New technologies that disrupt our complete ecosystem and their limits in the race to Zettascale

Patrick DEMICHEL

HPC & Hyperscale EMEA

Confidential disclosure reminder

- HP makes no warranties regarding the accuracy of this information. HP does not warrant or represent that it will introduce any product to which the information relates. It is presented for evaluation by the recipient and to assist HP on defining product direction

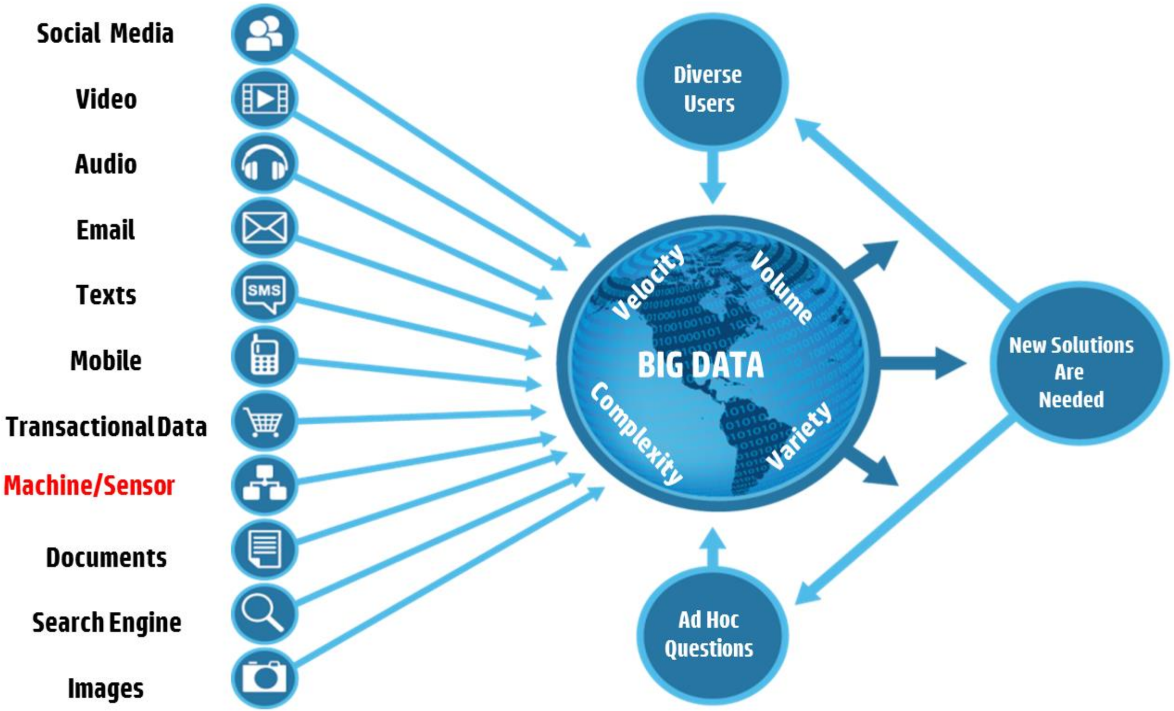
Agenda

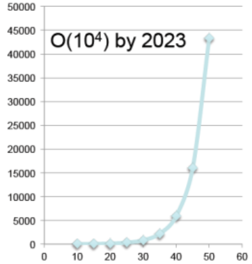
Exascale in ~2020

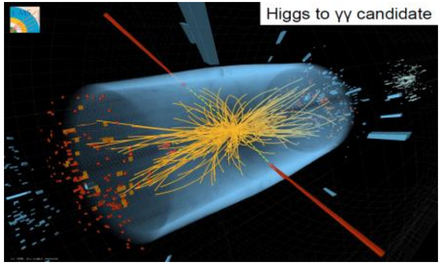
Challenges to the race to Zettascale

Tsunami of data on the horizon

202X will be the decade of Extreme Data; massive compute is required for Extreme Analytics



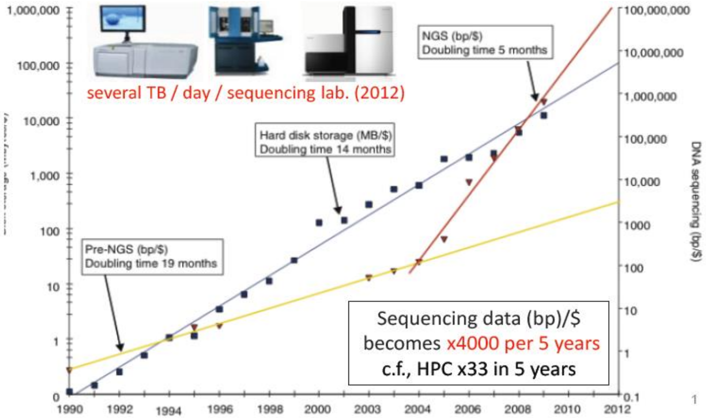




Problem: Hoped for solution:

$O(10^4) \sim O(10) \times O(10) \times O(10) \times O(10)$

Moore's law — New hardware architectures — New algorithms — Built a better detector

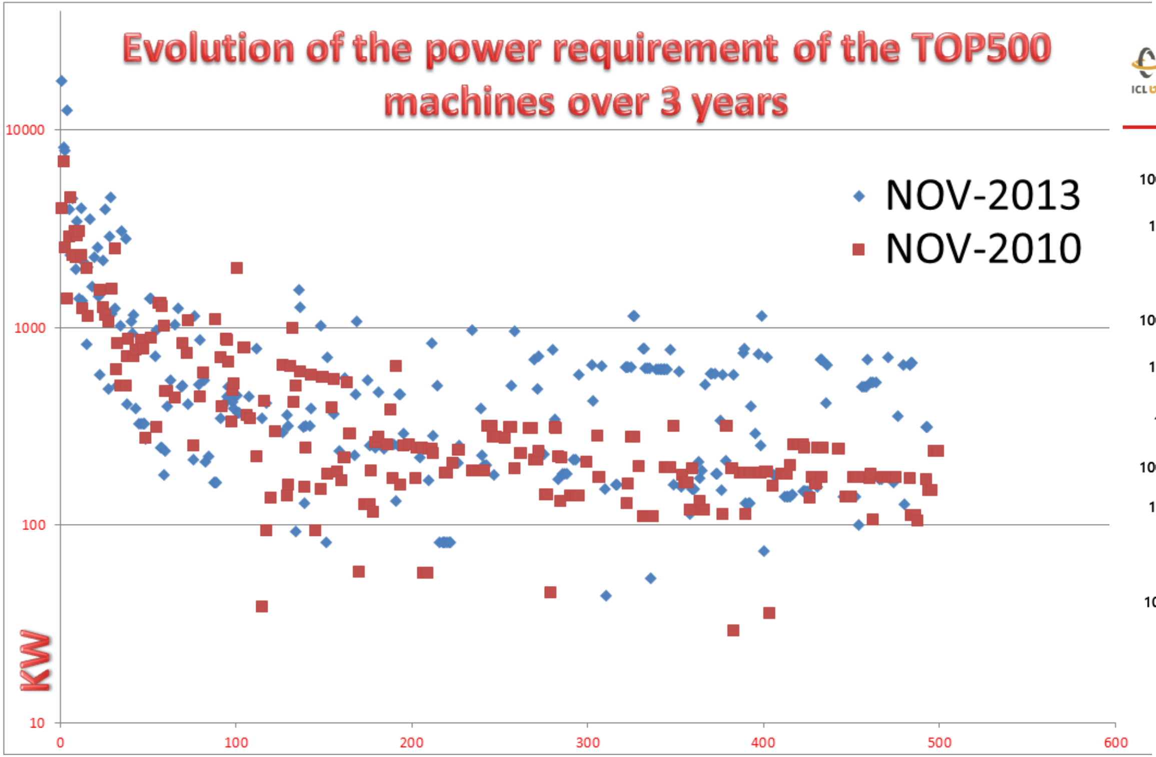


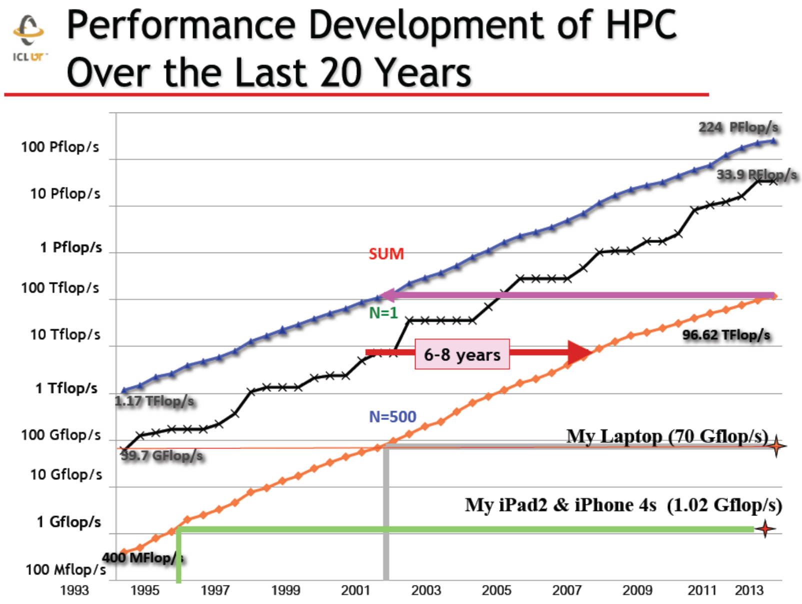
Lincoln Stein, Genome Biology, vol. 11(5), 2010

