Storage Class Memory in Scalable Cognitive Systems

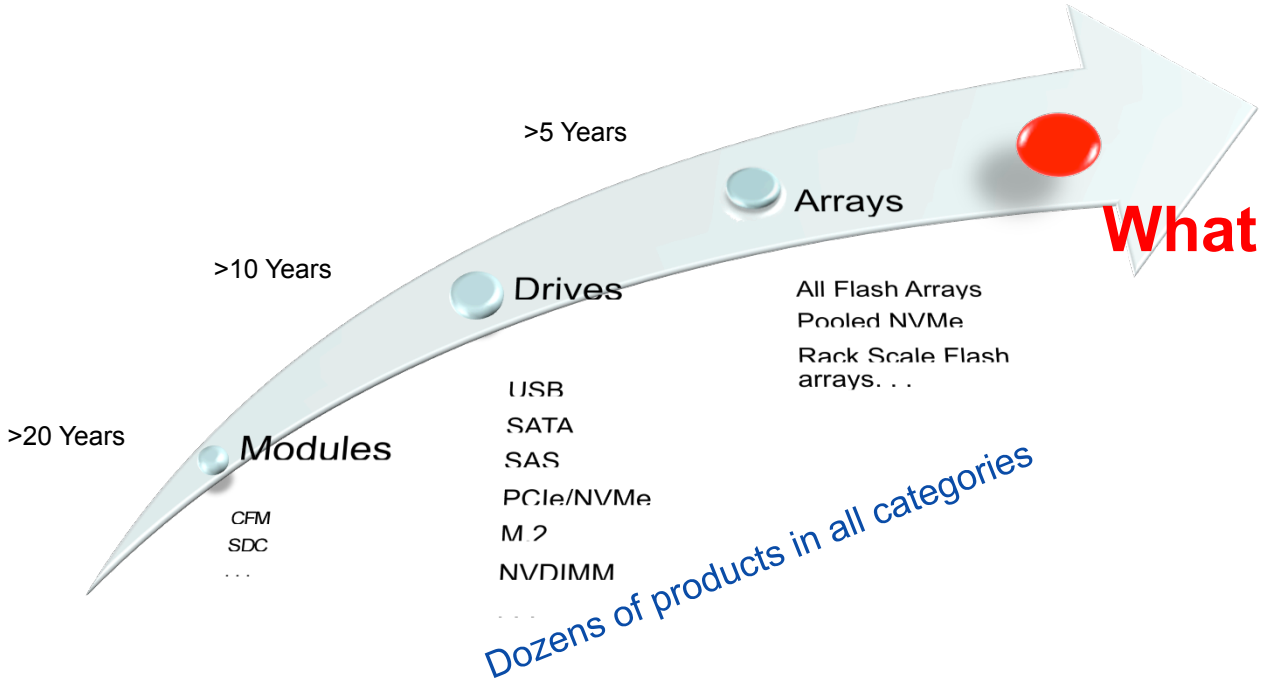
Balint Fleischer

Chief Research Officer

The impact of NVM on the Application/Data architecture

- Accelerated demanding applications
  - OLTP, Big Data, Etc.
- Changing scaling economics
  - Social Networks, Search, HPC
- Improved operational characteristics
  - Notebooks, tablets, Power efficiency
- Enabling new use cases/architectures
  - Hyperconvergent systems, Cold Storage, Streaming platforms,
    Data Virtualization systems, Etc.

NVM journey



What comes next?

Evolving IT focus

	Enterprise Automation	Online Transactions	Cognitive Computing
"Killer" use cases	OTLP ERP Email	eCommerce Messaging Social Networks Content Delivery	Discovery of solutions, capabilities Risk Assessment Improving customer experience Comprehending sensory data
Key functions	RDBMS BI Fraud detection	Databases Social Graphs SQL and ML Analytics Streaming	Natural Language Understanding Object Recognition Probabilistic Reasoning Content Analytics
Data Types	Structured Transactional	Structured Unstructured Transactional	Streaming Mixed Graphs, Matrices
Storage Types	Enterprise Scale Standards driven SAN/NAS, etc	Cloud Scale Open source File/Object	Application Scale Highly Optimized Intelligent

