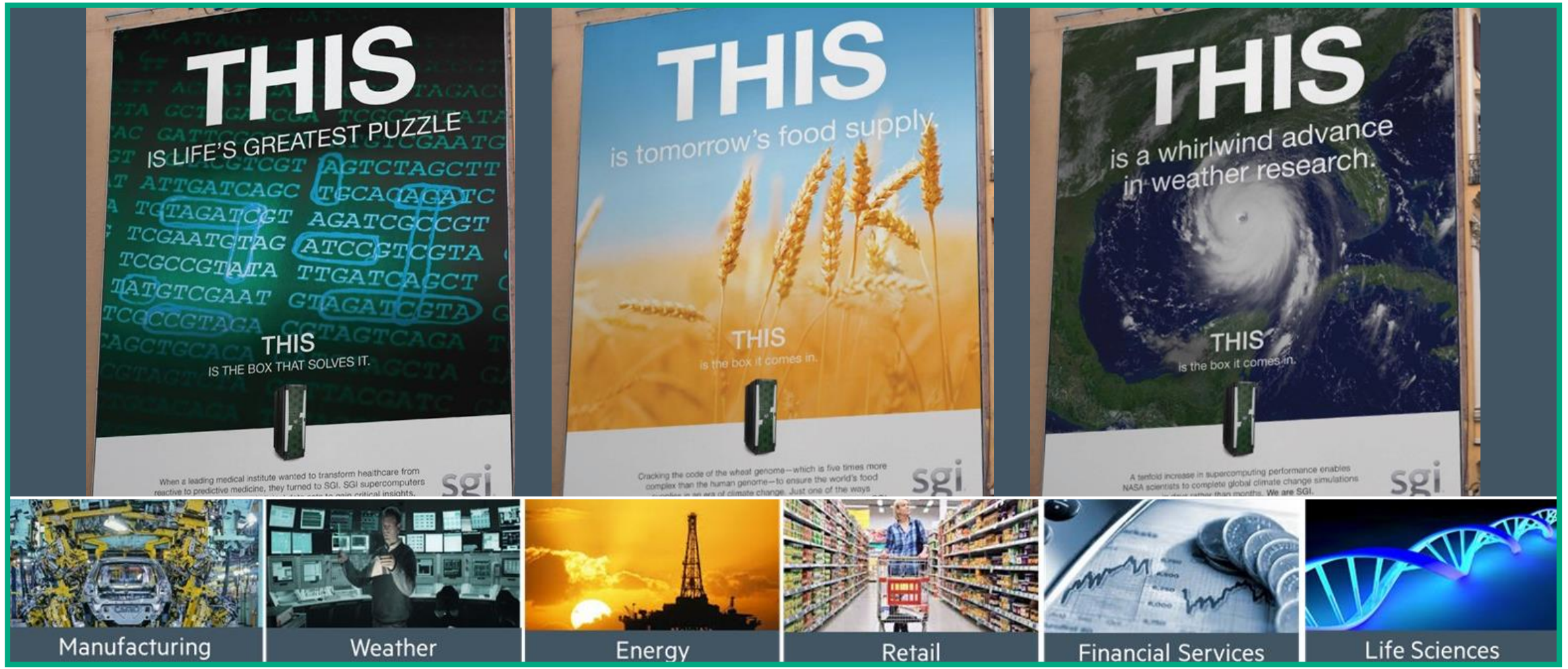
Hewlett Packard Enterprise

Progress Toward Exascale Computing

Mike Vildibill
VP Advanced Technologies & Exascale Development
October 2017

Exascale in our daily life

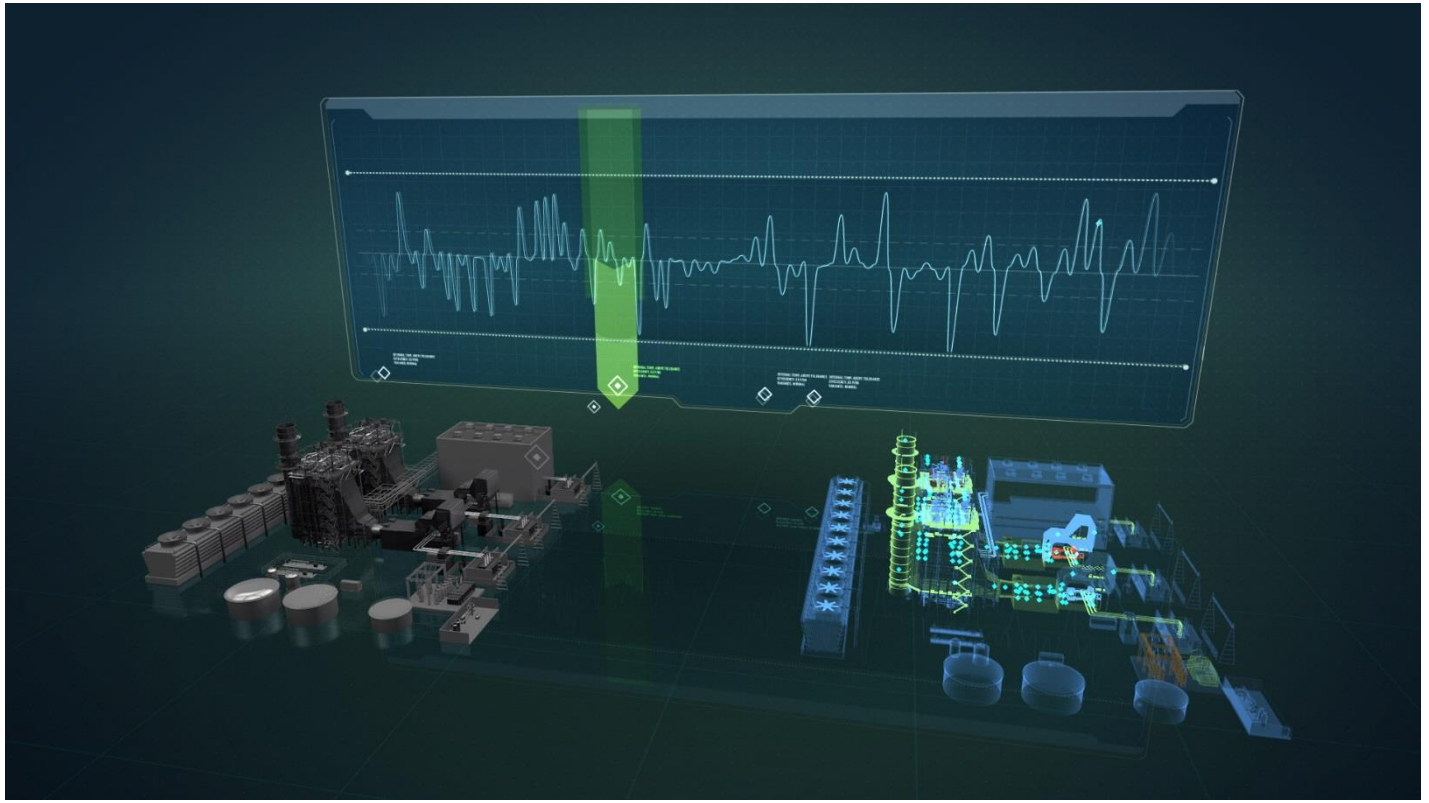


The Digital Twin

Exemplifies the Insatiable Demand for Affordable and Accessible Computing

Digital twin refers to a **digital** replica of physical assets, processes and systems that can be used for various purposes. The **digital** representation provides both the elements and the dynamics of how an Internet of Things device operates and lives throughout its life cycle.

Digital twin - Wikipedia
https://en.wikipedia.org/wiki/Digital_twin



Hewlett Packard Enterprise | intel

What Is Exascale?

