

Extreme scaling, without limits: **AI enabled Content Aware Storage**

IBM Storage Scale System 6000 & IBM Fusion CAS
Scaling Vector Database & KV Cache

IBM Storage Scale User Group Event

June 23rd, 2026 | Hamburg, Germany ISC

Ted Hoover (IBM)

<mailto:hoov@us.ibm.com>

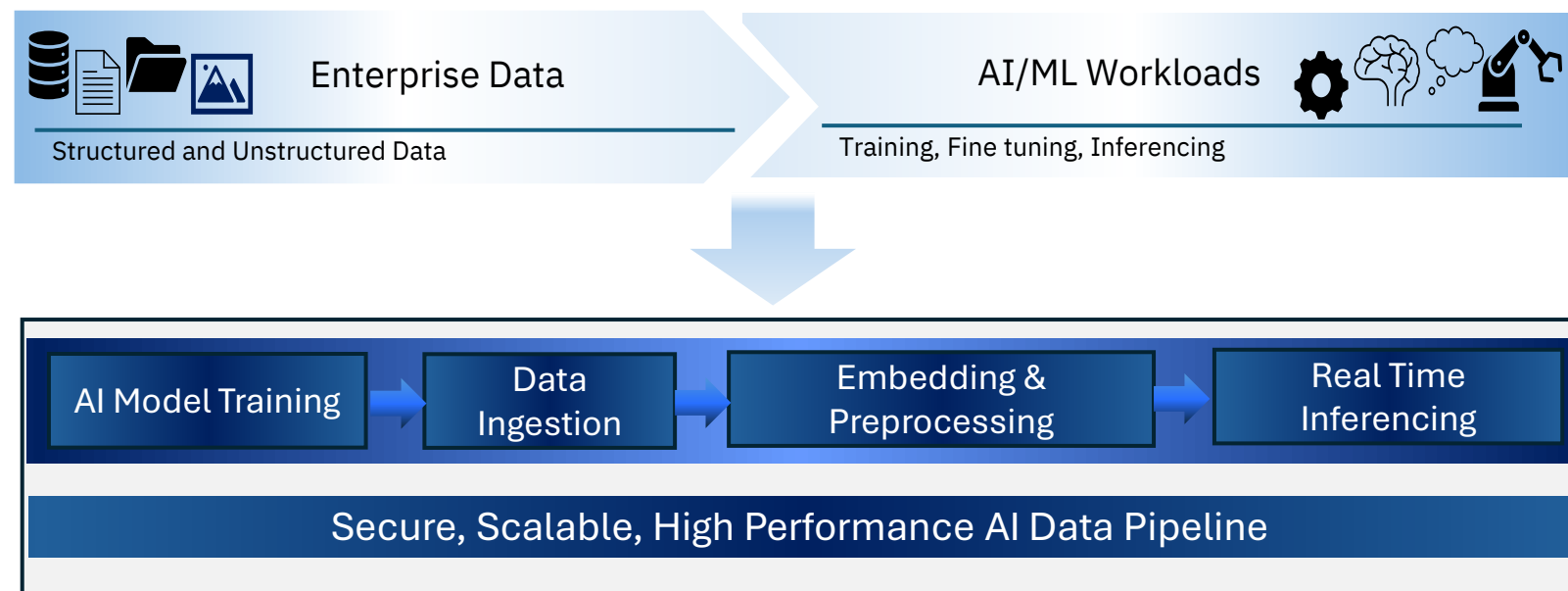
Simon Lorenz (IBM)

<mailto:simon.lorenz@de.ibm.com>



AI Data Inferencing Challenges

- **Data is Distributed:**
Too many copies of data and unnecessary data movement
- **Data Pipelines are Complex:**
Too many technologies to integrate and data types to support
- **Data is Critical Business Asset:**
Too many replicas of data with inconsistent ACL's between raw data and embedding
- **Data is Constantly Changing:**
Long latency between data change and updated vectors
- **Data is Growing:**
Limited scalability, Performance bottlenecks reduce time to value



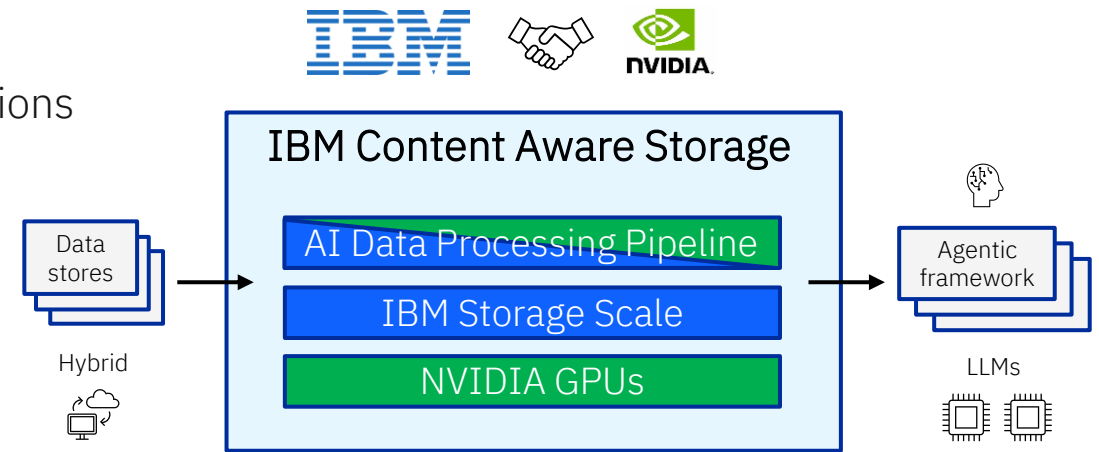
A new type of storage (Content Aware) is needed to address these challenges

What is IBM Content-Aware Storage (CAS)

 “CAS is the semantic interface to AI-ready storage”

- ✓ Extracts the meaning of unstructured enterprise data
- ✓ Prepares data for use by agentic & AI inferencing applications
- ✓ Enables contextual and lexical search of data in storage

 Fundamentally, CAS is a general **knowledge base** for generative **AI inferencing** applications



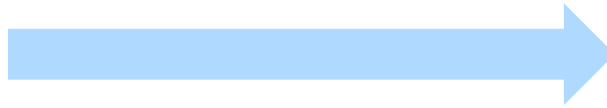
“CAS solves the 99% problem”

- Clients need to grant AI models access to their data so that they can perform useful work, and organizations are swamped with this unstructured data stored disparately across the enterprise...
- ...but 99% of this data isn't represented in AI model training sets or in a format that is useful to AI inferencing applications

What are we solving for?

Solving for how to get your data to AI applications **safely, **securely**, at **scale**, for the lowest **cost**!*

Enterprise Data

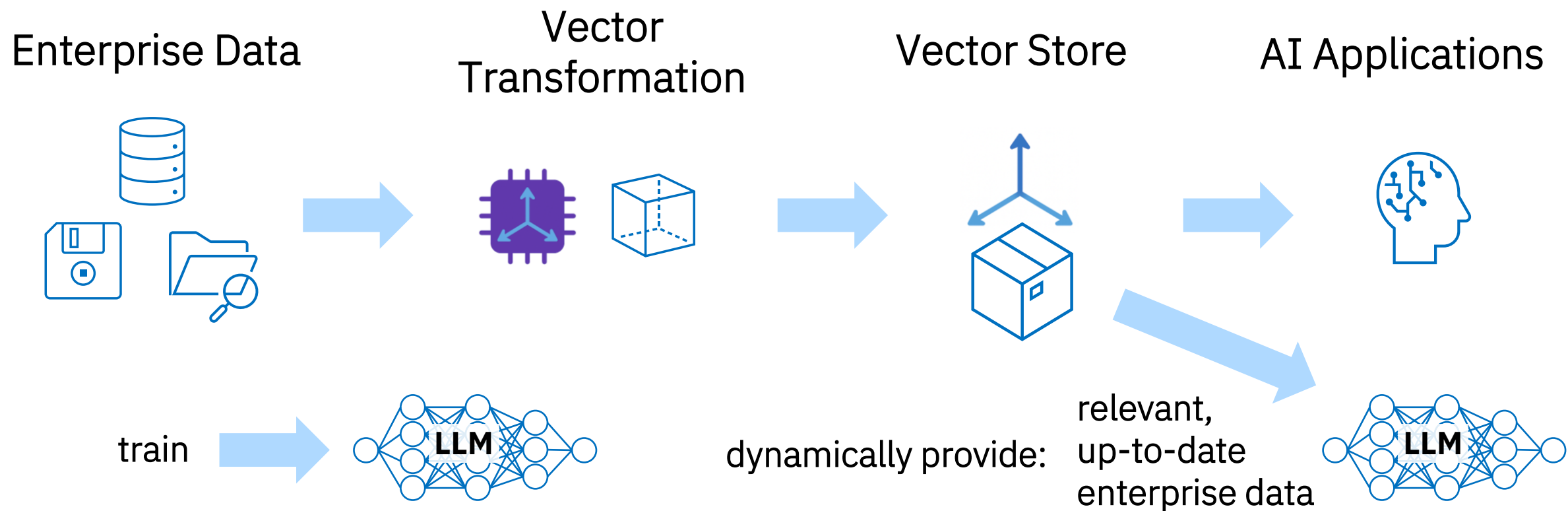


AI Applications



What's the basic answer?

A Retrieval-Augmented Generation (RAG) framework that **transforms your **data** into **vectors**, **stores** those vectors, and **serves** them **to AI Apps***



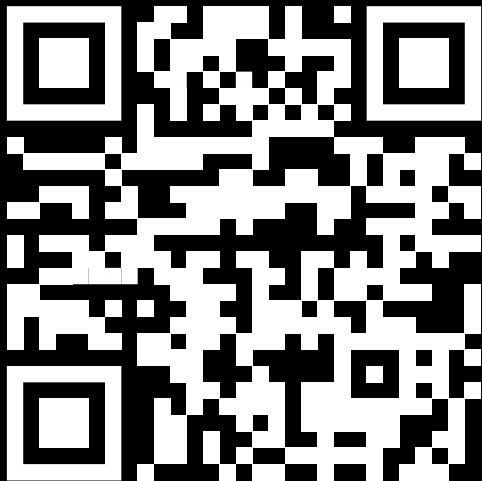
What is Retrieval-Augmented Generation (RAG)? And what features are added by Content Aware Storage (CAS)?



Retrieval-Augmented Generation



<https://www.youtube.com/@IBMTechnology>



<https://www.youtube.com/watch?v=T-D1OfcDW1M> 

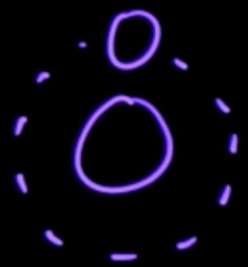
Retrieval-Augmented Generation



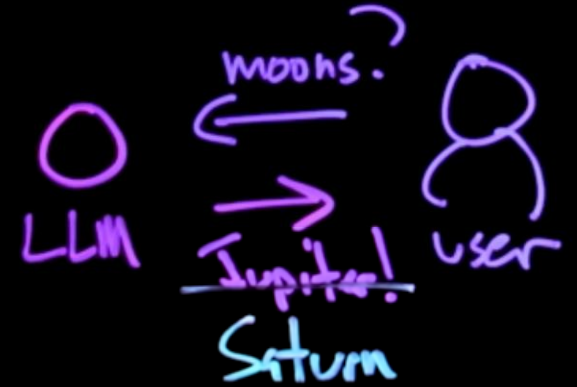
planet?
moons?



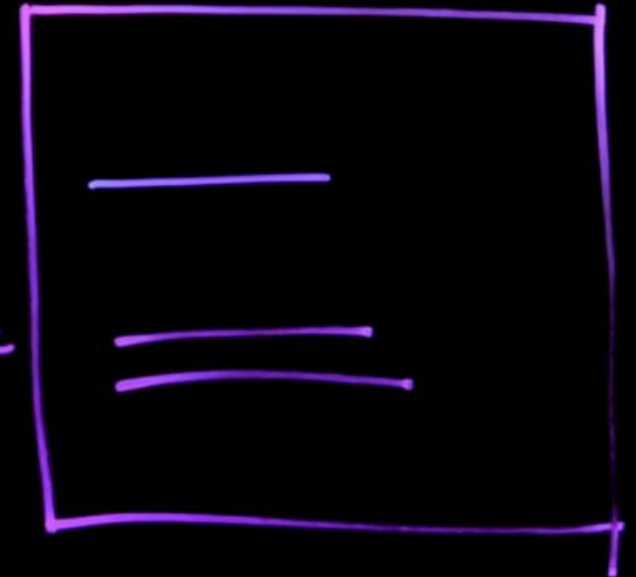
Retrieval-Augmented Generation



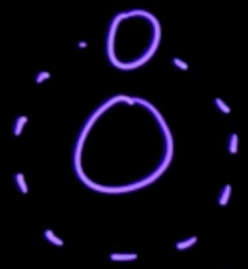
planet?
moons?



prompt
response



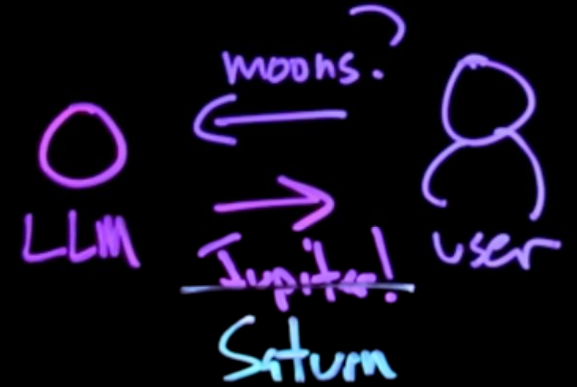
Retrieval-Augmented Generation



planet?
moons?

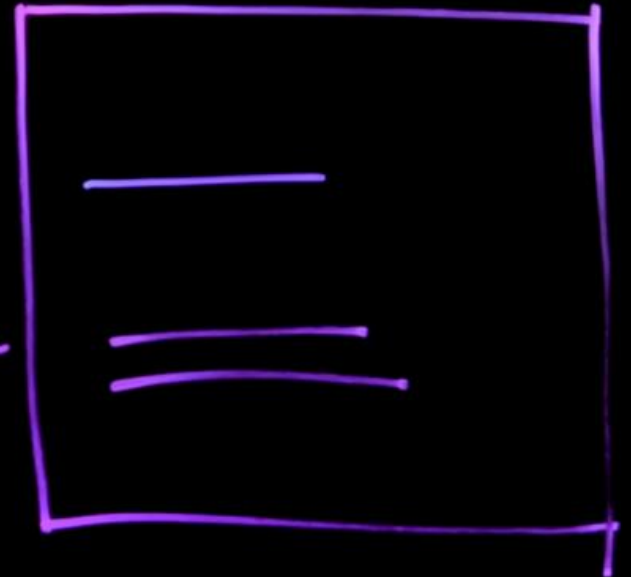
LLM challenges

1. No source
2. Out of date



prompt

response



Retrieval-Augmented Generation



planet?
moons?

LLM challenges

1. No source
2. Out of date

NASA



data



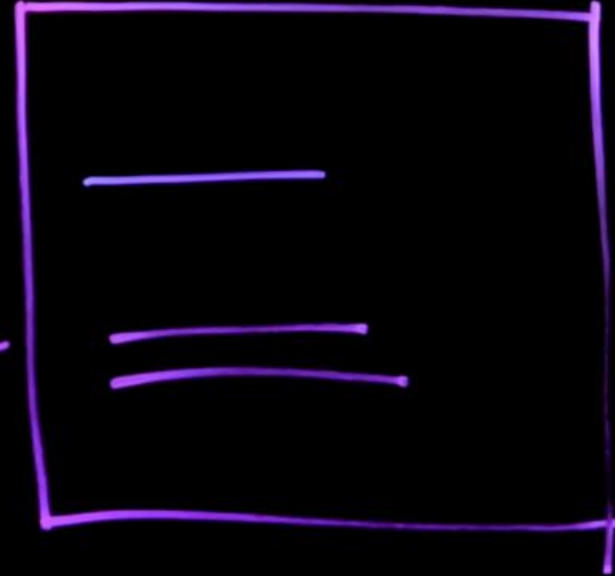
LLM



user

prompt

response



Retrieval-Augmented Generation

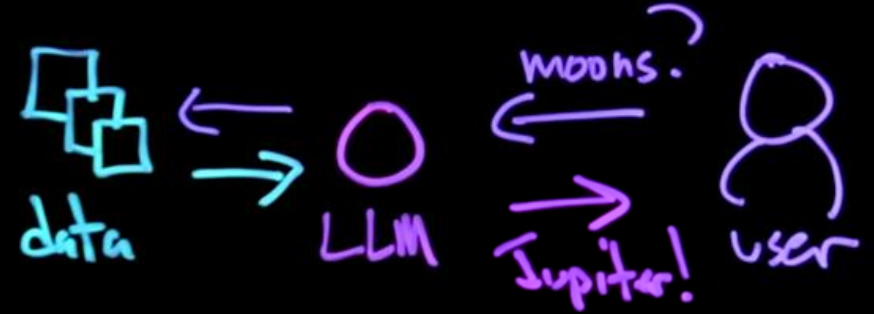


planet?
moons?

