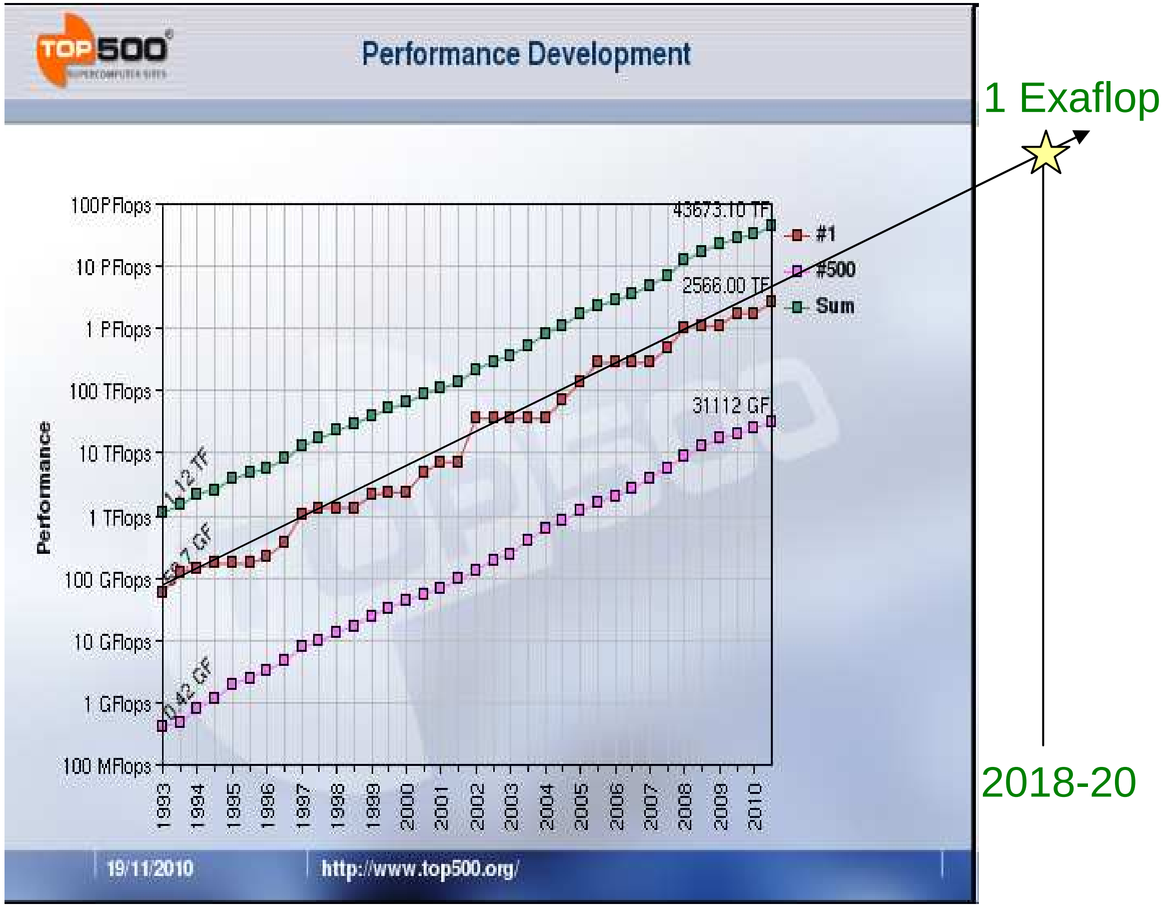
Future Trends towards Exascale

Jean-Pierre Panziera –March 21, 2011

Architect of an Open World™

Continuous growth of HPC systems performance

- HPC systems performance outruns Moore’s Law (>1000x/10years vs 32x)
- CPU performance increases by Moore’s Law
- To reach higher system performance, system parallelism (# CPUs) has increased