–One quintrillion

–10**18

–1,000,000,000,000,000,000

–1 billion * 1 billion

–1 Peta * 1,000

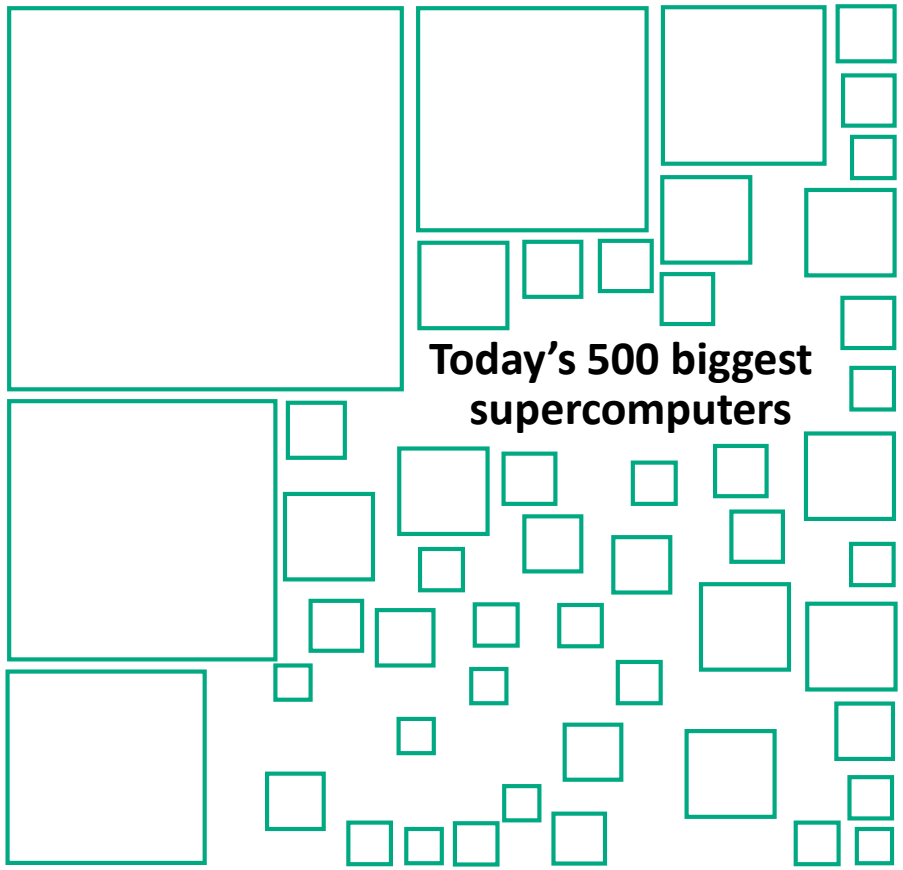
In FLOPs, the sum of all Top500 HPC systems today

If a single calculation is represented by a single piece of paper, then a quintrillion pieces of paper would amount to a stack of papers as tall as…

65,000 round trips to the moon

Hewlett Packard Enterprise | intel

How do we meet the exascale challenge?



650 MW

=

the MACHINE

One exascale supercomputer

1,000,000,000,000,000,000 math operations per second

20-30 MW

Hewlett Packard Enterprise | intel

Exascale is a Global Race

U.S.

- Sustained: 2023
- Peak: 2020-21
- Vendors: U.S.
- Processors: x86 and ARM
- Initiatives: NSCI, ECP
- Cost: $250-300M

EU

- Sustained: 2023-24
- Peak: 2021
- Vendors: U.S., Europe
- Processors: x86 and ARM
- Initiatives: PRACE, ETP4HPC
- Cost: $300-$350

China

- Sustained: 2023
- Peak: 2020
- Vendors: China
- Processors: China, x86, ARM
- Initiatives: 13th 5-Year Plan
- Cost: $350-500M

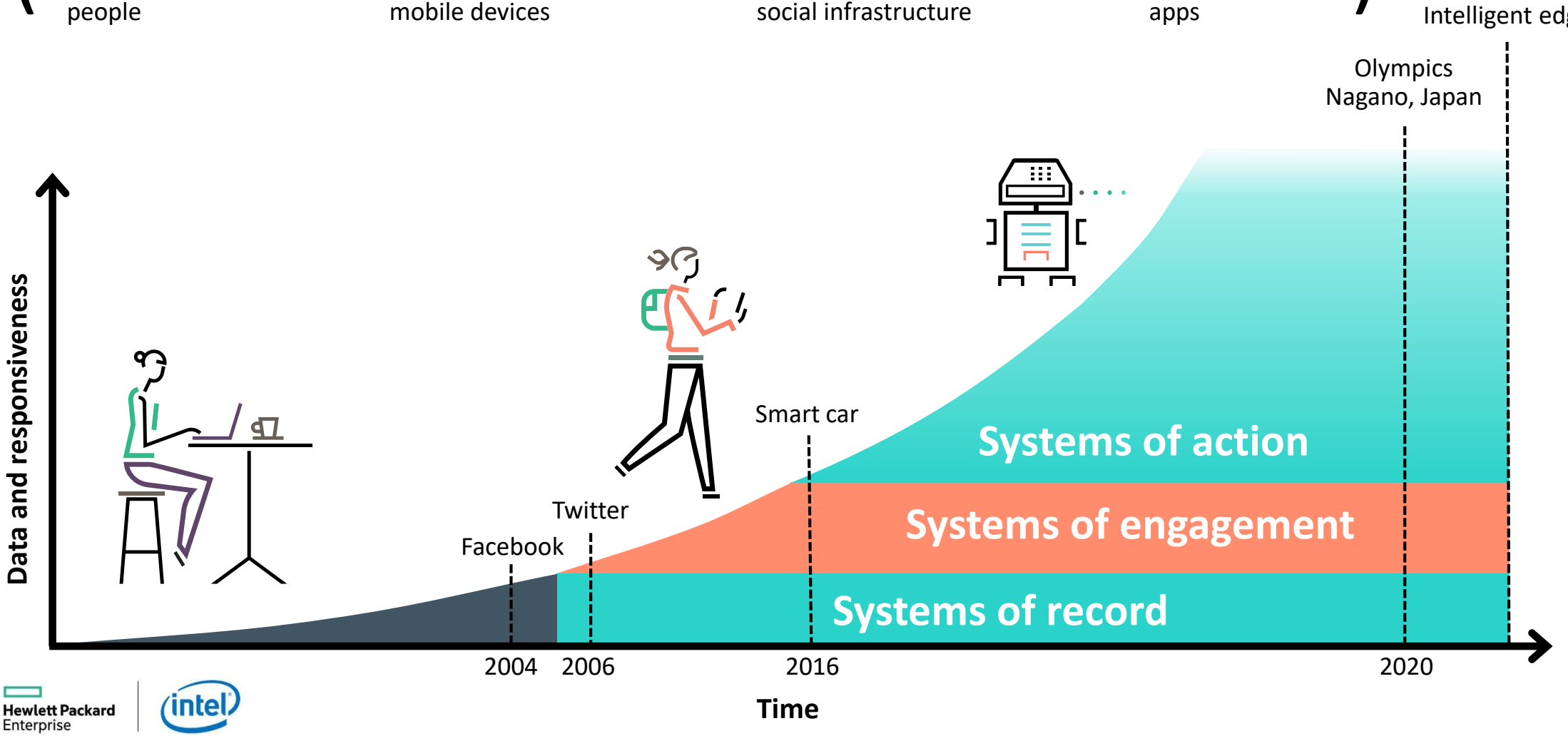
Japan

- Sustained: 2023-24
- Peak: Not planned
- Vendors: Japan
- Processors: ARM, Japan
- Initiatives: MEXT
- Cost: $600-800M

Source: Hyperion Research (IDC)

