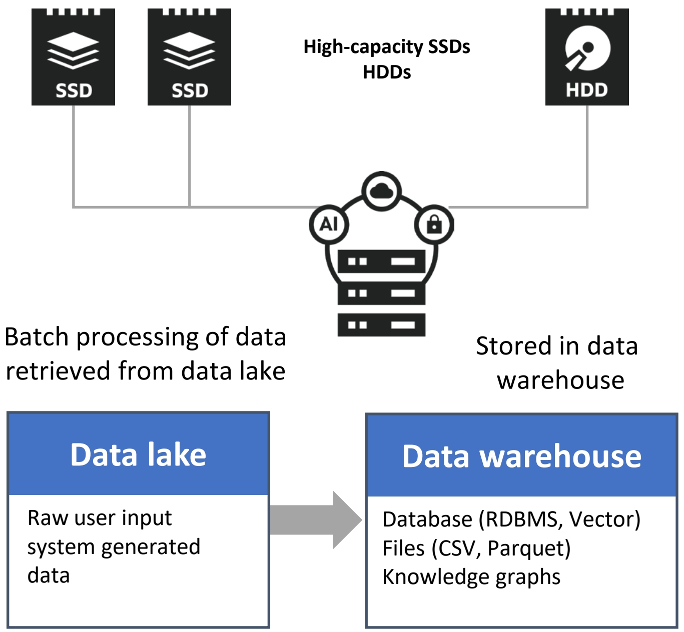
Flash Controller for the AI era

Vasanthi Jagatha, Senior Manager, Marvell

the **Future** *of* **Memory** *and* **Storage**

AI workloads using storage

Data processing



High-capacity SSDs
HDDs

Batch processing of data retrieved from data lake

Stored in data warehouse

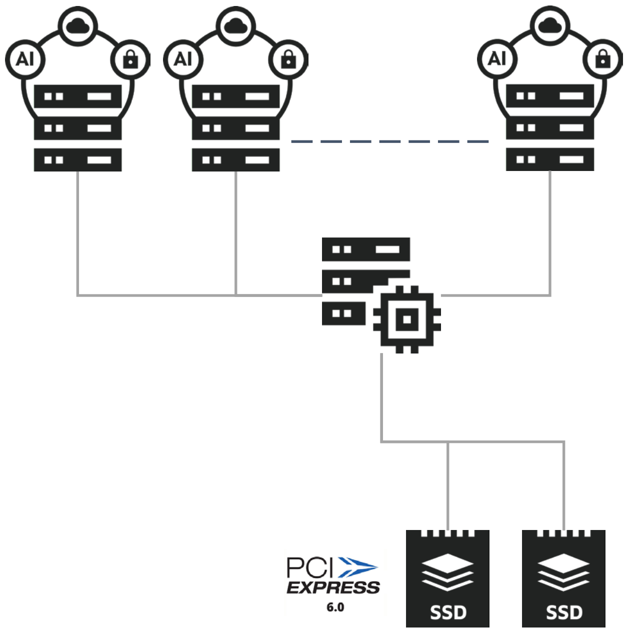
Data lake