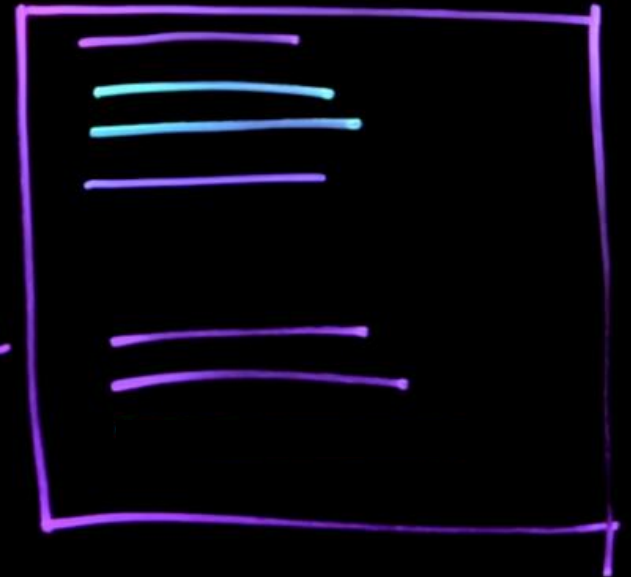
LLM challenges

1. No source
2. Out of date

NASA

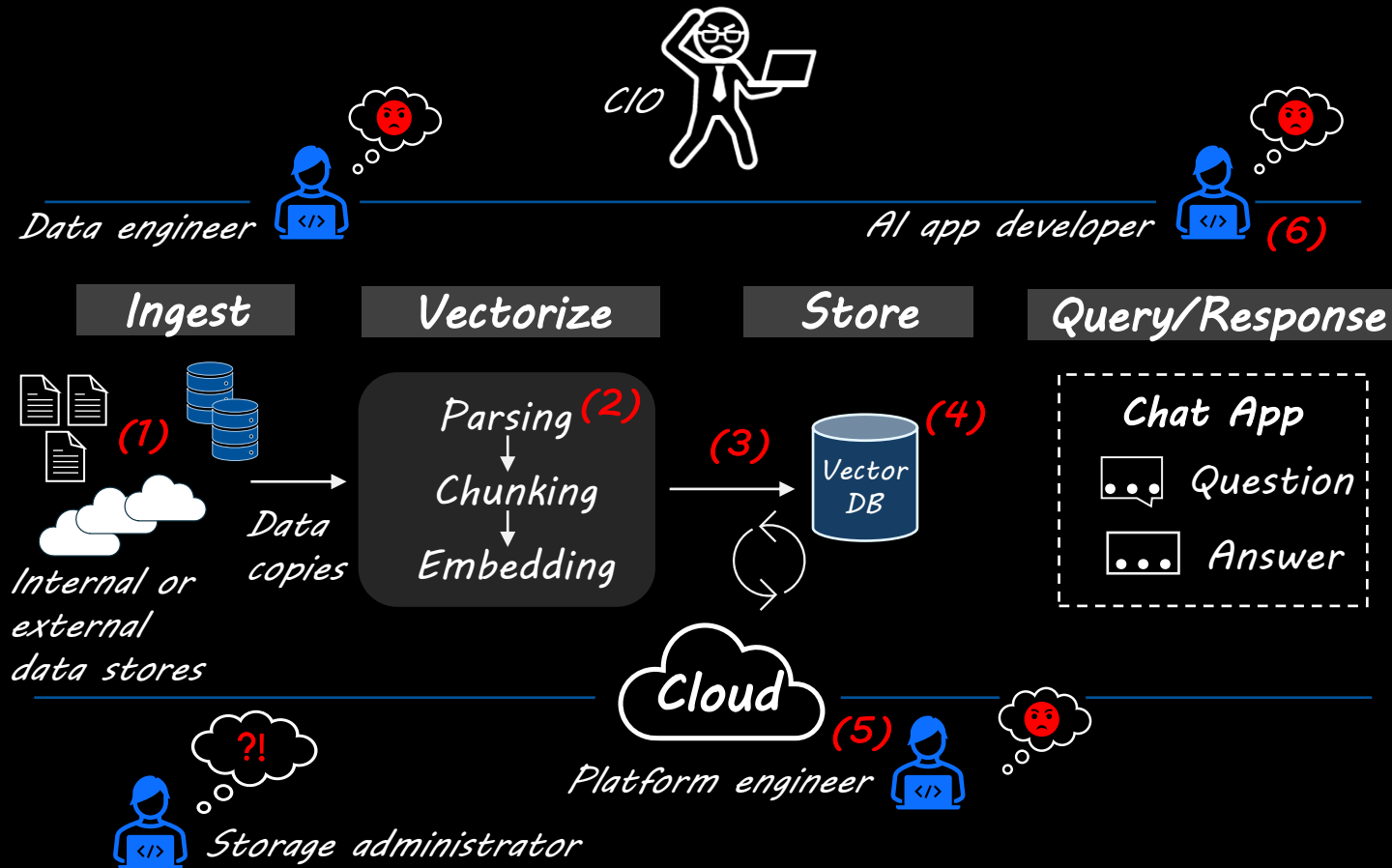


prompt
response



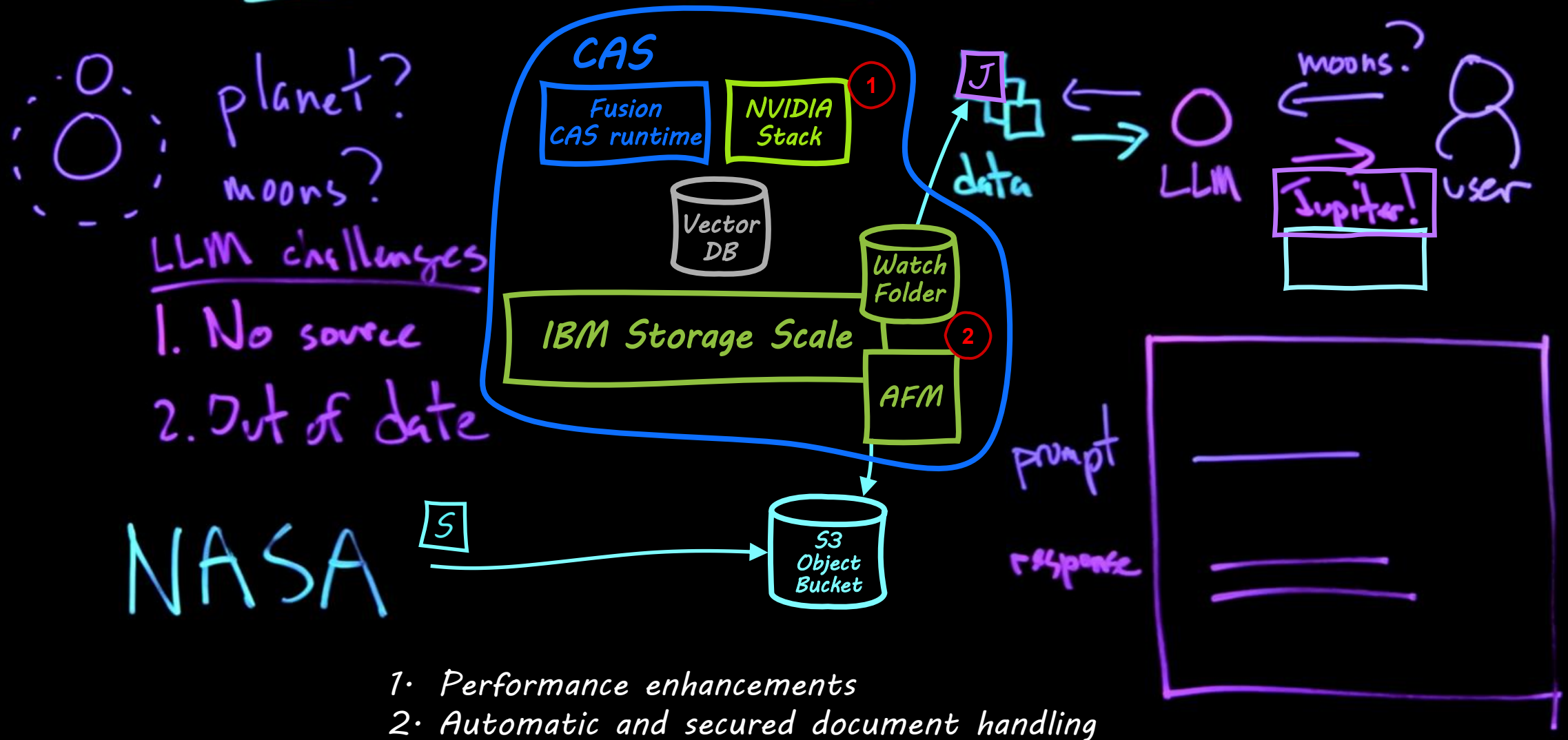
Retrieval-Augmented Generation

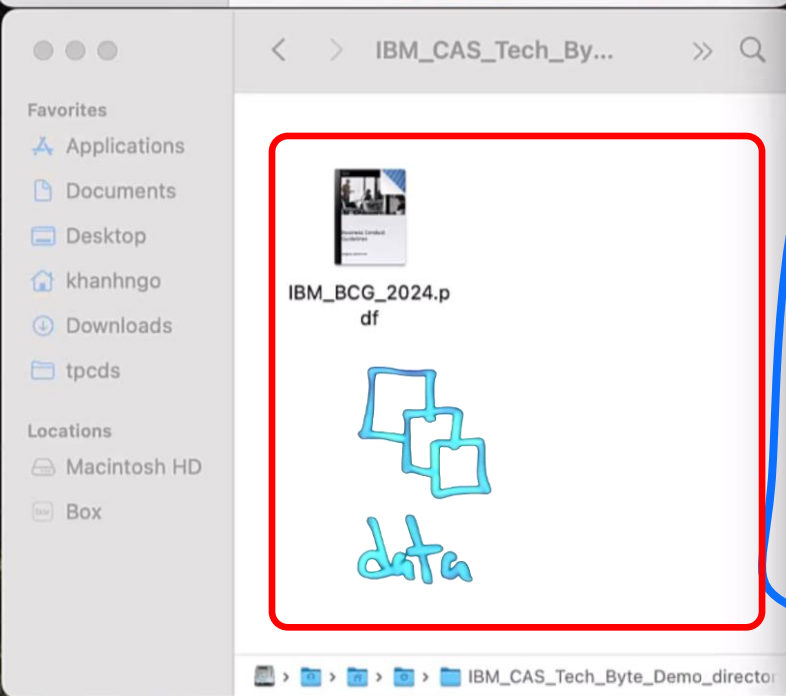
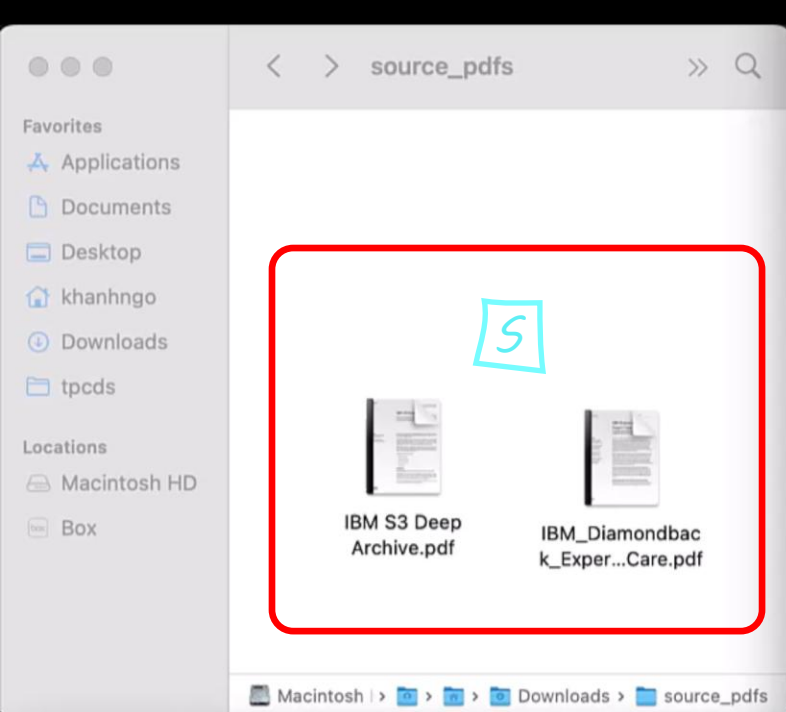
Limitations in existing RAG solutions



- 1) Data preparation, cleaning; disconnected data stores
- 2) Data copied multiple times in unsecure environment
- 3) Weekly batch refresh and rebuild of vector DB
- 4) Small fraction of data stored in vector DB
- 5) High, unpredictable monthly costs
- 6) Complexity saps FTE working hours

Retrieval-Augmented Generation





Docs processed for chat:
IBM_BCG_2024.pdf

Refresh

RAG with IBM Content Aware Storage

Granite 3.2 8b instruct is an open-source LLM supporting a 128k context window. Start with one of the sample prompts or enter your own. Keep in mind that AI can occasionally make mistakes. [View Documentation](#)