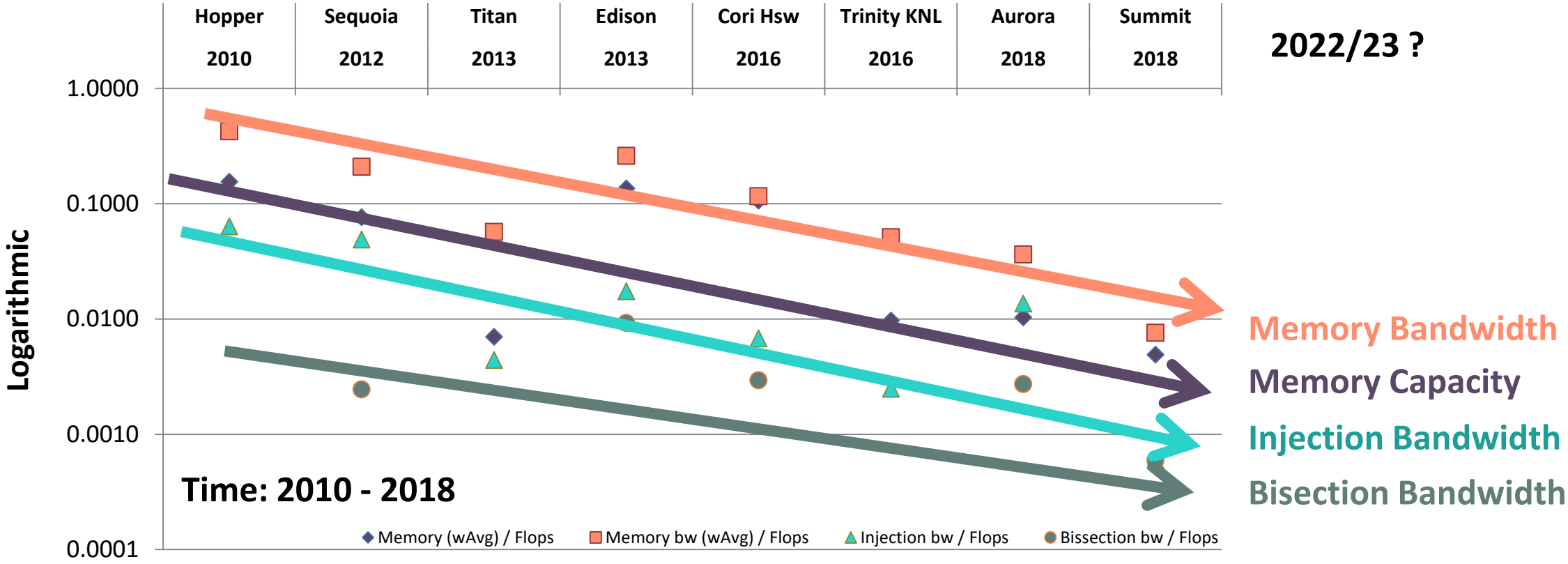
Scale of Computation is Inevitable

(**8B** people x **20B** mobile devices x **100B** social infrastructure x **1T** apps)



Hewlett Packard Enterprise | intel

Exascale Must Return Us to Reasonable System Balance



2022/23 ?

Memory Bandwidth

Memory Capacity

Injection Bandwidth

Bisection Bandwidth

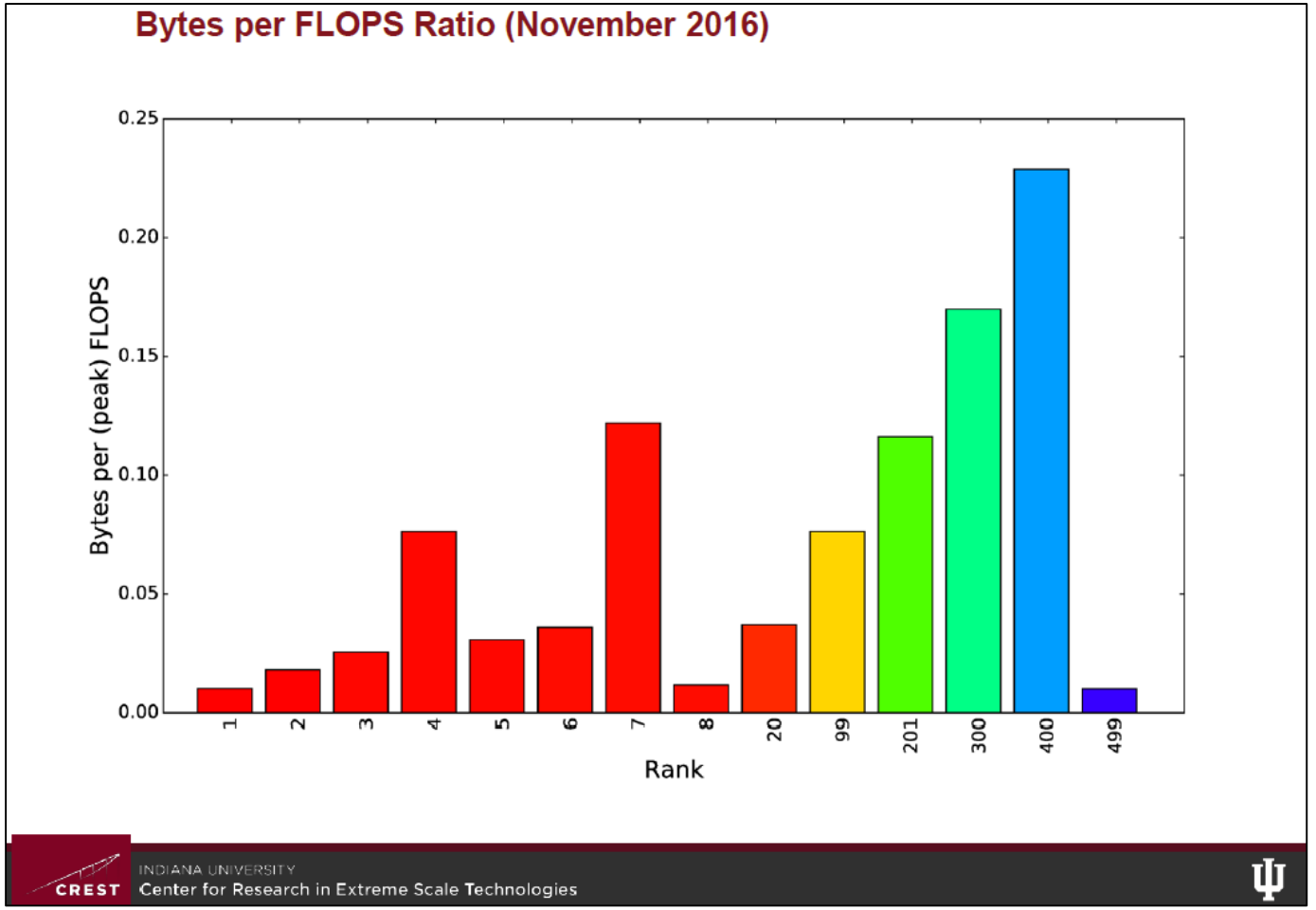
HPE's architecture objective: reignite progress in large scale HPC

Hewlett Packard Enterprise | intel

Balance + Scale is Really Difficult

- Bandwidth & capacity as a function of computation, at large scale, is difficult and expensive
- The system balance ratios continue to drop as systems have become laden with accelerators; this is great for some applications and not so great for others
- The system balance required in many Enterprise environments far exceed HPC systems
- Exascale technology advancements will address the challenges faced broadly across the entire Enterprise market