Trend in the datacenter power usage

In average every 3 years the datacenters increase their capacity by 3

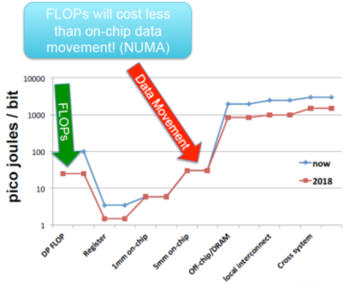
TOP500 systems moved from an average of 200 KW in 2010 to 600 KW in 2013 : an unsustainable trend





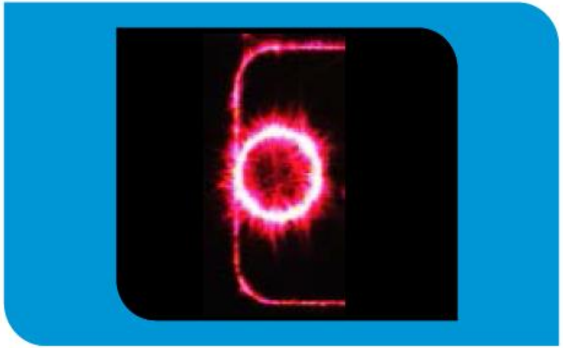
3 disruptive technologies to the rescue

But need holistic redesign for big impact

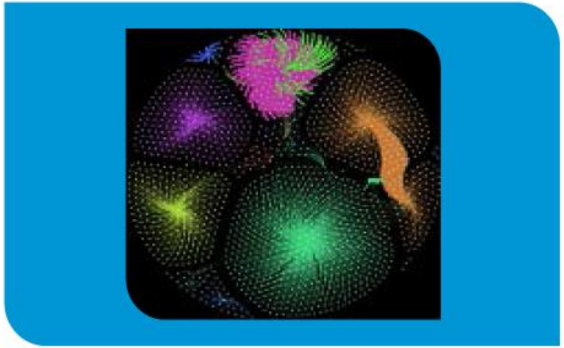


Will work for Exascale; but Zettascale?

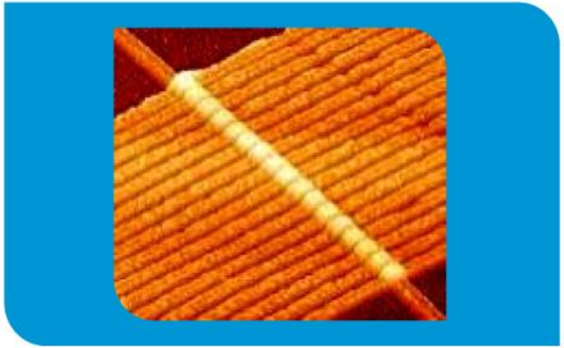
At Exascale 1pj*10^18=1MWatts ; At Zettascale 1fj*10^21=1MWatts



Breakthroughs in photonics transmit data via light, delivering quantum leaps in speed and power-efficiency



Powerful, intuitive tools to analyze, visualize and convert Big Data into actionable intelligence



Massive, universal memory enables software-defined computing from the personal to the zettascale

Emerging Memory Technologies

New memories are critical for the feasibility of extreme sciences

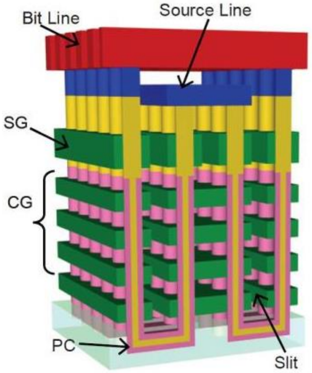
Flash Memory

Reaching the physical limits of charge storage

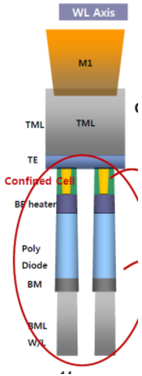
DRAM

Reaching the physical limits of charge storage

3D Flash

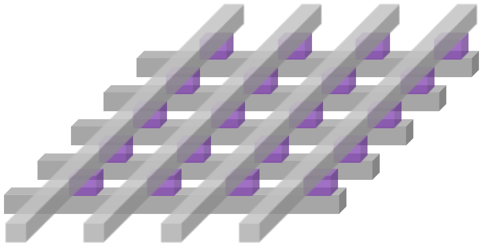


PCRAM



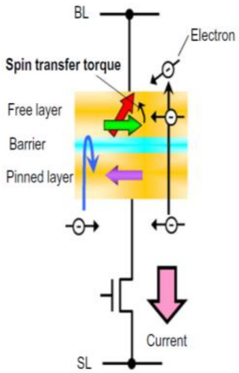
Flash Replacement

RRAM

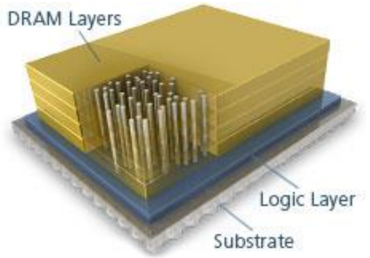


Storage Class Memory

STT-RAM



Hybrid Memory Cube

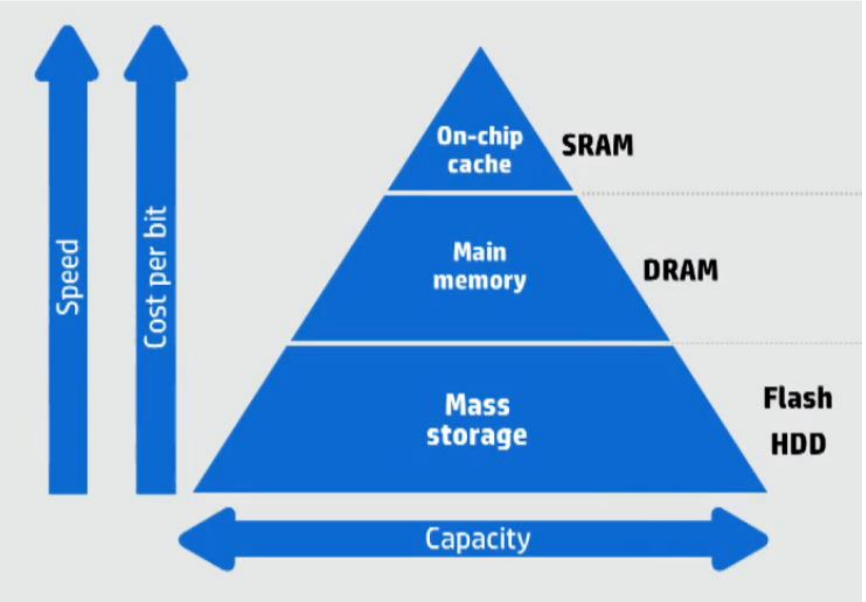


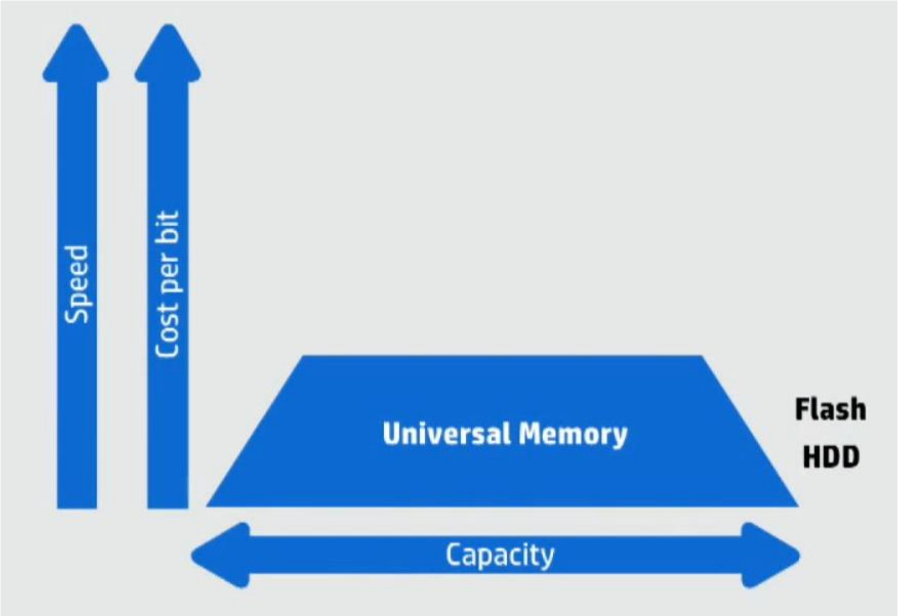
DRAM Replacement

UNIVERSAL MEMORY

A drastic reduction of the memory stack complexity and cost

But requires a complete software stack redesign to leverage the full potentiality of the new architecture