Raw user input
system generated
data

Data warehouse

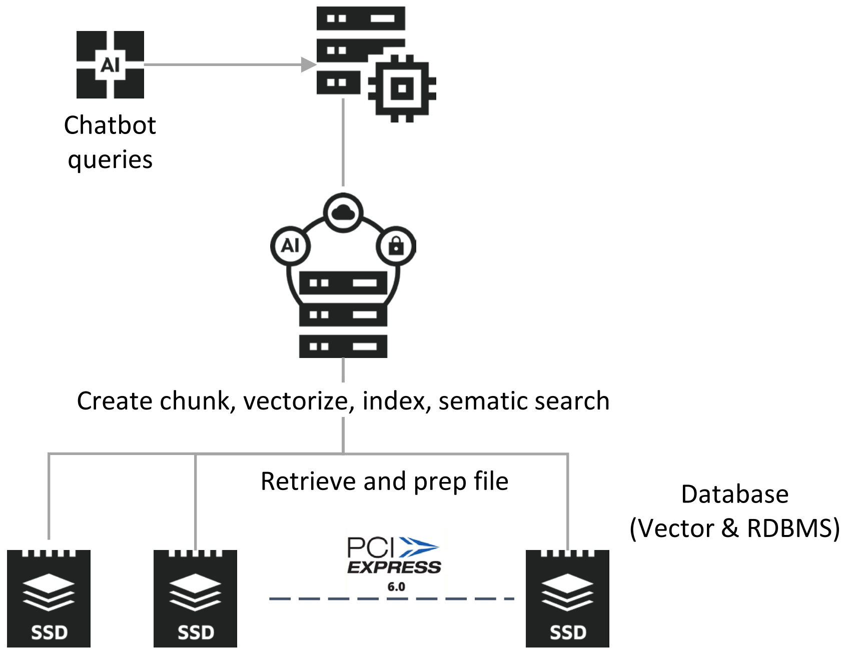
Database (RDBMS, Vector)
Files (CSV, Parquet)
Knowledge graphs

Training



Inference

Retrieval augment generation (RAG)



Chatbot queries

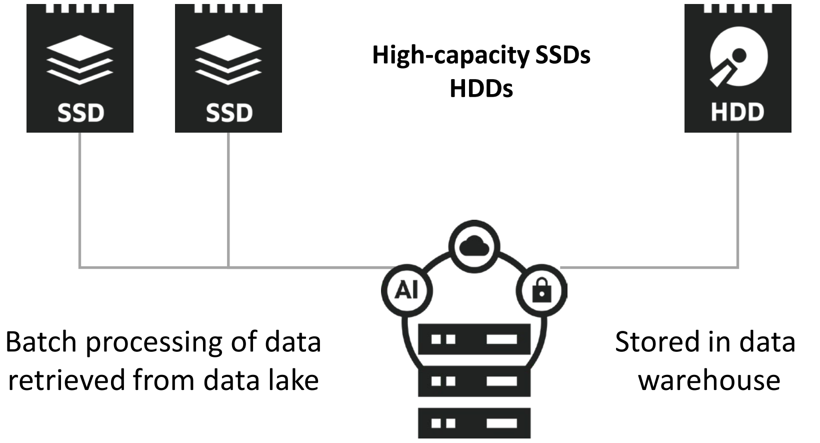
Create chunk, vectorize, index, sematic search

Retrieve and prep file

Database (Vector & RDBMS)

