SNIA DSI '16 DATA STORAGE INNOVATION CONFERENCE
Innovation in Storage Products, Services, and Solutions
June 13-15, 2016 | Marriott San Mateo | San Mateo, CA

Storage Implications of Cognitive Computing

Balint Fleischer & Jian Li

Huawei

6/13/2016

What do we do with all the Data?

Some examples of Unstructured Data

Financial and legal documents
Email, Blog
Research papers
Video and Social Network postings
Patient records
Industry reports, Market Studies
Regulatory publication

What do we do with all the Data?

Some examples of Unstructured Data

Financial and legal documents
Email, Blog
Research papers
Video and Social Network postings
Patient records
Industry reports, Market Studies
Regulatory publication

Store it

Build a Highly Scalable Storage

What do we do with all the Data?

Some examples of Unstructured Data

Financial and legal documents
Email, Blog
Research papers
Video and Social Network postings
Patient records
Industry reports, Market Studies
Regulatory publication

Process it with an Analytic Platform

Analyze it → Make predictions

Content

Sentiment

What do we do with all the Data?

Utilize it on a Cognitive Computing Platform

Some examples of Unstructured Data

Financial and legal documents
Email, Blog
Research papers
Video and Social Network postings
Patient records
Industry reports, Market Studies
Regulatory publication

Integrate with our knowledge base

Cognitive Computing (AI System)

Augmenting (Enhancing, Scaling and Accelerating) 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

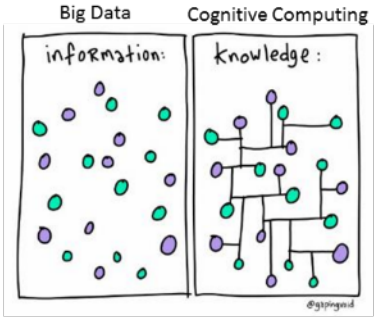
- Evolve with goals and respond to changes
- Participate in the shared discovery process and problem refinement iteration
- Understand meaning, goal, syntax, regulation, time, etc.
- Utilizes real time sensory and behavioral inputs as well as contextual data

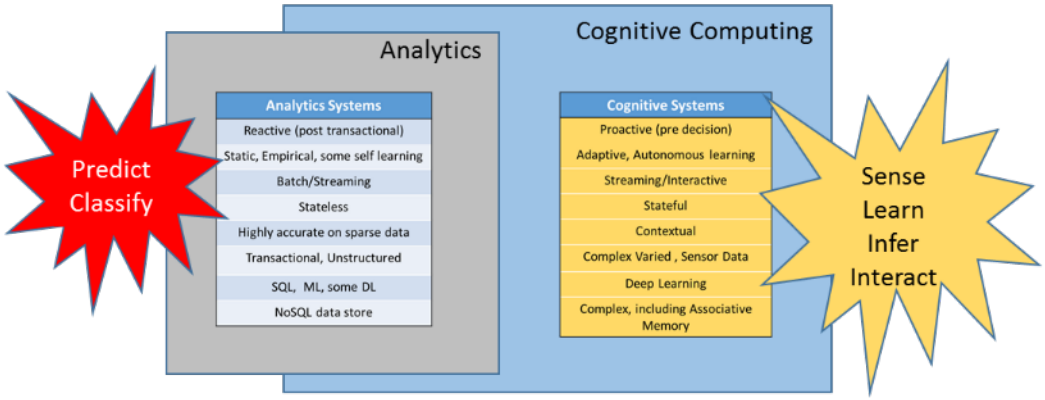
A growing number of use cases benefiting from Interactive Cognitive Systems

Analytics vs. Cognitive Computing

Based on past history:

What restaurant will I choose today





Where should we go:

Everyone will find a choice
Ample parking
The wait is short

Responsive & Accurate

gapingvoid art

Example use case:
Enhancing Mobile User Experience

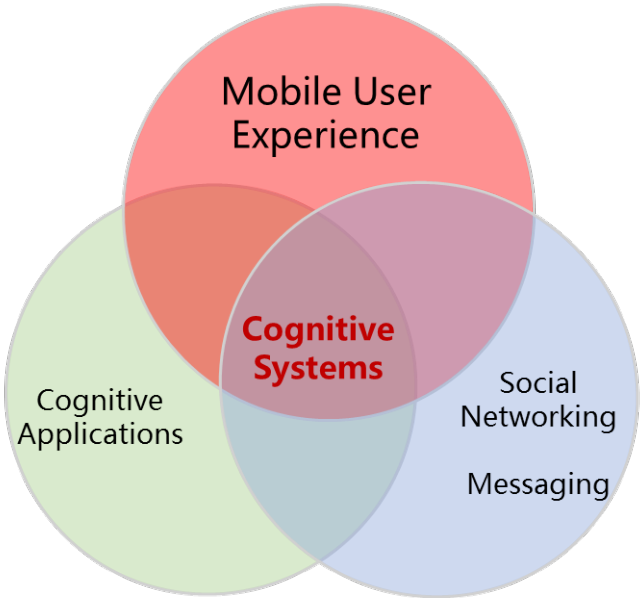
Personal Sensors

Tracking Actions & Surroundings

Location
Usage
Health
Image
Voice
Movement

Mobile User

New customer experiences
Understanding intentions
Anticipating needs



Personal Assistants
Health
Direction
Communication
Shopping
Entertainment
Transportation

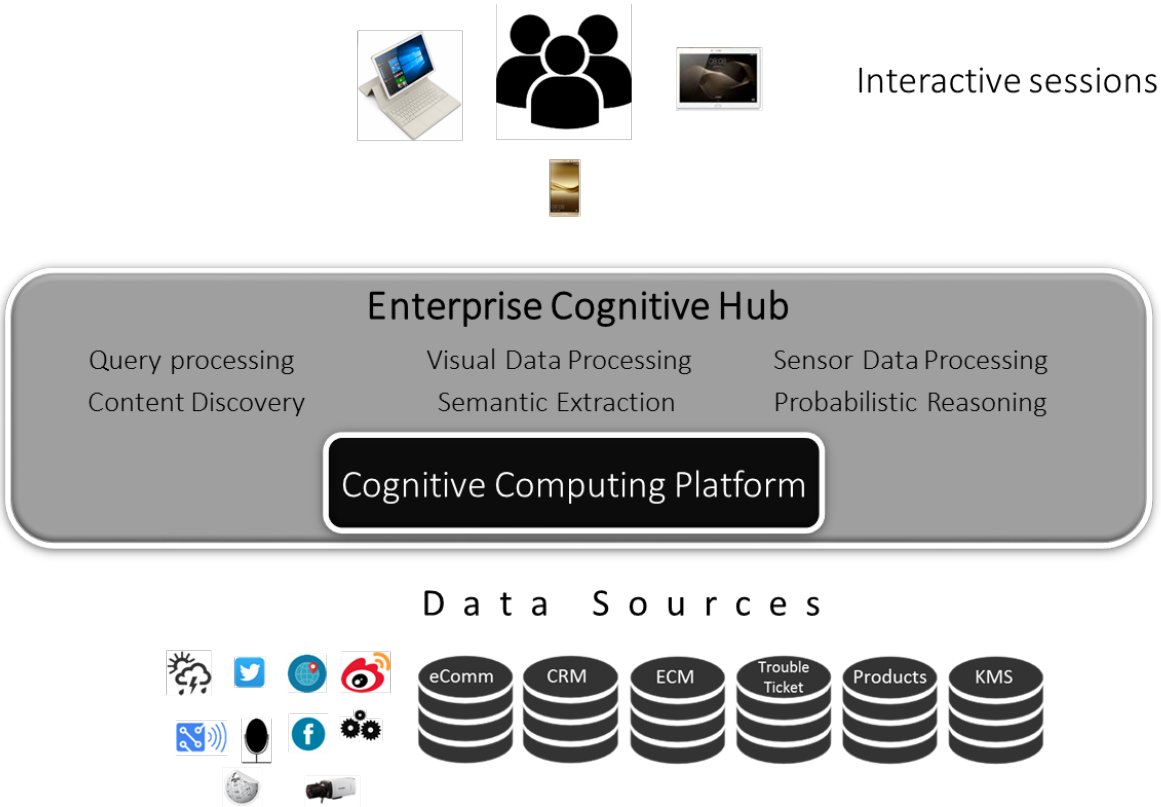
waze
Google Fit
Google Now
Apple Siri
Skype Translator
skype

