Hewlett Packard Enterprise

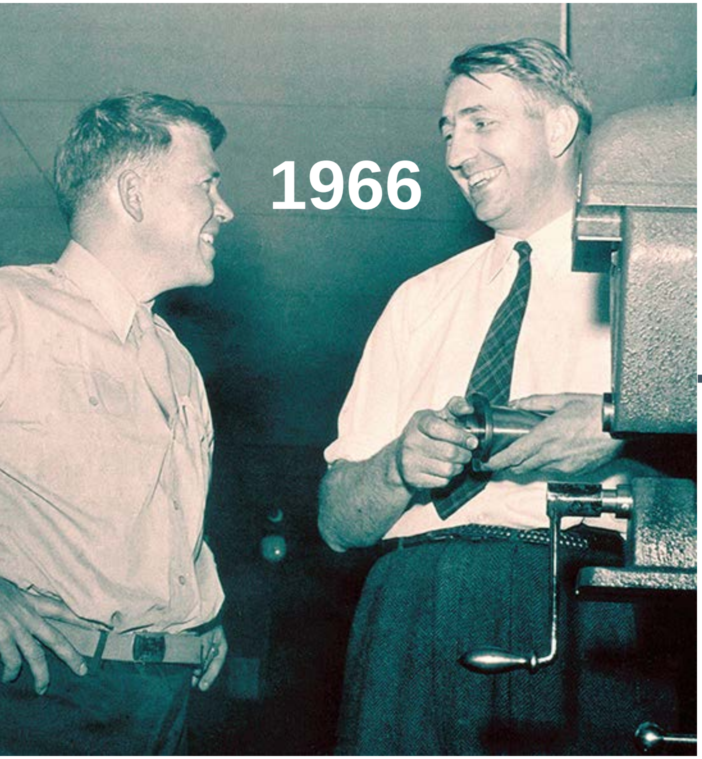
Memory-Driven Computing

A vision for the future of computing

Dejan Milojicic, Distinguished Technologist
Hewlett Packard Labs

With contributions with many, many people from HPE

Innovation is our legacy and our future



HP Labs

1966
Light-Emitting Diode (LED)

1967
Cesium-beam atomic clock

1968
Programmable Desktop Calculator

1972
Pocket Scientific Calculator

1975
Standard for Interface Bus

1980
64-channel Ultrasound

1980
Office Laser Printer

1984
Inkjet Printer

1986
3D graphics workstations

1986
Commercialized RISC chips

1989
Digital Data Storage Drive

1994
64-bit architecture

1999
Molecular Logic Gate

2001
Utility Data Center

2002
Rewritable DVD for standard players

2003
Smart Cooling

2005
Virus Throttle

2008
Memristor discovered

2010
ePrint

2011
MagCloud

2011
StoreOnce

2011
3D Photon Engine

2012
StoreAll

2012
OpenFlow switches

2013
HP Moonshot

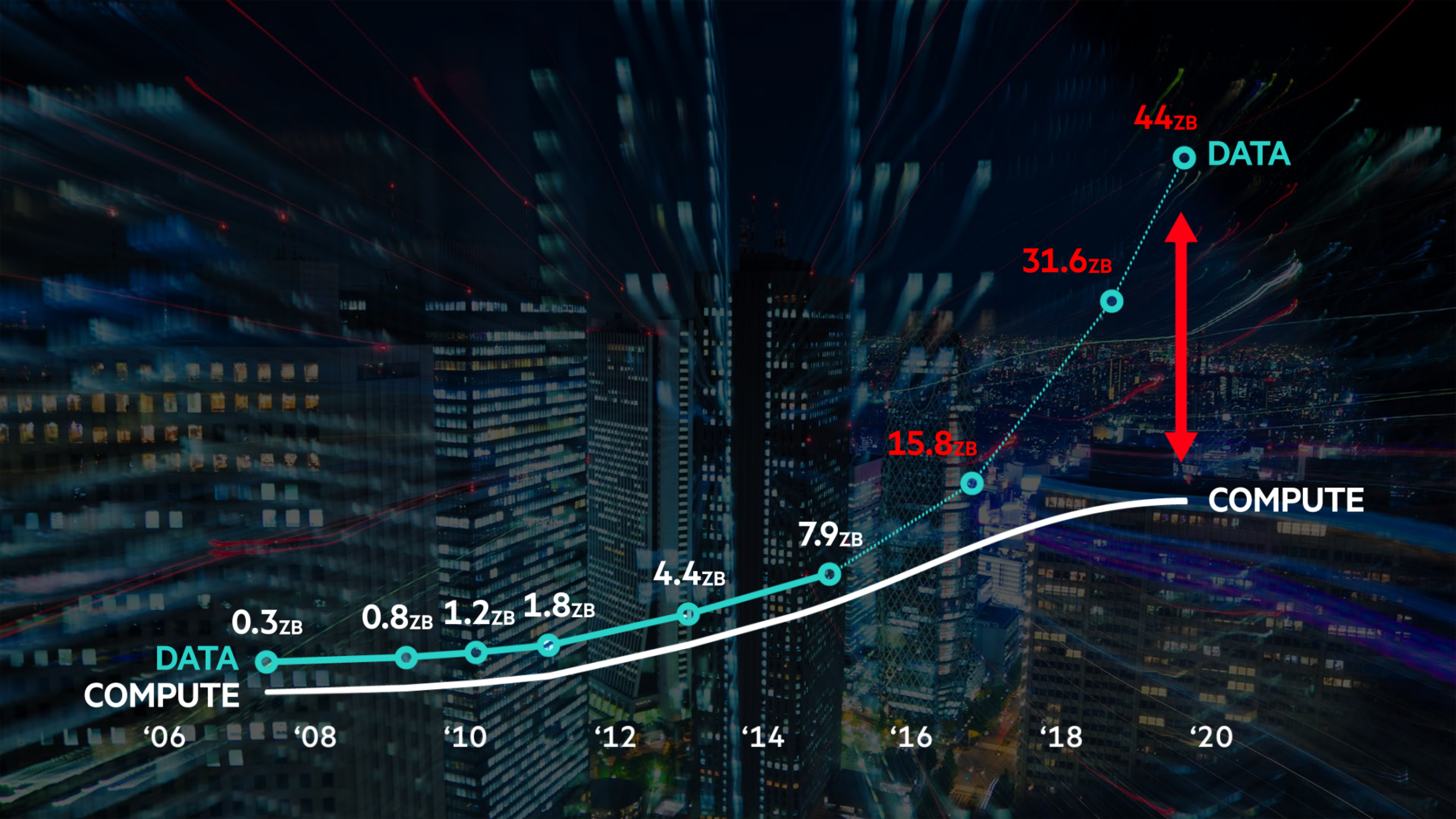
2013
Threat Central

2013
SureStart

2014
DNS Logging
Trafodion
Location Aware

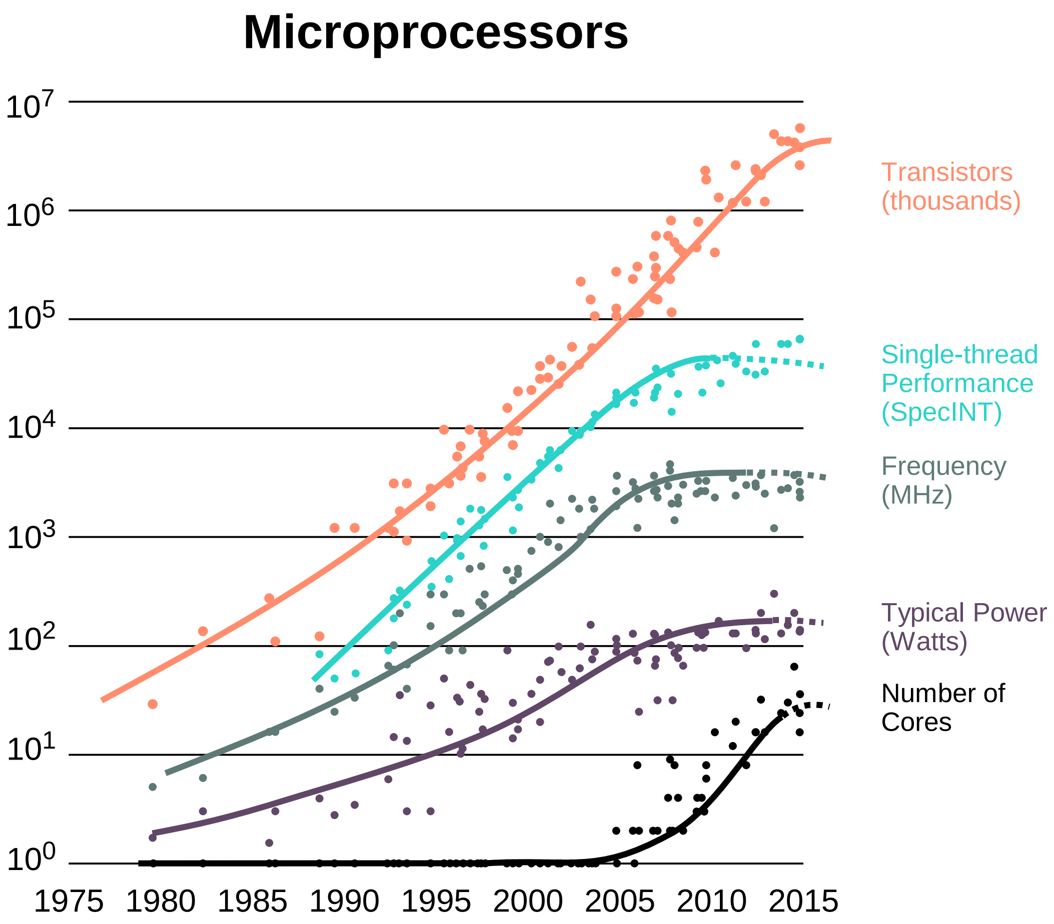
2014
3D Printing Technology

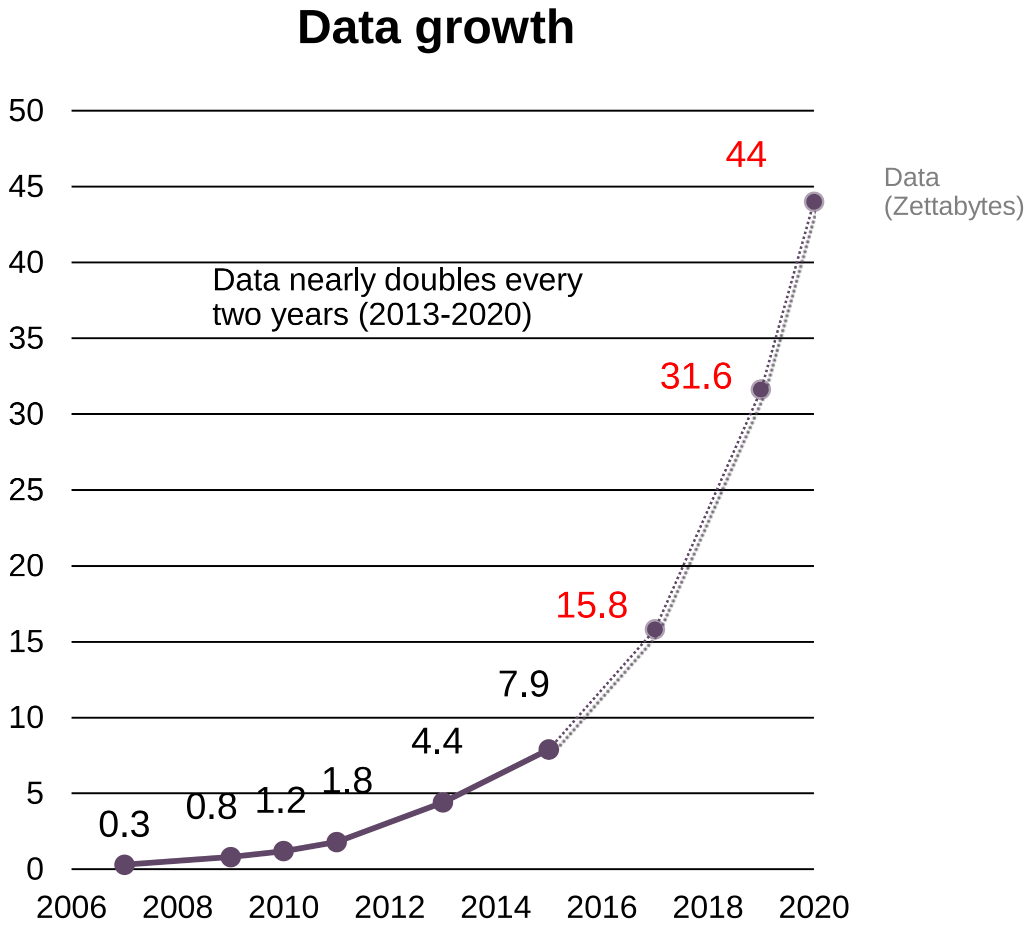
2015
Distributed R



44ZB
DATA
31.6ZB
15.8ZB
COMPUTE
7.9ZB
4.4ZB
0.3ZB
0.8ZB
1.2ZB
1.8ZB
DATA
COMPUTE
‘06
‘08
‘10
‘12
‘14
‘16
‘18
‘20

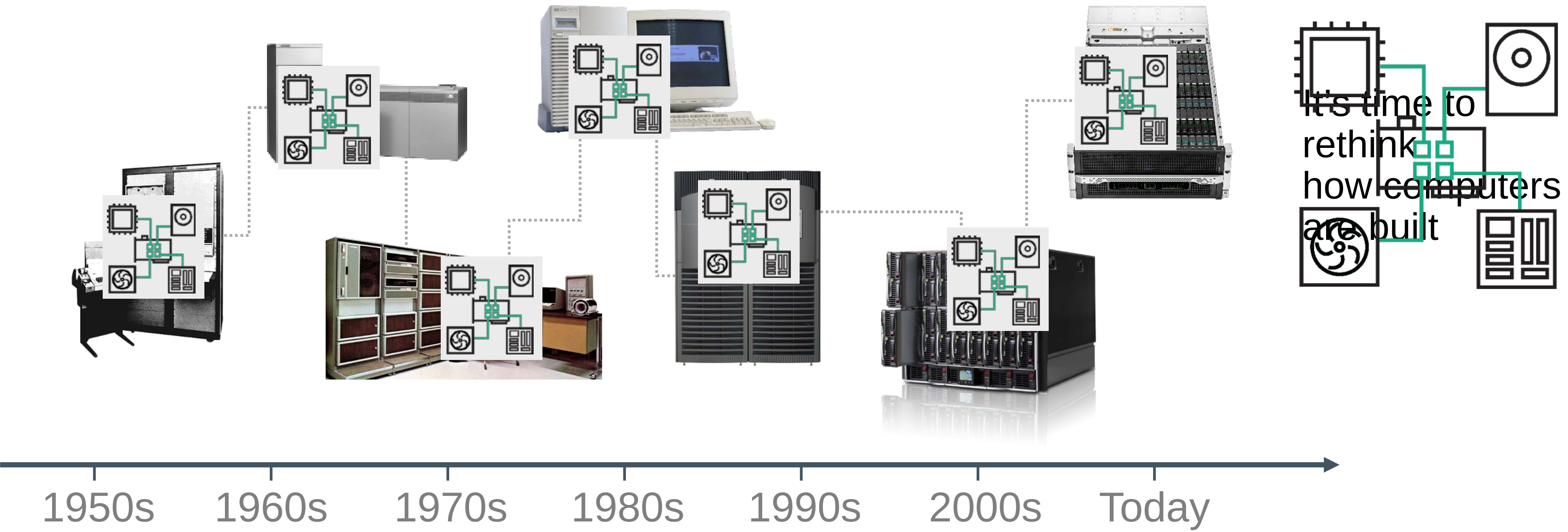
The New Normal: Compute is not keeping up

