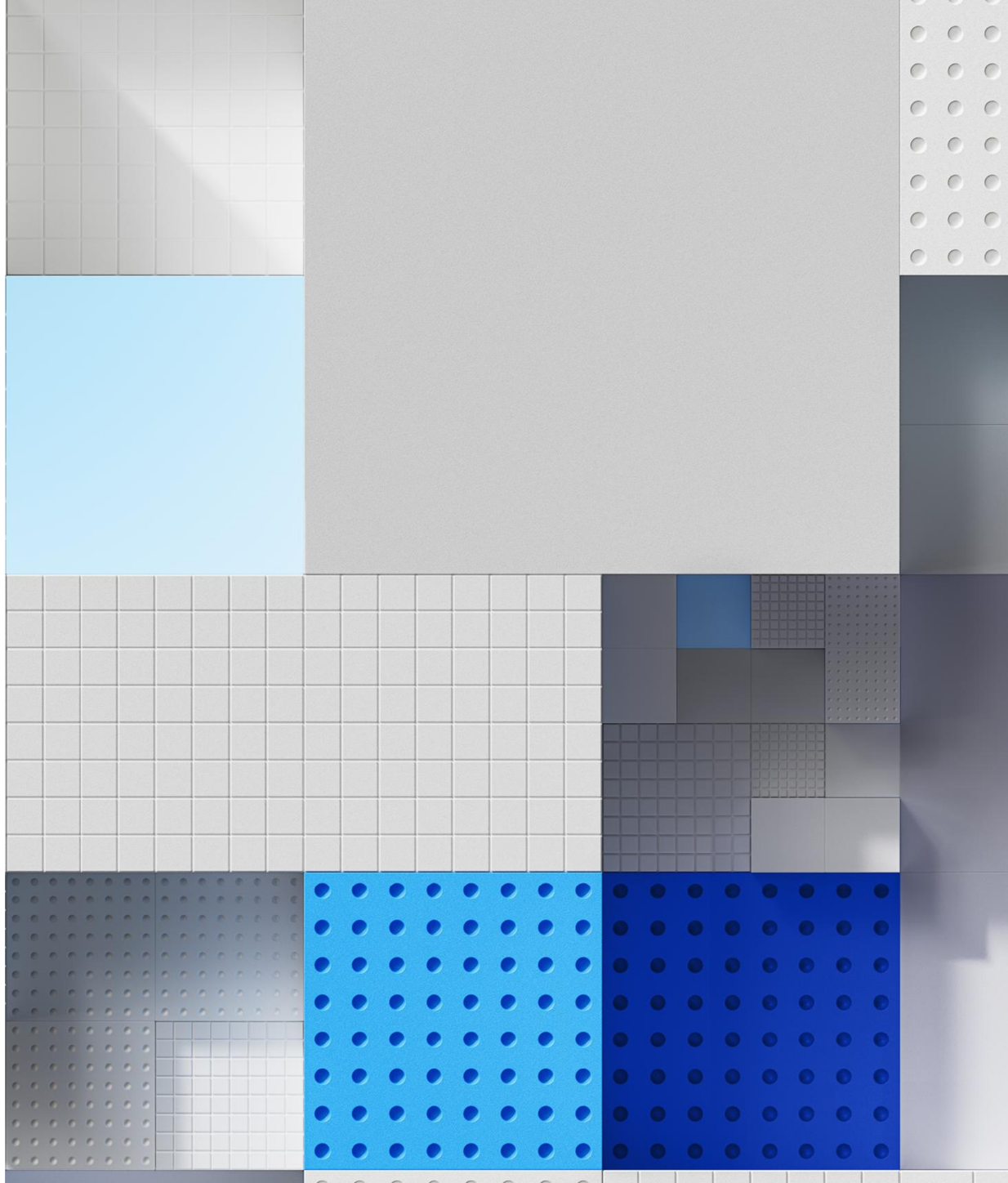
IBM Storage 2025

ISC | 10 Jun 2025



Top trends in enterprise data storage for 2025

Storage optimized for GenAI applications	Nearline SSD flash storage	Cyber storage	Integrated data intelligence	Hybrid cloud storage
• Prepare data for AI use cases • High speed key-value repositories • All flash storage	• QLC flash replacing HDD capacity • Offload data functions to FPGA-based flash • Reduce rack space	• Threat detection • Inline vs. post process • For both structured and unstructured data	• Interpret data to perform action • Functions to enrich object metadata • Eliminate separate data lake	• Connect storage across locations • Burst for capacity or performance • Storage standardization