Cognitive Computing

Augmenting human expertise

&

Transforming human <-> Computer interaction

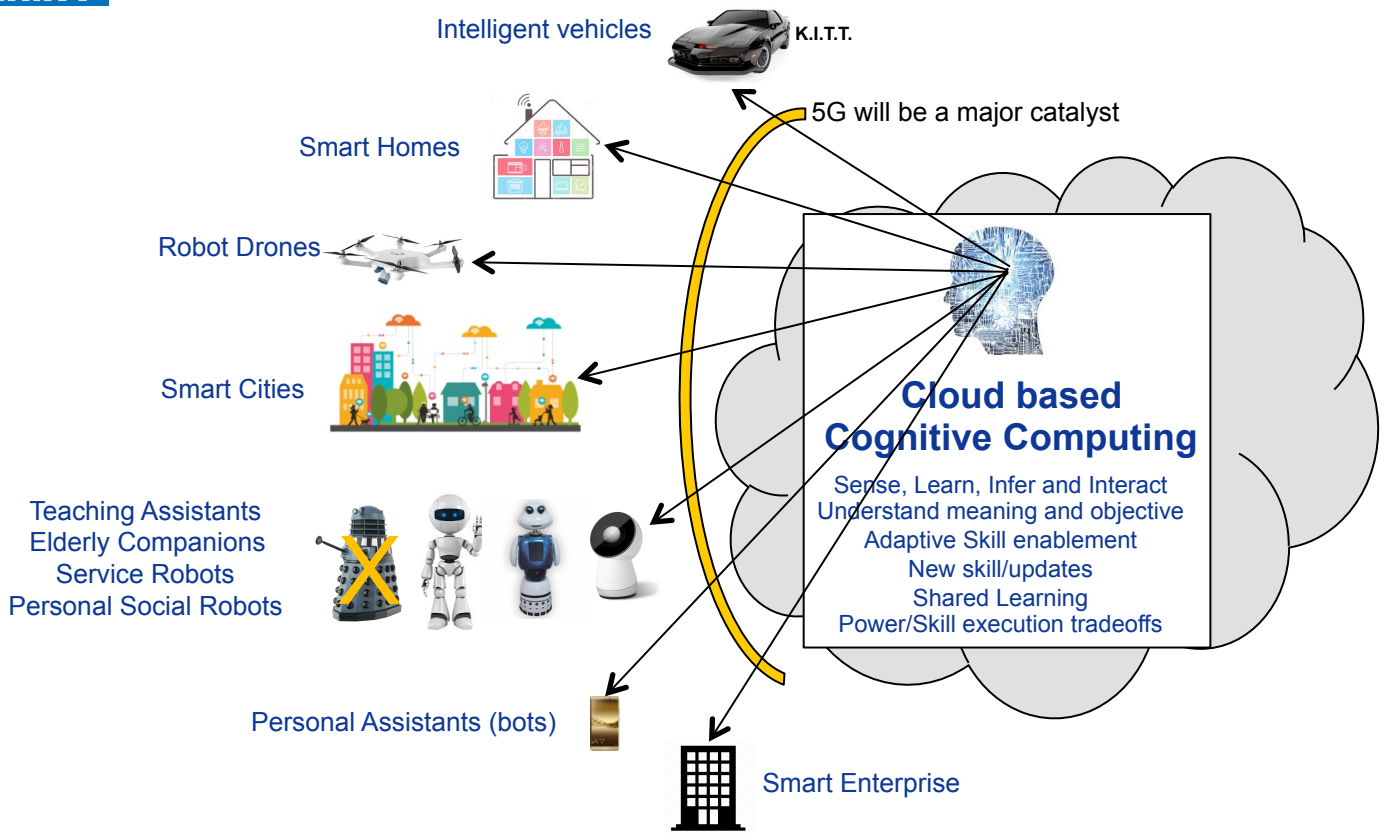
Business Benefits

- Identify connections between events, people and trends
- Discovery of new insights, uncover breakthroughs and predict trends through real time understanding of current and historical data
- Enabling new customer experience via service personalization
- Reinvention of business models and operations

Supporting Functionality

- Evolve with goals and respond to changes
- Participate in the shared discovery process and problem refinement iteration
- Understand meaning, goal, syntax, regulation, time, etc.
- Utilizes real time sensory and behavioral inputs as well as contextual data