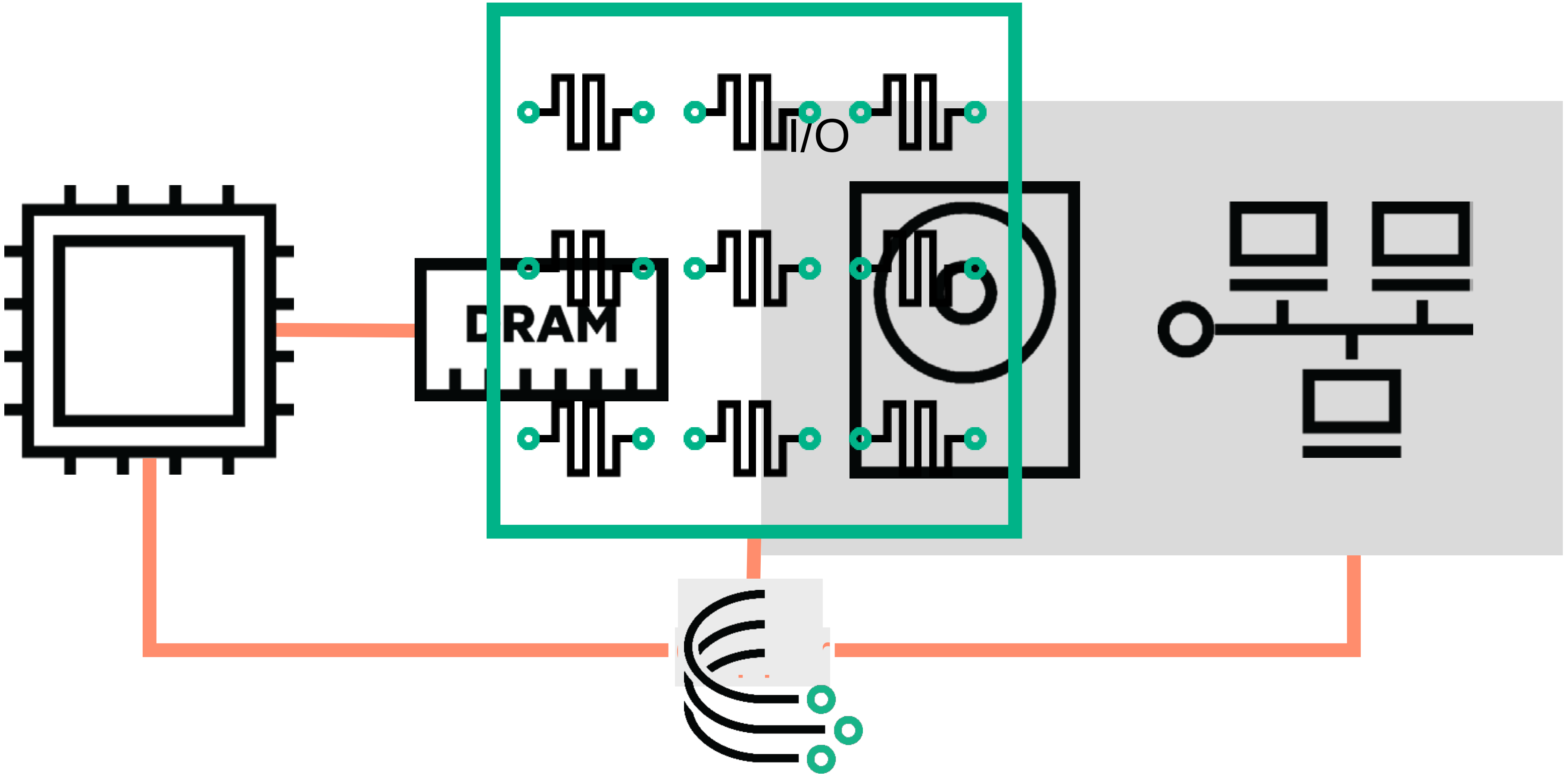




Hewlett Packard Enterprise

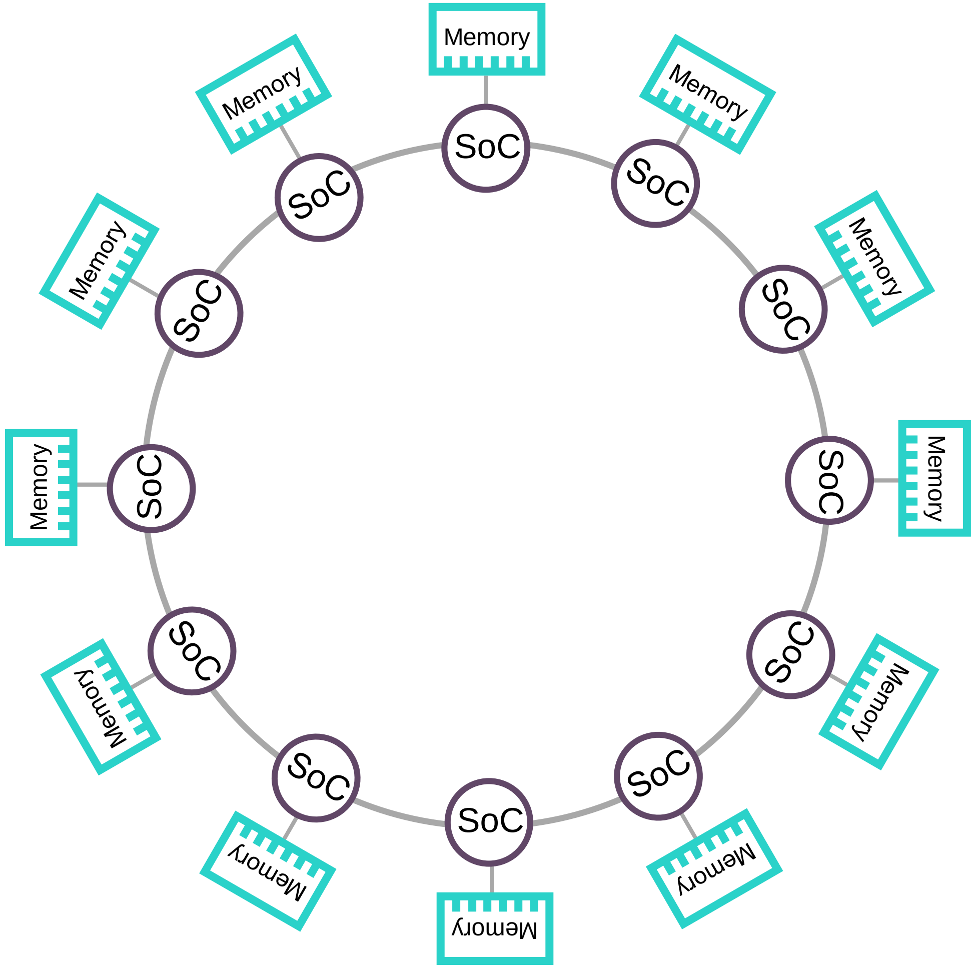
The Past 60 Years




I/O
DRAM

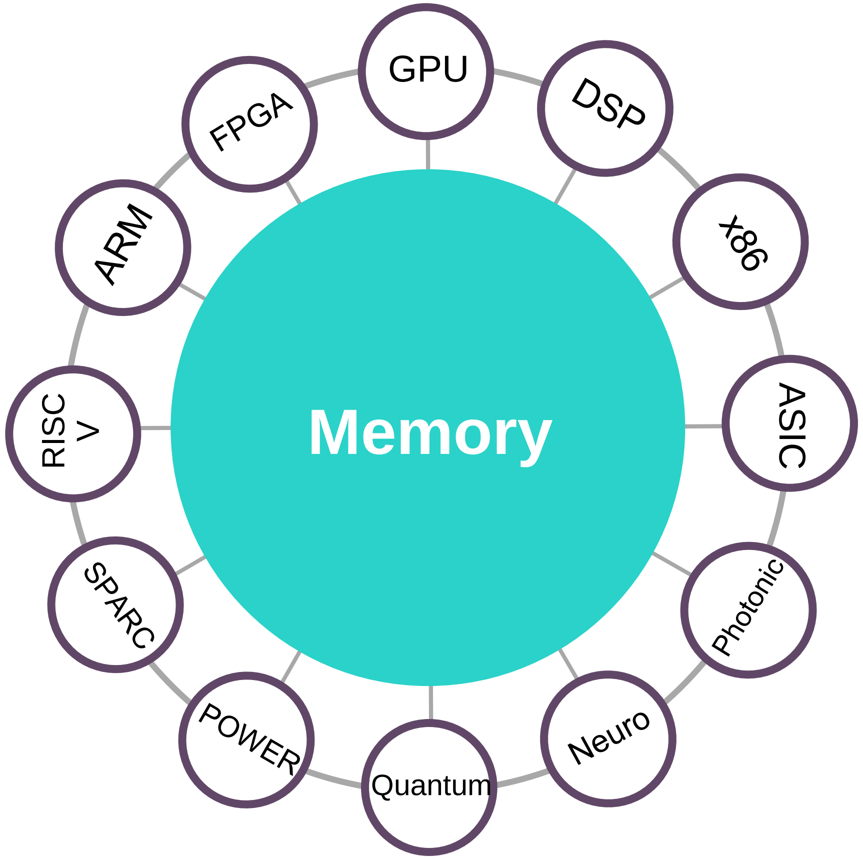
Today's architecture

From processor-centric computing



Future architecture

Memory-Driven Computing



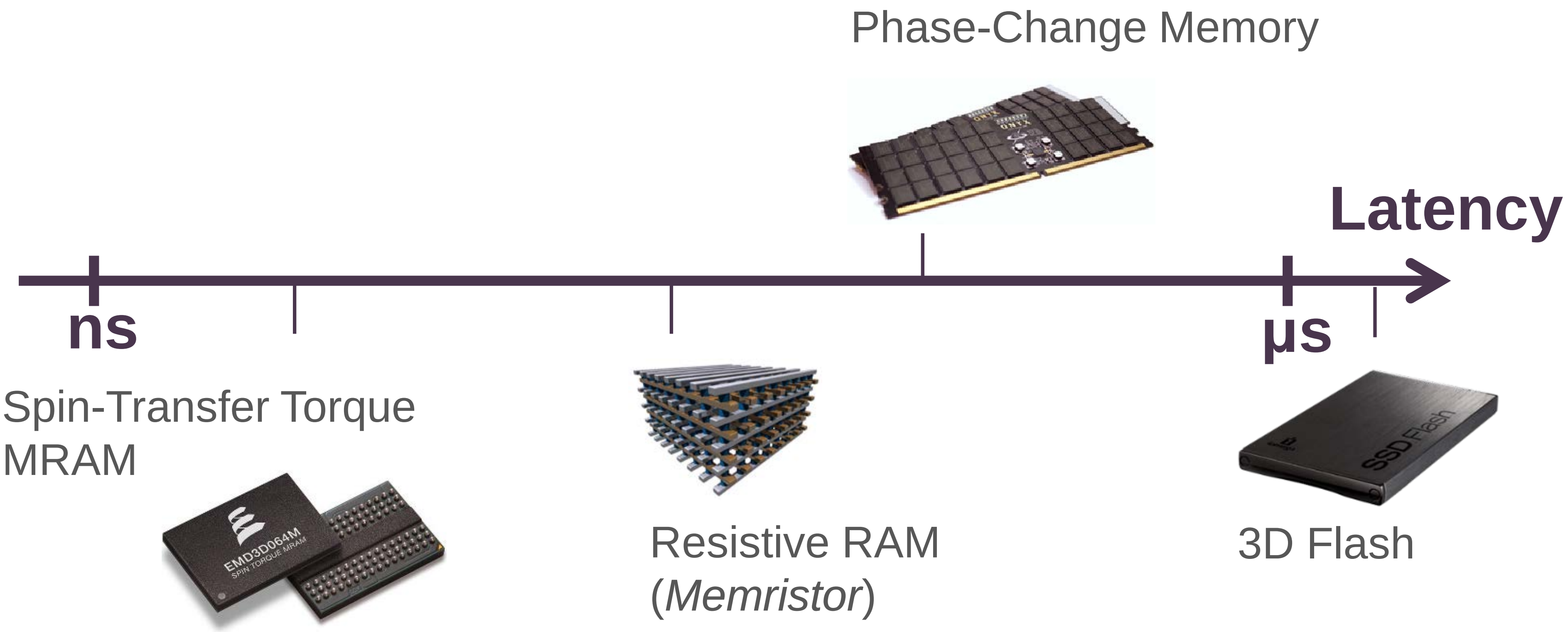
Hewlett Packard Enterprise

Core Memory-Driven Computing components

Fast, persistent memory	Fast memory fabric	Task-specific processing	New software

Hewlett Packard Enterprise

Non-Volatile Memory (NVM)



– Persistently stores data

– Access latencies comparable to DRAM

– Byte addressable (load/store) rather than block addressable (read/write)

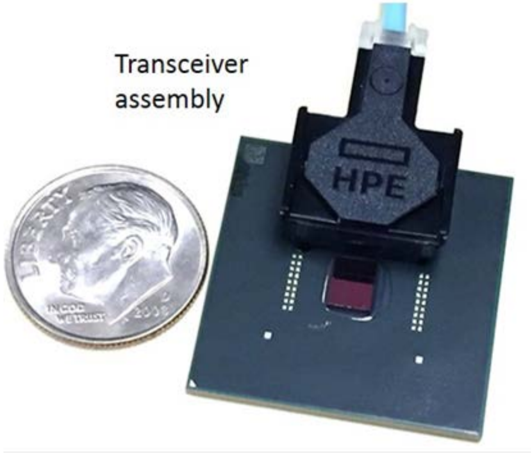
– More energy efficient and denser than DRAM

Haris Volos, et al. "Aerie: Flexible File-System Interfaces to Storage-Class Memory," *Proc. EuroSys 2014*.

