European pole of competence in high performance simulation

Exascale challenges

June 27,28 2012 Ecole Polytechnique Palaiseau France
patrick.demichel@hp.com

HP Labs around the world

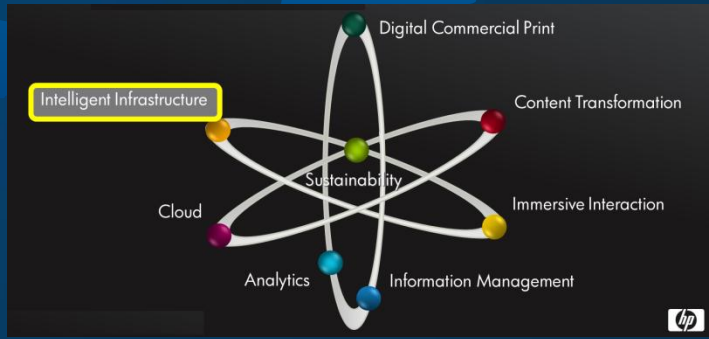
Beijing
Tokyo
Palo Alto
Bristol
St. Petersburg
Bangalore

7 locations

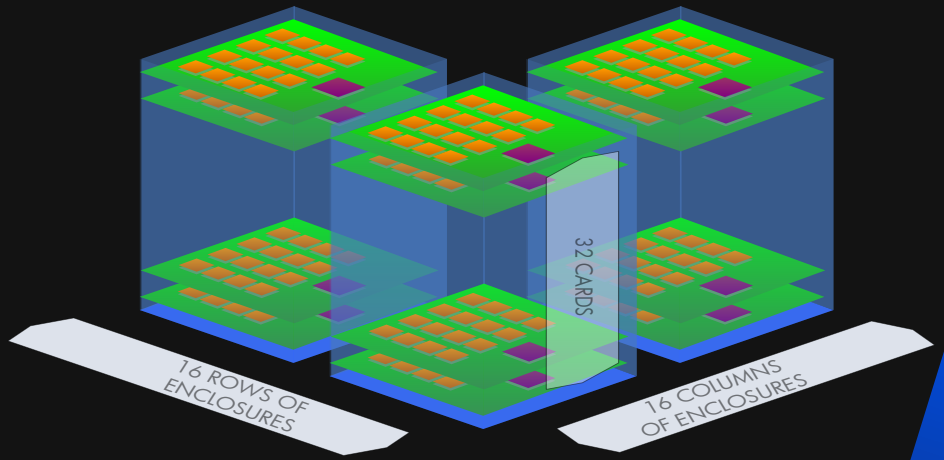
600 researchers in 23 newly formed labs

5 research themes with 20-30 projects at a time

Exascale Computing laboratory directed by Norm Jouppi



Vision



- Vision
- Photonics
- System
- Q&A

INTELLIGENT INFRASTRUCTURE

END STATE: Capture more value via dramatic computing performance and cost improvements

HP LABS' RESEARCH CONTRIBUTION: Radical, new approaches for collecting, storing and transmitting data to feed the exascale data center

BIG BETS:

NEXT-GENERATION DATA CENTERS
Exascale, photonic interconnects

NETWORKING
Open, flexible, programmable wired and wireless platform

NEXT-GENERATION SCALABLE STORAGE
Cloud-scale, dynamic, secure

NON-VOLATILE MEMORY AND STORAGE
Memristor

CeNSE
Nano-scale sensors creating a Central Nervous System for the Earth

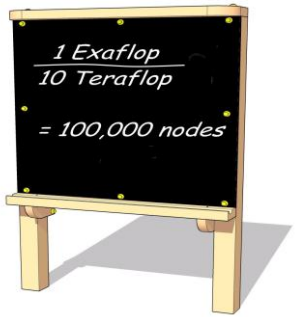
Vision for Exascale

Improve Performance/TCO by 10X

- Efficiency:
  - *Interconnects* using photons
    - 5x (short term: 5years) optical links between nodes
    - 10x (long term) with nanophotonics (+10x bandwidth)
  - *Nodes with 256 cores : 10TFlops/200Watts*
  - *Memory hierarchy extended with memristors*
- Manage: 1 operator for100K nodes
- Autodetec and autorepair failures:
  - *Check-point Restart integrated and transparent*