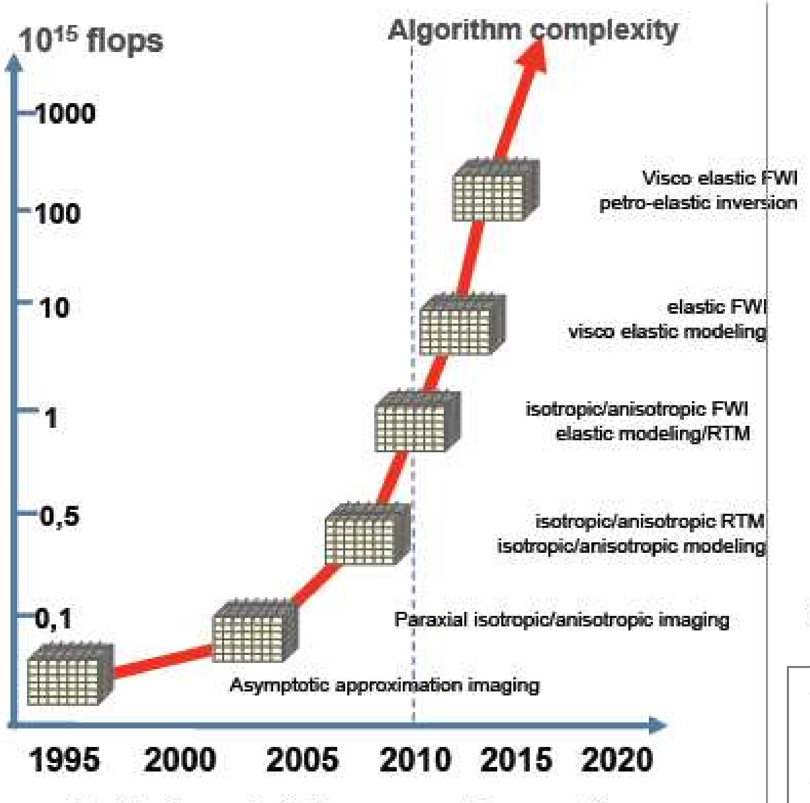


1 Exaflop

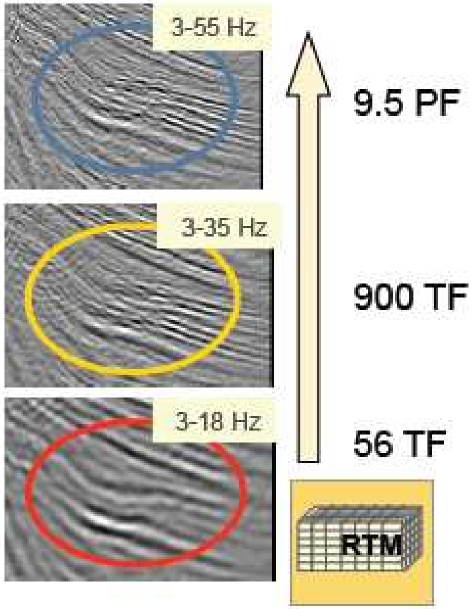
2018-20

Increasing demand for HPC performance

Industrial challenges in the Oil & Gas industry: Depth Imaging roadmap



Algorithmic complexity Vs. corresponding computing power

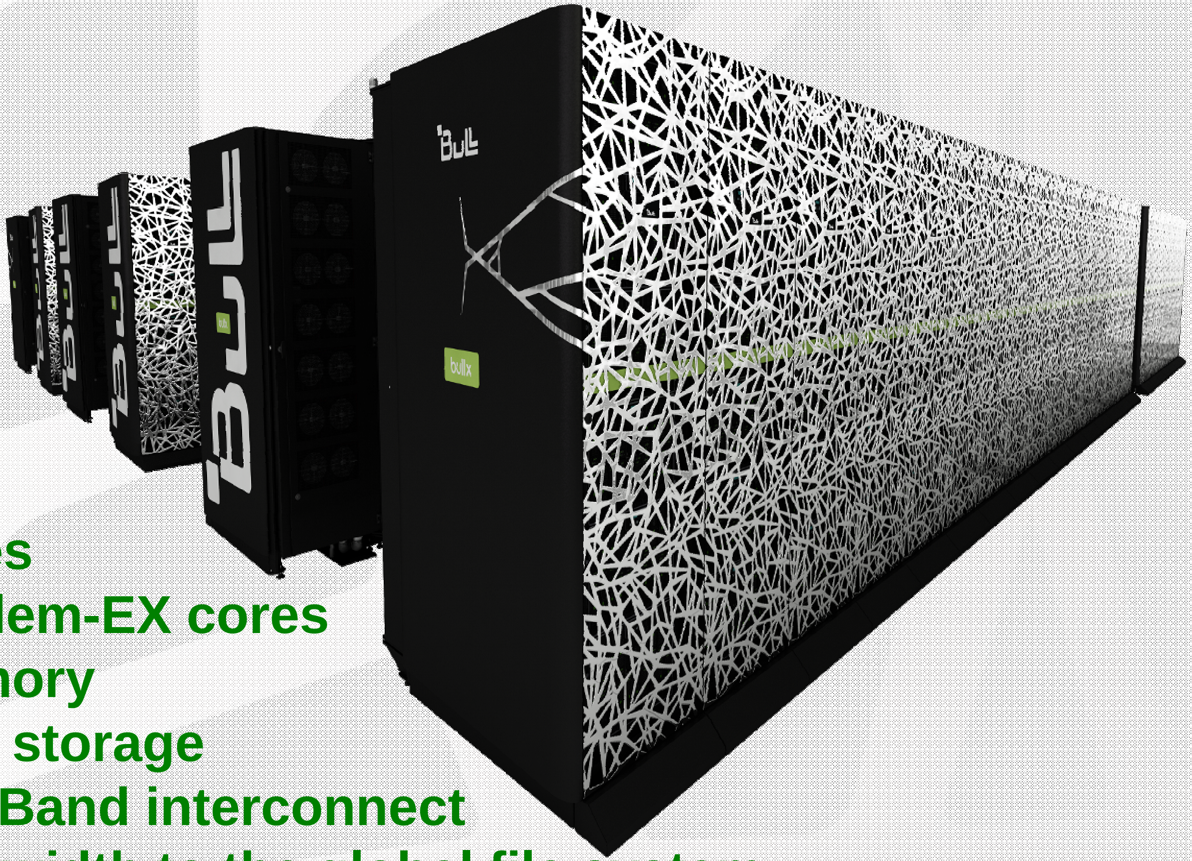


Substained performance for different frequency content over a 8 day processing duration