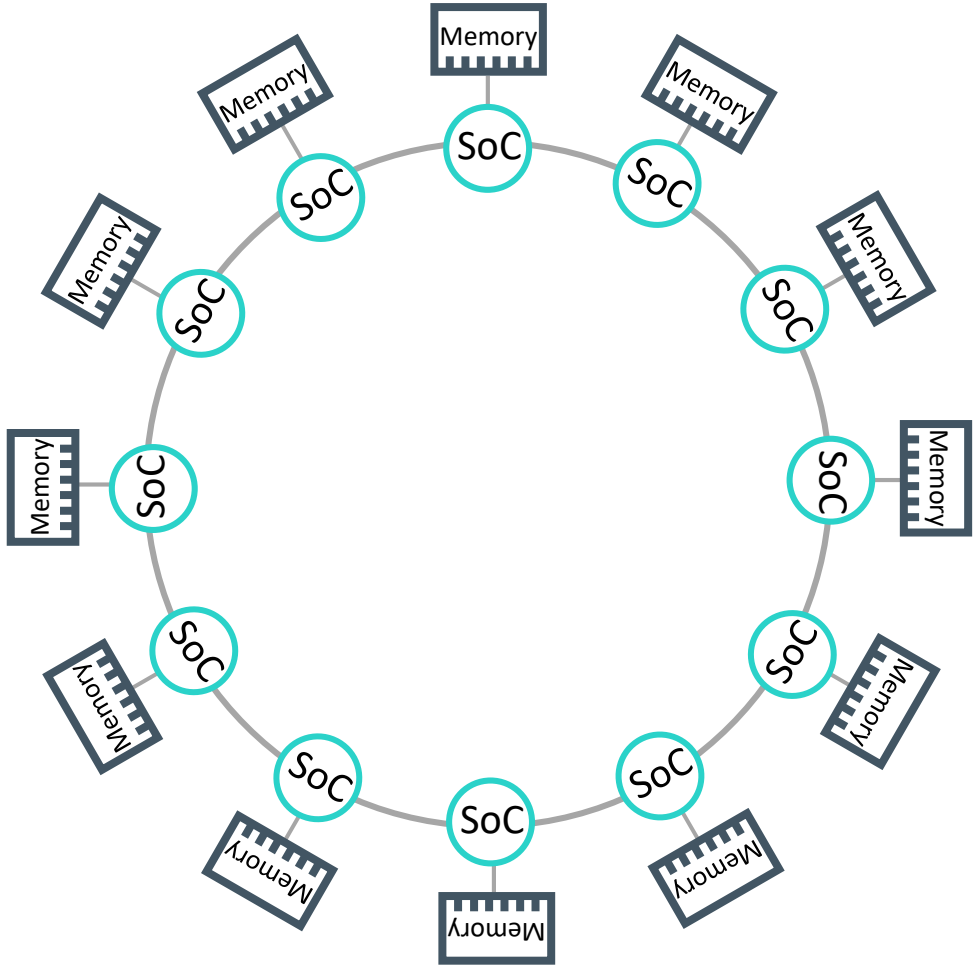
More scalability, less power, more bandwidth… while data movement consumes >10x more power than computation in future