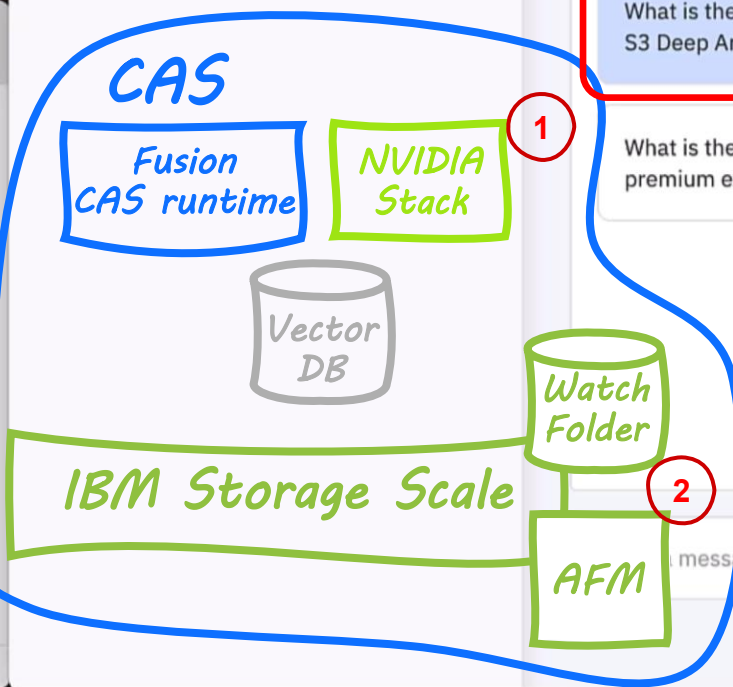
Chatbot

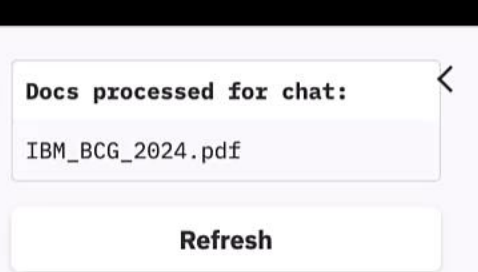
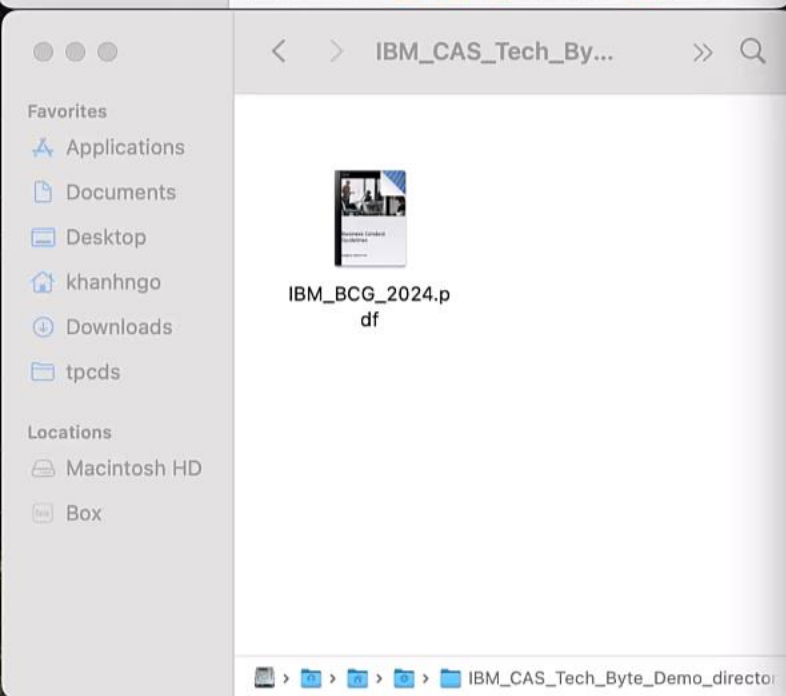
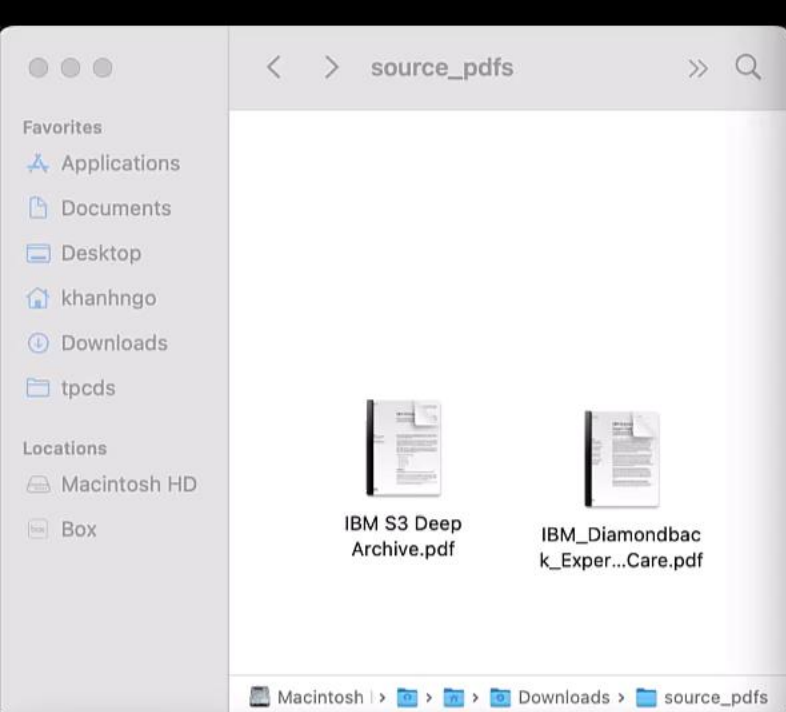
What is the energy consumption of the IBM S3 Deep Archive entry configuration?

What are the highlights of the IBM S3 Deep Archive solution?

What is the difference between basic vs premium expert care for IBM Diamondback?

What are the highlights of IBM Diamondback Expert Care?





RAG with IBM Content Aware Storage

Granite 3.2 8b instruct is an open-source LLM supporting a 128k context window. Start with one of the sample prompts or enter your own. Keep in mind that AI can occasionally make mistakes. [View Documentation](#)

Chatbot

What is the energy consumption of the IBM S3 Deep Archive entry configuration?

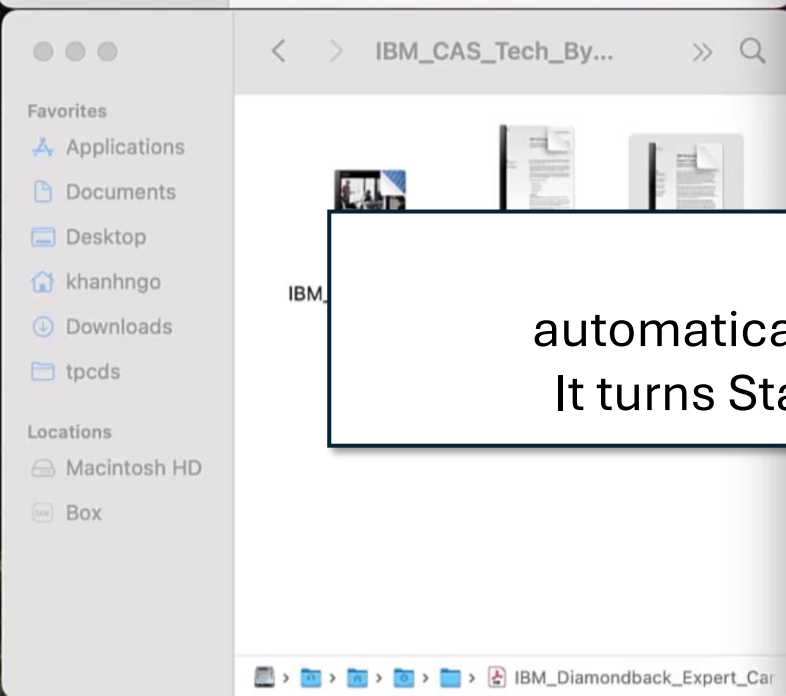
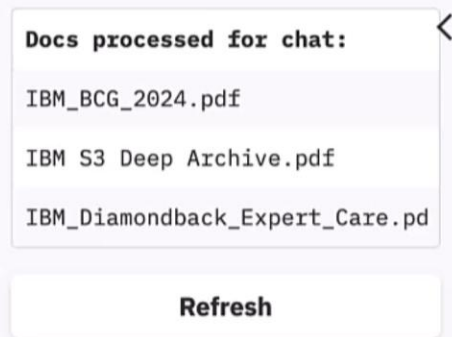
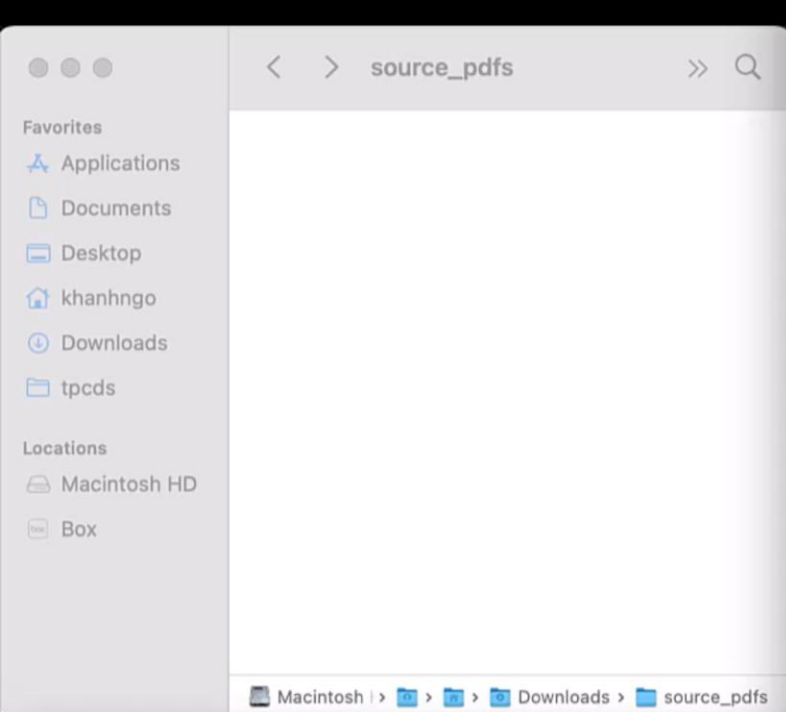
What are the highlights of the IBM S3 Deep Archive solution?

What is the difference between basic vs premium expert care for IBM Diamondback?

What are the highlights of IBM Diamondback Expert Care?

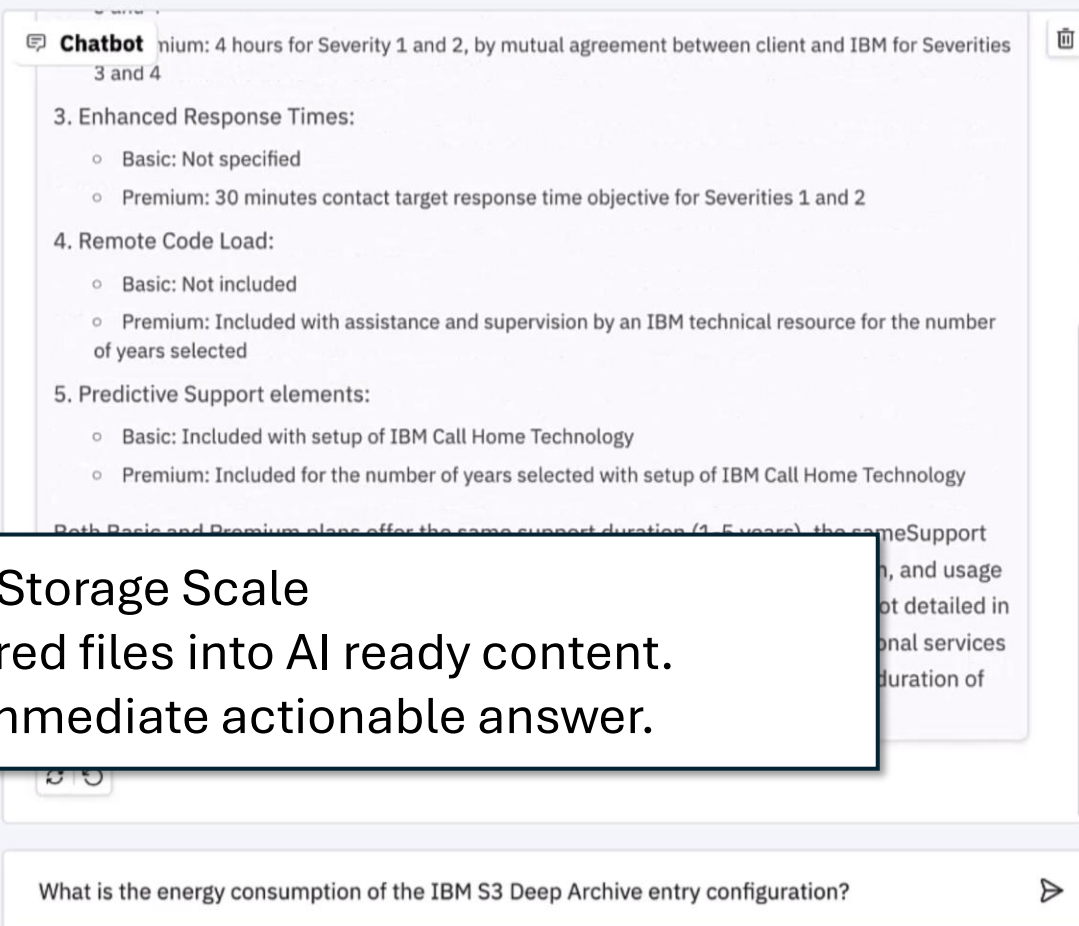
Type a message...





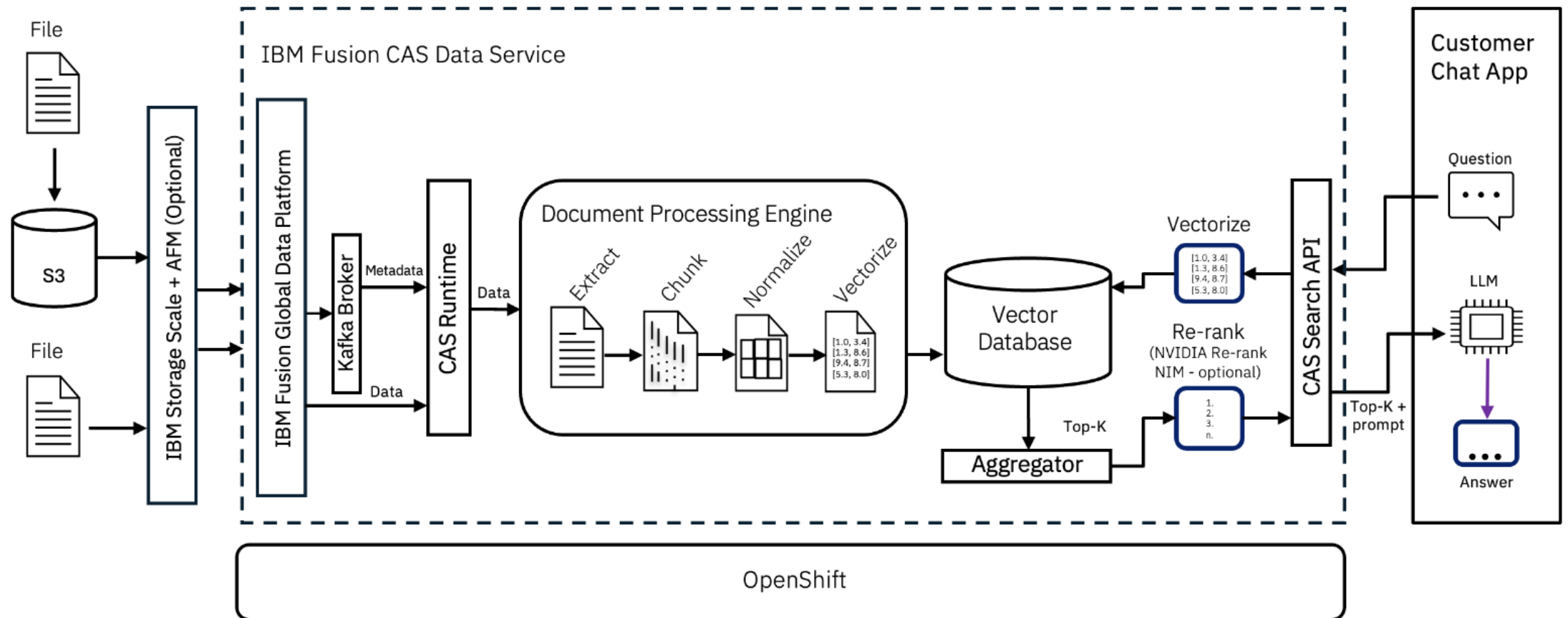
RAG with IBM Content Aware Storage

Granite 3.2 8b instruct is an open-source LLM supporting a 128k context window. Start with one of the sample prompts or enter your own. Keep in mind that AI can occasionally make mistakes. [View Documentation](#)



Content Aware IBM Storage Scale
automatically transforms unstructured files into AI ready content.
It turns Static documents into an immediate actionable answer.

Accelerating AI and Analytics with IBM Fusion CAS and IBM Storage Scale Data Flow

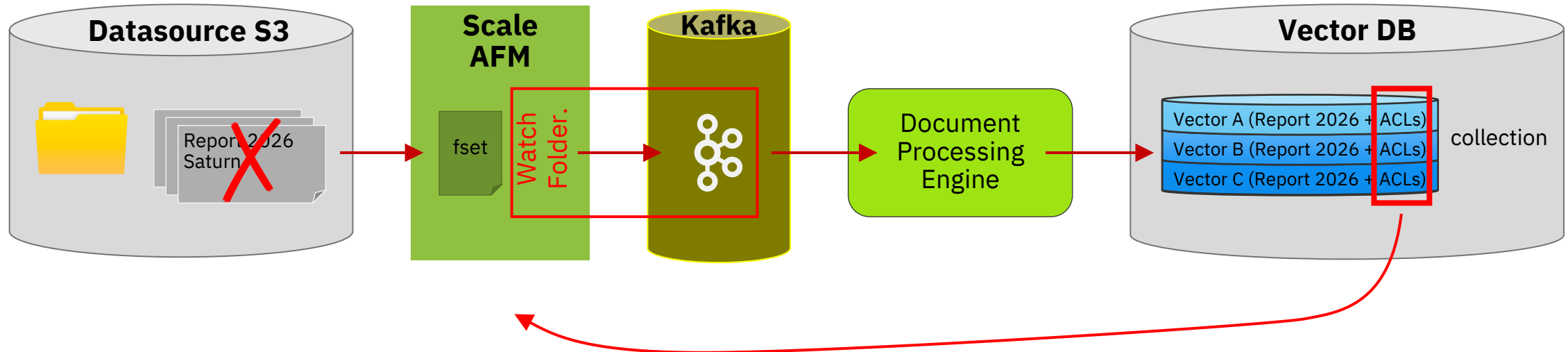


Accelerating AI and Analytics with IBM Fusion CAS and IBM Storage Scale

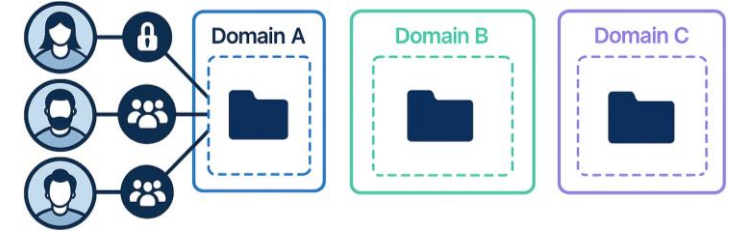
Features: AFM - with watch folders

Data Sovereignty & In-Place Processing

- CAS indexes data in place without copying it, respects existing enterprise security policies and storage systems (e.g., IBM Storage Scale, S3 (via AFM)).
- CAS uses watch folders and active file management to incrementally detect and index new or changed files for RAG.



