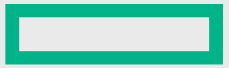


Hewlett Packard
Enterprise

December, 2015

Adapting to Thrive in a New Economy of Memory Abundance

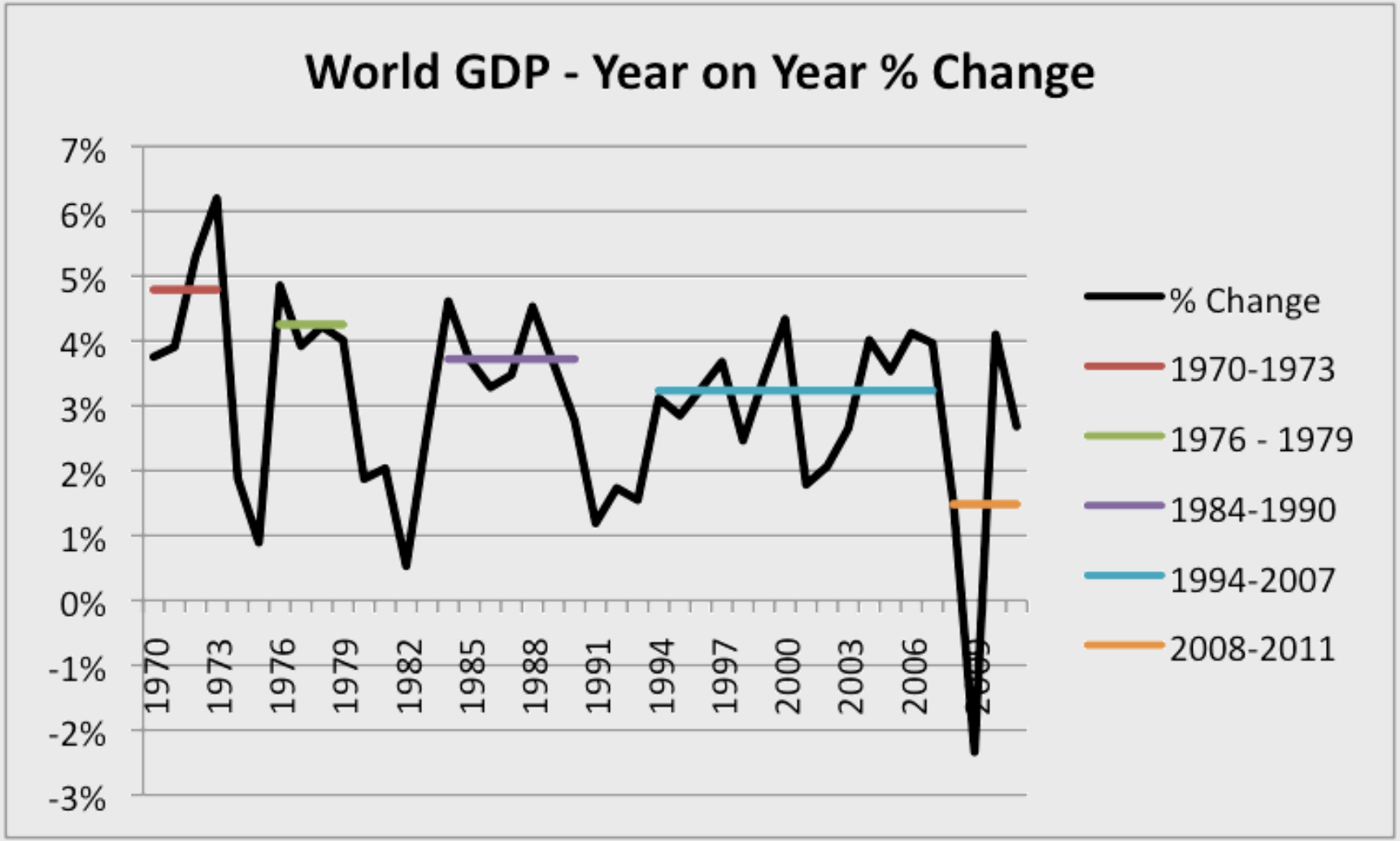
Kirk Bresniker, Hewlett Packard Labs Chief Architect, HPE Fellow

This result deserves to be noted. It shows in a most striking way where the real difficulty, the main bottleneck, of an automatic very high speed computing device lies.

Clearly the practicality of a device as is contemplated here depends most critically on the possibility of building such an M[emory], and on the question of how simple such an M[emory] can be made to be

John von Neumann (1945), First Draft of a Report on the EDVAC

Moore's Law kept up with data creation in the traditional economy



Source: http://ourfiniteworld.com/2012/07/18/how-much-oil-growth-do-we-need-to-support-world-gdp-growth/

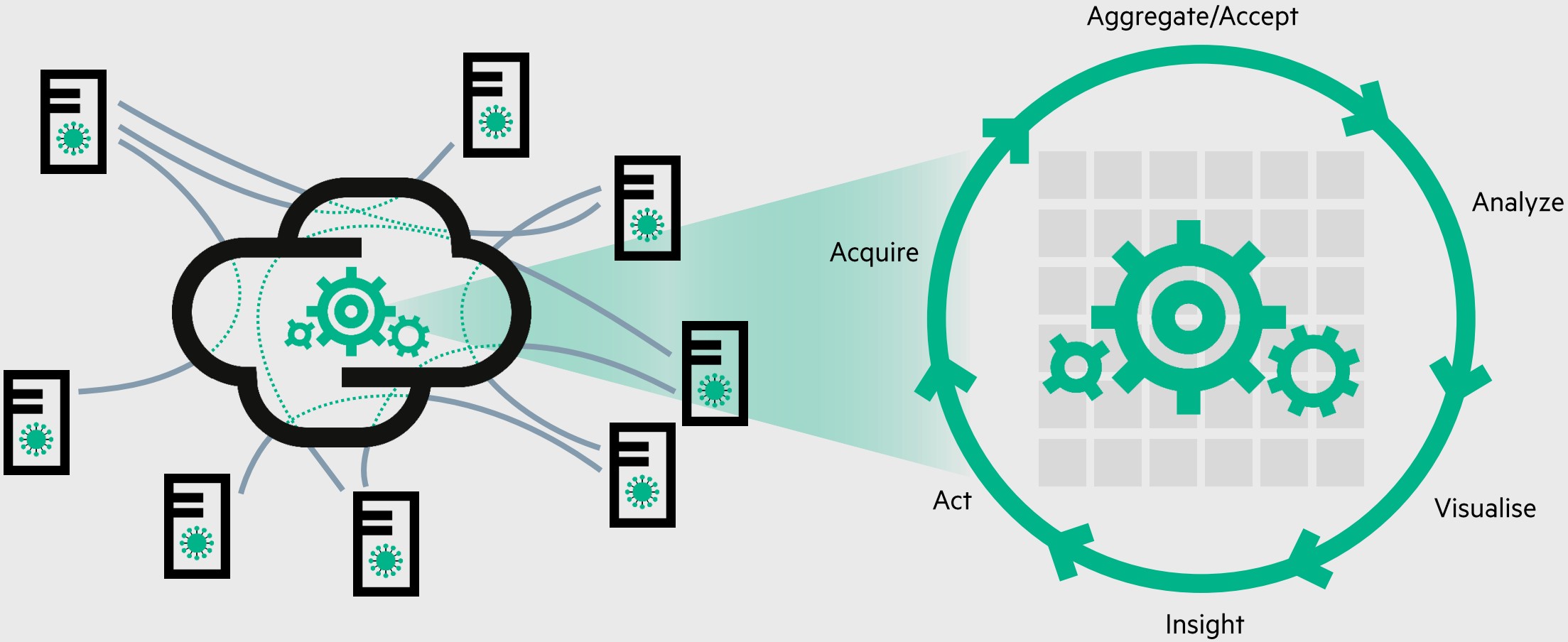
The end of scaling at just the wrong time …

8B × 20B × 100B × 1T

People | Mobile Devices | Social Infrastructure | Apps

Systems of Record ➔ Systems of Engagement ➔ Systems of Action

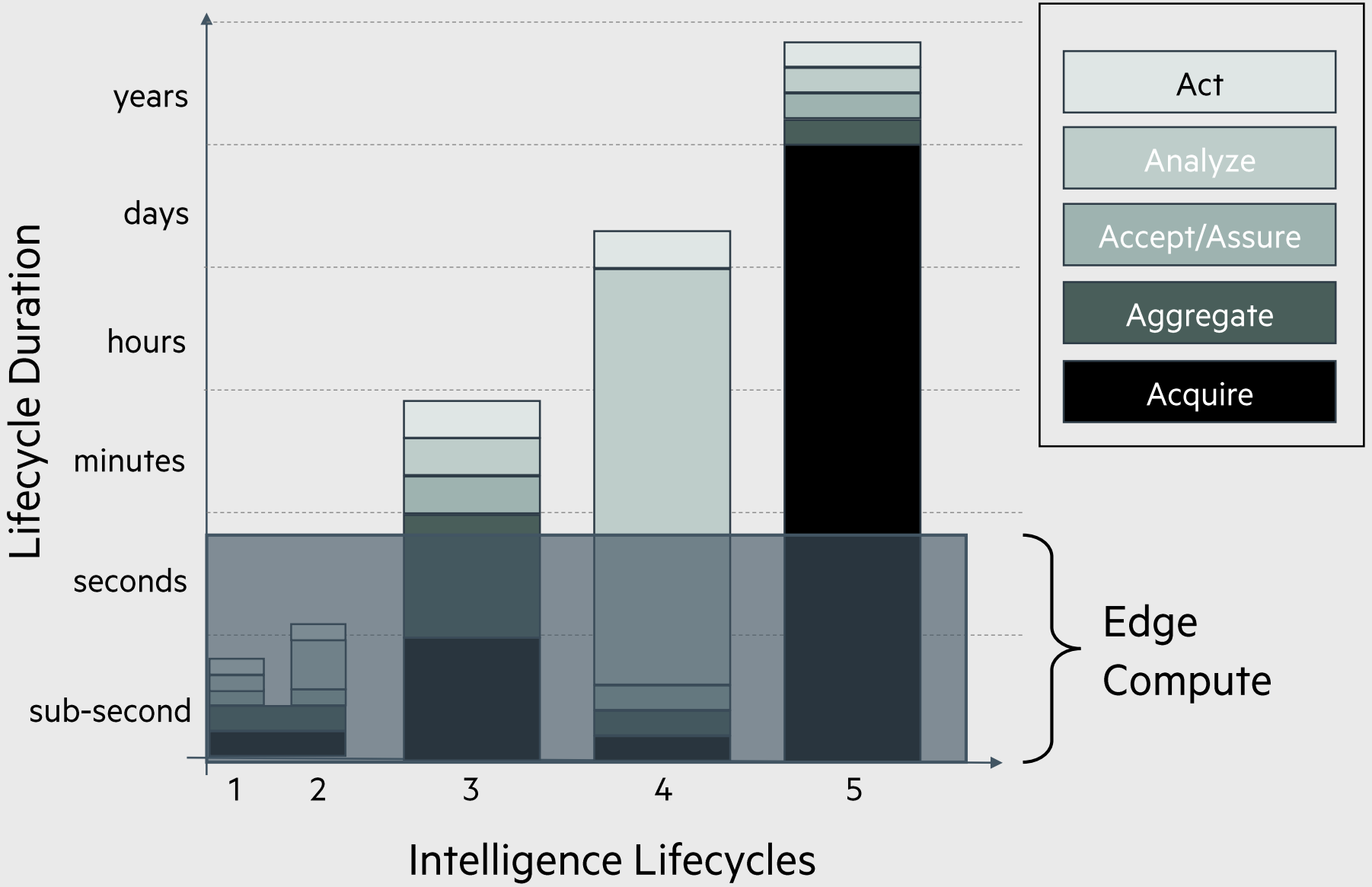
Analytics + visualization = actionable insight