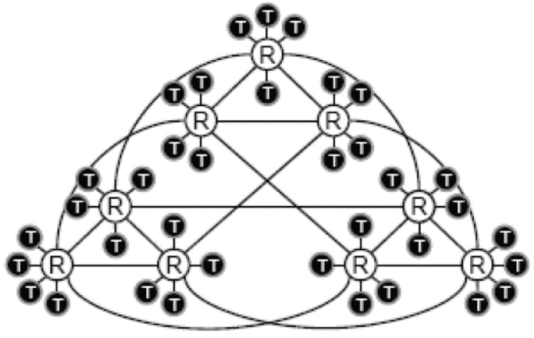
Interconnect advances

- Photonic interconnects
  - Ex: Vertical Cavity Surface Emitting Lasers (VCSELs)
  - 4 λ Coarse Wavelength Division Multiplexing (CWDM)
  - 100Gbps/fiber; 1.2Tbps with 12 fibers
  - Low power ~ < 5pJ/bit (target)
  - Low cost << $1/Gbps

- High-radix switches enable low-diameter network topologies
  - Pooled NVM will appear at near-uniform low latency



VCSEL optics



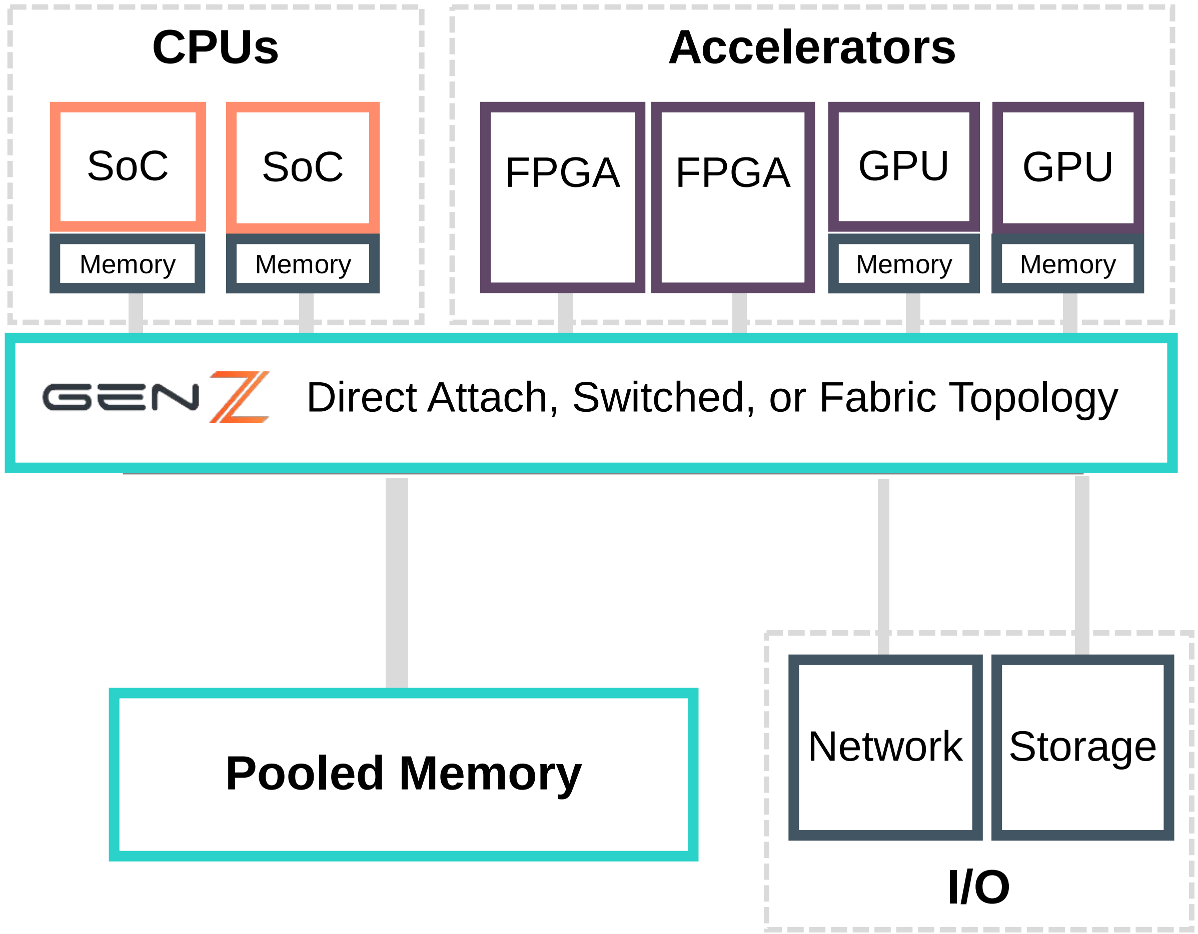
(b) $L = 2, S_1 = 3, S_2 = 3, T = 4$

HyperX topology

Source: J. H. Ahn, et al., "HyperX: topology, routing, and packaging of efficient large-scale networks," *Proc. SC*, 2009.

Gen-Z: open systems interconnect standard
http://www.genzconsortium.org

- Open standard for memory-semantic interconnect
- Members: ~50 companies covering SoC, memory, I/O, networking, mechanical, system software, etc.
- Motivation
  - Emergence of low-latency storage class memory
  - Demand for large capacity, rack-scale resource pools and multi-node architectures
- Memory semantics
  - All communication as memory operations (load/store, put/get, atomics)
- High performance
  - Tens to hundreds GB/s bandwidth
  - Sub-microsecond load-to-use memory latency
- *Spec available for public download*



Hewlett Packard Enterprise

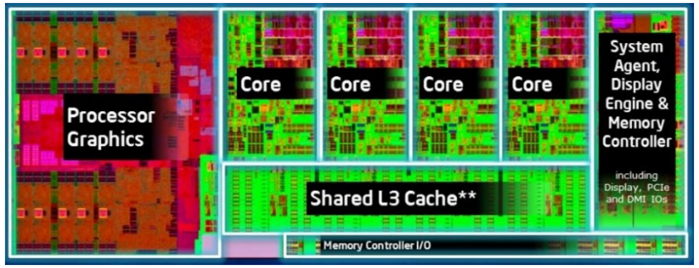
Open Consortium With Broad Industry Support (48)

Gen-Z Consortium Members

- Alpha Data
- AMD
- Amphenol
- ARM
- Avery Design Systems
- Broadcom
- Cadence
- Cavium
- Cray
- Dell EMC
- Everspin
- FIT
- Hirose
- HP Enterprise
- Huawei
- IBM
- IDT
- IntelliProp
- Jabil
- Jess Link
- Keysight
- Lenovo
- Lotes
- Luxshare-ICT
- Mellanox
- Mentor Graphics
- Micron
- Microsemi
- Mobiveil
- Molex
- NetApp
- Nokia
- Numascale
- PLDA Group
- Qualcomm
- Red Hat
- Samsung
- Seagate
- Senko Advanced Comp
- SK hynix
- Smart Modular
- Spin Transfer Tech
- TE
- Toshiba Memory Corp
- VMware
- Western Digital
- Xilinx
- Yadro

Heterogeneous compute

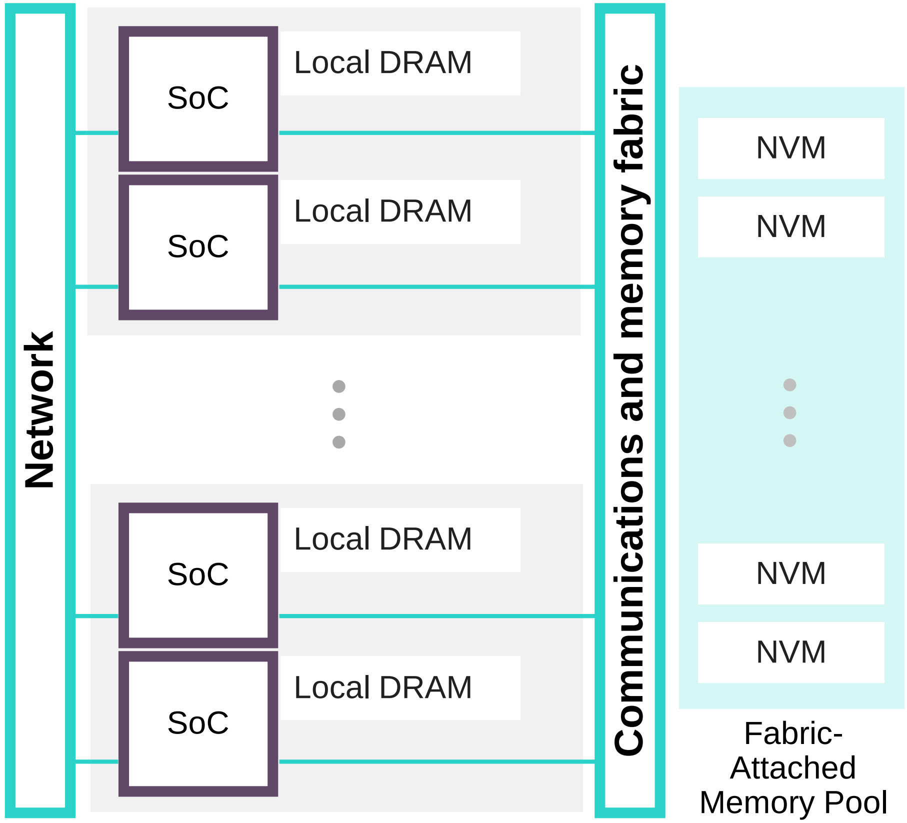
– Dark silicon effects
  – Microprocessor designs are limited by power, not area
  – Solution: combination of function blocks that are selectively activated



– Task-specific accelerators augment CPU compute
  – Examples: GPUs, FPGAs, ASICs
  – Enables higher energy efficiency

HPE Edgeline ProLiant m710x

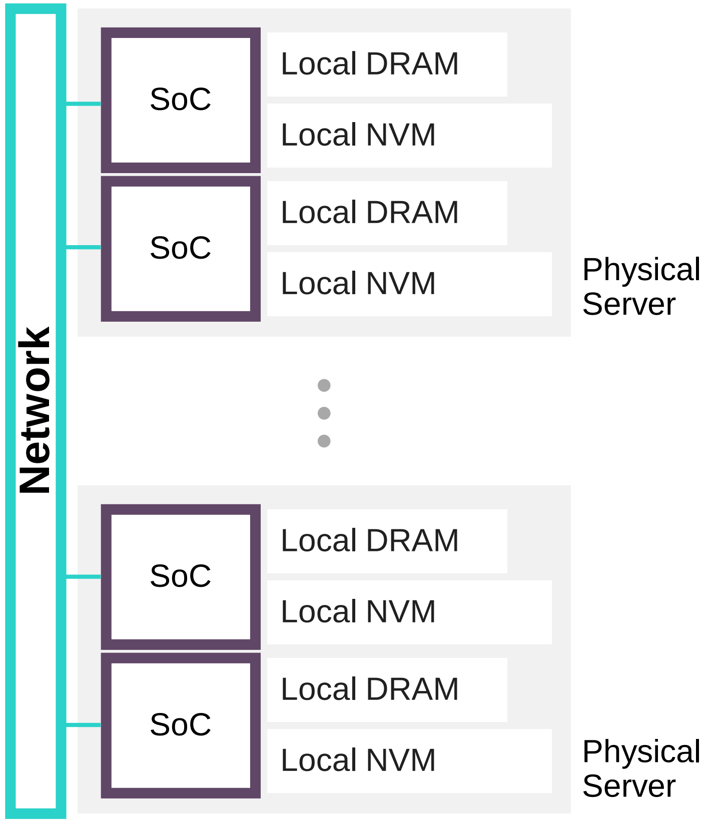
Putting it all together: Memory-Driven Computing