HP photonics technologies

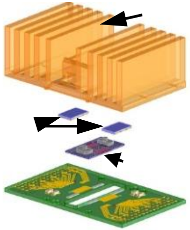
System-level architecture to large-scale integration

Devices

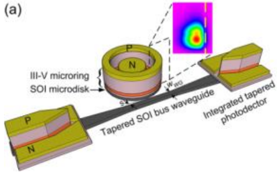
Architectures

Active cable

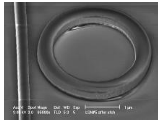
Low cost VCSEL



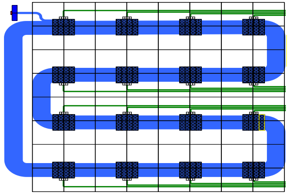
Hybrid laser



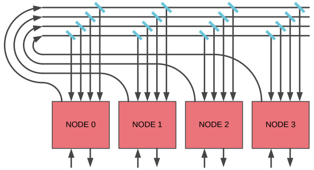
Silicon PIC



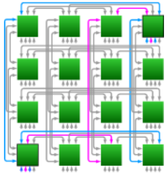
On-chip interconnect



Optical backplane



HyperX & networking



Optically connected memory

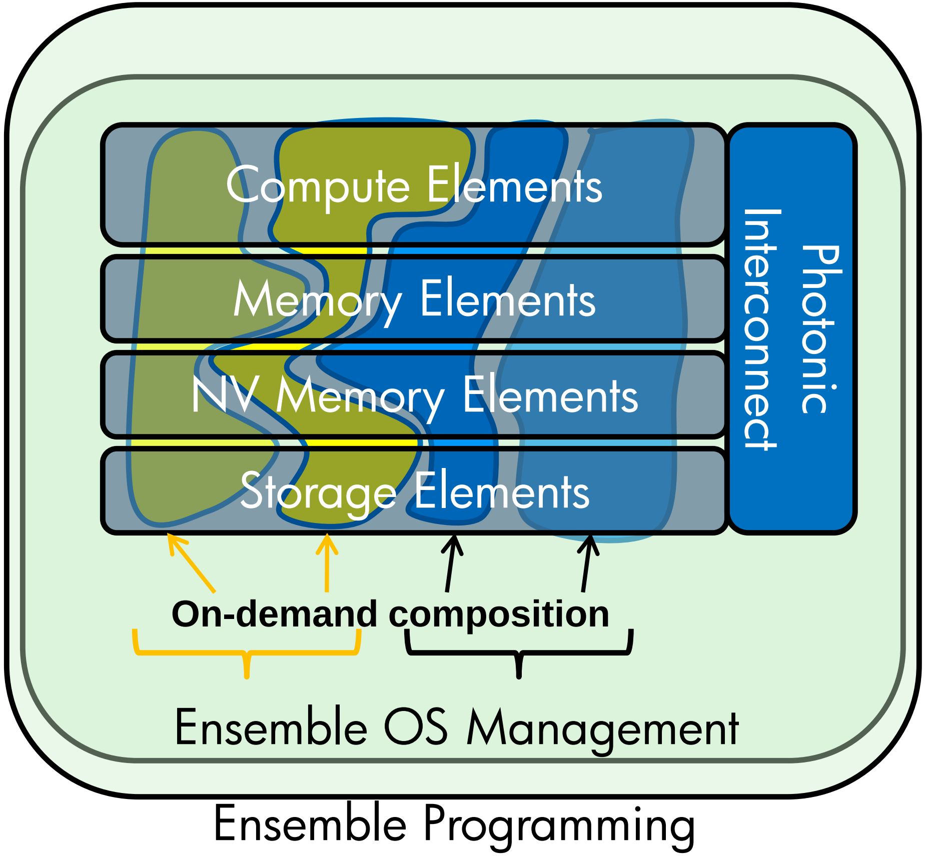
Corona

Now | 1 Year | 10 Years

Single wavelength | CWDM | DWDM

100pJ/bit | >.1 pJ/bit

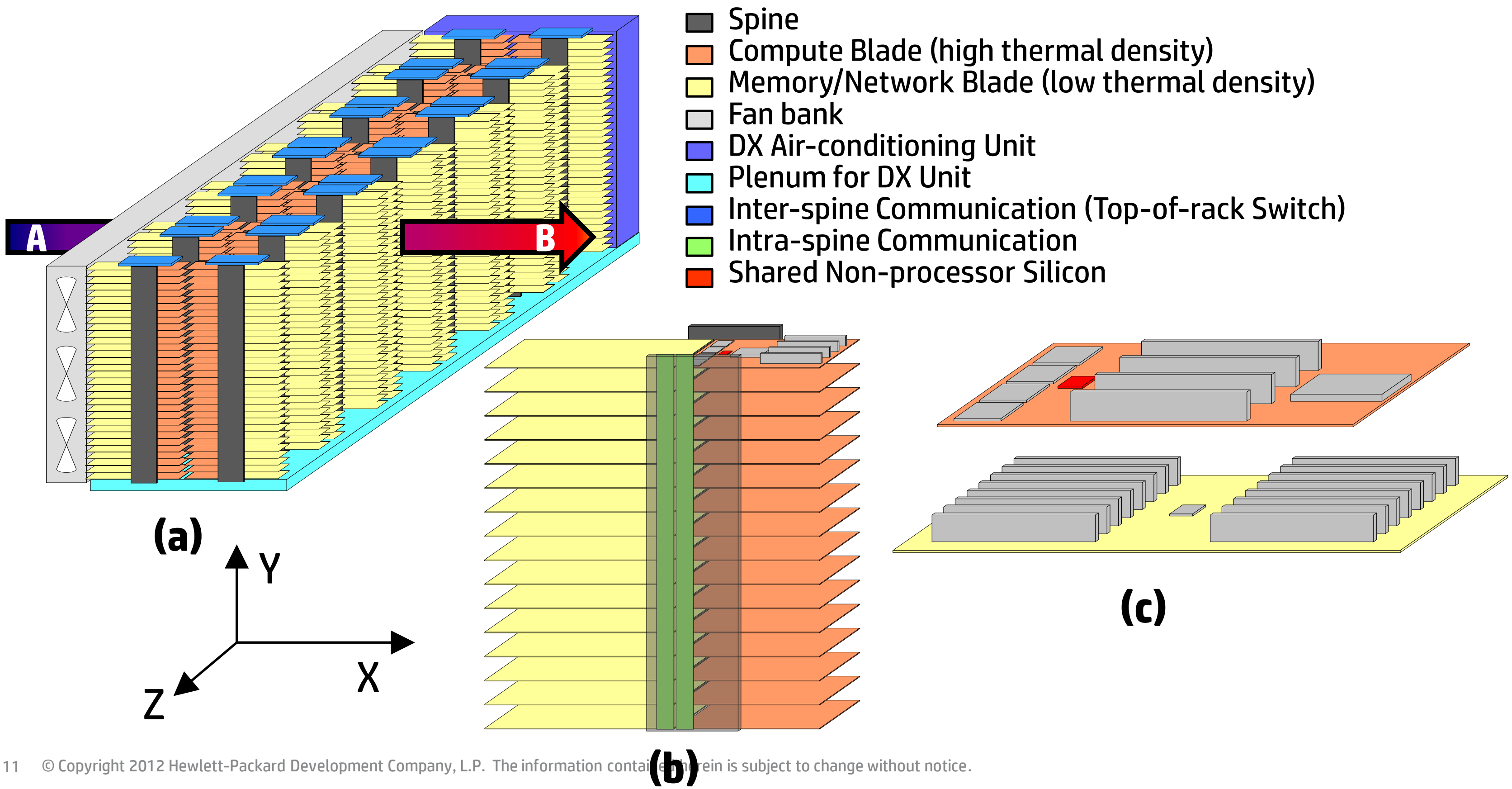
Architecture evolution/revolution



"Computing Ensemble": bigger than a server, smaller than a datacenter, built-in system software

- Disaggregated pools of uncommitted compute, memory, and storage elements
- Optical interconnects enable dynamic, on-demand composition
- Ensemble OS software using virtualization for composition and management
- Management and programming virtual appliances add value for IT and application developers