4 research axes as priorities:

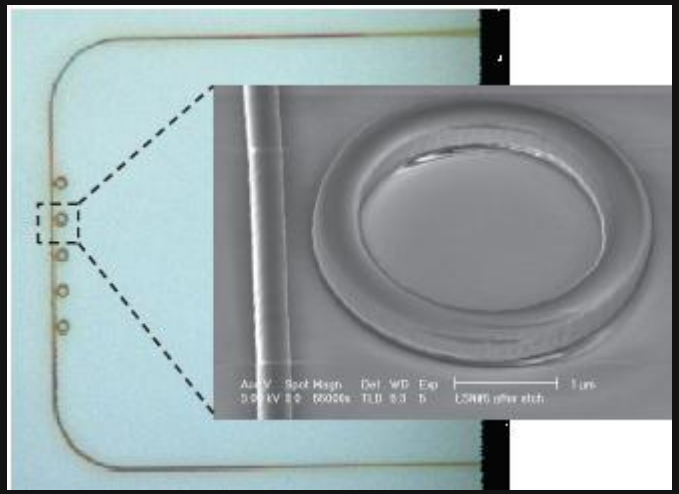
- Optical interconnects: Scalability up to 1M nodes
- Basic blocks for compute: Corona project
- System software: 1 operator for100K nodes
- Programmability: Reliability, efficiency



U.S. DEPARTMENT OF ENERGY

System attributes	2010	"2015"		"2018"	
System peak	2 Pflop/s	200 Pflop/s		1 Eflop/sec	
Power	6 MW	15 MW		~20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
Total Concurrency	225,000	$O(10^8)$		$O(10^9)$	
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

Photonics



- Vision
- Photonics
- System
- Q&A

Optical Interconnect at All Scales

The Optically Interconnected Datacenter

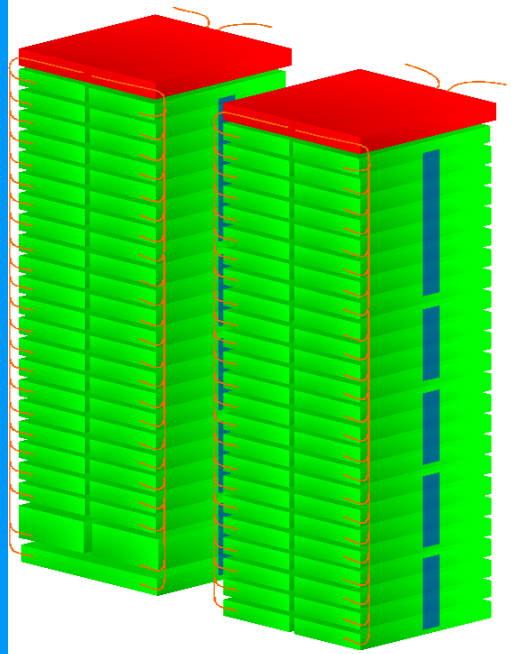
Networking

Research Challenges

- High radix switches
- Optical fabrics
- High radix routers
- Connectors, engines, media

Opportunities

- Uniform bandwidth and latency → high flexibility, new programming models
- Lower TCO through power saving, ease of installation, flexibility



Memory/CPU Interconnect

Research Challenges

- Low cost optical bus structures
- New memory architectures
- Large Scale Integrated Nanophotonics

Opportunities

- New architectures
- Flexibility, re-configurability

Metrics

Discrete optics

- 5x higher BW density
- 5x lower power

Integrated photonics

- 20x higher BW/pin
- 5x further power reduction

HP photonics technologies

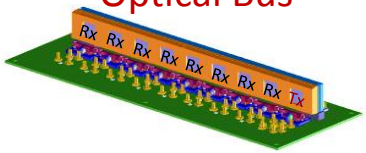
System-level architecture to large-scale integration