Lifecycle of information

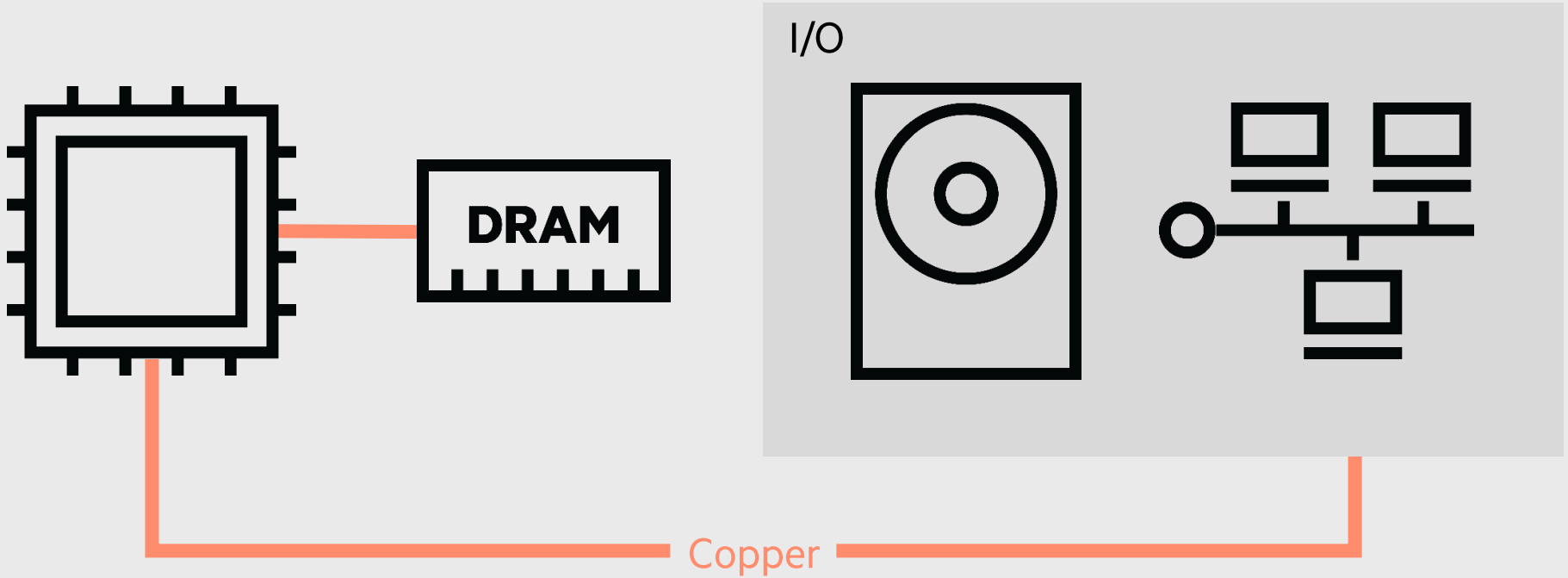
This graph is used to illustrate that:

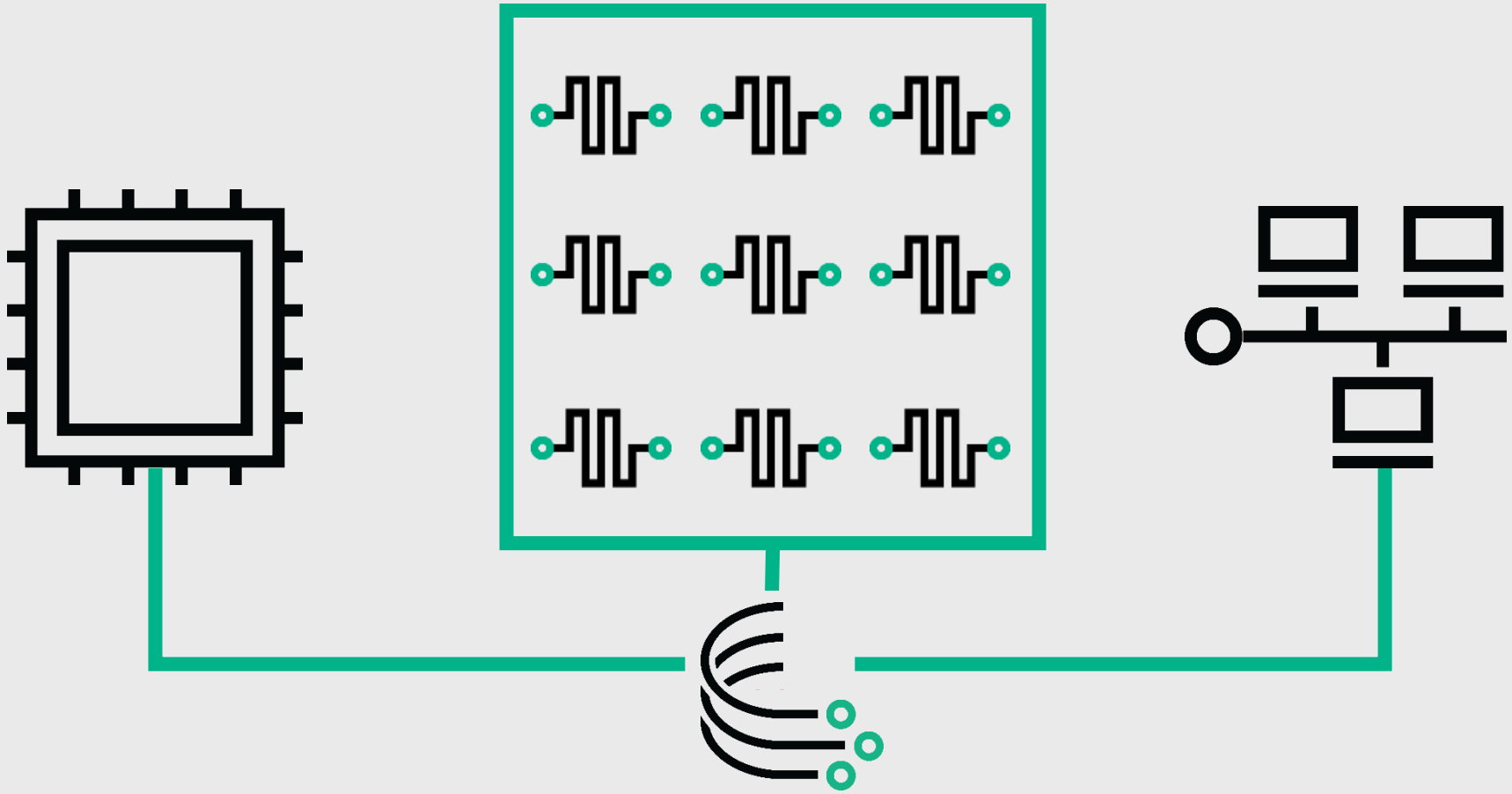
- The duration of intelligence lifecycles can differ
- Intelligence “apps” can overlap on one or more stages
- An organization can have multiple co-existing intelligence apps of varying durations

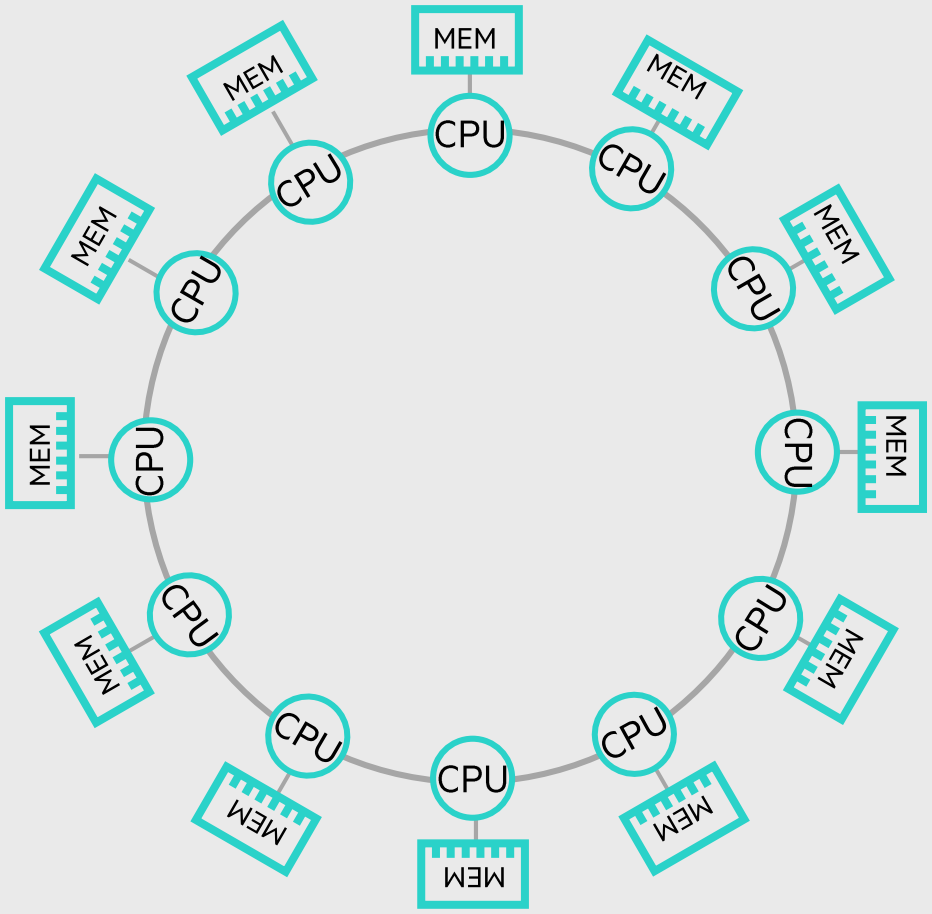


I must Create a System, or be enslav'd by another Mans I will not Reason & Compare: my business is to Create

