MemVerge | intel | NetApp | IDC | PENGUIN COMPUTING

Opening the Door for Big Memory

Presentation Team

Joe Barnes
Director
MemVerge

Eric Burgener
VP
IDC

Kristie Mann
Sr. Director
Intel

Charles Fan
CEO
MemVerge

Brett Roscoe
VP
NetApp

Kevin Tubbs
Sr. VP
Penguin Computing

Darrell Westbury
Sr. Director
Credit Suisse

Agenda

Time	Topic	Presenter
9:00 – 9:05	Introductions	Joe Barnes, MemVerge
9:05 – 9:15	New “Big Memory” Category	Eric Burgener, IDC
9:15 – 9:25	Re-Architecting The Data Landscape	Kristie Mann, Intel
9:25 – 9:35	Big Memory Software: Memory Machine	Charles Fan, MemVerge
9:35 – 9:45	Why MemVerge?	Brett Roscoe, NetApp
9:45 – 9:55	Big Memory: A Case Study	Kevin Tubbs, Penguin Computing
9:55 – 10:05	Customer PoV: Where Big Memory Fits	Darrell Westbury, Credit Suisse
10:05 – 10:15	Q&A	Joe Barnes, MemVerge

Digital Transformation Driving New “Big Memory” Requirements

Eric Burgener, Research Vice President

Infrastructure Systems, Platforms and Technologies Group

May 2020

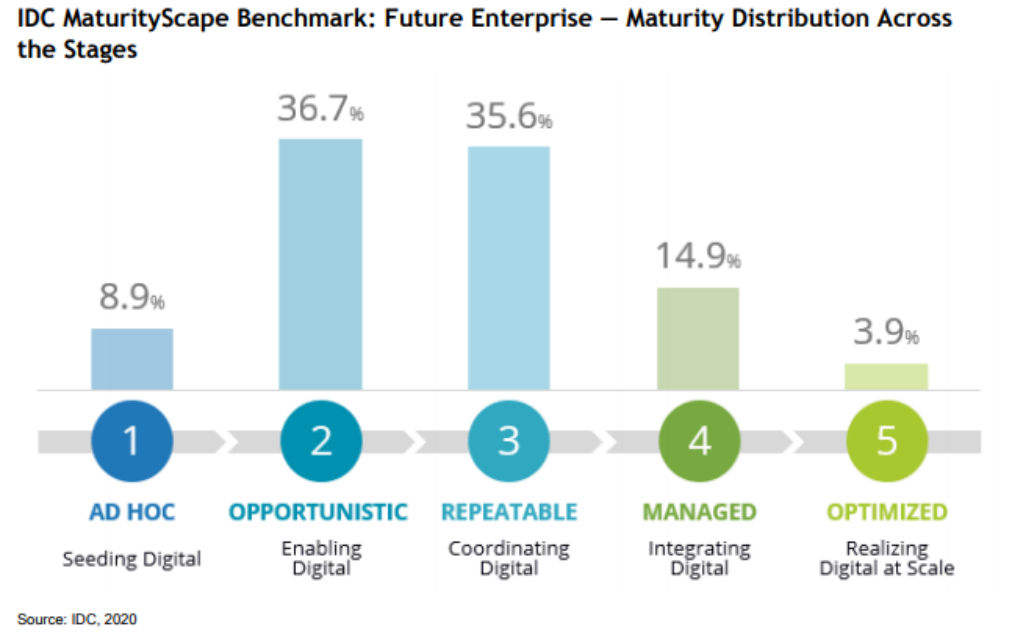
What Is Digital Transformation (DX)?