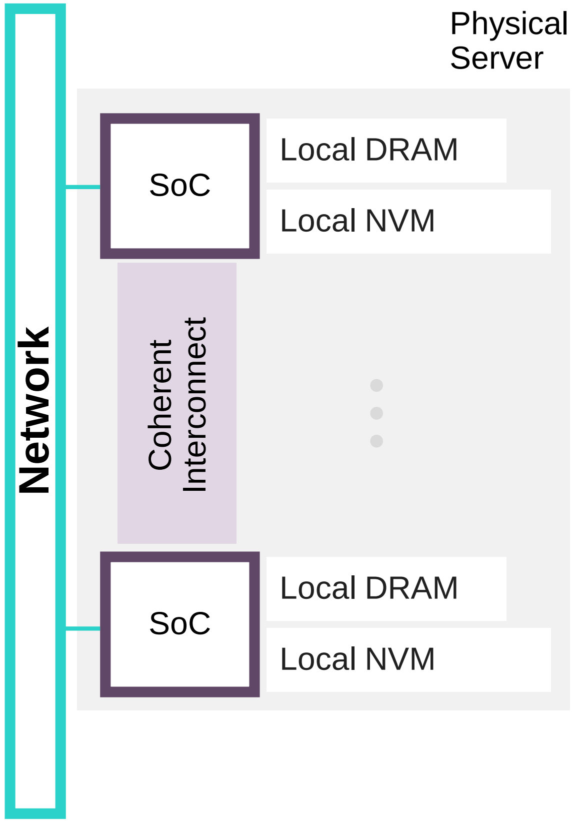
- **Converging memory and storage**
  - Byte-addressable NVM replaces hard drives and SSDs
- **Resource disaggregation leads to high capacity shared memory pool**
  - Fabric-attached memory pool is accessible by all compute resources
  - Low diameter networks provide near-uniform low latency
- **Distributed heterogeneous compute resources**
- **Local volatile memory provides lower latency, high performance tier**
- **Software**
  - Memory-speed persistence
  - Direct, unmediated access to all fabric-attached NVM across the memory fabric
  - Non-coherent accesses between compute nodes

Hewlett Packard Enterprise

Memory-Driven Computing in context

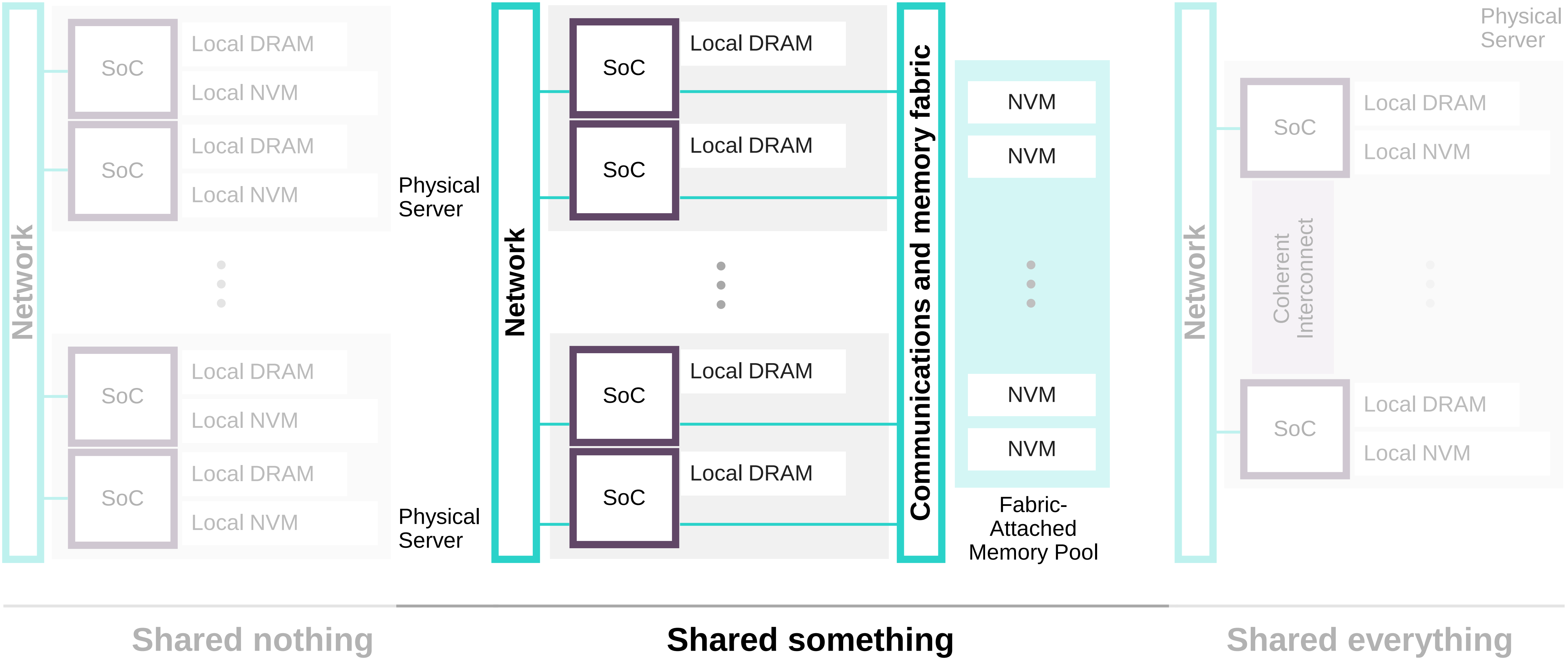


Shared nothing

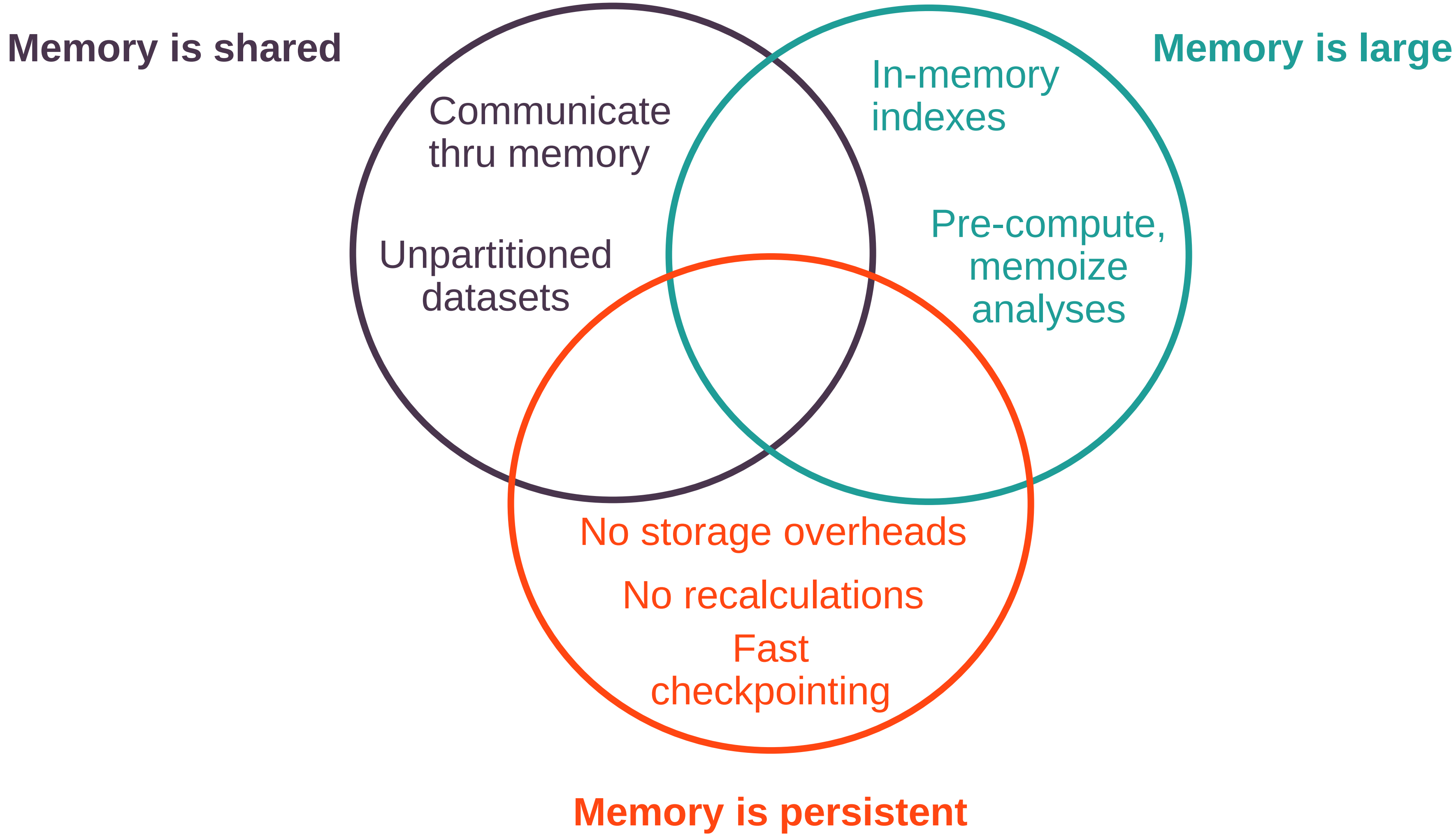


Shared everything

Memory-Driven Computing in context



Memory-Driven Computing benefits applications



HPE introduced the world's largest single-memory computer

The prototype contains 160 terabytes of memory

- 160 TB of shared memory spread across 40 physical nodes, interconnected using a high-performance fabric protocol.

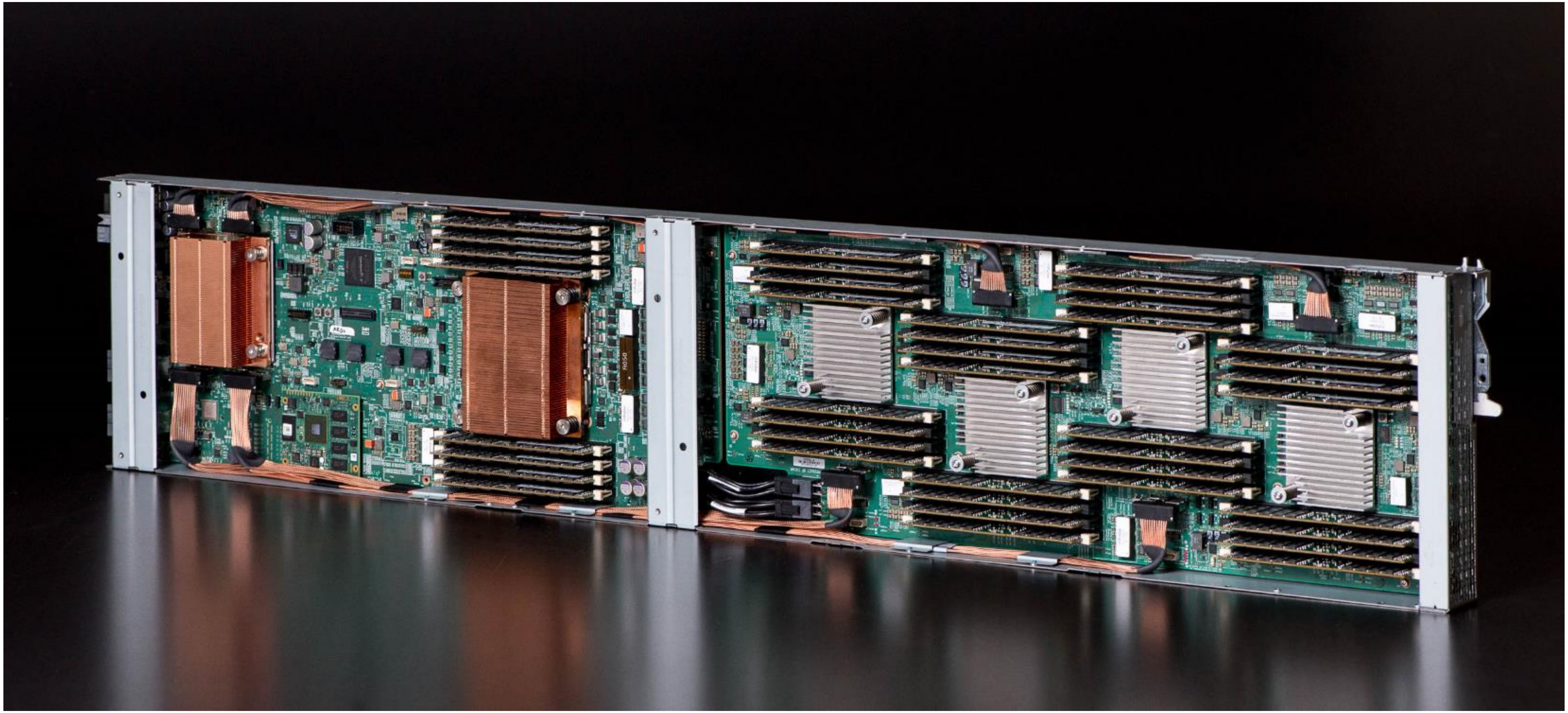
- An optimized Linux-based operating system running on ThunderX2, Cavium's flagship second generation dual socket capable ARMv8-A workload optimized System on a Chip.

- Photonics/Optical communication links, including the new X1 photonics module, are online and operational.