Accelerating AI and Analytics with IBM Fusion CAS and IBM Storage Scale Security (ACL preservation), Multi tenant



Domain-Level Access & CAS Resource Access Control (CRAC)

- CAS organizes **content** into **domains**, which act as logical **boundaries for tenants or projects**.
- **CRAC (CAS Resource Access Control) policies** define which **users** or **groups** can **access** a domain.
- Ensures tenants are isolated at domain level, preventing cross-tenant data exposure.

File-Level Security via LDAP + Identity Provider (IDP)

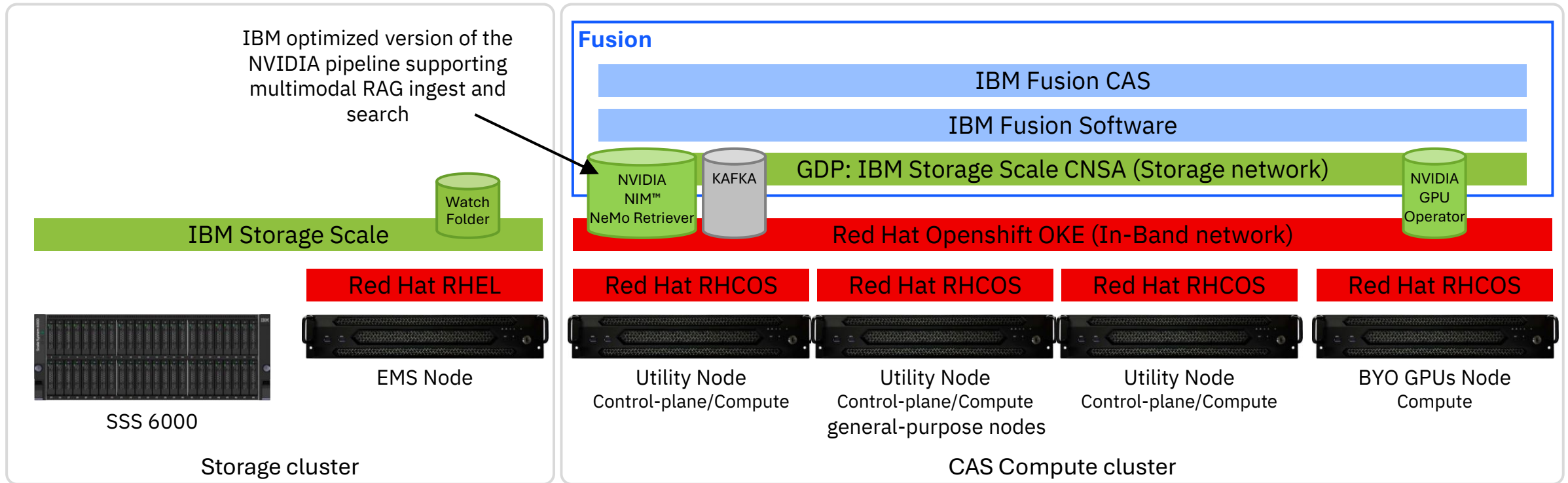
- Beyond domain-level control, CAS supports **file-level access enforcement**.
- Integrates with:
 - **LDAP directory services** (e.g., OpenLDAP) for user and group attributes.
 - **Identity Providers** (IDPs) using OpenID Connect (e.g., Keycloak) for authentication and claims.
- Storage **ACL's** are directly **applied to the vectors**, CAS can:
 - keep up with any changes that made to file access.
 - apply them in real-time to the associated vectors.

Accelerating AI and Analytics with IBM Fusion CAS and IBM Storage Scale Security (ACL preservation), Multi tenant

Query Access Control

- A user queries CAS → the system checks **real-time ACLs** and filters results so only files the user has permission to read are returned.
- CAS applies **query-time filtering** based on **user identity** and **group memberships**.
- Prevents unauthorized retrieval of sensitive documents during RAG operations.
- The configuration involves mapping LDAP attributes and IDP claims to CAS security policies.

Accelerating AI and Analytics with IBM Fusion CAS and IBM Storage Scale Core Building Blocks

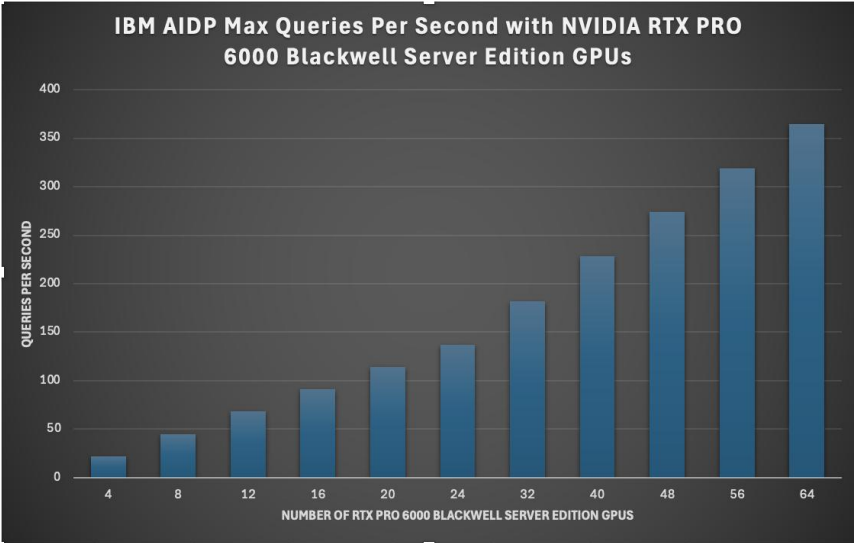
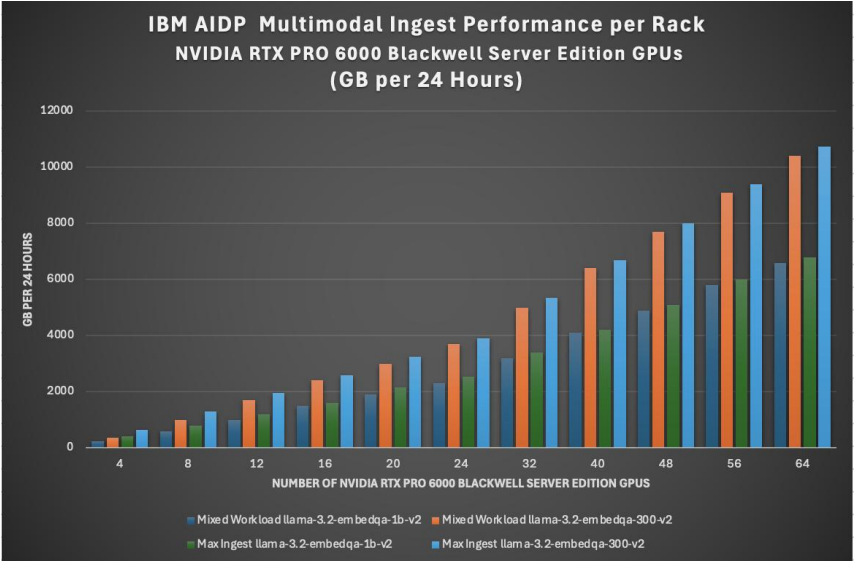


IBM AIDP with CAS and Storage Scale Performance Multimodal Data

- 1.3TB / day max multimodal ingest per IBM AIDP GPU server with NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPUs^{1,2}
- 10.75TB / day max multimodal ingest per AIDP rack ^{1,2}

	Max Multimodal Ingest ^{1,2} (TB/day Pages/sec)	Max Ingest Queries per second ³	Mixed Workload ⁴ (TB/day Pages/sec)
Single GPU Server with 8x NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPUs	1.3 116	45	0.35 31
Max AIDP Rack with 8x GPU Servers and 8x NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPUs	10.75 962	365	10.4 930

- 1) Multimodal content consisting of text, tables, charts & infographics based on Digital Corpora Dataset; No queries during ingest period
- 2) Using embedqa-300-v2 embedding NIM, 15% raw text to pdf file size ratio
- 3) No parallel ingest – all GPUs used for queries using llama-3.2-rerankqa-1b-v2 Reranking NIM
- 4) 10 queries per second and 40 re-ranks per query with parallel ingest



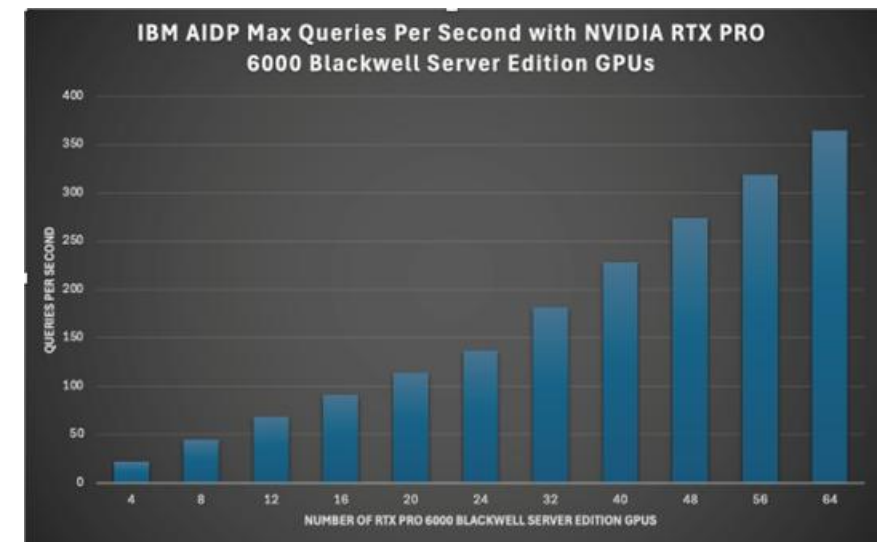
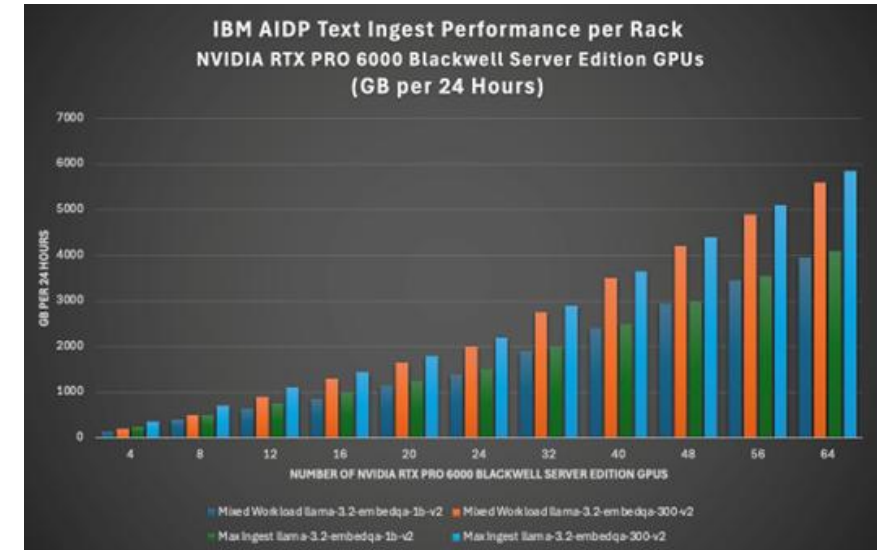
IBM AIDP with CAS and Storage Scale Performance

Text Data

- Max 4,420 Pages per second (5.7TB per day) text ingest per IBM AIDP GPU server^{1,2}
- 10.75TB/day max multimodal ingest per AIDP rack^{1,2}

	Max Multimodal Ingest ^{1,2} (TB/day Pages/sec)	Max Ingest Queries per second ³	Mixed Workload ⁴ (TB/day Pages/sec)
Single GPU Server with 8x NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPUs	0.7 543	45	0.5 388
Max AIDP Rack with 8x GPU Servers and 8x NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPUs	5.7 4420	365	5.6 4343

- 1) Text PDF content; 67,000 text pages per GB; based on Digital Corpora dataset; No queries during ingest period
- 2) Using embedqa-300-v2 embedding NIM, 15% raw text to pdf file size ratio
- 3) No parallel ingest – all GPUs used for queries using llama-3.2-rerankqa-1b-v2 Reranking NIM
- 4) 10 queries per second and 40 re-ranks per query with parallel ingest



CAS Configuration – Multimodal Data Ingest and Query (Mixed Workload) Sizing

Up to 8 GPU servers per rack

- 4/8 x RTX Pro 6000 GPU per G04 (runs NVIDIA NeMo Retriever, NVIDIA NeMo Reranking NIM)
- 3 x ESS Utility Node (CAS core components)

GPU Servers	NeMo Retriever Instances	Max RTX 6000 GPUs	CPU Worker Node	Physical Cores	vCPU ¹	Memory (GB)	RAG Ingest (I) llama-3.2-embedqa-1b-v2 Embedding Model		RAG Ingest embedqa-300-v2 Embedding Model	
							TB/day	Pages / sec	TB/day	Pages / sec
1	4	4	3	128	256	2304	0.24	21	0.35	31
	8	8	3	128	256	2304	0.6	54	1	89
2	12	12	3	256	512	4608	1	89	1.7	152
	16	16	3	256	512	4608	1.5	134	2.4	215
3	20	20	3	384	768	6912	1.9	170	3	268
	24	24	3	384	768	6912	2.3	206	3.7	331
4	32	32	3	512	1024	9216	3.2	286	5	447
5	40	40	3	640	1280	11520	4.1	367	6.4	573
6	48	48	3	768	1536	13824	4.9	438	7.7	689
7	56	56	3	896	1792	16128	5.8	519	9.1	814
8	64	64	3	1024	2048	18432	6.6	590	10.4	930

Factors	
GPU	NVIDIA RTX PRO 6000
Embedding model	llama-3.2-embedqa-1b-v2
	llama-3.2-embedqa-300-v2
Reranking model	llama-3.2-rerankqa-1b-v2
Document Modality	Multimodal
Average # of Images Per Page	2
Raw text data volume percent ³	15%
Ingestion SLA (hours)	24
Queries per second	10
# of chunks to rerank	40

- 1) Available vCPU is presented at SMT=2
- 2) 1GB of Multimodal content equates to ~7,730 pages consisting of text, tables, charts & infographics based on Digital Corpora dataset
- 3) 215,000,000 words per GB of text; 1.33 tokens per word; 512 chunk size (tokens)