Storage critical across data processing, training and inference workloads

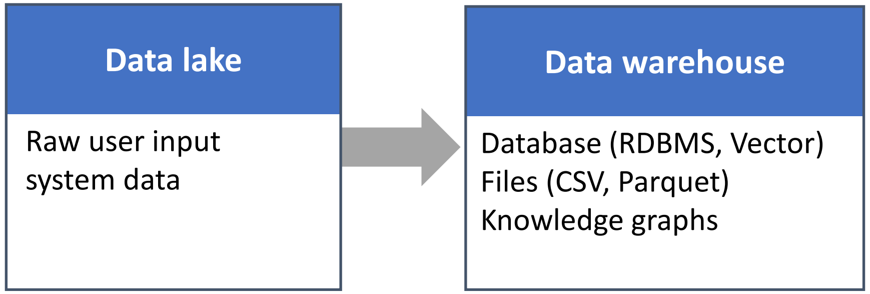
Data processing



High-capacity SSDs
HDDs

Batch processing of data retrieved from data lake

Stored in data warehouse



Data lake

Raw user input
system data

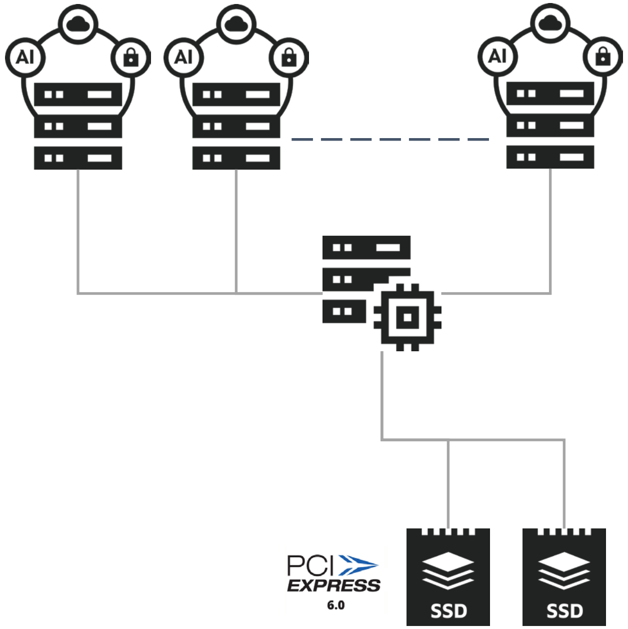
Data warehouse

Database (RDBMS, Vector)
Files (CSV, Parquet)
Knowledge graphs

- Dominates AI/ML development lifecycle
- Input data fidelity has outsize impact on resulting model performance
- Large data-sets typically stored in
  - Data lakes (unstructured)
  - Database, CSV, Parquet, JSON (structured)
- Bursty reads, write heavy workload

Key Storage considerations- capacity, data reliability, and performance

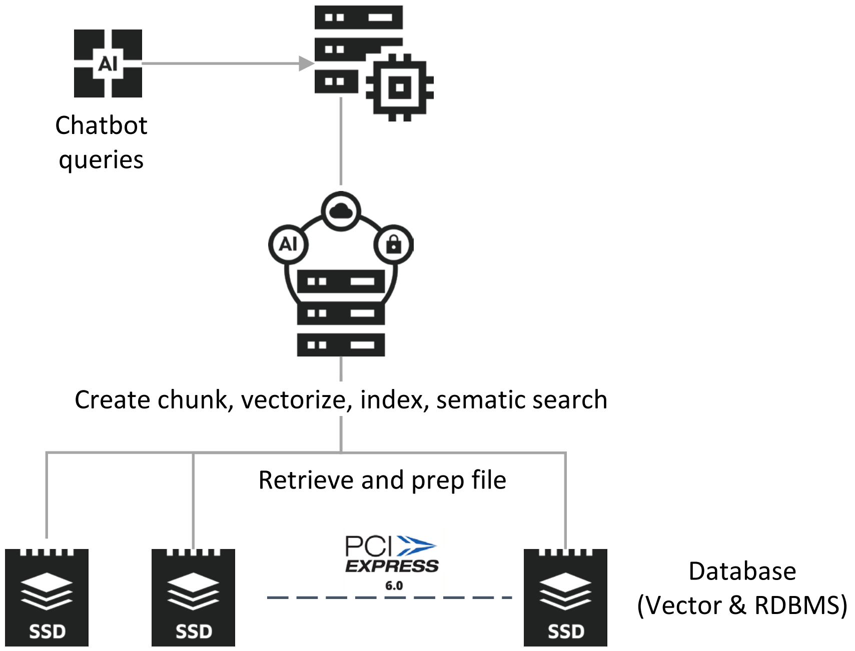
Training



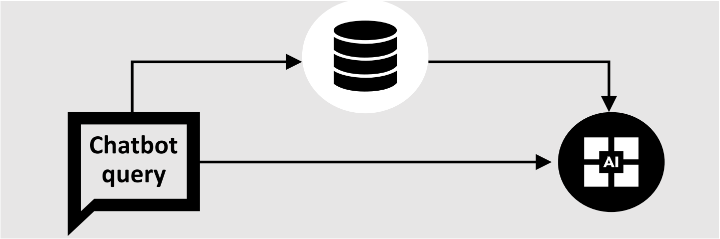
- Fast data retrieval -Reduce GPU idling
- Regular checkpointing for observability
- Mixed reads and bursty writes