Storage platform innovation

Gartner

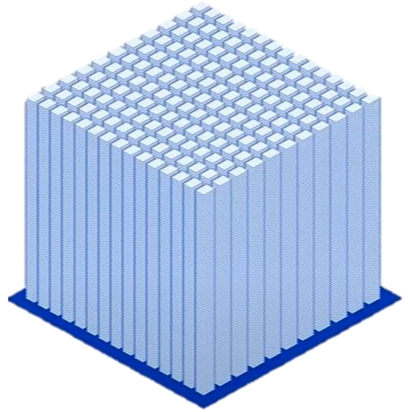
Source: Gartner, Top Trends in Enterprise Data Storage for 2025, Figure 1, 7 April 2025

Value of enterprise data for AI

To close the last mile of model efficacy, enterprises are looking to leverage their enterprise data

In 2024, 47% of enterprises said they are looking to do heavy model customization with enterprise data[1]

Current nearly all available *public data* is now represented in foundation models[2]



Less than 1% of all *enterprise data* is represented in foundation models[2]

52% of enterprise data is still in data centers[3]

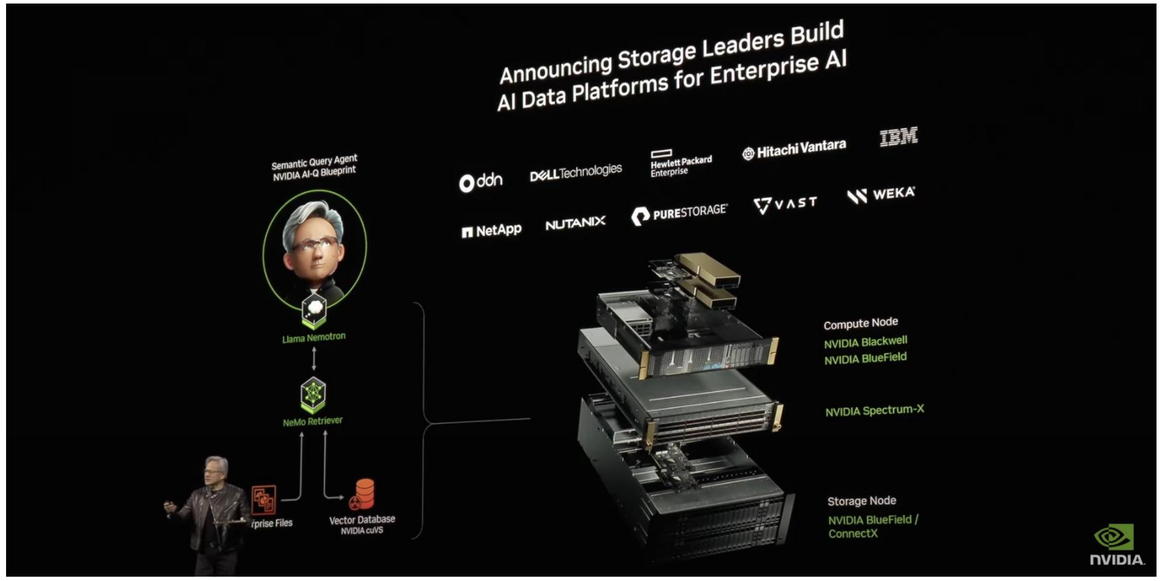
Sources:1) McKinsey & Company. (2023). The state of AI 2)IBM Research 3) Flexera. (2024). 2024 State of the Cloud Report

“2025 is the year of inferencing” – NVIDIA CEO Jensen Huang

- AI infrastructure focus is shifting from *training* AI models to *running* them – inferencing.

- Inferencing is driving major infrastructure investments. Inferencing market size:
  2024 – $20B
  2027 – $55B (moderate case)

- Storage is a critical part of inferencing solution.