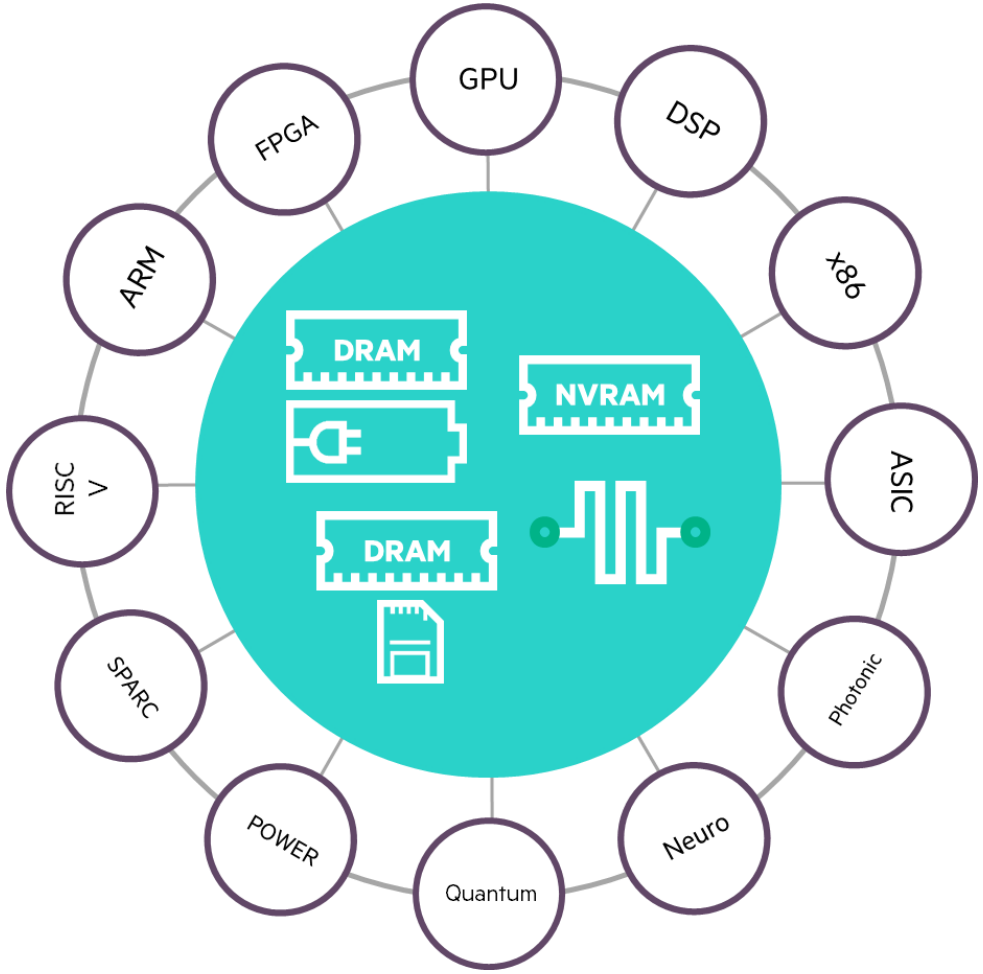
Source: Thomas Sterling, ISC'17 Keynote

Hewlett Packard Enterprise | intel

From processor-centric computing to Memory-Driven Computing



From processor-centric computing…

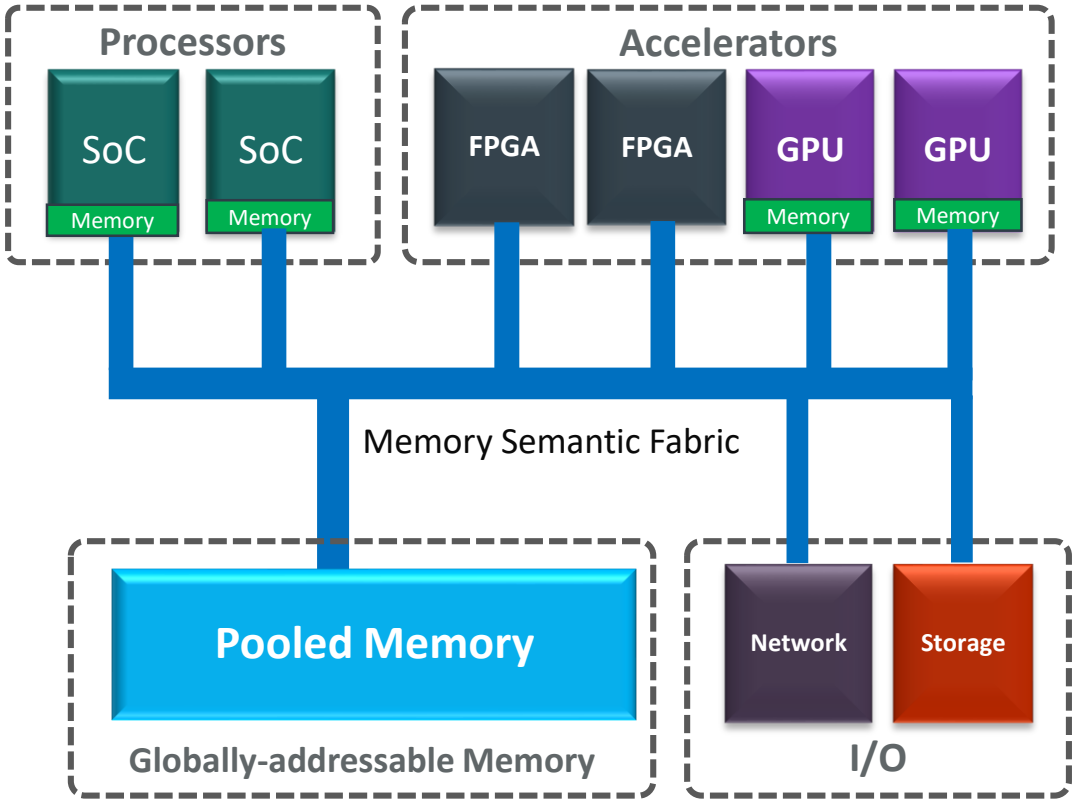


…to Memory-Driven Computing

Hewlett Packard Enterprise | intel

What is a “Memory Semantic” fabric?

- A communication protocol that speaks the **same language of the CPU ISA**:
  load and **store**, puts and gets, and atomic operations
- Today’s storage or network accesses are block based and managed by complex, code intensive, software stacks
- Memory semantics are optimal at random accesses, zero-copy, sub-microsecond operations, directly to CPU caches and registers



Hewlett Packard Enterprise | intel

Gen-Z Broad Industry and Device Support Ecosystem

GEN Z Consortium Members

ALPHA DATA, AMD, Amphenol, ARM, avery design systems, BROADCOM, cadence, CAVIUM, CRAY, DELL EMC, EVERSPIN TECHNOLOGIES, FIT, Hewlett Packard Enterprise, HUAWEI, IBM, IDT, IntelliProp, JABIL, Lenovo, LOTES, LUXSHARE ICT, Mellanox TECHNOLOGIES, Mentor Graphics, Micron, Microsemi, molex, NetApp, NOKIA, NUMASCALE, PLDA GROUP, redhat, SAMSUNG, SEAGATE, SK hynix, SMART Modular Technologies, Spin Transfer Technologies, TE connectivity, WD Western Digital, XILINX, YADRO

- Components
- Intellectual Property
- Connectors
- Subsystems
- Systems
- Software

GEN Z Component Categories

CPUs

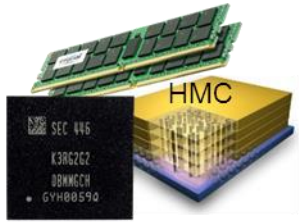

GPUs

Special Purpose Processors

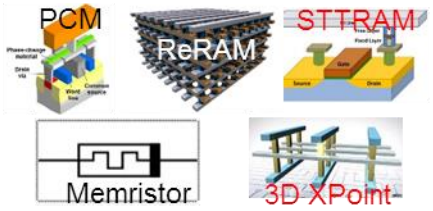
Storage Controllers

Switches/ Gateways

FPGAs



Volatile Memory



Storage Class Memory

http://www.genzconsortium.org

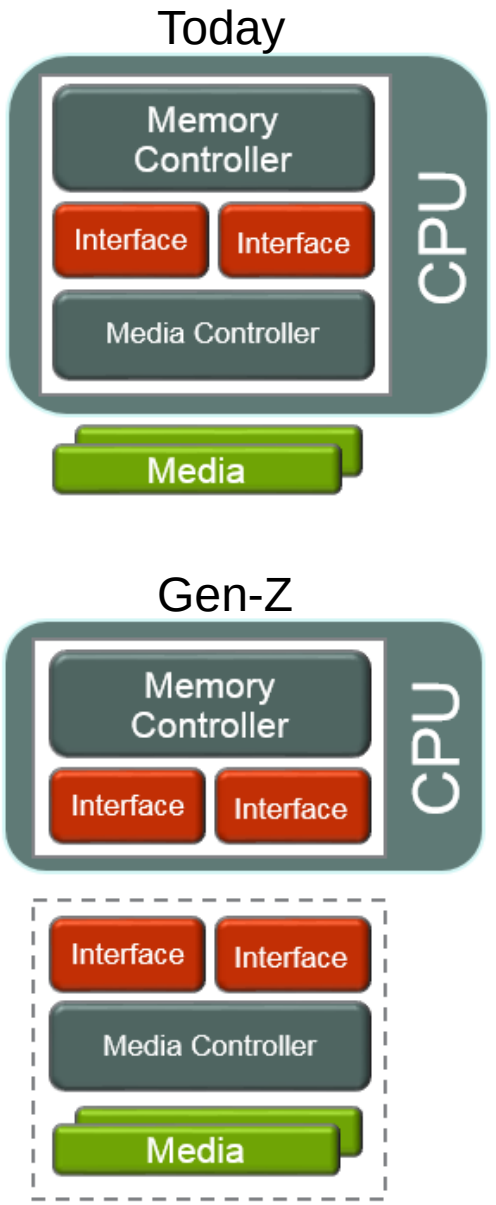
Hewlett Packard Enterprise | intel

Gen-Z Breaks Processor-Memory Interlock

A “split controller” model

- Memory controller (gets simplified)
  - Initiates high-level requests—Read, Write, Atomic, Put / Get, etc.
  - Enforces ordering, reliability, path selection, etc.
  - Memory semantic protocol routable over networks
- Media controller (handles the media specific detail)
  - Abstracts memory media
    - Supports volatile / non-volatile / mixed-media
  - Performs media-specific operations
  - Enables data-centric computing (accelerator, compute, etc.)
  - Enables open ecosystem for memory/storage and I/O