High performance fast data retrieval and checkpointing

Inference – Retrieval Augmented Generation (RAG)



AI
Chatbot queries
AI
Create chunk, vectorize, index, sematic search
Retrieve and prep file
SSD
SSD
PCI EXPRESS 6.0
SSD
Database (Vector & RDBMS)

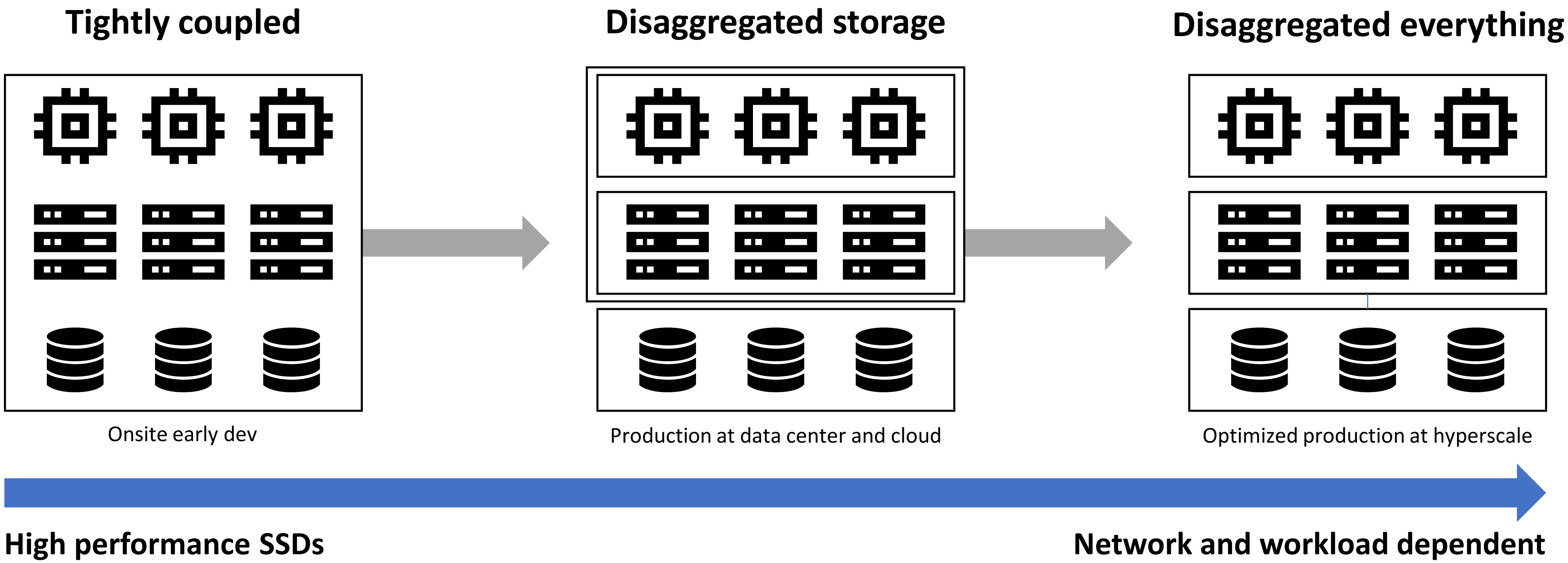


RAG enhances LLMs by integrating external data retrieval

- I/O Intensive (data prep, high user traffic)
- Large capacity vector storage

High performance, large capacity for large scale RAG

AI infrastructure



Tightly coupled

Onsite early dev

Disaggregated storage

Production at data center and cloud

Disaggregated everything

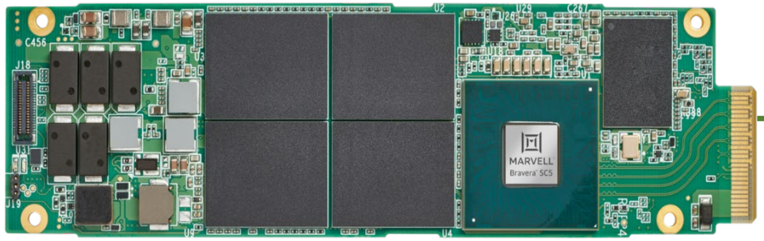
Optimized production at hyperscale

High performance SSDs

Network and workload dependent

AI infrastructure needs evolves with AI/ML development phase - impacts storage requirements