Smart Connections

Sina Weibo
Baidu
WhatsApp
facebook

SNIA DSI '16 DATA STORAGE INNOVATION CONFERENCE

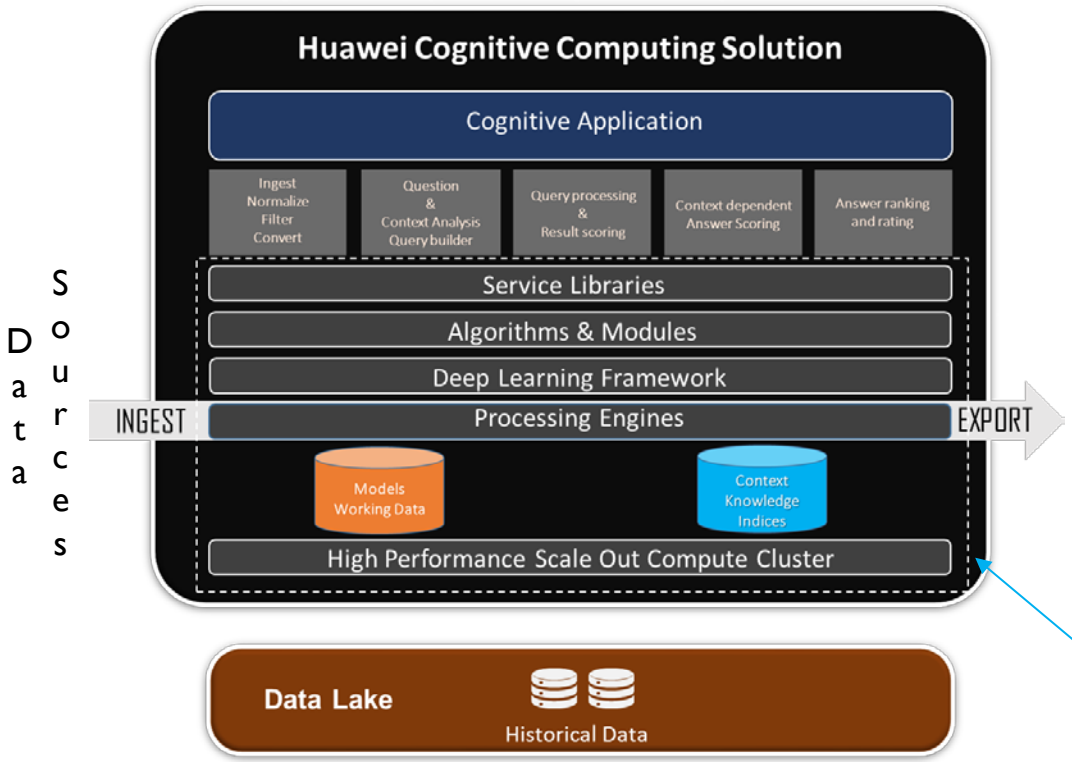
Example use case:
Future Enterprise Management System



Why is Cognitive Computing is different?

- Focusing on Human-Computing Interaction
- Computationally intensive
- Cognitive Computing typically has a response time constraint (QoS)
- Working on Large Data sets
- Data access latency is critical to performance
- Data Access Patterns differ from Analytics (Sparse matrix, Graph etc.)
- Context (state) matters!
- Data Needs to be sharable across applications implementing pipelines
- Multi tenancy with QoS is key to achieve economics

Huawei's Vision for Cognitive Computing



Highly Optimized for Interactive usages

AI based architecture developed for Human-Computer Symbiosis.