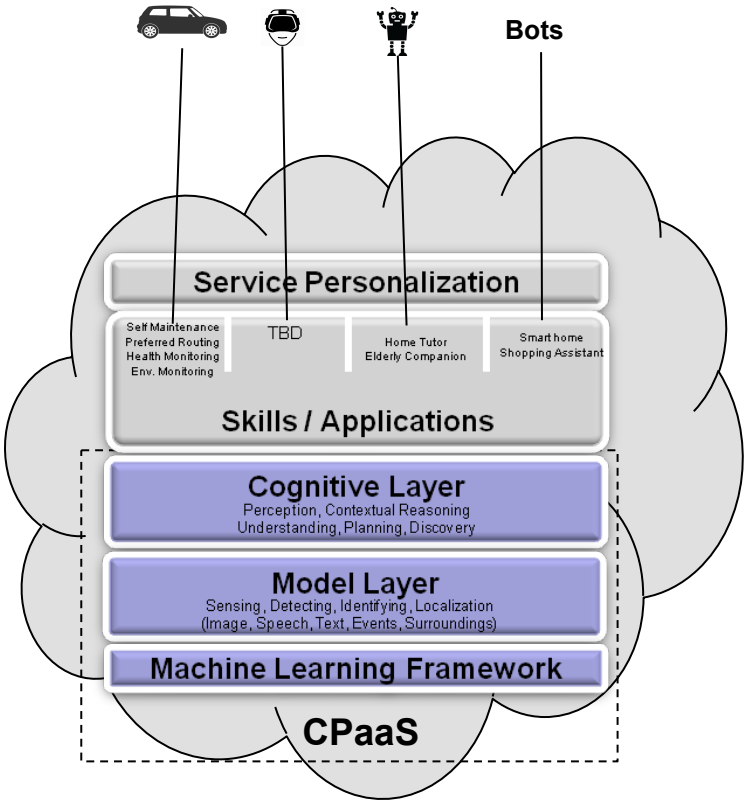
Cognitive Computing use cases



Cognitive Application Platform (CPaaS)

Overview

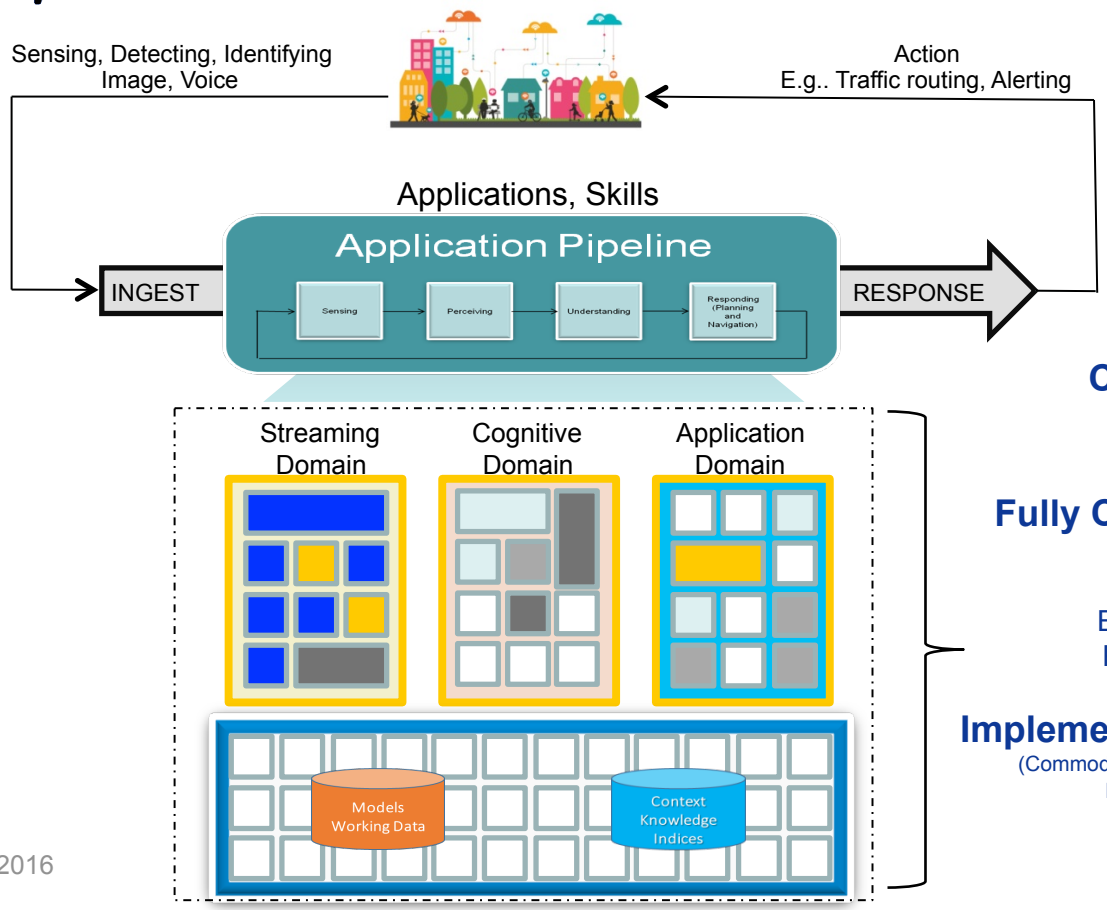
Complex SW stack
Executed as Dataflow
Optimized for fast response time
Highly scalable
Extensive use of Advanced Algorithms
Historical & Real Time Data processing
Proximity to data is key