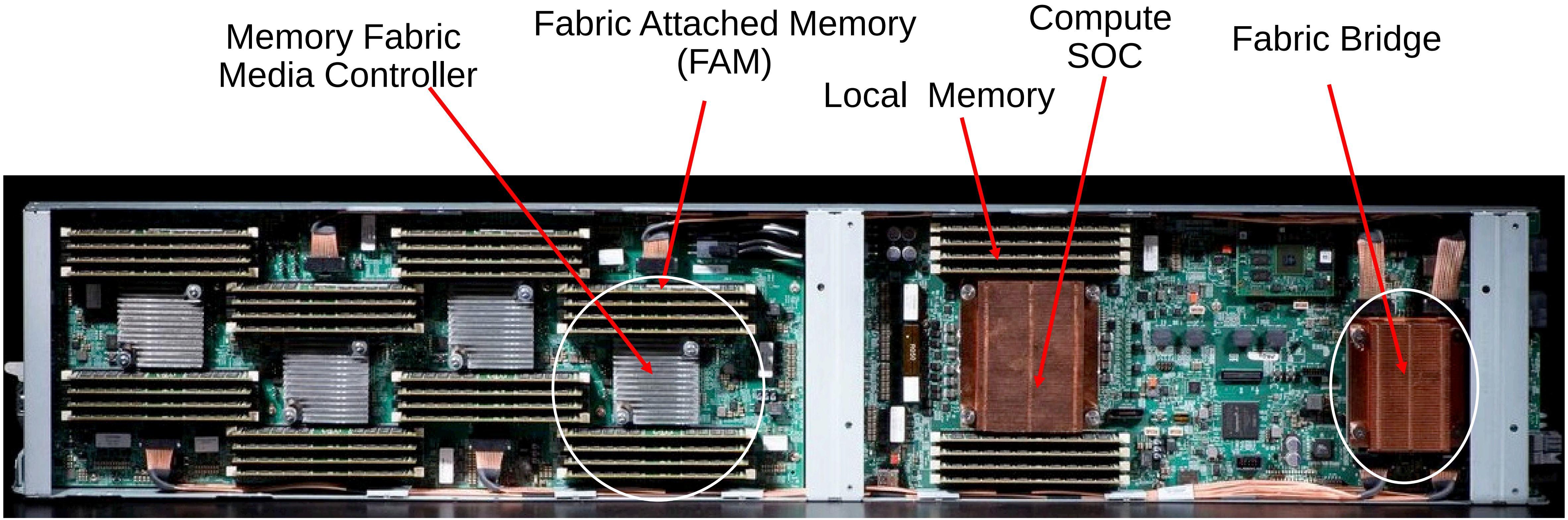
- Software programming tools designed to take advantage of abundant of persistent memory.

Hewlett Packard Enterprise

The Machine program: Memory fabric testbed

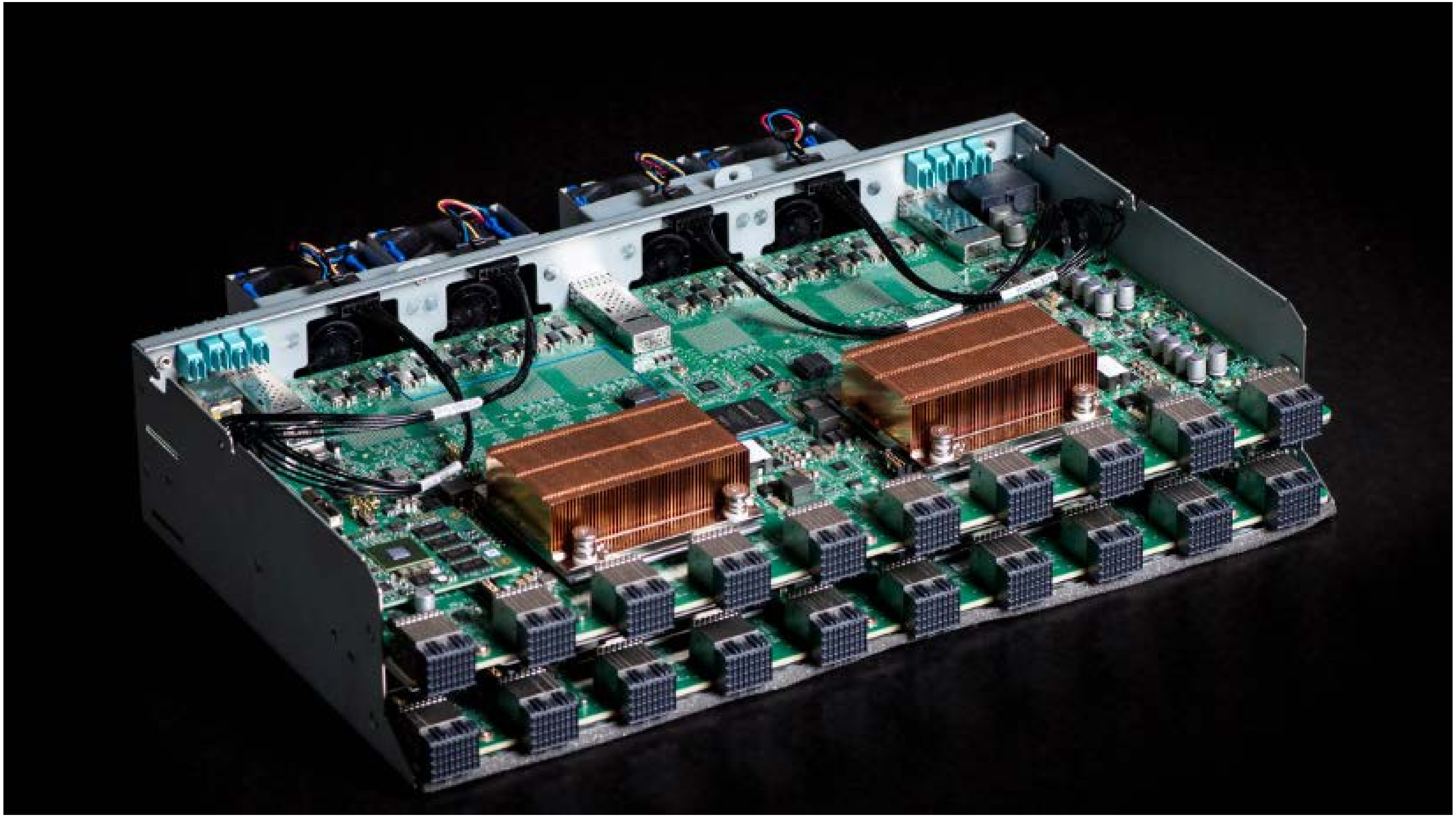


The Machine program: Memory Fabric Testbed

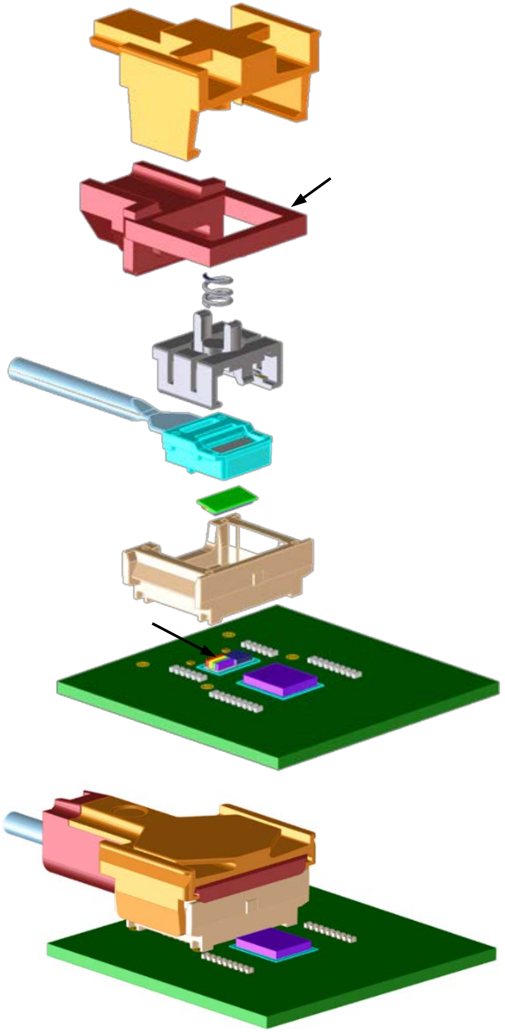


Hewlett Packard
Enterprise

The Machine program: Memory fabric testbed

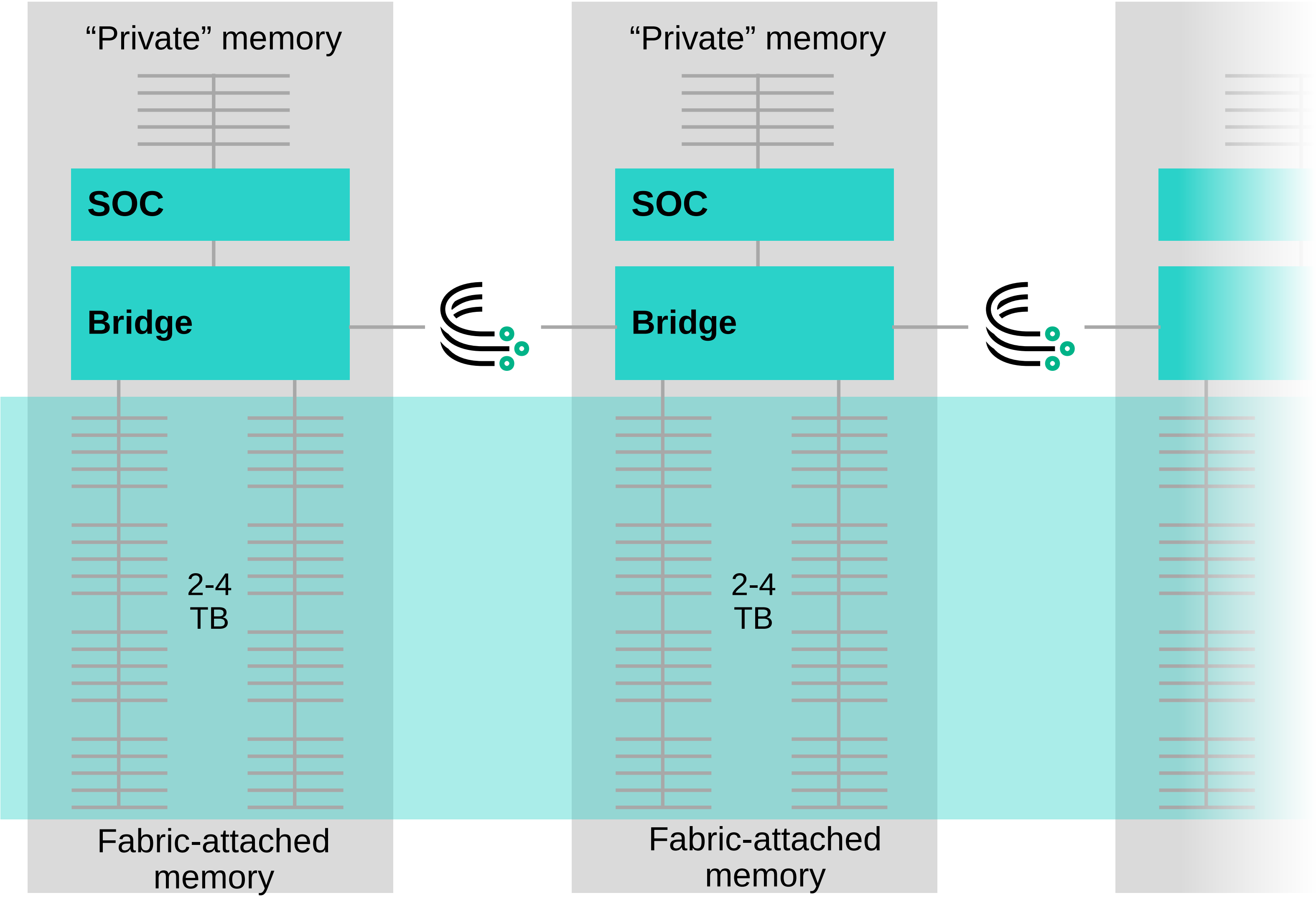


HPE's X1: Fully integrated photonics interconnect chip module



How fabric-attached memory works

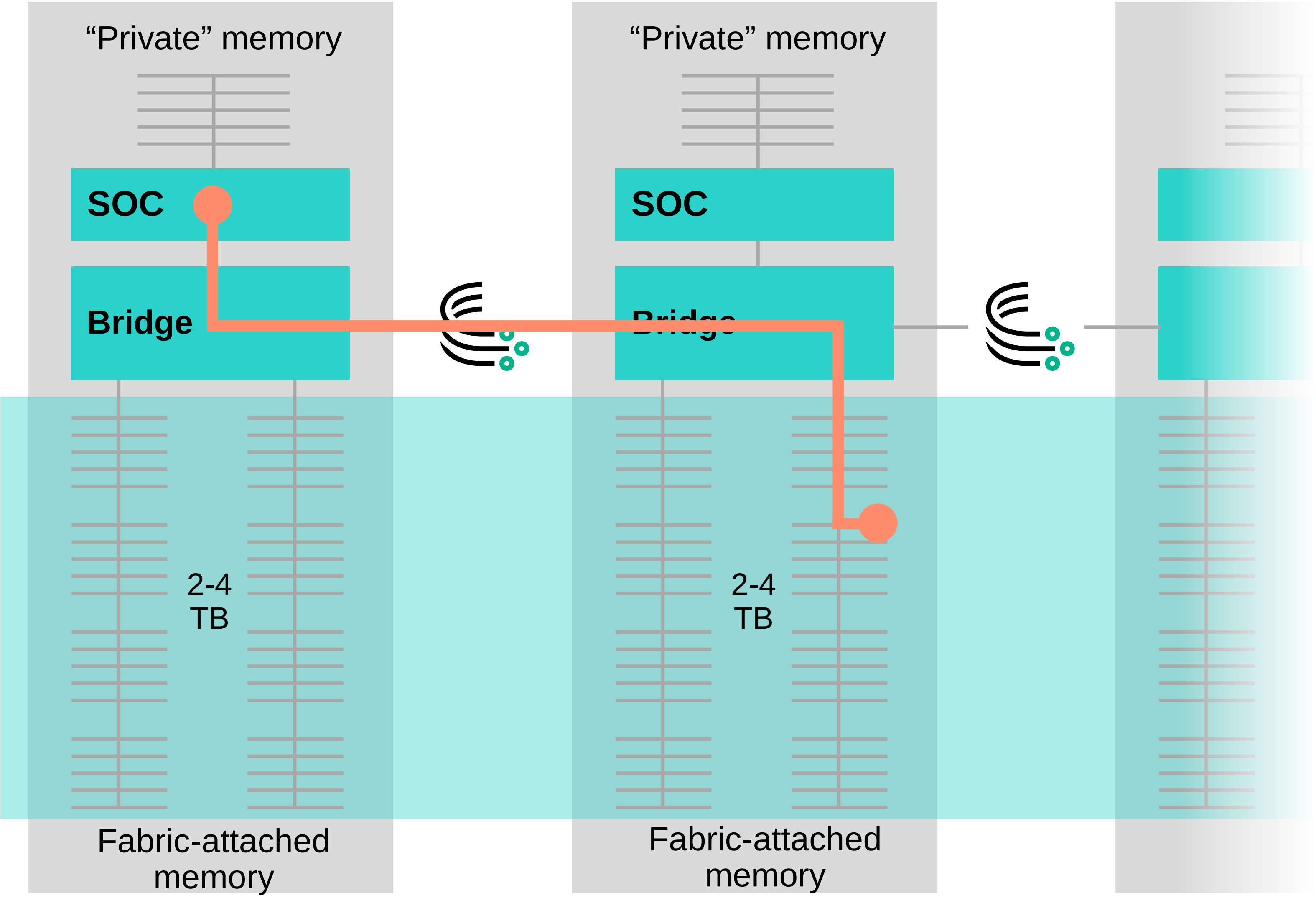
Allows a compute node to access any part of the fabric-attached memory pool



