SAMSUNG

FMS the Future of Memory and Storage

RAG pipeline optimization leveraging HC SSD and CXL memory

Name

Adam Manzanares, Hui Qi, Arun George – GOST, NAND AE
Samsung Semiconductor

AI/ML is Changing The World

- Infrastructure spending on the rise[1]
- Data set sizes increasing rapidly[2]
  - Especially LLM
- GEN AI ROI
  - IT Support, Data Analysis, Generating Content, Code[3]

[1] The cost of compute power: A $7 trillion race | McKinsey
[2] Scaling up: how increasing inputs has made artificial intelligence more capable - Our World in Data
[3] https://blog.purestorage.com/perspectives/how-youll-use-generative-ai/

What is a RAG Pipeline

- Core components
  - Data
  - Model
  - Embeddings
  - Query

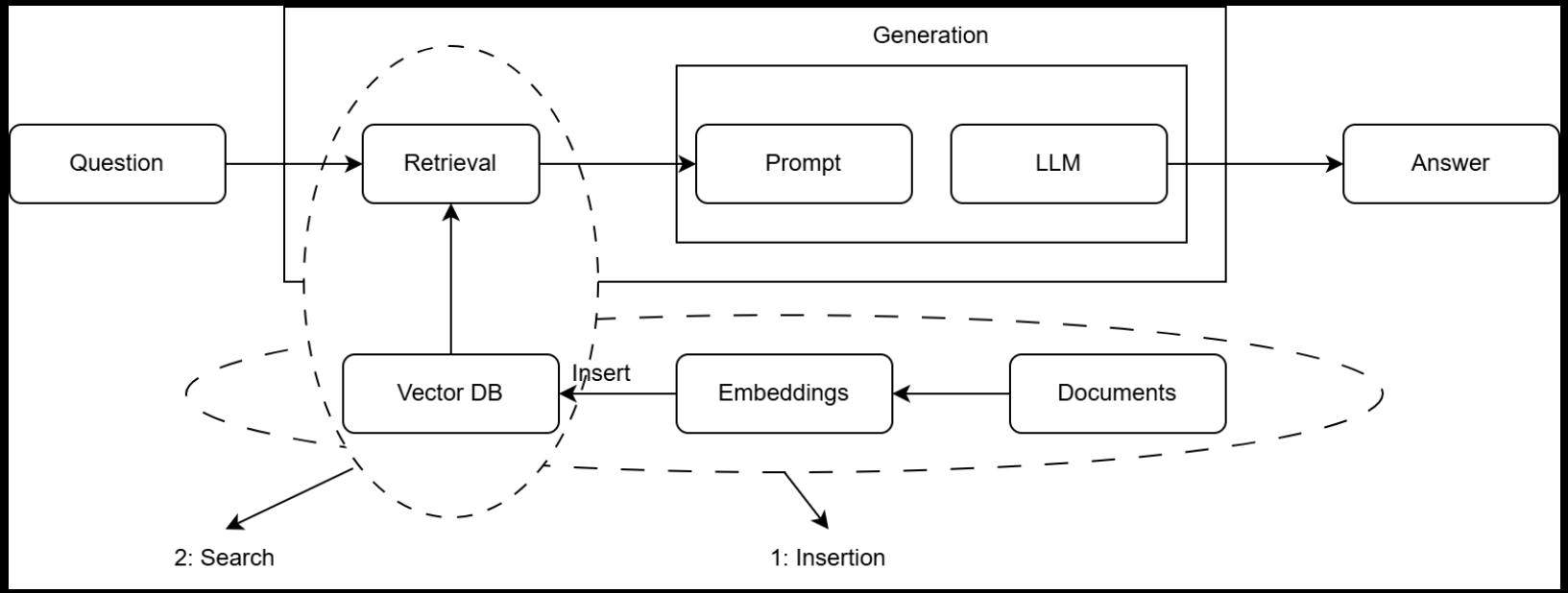


Figure based on[1]