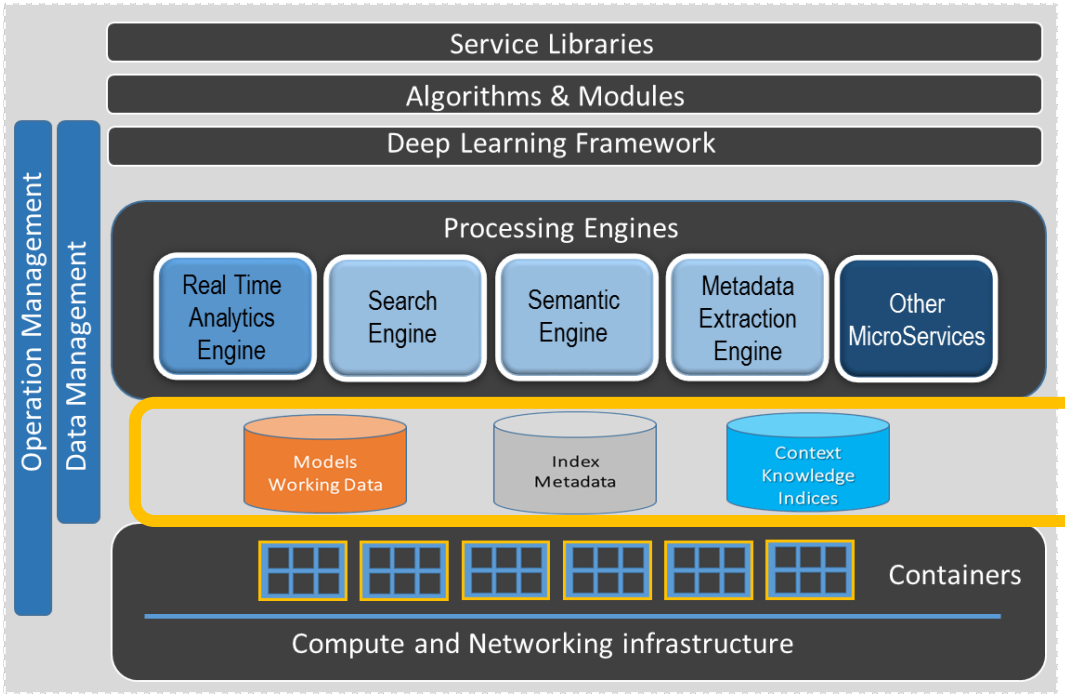
Tight integration of Event and Context Data

Processing engines tuned for performance and real time execution

Extensible Libraries

Product recommendation, Media Sentiment Analytics, Fraud detection, Ranking systems, NLP, Speech to Text and Text to Speech, Tradeoff analytics, Visual Recognition, Cognitive insight, Etc