Major inferencing opportunities for storage

1. Slow and costly data ingestion and data processing

- Data is copied multiple times – from source to lakehouse to data processor to vector database
- Too many copies of the data, too much data transfer, loss of security access control
- *All the data* gets reprocessed every time – no awareness of data changes
- Only a small fraction (<1%) of the enterprise data is used in gen AI

2. Poor inferencing performance due to limited memory space for models and tokens

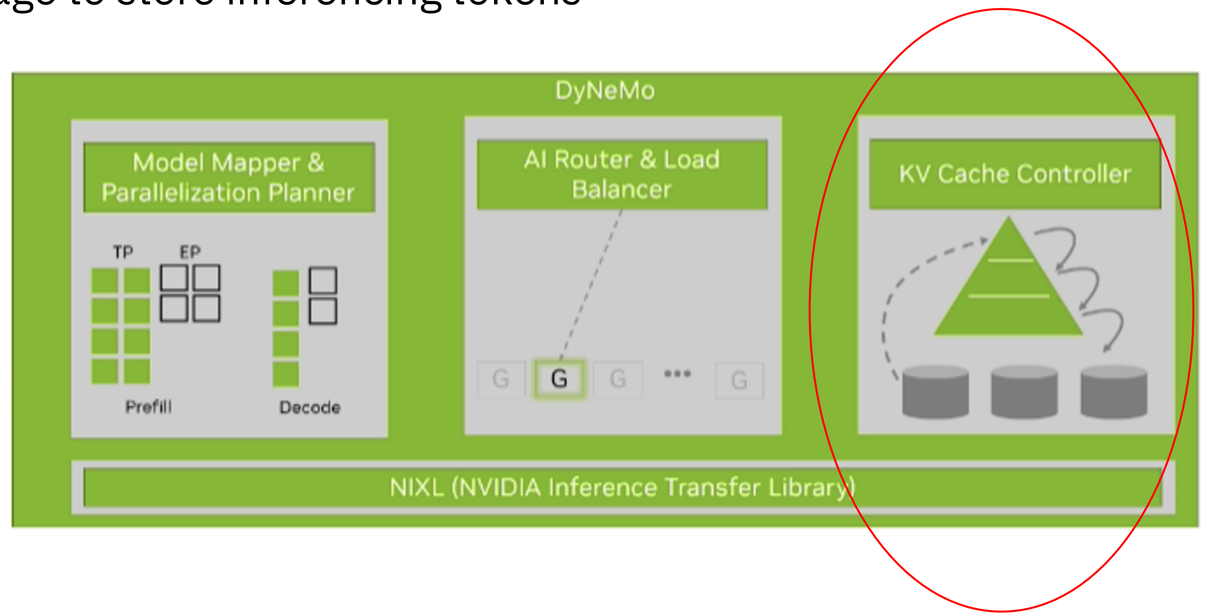
➢ Need high-performance/multi-tier/hybrid cloud storage to store inferencing tokens

- Simple calculation using HuggingFace config files (config.json)
- For ibm-granite/granite-3.1-8b-instruct
  - Each token's key-value embedding:

    Key = 81,920 bytes

    Value = 81,920 bytes

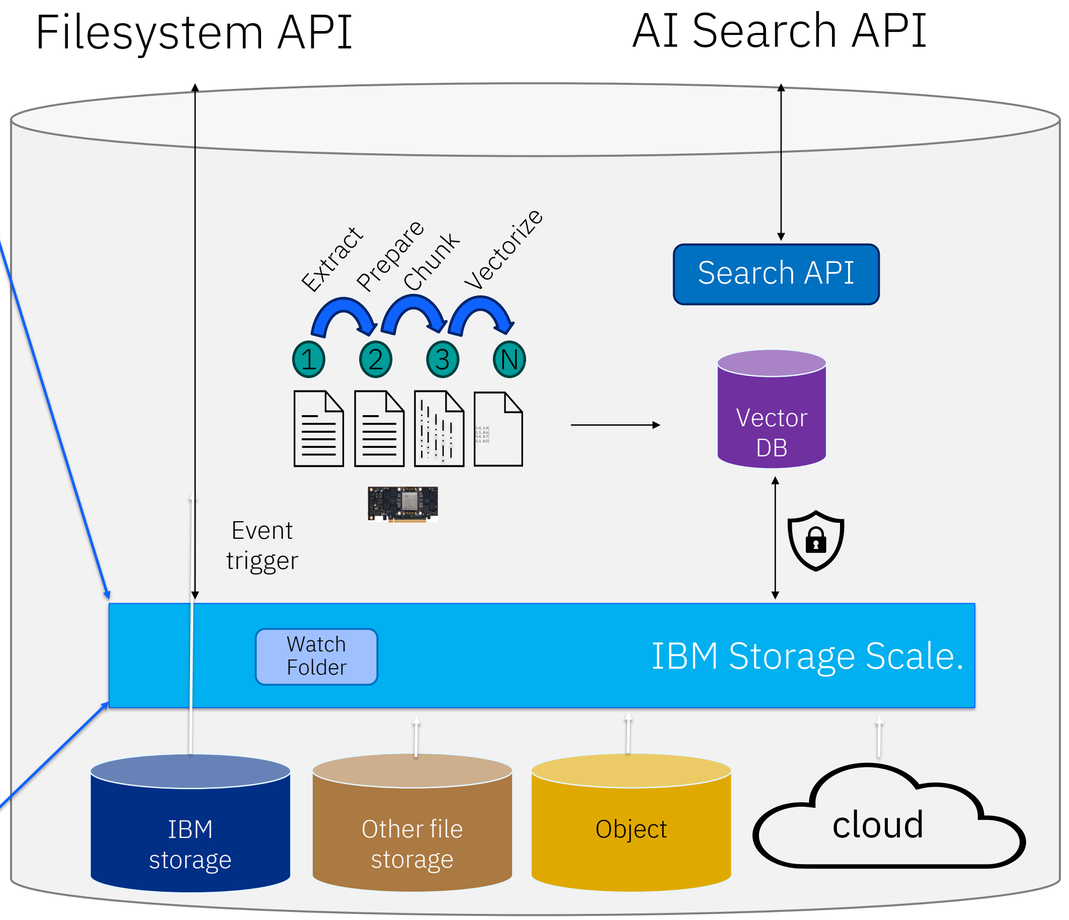
    Total = 163,840 bytes ~160 KiB/token



