023 Is Resilient report found that 66% of organizations had been victims of malicious attacks and more than 52% of them had experienced data leakage. Data resilience is now a pressing issue.

The ever-changing data resilience landscape has driven countries and regions to enact and revise laws and regulations in order to better address data resilience and privacy protection. In 2021, Singapore updated its original data protection legislative framework and released their Personal Data Protection Regulations 2021. In fact, the EU's Network and Information Systems Directive, the US's Federal Trade Commission Act, and Australia's Privacy Act have all been updated to include regulations on data resilience. Data resilience has become a core indicator of an enterprise or country's competitiveness.

Storage devices are an indispensable part of the data lifecycle, from generation and collection, to transmission, use, and destruction. They typically come with near-data protection and near-media control capabilities. Storage devices act as both a home for data and a safe. As such, they play a critical role in data resilience protection, backup and recovery, and secure destruction.

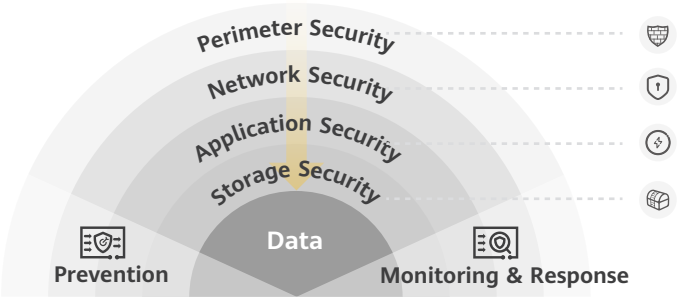


Figure 10: The defense in depth (DiD) model for data resilience

When it came to data resilience in the past, enterprises often prioritized security gateways and the application-layer security software, and chose to overlook storage and its important role in safeguarding data. It would be comparable to a jewelry store hiring guards and reinforcing their doors and windows, only to leave priceless jewels unprotected on a table inside. Therefore, it is necessary to prioritize storage to ensure data resilience.

The intrinsic resilience of storage builds the last line of defense for data resilience

The intrinsic resilience of storage is built upon an inherently resilient architecture. It has been designed to enhance both storage device resilience and data resilience.

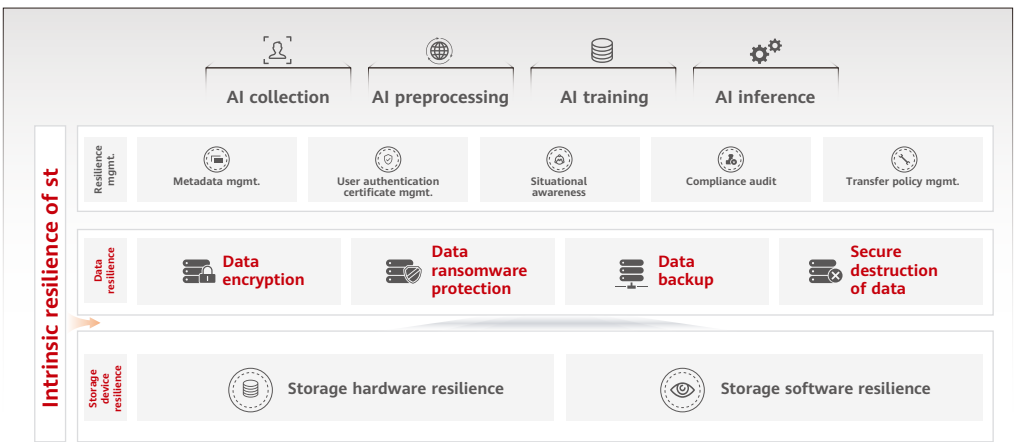


Figure 11: Intrinsic resilience of storage

Storage hardware resilience

A root key which functions similarly to an ID card is implanted on the hardware chip, so that each program in the system is authenticated before running, thereby ensuring that the system remains resilient and trustworthy.

Storage software resilience

Storage device resilience can be improved by building security R&D capabilities in compliance with applicable laws and regulations. Selecting high-value open-source software and using standard open-source software facilitate the reliable lifecycle maintenance of all software. Active community feedback and collaborative maintenance ensure the compliant and secure use of software.

Data encryption

Data encryption can be implemented at different layers, such as application software, databases, and storage systems. Application-layer encryption requires a lot of service reconstruction, and database encryption highly compromises performance. Storage devices can use self-encrypting drives (SEDs) to encrypt data, without affecting production services. This encryption method has the best cost-benefit trade-off.

Data ransomware protection

First, the production storage detects and intercepts ransomware by identifying abnormal reads/writes and calculating information entropy. Second, the write once read many (WORM) and secure snapshot features of the production storage protect data from tampering or unauthorized deletion. Third, the ransomware protection storage solution supports rapid recovery using local backups, and this prevents data loss. Fourth, an off-line data copy is retained in the air-gapped isolation zone, which shields access against ransomware attacks.

Data DR and backup

Data backup: Periodically copying important data to other storage systems in different backup locations ensures that this data can be recovered up to a specific point in time. Backup

capabilities can be improved based on service requirements to make the backup system compatible with new core ecosystems such as big data and data warehouses. Full backups of critical data defend against threats caused by human error, hardware faults, and ransomware.

DR: In scenarios without DR, self-built data centers can be used to implement dual DR for data and services so that the DR system can take over production data and services at any time. In scenarios which include existing DR systems, the active/standby DR setup can be changed to the active-active setup based on service continuity requirements. This ensures zero loss of critical data and defends against a series of threats such as natural disasters, power failures, and computer viruses.