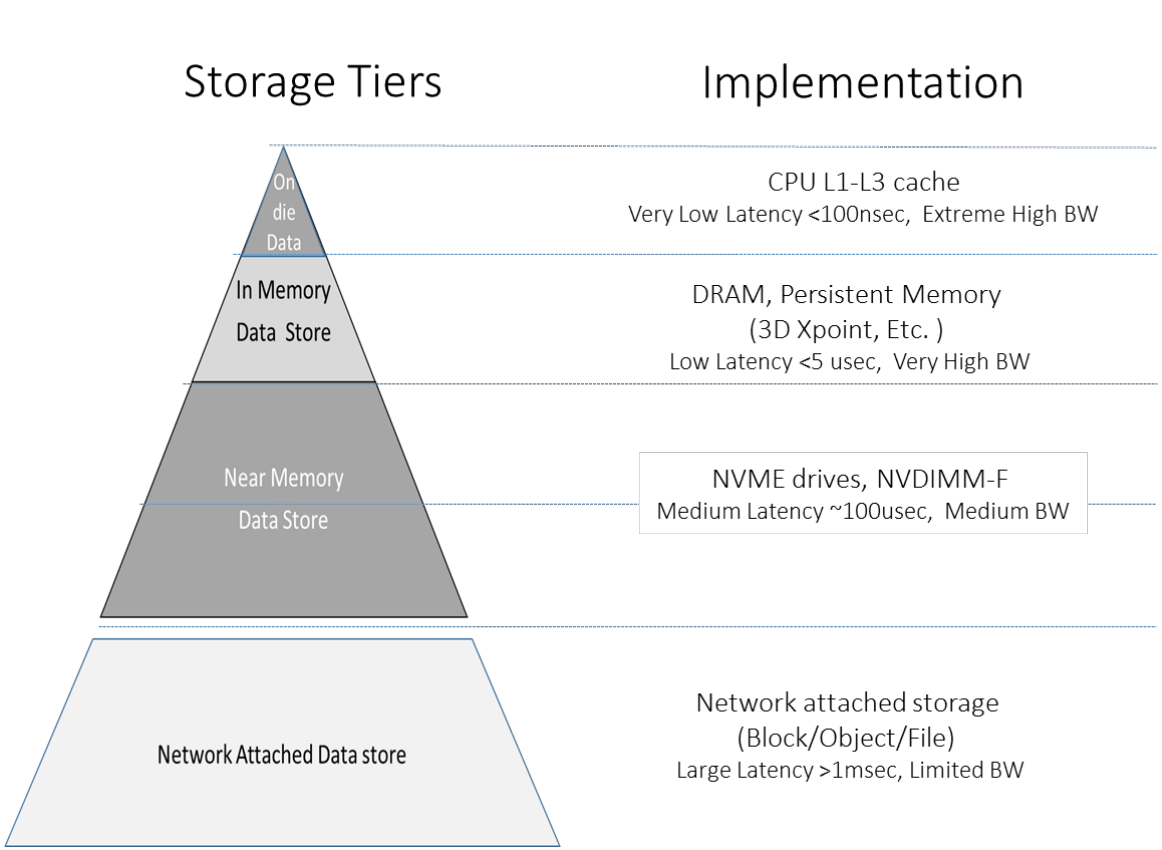
Cognitive Computing Platform



Operational Data Store

High Performance, Low latency
Cognitive Computing Optimized
Scale out, Resilient, Elastic
Sharable across Pipeline stages and Apps