- Algorithms complexity
  → 100-1000x
- Better Resolution (higher frequency)
  →100-200x
- Overall computation requirements
  → **10 000 - 200 000x**

source: exascale.org

2010: TERA 100



1.25	PFlops
4 300	bullx nodes
140 000	Intel Nehalem-EX cores
300	TB of memory
20	PB of disk storage
	QDR InfiniBand interconnect
500	GB/s bandwidth to the global file system

From Petascale to Exascale x1000 in <10 years

Extrapolating today's Petascale systems to the Exascale …

	2010:	<2020:	
Flops	1 PFlop	1 EFlop	1,000x
nodes	4,000	>128,000	32x
cores	>100,000	>100,000,000	1,000x
Memory Capacity	300 TB	150 PB	500x
Memory Bandwidth	>500 TB/s	> 250 PB/s	500x
Storage Capacity	20 PB	20 EB	1,000x
Interconnect BW	40 Gb/s	8 Tb/s	200x
Storage Bandwidth	500 GB/s	100 TB/s	200x

Exascale Technology Challenges

- Processor design : architecture and frequency

  Multi/Many-cores, Accelerators, …

- Memory Capacity & BW → MCM, 3D Packaging ?

  Feeding enough Bytes to the FP engines, fast enough

- Network bandwidth, latency, topology and routing

  Optical connections/cables , fewer hops, compact packaging

- I/O scalability and flexibility

  XXXLarge datasets + faster computations → data explosion

- System-level resiliency and reliability

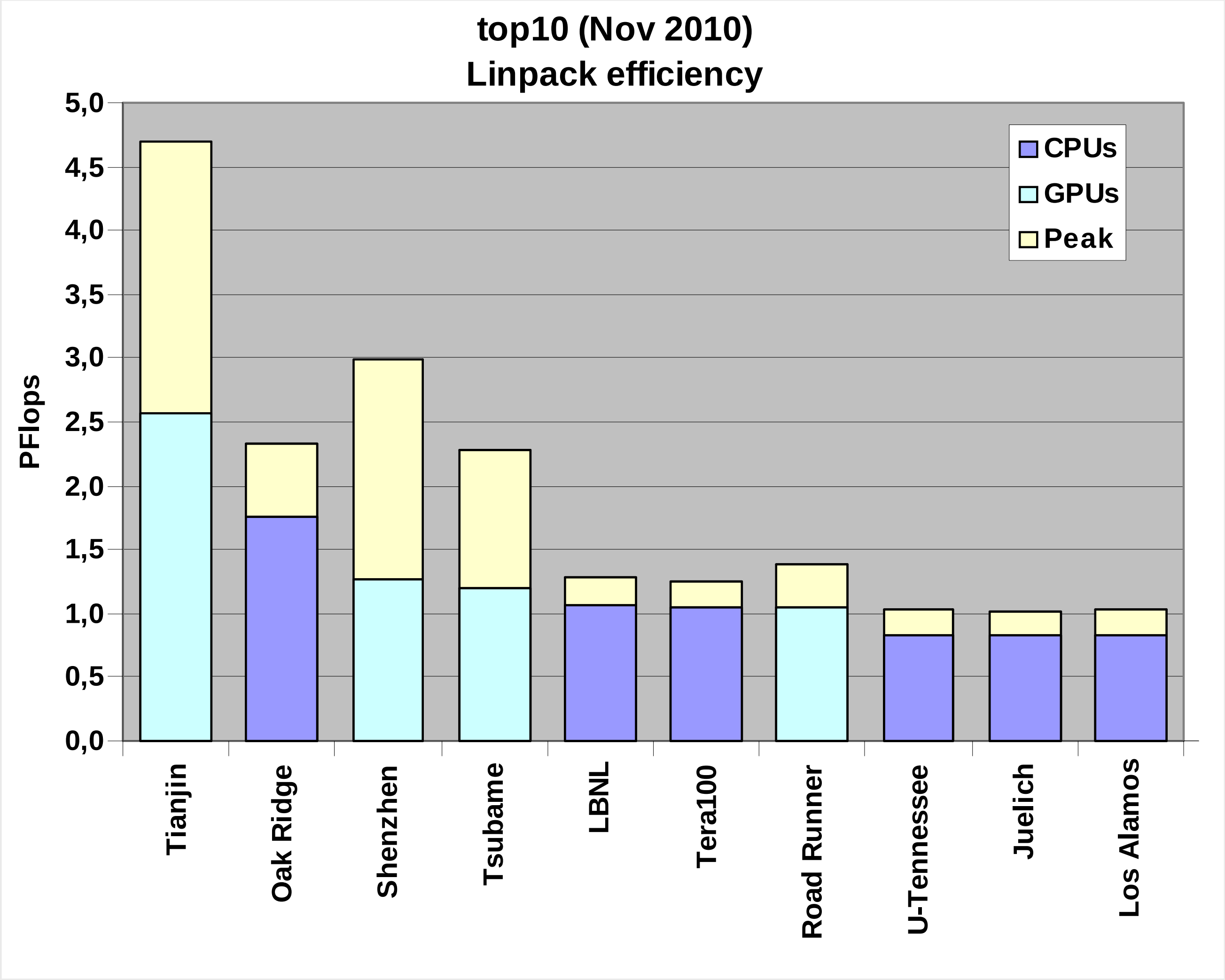
  Month(s) long jobs getting through HW failures