Secure destruction of data

The data on storage devices can be permanently deleted so that it cannot be recovered, and this prevents sensitive data leakage during the resale or discarding of storage devices.

Suggestions

Include storage resilience capabilities in enterprise construction plans when working on network resilience protection projects

Storage has a natural advantage when it comes to resilience as it is deeply connected to the data itself. It can further provide unique data resilience capabilities such as isolation and recovery after network interception. Enterprise resilience teams mainly consist of network teams at present. These teams enact strict resilience policies by deploying network resilience devices like firewalls and blocking high-risk ports. In contrast, storage teams tend to focus more on the operations of normal storage resilience services and the planning and development of new storage resilience service technologies. It is recommended that enterprises add storage resilience capabilities to the resilience system construction projects.

Enhance the software and hardware resilience capabilities of storage devices to improve their overall protection capabilities

Storage devices have to be able to defend against attacks at the underlying storage system layer. By focusing on the structure and design of storage systems and enhancing the resilience capabilities of the hardware and software, enterprises can enjoy higher resilience for data and assets with the protection and recovery capabilities of the storage.

Prioritize the deployment of data resilience capabilities such as encryption and ransomware protection for storage devices

Storage devices leverage their advantageous connection to the data and support data-related intrinsic resilience features, which are deployed with optimal reliability, performance, and cost. Data can be encrypted using SEDs to comply with industry requirements and compromise less performance with lower service reconstruction costs. To cope with ransomware attacks, an E2E ransomware protection system must be established to ensure that data can be accurately detected, the system can quickly respond to ransomware attacks, and data can be recovered in a timely manner. In addition, DR resilience must be enhanced, and DR compatibilities with new core ecosystems must be improved on the basis of full DR for mission-critical services and full data backups.

New Technologies

Outlook 7

All-Scenario Inclusive
All-Flash Storage

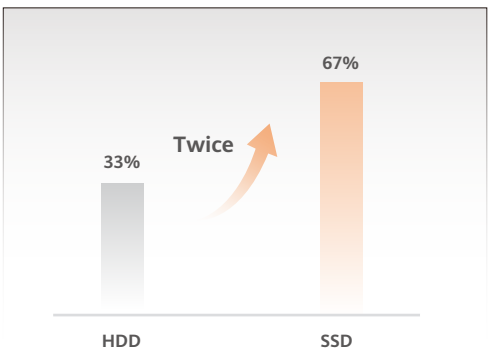
All-flash storage is here

According to Gartner, solid-state drives (SSDs) have surpassed hard disk drives (HDDs) in market share and shipments since 2022. The implementation of mass unstructured data into production decision-making systems signals the new era of all-flash storage that we are embracing.

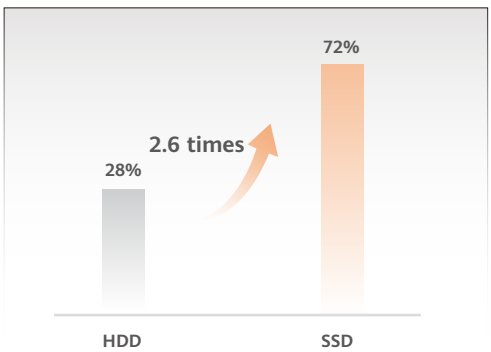
Trends

The global shipment of SSDs has far exceeded that of HDDs

In 2022, SSDs had over double the market share (over 65%) and shipment of HDDs, which strongly illustrates that enterprises are embracing all-flash storage.



Twice market share in 2022



2.6 times shipment proportion in 2022

Figure 12: SSD market share and shipment proportion

Higher-performance all-flash storage significantly improves enterprise efficiency and service experience

SSDs far outperform HDDs in storage performance. A single SSD has thousands of times higher IOPS than an HDD and can deliver millisecond- or even microsecond-level latency and high throughput. These characteristics make SSDs more suitable for the high-requirement scenarios associated with emerging services.

As data volumes increase sharply, enterprises will find it difficult to complete scheduled backup during backup windows (usually at night), even if it requires average performance. Compared with HDD-based backup systems, all-flash backup storage systems can deliver twice higher backup performance and four-fold higher recovery performance. In the past, backup systems mainly used HDDs to store cold data. Now these systems are gradually switching to all-flash backup storage for quicker backup and recovery.

All-flash storage has an obvious advantage in TCO over HDD storage

A higher number of cell layers, quad-level cells (QLCs), and penta-level cells (PLCs) will not only drastically reduce the price of a single SSD, but also bring a continuous decline to the storage cost of the same physical capacity.

NAND cells are a core component of enterprise-level SSDs, and often determine the cost of SSDs. QLCs and extra cell layers in 3D NAND are driving a steady decline in the equipment costs of all-flash storage. Most mainstream vendors now mass produce 176-layer 3D NAND cells, and multiple 200-layer (nearly double that of 2018) design roadmaps have been released. In addition to prioritizing stacking layers, triple-level cells (TLCs) are becoming a mainstream choice for enterprise-level SSDs, which have given rise to QLC SSDs.

Data reduction technologies for SSDs evolve rapidly, driving down the effective capacity cost.

In typical unstructured data scenarios like satellite remote sensing, data

reduction ratios can reach 2:1. For autonomous driving and PACS imaging scenarios, the ratios reach 1.5:1 and 3:1, respectively. Data reduction technologies have reduced the purchase cost of all-flash storage considerably. In backup scenarios, all-flash backup storage also provides a 50% higher data reduction ratio than benchmark HDD backup storage thanks to global deduplication, similarity-based inline deduplication, and semantic-level deduplication technologies.