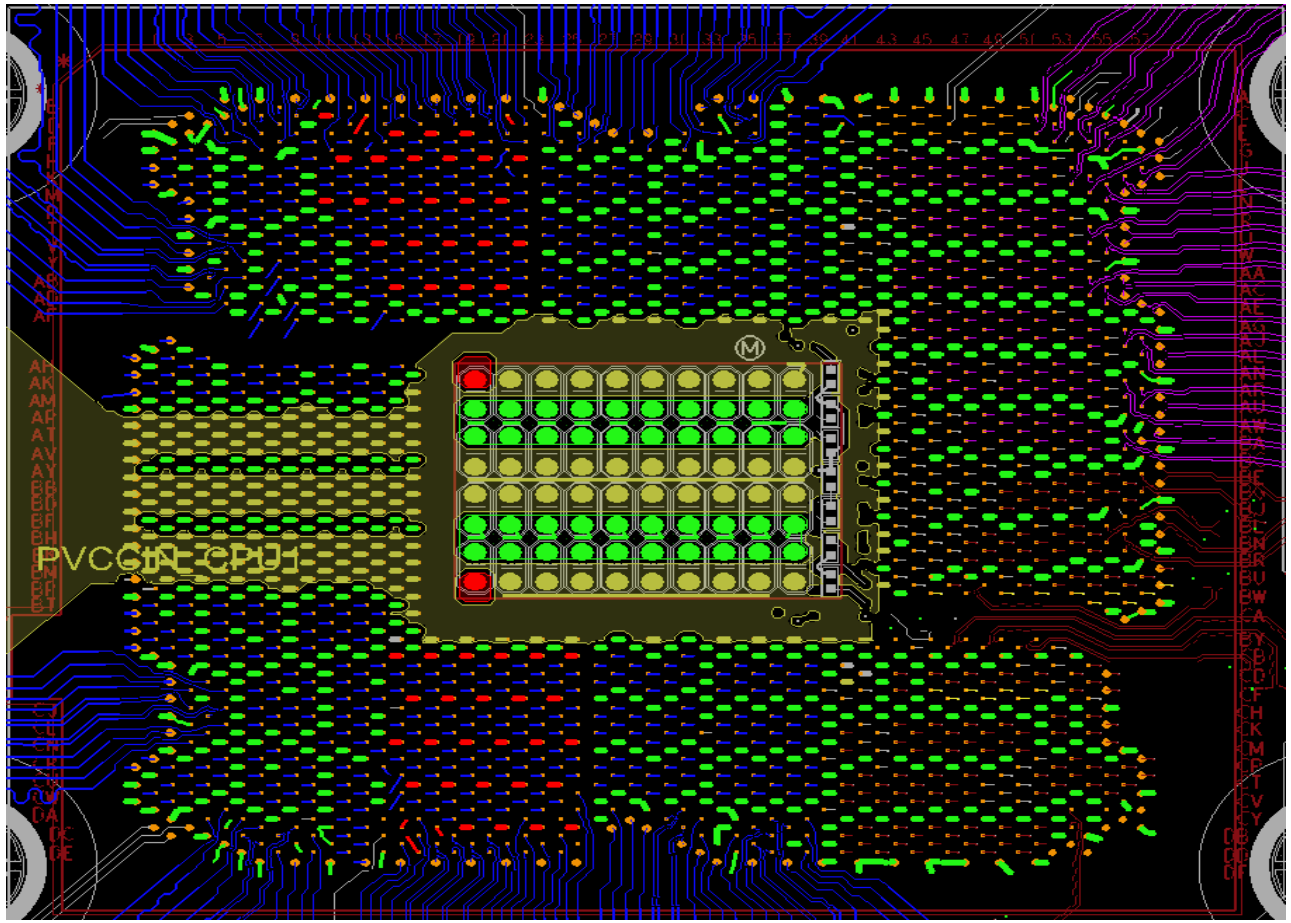
DDR, PCIe, SAS, SATA and other dedicated pins on the CPU package replaced with Gen-Z pins, providing greater configurability



Hewlett Packard Enterprise | intel

Gen-Z's Serial Links Make More Efficient Use of Pins

30x more bandwidth than DDR4 with fewer pins



Pin Type	Color	Count
DDR Signals	Blue	660
DDR Power	Red	58
DDR Ground	Green	342
DDR Total		**1060**
PCIe Data	Violet	192
PCIe Ground	Green	120
PCIe Total		**312**
Proprietary	Dark Red	175
Proprietary Ground	Green	110
Proprietary Total		**285**
PCH/Power/Ground/Misc	Orange/Green	354
Grand Total		**2011**

Hewlett Packard Enterprise | intel

HPE Exascale strategy centered around Gen-Z

High Bandwidth Low Latency

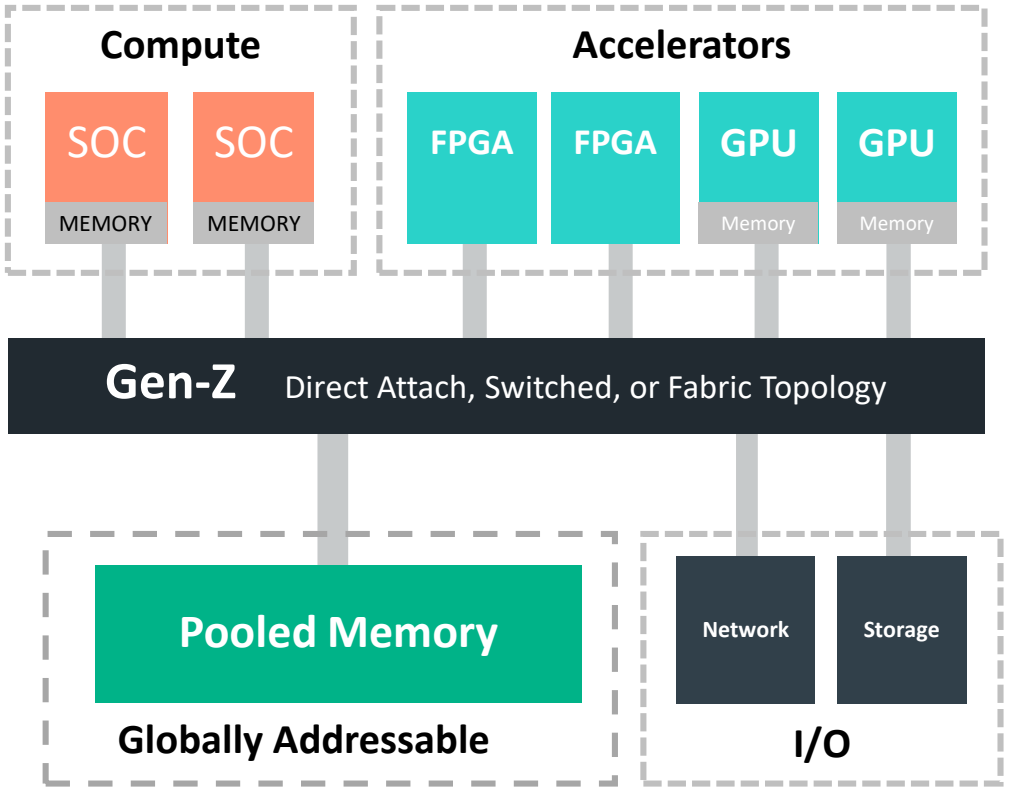
- Memory Semantics – simple Reads and Writes
- From tens to several hundred GB/s of bandwidth
- Sub-100 ns load-to-use memory latency

Advanced Workloads & Technologies

- Real time analytics
- Enables data centric and hybrid computing
- Scalable memory pools for in memory applications
- Abstracts media interface from SoC to unlock new media innovation

Secure Compatible Economical

- Provides end-to-end secure connectivity from node level to rack scale
- Supports unmodified OS for SW compatibility
- Graduated implementation from simple, low cost to highly capable and robust
- Leverages high-volume IEEE physical layers and broad, deep industry ecosystem



Hewlett Packard Enterprise | intel

“The Machine” Project

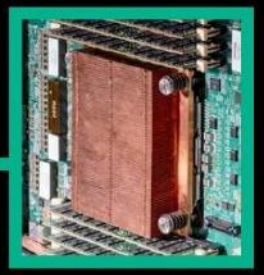
Nov 2016: Building blocks demo’d
Compute, Memory, Fabric, Switch
May 2017: 160 TiB 40-node prototype
Jun 2017: The Machine user group +
PathForward announcement

- Compute nodes accessing a shared pool of Fabric-Attached Memory;
- An optimized Linux-based operating system on a ARM-based SoC (Cavium Thunder-X2)
- Optical communication links, including the new X1 photonics module online and operational
- New software programming tools designed to take advantage of abundant persistent memory

https://www.labs.hpe.com/the-machine/user-group
https://www.labs.hpe.com/the-machine/developer-toolkit

Hewlett Packard Enterprise

MEMORY FABRIC SWITCH
Enables processors to access Fabric-Attached Memory across any node on the system.

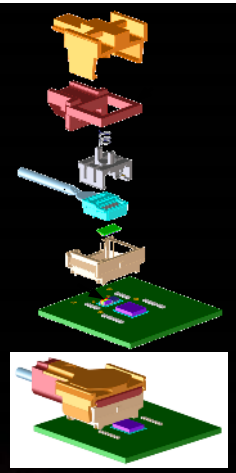


TASK-SPECIFIC PROCESSING
Flexible Memory-Driven Computing architecture can match compute tasks to different types of processor to optimize performance and efficiency.