Devices

Architectures

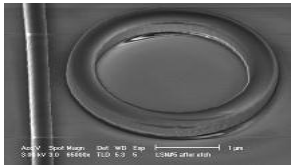
Active cable

Optical Bus

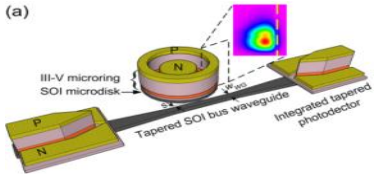


Hybrid laser cable

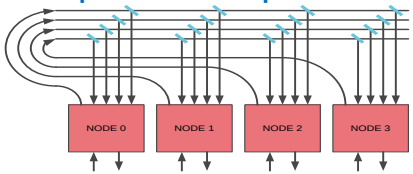
Silicon PIC



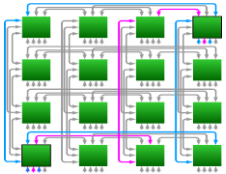
On-chip interconnect



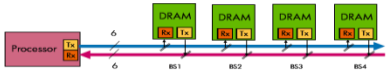
Optical backplane

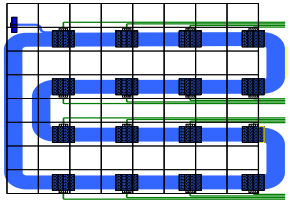


HyperX & ensemble



Optically connected memory





Now | 1 Year | 3 Years | 5 Years | 7 Years | 10 Years

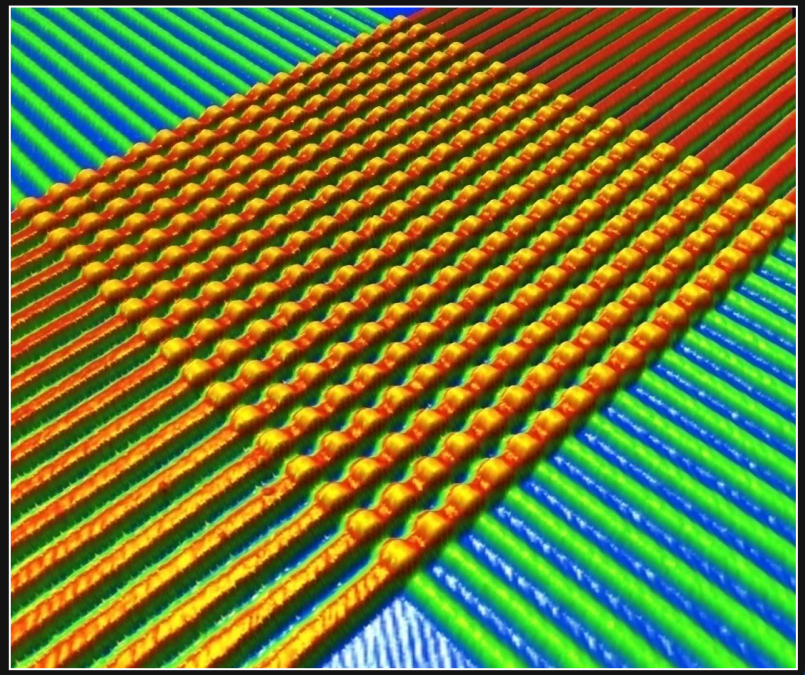
Single wavelength | CWDM | DWDM

100pJ/bit | >.1 pJ/bit

Benefits of photonics:

- Integrated photonics has the potential to:
  - Dramatically improve memory bandwidth
  - Significantly improve many-core performance
  - Reduce power
  - Simplify programming
  - All at the same time!
- Near term applications such as optical buses
  - Add significant system flexibility
  - Save latency and power
- Longer term give opportunity to rethink system architecture
  - New architectures & flexibility (e.g., optical buses)
  - Disaggregation and dematerialization enablement