Accelerating AI and Analytics with IBM Fusion CAS and IBM Storage Scale

High-Performance Storage Reference Architecture

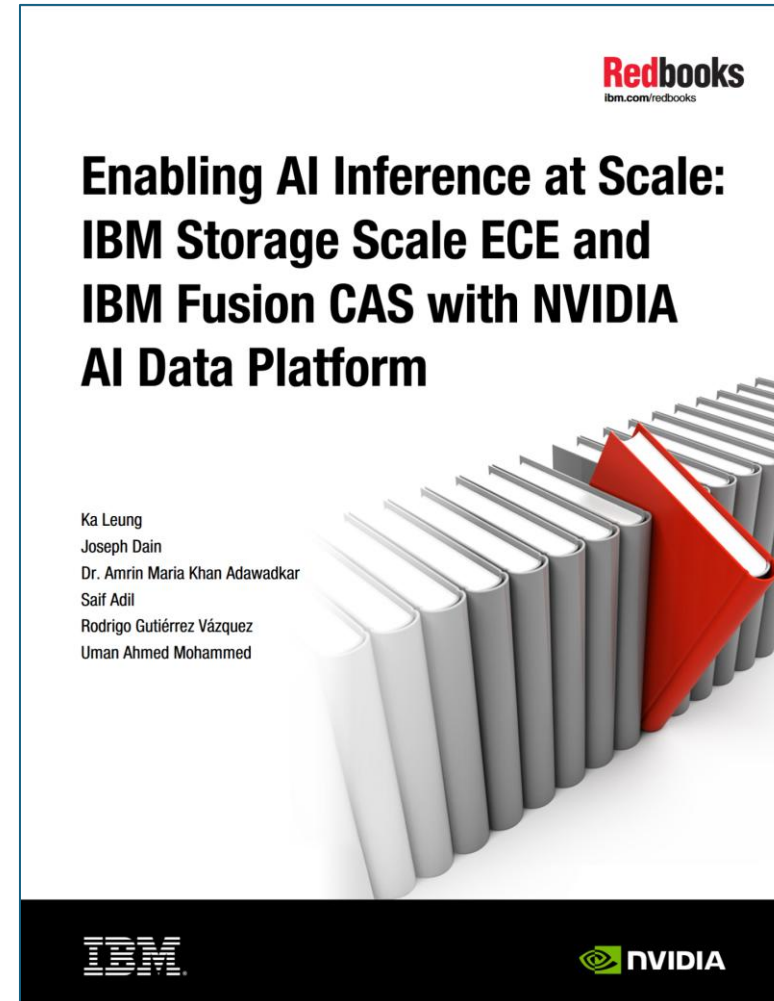


**Accelerating AI and Analytics with
IBM Fusion CAS and IBM Storage Scale**

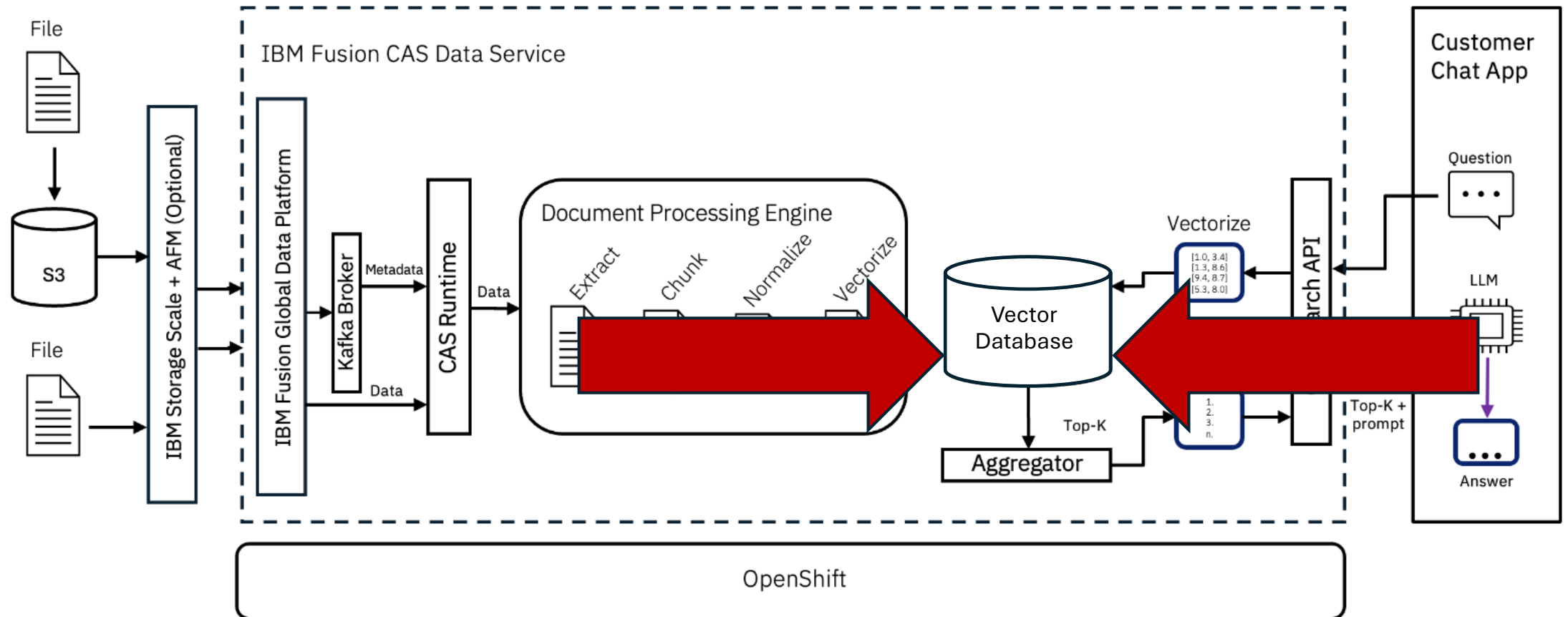
IBM Storage Scale and
IBM Storage Scale System 6000 and
IBM Fusion Content-Aware Storage (CAS)

High-Performance Storage Reference Architecture

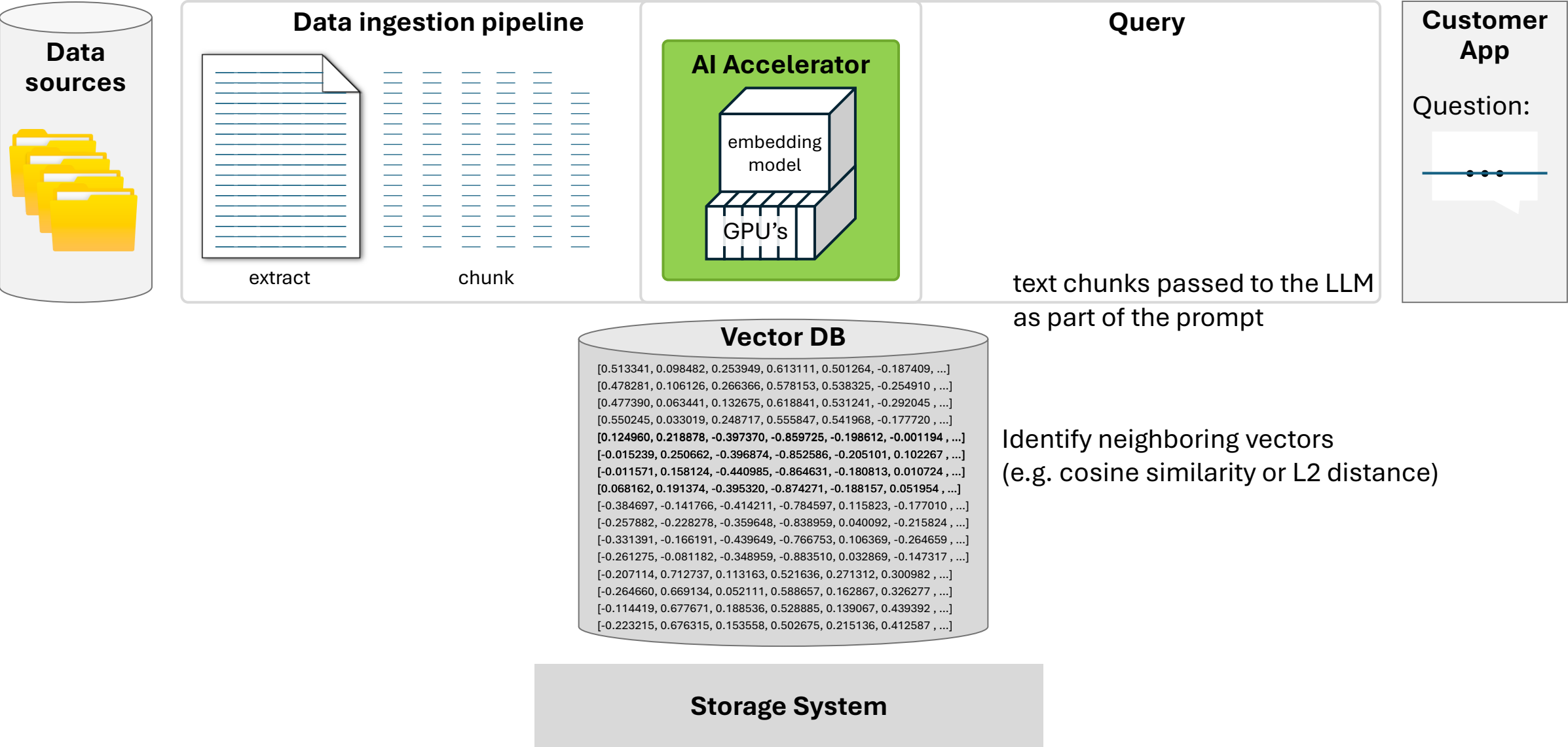
Version 1.0



Accelerating AI and Analytics with IBM Fusion CAS and IBM Storage Scale Data Flow



RAG, the Process



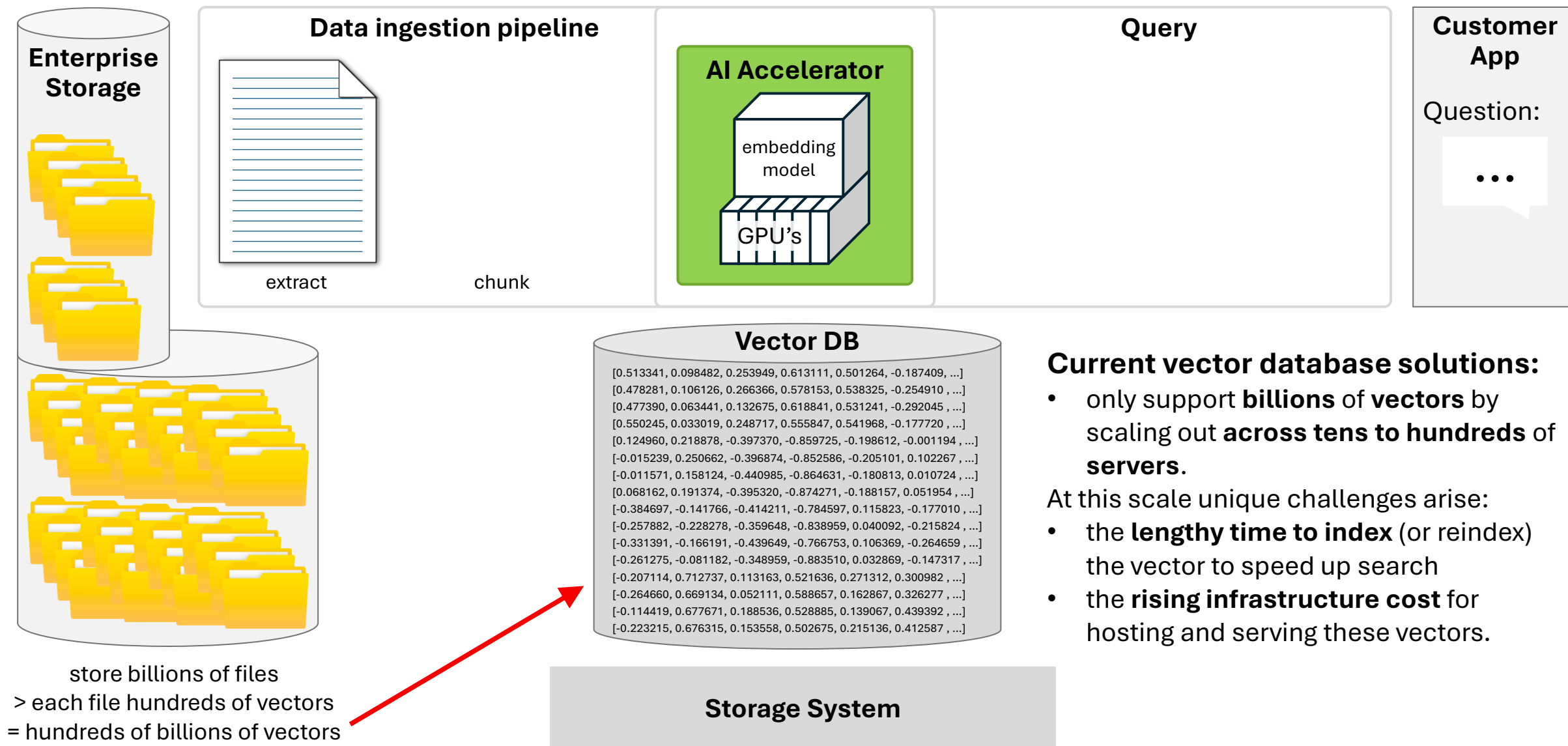
RAG, the Process

Embedding

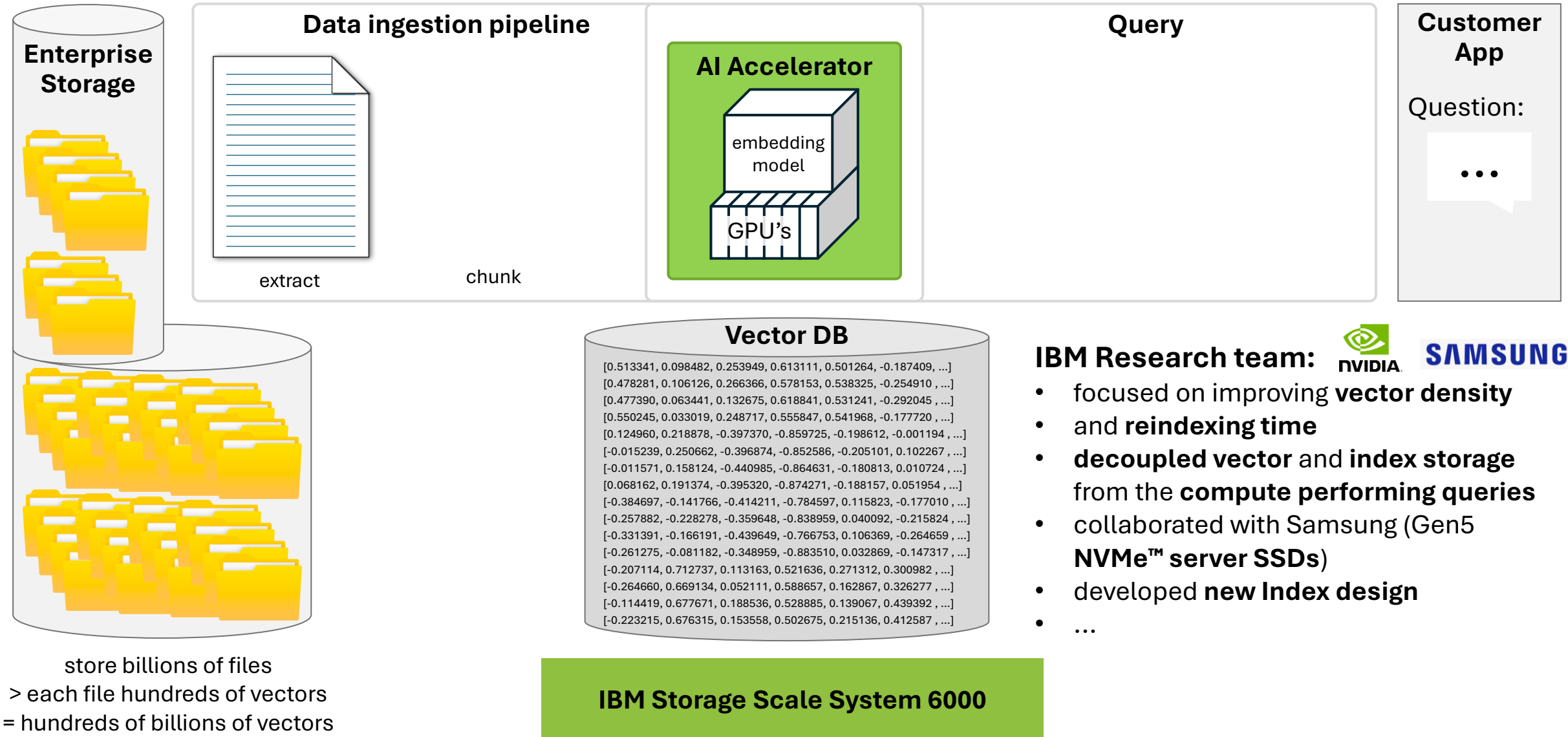
Context / Features	[0.00, 0.00, 0.00, 0.00, 0.95, 0.90, 0.85, 0.75]		[0.95, 0.90, 0.90, 0.70, 0.00, 0.00, 0.10, 0.00]	
	[0.00, 0.00, 0.00, 0.00, 0.90, 0.95, 0.95, 0.80]		[0.90, 0.90, 0.85, 0.80, 0.00, 0.00, 0.10, 0.00]	
				
Edible / Taste	0.0	0.0	0.9	0.95
Biological / Natural	0.0	0.0	0.9	0.9
Grows in Nature	0.0	0.0	0.8	0.9
Sweet / Flavor-related	0.0	0.0	0.8	0.7
Technology / Devices	0.9	0.9	0.0	0.0
Software / Digital	0.9	0.9	0.0	0.0
Business / Enterprise	0.95	0.8	0.1	0.1
Industrial / Manufactured	0.8	0.7	0.0	0.0

Embedding: Numeric representation of text

RAG, challenges of scaling



RAG, challenge solved: IBM Storage Scale System 6000



IBM Research team:  **NVIDIA** 

- focused on improving **vector density**
- and **reindexing time**
- **decoupled vector** and **index storage** from the **compute performing queries**
- collaborated with Samsung (Gen5 **NVMe™ server SSDs**)
- developed **new Index design**
- ...

