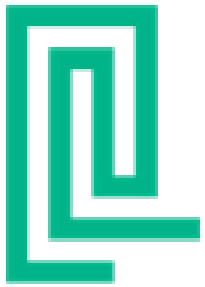


Hewlett Packard
Labs

The Future of Extreme Scale Computing

– Patrick.demichel@hpe.com

– Distinguished Technologist

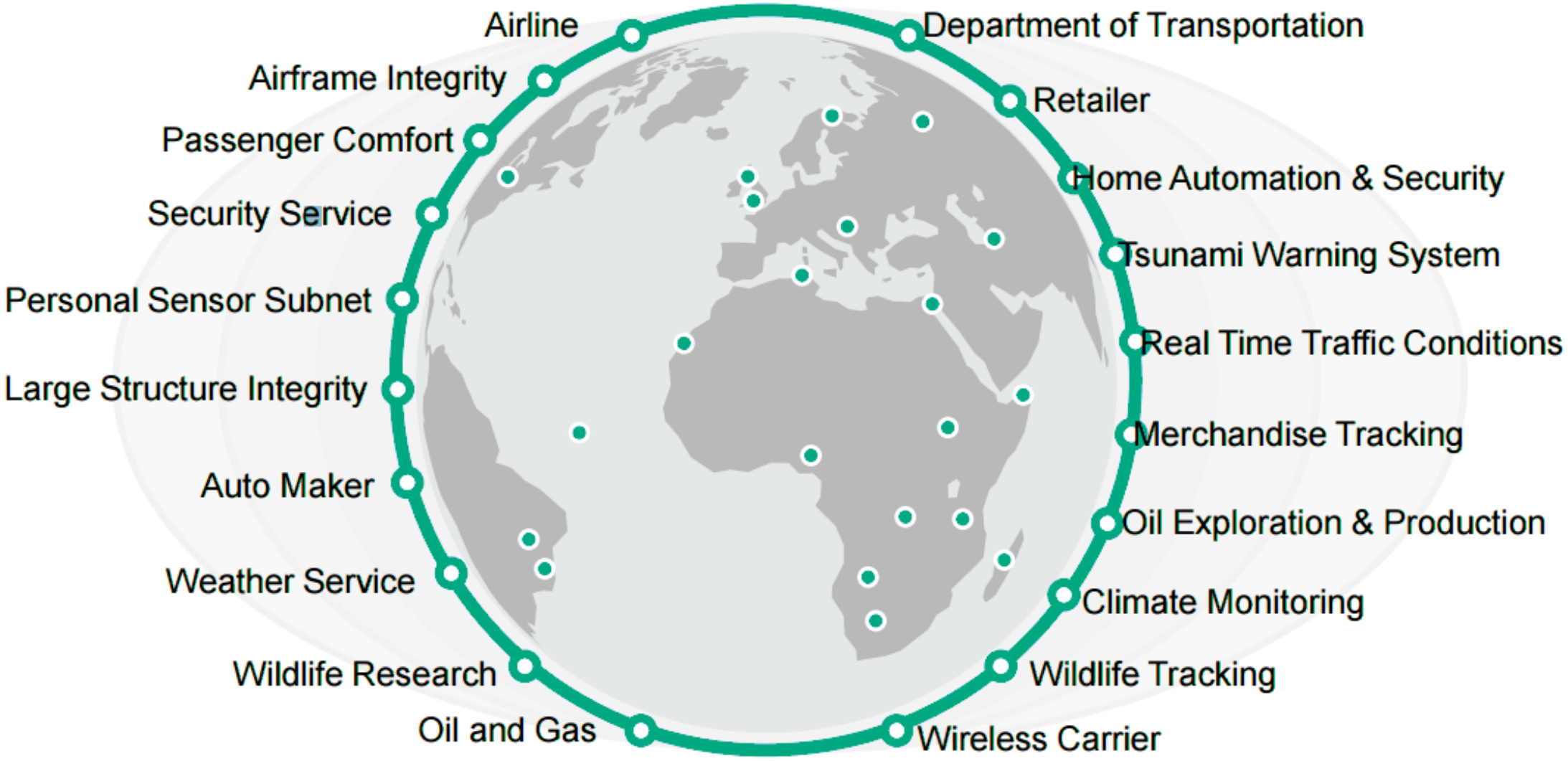
Hewlett Packard
Enterprise

Hewlett Packard
Enterprise

- MegaTrends and Vision
- Technologies for Extreme Scale Computing
- Gen-Z
- Software Implications and Benefits

MegaTrends and Vision

Sensors Big Data will be everywhere and massive



Hewlett Packard Enterprise

How to make sense of that tsunami?

Connected Devices Are Changing Our World

An unlimited potentiality of new business, of high value services to end users

Magnitude of the data

Velocity of data

Unable to secure

Real-time insight needed

Need intelligent actions

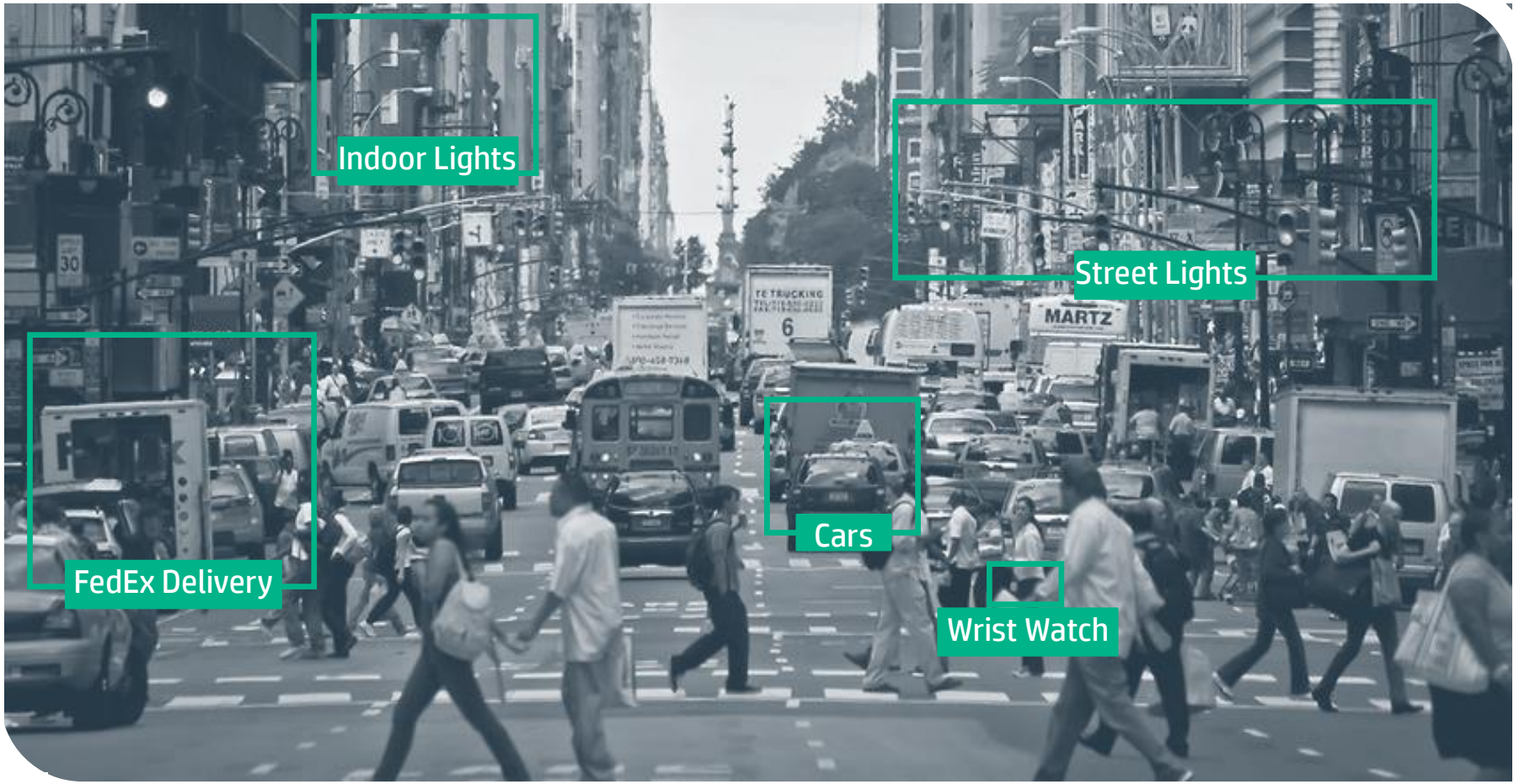
but:

Insufficient resources

End of cheap hardware

Exponential complexity

How to collect and process 10^x more data at iso-cost and iso-energy?