Target Applications

Cloud based Assist Bots
Intelligence “booster” for Robots
Enhancing awareness context
Shared knowledge/learning
Robot Power optimization
Robot Cost reduction
Robot re targeting (SDR)
Others

Running Cognitive Applications



Cognitive Platform (CPaaS)

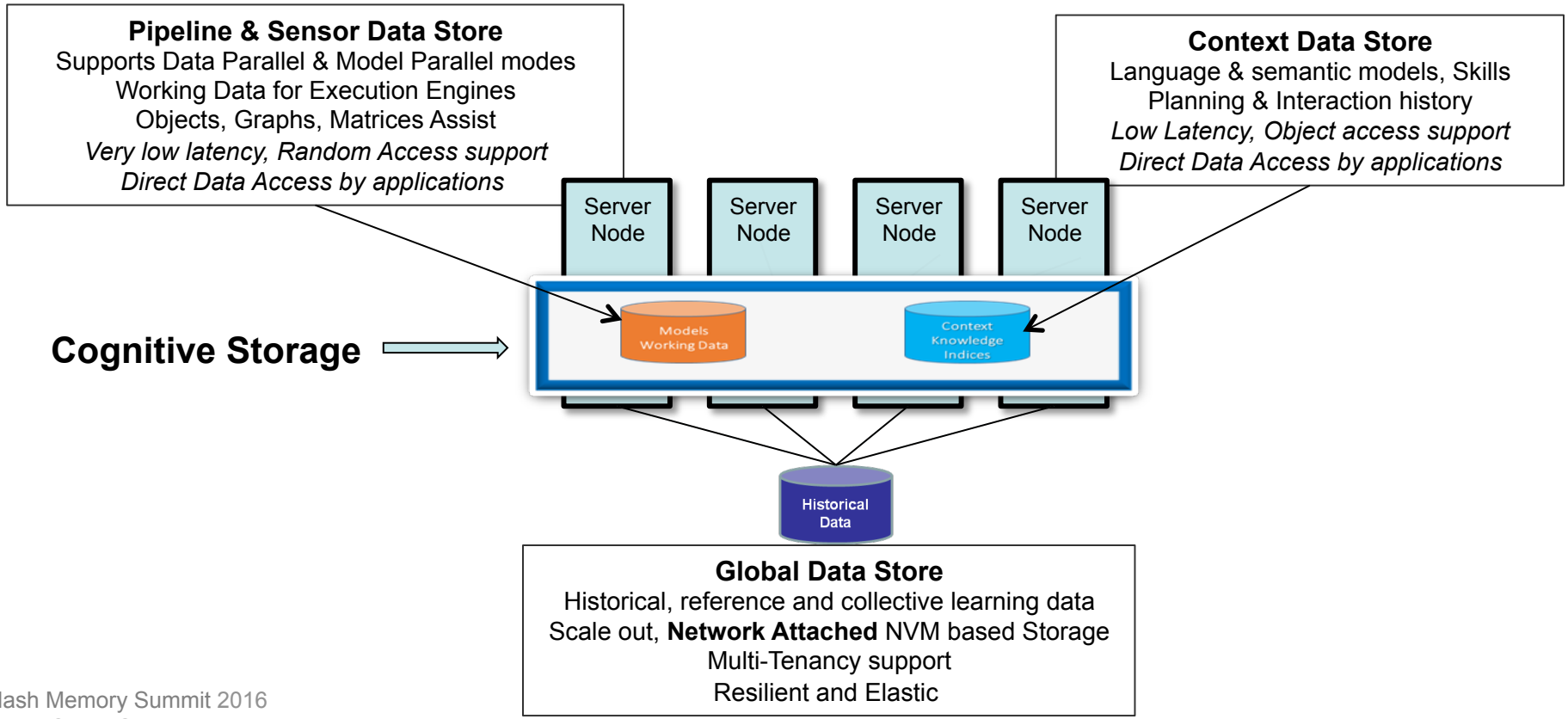