[1] Accelerating Data Retrieval in Retrieval Augmentation Generation (RAG) Pipelines using CXL - MemVerge

RAG Pipeline Storage Demand

- Training data for model
- Training checkpoints
- Database storage for embeddings
- Query/result storage

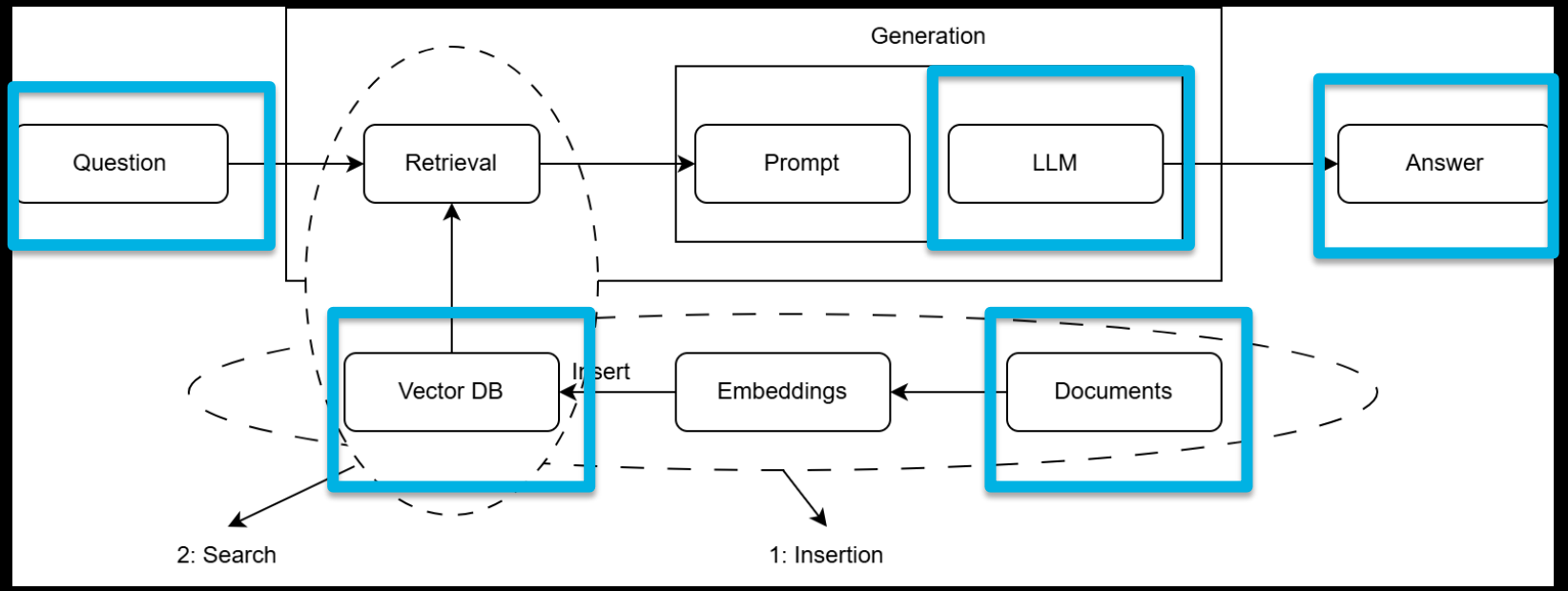
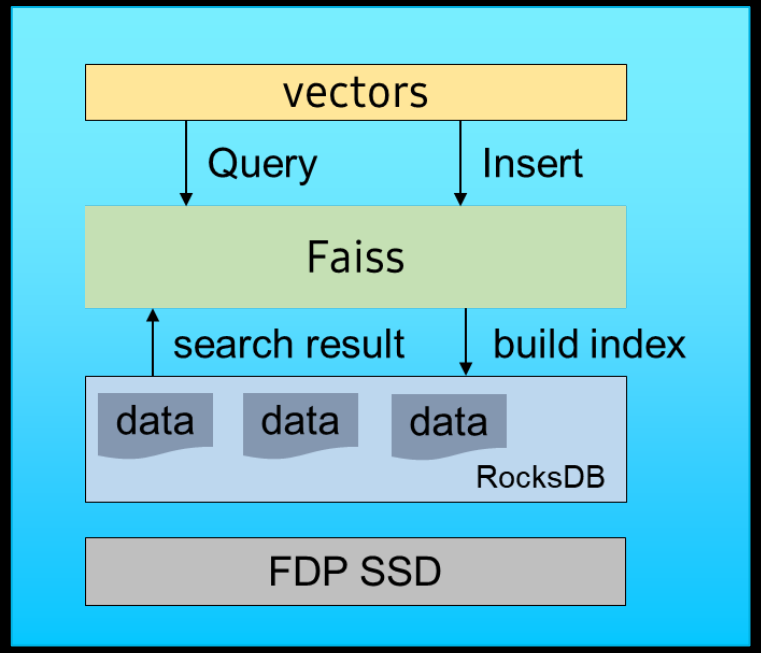
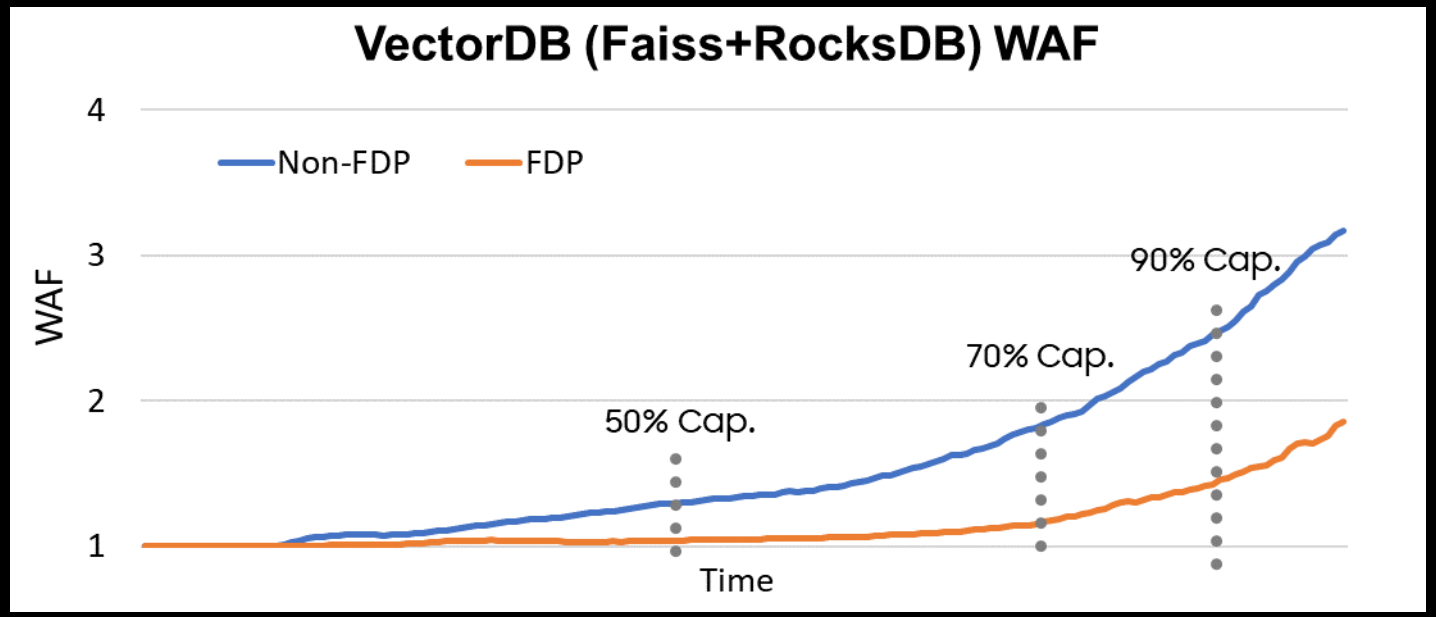