Extreme scaling Vector Database

able to demonstrate loading, indexing, and querying of
(384 dimensions, full precision floating point)

100

BILLION Vectors

QUERY LATENCY

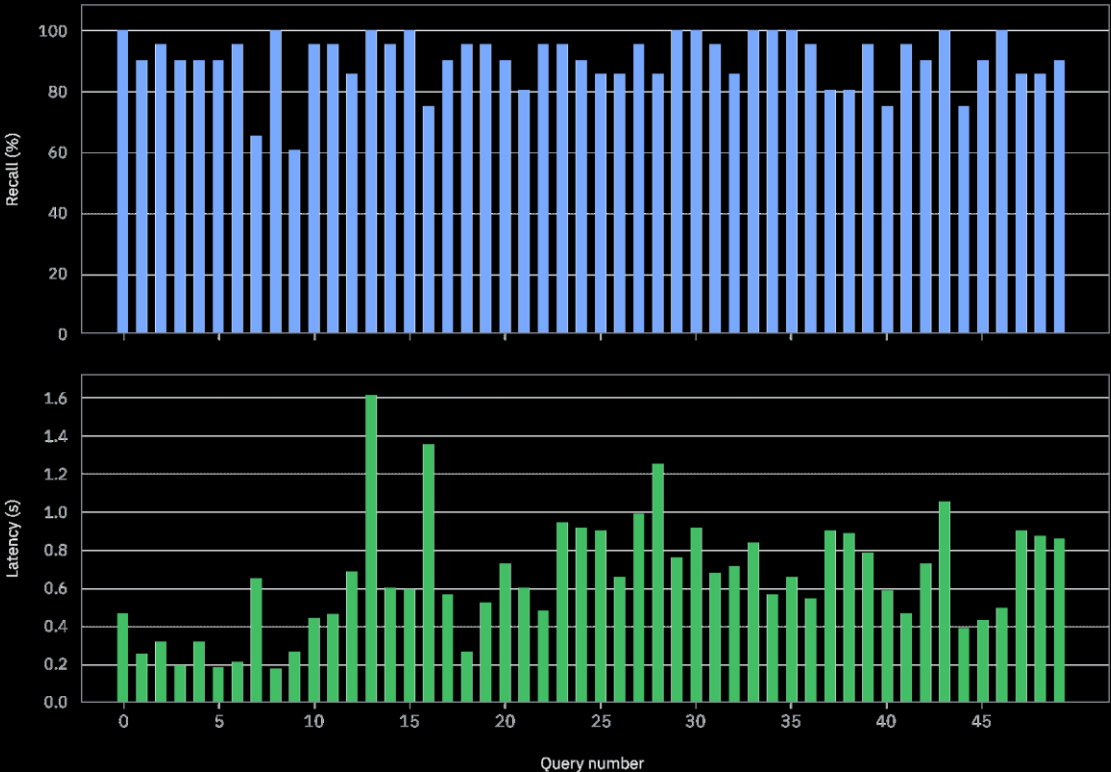
RECALL PRECISION

<700ms 90%+

Single 153 TiB
IBM Storage Scale System 6000



Query recall and latency for 100B vectors



The result: a mean query latency of 694 milliseconds with a recall precision of 90%.

Extreme scaling Vector Database

able to demonstrate loading, indexing, and querying of
(384 dimensions, full precision floating point)

100

BILLION Vectors

QUERY LATENCY

RECALL PRECISION

<700ms **90%+**

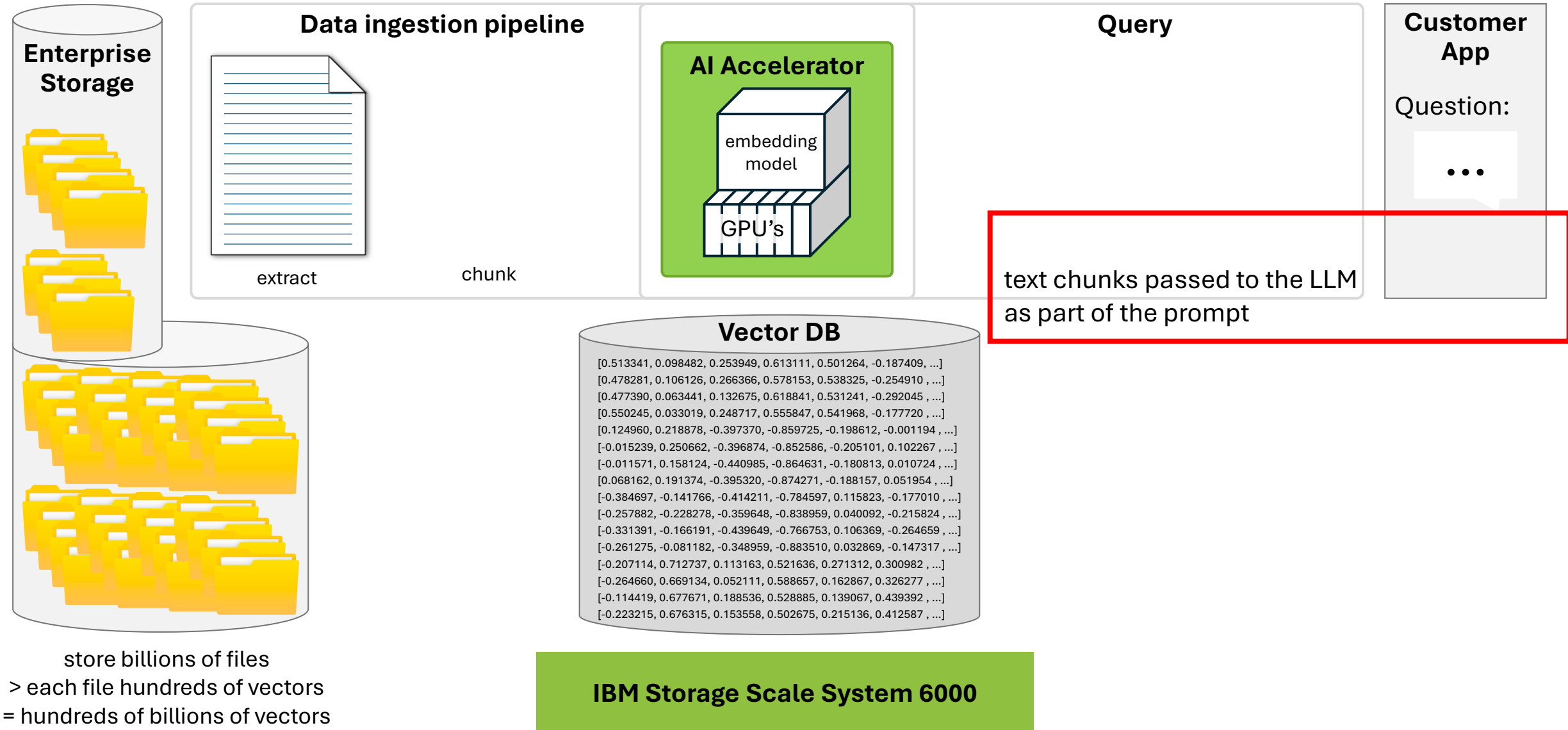
Single 153 TiB
IBM Storage Scale System 6000



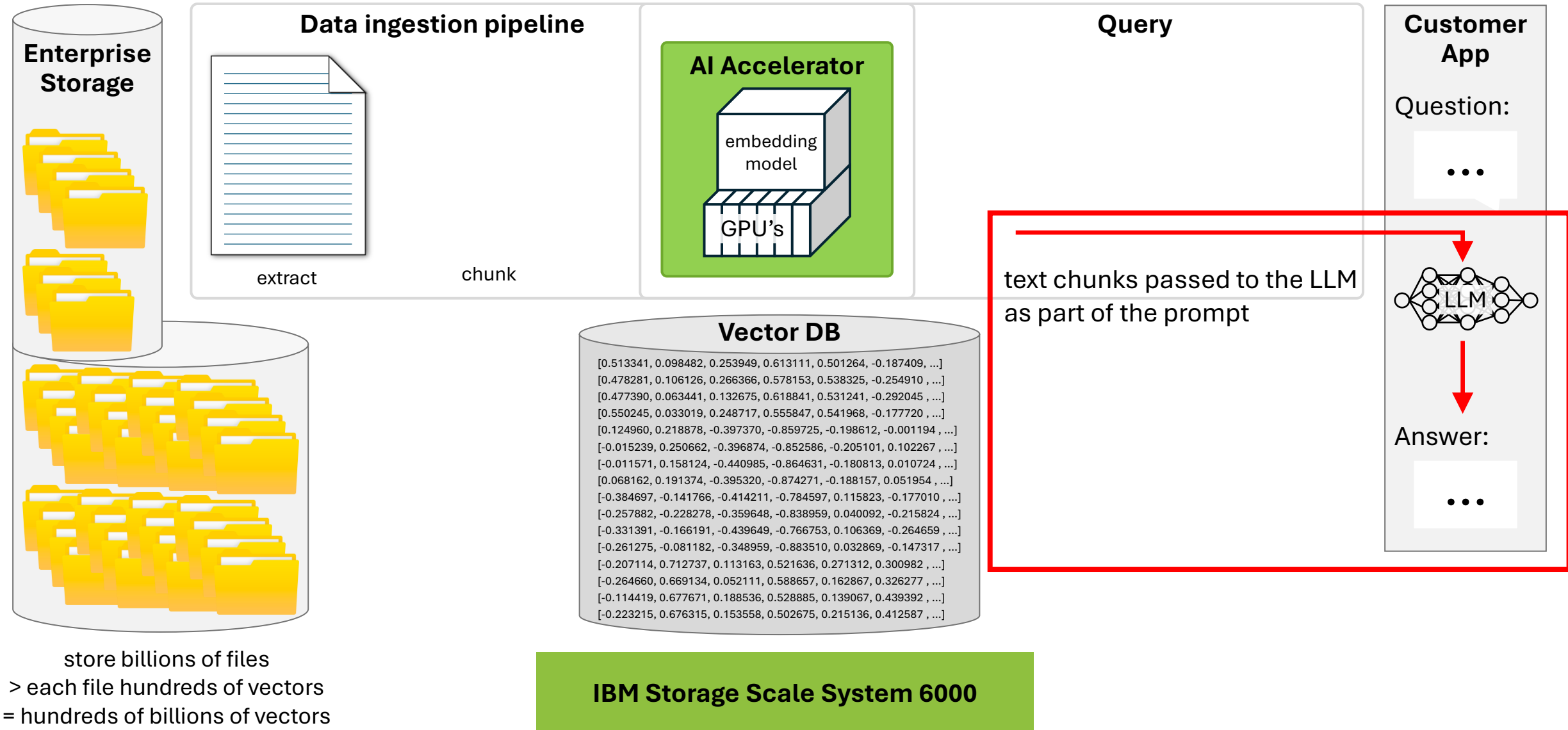
<https://research.ibm.com/blog/cas-100-billion-vector-storage-ai>

Further enhancements ongoing!

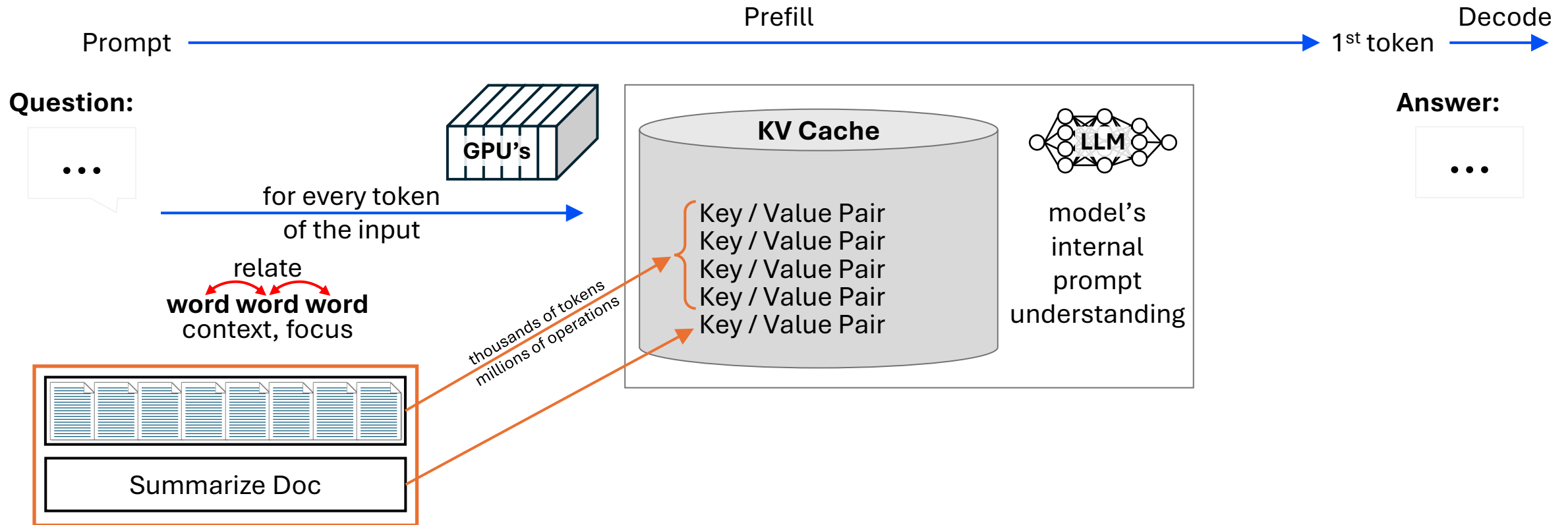
RAG, challenge solved: IBM Storage Scale System 6000



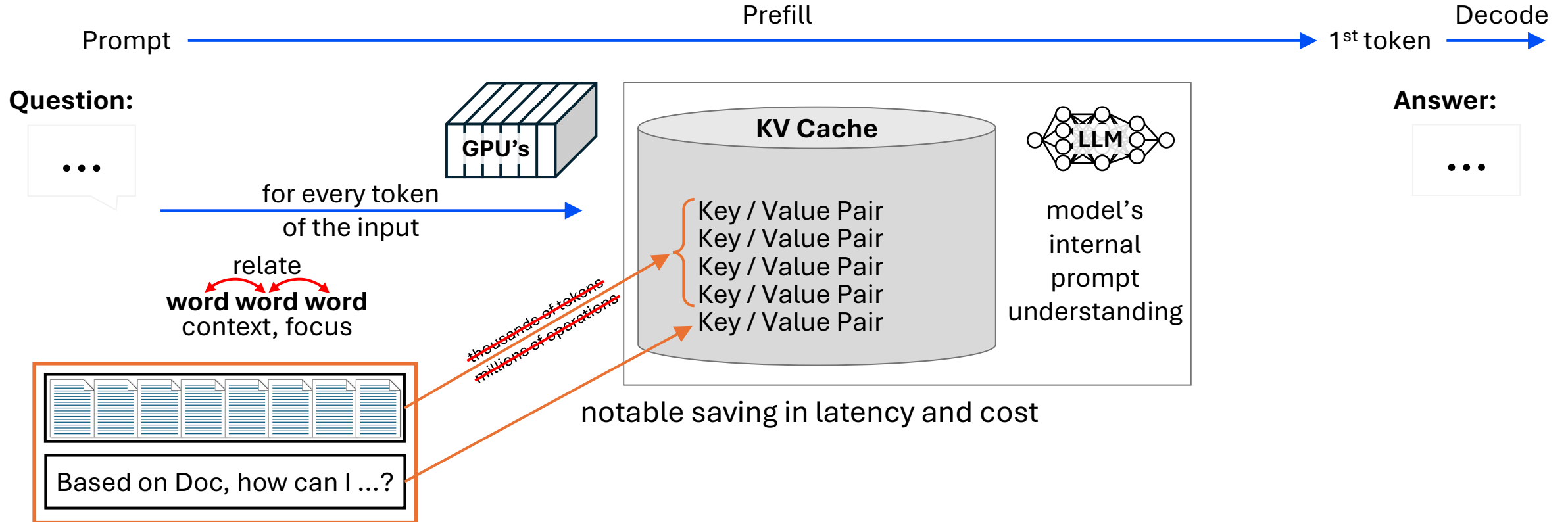
RAG, challenge solved: IBM Storage Scale System 6000