Dematerialized Data Centers

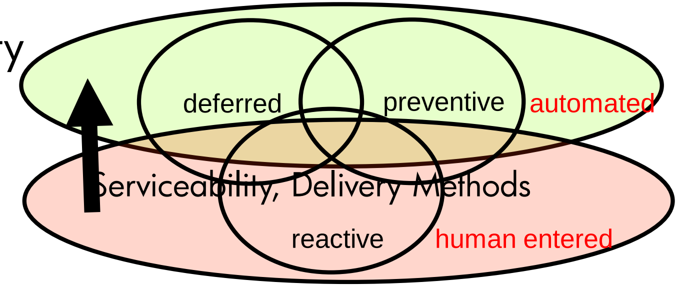


EXASCALE SYSTEM SUPPORT

Net zero

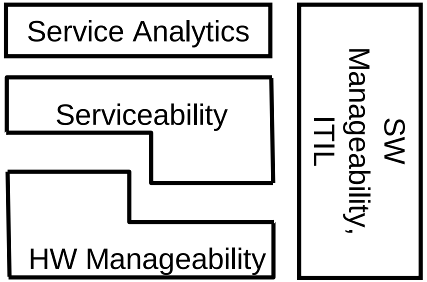
– Trends

- From hardware break-fix to higher levels (software, services)
- Significant integration between serviceability & manageability
- Level of automation is critical, move to lower cost deliveries
- Self-healing at lower levels (function of cost)
- Failures in infrastructure transparent to the service customer



– Challenges

- e2e automation, noise in data, no faults found
- Knowledge hard to search, store, share, use
- Back-end analysis (forecast, trend), global knowledge, closed loops

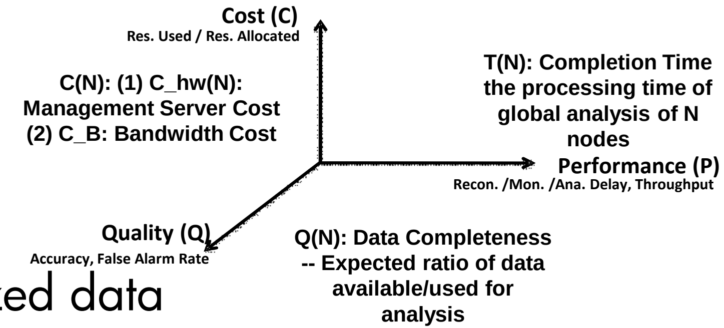


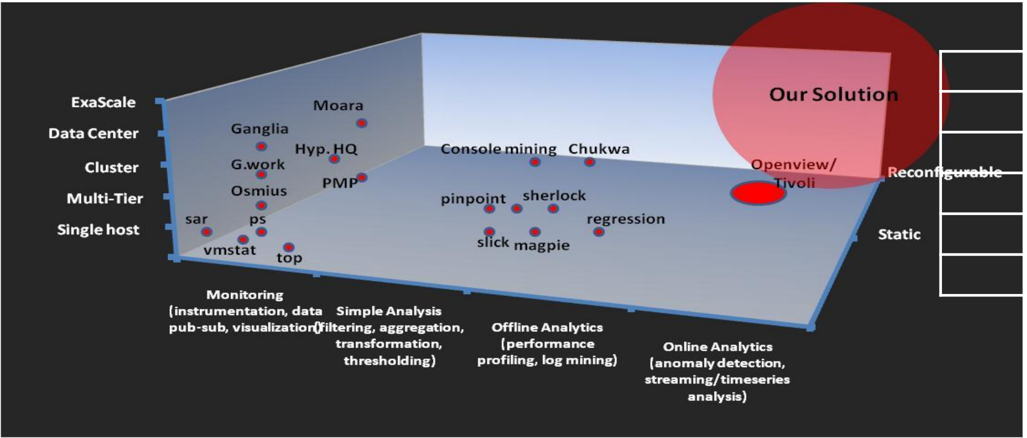
– Opportunities

- Clean data: resulting from e2e unified serviceability and self-healing
- Actionable knowledge: transparently captured, enabled by clean data
- Backend analysis: simplified by clean data and actionable knowledge

EXASCALE SYSTEM MANAGEMENT

– Monalytics – on-line management `at scale’
- Combine monitoring with analysis for scalability and fast response
- Lightweight, *dynamic*, and distributed
- Enable `local’ control loops for fast actions on analyzed data
- Adjust power states across a million nodes in a “hierarchical m-broker/channel system” in a matter of microseconds to achieve “no power struggles” by extending our existing iLO system which runs on its own management core





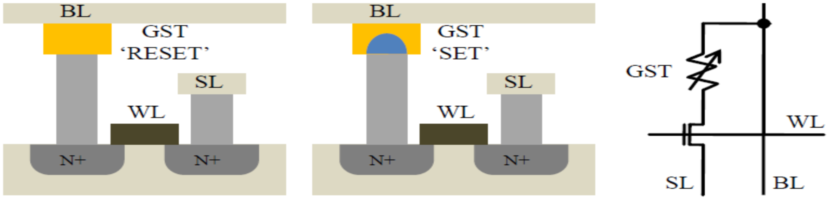
# of Nodes	Data in 1 Second	Data in 15 mins	Data in 30 mins
1	100KB	90MB	180MB
100	10MB	9GB	18GB
1,000	100MB	90GB	180GB
10,000	1GB	900GB	1.8TB
100,000	10GB	9TB	18TB

Technologies for Check-point Restart

www.nd.edu/~rich/SC09/tut157/SC2009_Jouppi_Xie_Tutorial_Final.pdf

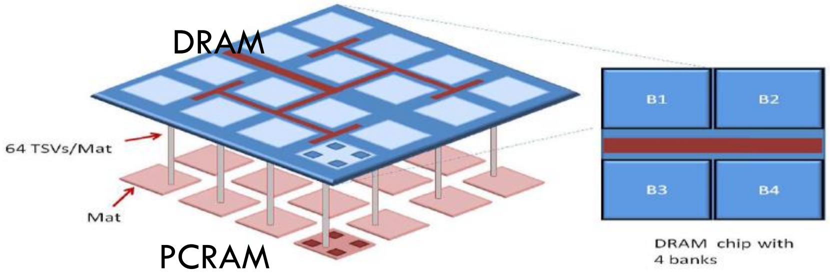
Architecture

PCRAM



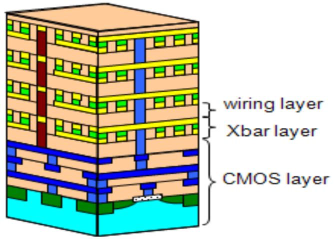
The schematic view of a PCRAM cell with NMOS access transistor (BL=Bitline, WL=Wordline, SL=Sourceline)

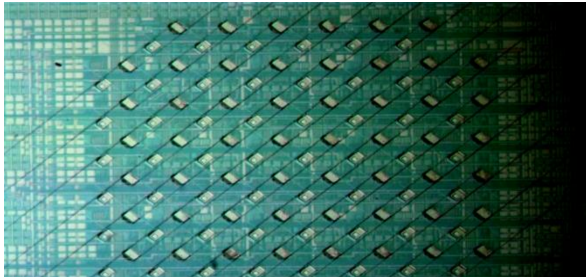
	HDD	NAND Flash	PCRAM
Taille cellule	-	4-6F^2	4-6F^2
Cycle lecture	~4ms	5us-50us	10ns-100ns
Cycle écriture	~4ms	2ms-3ms	100-1000ns
Watt à arrêt	~1W	~0W	~0W
Endurance cycles	10^15	10^5	10^8



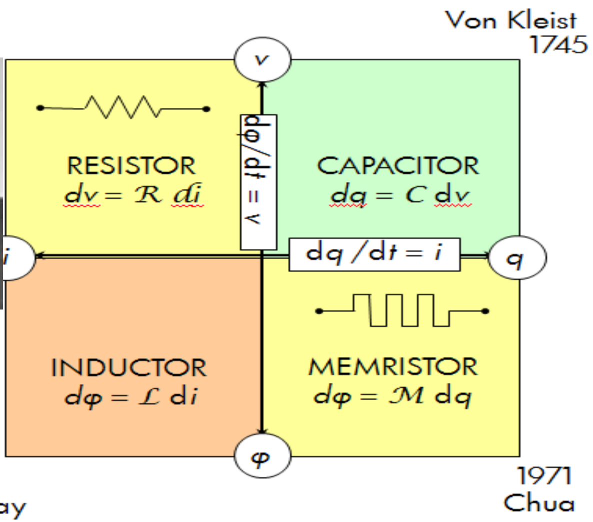
Memristor

CMOS chip avec des composants memrésistifs





L. O. Chua, (1971)



From microprocessors to nanostores for extreme efficiency

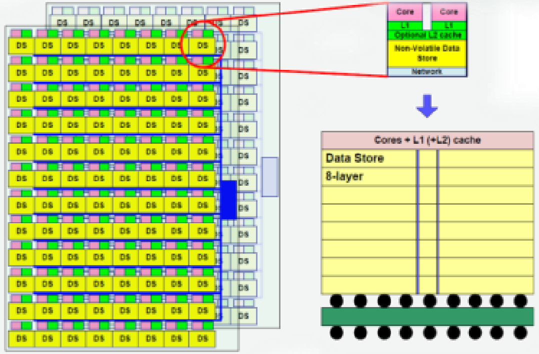
Game-changing differentiation for the data-centric data center

Enabled by HP **Memristors** technology,

HP Nanostores provide flat converged storage hierarchy with compute colocation for

10-100X better performance/watt

- **More efficient insight extraction from cold data**
- **Fast insights on hot data**