Large-capacity SSDs help continuously decrease equipment room footprints and the energy consumption of data centers.

In the next two to three years, the capacity of a single SSD will be 1.5 to 2 times higher than that of an HDD, or potentially even more, with comparable power consumption. Therefore, large-capacity SSDs are critical for enterprise data centers to reduce the energy needed. Furthermore, SSDs are more reliable than HDDs. The 5-year return repair rate of SSDs is only 1.75%, four times lower than HDDs.

To sum up, as SSDs help enterprises reduce their CAPEX, footprint, and power consumption, they have become an alternative for replacing HDDs in high-performance production and transaction systems. In addition, SSDs can also be used to store warm and cold data such as backup data and mass unstructured data. As a result, the TCO for all-flash backup storage over a five-year period is 50% to 60% lower than that of benchmark HDD backup storage. In terms of scale-out storage designed for mass unstructured data, an SSD can be used to replace an HDD, to offer the same available capacity. Notably, the TCO of SSD-based scale-out storage remains the same as that of HDD-based models over a five-year timeframe.

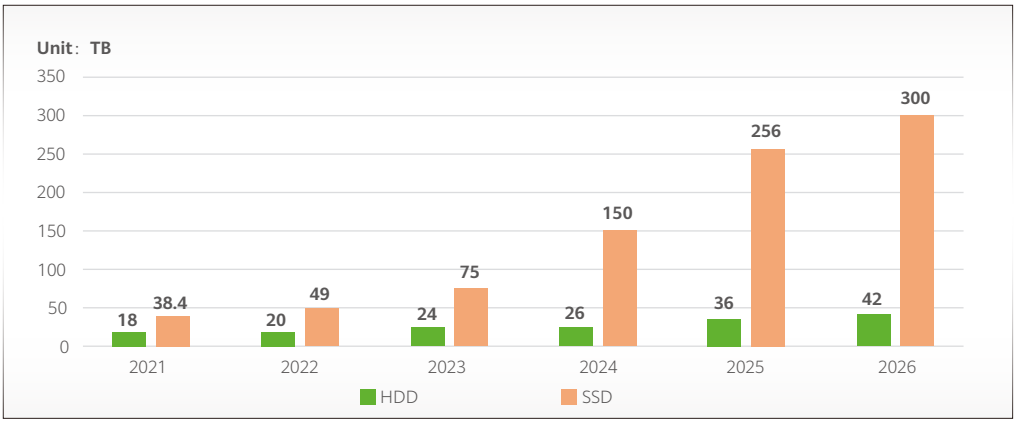


Figure 13: Maximum capacity of a single HDD or SSD

Suggestions

Tailor all-flash storage plans to current and future enterprise data volumes and requirements

Enterprises should work with storage providers to evaluate future data volumes and service pressure trends to formulate all-flash storage strategies and analyze benefits and O&M cost changes after the strategies are implemented.

Seize opportunities to replace legacy storage with all-flash models

HDDs are common in many enterprise storage environments, and most will soon reach the end of their warranty periods. Enterprises undergoing digital transformation are in urgent need of better, more performant storage devices. This presents an excellent opportunity for storage enterprises to promote all-flash storage.

New Technologies

Outlook 8

Data-Centric Architecture

The move from CPU-centric towards data-centric

The recent boom in AI and real-time big data analytics applications has displaced the convention of centering computing power on a CPU and popularized a heterogeneous computing method across CPUs, GPUs, NPUs, and DPUs. This shift has placed higher requirements on memory capacity and bandwidth, which traditional CPU-centric architectures cannot meet.

Trends

The CPU-centric server architecture is evolving into a data-centric composable architecture

CPU-centered computing alone is now insufficient for increasingly diverse applications that demand quick real-time data processing, therefore a heterogeneous computing approach represented by GPU is emerging.

New computing hardware has seen a boost in hot data processing efficiency of I/O-intensive applications, but causes extra memory access pressure. Current levels of capacity and bandwidth of local memory cannot meet the requirements for data processing.

In 2019, Intel launched the open interconnection standard, Compute Express Link (CXL). The new memory-semantic bus running on the CXL enables quick access to external memory and memory capacity expansion, making it possible to decouple memory from CPUs and allowing external high-speed storage devices and heterogeneous computing

resources to form a memory pool.

The decoupling extends to the CPU-centric server architecture which is evolving into a data-centric composable architecture. Computing, memory, and storage resources from different computing units can be combined on demand, and the heterogeneous computing can directly access memory and storage resources through the high-speed bus.

The new server-based architecture has revolutionized the storage positioning, which will no longer be limited to managing disks, but will also incorporate in-memory storage in the future.

At the Flash Memory Summit 2023, MemVerge, Samsung, XConn, and H3 released 2 TB CXL memory pooling systems suited for AI, and also the South Korean company Panmnesia