- Power and Cooling

  Fewer less consuming components, improved PUE

- Price ?

Accelerated HPC has reached the top



Traditional sources of performance improvement are Flat-Lining

- # transistors keeps increasing
- Processor frequency stopped at 2-5GHz
- Power per processor socket capped at ~100W
- Processor efficiency not improving anymore. Instruction Level Parallelism (ILP)

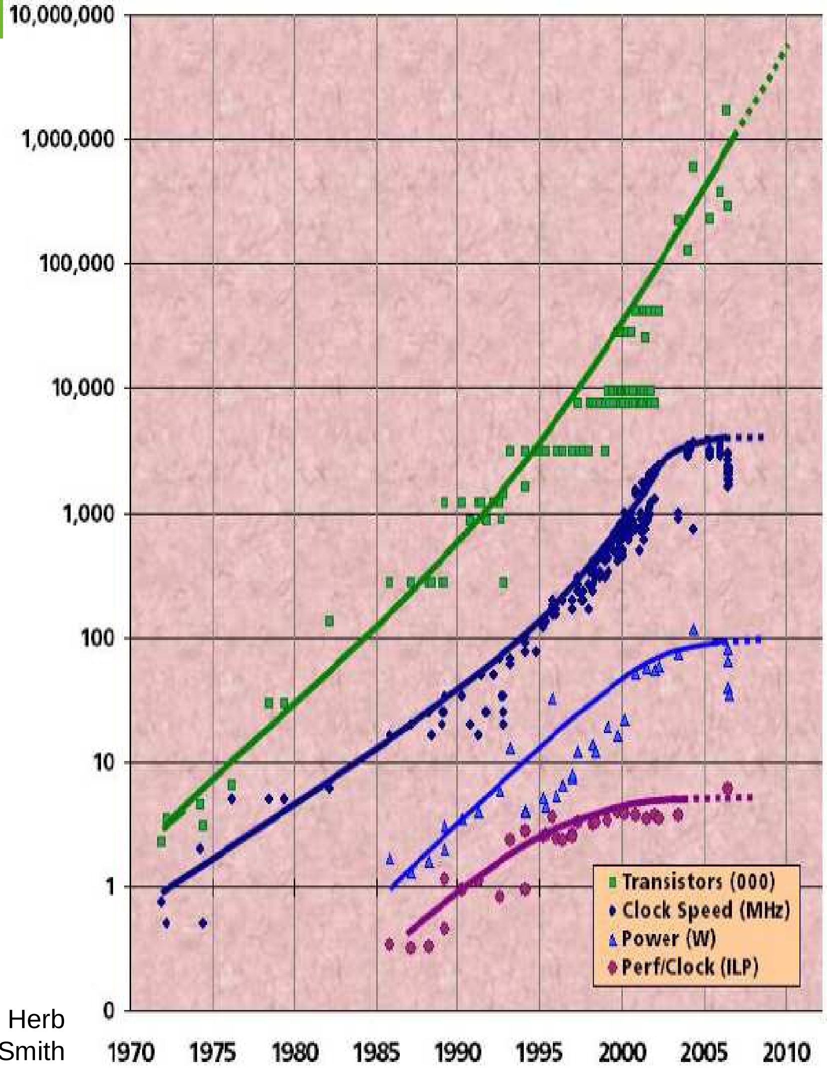
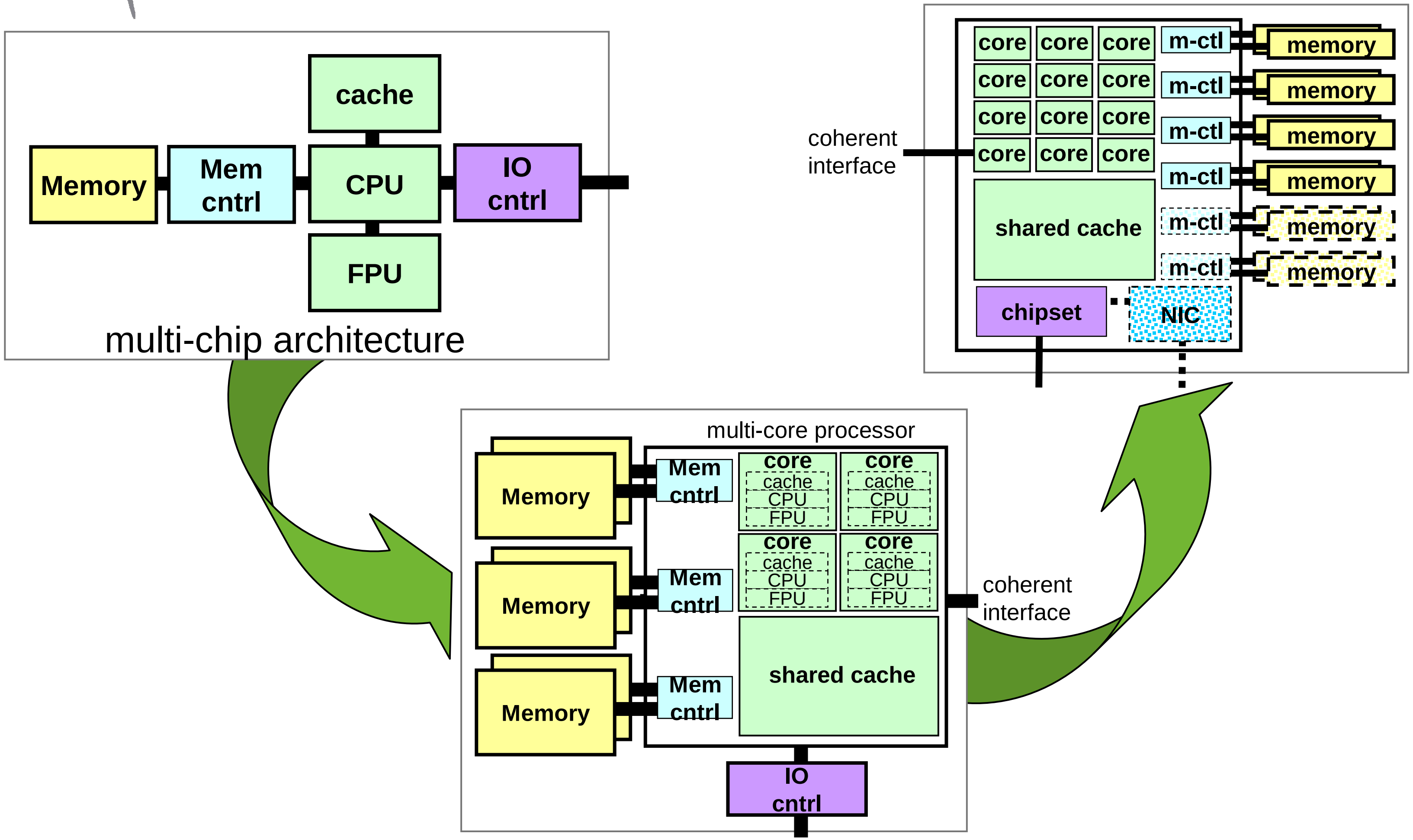