A big challenge for our infrastructures

Current technologies will not respond to the challenges

Need a systematic holistic rethinking

Hewlett Packard Enterprise

Deep Learning and Edge Computing is our future

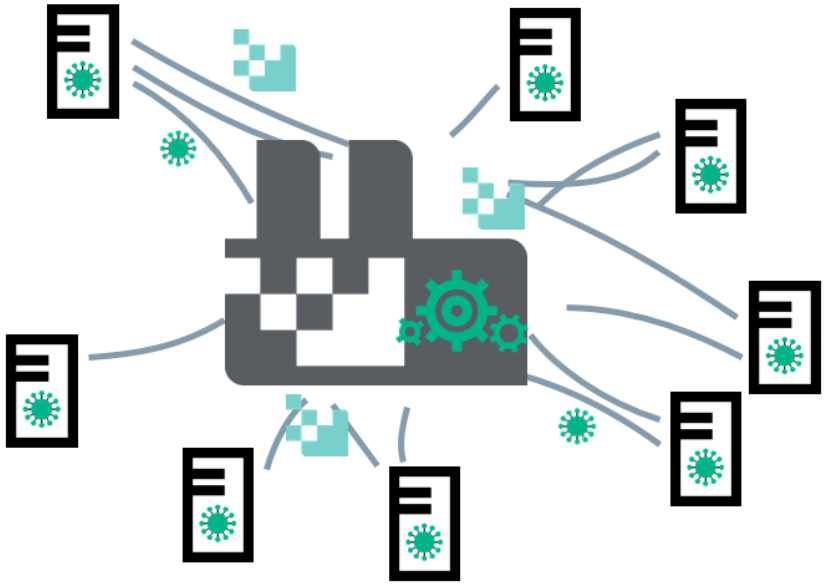
Edge Node : inference

Gets trained model

Uses the model in real-time

Collects data

Sends some data to center



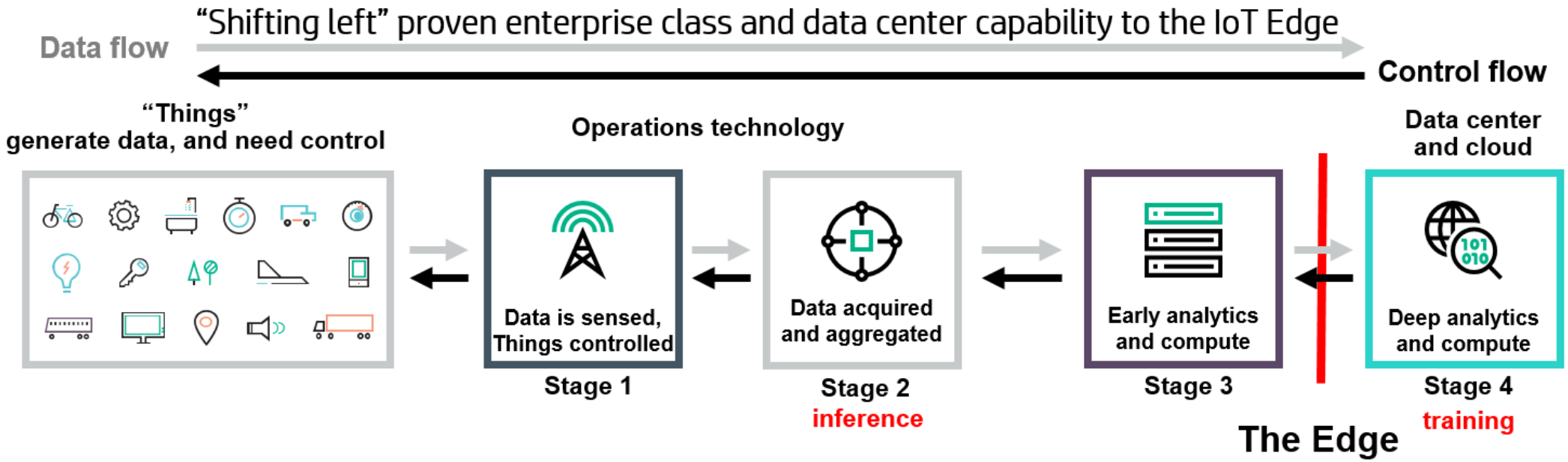
Center : training

Collects all data

Continuously trains models

Sends model to edge nodes

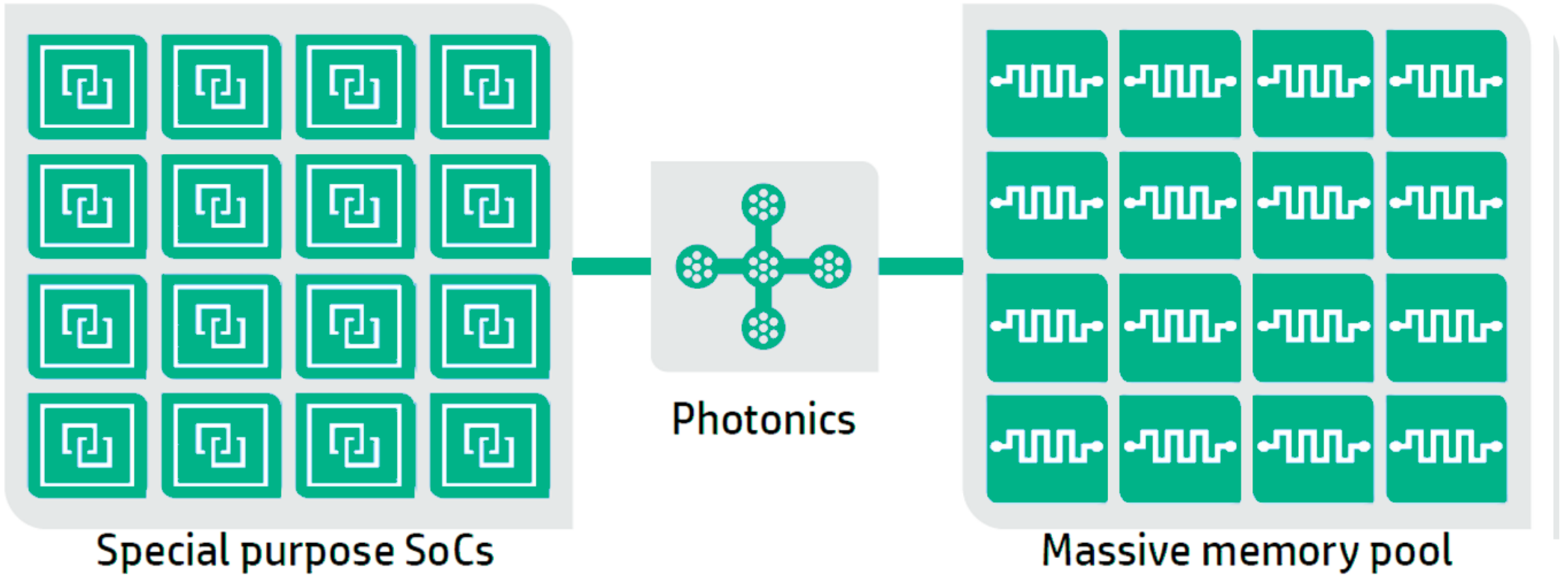
Large scale simulations



Hewlett Packard Enterprise

Technologies for Extreme Scale Computing

3 disruptive technologies to the rescue

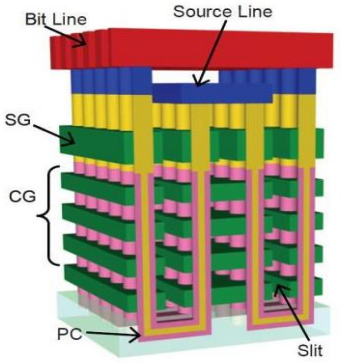


Electrons **Photons** **Ions**

Hewlett Packard Enterprise

Disruption #1: Storage Class Memories

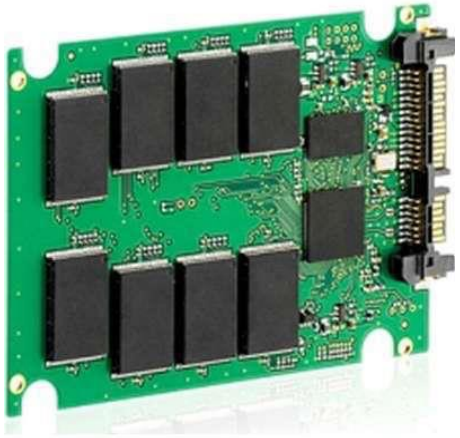
NVM and high speed memories are critical for extreme computing

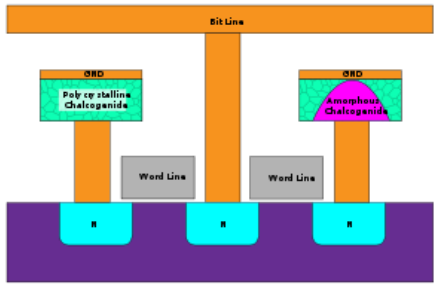


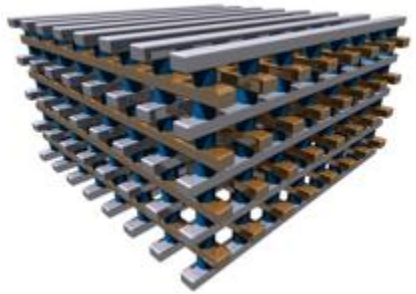
Reaching the physical limits of charge storage

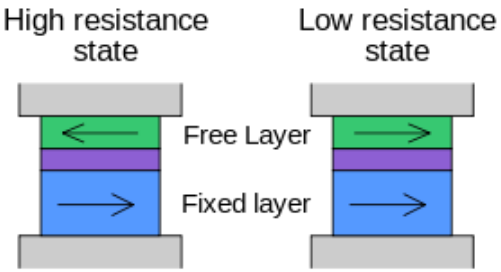
Non-Volatile memories – forms of memristor (Type 4)

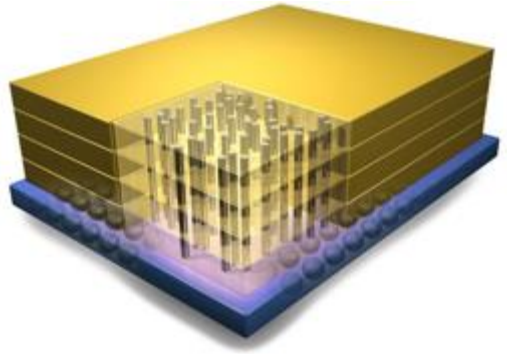
PCRAM **RRAM** **STTRAM** **HMC**











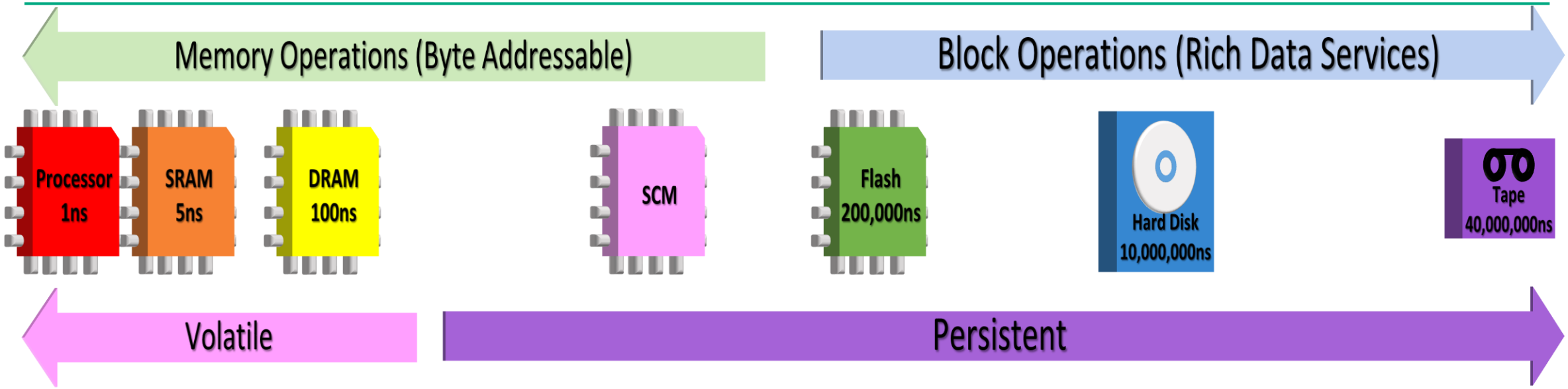
Technology	Density (μm^2/bit)	Bandwidth (GB/s)	Latency Read (ns)	Latency Write (ns)	Energy Read (pJ/b)	Energy Write (pJ/b)
Hard Disk	N/A	0.5	3,000,000	3,000,000	2500	2500
Flash SSD [3] [6]	0.0021	1.0	25,000	200,000	250	250
DRAM [6] [30]	0.0038	51.2	55	55	24	24
PCRAM (22nm) [30]	0.0058	variable	48	150	2	19.2
Memristor (22nm) [8]	0.0048	variable	100	100	1-3	1-3

3D Flash

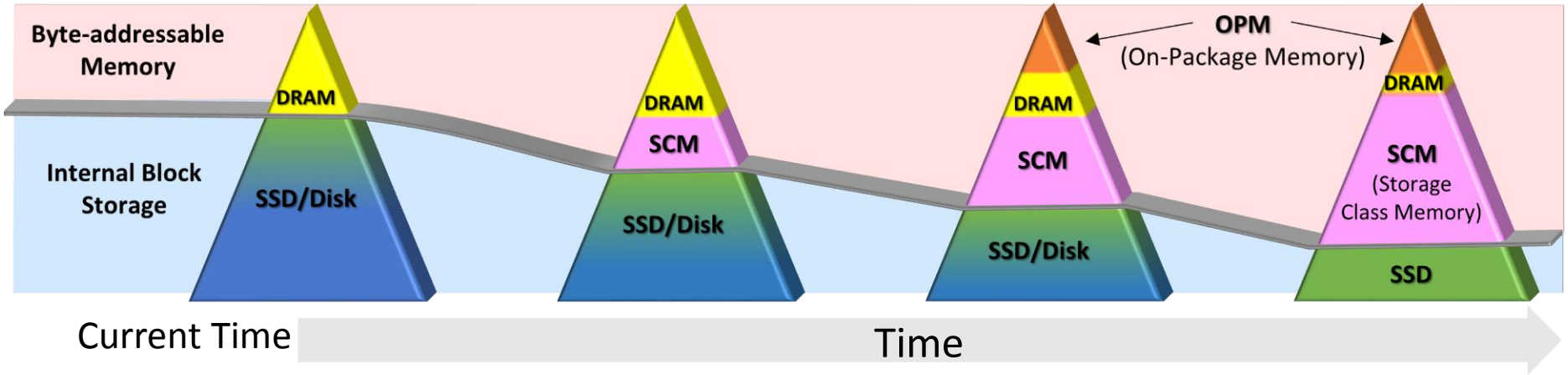
DRAM

Hewlett Packard Enterprise

Memory/Storage Convergence – The Media Revolution



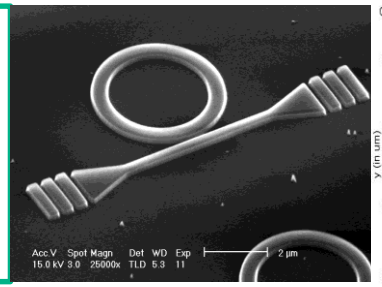
With memory/storage convergence, memory semantic operations become predominant (volatile & non-volatile)