Fully Containerized architecture

Static Containers
Dynamic Containers
Ephemeral Data Containers
Persistent Data Containers

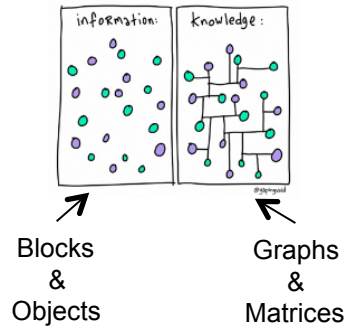
Implemented on a Scale out Cluster

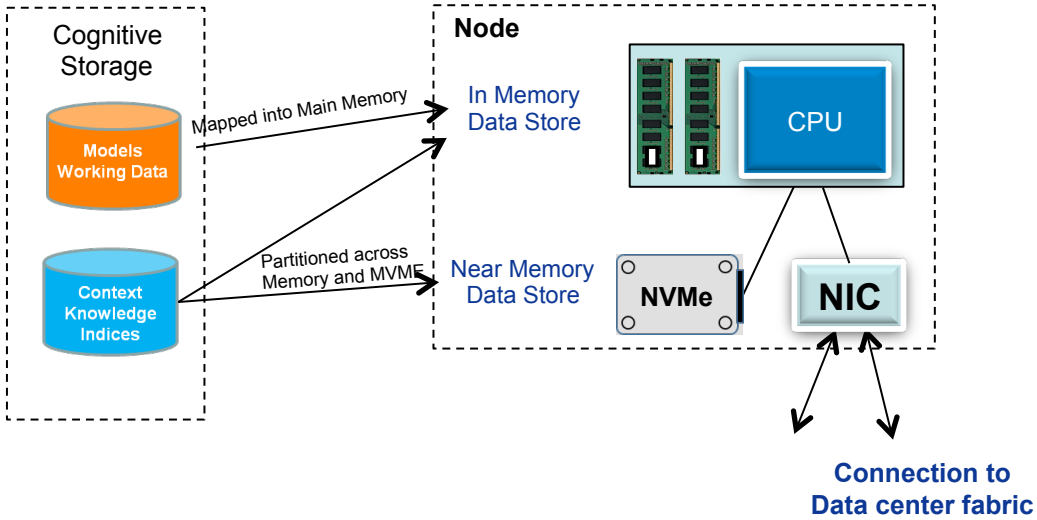
(Commodity Servers, Large Memory, Fast, Low latency Interconnect and Fast NVMe drives)