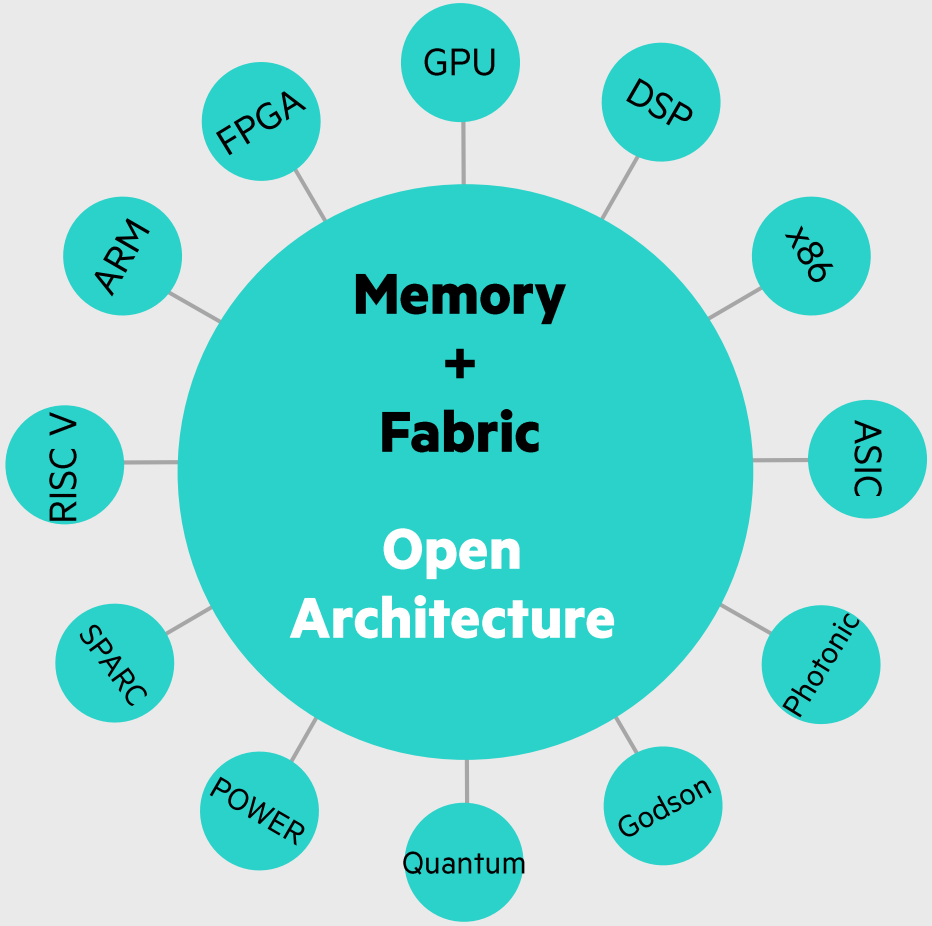
William Blake, "Jerusalem The Emanation of the Giant Albion", 1804-1820