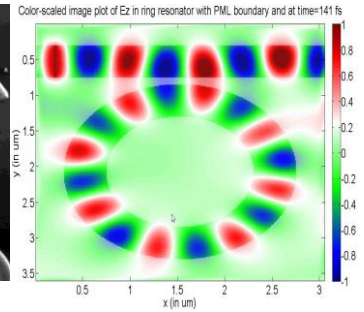


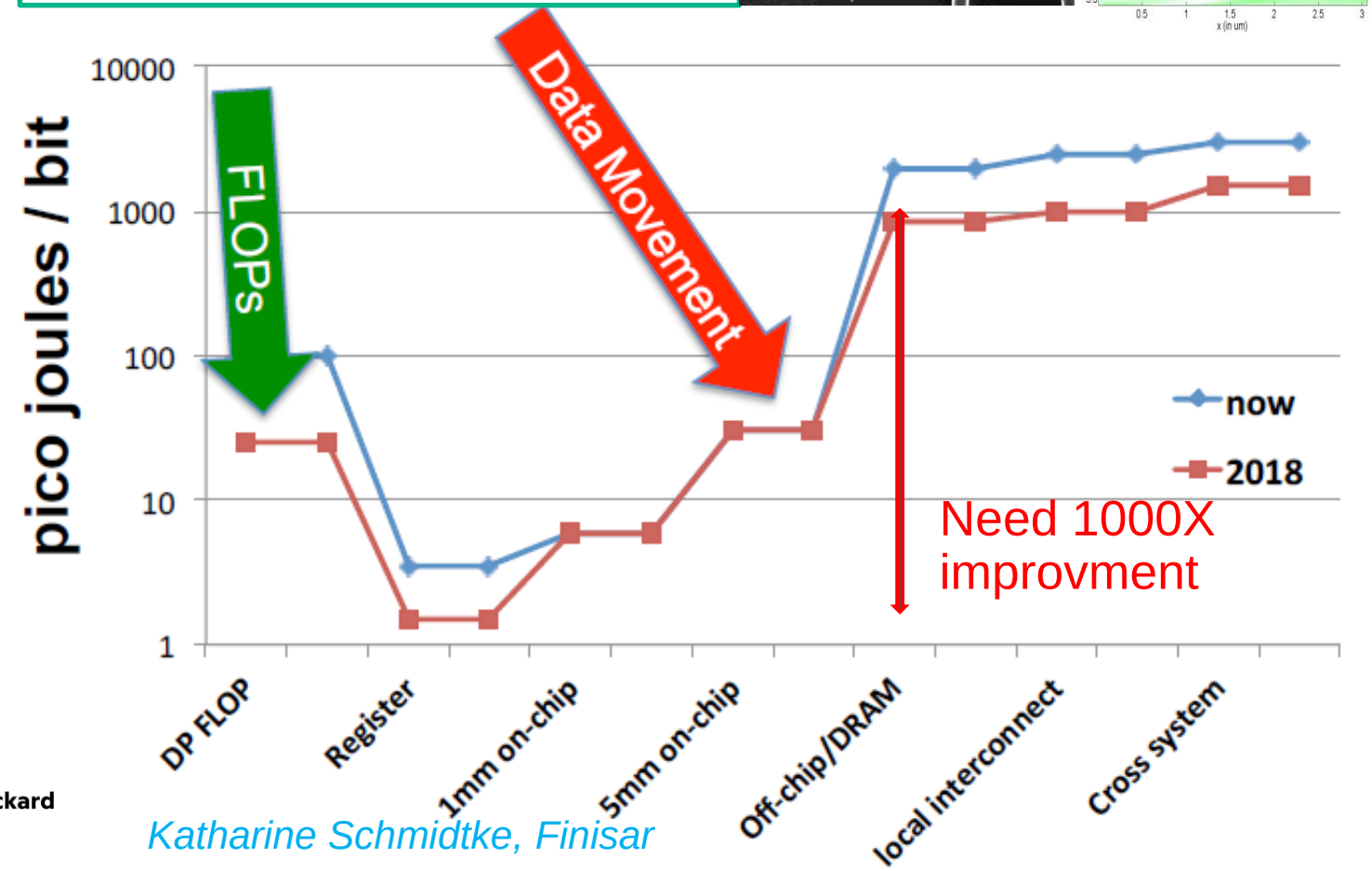
Hewlett Packard Enterprise

Disruption #2 : Photonics

FLOPS will cost less power than on-chip data movement







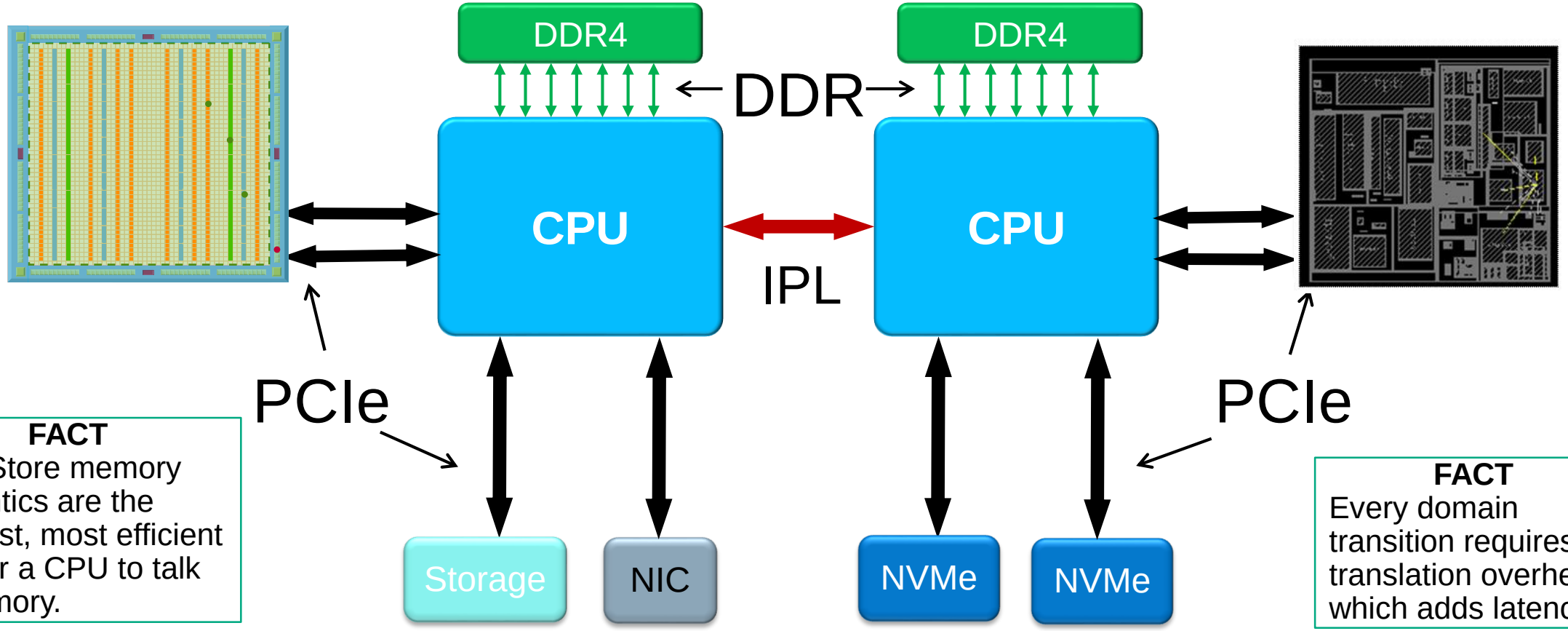
Katharine Schmidtke, Finisar

10^18 ops
*****
1Byte/ops
=
10^19bits
*****
1pj / bit
=
10MWatt

- ultra low energy
- low uniform latency
- high bandwidth
- low cost

Hewlett Packard Enterprise

Extreme scale computing cannot be that architecture

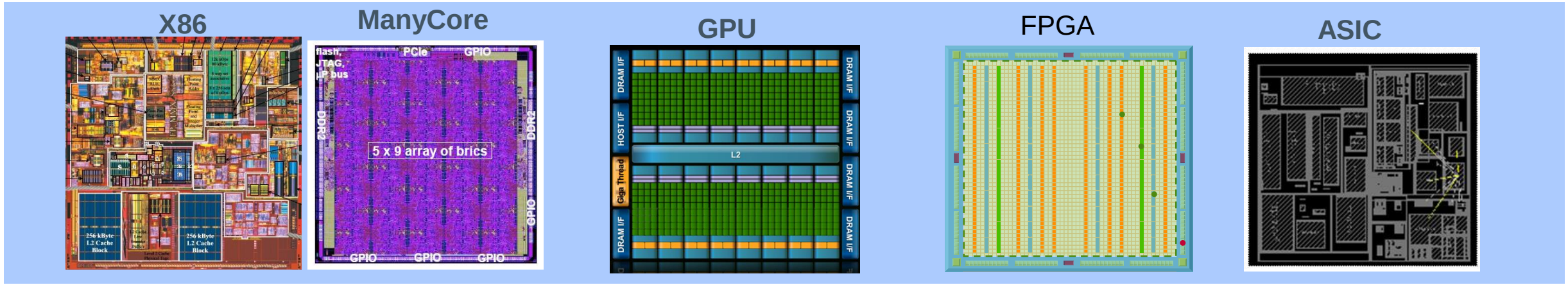


FACT
Load/Store memory semantics are the simplest, most efficient way for a CPU to talk to memory.

FACT
Every domain transition requires translation overhead which adds latency

In a data centric world how to feed our accelerators?
We need a new architecture to deliver TB/s

What technology is optimal for my code?



Flexibility, Easiness of Programming

Performance, Energy and TCO Efficiency

- **X86+ManyCore**
- best for 99.9% lines of code but 1% of cycles
- Large ecosystem
- Easy to debug

- **GPU**
- Better perf for few codes
- Harder to code
- Hard to debug
- Hard to tune
- Need to support 2 versions

- **FPGA**
- Very high performance on specific codes
- Very flexible
- Very energy efficient
- Hard to code and debug
- Very long compile cycle

- **ASIC**
- Only for ultra large market
- Only for ultra stable code
- Very expensive need high volume : like Bitcoin

Hewlett Packard Enterprise

New packaging technologies generate a wide range of cheap SOC

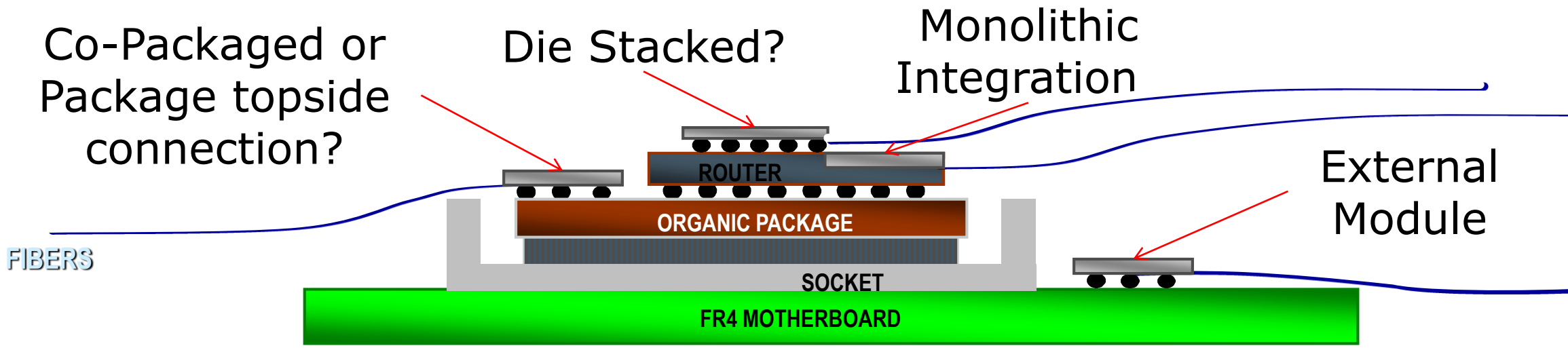


Figure courtesy
Drew Andrino, Intel

- External Module – board level connection
  - Current state of art – no bandwidth density advantage
- Top-side Package Electrical connector
- Co-packaged
- Die stacked
- Monolithic Integration with Router/CPU