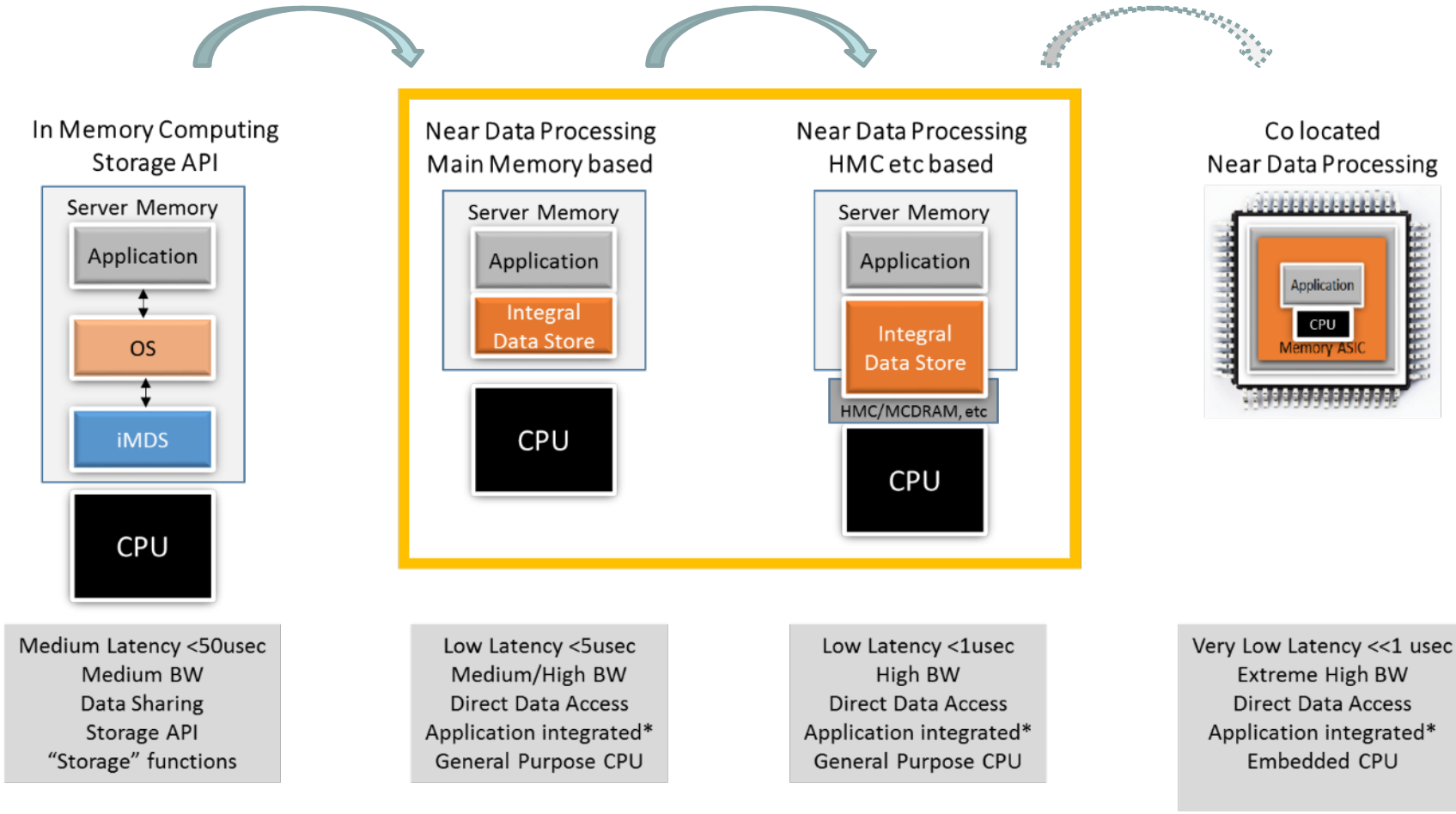
Compressing the Storage Hierarchy



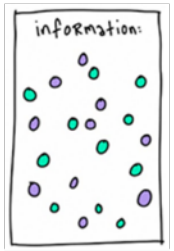
Access Latency Reduction
Storage Stack bypass
Enabling Specific Semantics

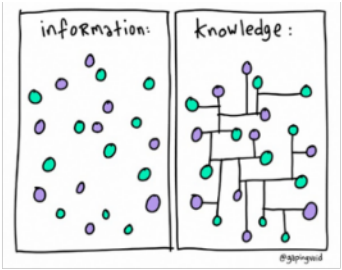
Moving Data
Closer to Processing

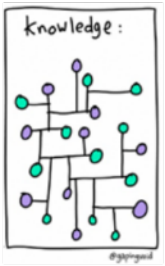
Further Reducing Application to Data Latency



Storage for Cognitive Computing



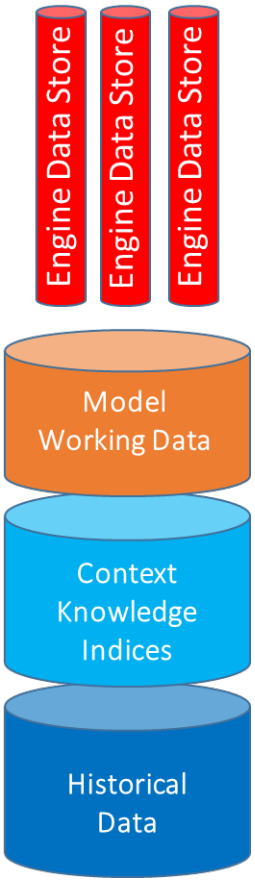




OLTP/OLAP	Big Data Systems	Cognitive Systems	Brain
RDBMS	Batch & Streaming Analytics	Deep Learning Systems Complex Pipelines	Knowledge, Recognition, Intuition, Approximation, Etc.
Structured Data Store Tables	Unstructured Data Store Files, Objects	Cognitive Data Store Objects, Graphs, Matrices	Associative Memory