I/O
DRAM
Copper





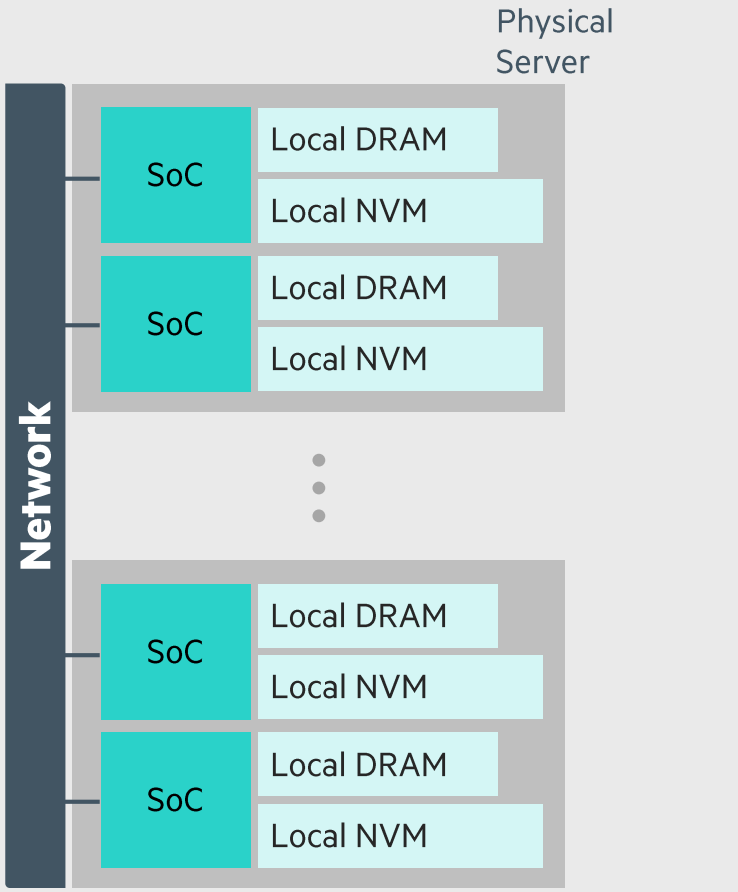
Processor-Centric Computing



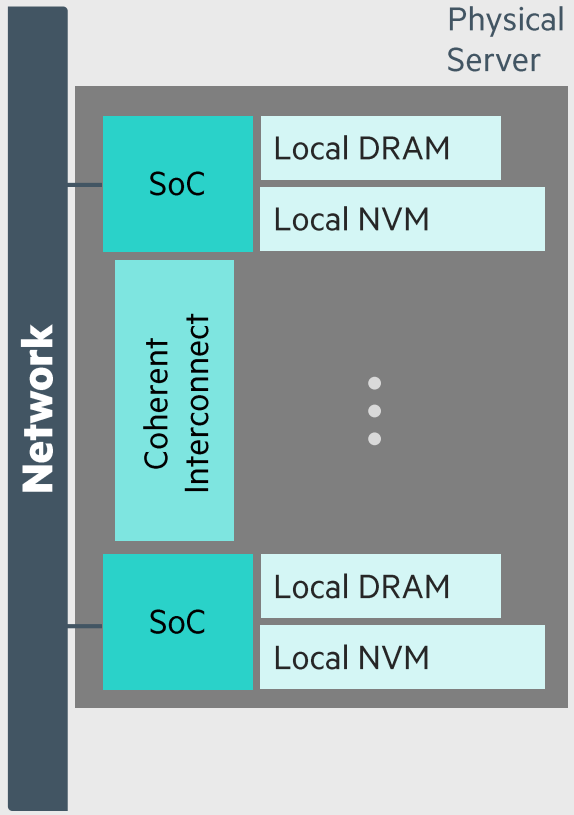
Memory-Driven Computing

Hewlett Packard Enterprise

The Machine in context



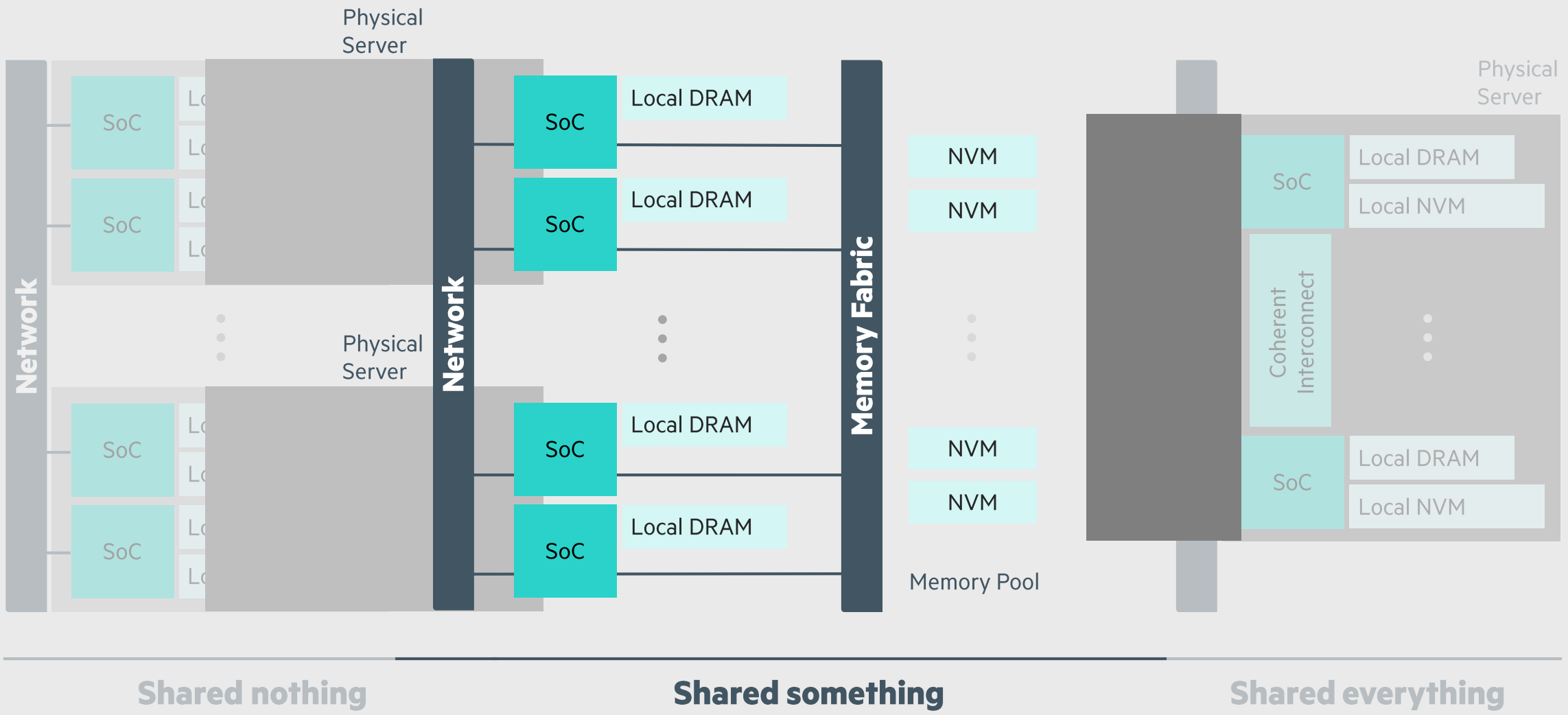
Shared nothing



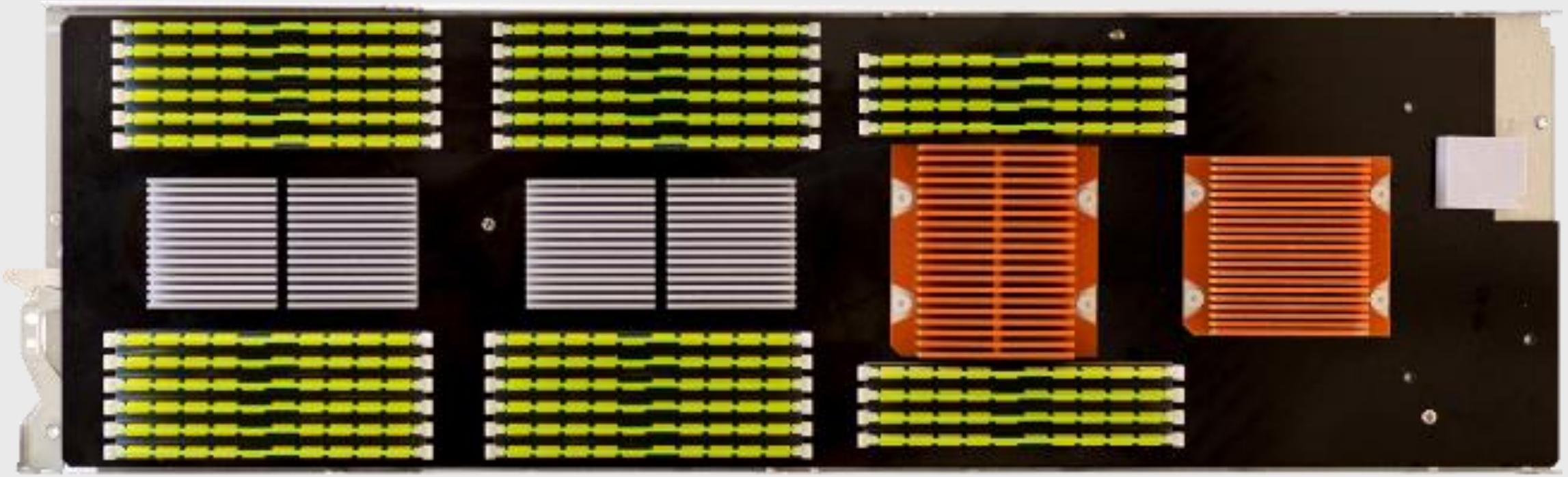
Shared everything

Hewlett Packard
Enterprise

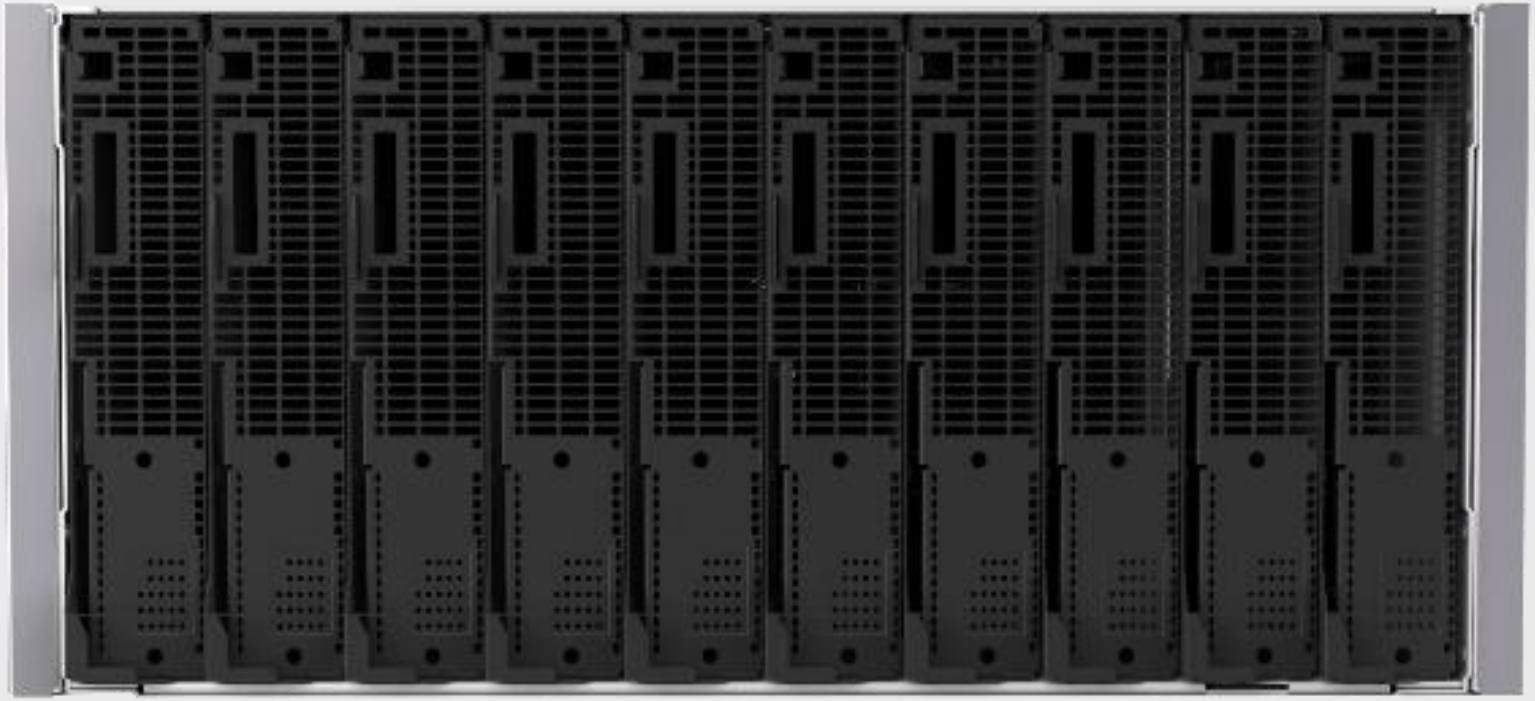
The Machine in context



Interesting at every scale from Node

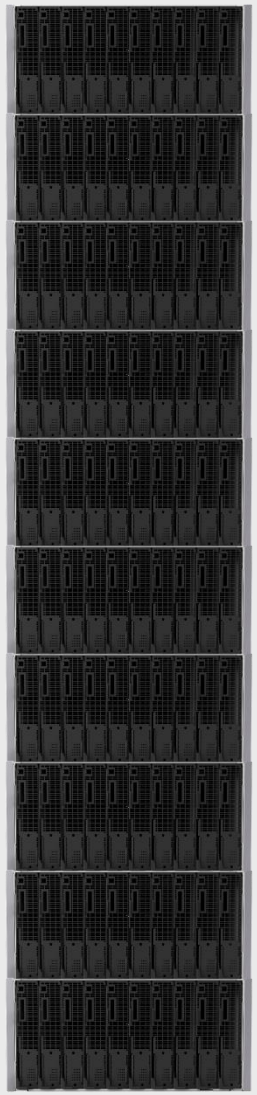


To enclosure



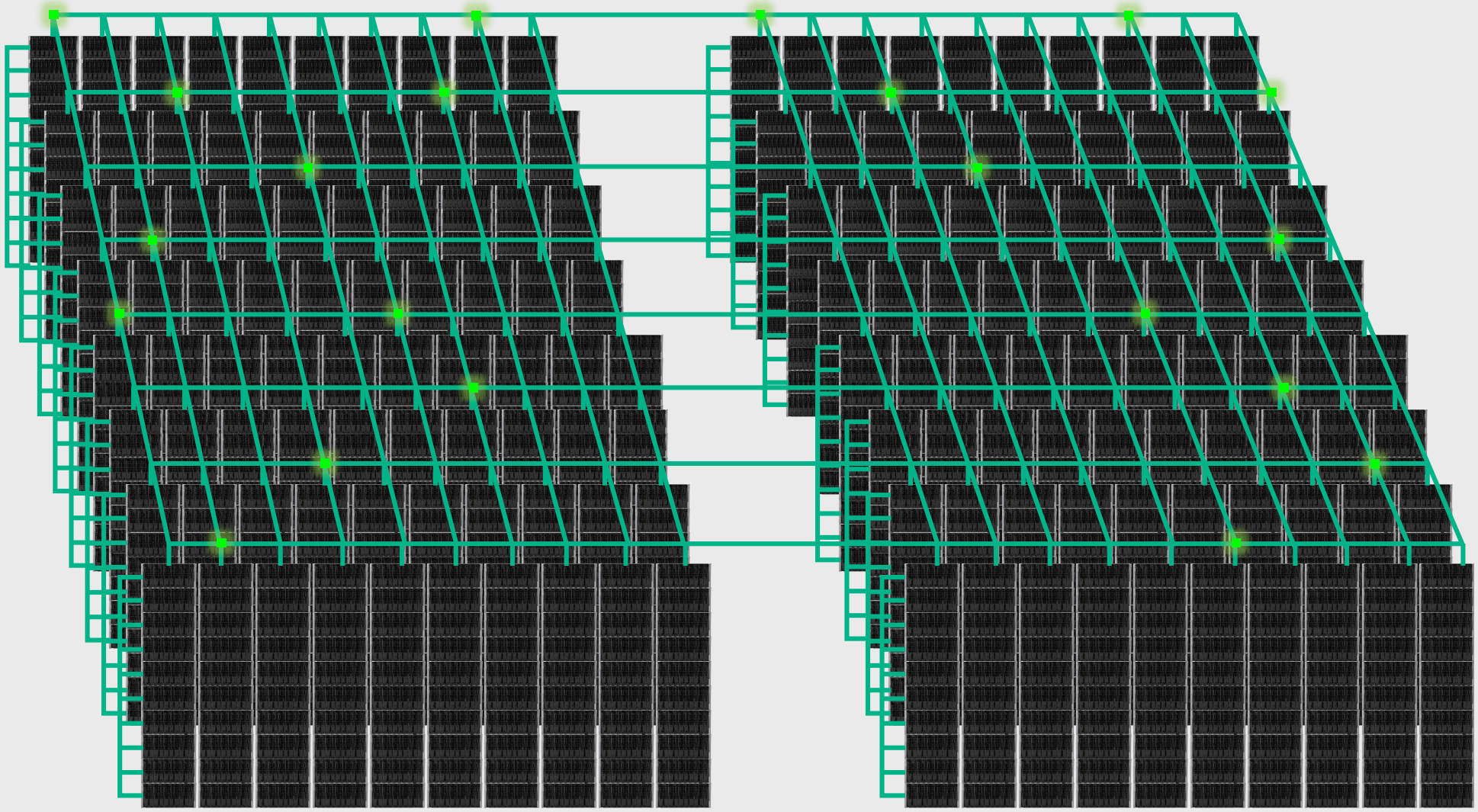
Hewlett Packard
Enterprise

To Rack



Hewlett Packard
Enterprise

To Data Center, hundreds of racks, hundreds of PBs of fabric attached memory



Hewlett Packard
Enterprise

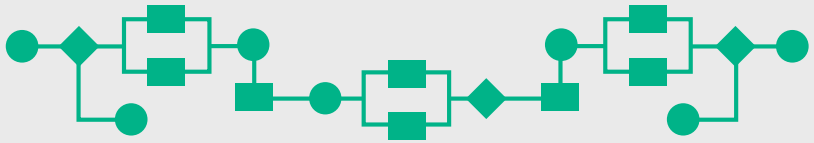