- Short time to design
- Easier to debug : chiplet
- Much cheaper even for short series
- Can buy and integrate external IP
- TCI can be a major disruption

Hewlett Packard Enterprise

homogeneity vs heterogeneity

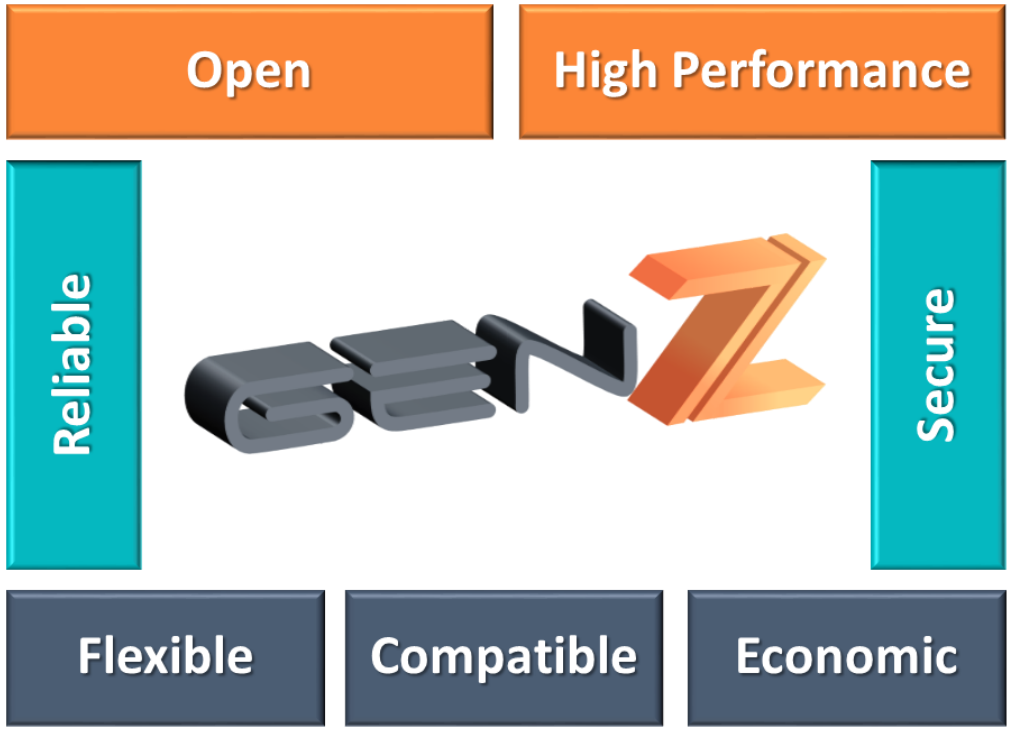
- Good enough for most workloads
- Simple to manage
- Good utilisation
- Long term stability

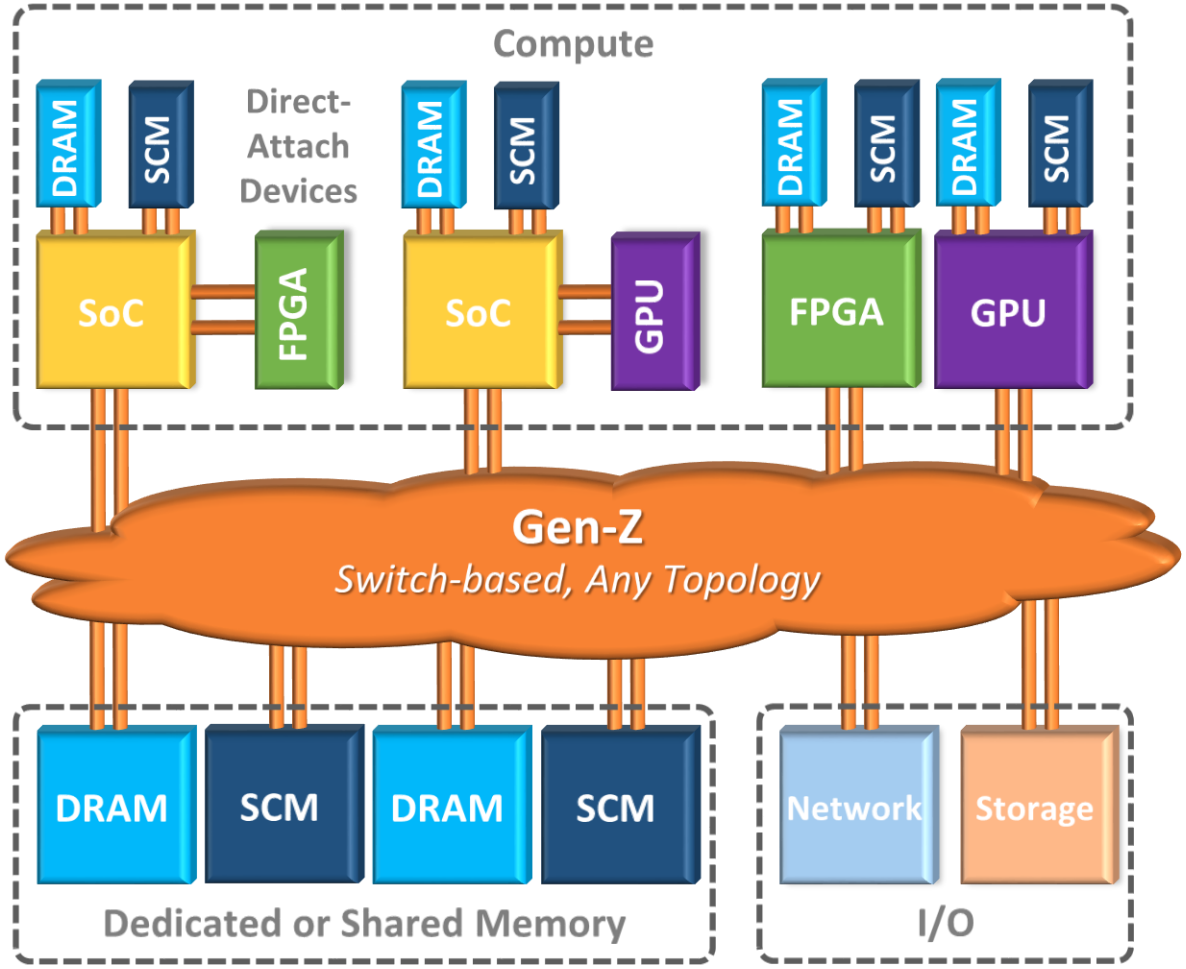
- 10+X better efficiency for few workloads
- Significantly cheaper when applicable
- Much better perf/Watt
- Hard to code and tune
- Hard to manage
- At extreme scale : NO CHOICE we need to find the best options

Gen-Z

The Solution: Gen-Z!

Memory Semantic Fabric (language of compute)





Hewlett Packard Enterprise

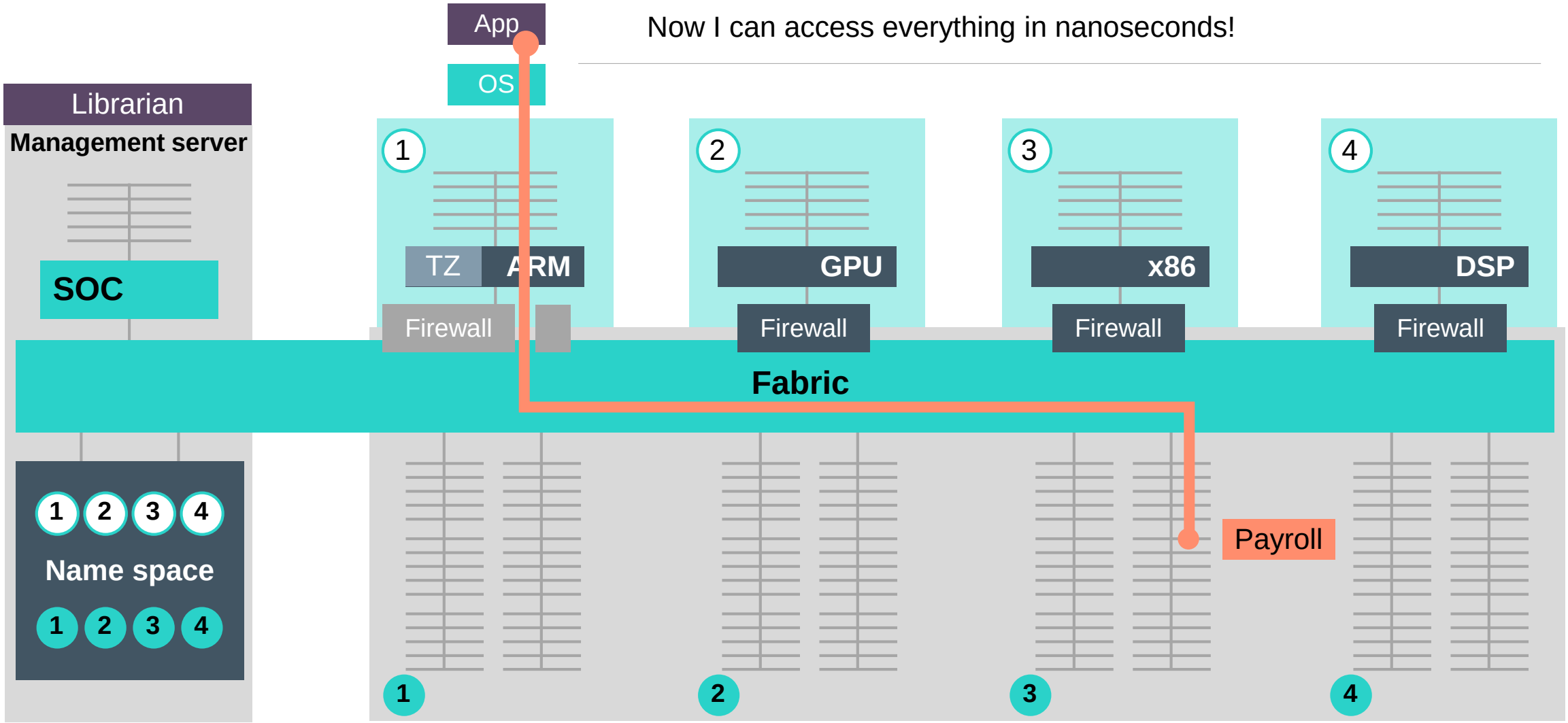
Open: Consortium with Broad Industry Support