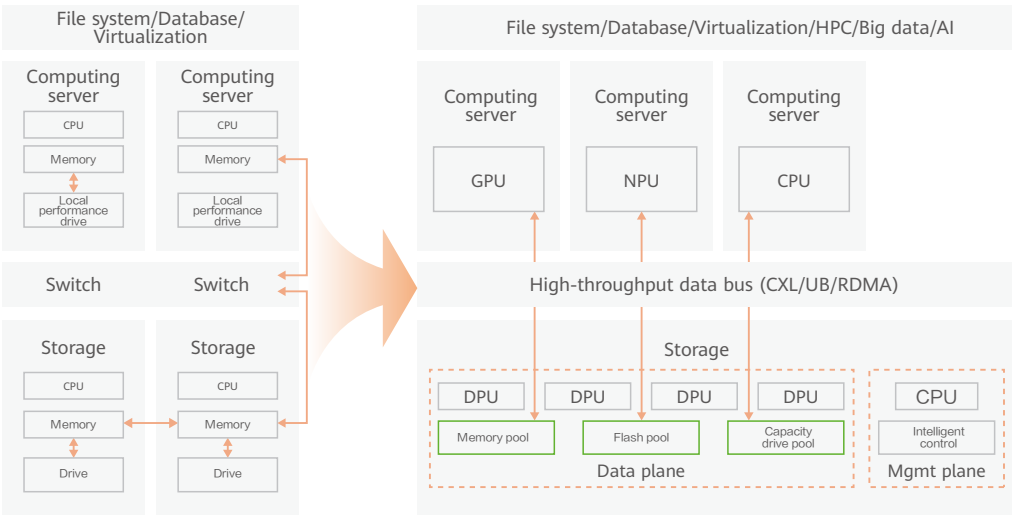


Figure 14: CPU-centric server architecture evolving into data-centric composable architecture

demonstrated its 6 TB CXL memory pooling system.

The CPU-centric storage architecture is evolving into a data-centric composable architecture

In the future, applications such as AI and big data will require higher performance and lower latency and CPU performance growth may slow down. With the development of the composable server architecture, the storage architecture will also evolve into a data-centric composable architecture to greatly improve the performance of storage systems. Various processors (CPUs and DPUs), memory pools, flash pools, and capacity disk pools of the storage system will be interconnected through new data buses. In this way, data can be directly stored in memory or flash memory after accessing the storage system, avoiding slow data access due to poor CPU performance.

Suggestions

Keep pace with the evolution of server and storage architectures, make timely adjustments, and seek opportunities from new storage

Data-centric architectures will be needed to cope with the sharp increase in data processing requirements. Enterprises should keep pace with the evolution of hardware in data centers, make timely adjustments, and build the optimal server and storage architectures to lay a solid data foundation for service development.

New Technologies

Outlook 9

AI-Powered Storage

Management intelligence→product intelligence: AI powers self-driving full-lifecycle data management

Artificial intelligence for IT operations (AIOps) is a popular method for enterprises to improve storage O&M automation. AI technologies are no longer limited to monitoring and O&M of storage devices, but also supercharging storage products from the bottom up with intelligence.

Trends

Storage vendors are adopting diverse disruptive innovations to optimize storage SLA management

The combination of conventional AI + foundation models helps optimize storage SLA management from diverse dimensions.

Optimized service rollout: Storage resource provisioning and changes are shortened from days to just minutes. Traditionally, service changes require manual solution planning, change script development, and execution. Conventional AI technology makes automatic service simulation capable of formulating optimal change solutions. When adding AIGC technology, change scripts can also be automatically generated to shorten change periods to mere minutes.

Optimized infrastructure availability: Annual average failure period of the data center is shortened from hours to mere minutes. Traditional AI can help predict performance, capacity,

and spare parts faults, reducing the probability of exceptions. In complex exception handling scenarios, storage management systems can also us AI foundation models to quickly strengthen interaction logic and facilitate manual fault location, thus greatly shortening troubleshooting periods.

Optimized cost management: Storage resource utilization is growing from 50% to 60%. Improper resource allocation has consistently been the primary cause for low resource utilization in data centers. AI-based intelligent identification and release of idle resources can protect storage investment. In addition, thermal analysis of global data optimizes data distribution in different media in the data center and migrates cold data in a timely manner, reducing storage costs. One large carrier in Asia Pacific, as an example, used Huawei's intelligent storage centralized management software to improve storage resource utilization from 30% to more than 60%.

The AI capability needs to be fully unlocked to build an AI management architecture

The AI capability needs to be fully unlocked to enablement AI management architecture

With the increasing complexity of enterprise IT technology stacks and the continued emergence of new applications such as big data, containers, and multi-cloud, storage must meet increasing usage and management requirements as it serves as the foundation of IT infrastructure. More and more enterprises use the AI management tools provided by storage vendors to build three-layer management architectures that feature intelligent device management, data centers, and clouds. This simplifies infrastructure management and optimizes management efficiency, all while creating a new AI process of incubation, release, and optimization to better cope with AI transformation.

Device management intelligence: Device management software collects basic information for cloud AI model incubation and obtains updated AI models in the cloud via online updates or offline imports. Software is then responsible for using and managing