Hardware + software stack

Insight

Analytics and visualization

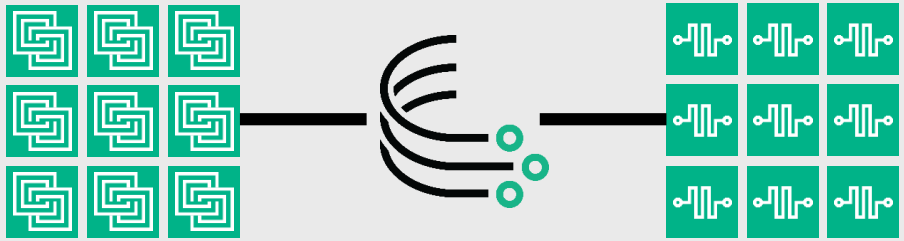
Exabyte-scale algorithms

Million-node management

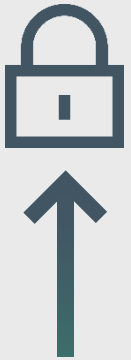


Operating systems and programming models

Ultra-efficient hardware

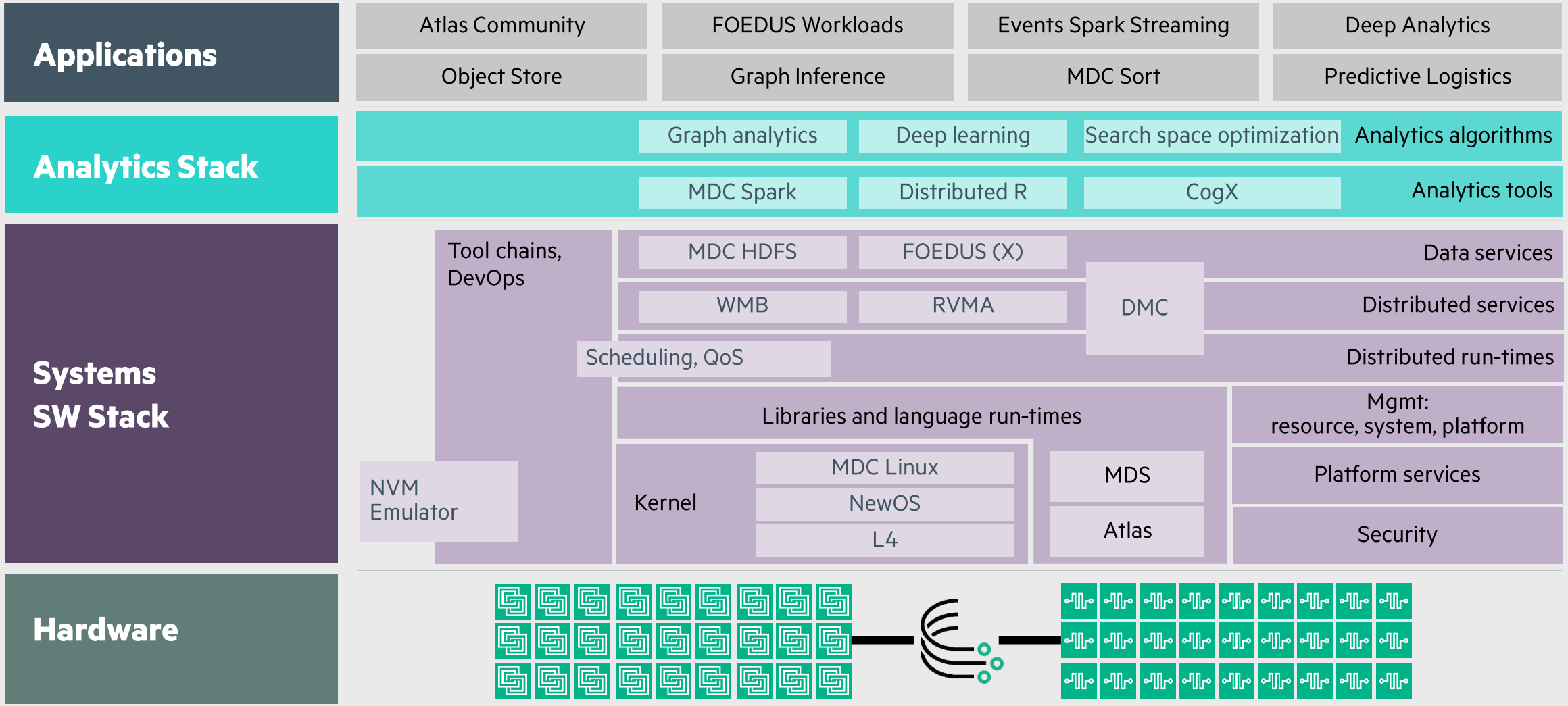


Data



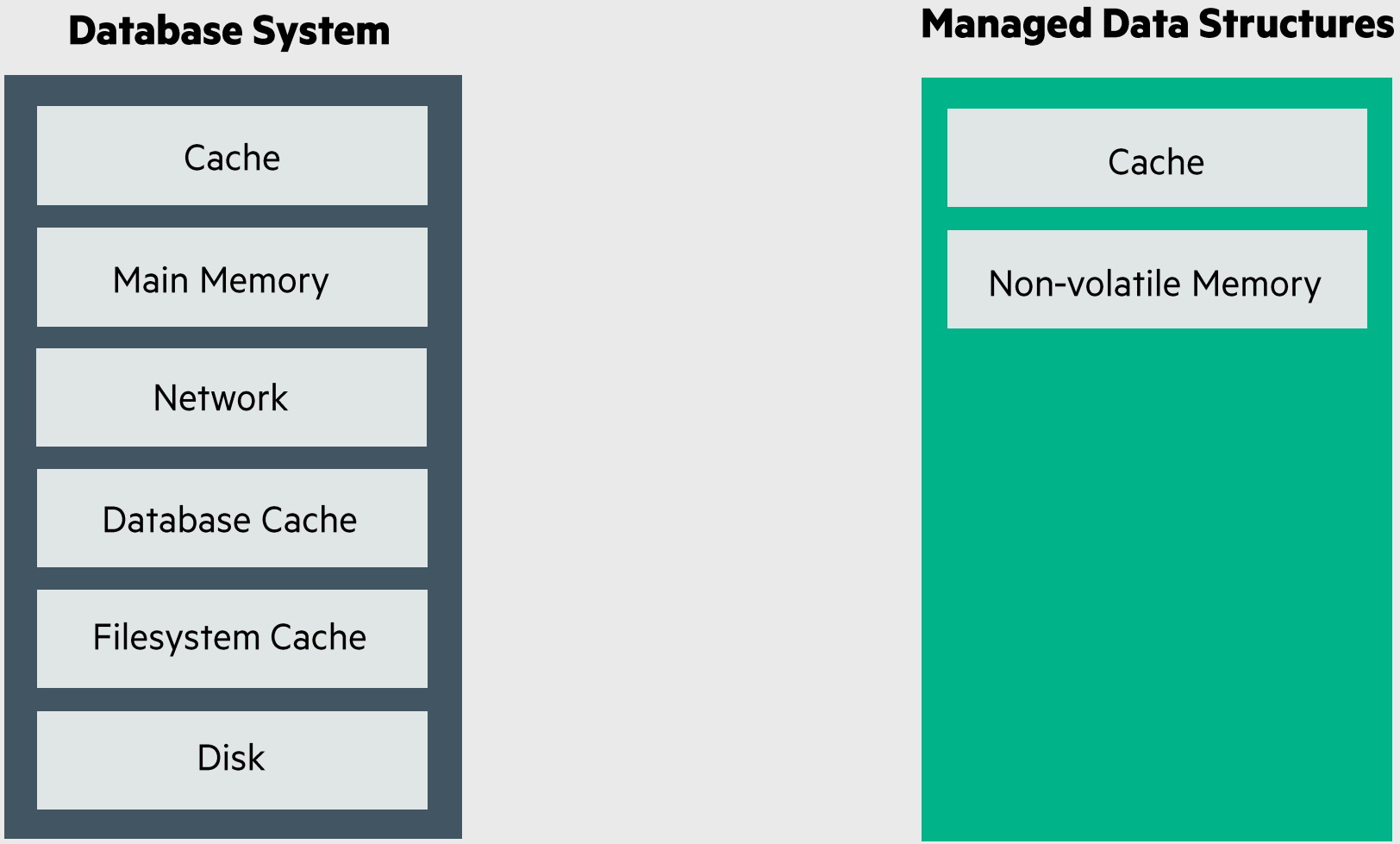
Security built-in from silicon upwards

The MDC Developer Experience



Hewlett Packard Enterprise

Developer Simplicity: Fewer data layers



Developer Simplicity: Fewer data layers

Conventional Data Formats

- Data structures
- Data format conversion
- Serialization/deserialization
- RPC, HTTP, message passing
- Disk communication latency
- Server
- Database
- File system
- Disk