System



- Vision
- Photonics
- System
- Q&A

Critical Component Technologies - NVRAM

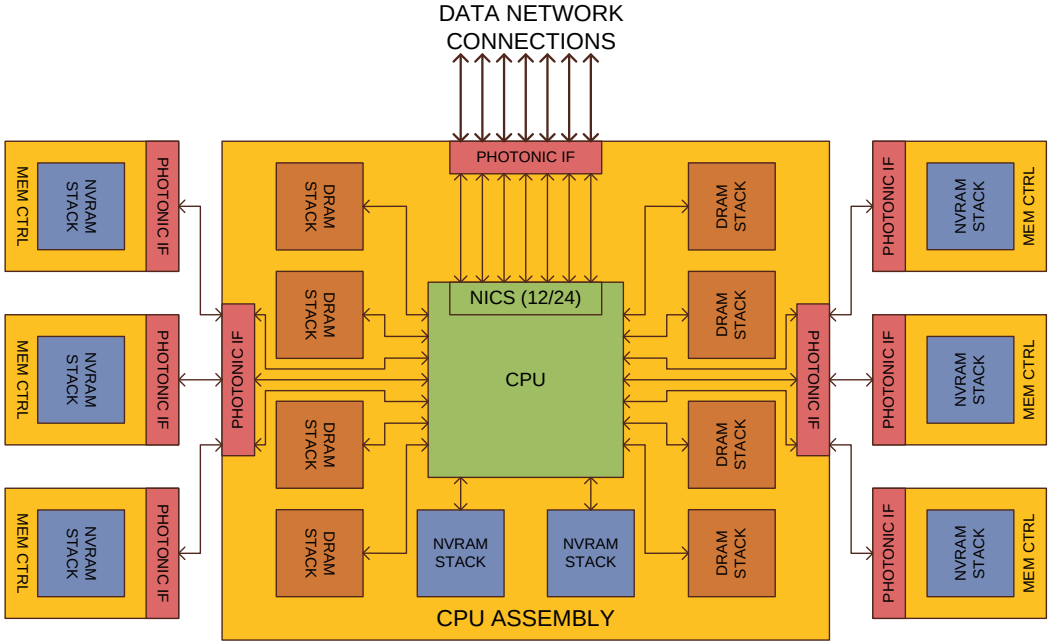
- Essential to provide adequate IO bandwidth
- Addresses shortfall in DRAM scaling
- Greatly superior power proportionality
- Rotating media still the lowest cost per bit

HP LABS

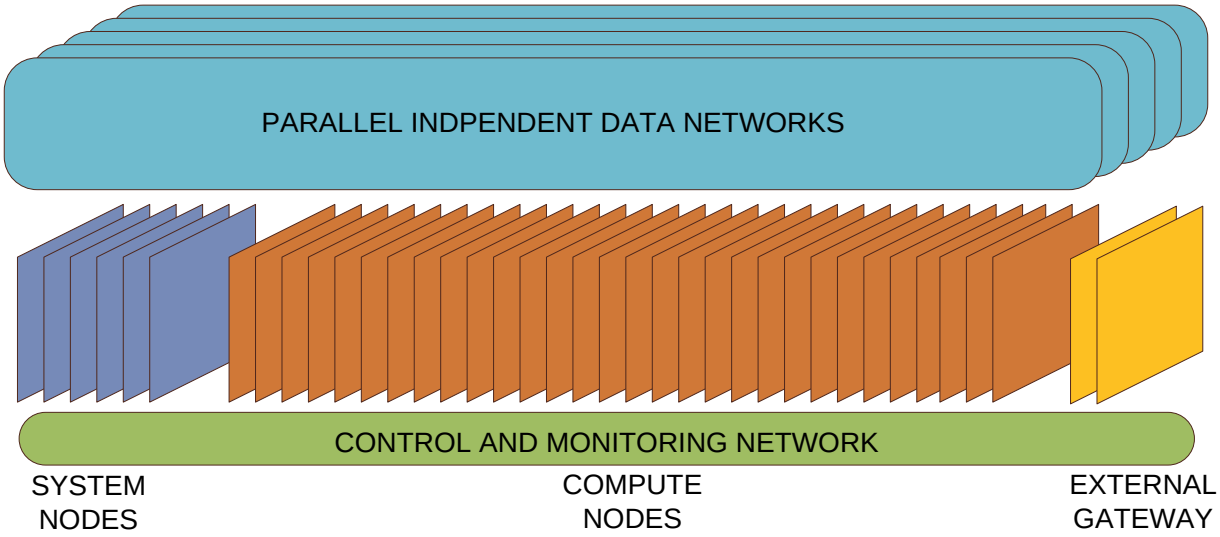
A Strawman Exascale System

What's on a node

- Single chip, highly parallel CPU
- Stacked or on-substrate "near" memory
- DRAM or NVRAM "far" memory
- Integrated network interface
- Multiple photonic links for off-node communications



