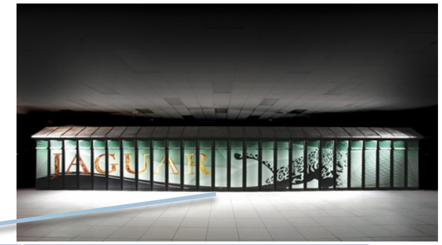
EXASCALE IN 2018 REALLY?

FRANCK CAPPELLO

INRIA&UIUC

What are we talking about?



Systems	2009	2018	Difference Today & 2018
System peak	**2 Pflop/s**	**1 Eflop/s**	**O(1000)**
Power	**6 MW**	**~20 MW (goal)**	
System memory	0.3 PB	32 - 64 PB	O(100)
Node performance	125 GF	1,2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s	2 - 4TB/s	O(100)
Node concurrency	12	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M) 100M cores	O(10) – O(100)
Total concurrency	225,000	O(billion) [O(10) to O(100) for latency hiding]	O(10,000)
Storage	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO	0.2 TB	60 TB/s	O(100)
MTTI	days	O(1 day)	- O(10)

12 cores/node

Power Challenges

Exascale Technology Roadmap Meeting
San Diego California, December 2009.

- **$1M per Megawatt per year → 20 MW Max (50 MW may be).**
- Flops are not really a problem:
  - FMA (fused multiply add) 100picojoules (Now), 10pj in 2018 (on 11nm lithography)
  - → Ok for architects

- Memory bandwidth is critical (biggest delta in energy cost is movement of data offchip):
  - CPU Reading 64b operands from DRAM costs ~2000pj (now), 1000pj in 2018
  - →2000W in 2018 (if 10TFfops/chip) for a ratio of 0.2 byte/flop. Not feasible
  - →200W OK but 0.02 byte/flop (BW → 0.5 byte/flop) → /25
    - → Need for more locality and less memory accesses in algorithms
  - Memory DDR3: 5000pj (read 64b word), DDR5 (2018): 2100 pj (JEDEC roandmap)
  - → At 0.2 B/flop, memory will need 70MW OR 0.02 byte/flop
  - → Need to develop new technologies for 0.2 B/flop but cost will be high

- Network power consumption is critical:
  - Optical links consume about 30-60pj/bit (Now), 10pj/bit in 2018
  - → globally flat bandwidth across a system: Not feasible
  - → topology choice based on power (mesh topologies have power advantages)
  - → algorithms, system software, applications will need to be data locality aware

Application Challenges

Application Programming:

Hybrid multi-core (100-1000 Accelerator cores + 2-2 general purpose cores)

→ hybrid programming will be required (MPI + threads, PGAS)

Less memory per core (could become less than 1GB → 512 MB/core)

→ End of weak scaling, disruptive transition to strong scaling

Less bandwidth for each core (0.02 Byte/flop could be required)

→ Communication avoiding algorithms

Applications candidates:

- Many demanding applications that will need development efforts (next slide)
- Uncertainty Quantification (UQ)

  Accurate model results are critical for design optimization and policy making

  Model predictions are affected by uncertainties: data, model param. (dust cloud…)

  UQ includes uncertainty information in simulations to provide a confidence level

  UQ investigations run ensemble of computational models of different configurations

  → UQ generates a "throughput" workload of O(10K) to O(100K) jobs ("transaction")

  However → UQ generate a vast quantity of data (Exa Bytes), files and directories

  → Database is required to keep the mapping between data, files, etc.

Application Challenges

Table 2. Algorithms expected to play a key role within select scientific applications at the exascale, characterized according to a seven dwarfs classification

Opportunity	Application area	Structured grids	Unstructured grids	FFT	Dense linear algebra	Sparse linear algebra	Particles	Monte Carlo
Material science	Molecular physics			X	X		X	X
	Nanoscale science	X			X		X	X
Earth science	Climate	X	X	X		X	X	X
	Environment	X	X			X	X	X
Energy assurance	Combustion	X			X		X	
	Fusion	X	X	X	X	X	X	X
	Nuclear energy		X		X	X		
Fundamental science	Astrophysics	X	X		X	X	X	
	Nuclear physics				X			
	Accelerator physics		X			X		
	QCD	X						X
Engineering design	Aerodynamics	X	X		X	X		

Resilience Challenge

Node architecture group Exascale Technology Roadmap Meeting San Diego California, December 2009:

•The current failure rates of nodes are primarily defined by market considerations rather than technology

•Because of technology scaling, transient errors will increase by factor of 100 x to 1000x.

→ Vendors will need to harden their components

•Market pressure will likely result in systems with MTTI 10x lower than today

→Today: 5-6 days for the hardware

→MTTI will be O(1 day).

However software is also a significant source of faults, errors and failures

→ Some studies consider that it is the main factor reducing the full system MTTI (Oliner and J. Stearley, DSN 2008, Charng Da lu, Ph. D thesis 2005):

→Bad scenarios consider full system MTTI of 1h…

Resilience Challenges

IESP Oxford
April 2010

	Critical Path	RollBack/ Reco	Fail. Avoid.
Uniquely Exascale:			
-Performance measurement and modeling in presence faults (Perf.)	X		
Exascale plus Trickle down (Exascale will drive):			
Application successful execution & correctness (Masking approach)			
-Better fault tolerant protocols (low overhead)	X	X	?
-Fault isolation/confinement + specific local management (software)	X	X	
-Use of NV-RAM for local state storage, cache of file syst.	X	X	
-Replication (TMR, backup core)	?		X
-Proactive actions (migration), automatic or assisted?	Pr.		X
Application execution and result correctness (Non masking approach)			
-Domain Specific API and Utilities for frameworks	X	X	
-Application guided (level) fault management	Pr.	X	
-Language, Libraries, compiler support for resilience	X	X	
-Runtime/OS API for fault aware programming (access to RAS, etc.)	X	X	
-Resilient Apps. + Numerical Libs & algo. (open question)	X?		X
Reliable System			
-Fault oblivious system software (and produce less faults)	X		X
-Fault aware system software (notification/coordination backbone)	X		X
-Prediction for time optimal checkpointing and migration		X	X
-Fault models, event log standardization, root cause analysis	X		X
-Resilient I/O, Storage and file systems	X		X
-Situational awareness	X		X
Experimental env. to stress & compare solutions	X	X	X
Debugging under the presence of errors/failures + considering faults	X		
Primarily Sub-Exascale (Industry will drive)			
-Fault isolation/confinement + local management (Hardware)	X		X
-Checkpoint of Heterogeneous architecture	X	X	

Exascale in 2018

Yes some hardware will probably be there

BUT

-what applications will be able to exploit even 5-10% of it with

+Strong Scaling (lower memory per core)

+Mesh topology

+0.02 Bytes / Flop (0.2 if we are lucky)

+MTBF of 1 hour (5h-10h if we are lucky)

May be ensemble calculation (UQ) is the most likely “applications” to run first at Exascale

→problem: this is not an “Exascale” application in the sense of a single code running over the whole computer.