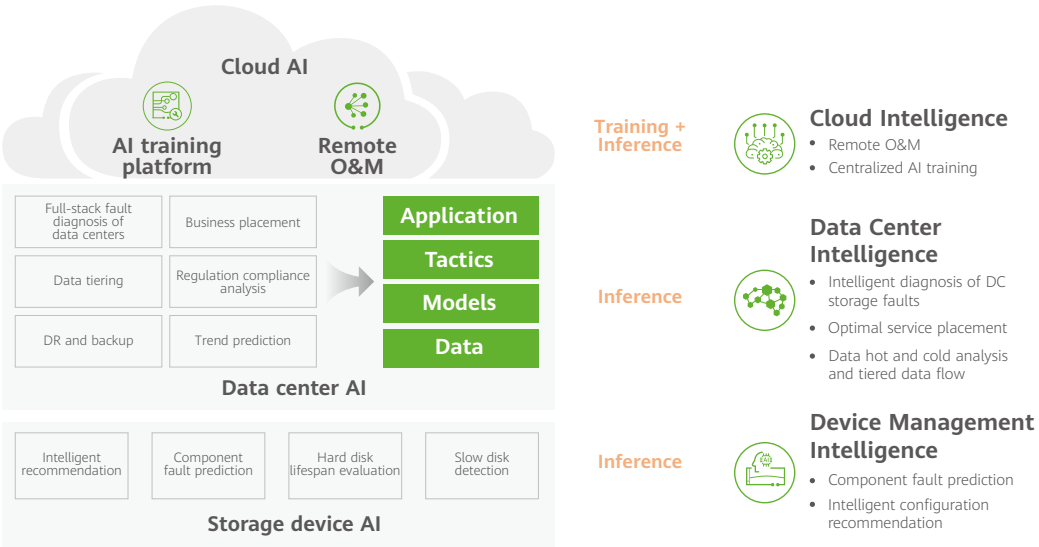


Figure 15: 3-layer AI architecture

individual storage devices, delivering features like optimal configuration recommendations, fault detection of optical modules, disks, and controllers, and slow disk detection.

Data center intelligence: Data center management software covers a wider scope than the device software. Unified management of multi-vendor storage devices simplifies O&M processes, and intelligent cross-device data scheduling and tiering optimize storage costs. By managing full-stack data center devices, management software can then intelligently analyze application, virtualization, network, and storage resources to diagnose problems in minutes. Unlike cloud intelligence, data center management software is deployed in data centers and therefore isolated from the extranet for more stringent data requirements.

Cloud intelligence: There are powerful computing and storage resources in the cloud, which can continuously perform AI model inference and training on the running data uploaded by a large number of devices, and distribute optimized AI models to data center management

software and devices on demand. Cloud management software also provides diversified O&M methods like mobile applications. However, compared to data center management software, cloud management does not have sufficient full-stack analysis capabilities for data center infrastructure and cannot support device change operations.

Storage vendors prioritize product intelligence to optimize storage device efficiency and reliability

To meet the diverse storage needs of emerging applications, storage vendors are focusing on product intelligence to boost performance and reliability. For example, Dell EMC storage products have built-in smart tuning and data reduction algorithms to achieve self-configuration optimization and optimal data reduction ratio; NetApp products deliver intelligent storage resource scheduling and fast data access; and Huawei storage systems intelligently provision hardware resources to improve read and write efficiency. In addition, Huawei storage systems intelligently adjust data reduction algorithms for a variety of data types, improving the compression ratio and reducing the average storage cost per-unit data.

Another point is the innovation of intelligent storage architectures. In traditional storage, algorithms and data are coupled. Multiple fixed algorithms are dispersed at the cache, scheduling layers, and storage pools, but these algorithms must be manually set to ensure the access efficiency of different types of data, resulting in poor flexibility. This brings us to the decoupled design of intelligent algorithms from data. Self-learning and adaptive algorithm libraries can independently determine the correct layout, scheduling, and reduction of different data types, ensuring optimal data access efficiency.

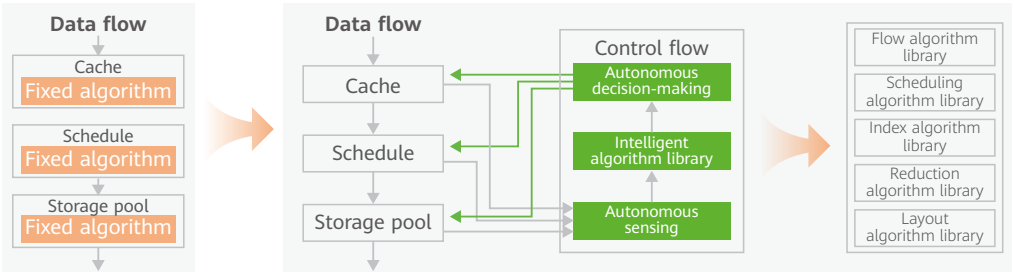


Figure 16: Algorithm-data decoupling with intelligent storage

Suggestions

Clearly define service model indicators and SLA requirements, and develop new evaluation standards systems once new platforms and technologies are introduced

Before enterprises plan to introduce AI-related products, they should first evaluate their current and future business needs and establish a multi-dimensional and quantifiable evaluation system that covers storage capacity, performance, reliability, energy efficiency, resilience, and ecological robustness requirements. If enterprises engage multiple storage suppliers, they should also establish supplier capability baselines and dynamically update those baselines as their product capabilities improve or deteriorate to ensure that intelligent storage devices and management platforms best suited to actual business needs are procured.

Leverage the AI capabilities provided by storage vendors and work together on continuous AI capability improvement

The application of AI in storage systems greatly improves storage SLAs.

Enterprises should pursue joint innovation with storage vendors to strengthen AI capabilities, thus incubating AI capabilities that are more closely related to their actual service characteristics.

Update the capability model of enterprise IT teams and provide comprehensive pre-training for employees

With the introduction of intelligent storage devices, employees will need to learn and adapt to an increasing number of new technologies. Enterprises need to proactively establish training plans and technical support mechanisms to help their employees better understand and utilize the functions of intelligent storage devices and their management tools. This will ensure intelligent storage devices deliver their promised functions, achieve better human-machine collaboration, and create more value to enterprises.

Energy Saving

Outlook 10

Energy-Saving Storage

Green data storage is essential for data centers to reach net-zero carbon emissions

Organizations around the world are working to meet carbon peak and neutrality goals, and this starts with data centers. More than 30% of a data center's energy consumption goes to storage. Therefore, to build sustainable data centers, we need to focus on reducing the energy consumption of IT equipment, particularly, storage devices, in addition to lowering power usage effectiveness (PUE).