GEN Z Consortium Members

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

*Board member

Semantic of access : load/store Gen-Z protocol

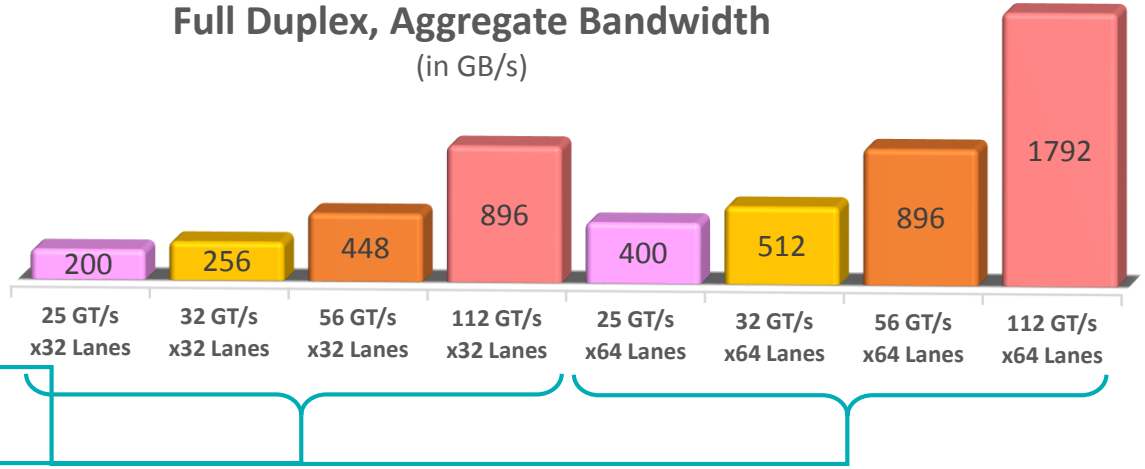


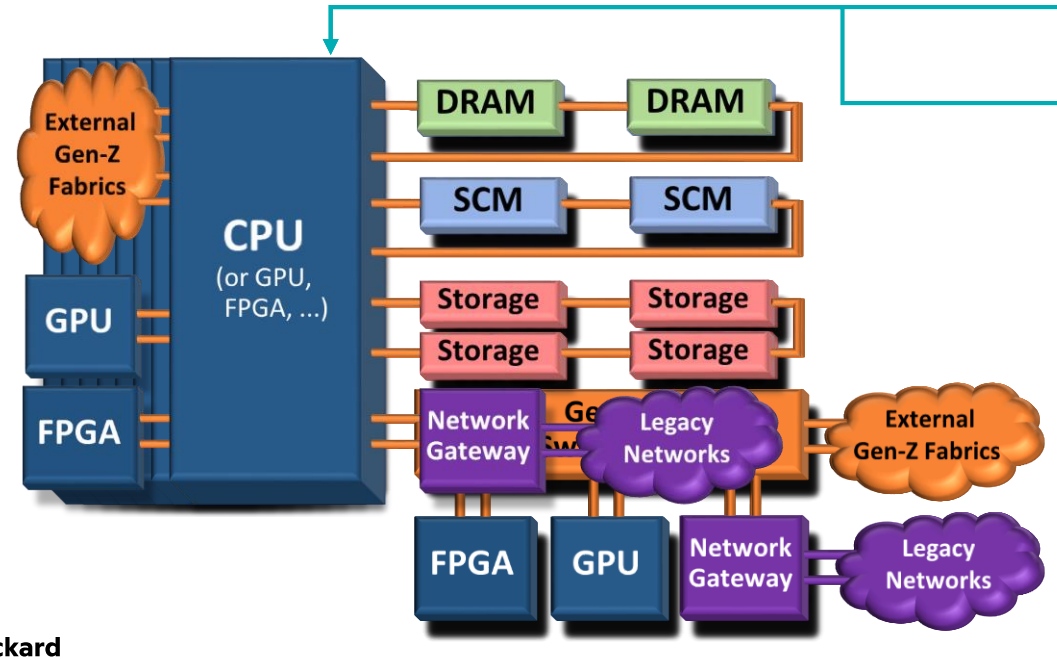
Hewlett Packard Enterprise

High Performance: Bandwidth

Leading Edge Bandwidth

- Powers Memory/Storage Convergence
- Platform for Data Access and Messaging

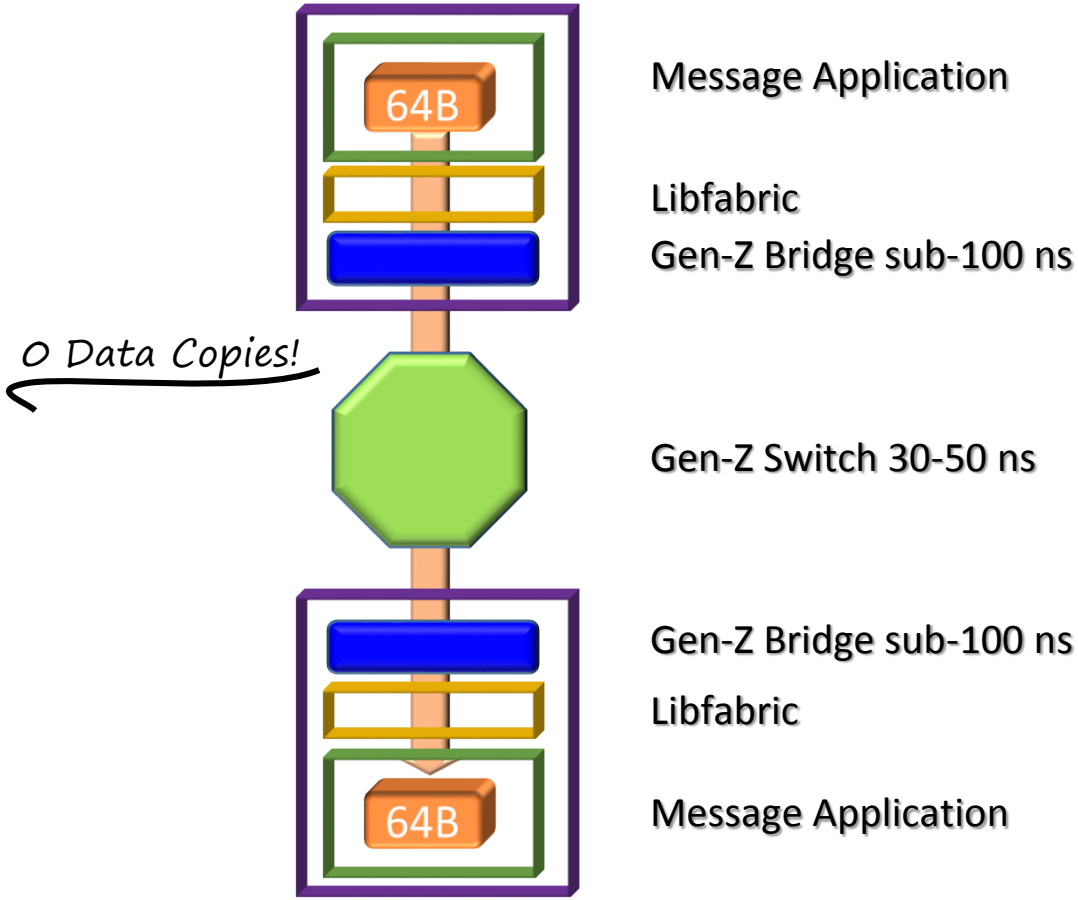




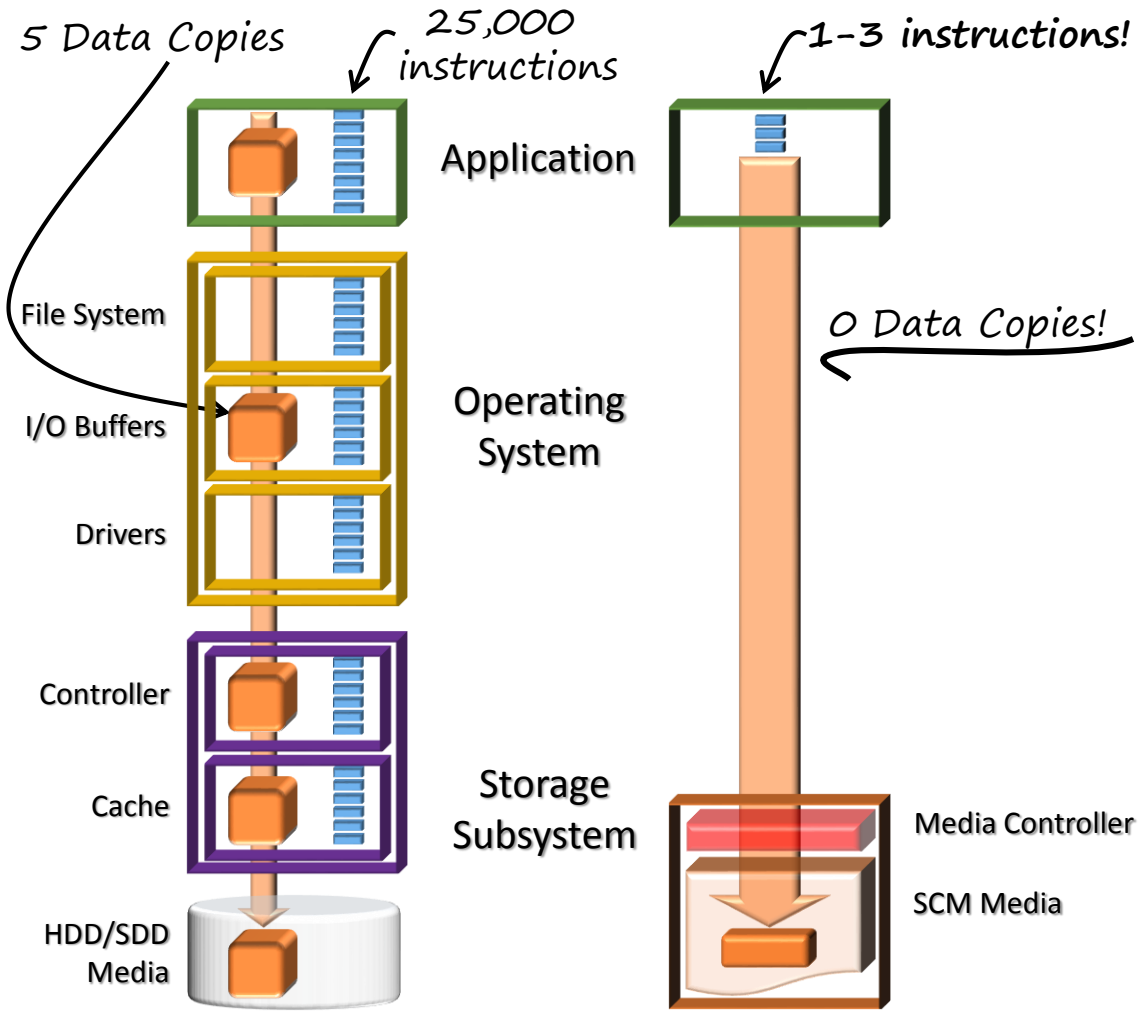
- Processors with Integrated Gen-Z
  - Highest bandwidth, low latency memory access
  - Can support 64 or more Gen-Z lanes per SoC

Hewlett Packard Enterprise

High Performance: Low Latency



- Ultra-low-latency messaging (Sub-250 ns one-way latency)
- Load/Store byte-addressable access to DRAM or SCM
  - Sub-100 ns load-to-use latency (DRAM media)
- Reduced CPU utilization = ***More workloads per core per CPU***



Traditional Storage Software Stack

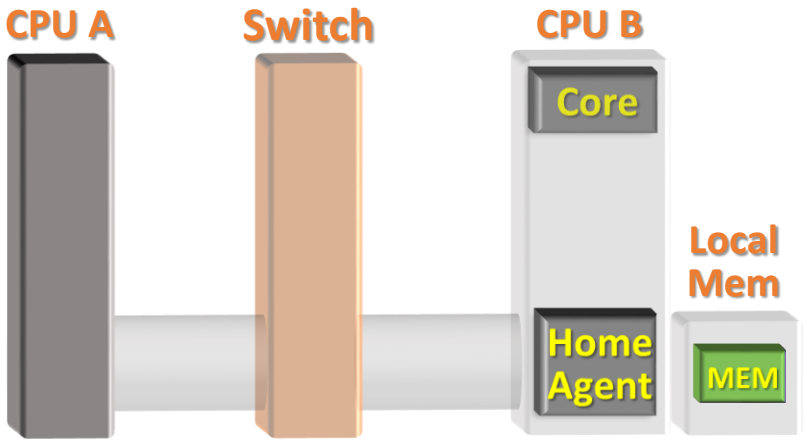
Gen-Z w/Persistent Memory Software Stack

Low Latency Messaging: Atomic Operations

- Designed from the ground up to support atomic operations
  - 18 atomics with extensible, powerful options, compatible with x86, ARM, and Power ISAs
- Atomics are not just for compute, but also fabric-attached memory
  - Enables multiple compute synchronization models