The digitization of business models, processes, products and services

Sets enterprises up to optimally take advantage of big data analytics

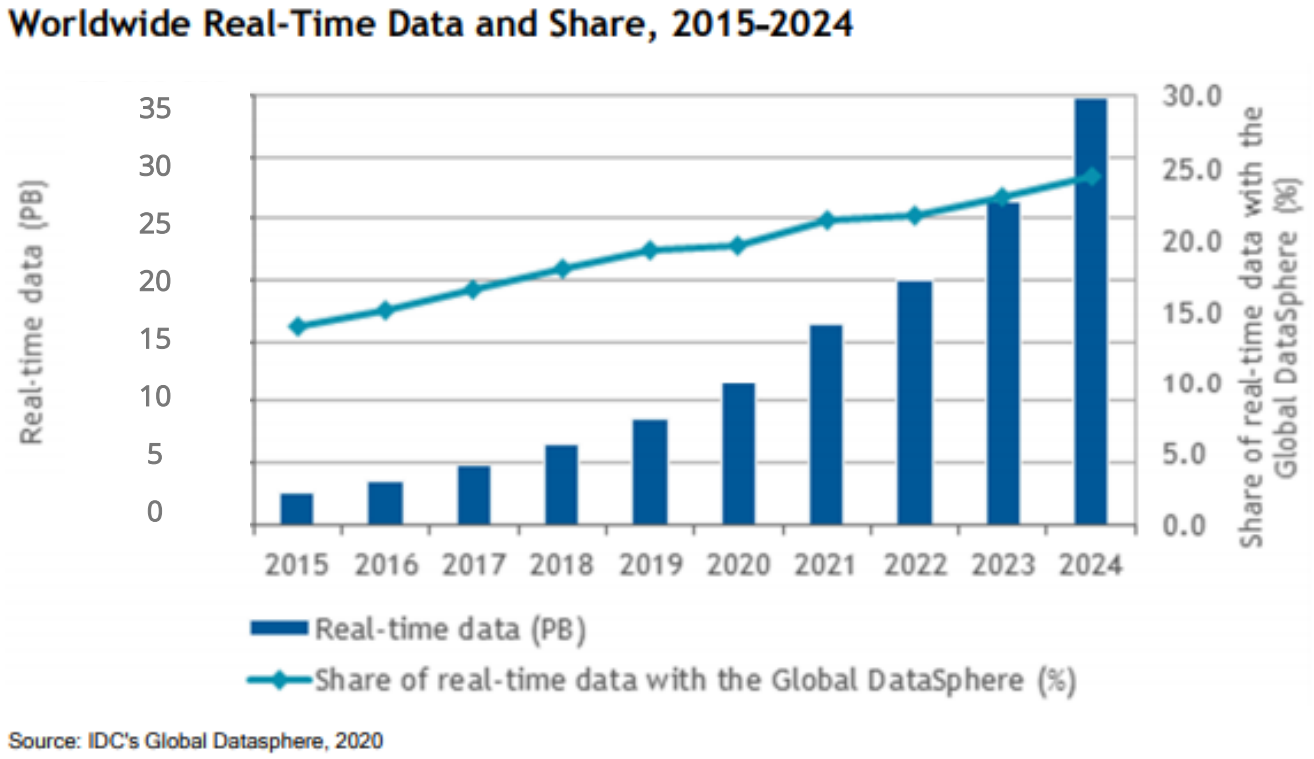
Digital Transformation is Widespread

DX Maturity Distribution



- 91.1% of enterprises undergoing DX in the next three years
- More data-centric business models will drive AI/ML-infused analytics
- Performance and availability implications for enterprise storage
- Market evolution will drive demand for persistent memory technologies

Real-Time Workloads Are On The Rise

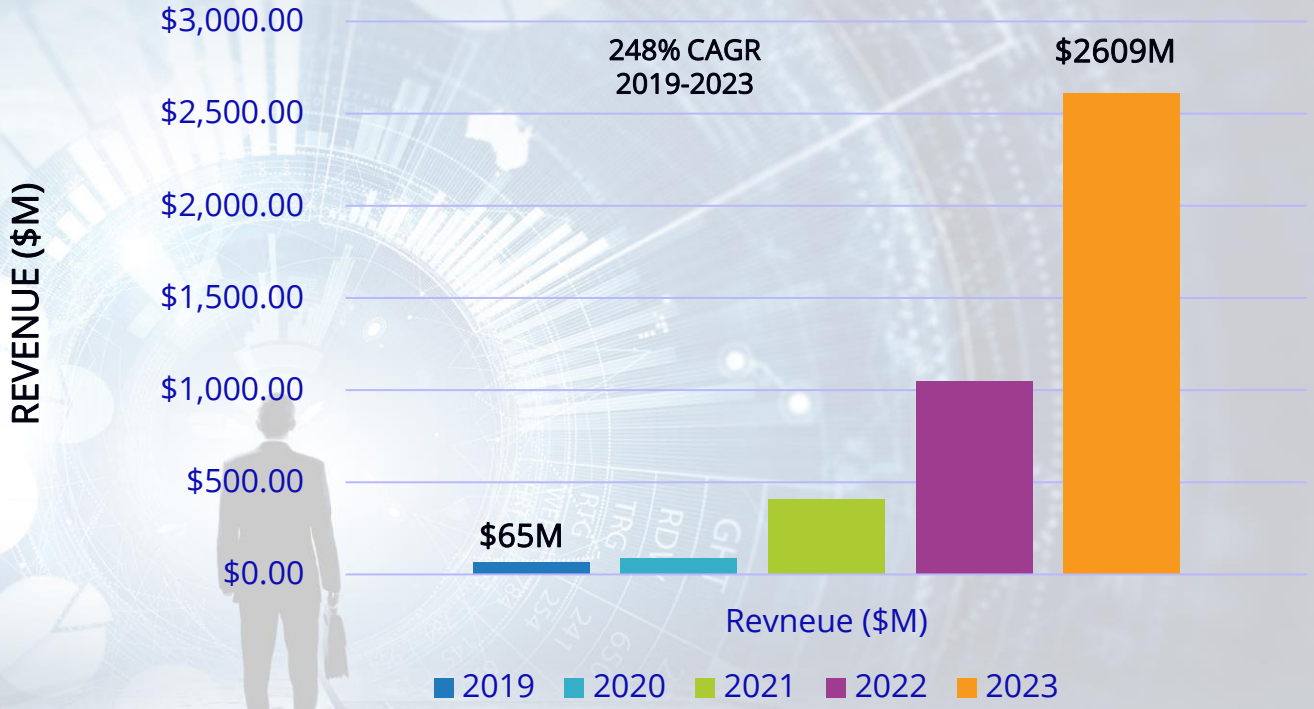


Source: IDC's Global Datasphere, 2020

Worldwide, data is growing at a 26.0% CAGR, and in 2024 there will be 143 zettabytes of data created

By 2021, 60-70% of the Global 2000 will have at least one mission-critical real-time workload

PM Revenue Forecast, 2019 - 2023



Business Drivers

MARKET EVOLUTION TO REAL-TIME

- Upping the ante: massive data sets, real-time orientation
- For accelerated computing, storage is still the bottleneck
- Value propositions include competitive differentiation, increased revenue

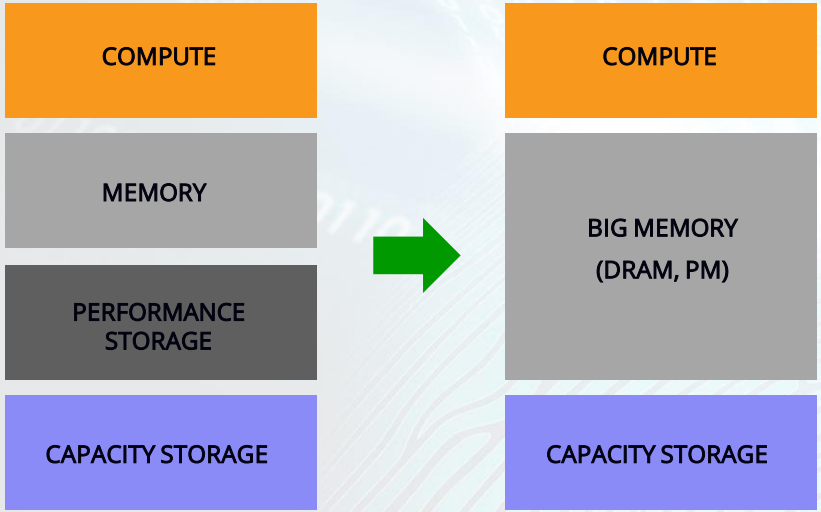
CONFLUENCE OF AVAILABLE TECHNOLOGIES

- Artificial intelligence and machine learning
- More concentrated compute power than ever before
- Emerging persistent memory technologies
- Memory virtualization with intelligent data placement

TARGET WORKLOADS

- Latency-sensitive transactional workloads (trading floor apps, etc.)
- Real-time big data analytics in financial services, healthcare, retail, etc.
- AI/ML analytics and inferencing (fraud analytics, social media, etc.)