Trends

Building a green data center requires energy-efficient data storage in addition to a lower PUE

Reduced PUE is just one step along the way to green data centers. Lowering the power consumption of IT equipment is arguably even more important. Storage devices are expected to be the main electricity-draining IT components. For example, the annual power consumption of 1 petabyte of data storage in a data center is 300,000 kWh, which is equivalent to emitting 235.5 tons of carbon. Without an effective green strategy, the carbon emissions attributed solely to storage in 2030 will easily exceed the total global carbon emissions recorded in 2019.

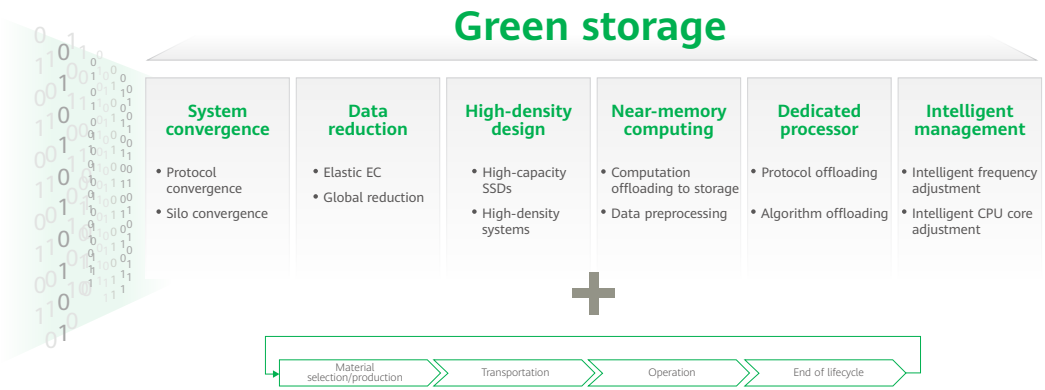


Figure 17: Lifecycle carbon footprint

Innovations in energy-saving storage technologies are a catalyst for the low-carbon development of data centers

In response to the mounting pressure to reduce storage energy consumption, storage vendors are proactively innovating and developing technologies to help data centers go green and contribute to sustainable development.

Energy-saving data storage technology

One system supports multiple workloads/application needs

Multi-protocol convergence and silo convergence enable one-for-all storage and improve resource utilization. One storage system can support multiple protocols like file, object, and HDFS to meet diversified requirements and integrate multiple types of storage. In addition, converged resource pools implement resource pooling to improve resource utilization.

Data reduction increases storage capacity

Data reduction technology squeezes as much data as possible into physical storage without distorting it, in order to maximize capacity. This lowers operation costs for enterprises and reduces the amount of energy data storage devices consume.

High-density design improves storage capacity density

Storage media accounts for 83% of a device's total power consumption. Compared with HDDs, SSDs consume 70% less power and save 50% space with the same capacity. A storage product equipped with large-capacity SSDs and high-density disk enclosures can store the same amount of data with less energy and space, meaning lower power consumption per unit capacity.

Near-memory computing reduces the energy consumed by data movement

Data movement increases energy consumption. Research shows that data movement consumes almost twice as much energy as computing on large-scale AI computing clusters. However, the energy it consumes on local compute nodes is only 5% of that of computing. Processing data close to where it resides reduces data movement. For example, processing tasks like vector retrieval on storage helps halve overheads in protocol and data conversion between storage and computing, and more than halve the energy required for data processing.

Accelerated dedicated processors reduce CPU overheads

Compared with storage devices with a single CPU computing architecture, devices equipped with dedicated processors perform better and use less power. Dedicated processors that take over data reduction and protocol processing tasks from general-purpose CPUs boost efficiency and reduce latency for data processing, and eliminate dependencies on CPU computing power, thereby making the storage solution more energy efficient.

Load-based intelligent frequency modulation and core adjustment slashes power consumption

Predictions and interventions based on AI models and software solutions make storage operations in data centers more energy efficient. Based on big data analytics of service I/O models, technologies like intelligent load prediction, dynamic resource

scheduling management, and timely shutdown and adjustment of resource usage frequency enable storage devices to have the lowest energy consumption during operation while also meeting SLA requirements for service workloads without affecting upper-layer applications.

Energy saving through storage lifecycles

Energy savings need to be made throughout the entire lifecycle of storage devices, including raw material selection, manufacturing, transportation, product use, and disposal. In the storage manufacturing phase, manufacturing plants widely use photovoltaic power, zero wave soldering, paperless labeling, and renewable materials such as aluminum and tin. Packaging made from FSC-certified paper and printed using soybean ink weighs less and so further eases the transportation burden. In the storage products' use phase, intelligent O&M based on AIOps enables on-demand use of storage resources. In addition, a proper recycling system is established to ensure that at the end of the product lifecycle, e-waste is handled in an environmentally friendly manner that optimizes recycling and minimizes environmental impact.

The current spotlight on carbon footprint is motivating the optimization of storage product design and the development of energy-saving technologies. With its industry-leading innovative green solutions, Huawei OceanStor All-Flash Storage has become the first to be awarded DEKRA Storage Product Carbon Footprint and DEKRA Seal certificates. DEKRA is one of the world's leading expert organizations in the testing, inspection, and certification sector.