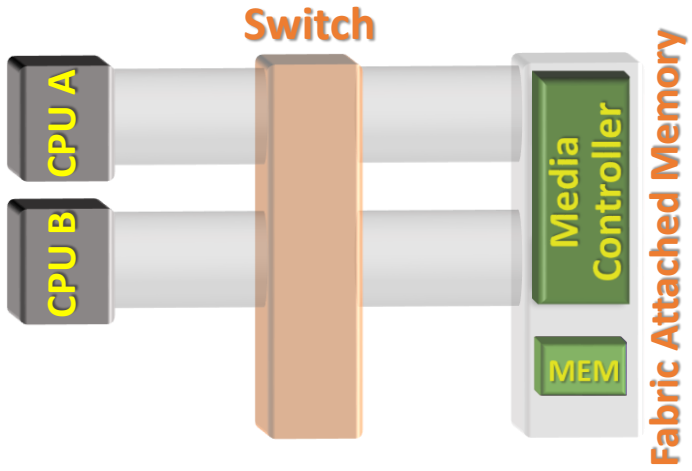
- **Add**
- **Sum**
- **Swap**
- **CAS**
- **CAS Not Equal**
- **Logical OR**
- **Logical XOR**
- **Logical AND**
- **Load Max**
- **Load Min**
- **Test Zero & Modify**
- **Increment Bounded**
- **Increment Equal**
- **Decrement Bounded**
- **Compare Store Twin**
- **Atomic Vector Sum**
- **Atomic Vector Logical**
- **Atomic Fetch**

Compute to Compute Shared Memory Atomics Operations



SWP **Atomic Swap Request/Response Packet**
(sets or returns a value of "Red")

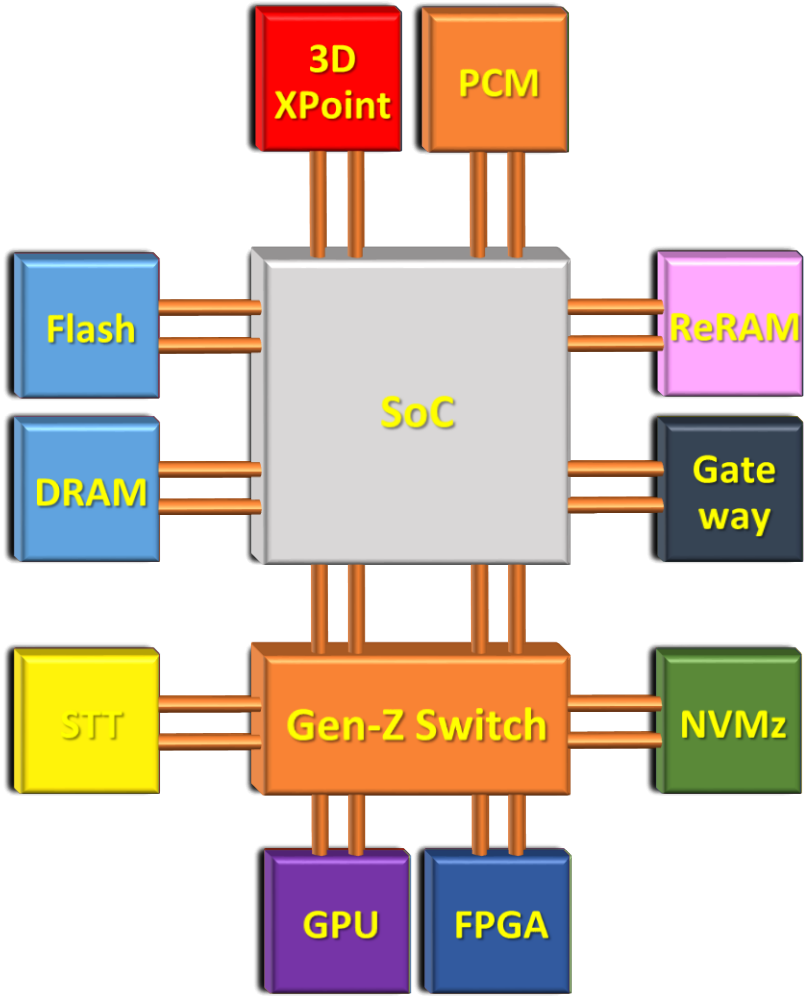
Compute to Shared Fabric Attached Memory Atomics Operations



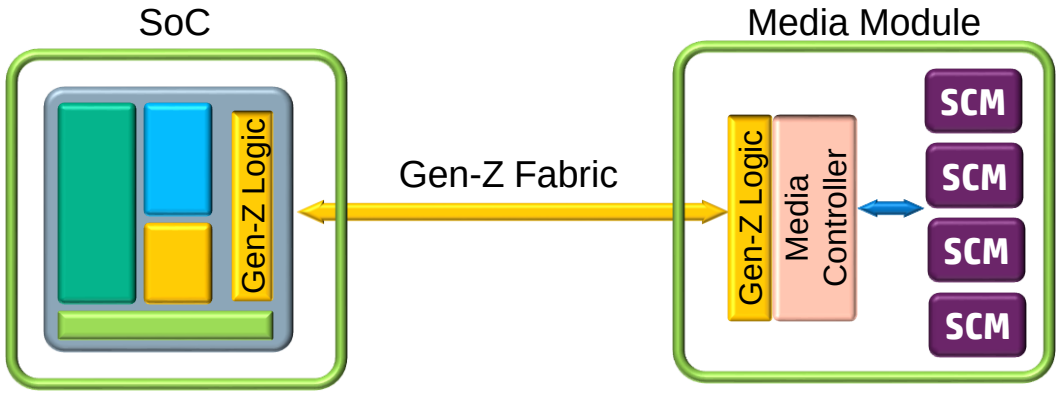
Hewlett Packard Enterprise

Flexible: Universal System Interconnect

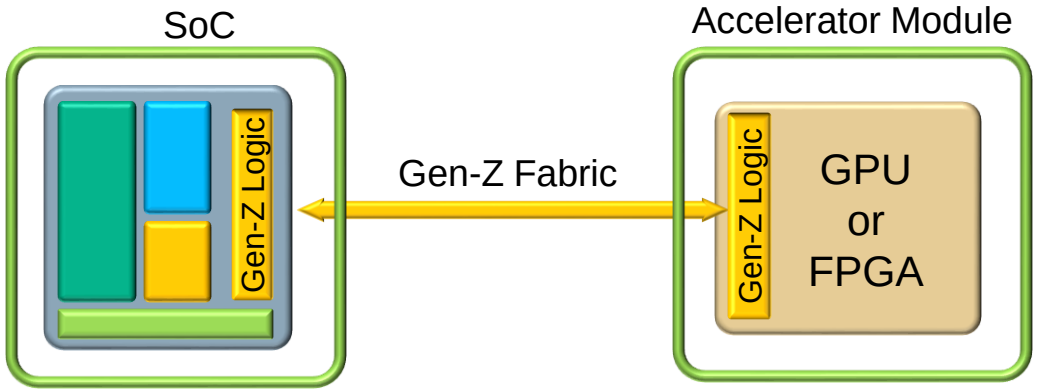
- Any device connected into any topology
- No dedicated memory, I/O, or storage links
- Enables fluid deployments
- Enables construction of “right-sized” infrastructure



Gen-Z is a high speed memory semantic fabric

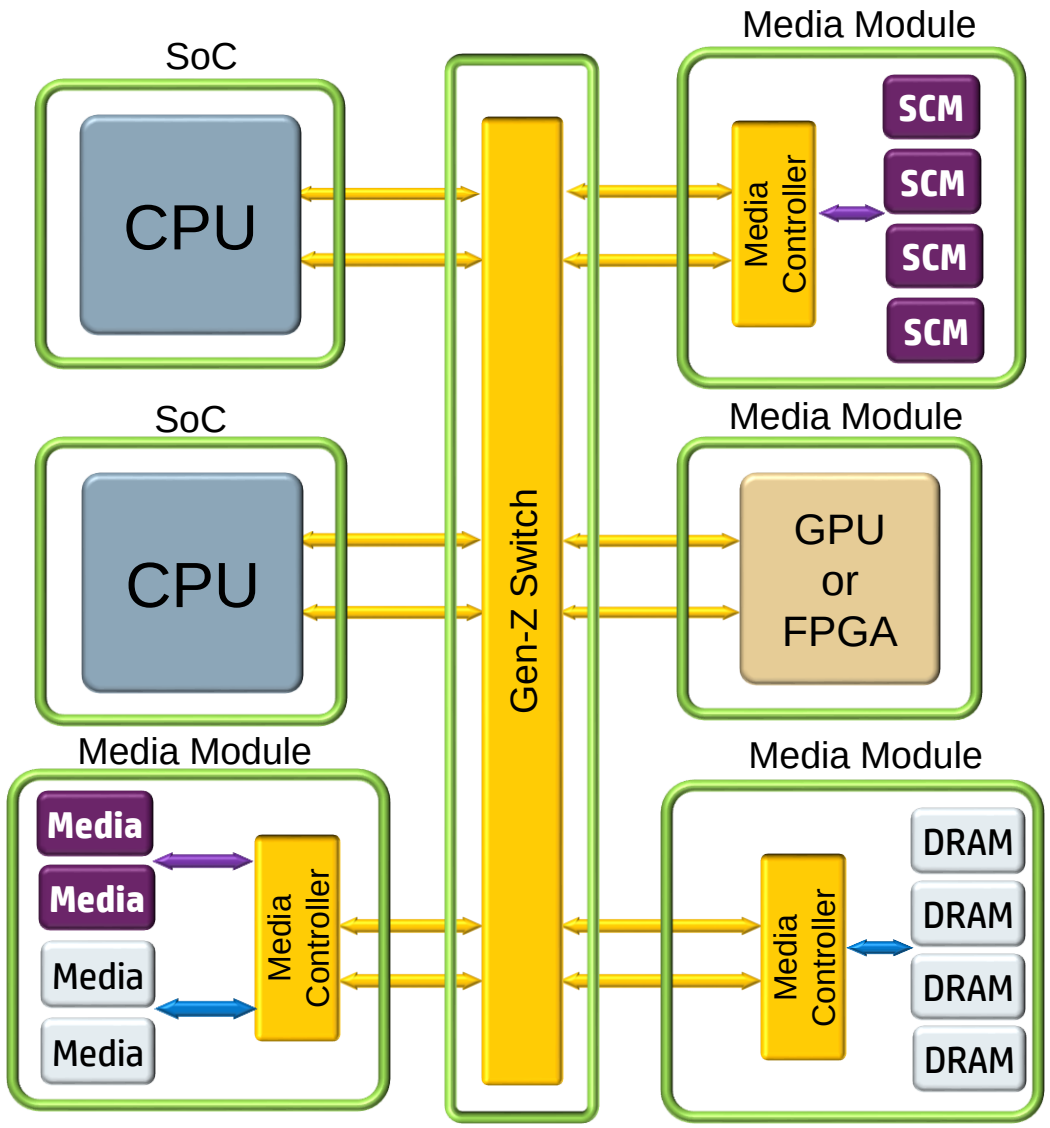


Storage Class Memory



GPU or FPGA

- Supports DRAM, Flash, Memristor, PCRAM, MRAM, 3D-Xpoint…**Universal Interconnect**
- Decouples CPU/memory design
- Enables independent innovation



Multiple resources enabled by Universal Interconnect

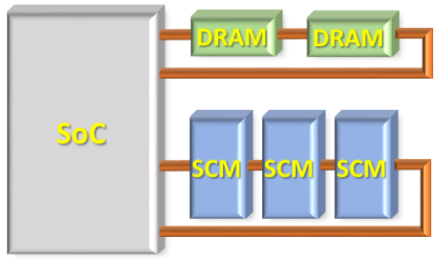
Hewlett Packard Enterprise

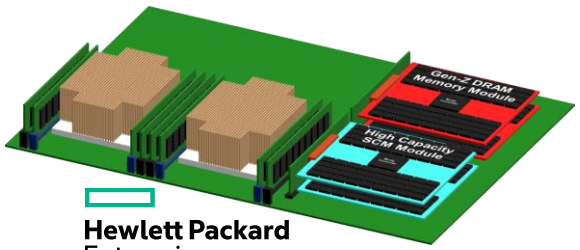
Flexible: Topologies of Various Scales

- Low cost, low power, copper direct-attached interconnects for local systems
- Optical, long range, switched interconnects for scale out and HPC
- Mixed deployment models for enterprise enclosure and rack scale