A Strawman Exascale System

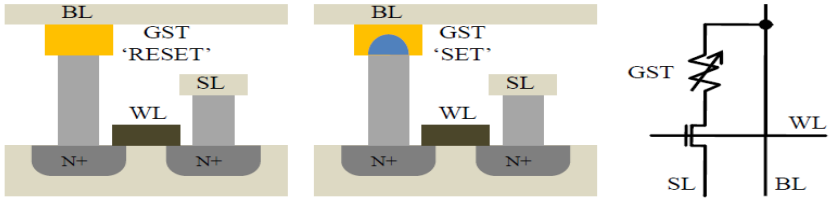


- 100,000, 10Tflop compute nodes (or 1,000,000 1Tflop processors)
- 32 to 64Petabytes of DRAM
- NVRAM capacity of at least 4x DRAM
- 40Petabyte/s network

Technologies for Check-point Restart

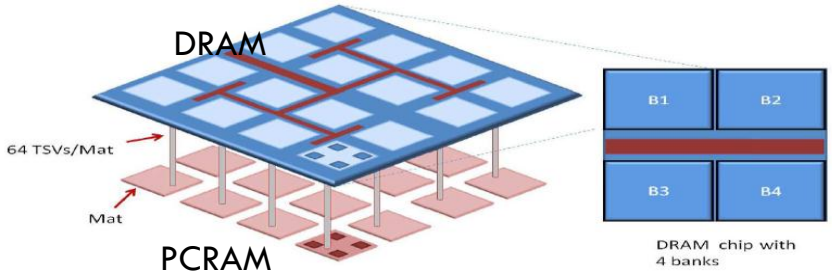
www.nd.edu/~rich/SC09/tut157/SC2009_Jouppi_Xie_Tutorial_Final.pdf

PCRAM



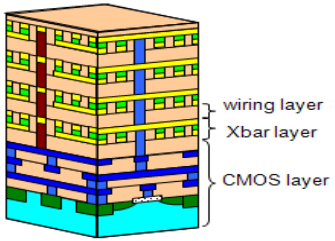
The schematic view of a PCRAM cell with NMOS access transistor (BL=Bitline, WL=Wordline, SL=Sourceline)

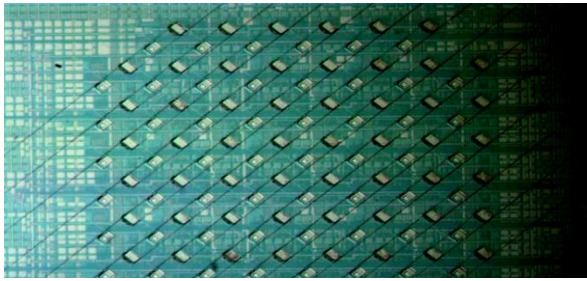
	HDD	NAND Flash	PCRAM
Taille cellule	-	4-6F^2	4-6F^2
Cycle lecture	~4ms	5us-50us	10ns-100ns
Cycle écriture	~4ms	2ms-3ms	100-1000ns
Watt à arrêt	~1W	~0W	~0W
Endurance cycles	10^15	10^5	10^8