Flash controller design



AI/ML workload
Fast and secure data movement and response times

System Infrastructure
Distributed and optimized infrastructure

NAND
speed, capacity, power and reliability

Flash controller features

Protocols

I/F speeds

PCIe lanes and NAND channels

Data management

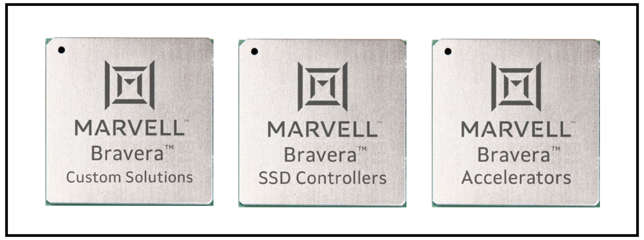
NAND support

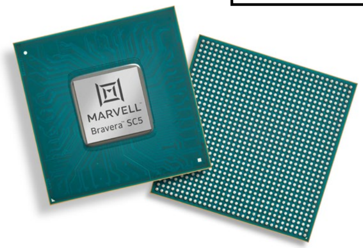
Compute offload

HW acceleration

Flash controller requirements driven by AI/ML workloads, system, NAND decisions

March to Gen6 SSDs





3.2GBps
NAND 800MT/s,256Gb
16TB SSD
LDPC

PCIe 3.0

7.4GBps
NAND 1600MT/s,512Gb
64TB SSD
LDPC+

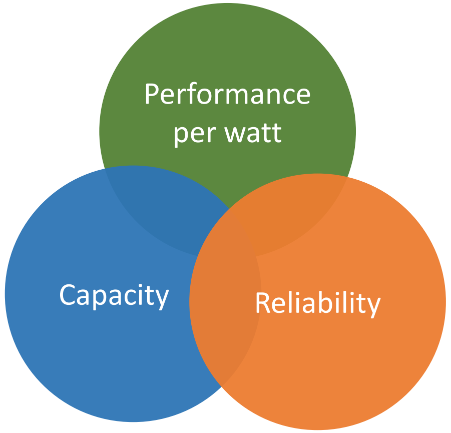
PCIe 4.0

14GBps
NAND 2400MT/s,512-1Tb
128TB SSD
LDPC++

PCIe 5.0

28GBps
NAND 4800MT/s,1-2Tb
256TB SSD
LDPC+++

PCIe 6.0

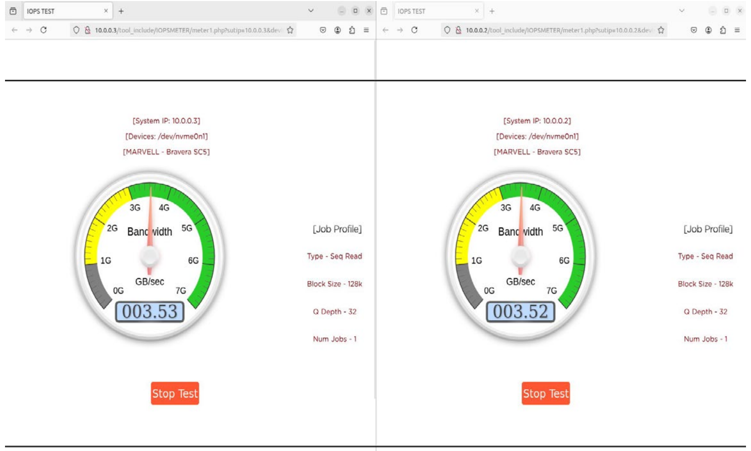


Critical Flash Controller AI workload needs