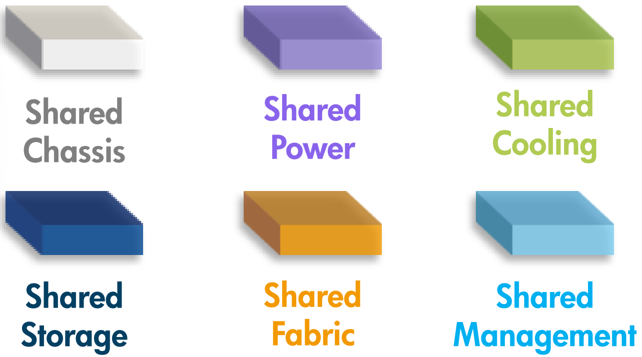


Technology

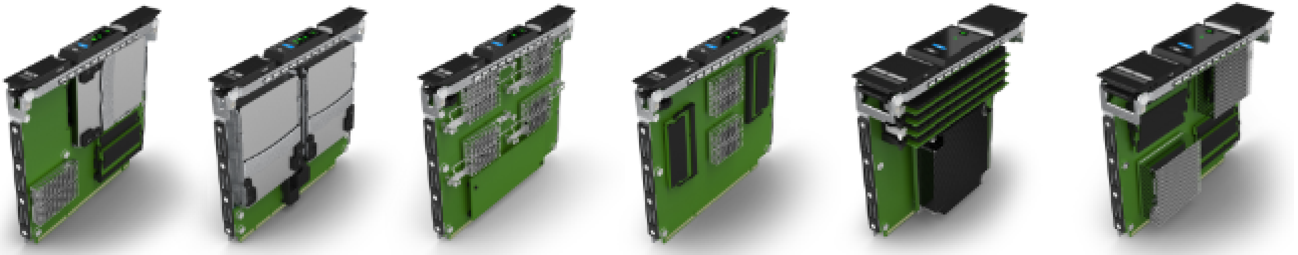
Moonshot for extreme efficiency

The new metric Gflops/Watt

Converged Infrastructure for extreme scale



… with a rich set of applications specific cartridges codesigned for extreme efficiency



At extreme scale no way to escape specialization and heterogenity

Project and roadmap

Holistic, systematic & step-wise roadmap to revolutionary impact

Architecture

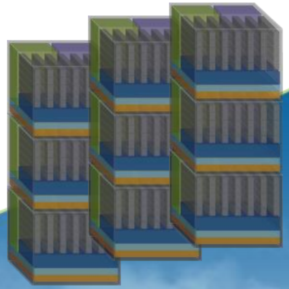
Converged infrastructure:
blades & modular datacenters

HP Labs: blades++, power & cooling, mchannels/mbrokers

Project Moonshot: Gemini, Discovery Lab, PathFinder

HP Labs: μblades, ensemble mgmt, SoC aggregation, fabric computing, new design models

Nanostores & compute hierarchies
in Data-Centric DataCenters



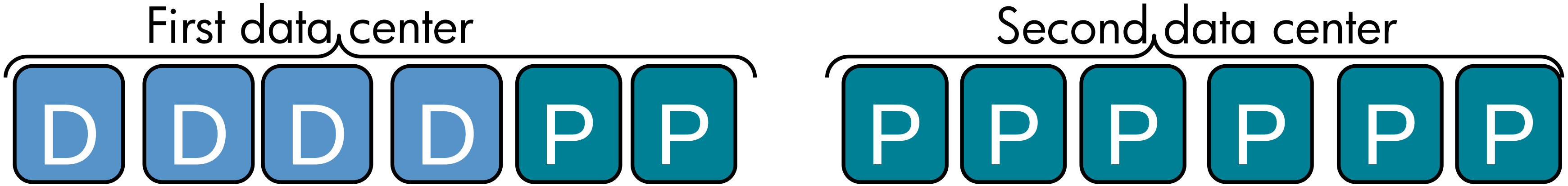
HP Labs innovations for 10-100X disruptions & new information-to-insight markets

Global Scale Storage

- Global infrastructure
- Global clients/applications
- Research challenges
  - Scalability
  - Availability
  - Low cost
  - Flexibility

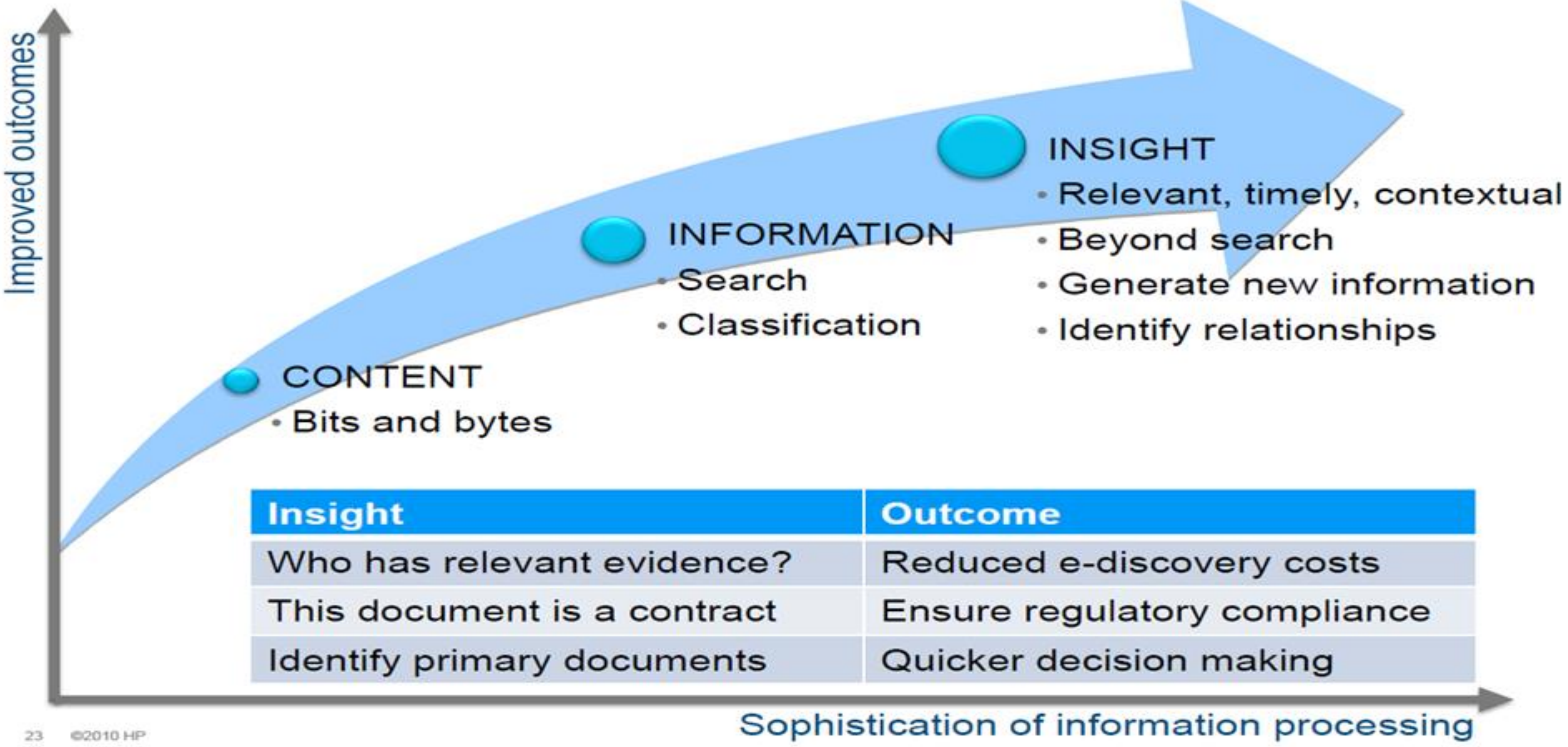
Erasure codes: Low cost availability

- Example: 4 data + 8 “parity” fragments, any 4 can recover



- Fault tolerance
  - Tolerates loss of one **entire** data center
  - Each data center independently tolerates any two disk failures
  - Eight disk failures tolerated across data centers
- Space efficient
  - Overhead of 3x replication with fault tolerance of 9x replication
  - Can tune the space efficiency-reliability trade off
- Costs
  - Computation for encode and decode
  - To recover on failed disk, 4 disks’ worth of data must be read
- Tunable tradeoff between storage efficiency and fault tolerance