MDS Data Formats

- Data structures
- Local function calls
- Non-volatile memory

Shorter path to persistence

Less code

Fewer errors

Faster development

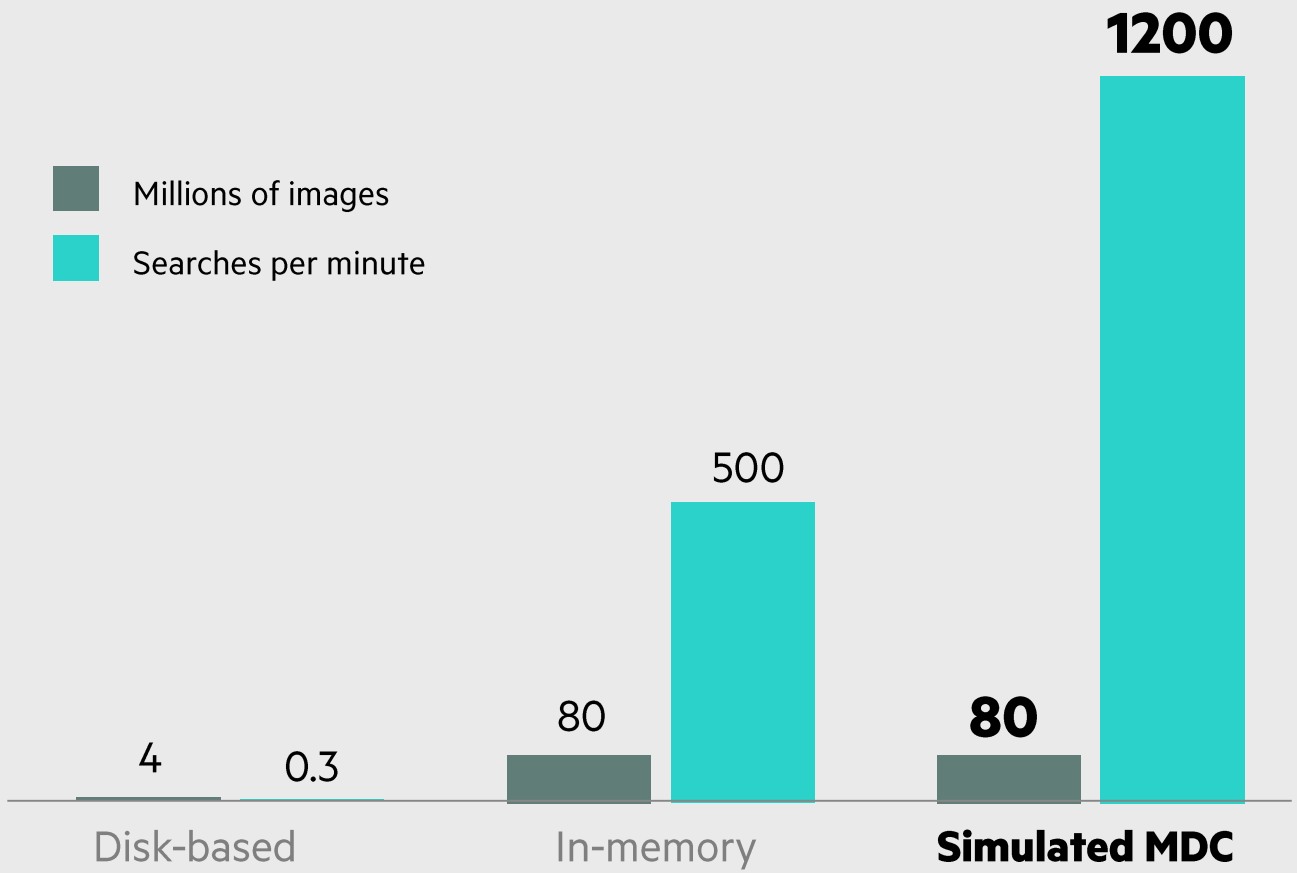
Decades of Zero Days Removed

Here is Edward Bear, coming downstairs now, bump, bump, bump, on the back of his head, behind Christopher Robin. It is, as far as he knows, the only way of coming downstairs, but sometimes he feels that there really is another way, if only he could stop bumping for a moment and think of it.

A. A. Milne, Winnie-the-Pooh

Performance demonstration – similarity search

From offline to decision time



Use cases:

Content-based image/video retrieval

Near-duplicate web page detection

Similar document retrieval

Outlier detection for e-commerce fraud mitigation

Fingerprint matching

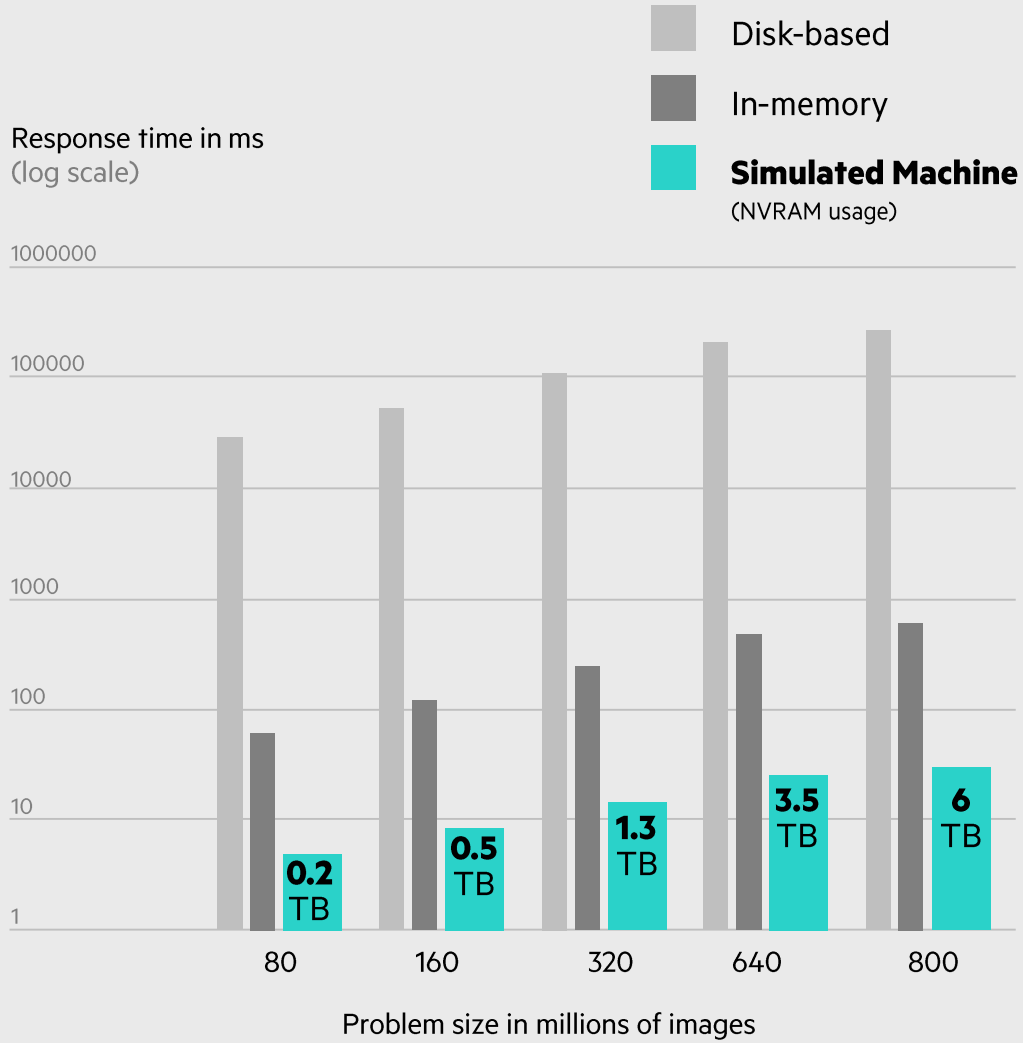
Scalable object recognition

Nearest-neighbor classification

Performance demonstration – similarity search

Buying speed with persistent memory

Parameter	Disk-based	In-memory	**Simulated Machine**
Processing	20 commodity server nodes		20 SOCs with associated NVM
Feature vector memory	16 GB/node with “unlimited” hard drive space	As required to hold image parameters	
Index memory	None		Simulated NVM with 8x DRAM latency
Threads	32 threads/server node		32 threads/SOC
Problem	4 million images	80 million images	
Typical search time (5 images)	3 minutes	2 seconds	50 milliseconds