Defining “Big Memory Computing”



OLD MODEL

BIG MEMORY COMPUTING

- Enables the ability to run applications in memory for improved performance and efficiency
  - Leverages byte addressable memory media
- Includes enterprise-class data services to handle tier 1 availability and management requirements
- Runs on a software-based memory virtualization layer on industry standard hardware without application modification
- The technology enabler for mission-critical real-time computing

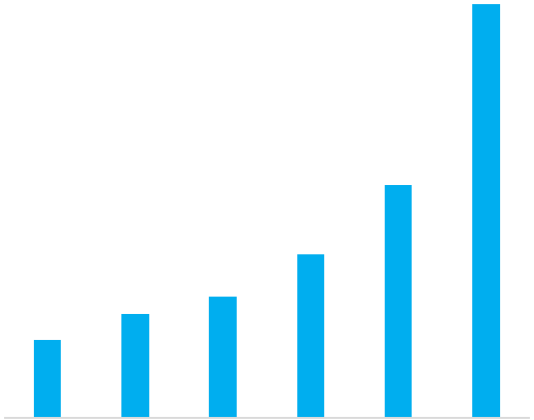
RE-ARCHITECTING THE DATA LANDSCAPE

Kristie Mann

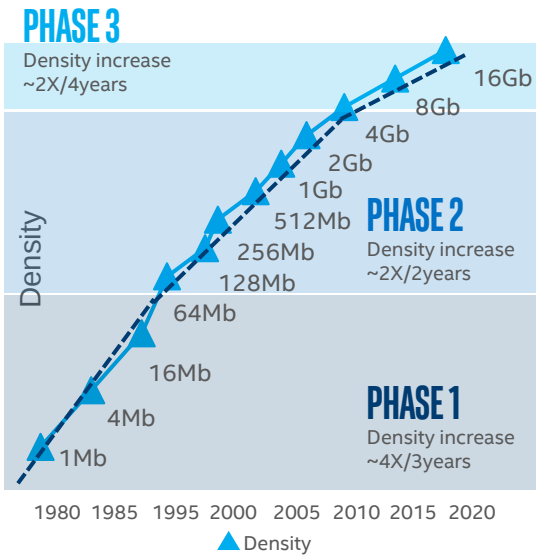
Sr. Director of Products

ADAPTING TO THE CHANGING DATA LANDSCAPE

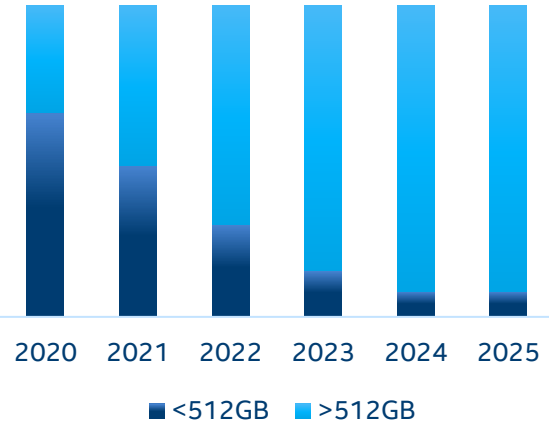
COMPUTE DEMAND ACCELERATING



DRAM IS NOT SCALING

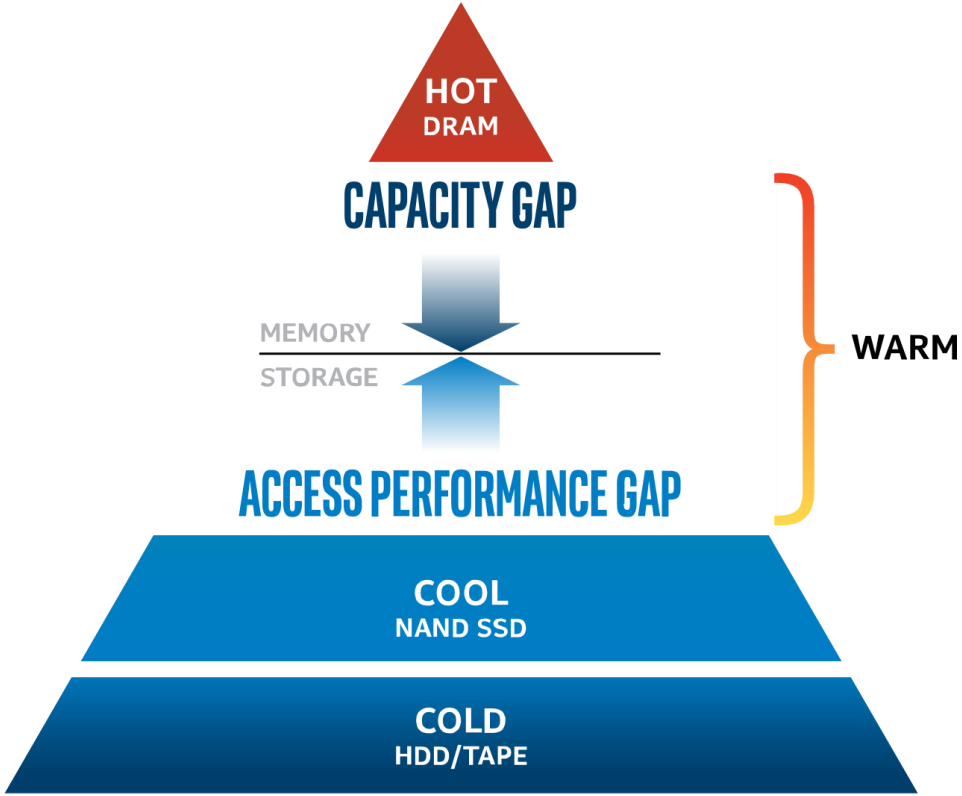


LARGE MEMORY SYSTEMS GROWING

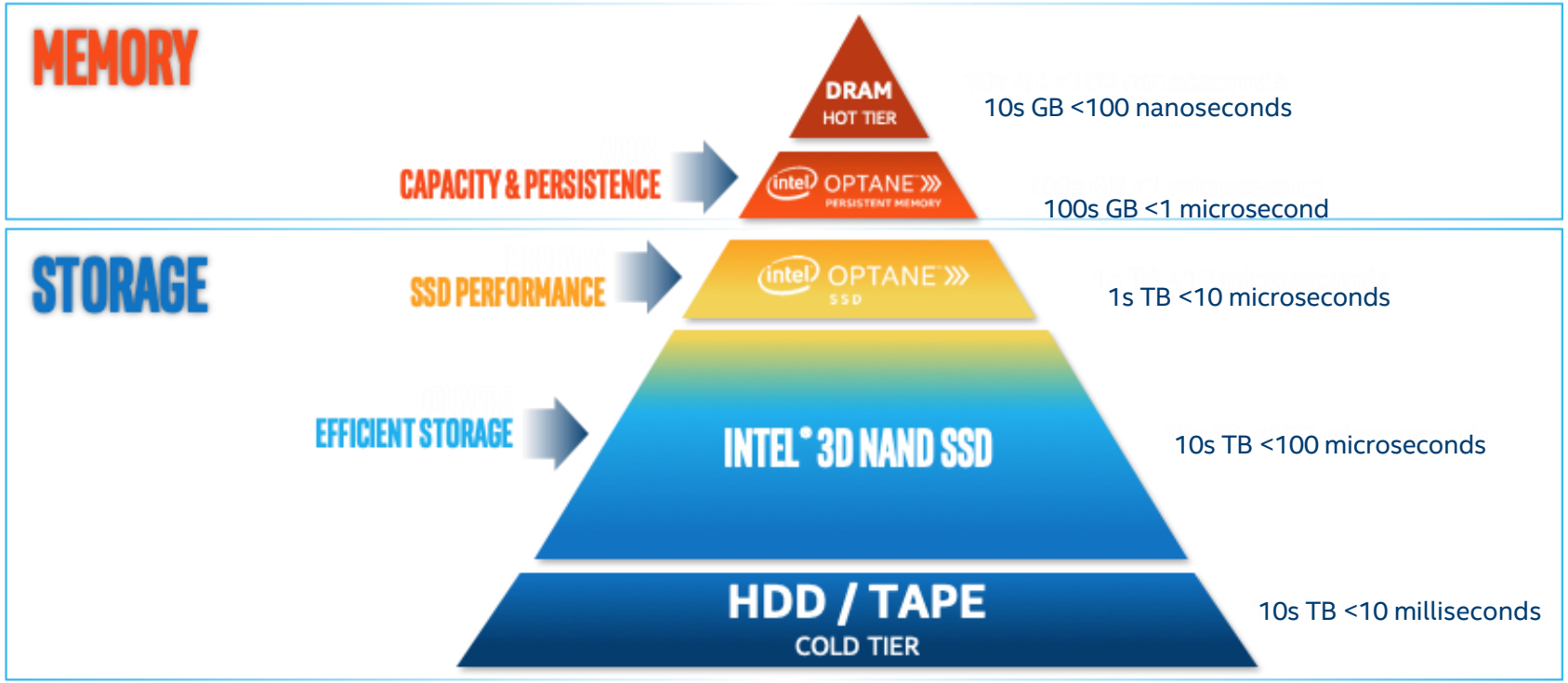


Source: Intel estimates; "3D NAND Technology – Implications for Enterprise Storage Applications" by J.Yoon (IBM), 2015 Flash Memory Summit; "DRAM Market Monitor Q1-20" by Yole Development