Trends

Shift focus from the current power consumption of a single device to carbon emissions throughout the device's entire lifecycle

An increasing number of global enterprise customers require equipment suppliers to provide carbon footprint reporting or carbon emissions reporting for the

entire lifecycles of their products. As carbon disclosure continues to draw widespread attention, carbon footprint reporting is becoming a mainstream requirement as we work towards a green future in international trade. Currently, carbon footprint evaluation standards and its accounting methods vary across the industry. The best course of action would be for enterprises to use certification from reputable third-party testing organizations.

Comply with unified energy efficiency evaluation standards

Standard organizations, enterprise customers, and storage vendors use different methods and indicators to assess how energy efficient storage solutions are because the field is still in its infancy. This means enterprises and vendors repeat tests based on various standards and evaluation methods. The establishment of a unified set of energy efficiency standards can reduce both resource waste and pressure on vendors, who have to get their products repeatedly tested. It could also help enterprise customers make more informed choices when it comes to green storage solutions, and facilitate the healthy development of energy-saving and low-carbon technologies in the storage industry.

Promote storage vendors to innovate for lower power consumption

Enterprises are encouraged to proactively deploy storage products powered by energy-saving technologies and push storage vendors to innovate, for example, in dedicated processors, data reduction, high-density all-flash design, and intelligent frequency modulation for higher capacity density and energy efficiency. Next-generation storage products can be equipped with near-memory computing to reduce the amount of energy used for data movement.

New Data Paradigm Unleashes the Power of AI

Data is a new factor of production in the digital economy, and has become a standard, strategic resource. As data scale surges to yottabytes and data applications boom, powerful, resilient, and reliable data infrastructures are needed to safeguard enterprises. Professional storage devices are responsible for protecting valuable data assets in various industries. Zero data loss, continuous access, and no latency are the major concerns of industries that are focused on professional storage. Huawei OceanStor storage believes that an era of uncertainty, complexity, and diversity entails inclusive capabilities that are changeable on demand to identify and integrate imperceptible data requirements.

For AI storage, Huawei launched OceanStor A310, which is oriented to the data lake scenarios of foundation models and industry-specific models. It can implement mass data management throughout the entire AI development lifecycle from data collection and preprocessing to model training, inference, and application. Another storage model, FusionCube A3000, is oriented to the training/inference scenarios of industry-specific models and model applications with 10 billion-level parameters. It integrates OceanStor A300's high-performance storage nodes, training/inference nodes, switching devices, AI platform software, and O&M software, so as to provide AI model vendors with a ready-to-use deployment experience and one-stop delivery.

Huawei's OceanStor Dorado all-flash storage enhances NAS capabilities and provides industry-leading SAN features. Its comprehensive functions for cyber resilience, such as Quota, QoS, and ransomware protection, facilitate resilient cross-agency file sharing. OceanStor Dorado boasts the industry's only integrated SAN and NAS active-active solution, ensuring service continuity. Additionally, customers can leverage OceanStor Dorado's high performance, reliability, and manageability to support the new application ecosystem such as distributed databases and containers and implement a smooth cloud-native transformation.

OceanStor Pacific scale-out storage configures high performing all-flash disks that have a large capacity and applies data reduction algorithms. It is the only device that supports hybrid workloads and lossless multi-protocol interworking in the industry, enabling one storage system to support diverse applications, including AI, HPDA, videos, and archiving. Huawei firmly advances all-flash scale-out storage and supports mass unstructured data entering the production decision-making systems. This has significantly increased enterprise efficiency and service experiences.

For data protection, OceanProtect Backup Storage provides up to 72:1 data reduction ratio, as well as 3x backup bandwidth and 5x recovery bandwidth higher than the industry average. Furthermore, the storage can integrate backup software and support automatic backup and recovery management for databases, virtualization, file systems, big data, and data warehouses. Comprehensive ransomware protection solutions help users achieve efficient backup and recovery, building the last line of defense for data protection.

FusionCube HCI integrates compute, storage, and network resources, as well as implementing on-demand loading of management, backup, disaster recovery, and more functions through software-defined architecture. Integrated delivery can also be realized thanks to pre-integration. Intelligent management software and algorithms that match various industries cover all scenarios from data centers to enterprise branches, providing FusionCube with flexible architecture, powerful performance, solid resilience, high reliability, simplicity, and high efficiency.

Data Management Engine (DME) offers a unified management GUI alongside AI-powered multi-dimensional risk prediction and optimization features to provide automatic management and intelligent O&M throughout the storage lifecycle, which includes planning, construction, O&M, and optimization. This streamlines storage management and enhances data center operation efficiency. Additionally, DME enables the fast retrieval, analysis, and tiering of mass unstructured

data, which leads to more efficient data access and lower data storage expenses.

The DCS Lightweight Data Center Solution is primarily focused on providing virtualization, container, and DR and backup suites. It is compatible with mainstream hardware infrastructure through its southbound APIs, and can also connect to third-party cloud management systems via the northbound side to offer cloud services. In addition, the PaaS service incorporates big data solutions and AI enablement platforms. It harnesses the power of the eDME unified management platform to coordinate the construction of IT infrastructure and applications and unify applications, hardware, and platforms respectively.

Huawei Storage has established a presence in more than 150 countries and regions, and has already garnered the trust of more than 25,000 enterprises. Notably, 47 of the world's top 100 banks have chosen to use Huawei storage solutions. Huawei has been named a Leader in the Gartner® Magic Quadrant™ for Primary Storage for seven consecutive years, with scores climbing higher and higher each year. Huawei has also been recognized as a Customers' Choice by the Gartner® Peer Insights ™ "Voice of the Customer" for its exceptional scale-out and primary storage.