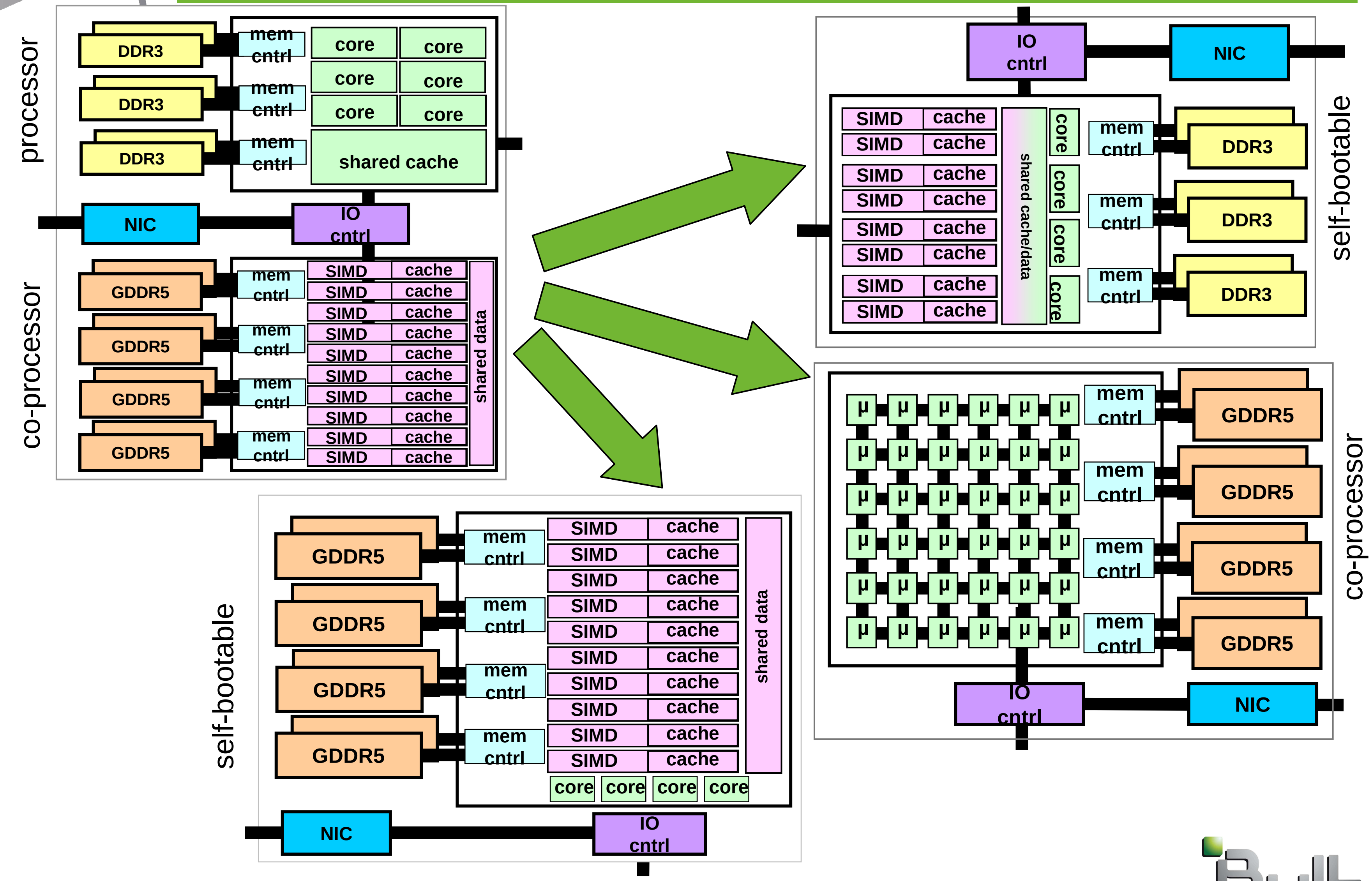
Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Architect of an Open World™