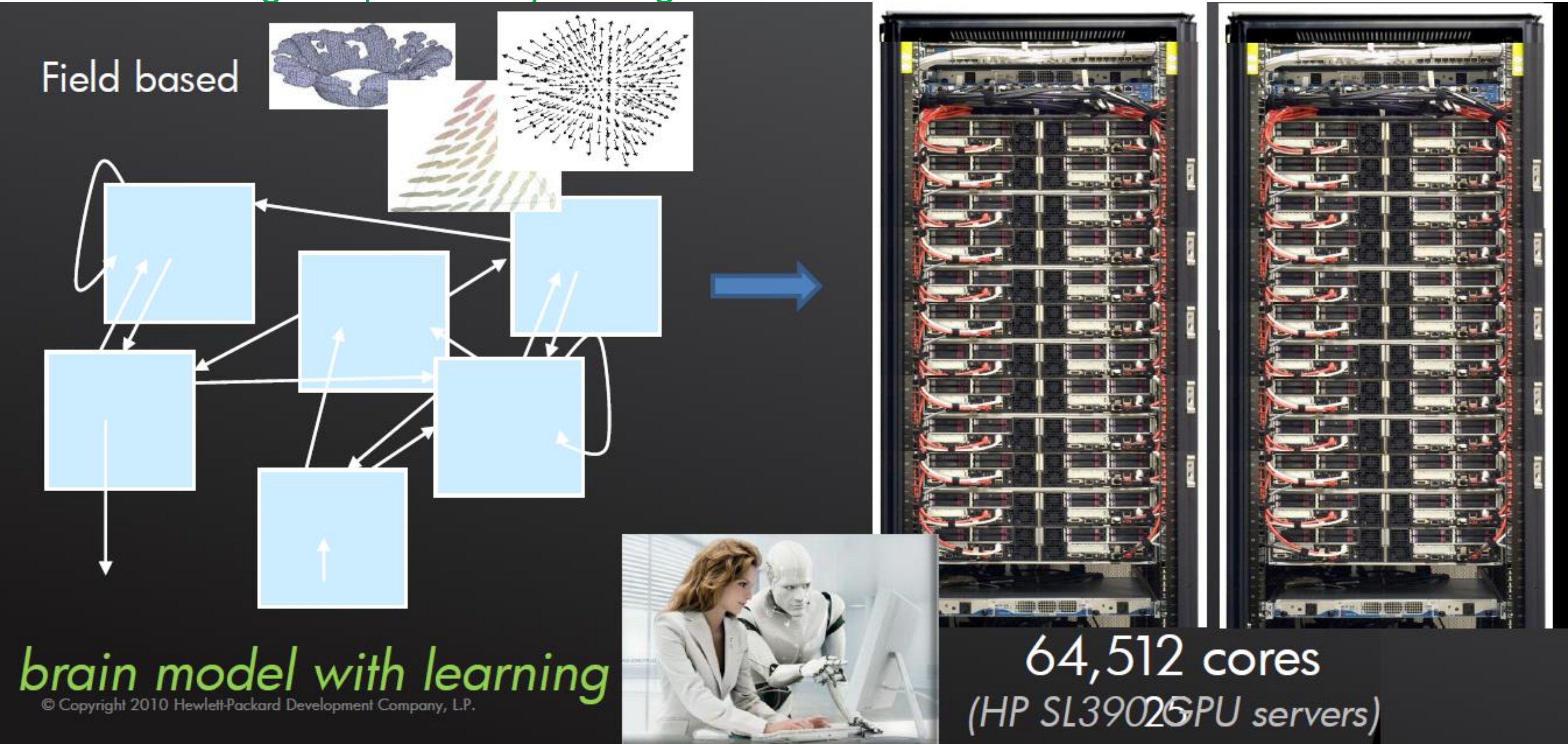
Vision : from content to insight



Insight	Outcome
Who has relevant evidence?	Reduced e-discovery costs
This document is a contract	Ensure regulatory compliance
Identify primary documents	Quicker decision making

Neuristors and cognitive systems

Self-learning adaptive analytics engine

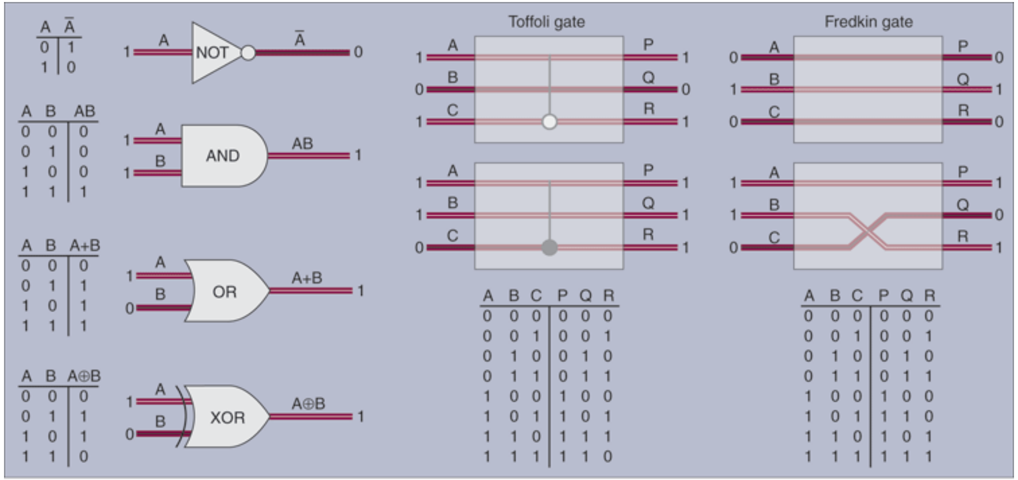


Toward Zettascale

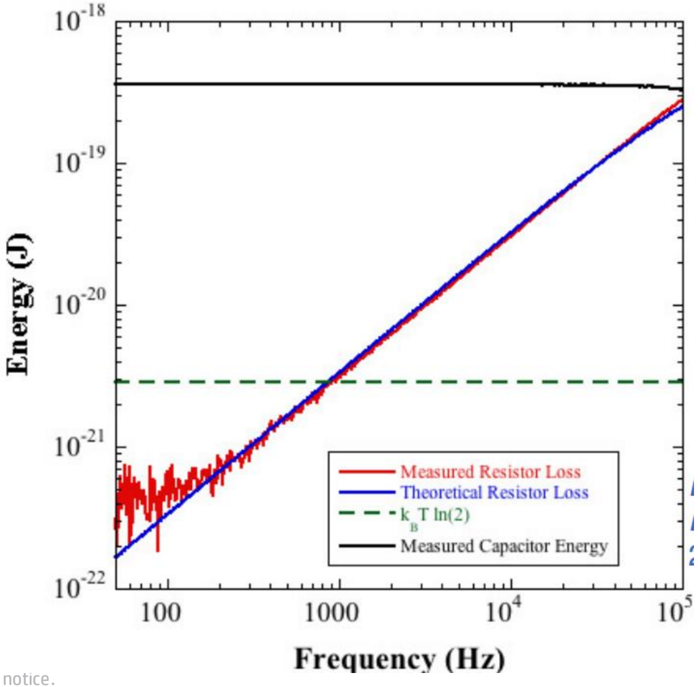
Any fundamental limit to compute?

Need to compute with reversible adiabatic or will not reach zettaflop

Landauer limit : 3 zeptojoules per erasure
Boltzmann's Distribution force us to consider
we can only do *at most* 6*10^18 erasures per Joule
We need to develop a reversible logic architecture
Maybe change the IEEE 754 format !!!



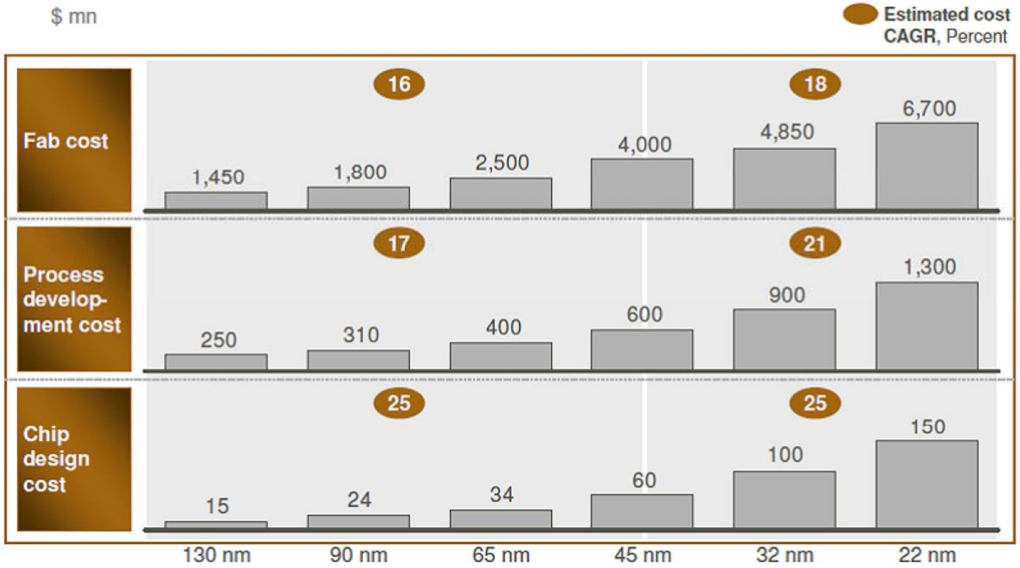
C.f., Boechler *et al.* (APL **97**:103502, 2010) measured dissipation for charging a capacitor through a resistor adiabatically