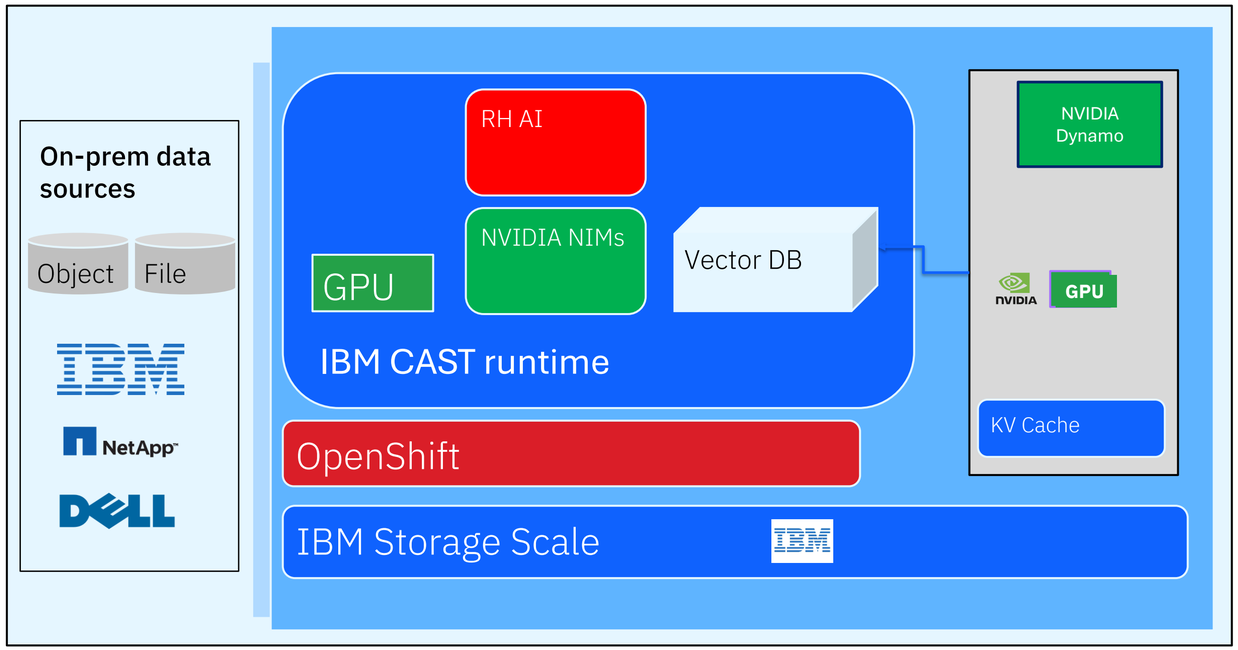
IBM AI optimized storage provides CAS differentiation