In the future, storage will be used with diversified data application acceleration engines, which can improve storage performance by a factor of 10. In terms of hardware, data flows across two buses: the double data rate (DDR) memory bus and the PCIe system bus. Huawei will use a global data bus with high throughput to integrate these two buses, greatly improving the efficiency of data flows from 5 GB/s to more than 50 GB/s. On the software front, employing a control and data plane separation architecture instead of an integrated architecture can significantly reduce data access latency from 100 microseconds to 10 microseconds. To further accelerate applications and enhance user experience, data application acceleration engines will be used to enable near-data

processing based on optimized data storage and access functions. This innovative approach can make processing 5-10 times more efficient.

Huawei Storage is fully committed to implementing a data-centric philosophy to build a more resilient, reliable, and efficient data foundation for customers and support the ever-changing, diversified data applications in the digital era.

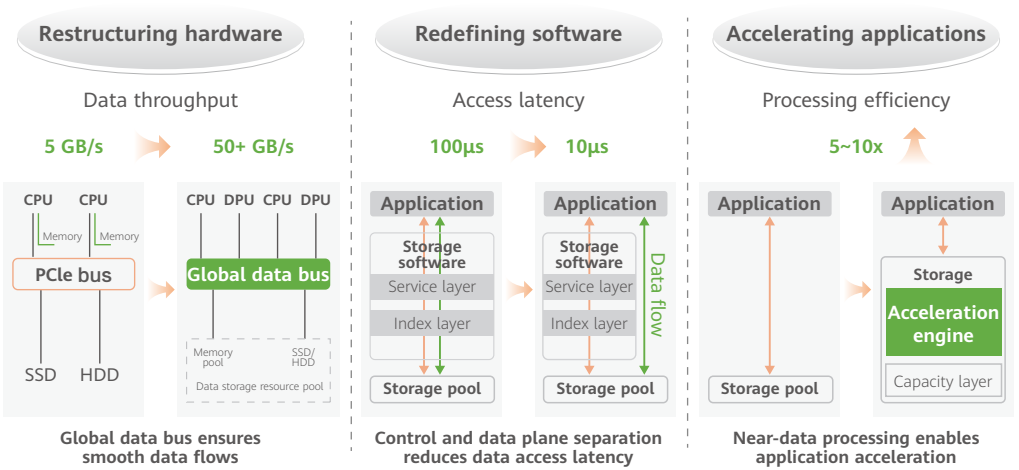


Figure 18: Reshaping storage architectures us ing a data-centric approach

Appendix

References

❶ **Striding Towards The Intelligent World White Paper 2022**
https://www.huawei.com/en/giv/industries/data-storage

❷ **Vikas Arora: How are Artificial intelligence and Big Data connected?**
https://dzone.com/articles/how-ai-and-big-data-connected

❸ **Gartner Survey Shows Generative AI Has Become an Emerging Risk for Enterprises**
https://www.gartner.com/en/newsroom/press-releases/2023-08-08-gartner-survey-shows-generative-ai-has-become-an-emerging-risk-for-enterprises

❹ **China Mobile: Big Data Lakehouse Technical White Paper 2022**
https://www.gartner.com/en/newsroom/press-releases/2023-08-08-gartner survey-shows-generative-ai-has-become-an-emerging-risk-for-enterprises

❺ **Forecast: Hard-Disk Drives, Worldwide, 2021-2027, 2Q23 Update**
https://www.gartner.com/en/documents/4568499

❻ **Splunk's Report: The State of Security 2023 Is Resilient**
https://www.splunk.com/en_us/campaigns/state-of-security.html

❼ **Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers**
https://www.sciencedirect.com/science/article/pii/S2096511720300761

❽ **Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025**
https://www.statista.com/statistics/871513/worldwide-data-created/

❾ **How To Identify And Break Down Tech Silos In IT**
https://www.advsyscon.com/blog/break-down-silos-in-it/

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P.R. China
Tel: +86-755-28780808
www.huawei.com

Editor-in-Chief: Zhang Guobin, Qiu Fangjia

Managing Editors: Pang Xin, Ding Zhibin, Zhang Yi, Wang Xuesong, Wang Zhen, Wang Qiang, Yang Chuanbin

Section Editors: He Yujin, Ding Jiangbo, Qiu Donghua, Qin Wei, Dai Fuxi, Li Yunli, Wu Lin, Zhao Le, Yang Dan, Yang Zesheng, Zheng Yong, Fu Chunhao, Cao Yu, Cai Jinrong, Hong Yue, Zhu Leshan, Zheng Chengming, Song Jiangxian, Jiang Huahu, and Zhu Lingyun, Zhang Yue, Guo Nan, Wu Zhuang, Gao Tan, Wang Guanghong, Wang Wei, Li Wenxiu, Shen Qi, Chen Youguang, Suo Haidong, He Changjun, Chen Aiping, Zhao Bonan, Cui Wenlin, Shu Changlin, Lin Peng, Tan Hua, Chen Hua, Lan Wenhai

Translators: Wang Jing,Yu Shanshan, Lu Shasha, Zeng Maolin, Mei Wuzhi, Yin Yingying, Guan Hanwen, Jiao Zhiling, Xiao Yue, Liu Xuan, Wu Mengni, Hu Wenxiang, Li Jingwen, Liu Jiawen, Chen Xin, Daniel Curran, CHRISTOVA EKATERINA ROUMENOVA, YOUNG MEGAN MARY, Jimmy Ding, Kyle Melieste, PETER PEIYU ZHAO, Gavin Wills, Nathanael Schneider