Multi-core CPU architecture evolution



Hybrid CPU-GPU architecture evolutions

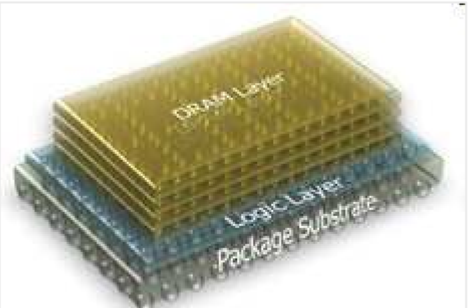


Memory capacity and bandwidth

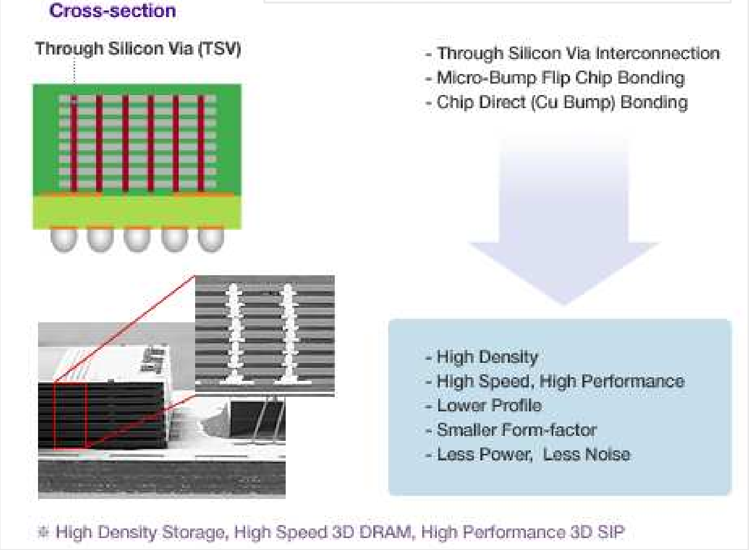
- Using more memory channels per socket is expensive
- Memory Speeding up slowly (DDR2→ DDR3 →DDR4 → …)
- Fast Memory is small and expensive (e.g. GDDR5)
- Speeding up memory + increasing capacity is a real challenge

- New packaging (3D stacking, Multi Chip Modules)
- Extra levels in memory hierarchy
- Smaller data footprint for full bandwidth access
- Select/Develop algorithms with smallest data footprint