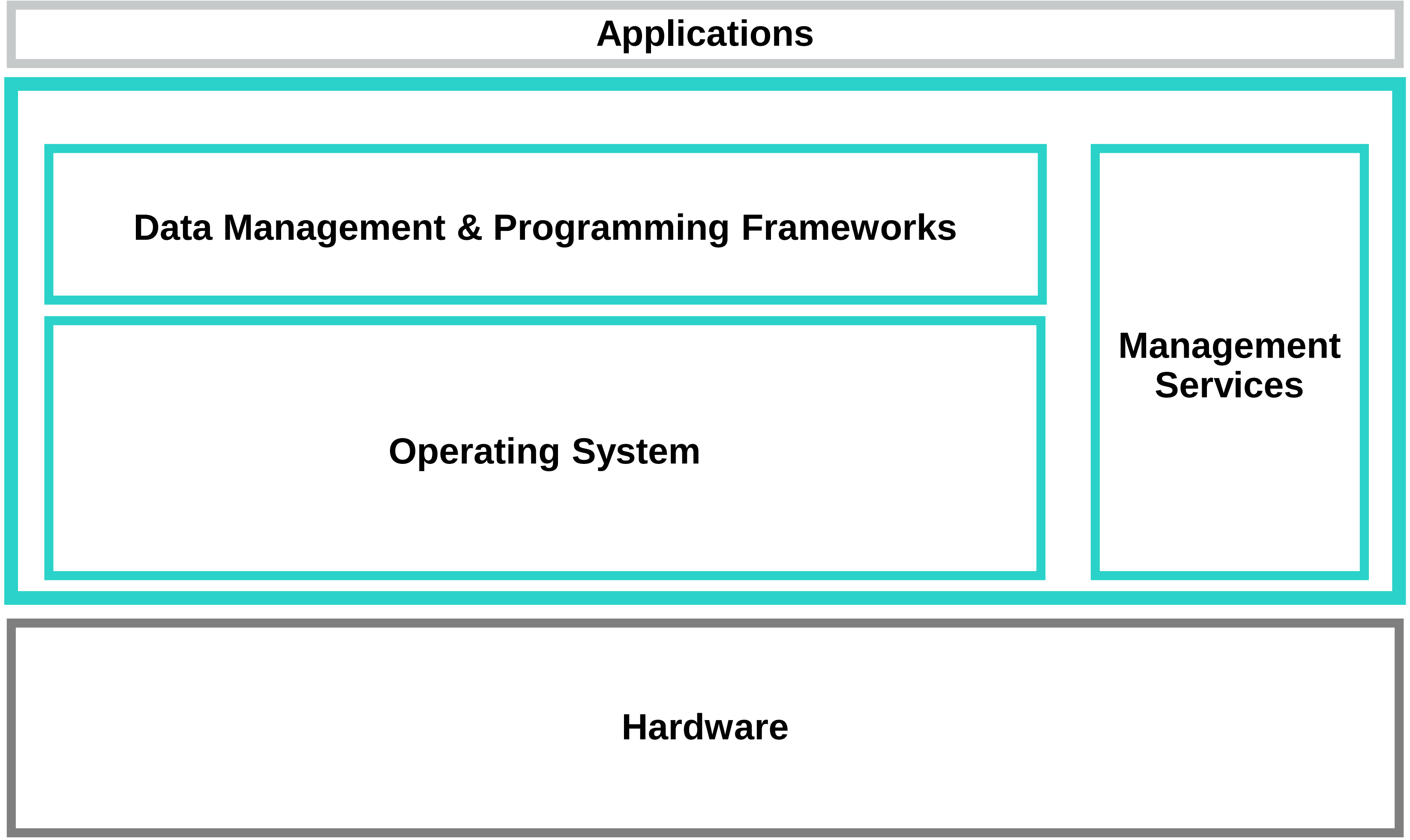
Hewlett Packard Enterprise

How fabric-attached memory works

Allows a compute node to access any part of the fabric-attached memory pool

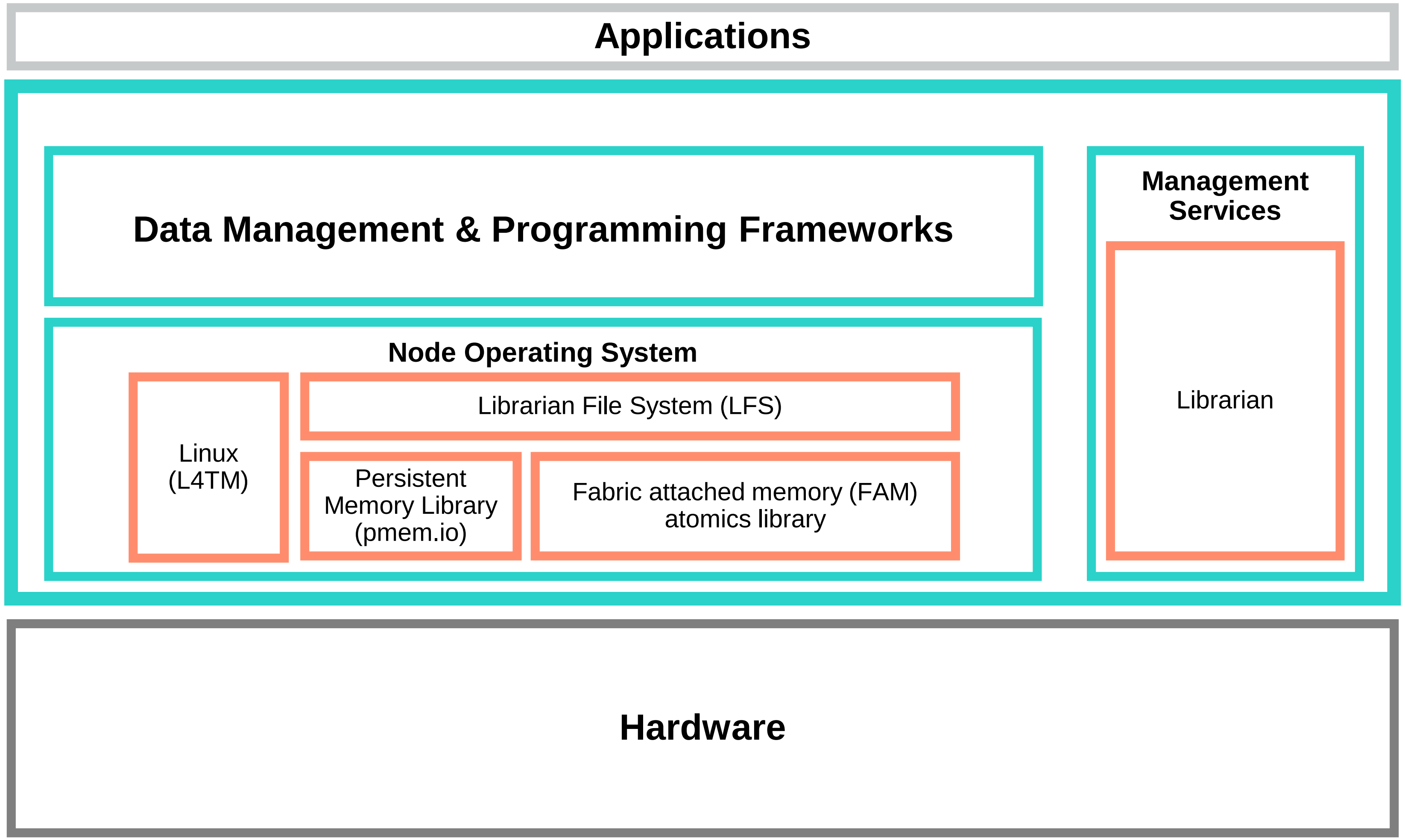


Opportunities to rethink the whole software stack



Hewlett Packard Enterprise

Linux for The Machine



- L4TM: Linux modifications to support fabric-attached persistent memory
- FAM atomics primitives to handle sharing across nodes
- Pmem.io modifications to support non-coherent access
- LFS exposes fabric-attached memory as mmap'd shared FS
- Librarian for cross-node fabric memory allocation

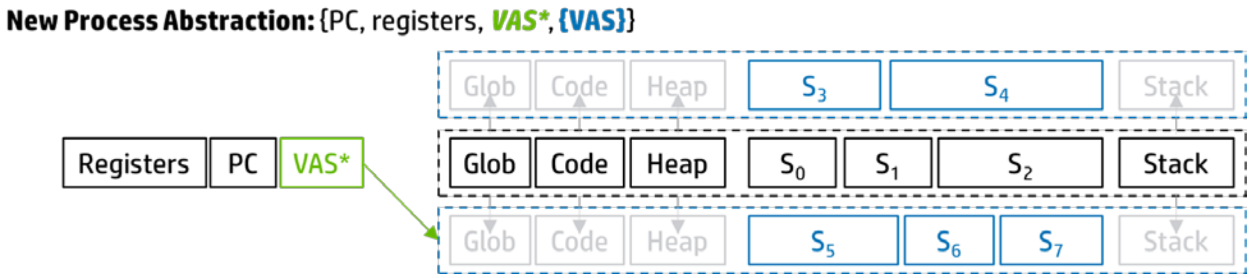
Open sourced components

https://github.com/FabricAttachedMemory

Hewlett Packard Enterprise

SpaceJMP: Programming with Multiple Virtual Address Spaces

- Virtual address space as first-class citizen
- Process can have multiple virtual address spaces



- Efficient safe programming and sharing for huge memories
- Data sharing and communication between processes
- Versioning and checkpointing
- Co-design between OS, programming languages, compilers, and runtimes
- Prototype implementations in BSD, Linux, and Barrelfish

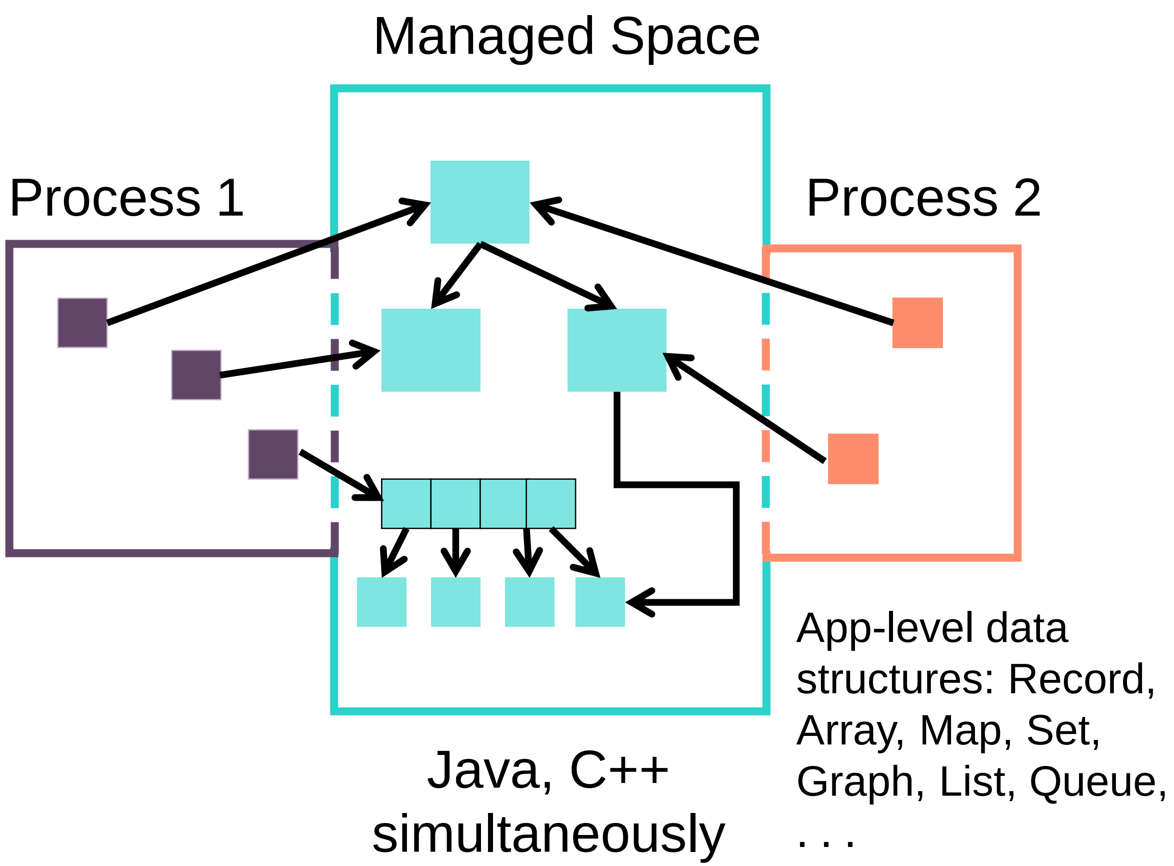
I. El Hajj, et al. “SpaceJMP: Programming with Multiple Virtual Address Spaces,” *Proc. Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2016.