Gen6 closes the gap for AI workload needs

Visit Marvell booth # 1046

Dual Port Demo



10.0.0.3/tool_include/IOPSMETER/meter1.php?sutip=10.0.0.3&dev

[System IP: 10.0.0.3]
[Devices: /dev/nvme0n1]
[MARVELL - Bravera SC5]

[Job Profile]
Type - Seq Read
Block Size - 128k
Q Depth - 32
Num Jobs - 1

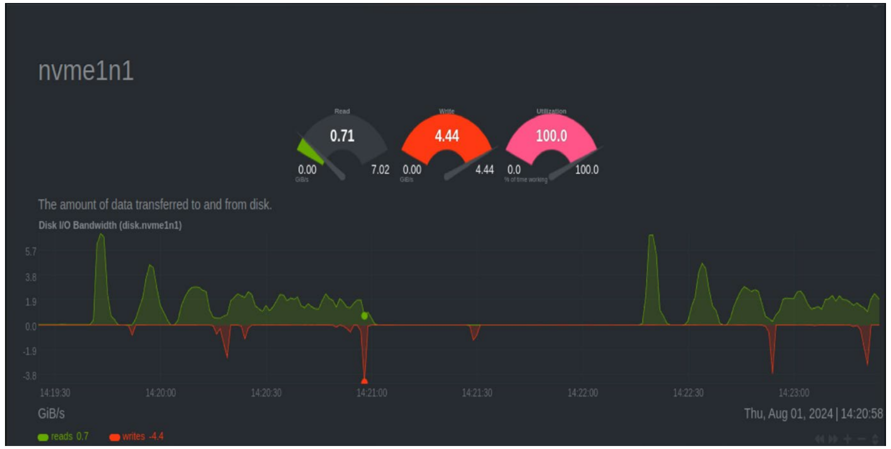
10.0.0.2/tool_include/IOPSMETER/meter1.php?sutip=10.0.0.2&dev

[System IP: 10.0.0.2]
[Devices: /dev/nvme0n1]
[MARVELL - Bravera SC5]

[Job Profile]
Type - Seq Read
Block Size - 128k
Q Depth - 32
Num Jobs - 1

MARVELL™

Accelerated Storage for RAG and AI Inference



nvme1n1

The amount of data transferred to and from disk.

Disk I/O Bandwidth (disk.nvme1n1)

Thu, Aug 01, 2024 | 14:20:58

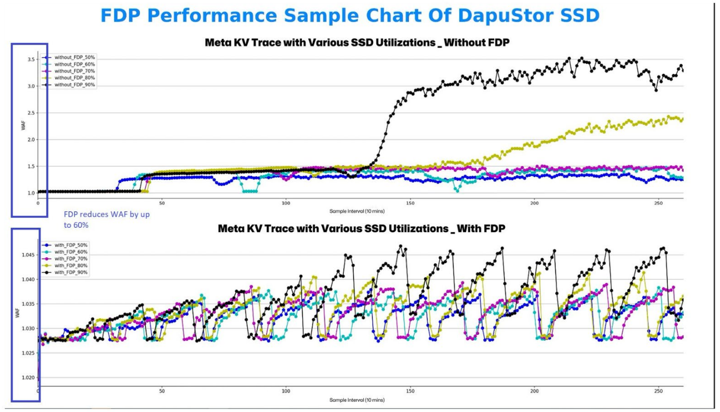
MARVELL™

Powered by

MEMBLAZE

FDP QLC with PCIe 5.0 NVMe™ SSD

Customer Demo



FDP reduces WAF by up to 60%

DapuStor