3D Memory, Memory-Processor packaging



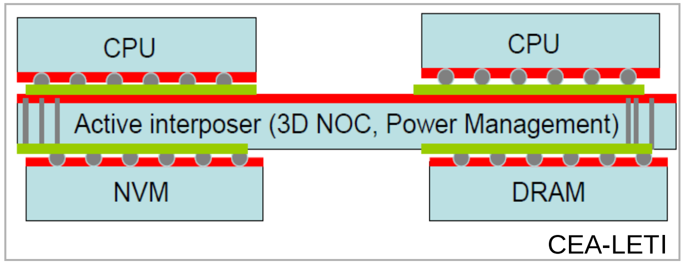
hybrid memory cube (Micron)



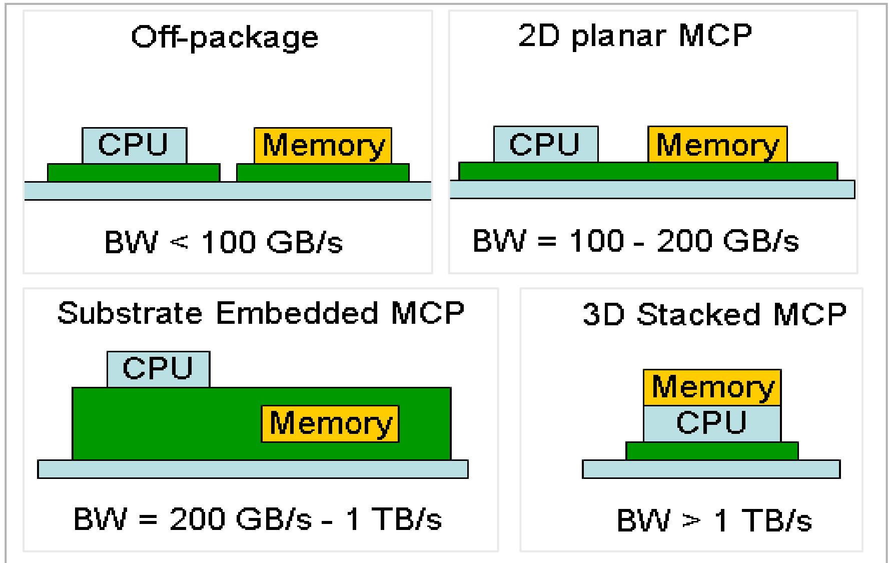
- Through Silicon Via Interconnection
- Micro-Bump Flip Chip Bonding
- Chip Direct (Cu Bump) Bonding

- High Density
- High Speed, High Performance
- Lower Profile
- Smaller Form-factor
- Less Power, Less Noise

※ High Density Storage, High Speed 3D DRAM, High Performance 3D SIP



CEA-LETI

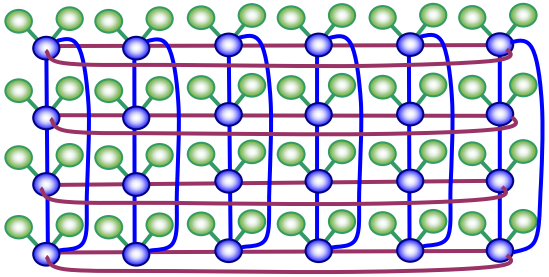


adapted from Polka et al.

higher BW, lower latency, integrated Interconnect

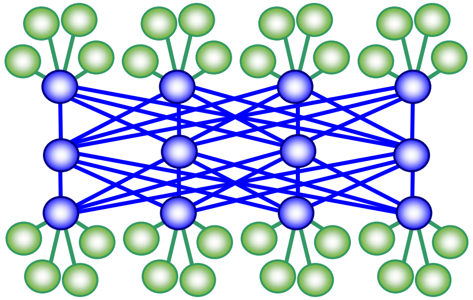
- Signaling frequency evolution is slow (10 → 25 → 50+? gb/s)
- Larger systems → more latency (#hops & wire length)
- Copper wire length gets shorter to keep noise level down

- Better electrical-optical interface for connectors
- More optical links: inter-rack → inter-board → inter-chips → …
- Better interconnect topologies (fewer hops)
- Higher density packaging (smaller distances)
- More efficient congestion control, Adaptive routing



Torus

vs



Fat-Tree

Increase system MTBF (Mean Time Between Failure)

- Current PFlop systems have MTBF ~day(s)
- Larger systems (more components) → MTBF ~hours or <1h
- Checkpoint/Restart frequency will increase

- fewer components → better efficiency
- Self-healing / redundant components
- Failure occurrence integrated into Application development
- Resilient network; multiple-failure resistant
- Local Checkpoint; remote access for Restart