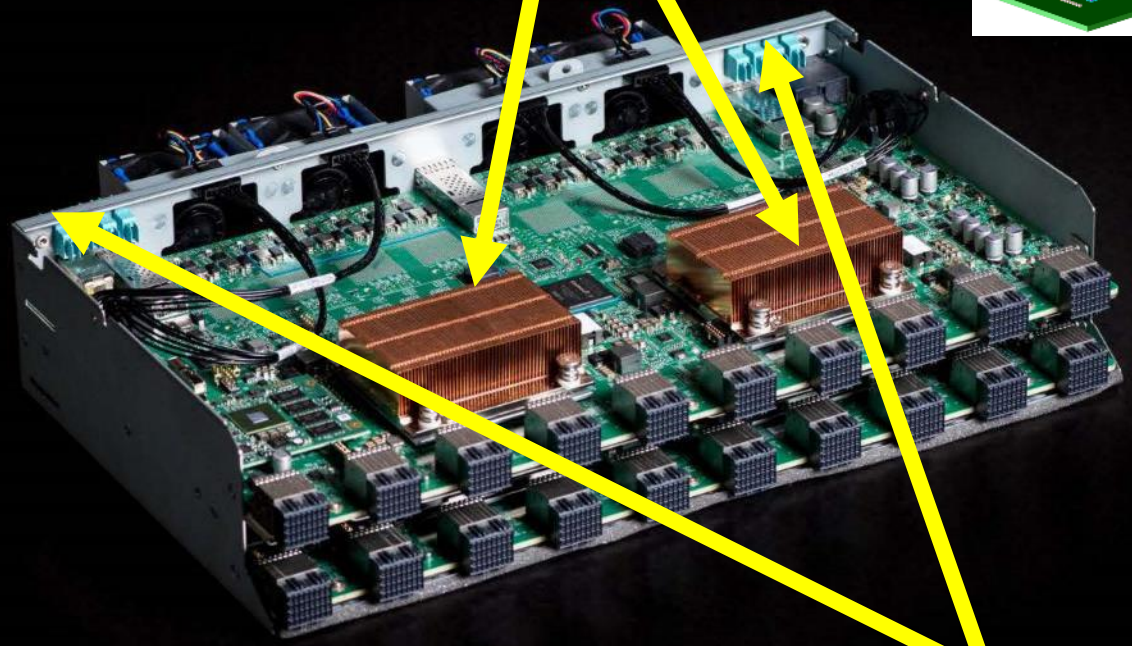
PHOTONICS INTERCONNECTS
Rapidly transfers data between enclosures with light instead of electricity to access shared memory.

MEMORY AT THE CENTER
Combines memory and storage into a vast pool of Fabric-Attached Memory to radically increase computing efficiency and speed by enabling multiple processors to share memory.

Memory Fabric
Switches



High-speed
High-density
Small form factor
Low power
Low cost
VCSEL
optics




ENERGY.GOV
SCIENCE & INNOVATION
ENERGY ECONOMY
SECURITY & SAFETY
SAVE ENERGY, SAVE MONEY
Department of Energy
Department of Energy Awards Six Research Contracts Totaling $258 Million to Accelerate U.S. Supercomputing Technology
JUNE 15, 2017

DOE PathForward Goals

- High Level Approach
  - Close gaps in vendor's technology roadmaps or accelerate time to market to address ECP performance targets
  - Provide an opportunity for Application Development and Software Technology to *influence* the design of future node and system architecture designs
  - Deliver hardware technology analysis and (where appropriate), demonstrations to increase confidence in node and system design performance benefit, programmability and ability to affect a 2019 Exascale System RFP
- Goals
  - Ensure that laboratory platform acquisition teams have quantitative information to identify the most promising technology options to include in the 2019 Exascale System RFP
  - Improve vendor's confidence in the value and feasibility of aggressive advanced technology options that they may propose for 2019 Exascale System RFP

Exascale systems require a new scalable and balanced architecture

Exascale Requirements

Scalability
10x greater performance than today's largest systems

Productivity
Increase system utilization by 6x

Power
20-30 megawatts of power consumption (10x more efficient than today)

Resiliency
failure reduced to <= 1week

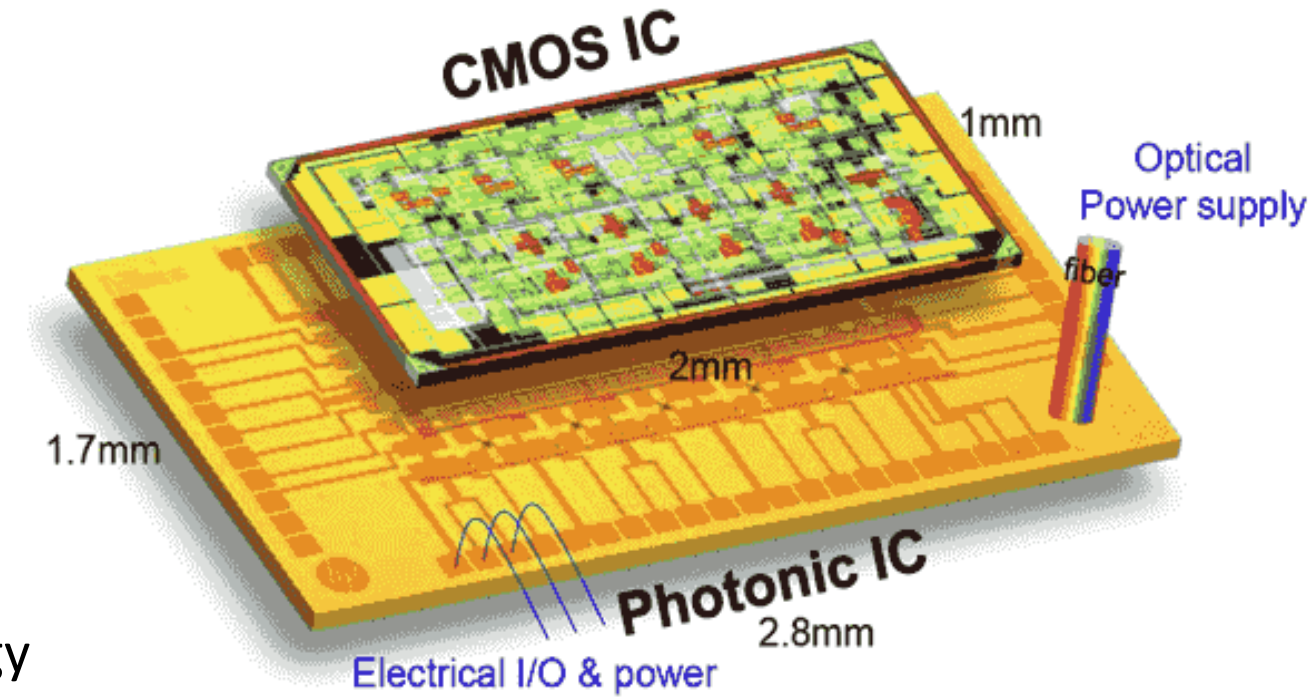
Applications
Optimized for a broad spectrum of workloads

Hewlett Packard Enterprise | intel

Example: Why are Better Photonics Needed?