Memory and information processing in neuromorphic systems Giacomo Indiveri and Shih-Chii Liu Senior, IEEE Proceedings 2015
Saffron Technology

Cognitive Computing Optimized Storage Stack



Model Data Store

- Cluster wide, Resilient and Elastic
- Very low latency to support random access*
- Pipeline data into on die Engine Data Store
- Direct Data Access by applications
- Mapped onto main memory and next gen fast NVME drives
- Data Parallel and Model Parallel modes
- Assist functions for Objects, Graphs, Matrices
- Fast access to **Context Store**

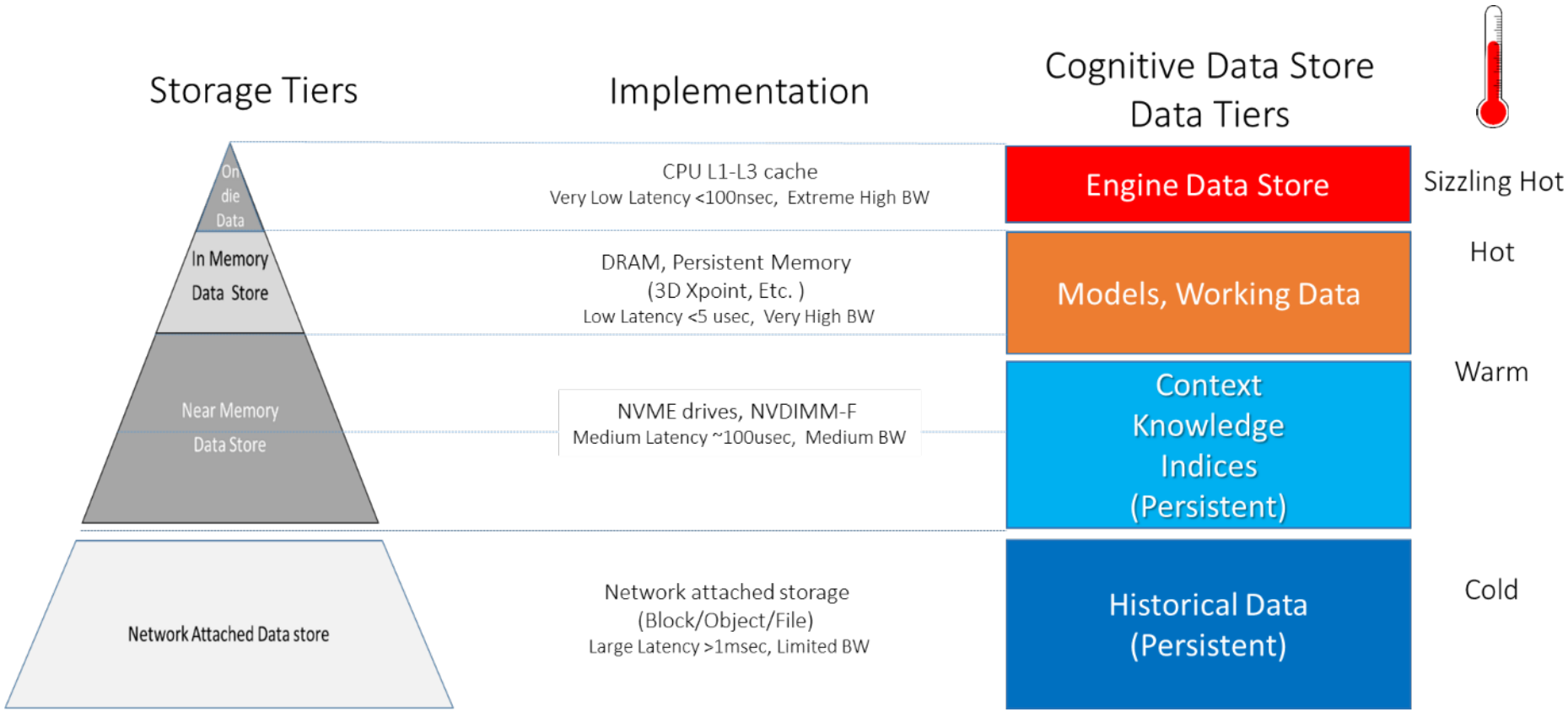
Context Data Store

- Cluster wide, Persistent, Resilient and Elastic
- Stores Pre computed data and Context info
- Mapped onto direct attached NVMe drives