Local System

- Copper, low cost
- PCIe or IEEE 802.3 PHYs
- P2P, Daisy-chain, switched

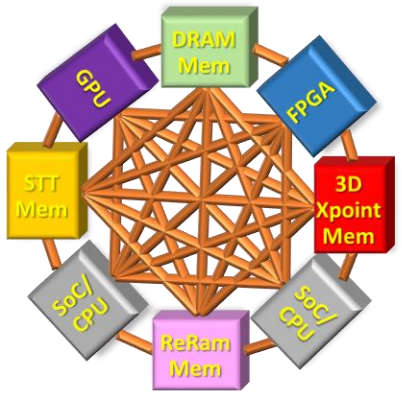


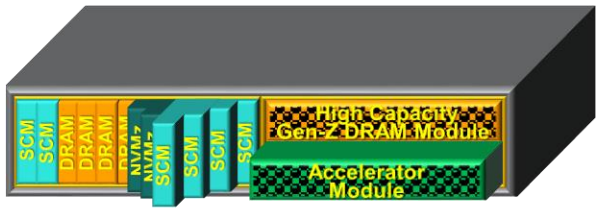


Hewlett Packard Enterprise

Chassis/Enclosure

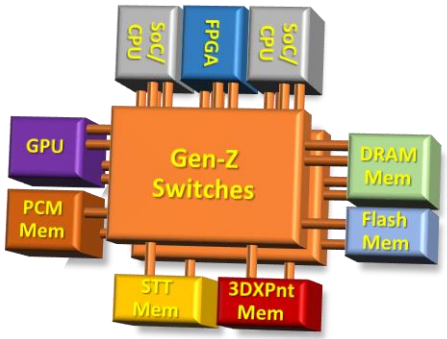
- Copper, low cost
- PCIe or IEEE 802.3 PHYs
- P2P, Mesh, Torus, switched

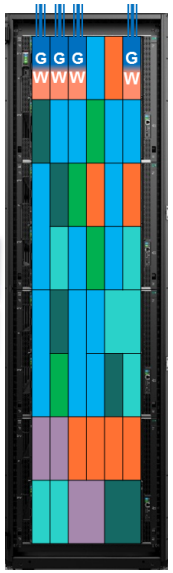




Rack-Scale

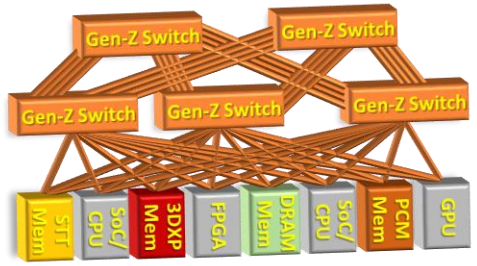
- Copper or optical
- IEEE 802.3 PHYs
- Switched, Torus, Spine/Leaf

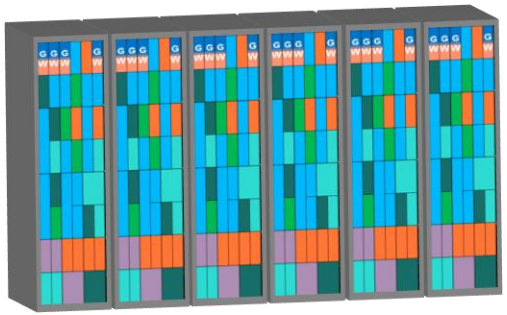




Row-Scale

- Optical
- IEEE 802.3 PHYs
- Switched, Fat Tree, Clos, Butterfly, Hyper-X, etc.

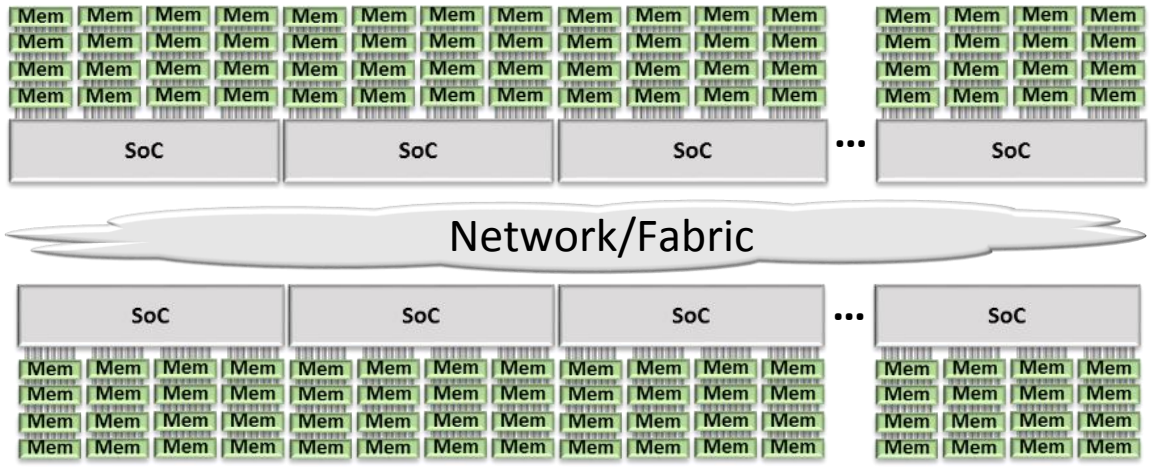




Memory: Composability

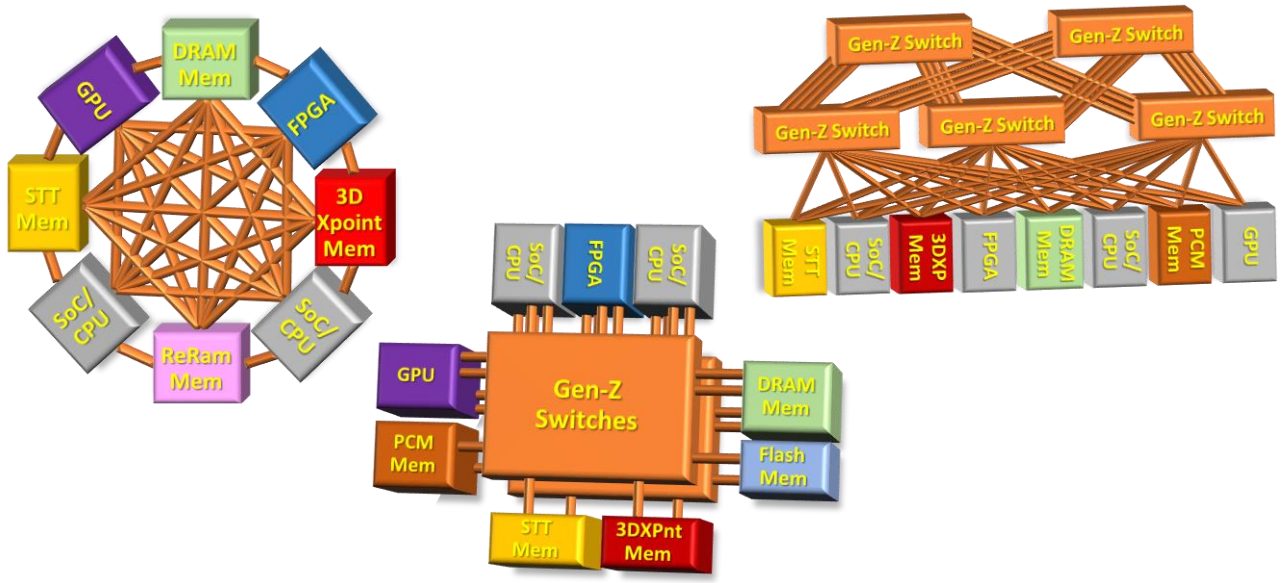
Today

- Memory is captive of the host device (processor)
  - Stranded memory channels and memory resources
- Can't scale memory independently of processing
- All accesses must traverse host processor



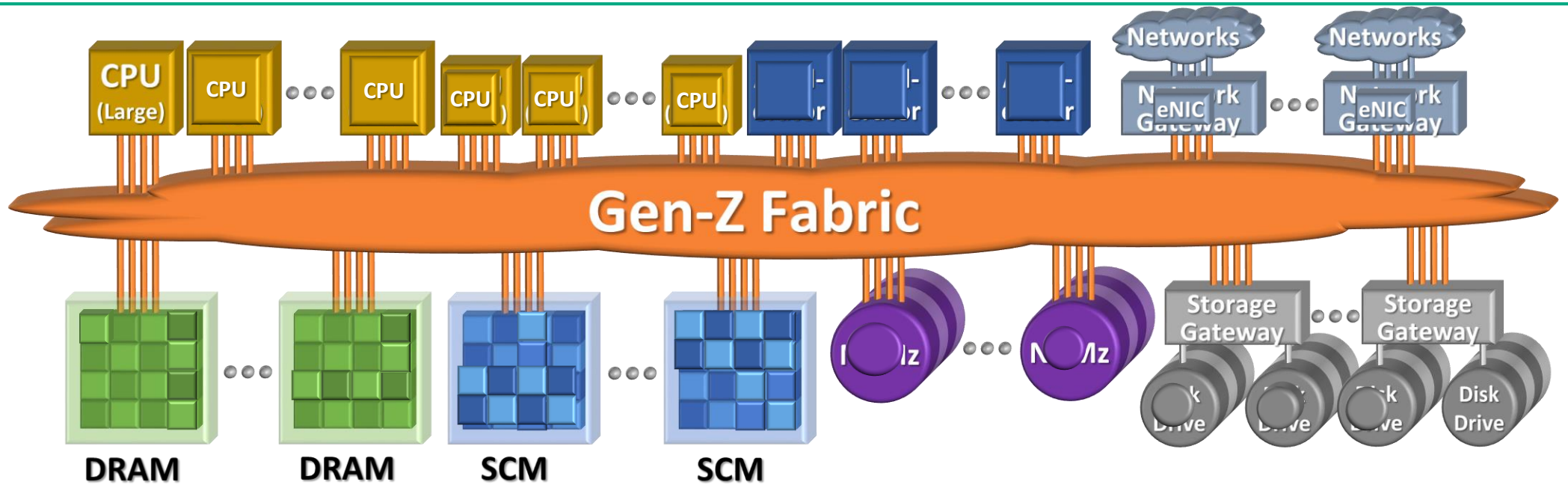
Gen-Z

- Memory and processing scale independently
- Heterogeneous compute & memory deployments
- Direct access to memory devices across fabric
- Memory can be dedicated or shared by processors
- Supports up to 64-way barber pole memory interleave
- Supports RAID / erasure code-based memory solutions
- Scales from motherboard to rack-scale

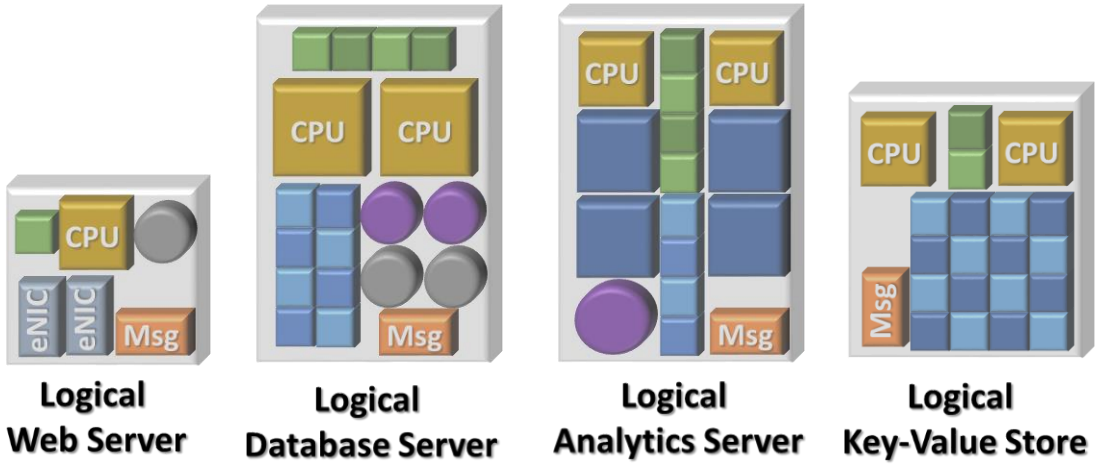


Hewlett Packard Enterprise

Economic: Right-Sized Solutions



- Logical systems composed of physical components
  - Or subparts or subregions of components (e.g. memory/storage)
- Logical systems match exact workload requirements
  - No stranded resources overprovisioned to workloads
- Facilitates data-centric computing via shared memory
  - Eliminates data movement: Do more with less, reduces cost