Figure based on[1]

[1] Accelerating Data Retrieval in Retrieval Augmentation Generation (RAG) Pipelines using CXL - MemVerge

RAG Pipeline Storage Optimization

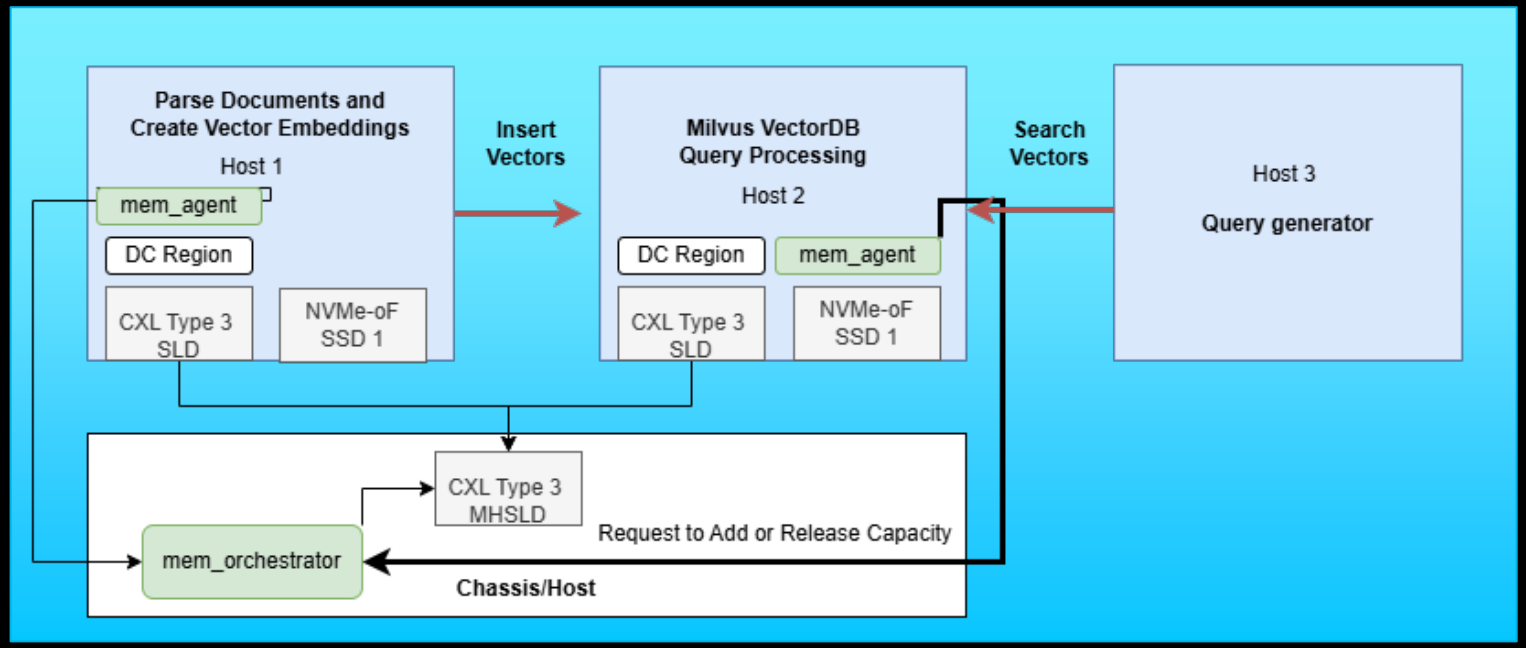




- FAISS supports similarity search and clustering of vectors
  - Used as vector DB (Langchain)[1]
- FAISS interface to storage can be RocksDB
  - Samsung has extensive RocksDB development experience
- FDP integrated into RocksDB lowers device measured WAF
  - FDP has improved device utilization for hyperscale applications