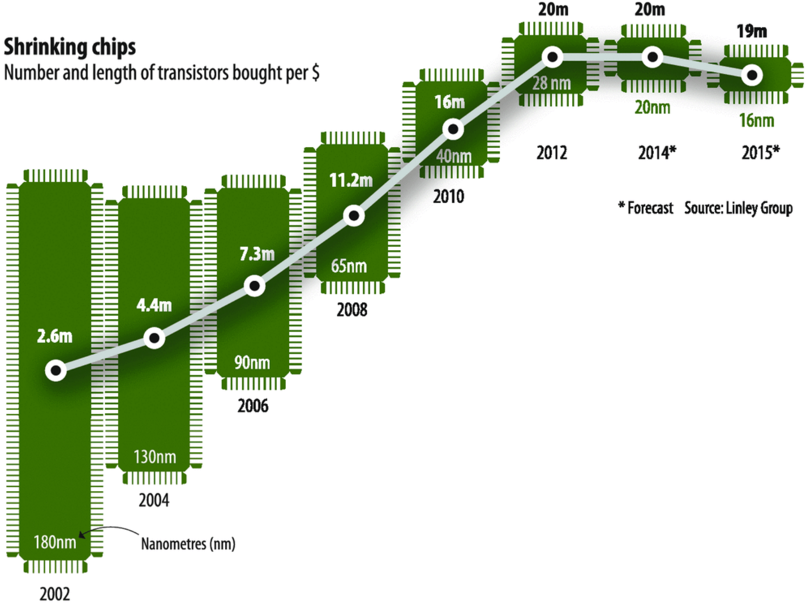
M. Frank, RevComp Cross-Disc. Intro for Beyond Moore group

Also economical limits

Is not Moore Law about economics? OOPS

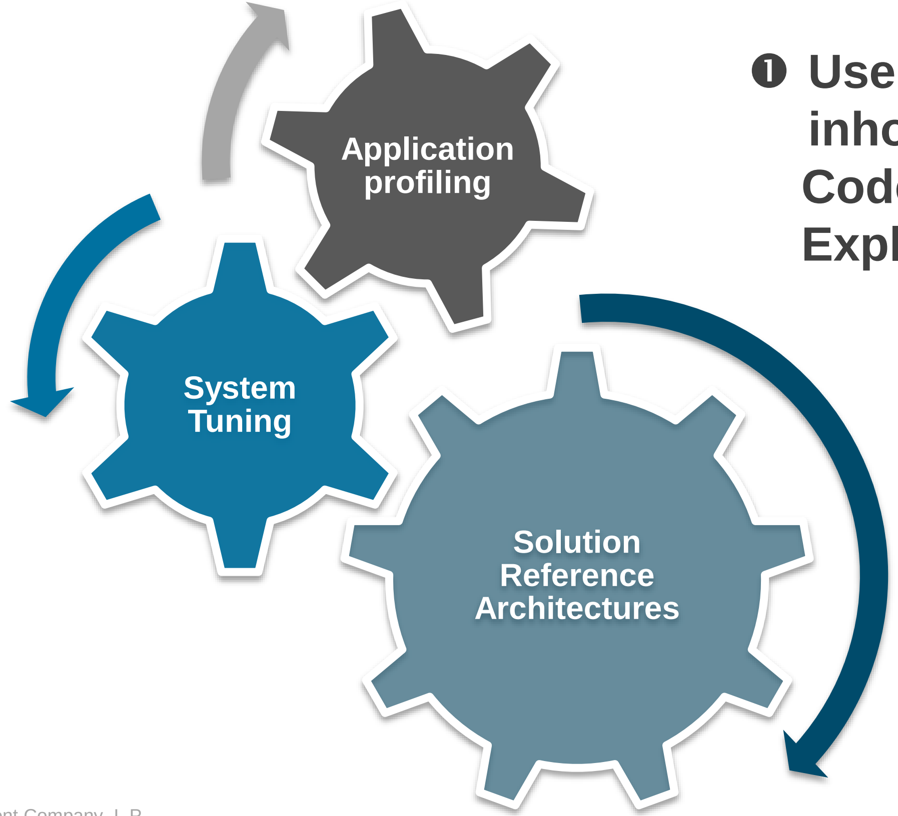


Source: press reports, iSuppli, ICKnowledge, WorldFabWatch, GSA, ITRS, internal analysis



Methodology for performance tuning

Clearly the best way to do more with less Joules ; need to invest in proportion of potentiality



❶ Use of OpenSource and inhouse tools
Codesign hard and soft
Exploit heterogeneity

❷ Determine best system setting options
- Libraries
- Memory topologies
- IOs tunings

❸ Performance optimized hardware configurations

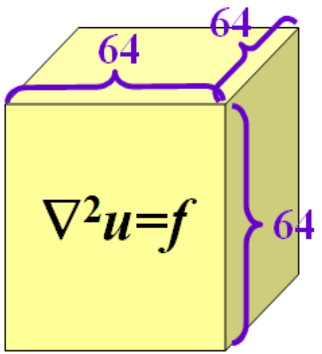
Better algorithms

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

No limit to human creativity ; could the intelligent machines beat us?

Exemple : Poisson's equation on a cube of size *N=n³*

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



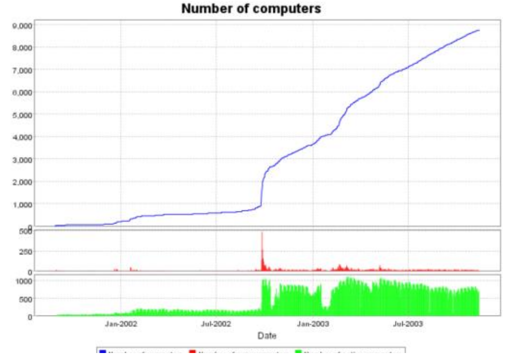
*www.siam.org/about/science/**keyes**.ppt*

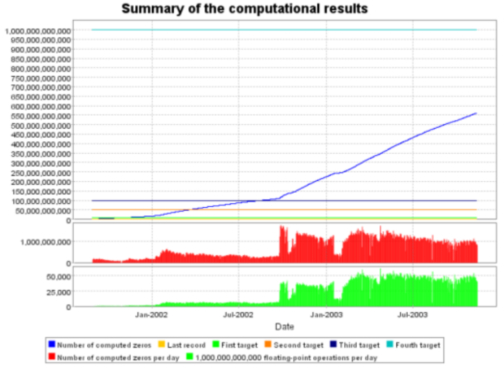
Zetagrid

- A project to compute 10^12 first zeros of Riemann Zeta function
- Highly tuned , assembly versions
- Reach 925 billions with 11900 computers after 2 years of efforts and a great sponsor
- But stop 2 months before the objective
- Why?
- Because a better algorithm gave a small team of 2 ninja programmers the capability to compute 10^13 «40 times more CPU» with 1 year of X86 CPU

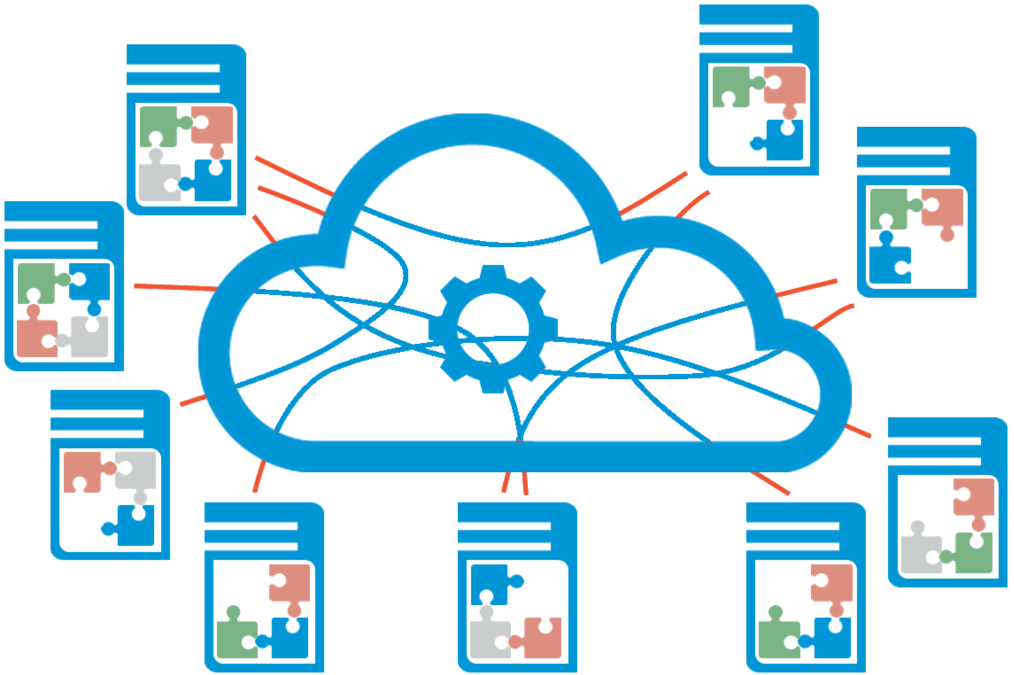
Performance characteristics

- **Participating in ZetaGrid (11/11/2003): 3,038 users and 7,899 computers**
- **1.8 x 10^{19} floating-point operations for calculating about 561 billion zeros of the Riemann zeta function in 805 days**
  - **~261 GFLOPS**
  - **~29 days maximal performance of IBM ASCI White, 8192 Power3 375 MHz processors (place 2, 06/2002, www.top500.org)**
  - **~2304 years maximal performance of one Intel Pentium 4 with 2 GHz processors, 250 MFLOPS**