Hewlett Packard Enterprise

Managed Data Structures (MDS)

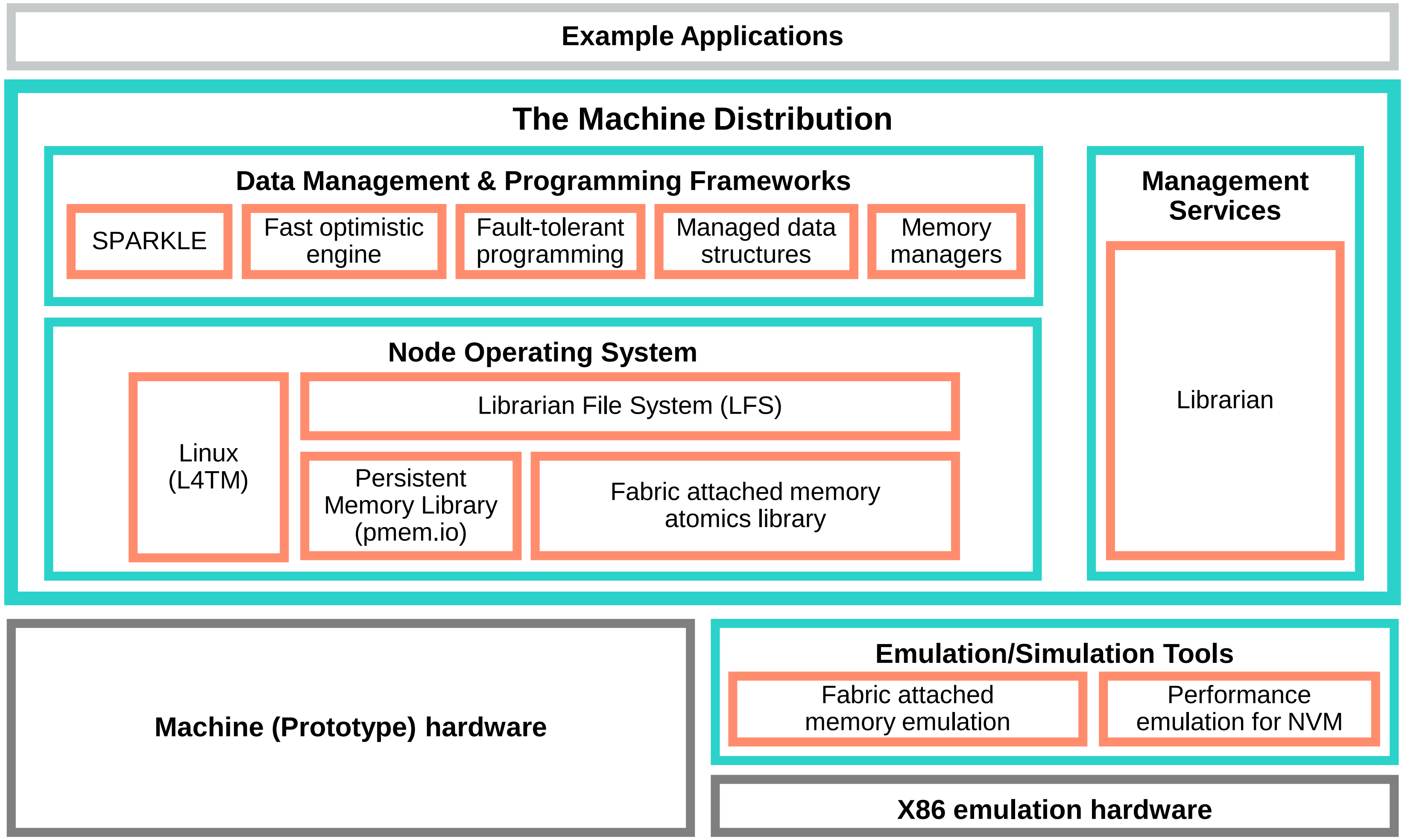
Simplify programming on persistent in-memory data

- Ease of Programming
  - Programmer manages only application-level data structures
    - MDS data structures are automatically persisted in NVM
  - APIs in multiple programming languages: Java, C++
    - Programmer access through references to data
    - Direct reads and writes
- Ease of Data Sharing
  - Just pass a reference
    - Each program treats the data as if it was local to the program
  - High-level concurrency controls
    - Ensure consistent data in the face of data sharing by multiple threads/processes



The Machine Distribution

Software stack for Memory-Driven Computing



Programming and analytics tools

Operating system support

Emulation/simulation tools

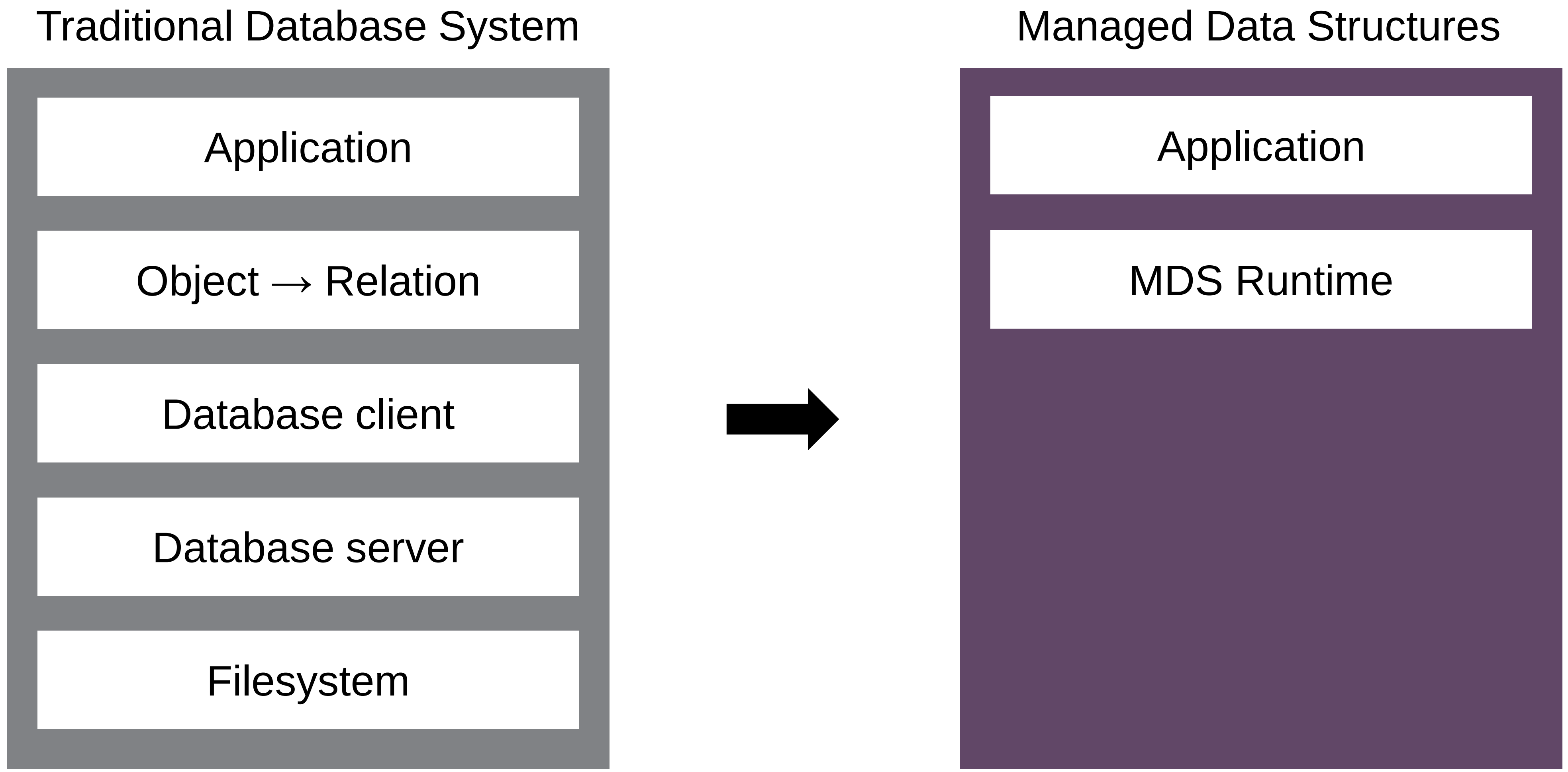
Open sourced components

https://github.com/HewlettPackard/mdc-toolkit

https://www.labs.hpe.com/the-machine/the-machine-distribution

Hewlett Packard Enterprise

Fewer software layers



Open source code at https://github.com/HewlettPackard/mds

Research publication highlights...

– R. Achermann, C. Dalton, P. Faraboschi, M. Hoffman, D. Milojicic, G. Ndu, A. Richardson, T. Roscoe, A. Shaw, R. Watson. “Separating Translation from Protection in Address Spaces with Dynamic Remapping,” *Proc. 16th Workshop on Hot Topics in Operating Systems (HotOS XVI)*, 2017.
– T. Hsu, H. Brugner, I. Roy, K. Keeton, P. Eugster. “NVthreads: Practical Persistence for Multi-threaded Applications,” *Proc. ACM EuroSys*, 2017.
– S. Nalli, S. Haria, M. Swift, M. Hill, H. Volos, K. Keeton. ”An Analysis of Persistent Memory Use with WHISPER,” *Proc. ACM Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2017.
– H. Kimura, A. Simitsis, K. Wilkinson, “Janus: Transactional processing of navigational and analytical graph queries on many-core servers,” *Proc. CIDR*, 2017.
– F. Chen, M. Gonzalez, K. Viswanathan, H. Laffitte, J. Rivera, A. Mitchell, S. Singhal. “Billion node graph inference: iterative processing on The Machine,” Hewlett Packard Labs Technical Report HPE-2016-101, December 2016.
– P. Laplante and D. Milojicic. "Rethinking operating systems for rebooted computing," *Proc. IEEE International Conference on Rebooting Computing (ICRC)*, 2016.
– D. Chakrabarti, H. Volos, I. Roy, and M. Swift. “How Should We Program Non-volatile Memory?”, tutorial at *ACM Conf. on Programming Language Design and Implementation (PLDI)*, 2016.
– K. Viswanathan, M. Kim, J. Li, M. Gonzalez. “A memory-driven computing approach to high-dimensional similarity search,” Hewlett Packard Labs Technical Report HPE-2016-45, May 2016.
– N. Farooqui, I. Roy, Y. Chen, V. Talwar, and K. Schwan. “Accelerating Graph Applications on Integrated GPU Platforms via Instrumentation-Driven Optimization,” *Proc. ACM Conf. on Computing Frontiers (CF’16)*, May 2016.
– I. El Hajj, A. Merritt, G. Zellweger, D. Milojicic, W. Hwu, K. Schwan, T. Roscoe, R. Achermann, P. Faraboschi. “SpaceJMP: Programming with multiple virtual address spaces,” *ASPLOS*, 2016.
– J. Izraelevitz, T. Kelly, A. Kolli. “Failure-atomic persistent memory updates via JUSTDO logging,” *Proc. ACM ASPLOS*, 2016.
– D. Milojicic, T. Roscoe. “Outlook on Operating Systems,” *IEEE Computer*, January 2016.
– K. Bresniker, S. Singhal, and S. Williams. “Adapting to thrive in a new economy of memory abundance,” *IEEE Computer*, December 2015.
– H. Volos, G, Magalhaes, L, Cherkasova, J, Li. “Quartz: A lightweight performance emulator for persistent memory software,” *Proc. of ACM/USENIX/IFIP Conference on Middleware*, 2015.
– J. Li, C. Pu, Y. Chen, V. Talwar, and D. Milojicic. “Improving Preemptive Scheduling with Application-Transparent Checkpointing in Shared Clusters,” *Proc. Middleware*, 2015.
– H. Kimura. “FOEDUS: OLTP engine for a thousand cores and NVRAM,” *Proc. ACM SIGMOD*, 2015.
– P. Faraboschi, K. Keeton, T. Marsland, D. Milojicic. “Beyond processor-centric operating systems,” *Proc. HotOS XV*, 2015.
– S. Gerber, G. Zellweger, R. Achermann, K. Kourtis, and T. Roscoe, D. Milojicic. “Not your parents’ physical address space,” *Proc. HotOS,* 2015.
– F. Nawab, D. Chakrabarti, T. Kelly, C. Morrey III. “Procrastination beats prevention: Timely sufficient persistence for efficient crash resilience,” *Proc. Conf. on Extending Database Technology (EDBT)*, 2015.
– S. Novakovic, K. Keeton, P. Faraboschi, R. Schreiber, E. Bugnion. “Using shared non-volatile memory in scale-out software,” *Proc. ACM Workshop on Rack-scale Computing (WRSC)*, 2015.
– M. Swift and H. Volos. “Programming and usage models for non-volatile memory,” Tutorial at *ACM ASPLOS*, 2015.
– D. Chakrabarti, H. Boehm and K. Bhandari. “Atlas: Leveraging locks for non-volatile memory consistency,” *Proc. ACM Conf. on Object-Oriented Programming, Systems, Languages & Applications (OOPSLA)*, 2014.
– H. Volos, S. Nalli, S. Panneerselvam, V. Varadarajan, P. Saxena, M. Swift. "Aerie: Flexible file-system interfaces to storage-class memory," *Proc. EuroSys*, 2014.