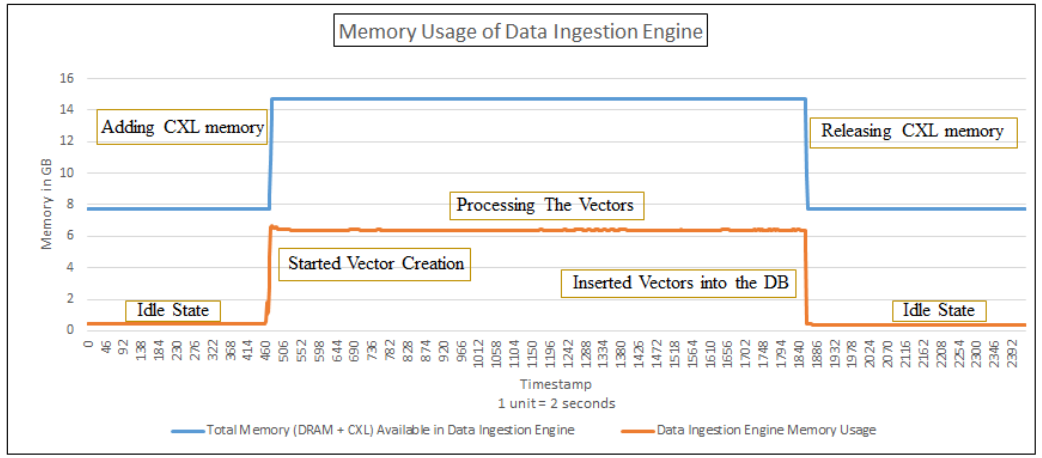
https://python.langchain.com/docs/integrations/vectorstores/faiss/