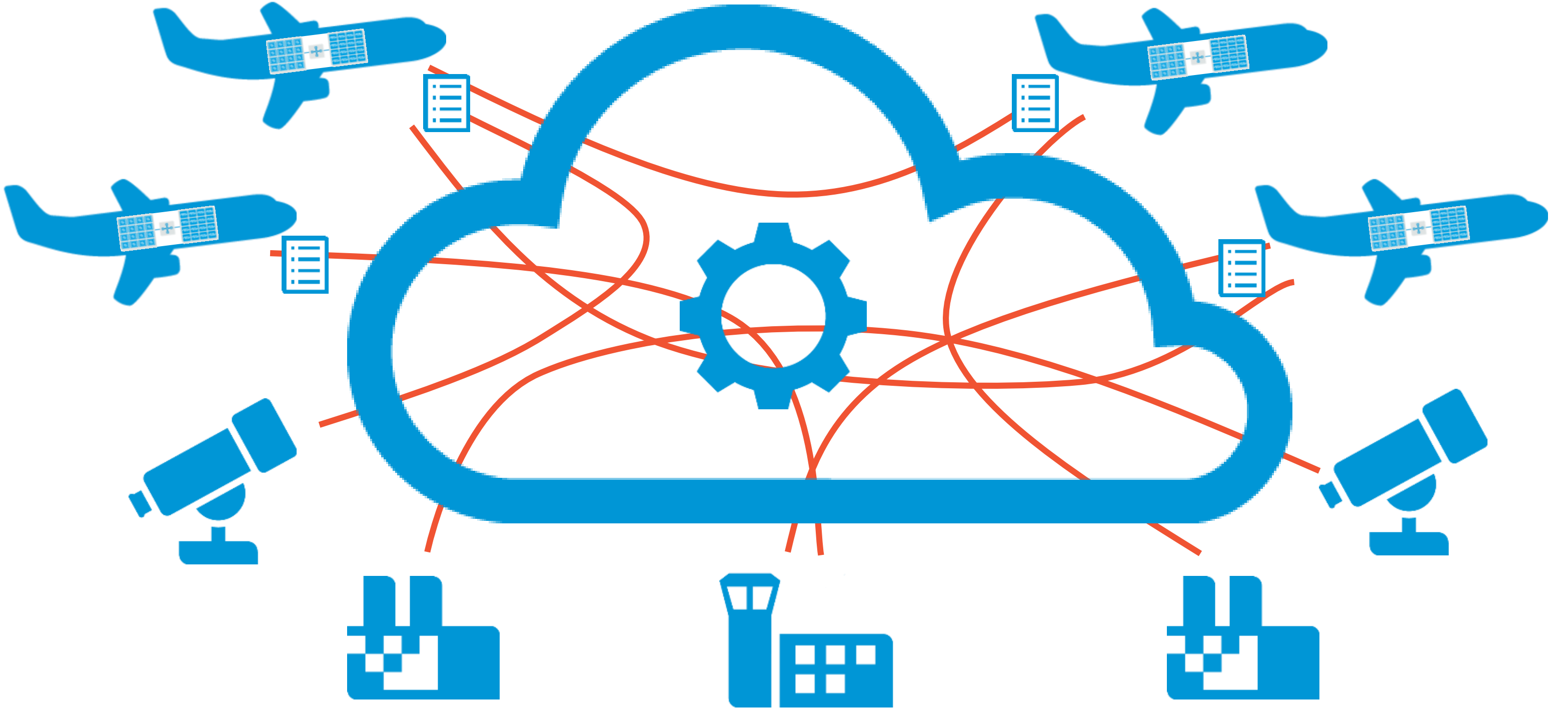


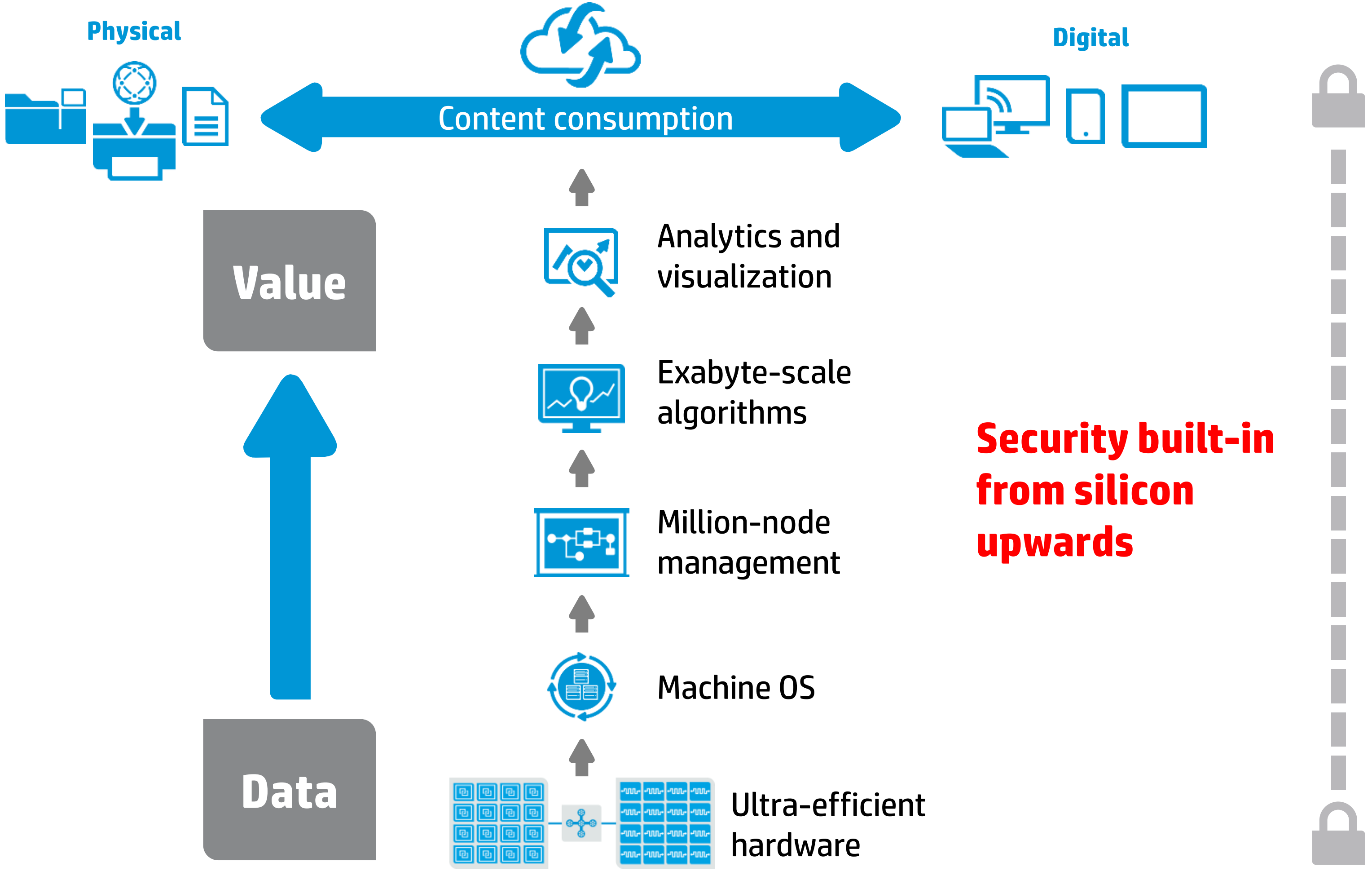
Distributed Mesh Compute



Translator
Coordinator
Orchestrator
Arbitrator
Aggregator
Replicator
Anonymizer
Border guard
Learning engine

A mesh of connected aircrafts …




Physical
Content consumption
Digital
Value
Data
Analytics and visualization
Exabyte-scale algorithms
Million-node management
Machine OS
Ultra-efficient hardware
Security built-in from silicon upwards

KEY MESSAGES

- Exascale will be hard but we have a solid plan « THE MACHINE »
- Zettascale will require even more drastic changes and many miracles
- Those *scale machines will be the brains of our highly engineered planet; they will manage thousands of tier-2 systems and trillions of intelligent objects
- There is an unlimited potientiality to solve many of the problems of our planet and its passengers, assuming we can deliver the promise of extreme scale analytics at low cost and low energy
- But we need to holistically redevelop all components from the CPU, memories, file system, OS, codes, tools, trainings, focus, etc … , with the obsession of extreme efficiency
- There is an imperative opportunity to rethink the security
- Still plenty room to do better usage of our Joules
- Disruption is everywhere; you like it or not; the physics impose it
- At least we have a solid business case « extreme scale IOT analytics »
- Big Sciences will be Sciences of Extreme Data
- Heterogeneity is imperative
- Dont expect THE magic langage ; we need plenty ninja programmers

Questions