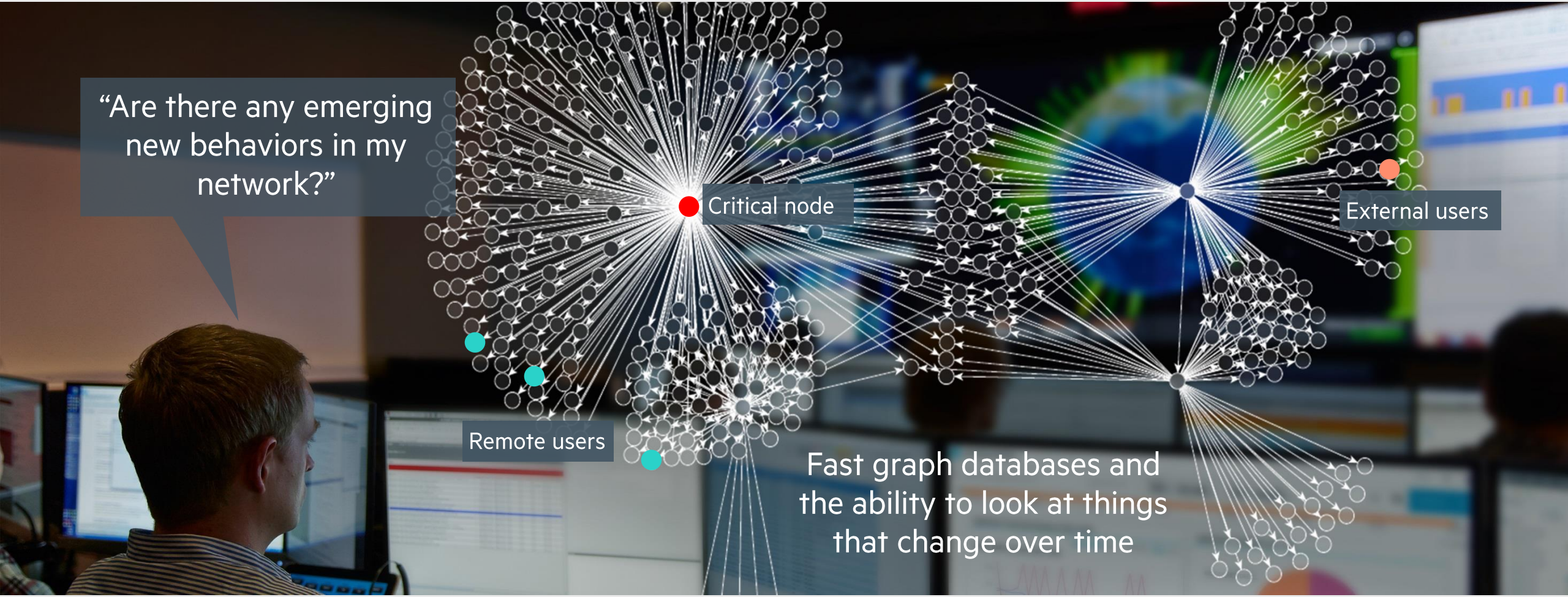
Hewlett Packard Enterprise

Machine Learning that can keep up with the now

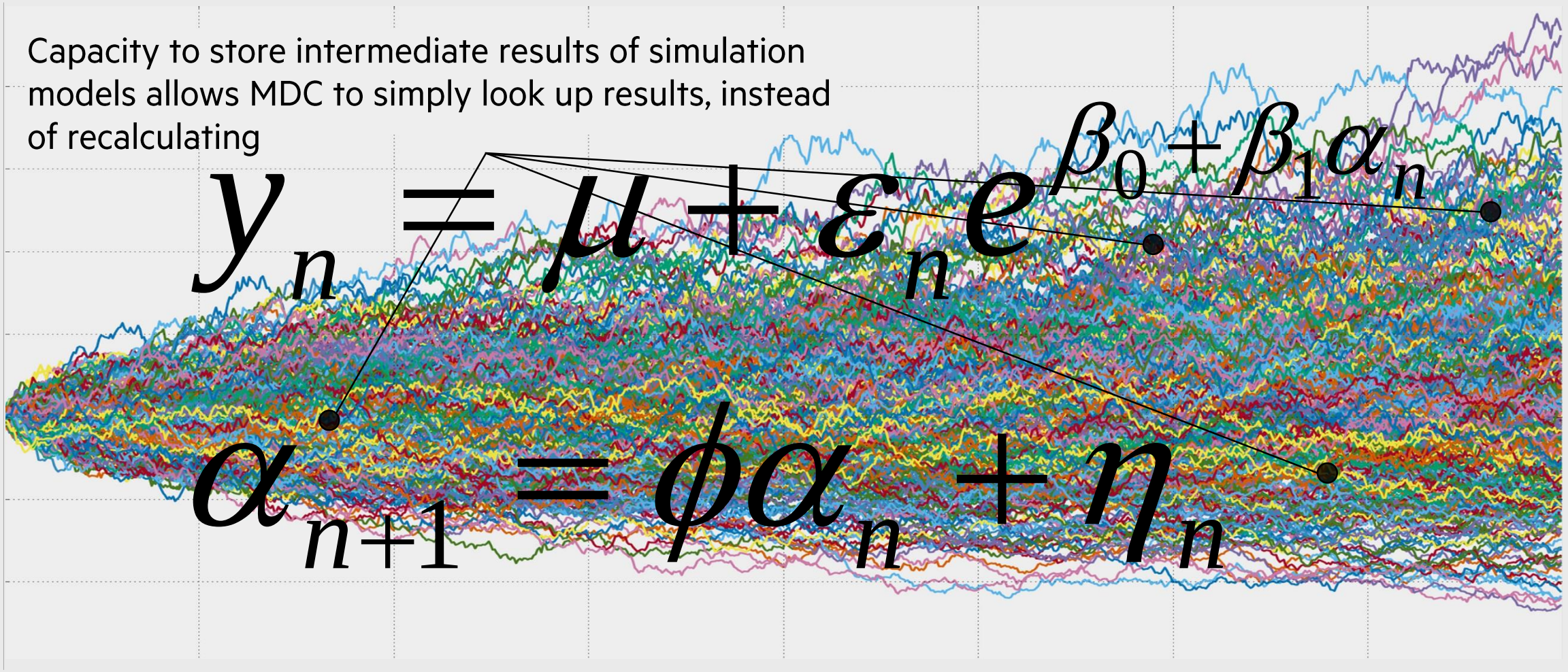
Training up-to-date deep neural networks in minutes, not weeks

Graph analytics time machine

Massive memory and fast fabrics to ingest all data



Complex models converge in minutes not days

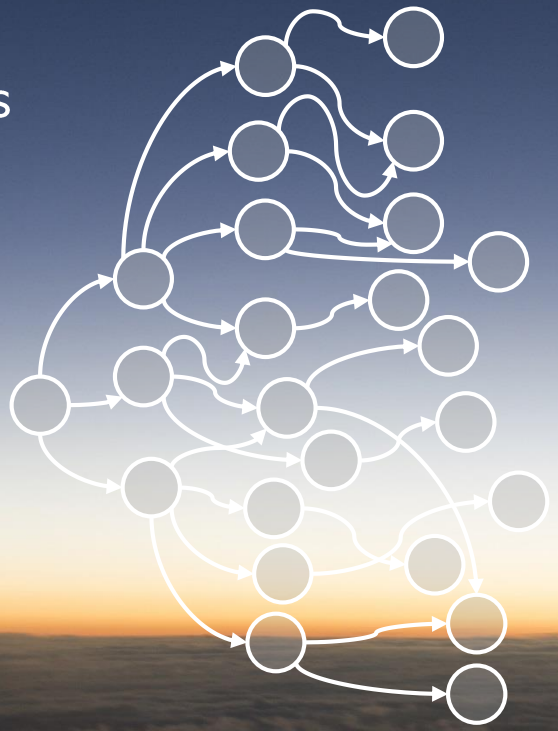


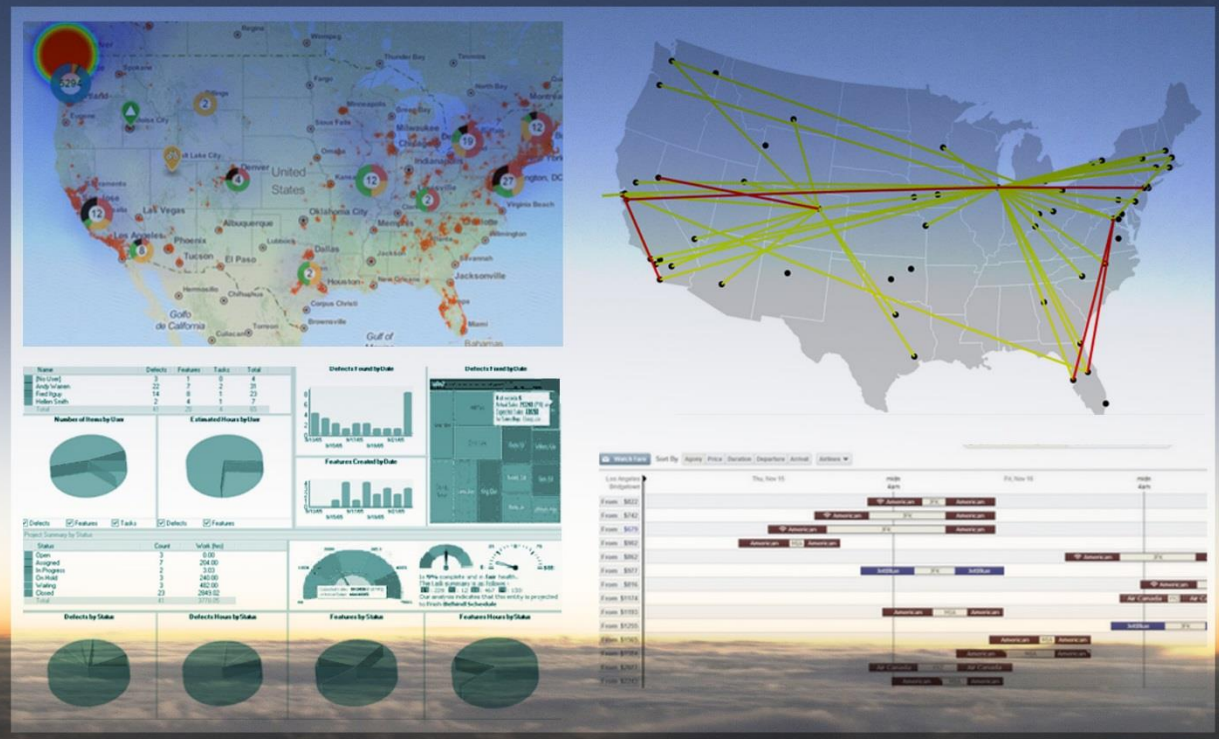
Hewlett Packard Enterprise

What if we could pre-compute an almost infinite set of “what ifs”?

Optimization over a large search space in real time becomes realistic

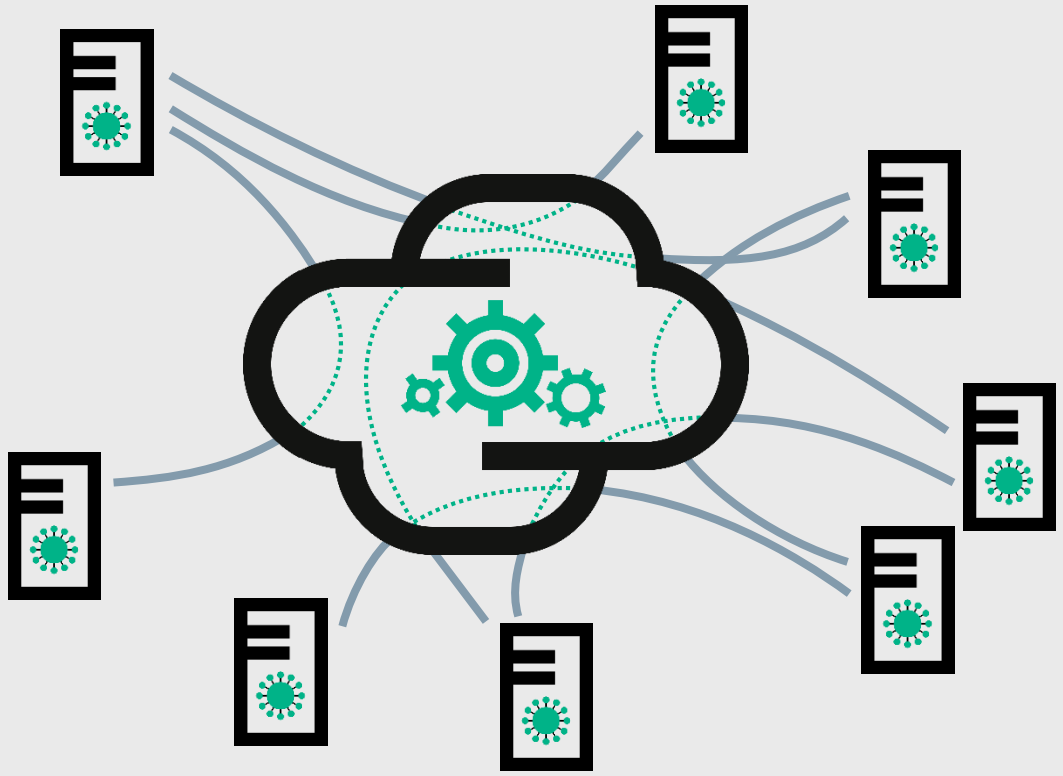
Solve complex problems
before they happen