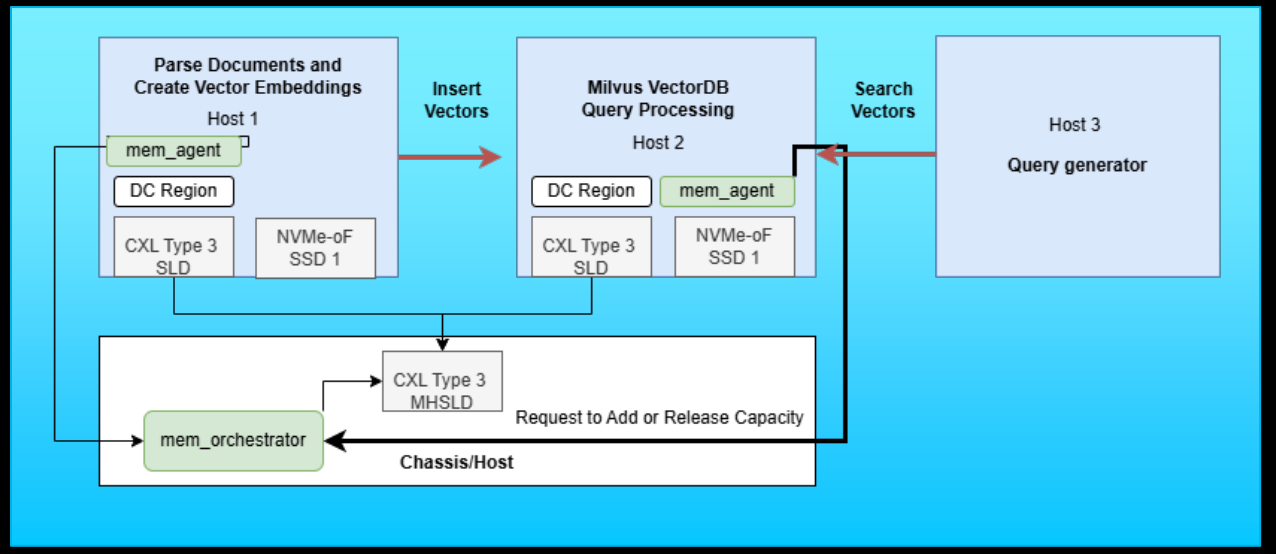
RAG Pipeline Memory Demand



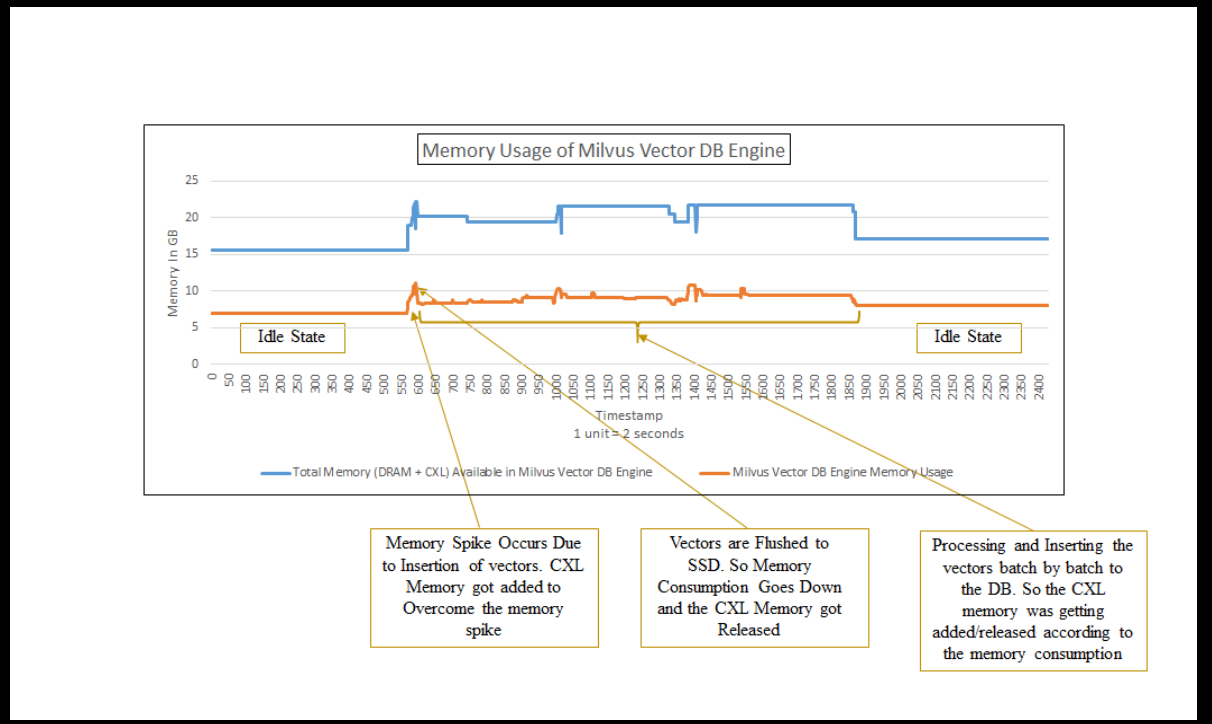
- Phased Approach
  - Generate Embeddings
    - Memory demand spikes
  - Running the pipeline
    - Based upon the app

RAG Pipeline Fabric Memory





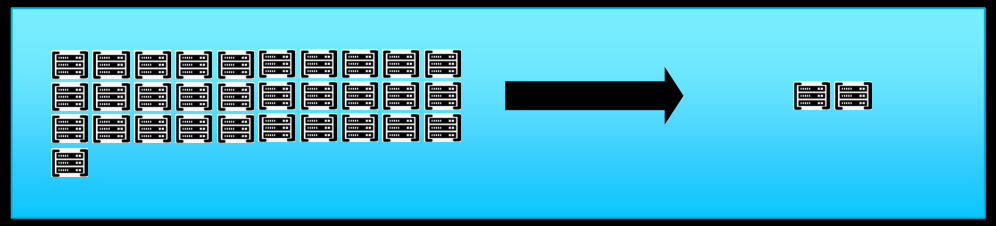
- CXL our example memory fabric
  - Capable of adding/remove host memory dynamically
    - Demo leveraging multi-head CXL device
- Standards and software useable today