IBM AI optimized storage and AI runtime

- **Enterprise grade secure:** Consistent ACL and encryptions
- **Efficient:** Detect data change for incremental data processing
- **Support your legacy storage:** Connect to heterogeneous storage systems, including legacy unstructured data storage
- **Accelerate and scale:** GPU-optimized storage solution
- Ecosystem is key:
  - Nvidia NIMs
  - Red Hat LLMd



The end-to-end inferencing solutions with IBM Content-Aware Storage (CAS): from intelligent data ingestion to high performance vLLM inferencing with KV cache offload



Full stacks Intelligent data ingestion

- *Flexible architecture supporting NVIDIA and RedHat*
- *Support legacy storage*
- *Only incremental update*
- *E2E data security*

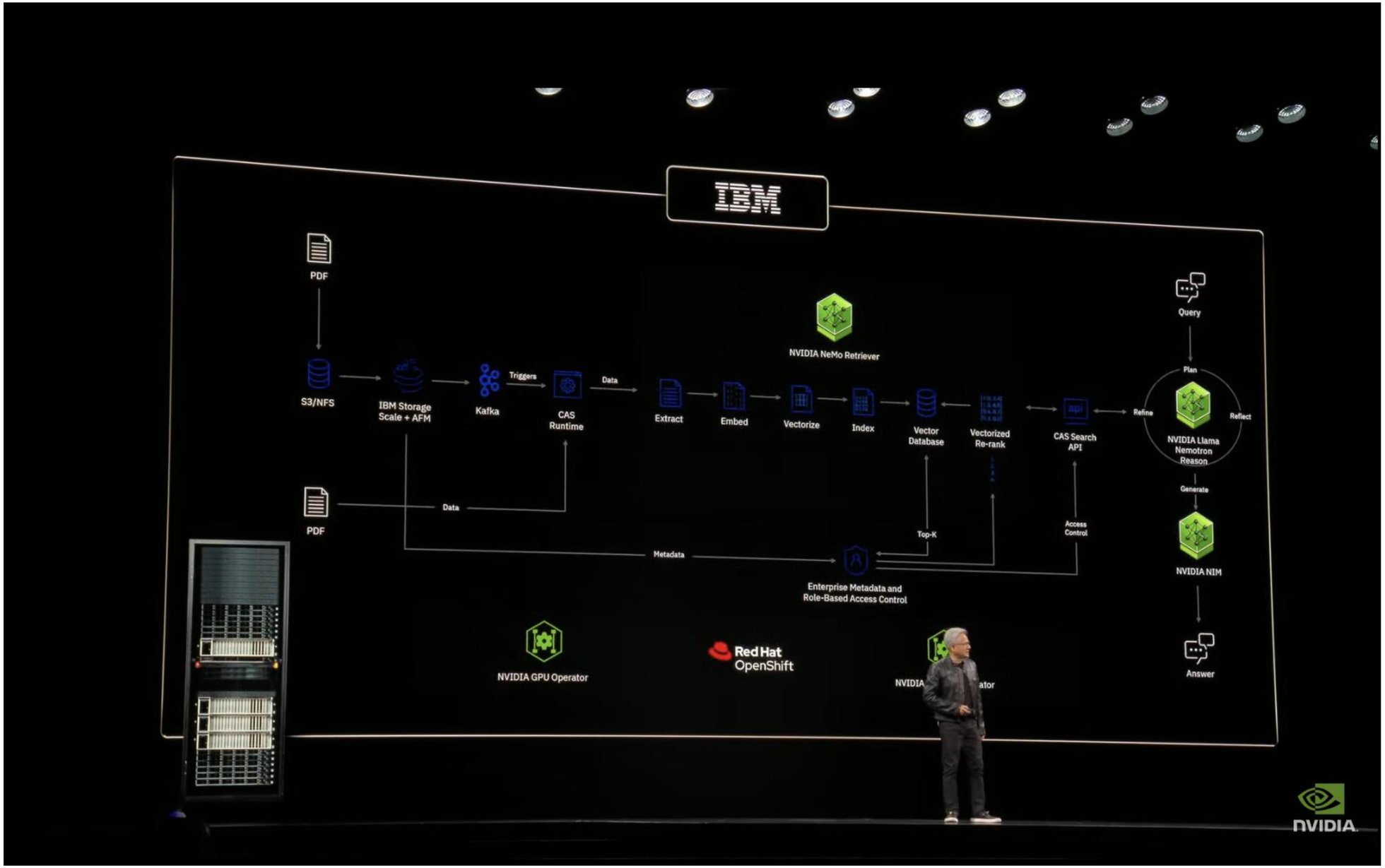
High performance model caching

- *Cache LLM models to high performance local storage*

Distributed KV cache storage

- *Efficient KV cache storage management*
- *High performance local storage*
- *Direct RDMA to GPU memory*
- *Intelligent tier storage and recalls*