Memristor

CMOS chip avec des composants memrésistifs

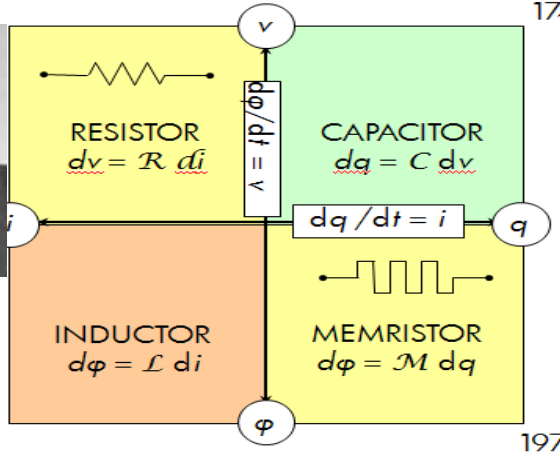




Ohm
1827

Von Kleist
1745

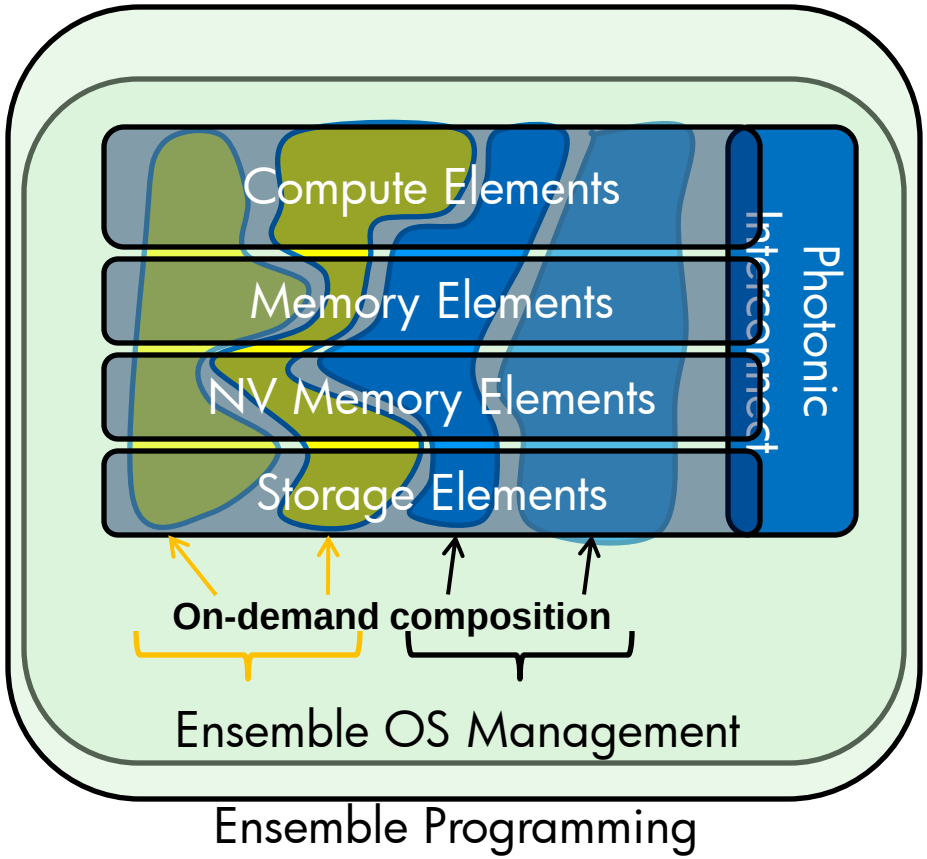
L. O. Chua, (1971)



1831
Faraday

1971
Chua

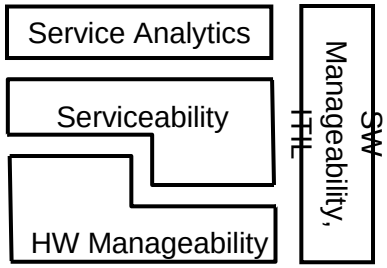
Architecture evolution



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

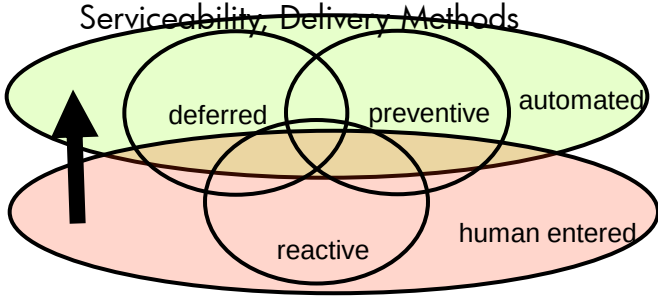
- Disaggregated pools of uncommitted compute, memory, and storage elements
- Optical interconnects enable dynamic, on-demand composition
- Ensemble OS software using virtualization for composition and management
- Management and programming virtual appliances add value for IT and application developers

EXASCALE SYSTEM SUPPORT



– 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

Questions?