Historical Data Store

- Data Center wide, Network Attached, Persistent, Resilient and Elastic
- Stores Historical and Reference data

Mapping to Classic Storage Hierarchy



Enabling Cognitive Computing with Cloud Services
Huawei's Data Function Virtualization Platform Vision

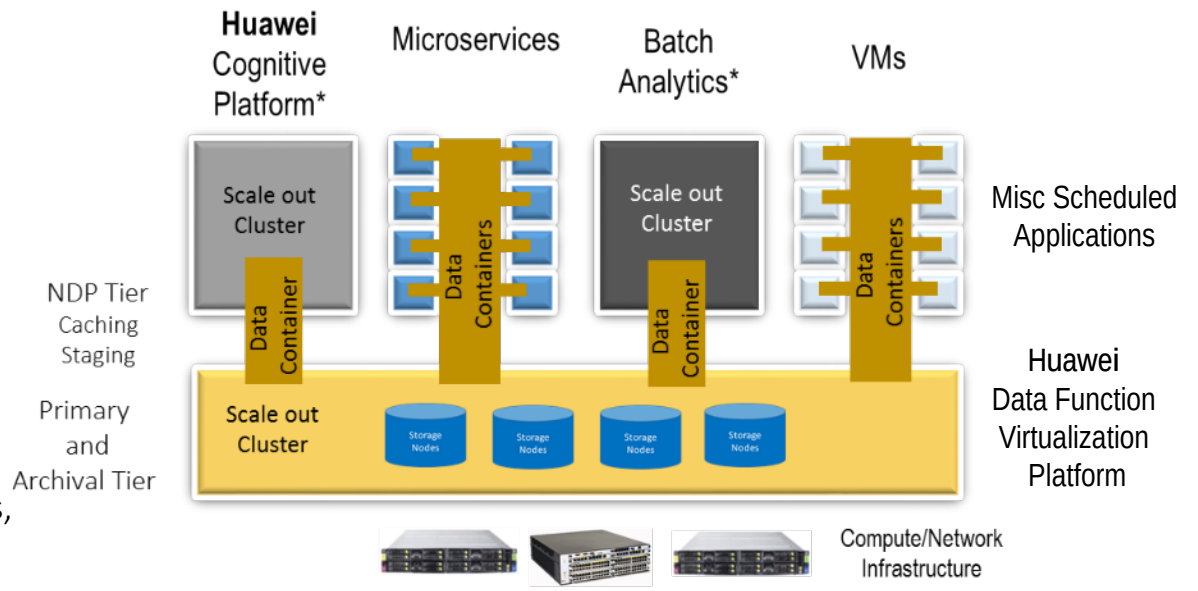
DFV Data Container model

Data Container per application
Application can be stand alone or clustered
Attributes are application specific
Locality management, Storage semantics,
Tiering policy, Resiliency Policy,
Security, Performance,
Sharing semantics, Etc.

DFV Control Plane functions

Pool management, Allocation, Data Services,
Data Pool classes, Migration, Staging,
Recovery Assist, Elasticity management,
Global replication, Etc.

Post Provisioning
Functional Examples



* Note, Analytics and Cognitive Platforms maybe running in containers

Summary

- Augmenting human expertise yield great business value
- New technologies, such as new Memory, high core count CPUs, fast Fabrics and various accelerators are critical HW ingredients of Cognitive Computing Platforms
- New Software innovations coupled built on Near Data Processing enables the delivery these High Performance, Responsive Cognitive Platforms
- Re architecting of the storage stack will make it possible to scale cloud architectures to support high performance solutions
- Huawei is developing a comprehensive vision addressing these changes