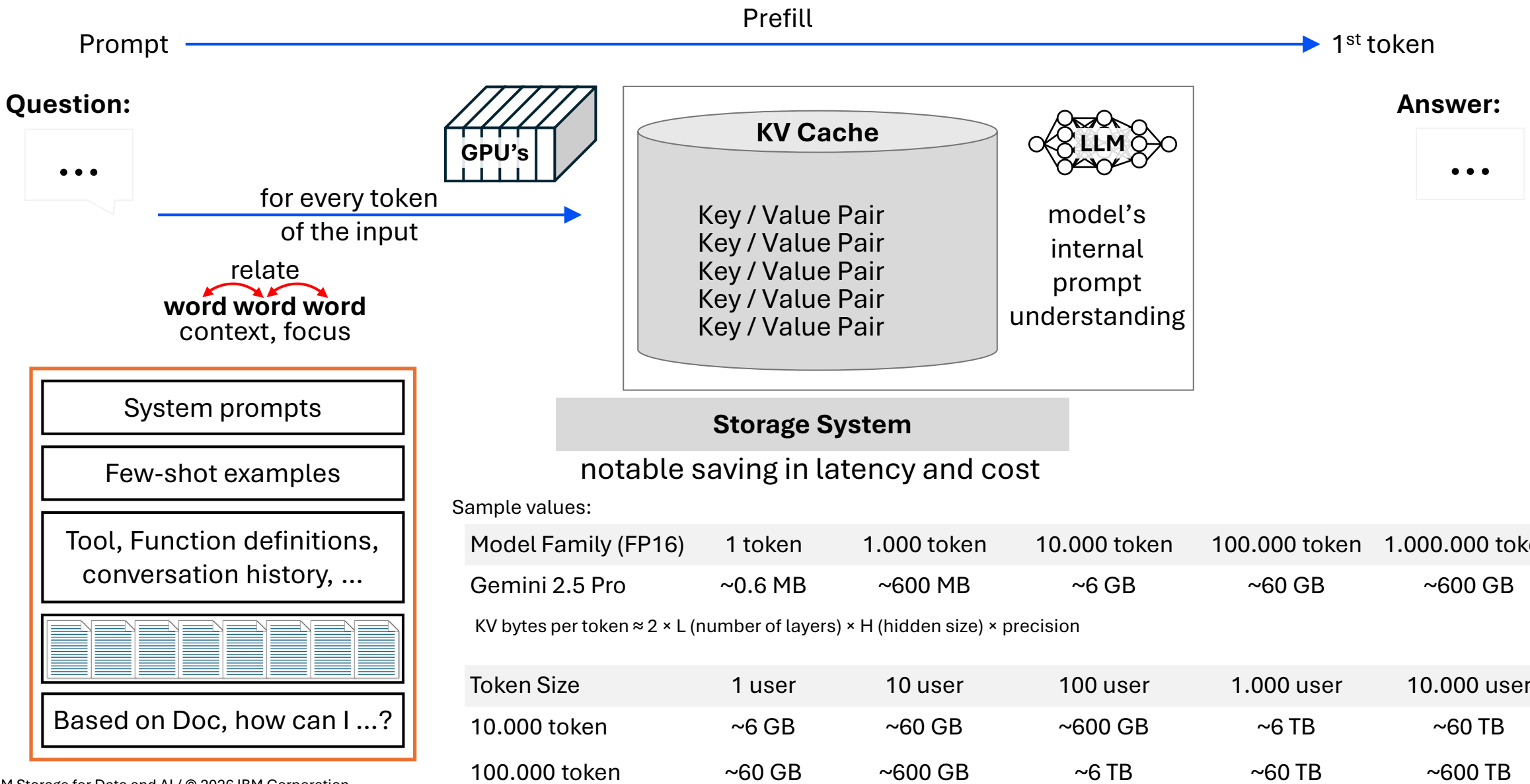
Time to first Token (Prompt caching example)



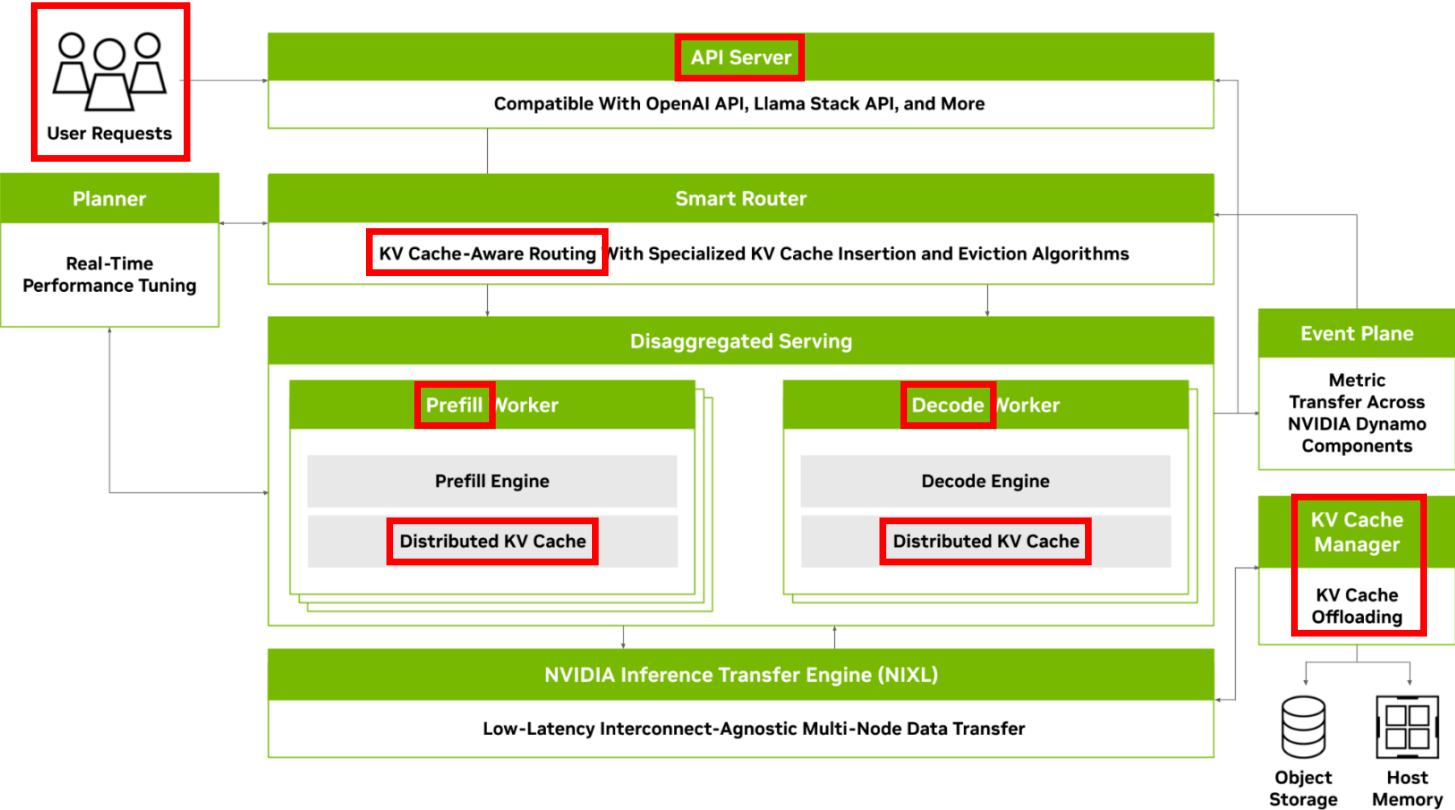
Time to first Token (Prompt caching example)



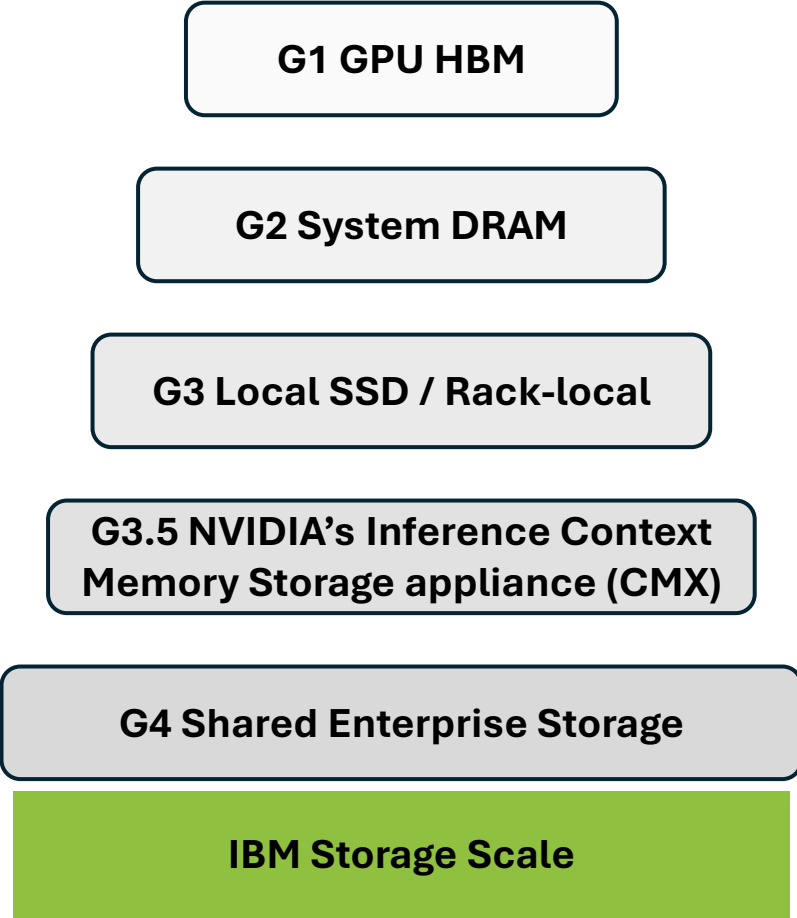
Time to first Token (Prompt caching example)



KV Cache < > Storage Management



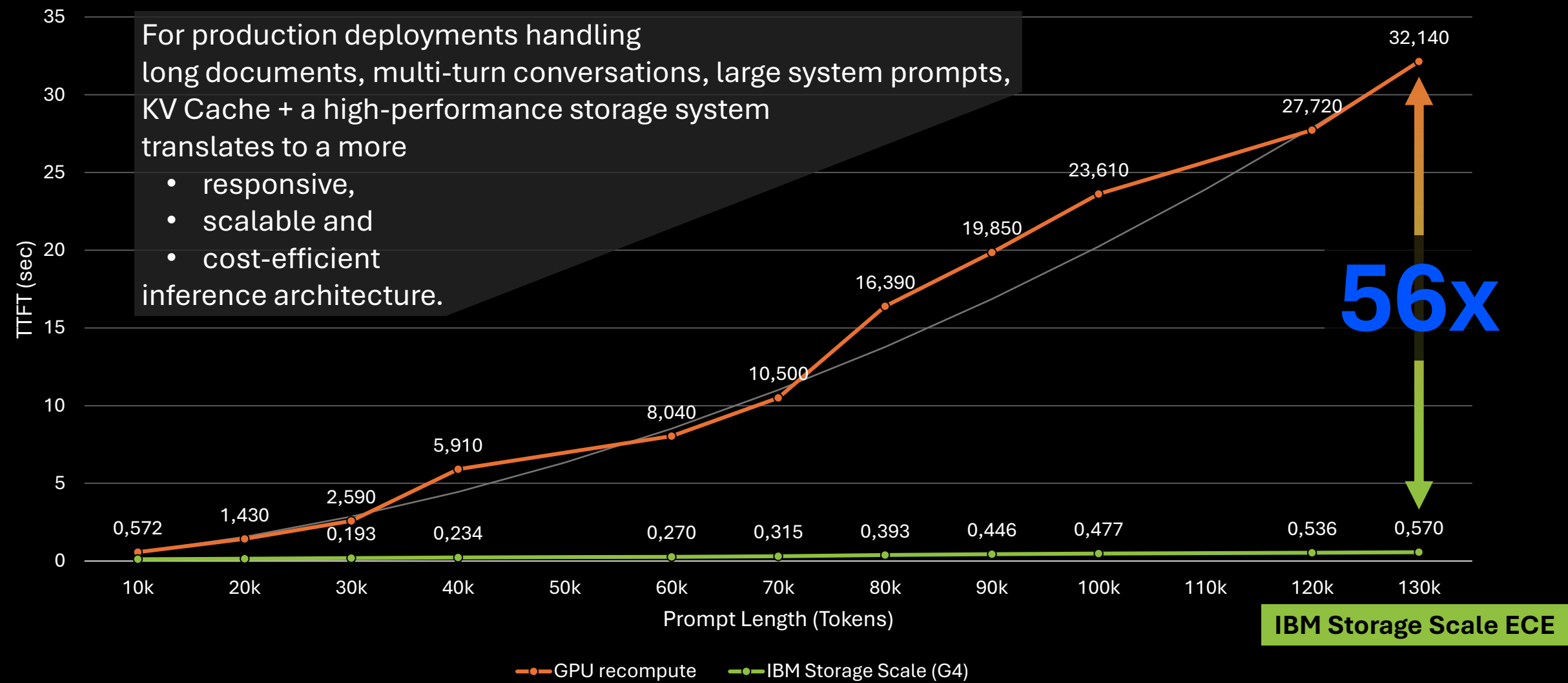
NVIDIA KV cache memory tier model



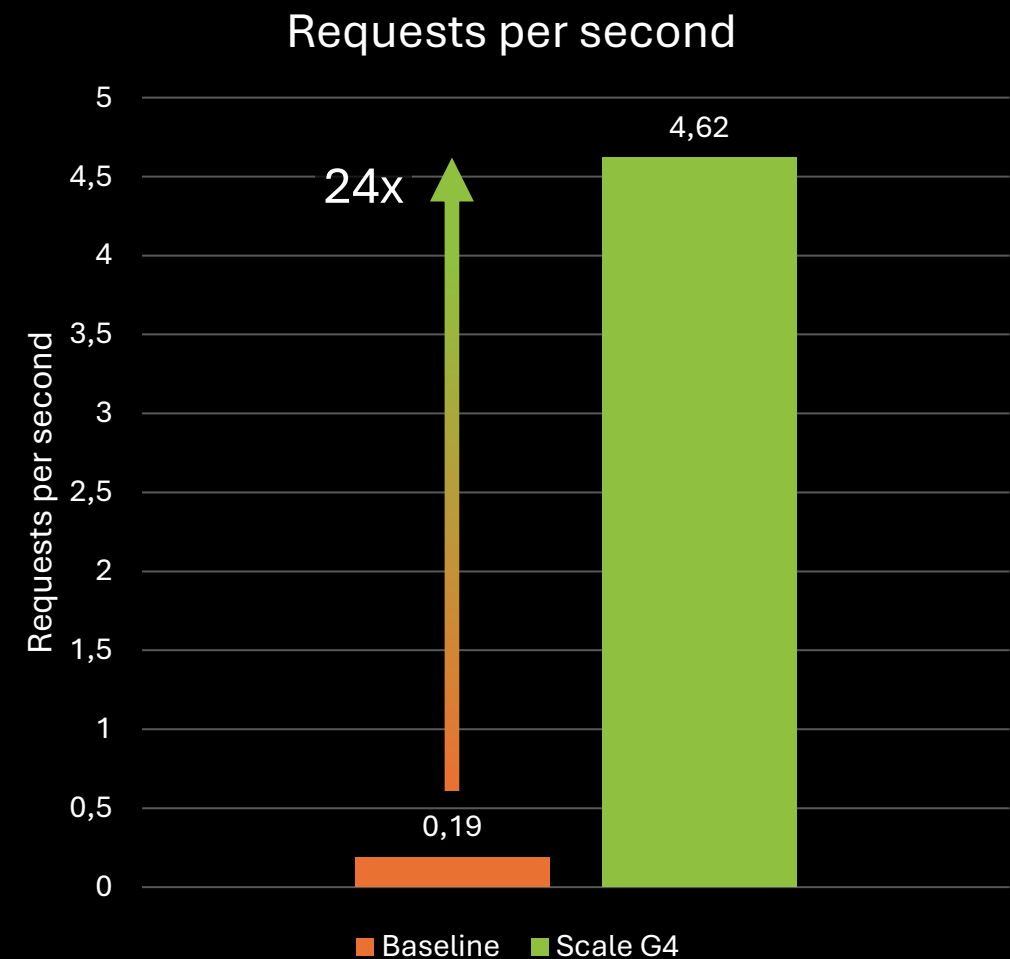
NVIDIA Dynamo for intelligent distributed KV cache Management