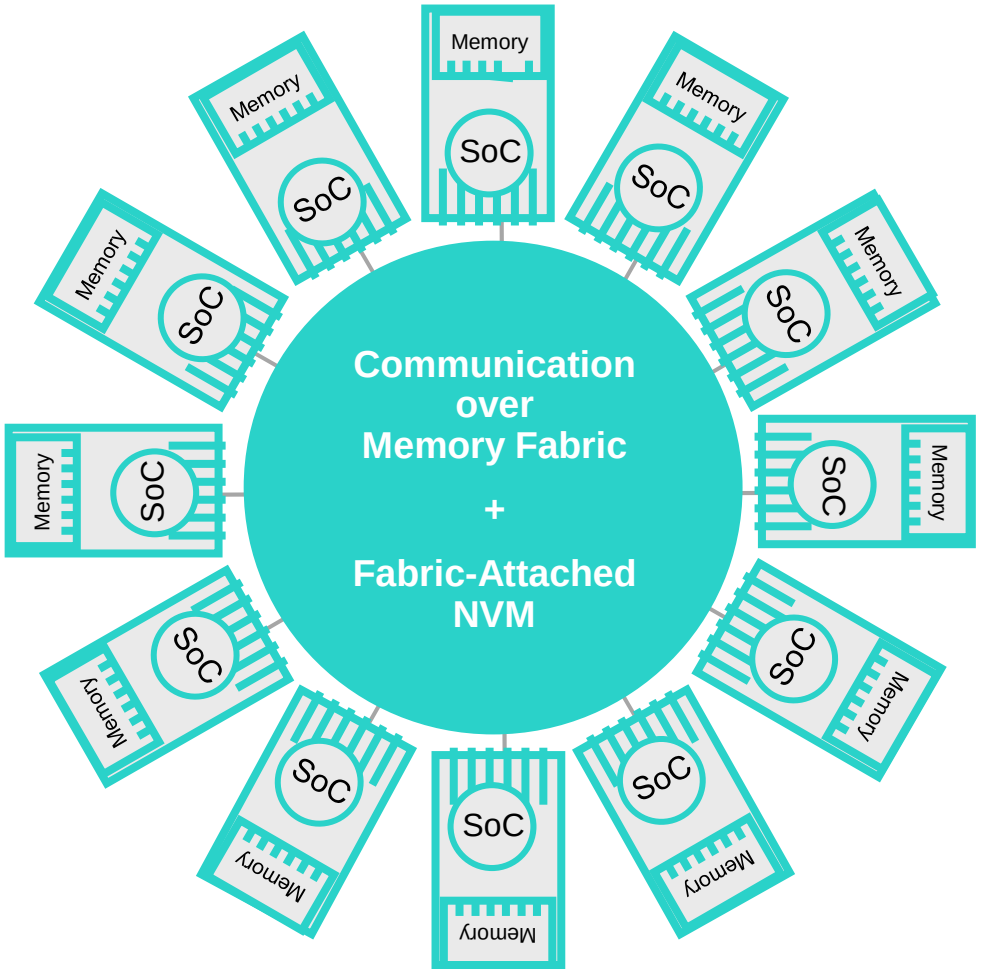
- **Today's active optical cables**
  - Discrete components
  - $1,000 for 100Gbps, at 4W ($10/Gbps, 40 pJ/bit)
  - At Exascale: ~$1B and >10MW (only the cables!)

- **Silicon Photonics**
  - Integrated CMOS and Photonic ICs
  - Target: 100x lower cost, 10x lower energy



Hewlett Packard Enterprise | intel

Gen-Z and The Machine → Exascale



Open Memory Semantics Gen-Z Fabric ❭ **Improve Data Movement Efficiency**

Fabric-attached non-volatile memory ❭ **Improve Resilience and System efficiency**

Silicon Photonics ❭ **Improve Energy and Cost efficiency**

HPE Exascale Architecture Innovation

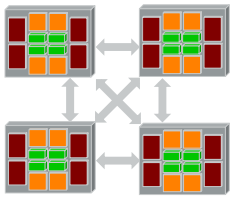
infrastructure optimized for running a broad range of applications

Design Tenets	Technology Improvements
Improved scalability and application performance at scale	Improve Latency
	Increase bandwidth - Reduced data movement at scale
Improve Resilience and System efficiency	Less hardware specific optimization and tuning
	Optimize system model (One address space for memory and storage)
Improved energy efficiency and reduced cost	Silicon Photonics for low latency and reducing data movement at scale
	Power efficient electronics
	Intelligent throttling based on electricity/rate

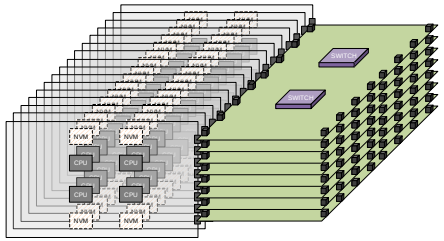
Optimized for a broad range of applications with minimum customization

Hewlett Packard Enterprise | intel

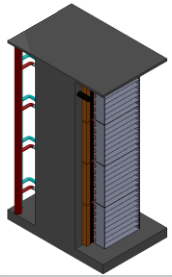
Notional Exascale System: 2020 - 2022



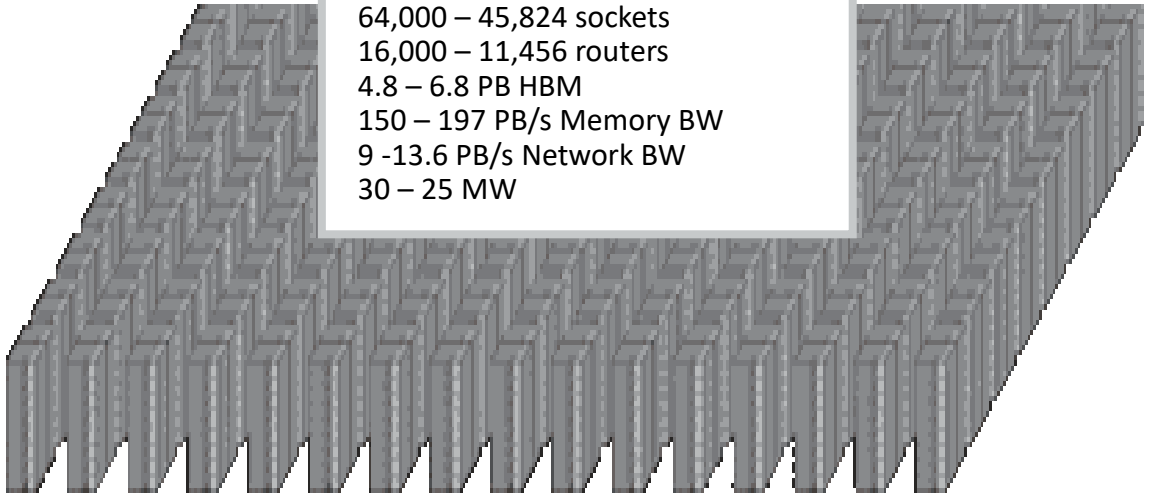
Notional Board
4 sockets / 1 node
60 - 90 TFLOPS
0.3 – 0.6 TB HBM
10 - 18 TB/s memory BW
0.6 – 1.2 TB/s network BW
1,500 – 1,700 W



Notional Chassis
8 trays/16 nodes
64 sockets / 16 routers
1 – 1.4 PFLOPS
4.8 – 9.6 TB HBM
160 - 288 TB/s Memory BW
9 – 19 TB/s Network BW
30 – 35 kW



Notional Rack
4 chassis / 32 trays / 64 nodes
256 sockets / 64 routers
4 – 5.6 PFLOPS
19 - 38 TB HBM
0.6 – 1.1 PB/s Memory BW
36 - 76 TB/s Network BW
120 – 140 kW / 100% Liquid-Cooled



1 ExaFLOPS Peak
250 – 179 Racks
16,000 – 11,456 nodes
64,000 – 45,824 sockets
16,000 – 11,456 routers
4.8 – 6.8 PB HBM
150 – 197 PB/s Memory BW
9 -13.6 PB/s Network BW
30 – 25 MW

Exascale R&D is Addressing Enterprise Challenges of the Future

the MACHINE

Faster to result

Faster to program

More compact

More energy-efficient

More scalable

More reliable

More cost-efficient

More secure

Hewlett Packard Enterprise

intel

eRESEARCH AUSTRALASIA 2017 | 16-20 OCTOBER BRISBANE CONVENTION AND EXHIBITION CENTRE

Thank You