Optimized Storage for Cognitive Computing



Implementing Cognitive Storage



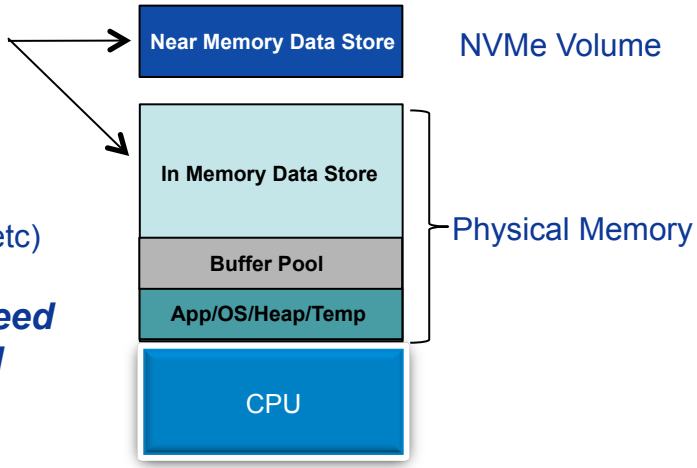


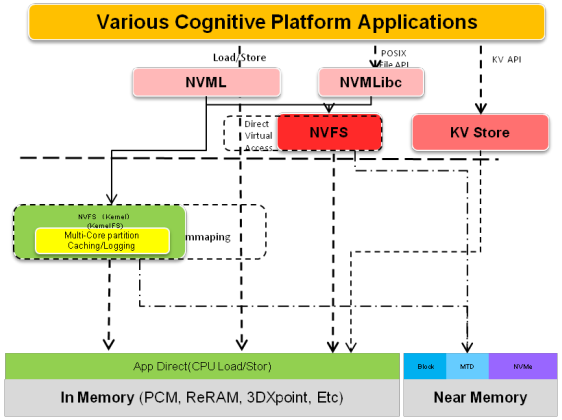
Partitioned into Node vs Cluster shared
Application Optimized (semantics consistent with data types)
Application Direct Accessed (use space IO)
Intelligent Data placement for scalability

Technology for Cognitive Storage

SCM is very promising
Almost as fast as DRAM
Higher capacity vs. DRAM
Envisioned to be less expensive
Read Performance biased
Selectable Attributes (Persistency, etc)

SCM latency highlights the need to further reduce overhead





Huawei SCM SW stack

"Memory Speed Data Store" (MSDS) would be better name to reflect new functionality

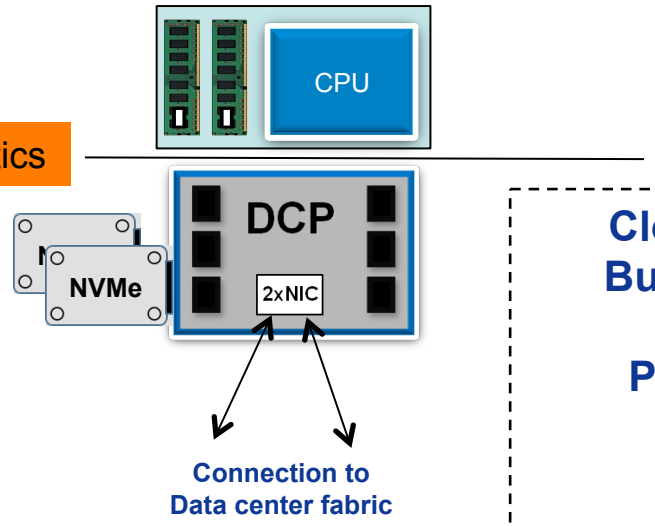
Optimizing for applications

Reducing IO Overhead

Application Processing

Application Specific IO semantics

IO Processing
&
Media



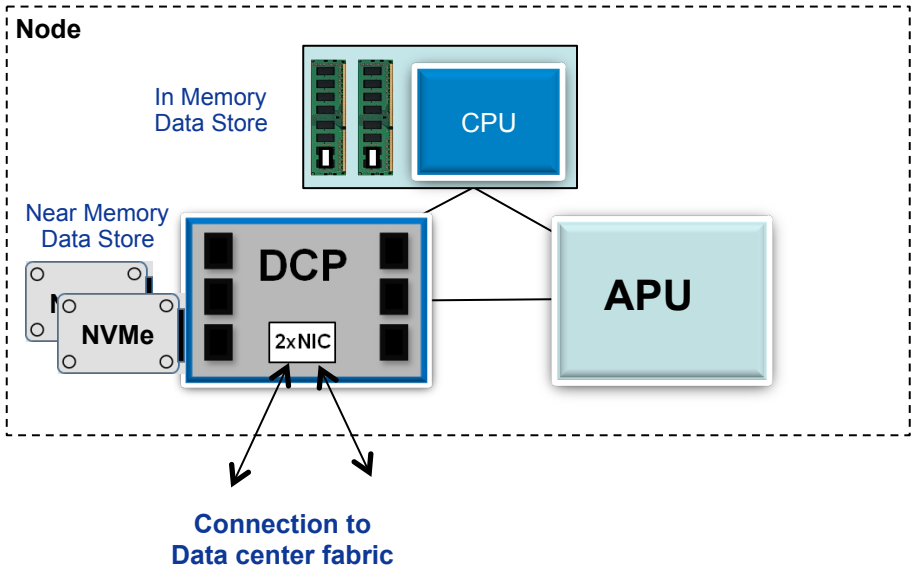
Closely coupled to NVM media
Built in High Speed networking
Low latency link to CPU
Programmable Data Engines