Memory-Driven Computing – Driving innovation to product

Scale	MDC testbed	Rack scale MDC development continues	Exascale computing research Commercial petascale systems	Prototype exascale systems MDC Edge computing
Phases	**Realized**	**Just Realized** (2017)	**Near to Longer Term** (2018-2019)	**Future State** (2020)
Non-volatile memory	**DRAM-based persistent memory technology launched** – Significant performance gains with today's apps	**Extended DRAM-based persistence** – Build on performance gains	**True non-volatile memory** – Enabling high-performance data-intensive analytics	**Non-volatile memory realized** – Used across multiple product categories
Fabric	**Demonstrated photonic interconnects** – Low-cost, high-performance – Future-proofing for HPE Synergy	**Select product integration** – Photonics enablement – Data Fabric for software-defined storage across any system	**Extending photonics** – Storage fabrics – Fabric-attached memory	**Photonics for short and long-distance applications realized** **Gen-Z fabric from Edge to data center**
Ecosystem enablement	**Building community for Memory-Driven Compute** – Joined Gen-Z consortium – Demonstrated improved performance using MDC software	**Large-scale MDC proof points** – Impressive performance gains using Memory-Driven software **MDC dev toolkit in open source** – Building developer community	**Next-gen analytics and applications** **MDC Ecosystem Thriving in Open Source**	**Memory-Driven Computing ubiquitous** **Gen-Z ecosystem established**
Security	**Secure fabrics and E2E integrity** – Gen-Z: security built-in, not bolted-on – Memory fabric with data encrypted in flight and at rest demonstrated	**Security, a first class requirement** – Systems with built-in security, data integrity, and resiliency – Integrity assurance, custom recovery, app integration	**Secure containers** – Security and agility for rapid app development	**Scalable security from Edge to Core** – Self-protecting data

How Memory-Driven Computing benefits applications

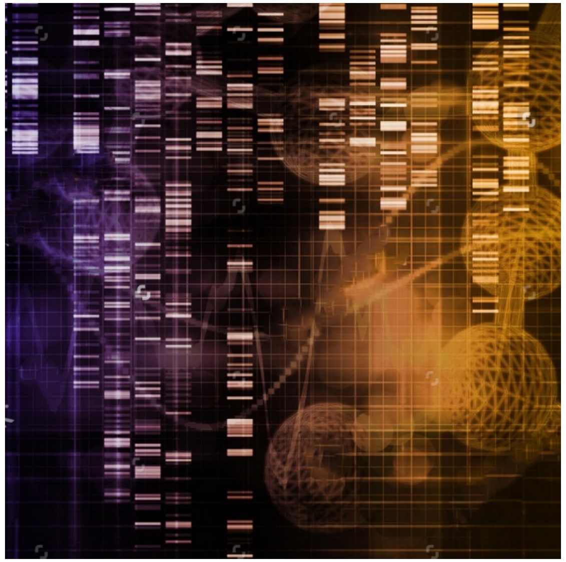
Hewlett Packard
Enterprise

Transform performance with Memory-Driven programming

Modify existing frameworks

New algorithms

Completely rethink



In-memory analytics — **Similarity search** — **Large-scale graph inference** — **Financial models**

In-memory analytics	Similarity search	Large-scale graph inference	Financial models
15x faster	**40x** faster	**100x** faster	**10,000x** faster

How Memory-Driven Computing influences HPE business and customers

Hewlett Packard Enterprise

U.S Department of Energy works with HPE to design a Memory-Driven SuperComputer

- Develop a reference design for an exascale supercomputer that will enable a broad set of modeling and simulation applications unachievable today

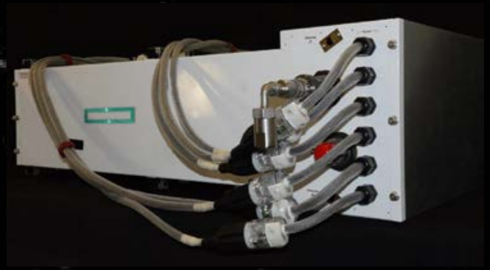
- Accelerating breakthroughs in science, medicine, technology, engineering and many other fields.

- Scientific applications would impact nearly every corner of research, from the physics of star explosions to precision medicine for cancer.

“We see this DOE grant as a vote of confidence in the ability of HPE and Hewlett Packard Labs to help overcome daunting technology challenges that are impeding everyone’s progress toward exascale computing,” - Steve Conway, IDC research vice president of high performance computing

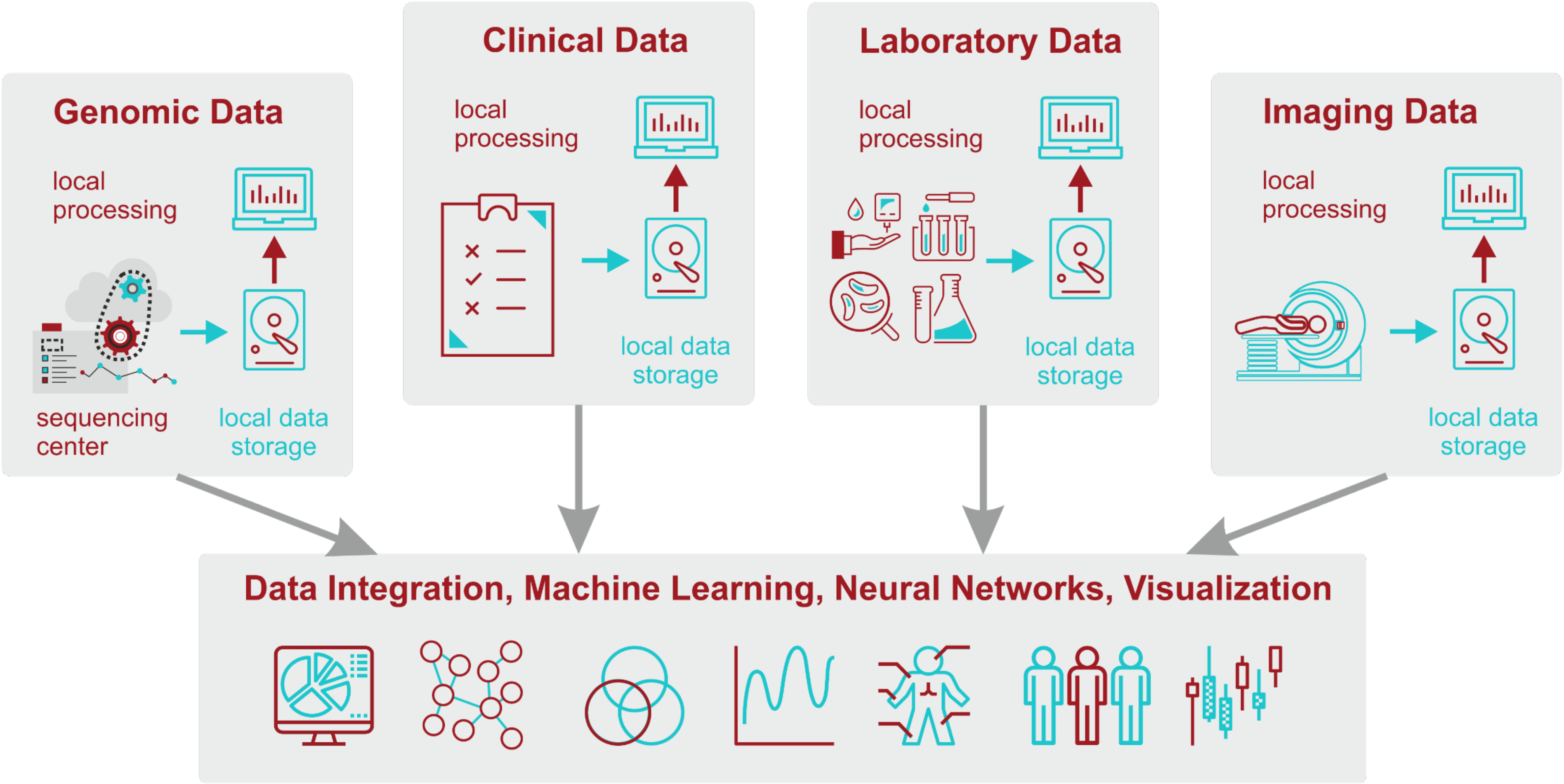
HPE Spaceborne Supercomputer to accelerate mission to Mars

The first step in developing energy efficient and environmentally sophisticated onboard computing resources for extended space travel



Spaceborne Computer – similar compute nodes as NASA's premier supercomputer, Pleiades

What we envision for one DZNE site



Memory-Driven Computing will help us to
- integrate different medical data locally

Memory-Driven Computing helps outpace the global time bomb of neurodegenerative disease

DZNE
Deutsches Zentrum für Neurodegenerative Erkrankungen in der Helmholtz-Gemeinschaft

DZNE discovered HPE’s Memory-Driven Computing — and saw unprecedented computational speed improvements that hold new promise in the race against Alzheimer’s

60%
power reduction cuts research costs

101x
increase in analytics speed blasts research bottlenecks, leading to shorter processing time — from 22 minutes to 13 seconds

HPE Superdome Flex

Turn critical data into real-time business insights

Turn data into actionable insights in real time

- Unparalleled scale 4-32 sockets, 768GB-48TB+ memory
- Highly expandable for growth ultra fast fabric

Keep pace with evolving business demands

- Unique modular 4-socket building block, 45% lower cost at 4s entry point
- Open management and hard partitioning for hybrid IT consumption

Safeguard mission-critical workloads

- Proven Superdome RAS with 99.999% single system availability
- Mission critical expertise with HPE Pointnext services

Designed with Memory-Driven Computing principles

Advancing the real-time enterprise journey

Empowering the data-driven enterprise...

SAP HANA in-memory

Project “Kraken”
Co-innovation with SAP

Extreme scalability for transactions & analytics

Superdome X

High performance Data analytics

Acquisition sgi

Platform for the next decade of in-memory computing

NEW

HPE Superdome Flex

Real-time Enterprise

Memory-Driven Computing advancements

the MACHINE

.... Accelerating real-time business insights

Hewlett Packard Enterprise

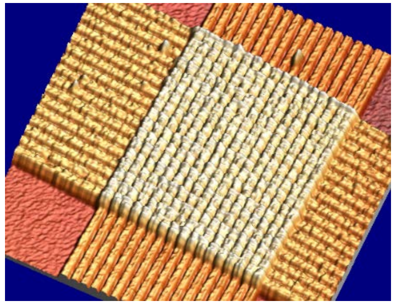
How to get started with Memory-Driven Computing

- To get the latest updates on The Machine project and Memory-Driven Computing, visit www.hpe.com/themachine
- Join The Machine User Group at https://www.labs.hpe.com/the-machine/user-group
  - For community discussions, sign up to our Slack group #themachineusergroup channel at https://www.labs.hpe.com/slack
  - Subscribe to "The Machine User Group" tab in the "Behind the scenes @ Labs" blog https://community.hpe.com/t5/Behind-the-scenes-Labs/bg-p/BehindthescenesatLabs/label-name/The%20Machine%20User%20Group#.WXZGN4jyscE. Register and click "subscribe to this label".
  - Questions? Contact themachineusergroup@hpe.com
- Get access to the Memory-Driven Computing Developer Toolkit at https://www.labs.hpe.com/the-machine/developer-toolkit
- Follow us on our Hewlett Packard Labs social handles:
  - Twitter: @HPE_Labs
  - LinkedIn: "Hewett Packard Labs"
  - "Hewlett Packard Enterprise" YouTube page – The Machine and Hewlett Packard Labs channels
  - Instagram: HPE
  - Facebook: Hewlett Packard Enterprise

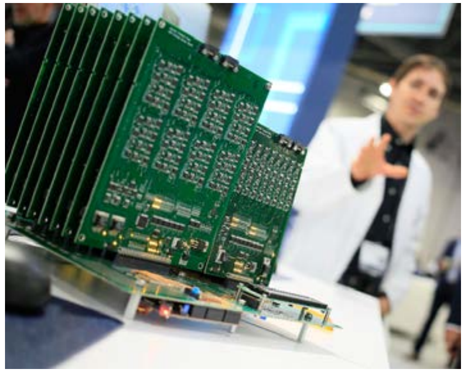
Beyond Moore’s Law

Further into the future: unconventional accelerators

Neuromorphic computing:
Dedicated hardware for brainlike computing

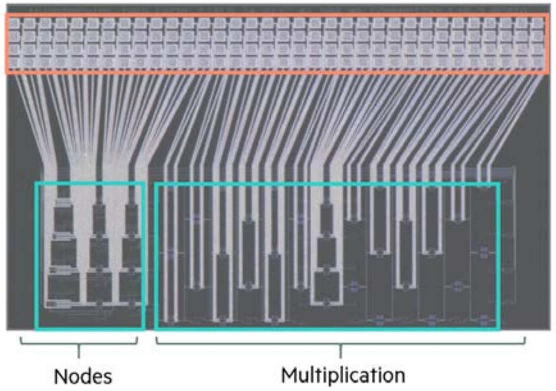


– Neuromorphic computing can **quickly handle tasks that take trained computers several hours**



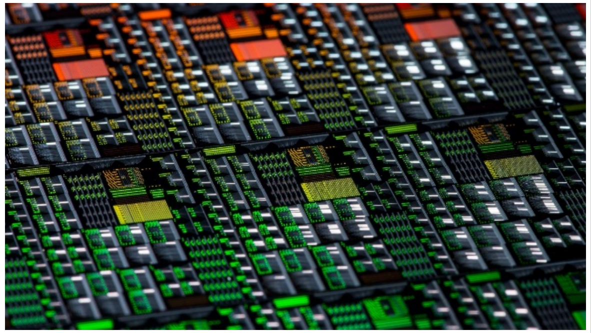
– **Dot Product Engine is our testbed using vector-matrix multiplication** and studying which algorithms and applications benefit the most from using this speedup architecture

Optical Computing:
Computing at the speed of light



– Pushing limits of photonic chip design

– Pushing complex computations through light to boost speed and save energy



– Typical circuits are <10 components. **We’re integrating over 1,000 optical parts in a chip** – the largest photonic components working together to compute.

Thank you!