Hewlett Packard Enterprise

HPE introduces 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

Software Implications and Benefits

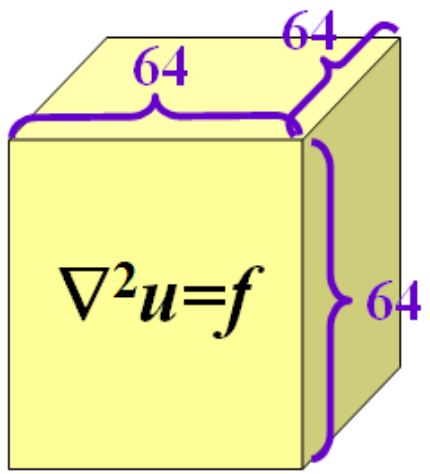
Better algorithms

Algorithmic efficiency is more critical than hardware architecture improvements at extreme scale

The new architecture offers a wide range of opportunities to make breaktrough transformations

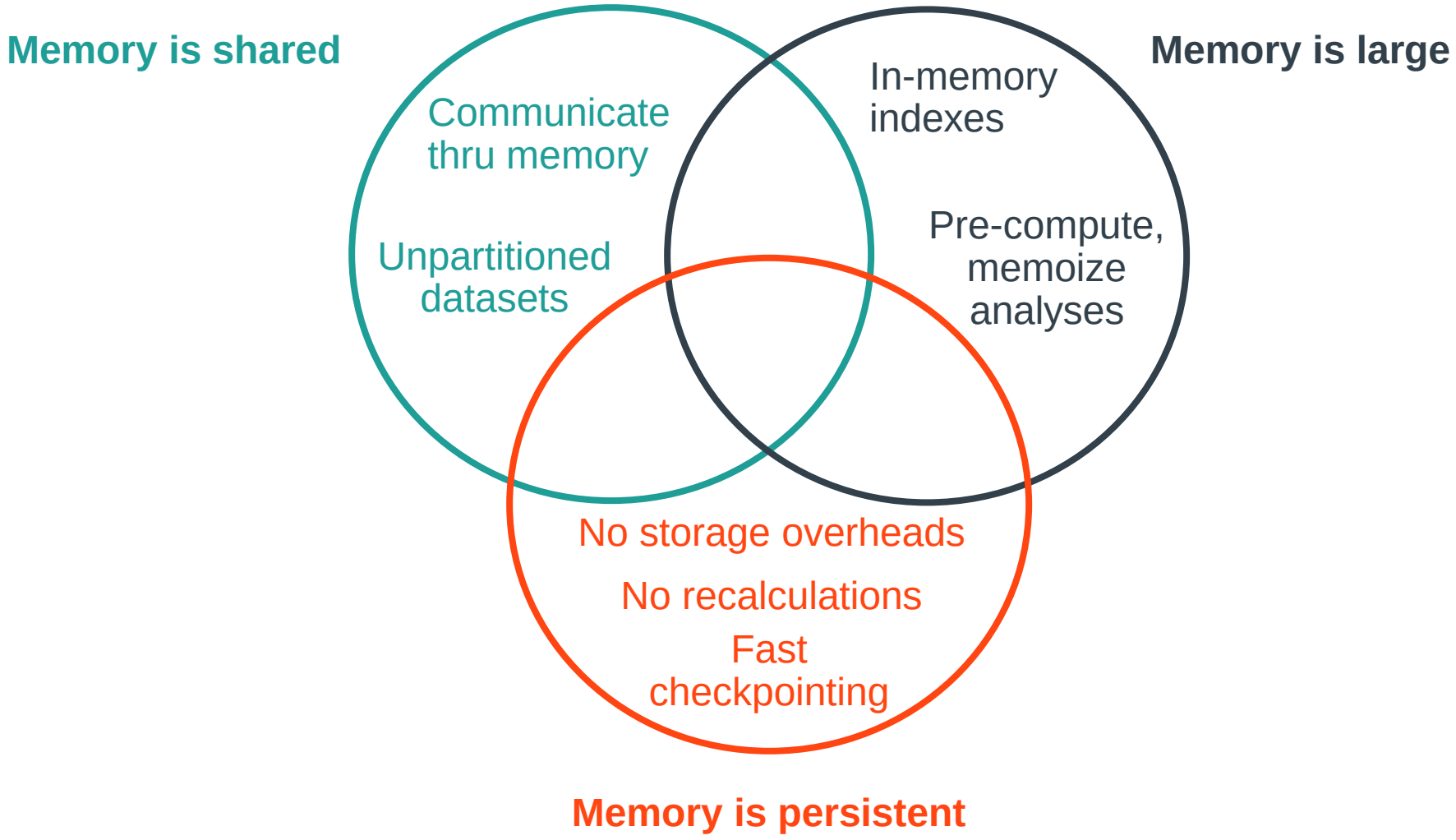
– Exemple : Poisson's equation on a cube of size $N=n^3$

Year	*Method*	*Reference*	*Storage*	*Flops*
1947	GE (banded)	Von Neumann & Goldstine	n^5	n^7
1950	Optimal SOR	Young	n^3	$n^4 \log n$
1971	CG	Reid	n^3	$n^{3.5} \log n$
1984	Full MG	Brandt	n^3	n^3



*www.siam.org/about/science/**keyes**.ppt*

Memory-Driven Computing benefits applications

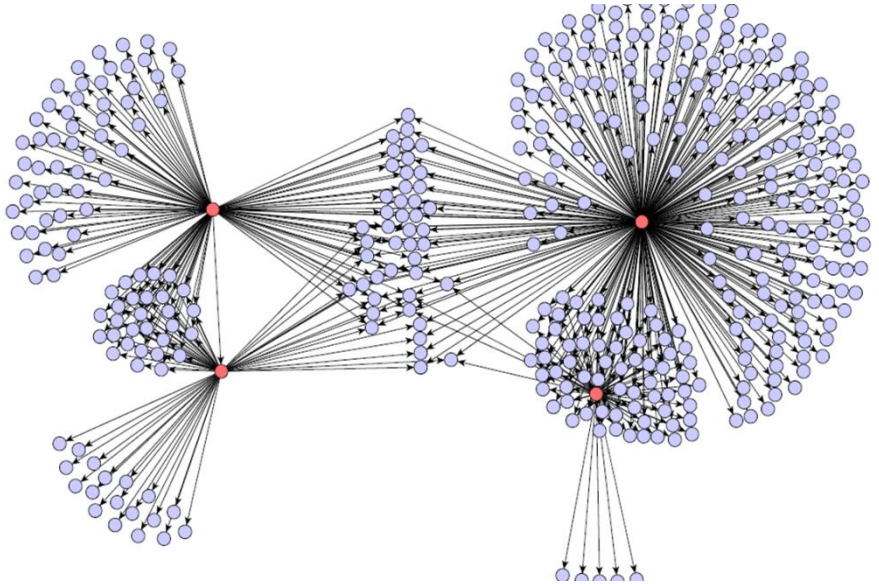


Large-scale Graph Inference

Memory-Driven Computing architecture is a “natural” fit for graph problems

Current architectures do not support efficient computations on large graphs

- Horizontally scalable clusters have too much communication overhead
- Vertically scalable machines have limited computation capacity



Memory-Driven Computing provides both

- **Large memory**, enabling the entire data to stay in-memory in the same place
- **Large number of cores**, enabling parallel computation

Large Scale Graph Inference (LSGI)

- Compute probabilities across whole graph based on a small known set of vertices
- Predict website safety, customer behavior, etc.

Memory-Driven Monte Carlo simulations

Traditional

Step 1: Create a parametric model $y = f(x_1,\ldots,x_k)$
Step 2: Generate a set of random inputs
Step 3: Evaluate the model and store the results
Step 4: Repeat steps 2 and 3 many times
Step 5: Analyze the results

Memory-Driven

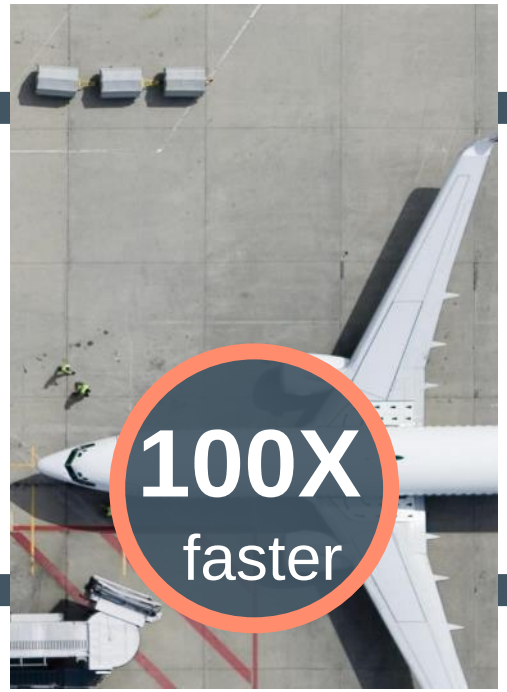
Capacity to store representative behaviors of pre-simulated model allows us to **replace** steps 2 and 3 with look-ups and simple transformations

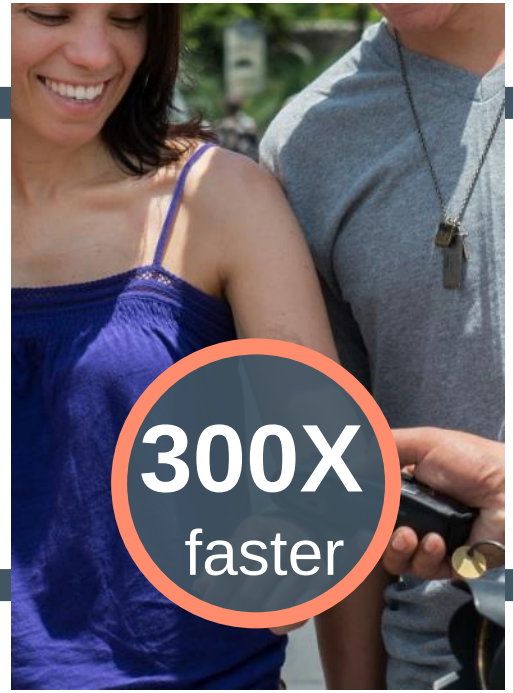
Transform performance with SCM

Modify existing frameworks

New algorithms

Completely rethink





In-memory analytics

Large-scale graph inference

Similarity Search

Financial models

Hewlett Packard Enterprise

First collaboration: German Center for Neurodegenerative Diseases (DZNE)

- Collaborating with **German Center for Neurodegenerative Diseases (DZNE)** needed a new kind of computer to manipulate their massive data sets to accelerate finding a cure for Alzheimer's, a disease that impacts one in 10 people age 65 or older in the world. The initial findings from our collaboration are powerful and promising:

- We've only started to scratch the surface with one component of their overall data analytics pipeline and are already **getting over 40X speed improvements.**

- We're getting **results that used to take more than 25 minutes in 36s.**

- We believe these **gains could increase up to 100X** when we expand our learnings to the other components of their pipeline. Saved time translates to save lives, these efficiencies could change the game.

- DZNE has never been able to work with so much data at one time, which means **finding hidden correlations and better answers than ever before** – ultimately resulting in new discoveries to help cure Alzheimer's.

DZNE
Deutsches Zentrum für Neurodegenerative Erkrankungen in der Helmholtz-Gemeinschaft

Hewlett Packard Enterprise

Memory-Driven, Data-Centric Computing

Powered by Gen-Z!

- Fabric-attached memory for active data
- Stop moving bulk data
  - Only access the data required
- Stop re-formatting data (block, file/object)
- Simplify the software (more CPU for workloads)
  - Create, Persist, Analyze all in one place, all in one format
- Enables dynamic workload placement
  - Move the workloads, not the data

Hewlett Packard
Enterprise

Thank you