Data Management
Security
Replication/Resiliency support
CODEC
Data Placement
Scanning, etc
Messaging/IPC engine

Architecture for Next gen Applications

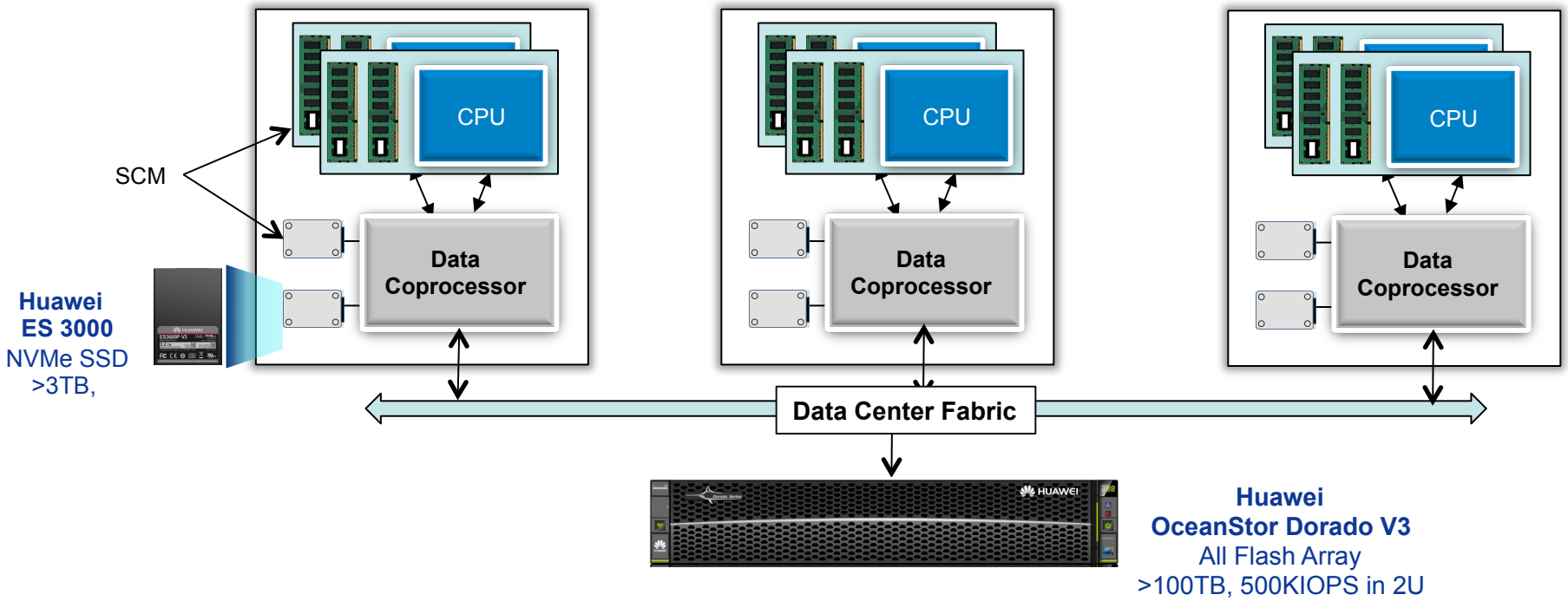
General Purpose Computing

IO Accelerator (Data Mgmt, Etc.)



Application Accelerator (Deep Learning, Etc.)

Adaptable Cluster



Summary

- By extending, scaling and accelerating human expertise, Cognitive Computing is rapidly becoming the most important next gen application
- SCM based Very Large Capacity, Low Latency, processor attached Data Store is one of the key ingredient to make Cognitive Computing ubiquitous
- Advanced Optimization of NVM storage for applications will be key to achieve Performance and TCO objectives

Thank you