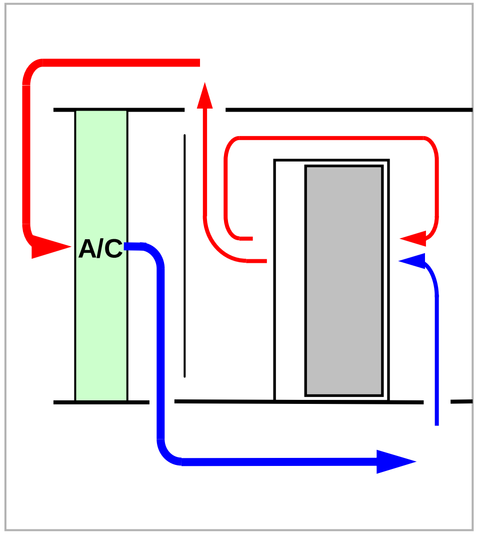
Power and Cooling

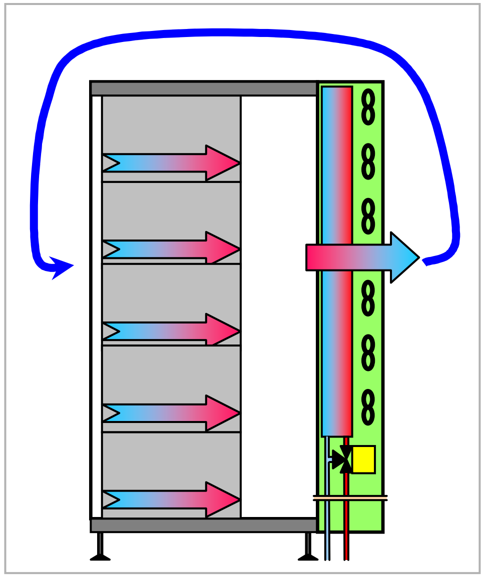
- Current PFlop systems power consumption is high (3-7 MW)
- EFlops systems would consume > 100MW
- Energy price is increasing: 50-100 → 150-200+ €/MWh

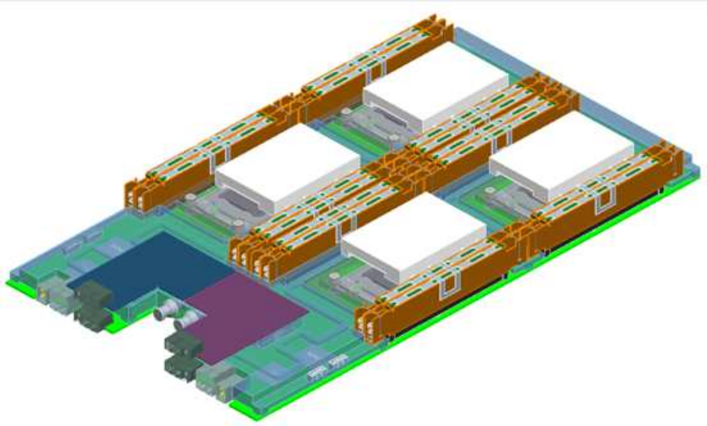
- Less power hungry components
- Better power supply transformation
- Better PUE (Direct Liquid Cooling)
- Cogeneration (re-use of heat produced)

Cooling & Power Usage Effectiveness (PUE)

Air-cooled		Water-cooled doors			Direct-Liquid-cooling	
10(-20) kW/rack		40 kW/rack			70 kW/rack	
Room	20°C	Room	20°C	27°C	Room	27°C
A/C water	7-12°C	Water	7-12°C	14-19°C	Water	ambient θ
PUE	1.8-1.9	PUE	1.6-1.7	1.4-1.5	PUE	1.1-1.2







Storage and Parallel File systems

- File Number + Size explosion
- Larger reconstruction times (performance degradation)
- Multiple failures resilience
- Security
- End to End data protection

- IO servers and RAID controllers integration
- pNFS as generic client protocol
- Non POSIX API
- Declustered RAIDs
- SSDs increase meta data efficiency (IOPs)
- Multi-tiers file systems

Programming models

- Challenges
  - Massive parallelism; Heterogeneity; Complex memory hierarchy
- Programming Languages
  - Hybrid models : MPI, OpenMP, MLP, PGAS, Cuda, OpenCL, new…
  - Expression of parallelism, locality, IO
- Numerical libraries
  - Hybrid libraries
  - Auto-tuning libraries (FFTW, MAGMA, …)
- Development tools: debugging, performance analysis
  - Multi-level analysis
  - Automatic detection of patterns

Data analysis, Visualization, data management

- Access to large data sets
- Statistical methods for exabyte data sets analysis
- Integration of pattern recognition into the simulation and/or I/O operation
- Real time analysis of computation
- Workflow and databases for large scientific data sets

conclusion

- HPC applications requirements keep increasing … well beyond the Petascale → Exascale → …
- HW Accelerators providing a performance boost to HPC applications
- More challenges for the Exascale systems (Memory & Interconnect Bandwidths/Latencies, Resilience, Power)
- Exascale Development tools are still being designed
- HPC applications will need to be modified / revisited / rewritten for Exascale
- Massive amounts of data to analyze

- Interesting times ahead

bullx

instruments for innovation

powered by Bull