IBM CAS at GTC Taipei on 05/19/2025



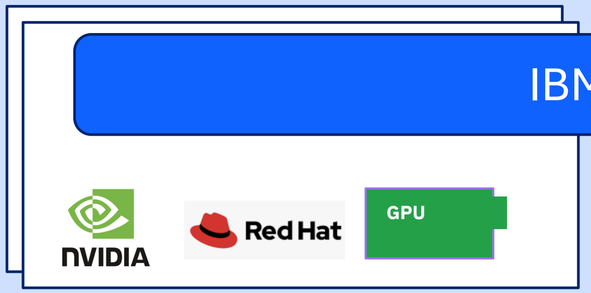
IBM Content Aware Storage

Hybrid Cloud Enterprise inferencing storage services

(EDGE, Fusion HCI, GPU servers, clouds)

Inferecning SDS :

- Running CAS SDS on GPU servers (including DGX, HGX)

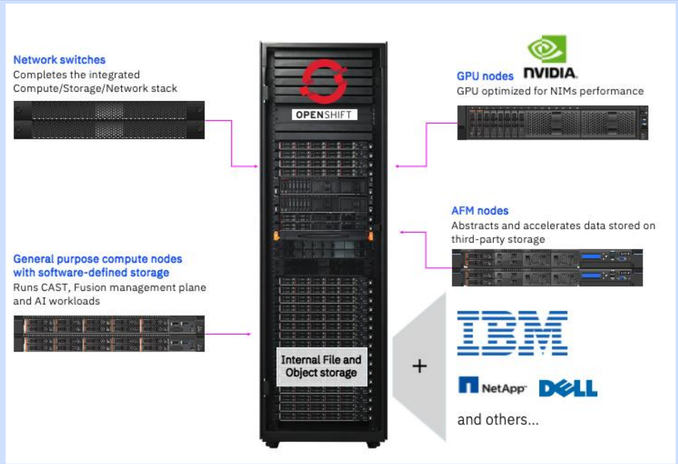


IBM Scale

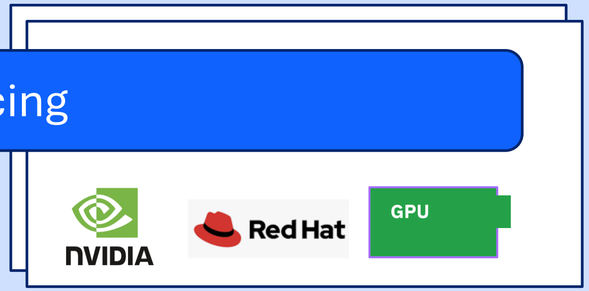
IBM Ceph

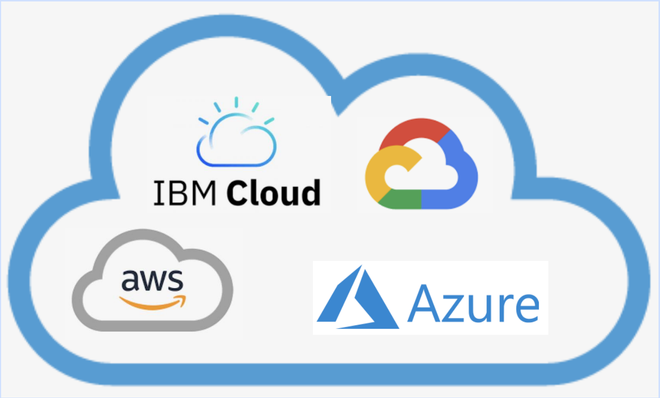
Inferening appliance

Turn-Key inferencing server
IBM Fusion HCI



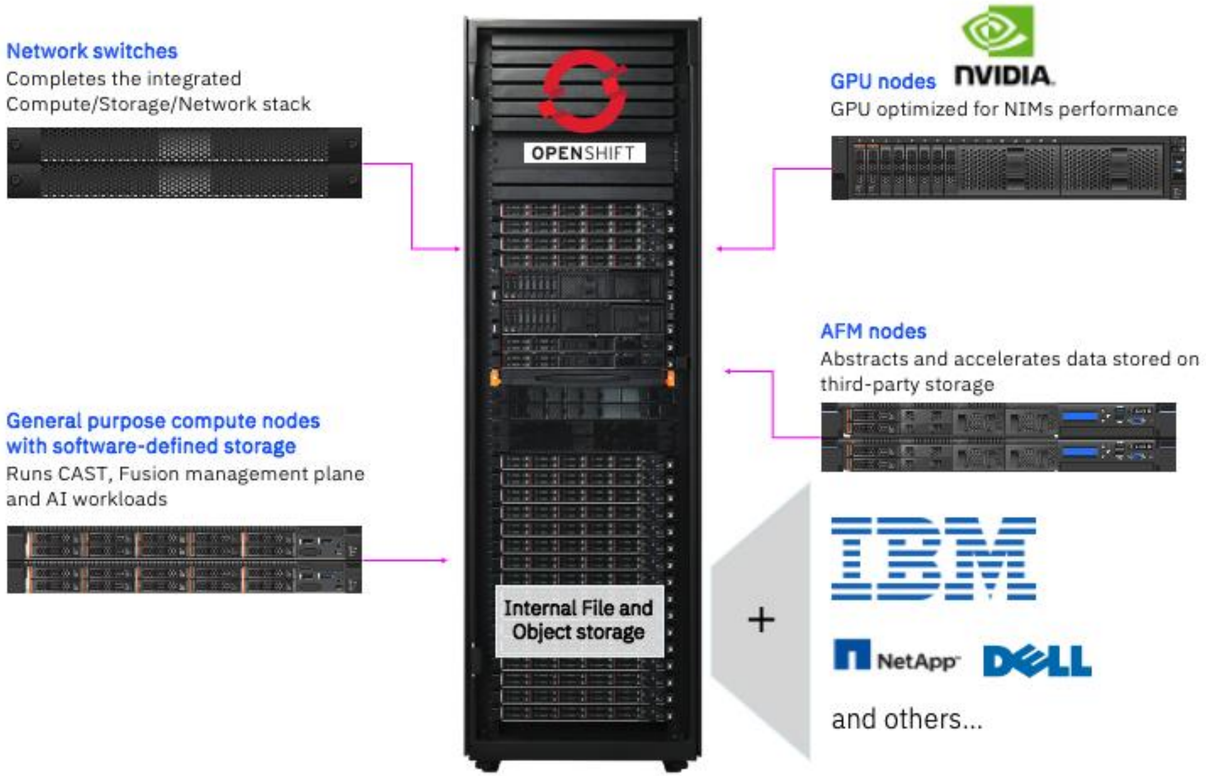
Inferencing storage aaS





IBM CAS – Single name space for distributed inferencing

Integrated Inferencing appliance



Simple

- Fully integrated with inferencing eco system: NVAIE and Openshift AI
- Turn-key, all-in-one
- Zero to inferencing in weeks not months

Works w/ legacy data

- Connects to IBM and third-party storage
- Unleash data without copying/moving

Enterprise hybrid cloud ready

- HA/DR/backup built-in
- Automated Day-2 operations
- Global data platform

Notices and disclaimers

This document is distributed “as is” without any warranty, either express or implied. In no event shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.

Case studies and client examples are presented as illustrations of how customers or IBM has used IBM products in production or test environments and the results they may have observed. Actual performance, cost, savings or other results in other operating environments may vary.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM.

Not all offerings are available in every country in which IBM operates.

Any statements regarding IBM’s future direction, intent or product plans are subject to change or withdrawal without notice.

Certain comments made in this presentation may be characterized as forward looking under the Private Securities Litigation Reform Act of 1995.

Forward-looking statements are based on the company’s current assumptions regarding future business and financial performance. Those statements by their nature address matters that are uncertain to different degrees and involve a number of factors that could cause actual results to differ materially. Additional information concerning these factors is contained in the Company’s filings with the SEC.

Copies are available from the SEC, from the IBM website, or from IBM Investor Relations.

Any forward-looking statement made during this presentation speaks only as of the date on which it is made. The company assumes no obligation to update or revise any forward-looking statements except as required by law; these charts and the associated remarks and comments are integrally related and are intended to be presented and understood together.