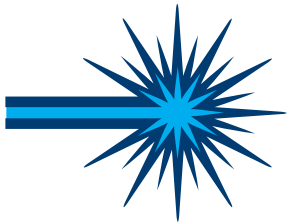
GAPS IN THE MEMORY/STORAGE HIERARCHY



THE BEST OF MEMORY AND STORAGE WITH PERSISTENT MEMORY



INTEL® OPTANE™ PERSISTENT MEMORY IS UNIQUE



REVOLUTIONARY MATERIAL

Most significant memory and storage advancement **in the last 20 years**

WRITE IN PLACE

Set or reset data as needed, **no need to erase media**

BYTE ADDRESSABLE

Every memory cell can be **individually addressed**

LOW LATENCY

...together delivering **remarkably fast media**

THE BEST OF MEMORY AND STORAGE

THE OPPORTUNITY IS IMMENSE

IN-MEMORY DATABASE

ADVANCED ANALYTICS

CLOUD WORKLOAD
Virtualized and Hybrid

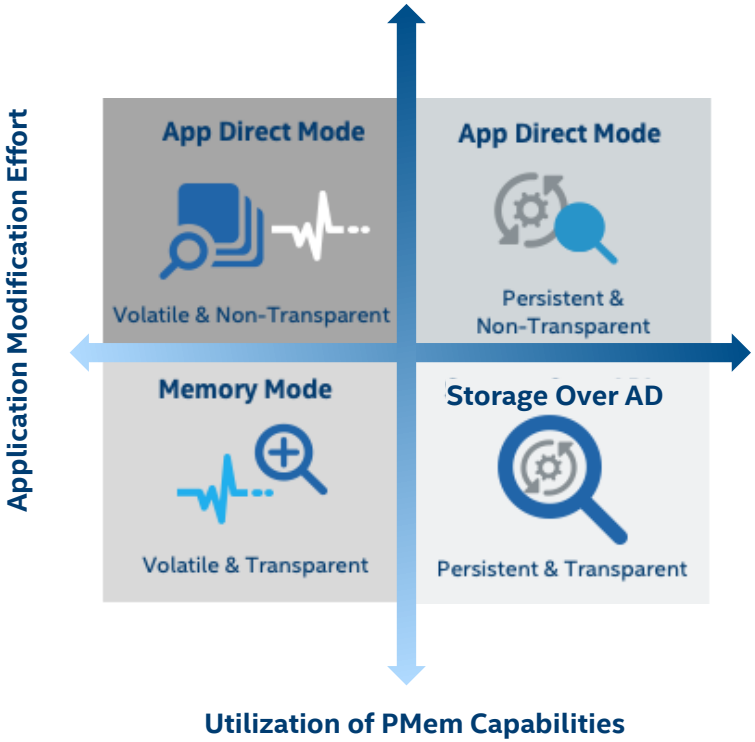
HCI COMPUTE

YET TO BE DISCOVERED

MemVerge

SPEED TIME-TO-VALUE AND INSIGHT FOR DATA-INTENSIVE WORKLOADS

SOFTWARE OPTIMIZATION DRIVES VALUE FOR CUSTOMERS



MemVerge

OPTIMIZED SOFTWARE – SO YOU DON'T HAVE TO

- Virtualized pools of memory
- Byte-addressable persistence without code changes
- Data Services for High Availability and Durability
- Super-charged performance and flexibility in memory

MemVerge | intel | NetApp | IDC | PENGUIN COMPUTING

Big Memory Software: Memory Machine

Charles Fan

MemVerge

What We Announced on May 12

- New Big Memory Computing category
- New round of $19M investment led by Intel Capital, joined by Cisco Investments, NetApp and SK hynix
- Memory Machine™ software available via Early Access Program

Our BIG MEMORY vision

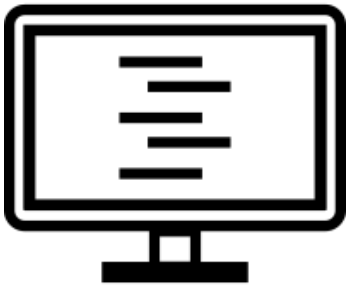
All applications live in memory

MemVerge


INTEL® OPTANE™ DC
PERSISTENT MEMORY
REVOLUTIONIZING MEMORY

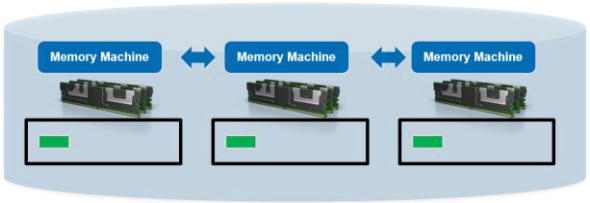
Bringing Mission Critical Apps to Big Memory

Plug-and-play
No App rewrite needed



Data Services
Quick crash recovery

Scale Memory
Beyond a single server

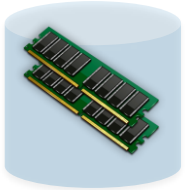


MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

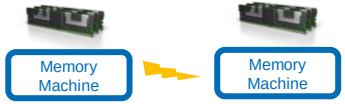
MemVerge Memory Machine™

Software
Subscription

Virtualizes
DRAM & PMEM



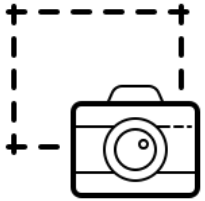
Low Latency
PMEM over RDMA



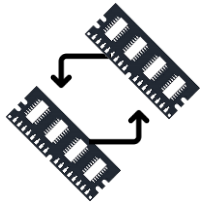
Plug Compatible
No re-writes

Memory Data Services

ZeroIO™ Snapshot



Replication



Tiering

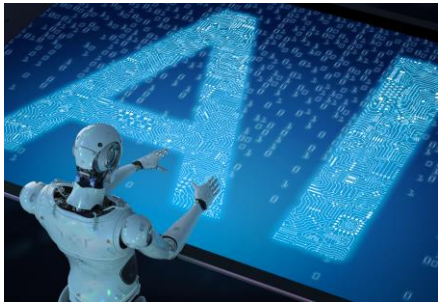
MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Uses Cases: Real-Time Workloads

According to IDC, by 2021, 60-70% of the Global 2000 organizations will have at least one mission-critical real-time workload. Below are just a few examples of use cases that are implementing Big Memory now.

Latency-sensitive transactional workloads such as trading applications

Real-time big data analytics in financial services, healthcare, and retail



AI/ML analytics and inferencing like fraud detection and social media

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Our mission

Open the door to Big Memory

A world of abundance,
persistence and high availability

MemVerge

NetApp

NetApp and MemVerge:

Continued innovation to accelerate business

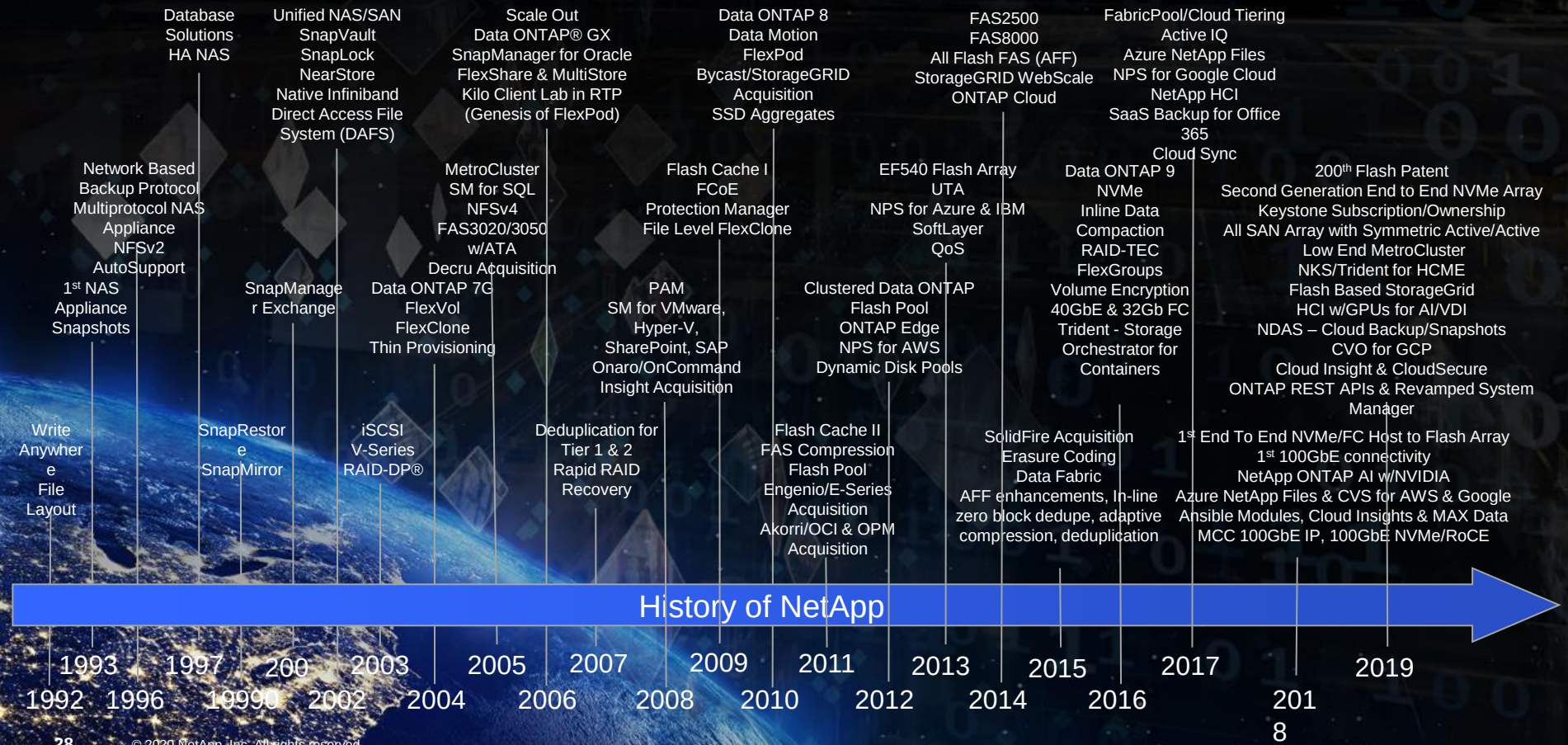
Brett Roscoe
Vice President of Product Management, HCI

May 19, 2020

Over 27 Years of Customer-Focused

RELENTLESS INNOVATION

27 Years of Customer-Focused Storage Innovation



NetApp's Innovation Continues into the Public Clouds

Only NetApp provides a cloud native shared storage experience on the cloud of your choice

Microsoft Azure

Google Cloud

aws

NetApp Cloud

The #1 Platform for **Shared Storage**, **Modern Workplace**, and **Continuous Optimization** on the cloud of your choice

- #1 Platform for SAP Hana with >200 deployments on Azure and Google Cloud Platform
- #1 platform for WVD Enterprise (deployments >200 users)
- #1 platform for High Performance Compute
- Azure NetApp Files (ANF) is faster than block based premium SSD's

NetApp and MemVerge: Continuing our history of innovation

The next evolution in innovation in application and data acceleration

- NetApp has a long-term industry view
- We constantly leverage in-house and industry innovations to accelerate the data we manage and the businesses it drives
- The market and competitive transitions are accelerating with COVID-19
- NetApp has been working to accelerate applications that are data centric including AI and machine learning
- Near zero latency will drive another round of innovation in data centric applications
- MemVerge is bringing lightning-fast, in-memory support of tier 1 applications that promises to deliver another disruption in performance and capabilities

NetApp

MemVerge | intel | NetApp | IDC | PENGUIN COMPUTING

Big Memory: Case Study

Kevin Tubbs

PENGUIN COMPUTING

Deep Learning Recommendation on Persistent Memory

Motivation

Facebook DLRM

DLRM: An advanced, open source deep learning recommendation model
https://ai.facebook.com/blog/dlrm-an-advanced-open-source-deep-learning-recommendation-model/

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Compute Anywhere – Edge to Core Technology

Market Trends

- Hardware Abstraction & Workload Driven Architectures
- Software Defined Architectures
- AI / Machine Learning, Advanced Analytics & Real-time workloads
- Workload Portability & Cloud Native Technologies

SMART GLOBAL HOLDINGS

PENGUIN COMPUTING

A subsidiary of SMART Global Holdings, Inc.

Datacenter and Cloud Scale Computing

+

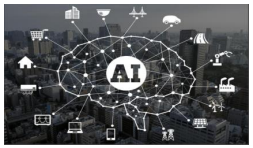
SMART™ Embedded Computing

Embedded Computing

+

SMART™ Wireless Computing

Wireless Computing

Core	Near Edge	Far Edge
←	 (AI / ML)	→
←	 (Advanced Analytics)	→
←	 (Real-time Workloads)	→

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Customer Dynamics

Current Trends & Requirements

Large model & embedding table size

- Model size to GB level, embedding table size to TB level
- Multiple models on single server

Model Size	GB
Embedding Table Size	TB

Fast for online inference service

- Need be finished in tens of μs

Solution

Put models and embedding tables into DRAM

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Customer Dynamics

Limitations

High TCO

Limited DRAM Space

Volatile

Solution

Models and embedding tables in DRAM + PMEM

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Case Study: AI / Machine Learning – Facebook's DLRM

Background

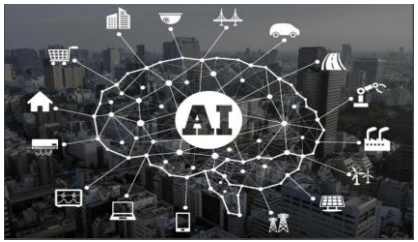
Customer Type: AL / ML Customer

Business Challenge:

- Dynamic and Scalable Production Inferencing

Context:

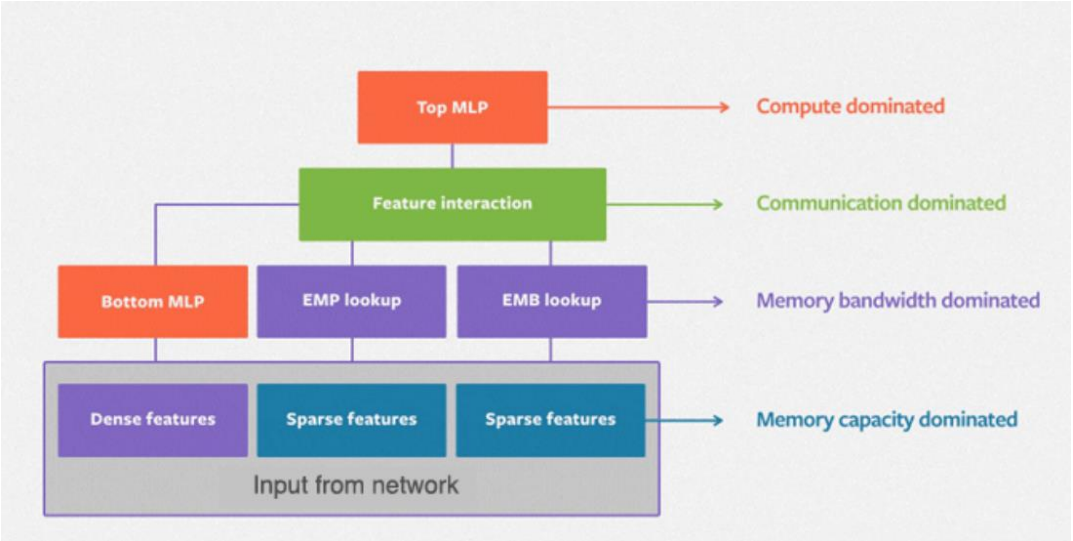
- Scalable Model and Data Capabilities
- Enterprise Class Data Services
- Big Memory Computing Performance Demands



Penguin, MemVerge and Intel Solution

➢ **Deep learning recommendation model for personalization and recommendation systems**

- Consists of dense and sparse features
- Dense feature: a vector of floating-point values
- Sparse feature: a list of sparse indices into embedding tables
- Open source:



MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Case Study: AI / Machine Learning – Facebook's DLRM

Background

Customer Type: AL / ML Customer

Business Challenge:

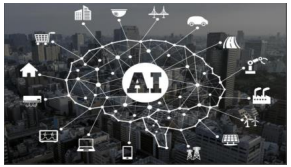
- Dynamic and Scalable Production Inferencing

Platform:

- Innovative Big Memory Computing platform for leveraging persistent memory for real-time, AI/ML and Advanced Analytics and extensible to all memory – Centric workloads.

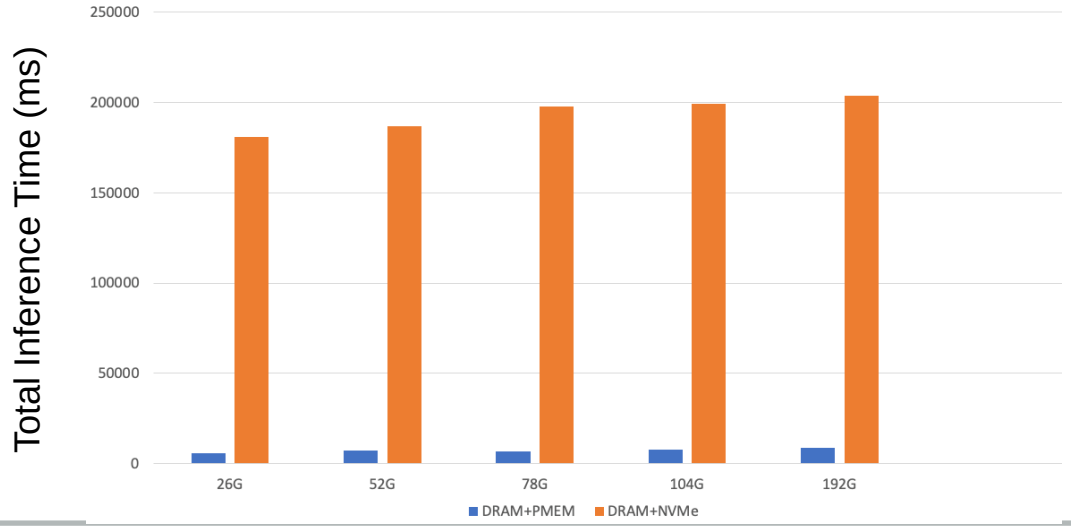
Software:

- Software Defined Architecture extracting performance benefits of cutting edge hardware supporting workload portability to truly compute anywhere with the memory speeds.



Penguin, MemVerge and Intel Solution

~400GB swap on NVMe is on, 20480 samples evaluated



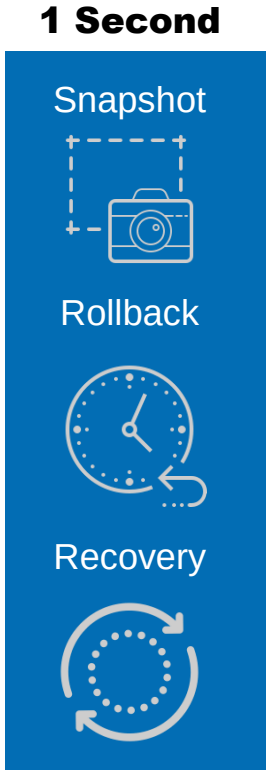
Result:

- Customer has state of the art AI/ML Big Memory Platform that is can scale and deliver performance when Data is Greater than Memory
- Achieved flexible software defined platform Big Memory Computing capabilities and poised for future dynamic model and data growth

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Instant Model Snapshot and Rollback/Recovery for Online Serving

- How to improve the fault tolerance of new model publishing?
  - Pushing new model into production is risky
  - If failed, revert to last workable version ASAP
  - Rollback/Model reloading takes time (for large models) due to slow I/O
- Leveraging PMEM's persistence
  - Take a snapshot of the model serving application
  - Restore a snapshot without reloading from disk or remote storage
  - Snapshot can be published to many serving nodes through our fast pub/sub service
- Solution
  - Instantaneous snapshot without interrupting online inference
  - Instantaneous rollback without loading and publishing time
  - Snapshot, rollback, and recovery are within 1 second



MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

MemVerge | intel | NetApp | IDC | PENGUIN COMPUTING

Customer PoV: Where Big Memory Fits

Darrell Westbury

CREDIT SUISSE

Gradients of Memory and Storage

Balancing the Cost of Performance and Density

Type	Approx. Latency	Density Scale	Relative Cost
CPU			
Registers	1 clock cycle	Bytes	N/A (within CPU die)
CPU Cache			
L1 Cache	4 clock cycles	Kilobytes (2^{10})	N/A (within CPU die)
L2 Cache	10 clock cycles	Kilobytes (2^{10})	N/A (within CPU die)
L3 Cache	20 clock cycles	Megabytes (2^{20})	Premium (Special CPU SKU)
Main Memory			
DRAM	100 nanoseconds (10^{-9})	Gigabytes (2^{30})	High
3D XPoint NEW!	250 nanoseconds (10^{-9})	Gigabytes (2^{30})	Medium-High
Primary Storage			
NVMe SSD	150 microseconds (10^{-6})	Terabytes (2^{40})	Moderate
SAS SSD	500 microseconds (10^{-6})	Terabytes (2^{40})	Reasonable
SAS HDD	3 milliseconds (10^{-3})	Terabytes (2^{40})	Low

BETTER PERFORMANCE

HIGHER DENSITY

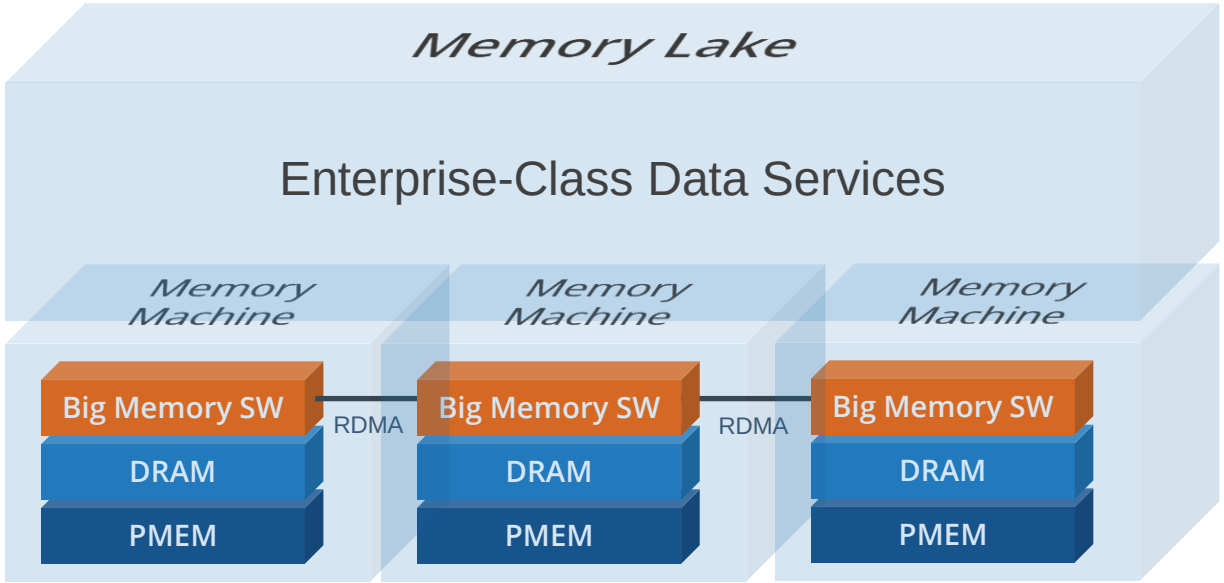
LOWER COST

CREDIT SUISSE

Big Memory Lakes are now possible

MemVerge software enables memory pooling and real-time replication

Aided by ultra-low latency network such as RDMA and/or PCIe bridging technologies



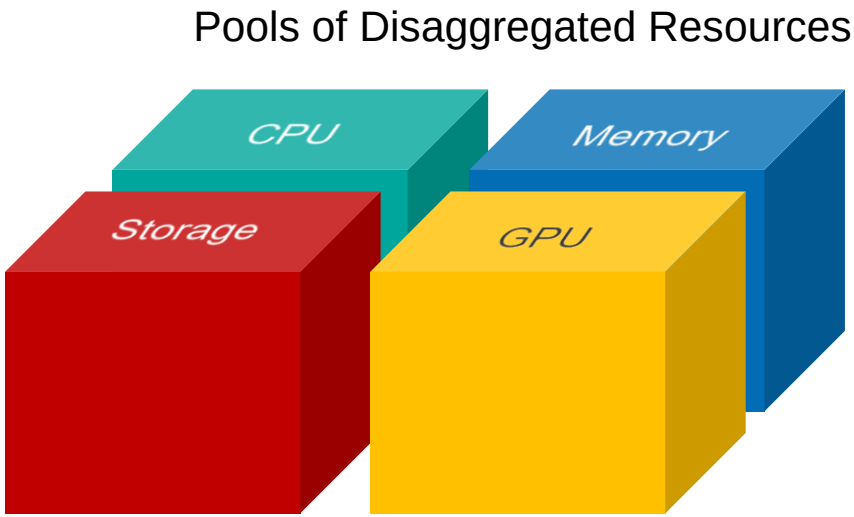
CREDIT SUISSE

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

What's magical and becomes possible in the near future:

Memory bus technologies, such as Intel's Compute Express Link (CXL) will further reduce inter-server latency

Instantaneous snapshots of memory, combined with CXL, will make the dream of dynamically composable infrastructure possible.



CREDIT SUISSE | MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Thank-you

Try it

Contact andrew.degnan@memverge.com to sign-up for a PoC

Download this presentation at:

https://www.memverge.com/opening-the-door-to-big-memory/

View this webinar on the MemVerge YouTube Channel

https://www.youtube.com/channel/UCLT4fehLcQiW4bfQgrHIlpg/featured

MemVerge | intel | IDC | NetApp | PENGUIN COMPUTING

Q&A

MemVerge

What happens in memory stays in memory…