Benchmark 1: TTFT from a single GPU

Recompute vs KV Cache on IBM Storage Scale ECE (G4)

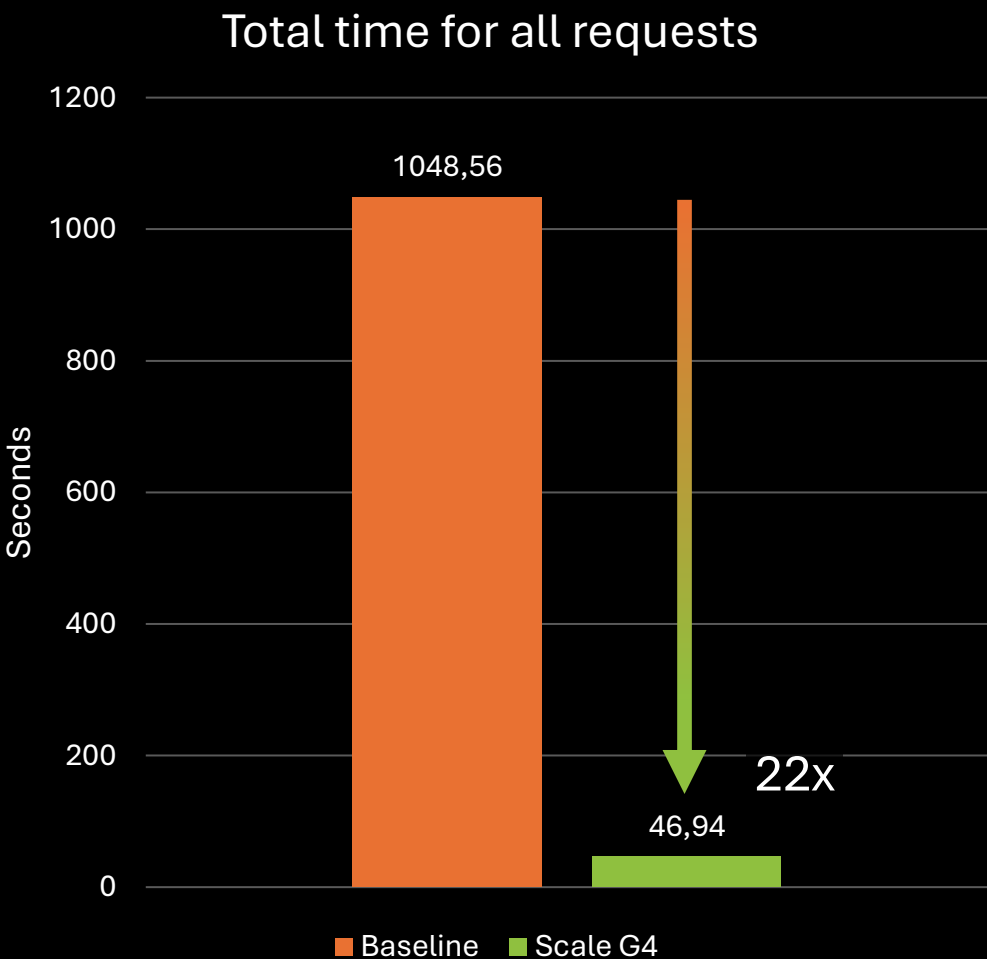


Benchmark 2: Throughput and response time efficiency



- 200 requests
- 120B model
- 28 concurrent connections

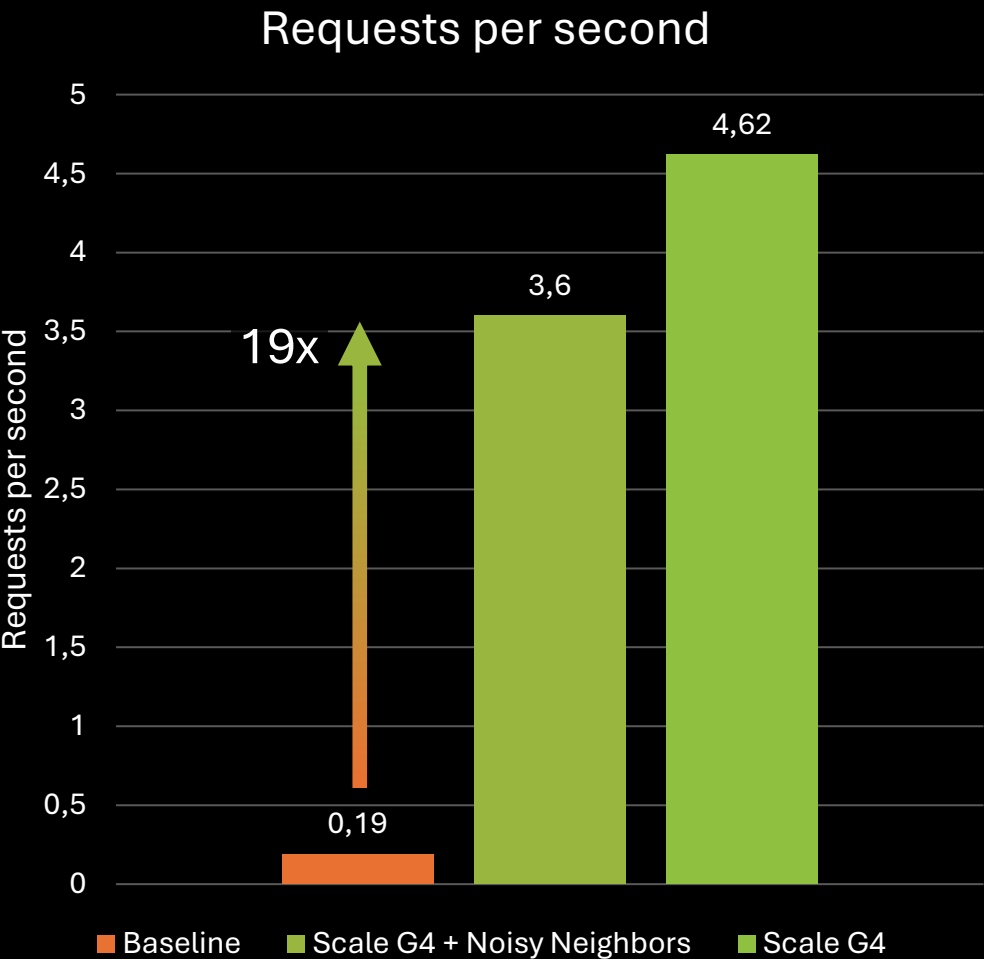
22x
more
efficient



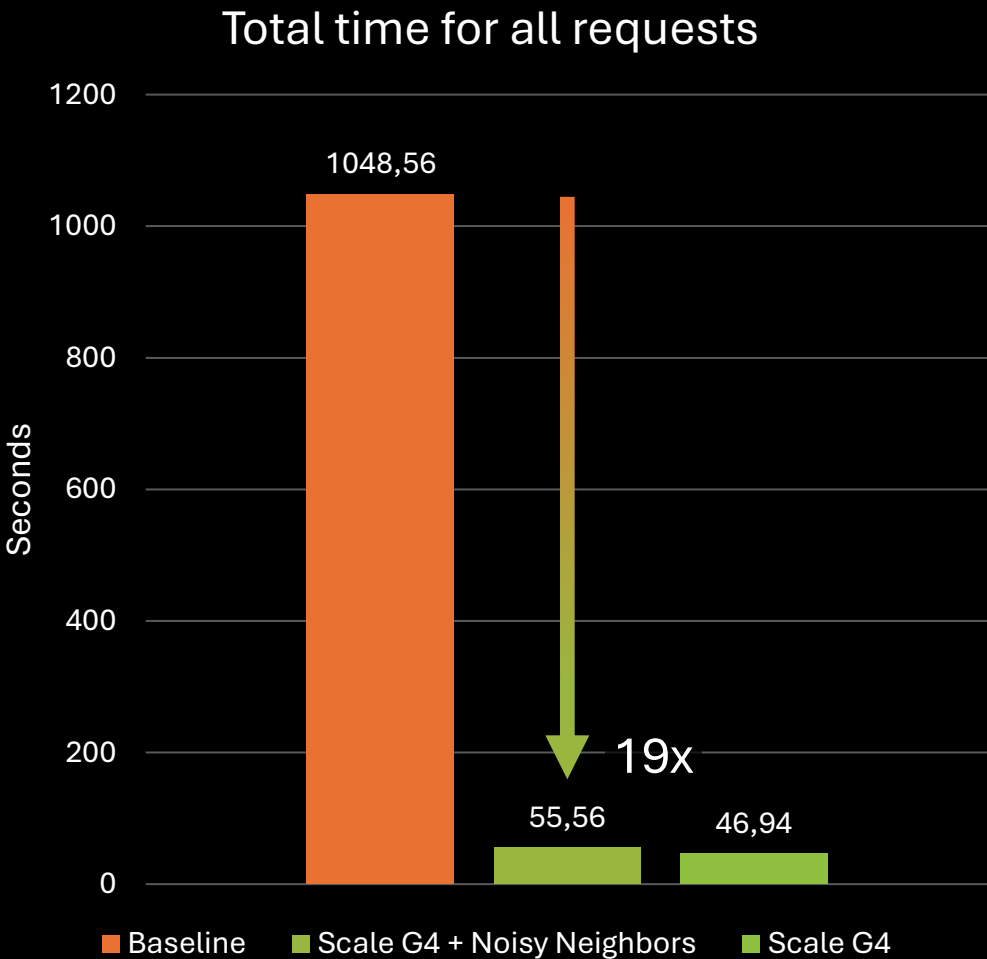
- rand. 100 unique prompts
= 24 million tokens
= 825 GB of KV cache
G4 cache started empty

Benchmark 3: Noisy neighbor test

Simulate shared, multi-tenant environment:
four concurrent clients continuously generating
200 GB per second of network I/O traffic



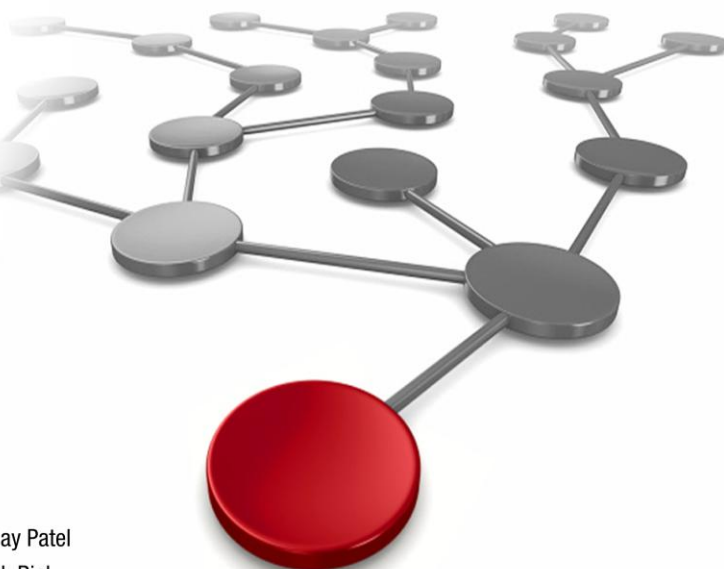
18x
more
efficient



- 200 requests
- 120B model
- 28 concurrent connections
- rand. 100 unique prompts
= 24 million tokens
= 825 GB of KV cache
G4 cache started empty

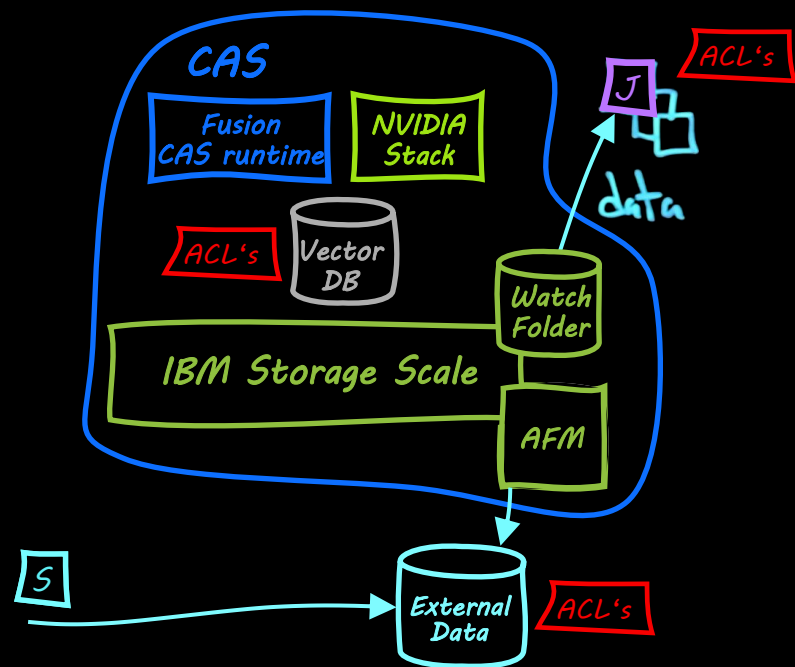
Context Without Limits: A High-Performance KV Cache Platform for Large-Scale AI Inference

Ka Wai Leung
Patrick Chiu
Hrishikesh Gaikwad
Ronan How
Anthony Hsu
Shaunak Karmarkar
Nikhil Khandelwal
Chen Li
William Li
Sherry Lin
Allen Liu
Kumar Mahesh
Guy Margalit
Jaya Mondal
Khanh Ngo
Tanmay Patel
Patrick Riel
Jeff Shao
James Tau



<https://www.redbooks.ibm.com/docs/MD260021/MD260021.html>

AI enabled Content Aware Storage based on IBM Storage Scale and IBM Fusion CAS



- Semantic interface to your data
- Fully automated document management
- Multi-tenancy, security

Content Aware Storage

100

BILLION Vectors

QUERY LATENCY

<700ms

RECALL PRECISION

90%+

Single 153 TiB
IBM Storage Scale System 6000



Vector Database

56x

Faster time-to-first-token

22x

Higher throughput

130k tokens

Flat latency at this context length

IBM Storage Scale

Context without limits

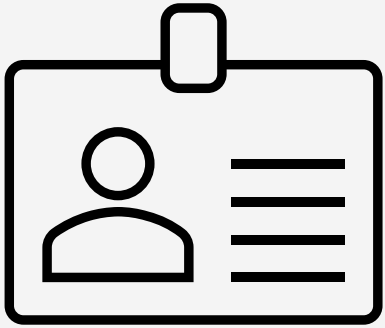


KV Cache

Agentic

- Do you use AI agents or autonomous workflows in your systems?
- Are you planning to use AI agents or autonomous workflows in your systems?
- Which use cases would you like to see with Scale for either prompt-driven inference, or also agent driven actions?
 - I would like to config x to be y
 - I need a filesystem a with size b create it
 - I need a filesystem a with size b. What are commands to create it?
 - I see an issue with... Help me understand where I should start checking...

Thank you



Ted Hoover (IBM)

<mailto:hoov@us.ibm.com>

Simon Lorenz (IBM)

<mailto:simon.lorenz@de.ibm.com>