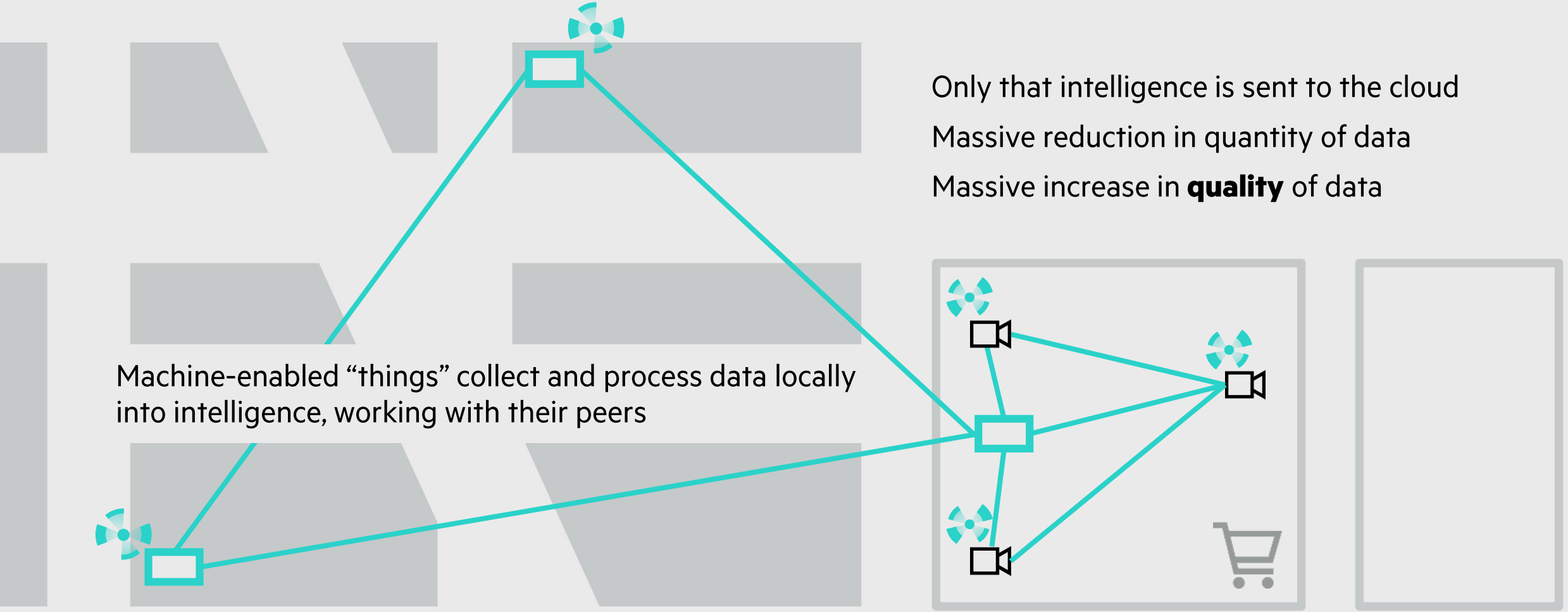
Distributed Mesh Computing

Translator
Coordinator
Orchestrator
Arbitrator
Aggregator



Replicator
Anonymizer
Border guard
Learning engine

An internet of intelligent things



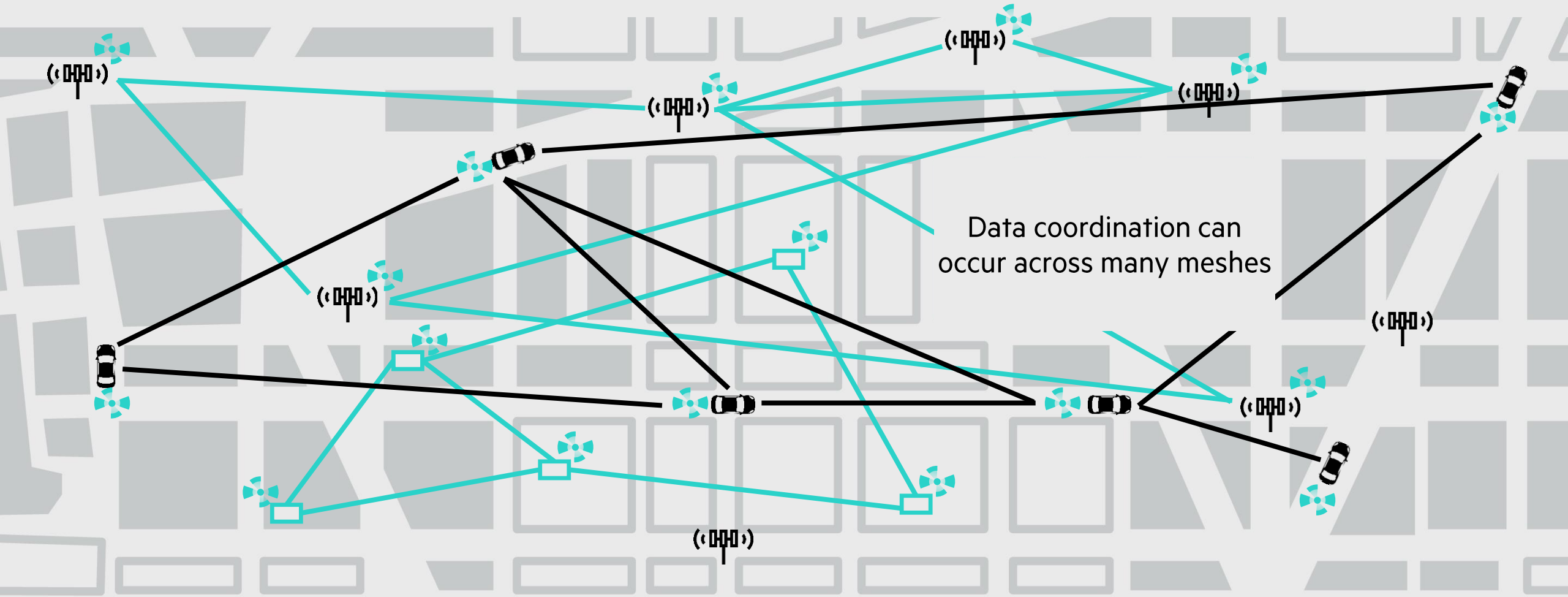
Machine-enabled “things” collect and process data locally into intelligence, working with their peers

Only that intelligence is sent to the cloud

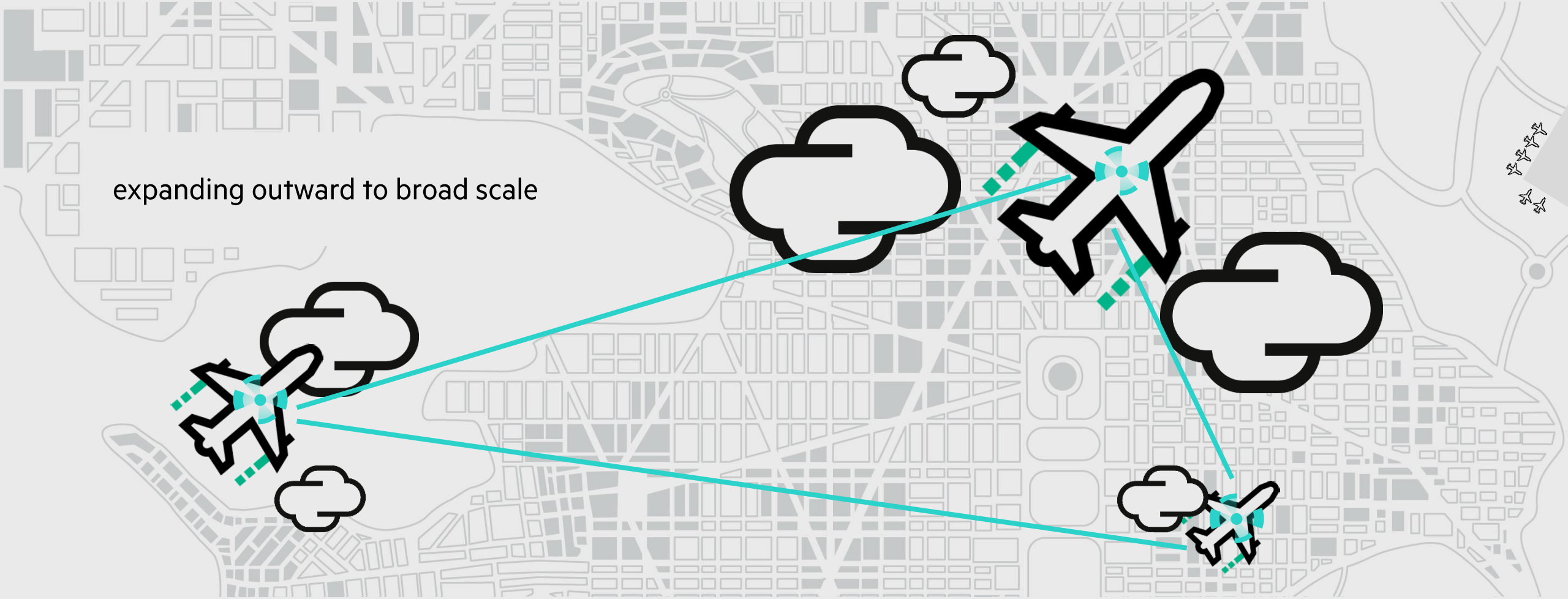
Massive reduction in quantity of data

Massive increase in **quality** of data

Broadening scale



A mesh of meshes



kirk.bresniker@hpe.com