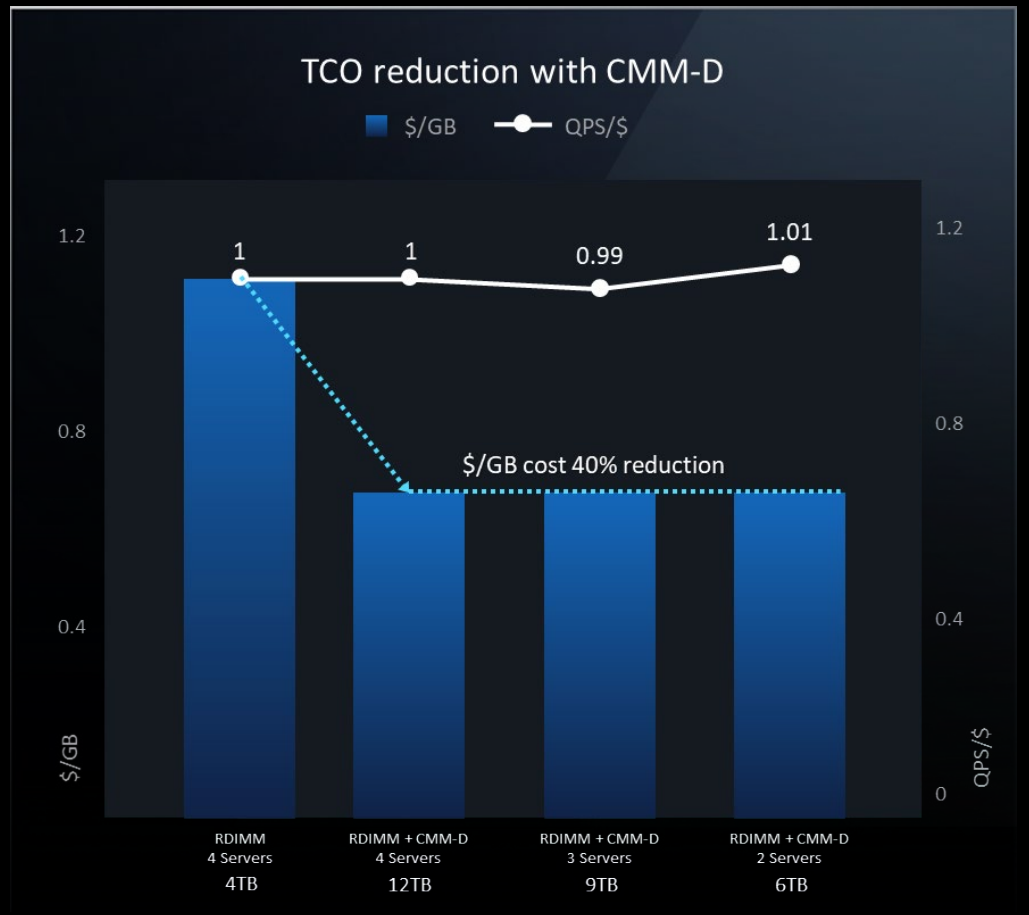
RAG Pipeline Potential TCO Gains

- Larger SSDs = Less Racks
  - Assumptions
    - Workload Scales
    - Power is available

Component	Low Capacity	High Capacity
SSD Capacity (TB)	7.68	128TB
Number of drives/server	24	24
Capacity/server (TB)	184.32	3072
Number of servers	543	33
Server/rack*	18	18
Number of racks	**31**	**2**



- Memory Utilization Increase
  - View memory utilization as a rack level problem
  - Not limited to local DIMMs



Key Takeaways

- Storage and Memory crucial to RAG
  - Long context on the horizon
- Samsung key producer of storage and memory
  - Increasing Ecosystem and Application Expertise
- Solve AI challenges together
  - Leverage our hardware/software expertise
  - Apply it to your view of the AI world

Thank You

SAMSUNG