Accelerating towards Exascale

Accelerated Computing For Fusion Workshop

Jean-Pierre Panziera

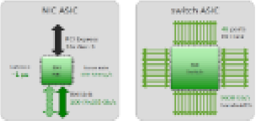
November 29, 2016

Bull
atos technologies

Atos HPC roadmap

2016	2017	2018	2019	2020	2021	2022

ARM

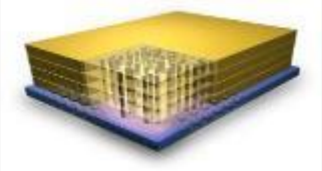


New Exascale
BXI
Interconnect

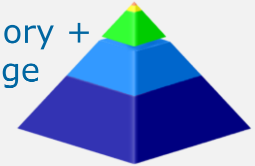
Sequana
New flexible
packaging

SCS

New integrated
computing architecture



Compute + Memory +
Network + Storage
integration



Extreme scale
packaging

SCS

Energy &
performance oriented

10^{18}
Exascale

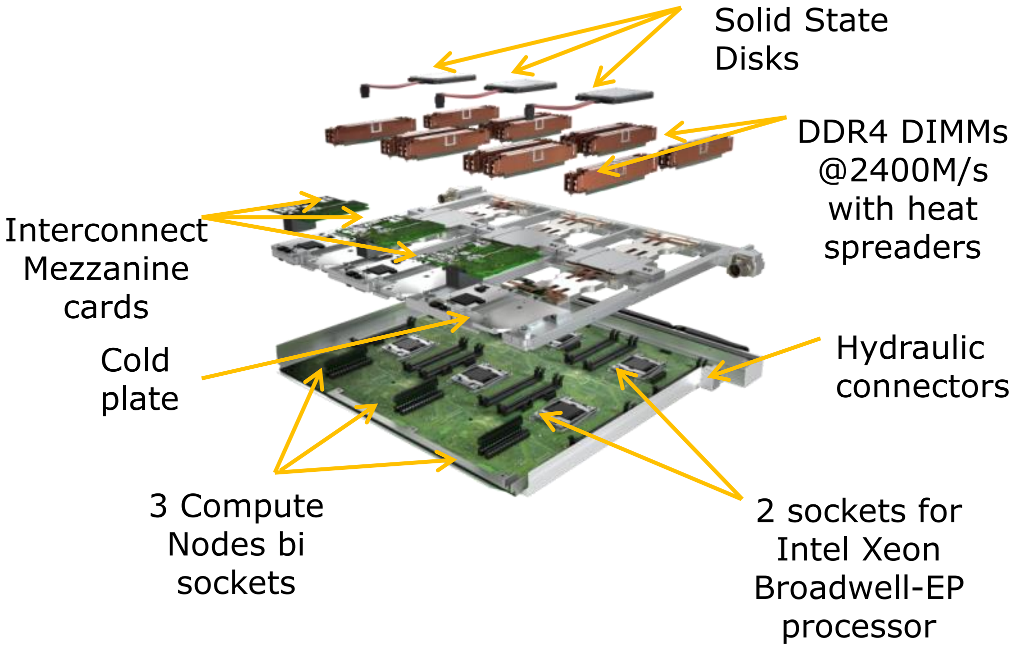
Bull sequana

the Bull exascale generation of supercomputer

- **Open and modular platform designed for the long-term**
  - To preserve customer investments
  - To integrate current and future technologies
  - Multiple compute nodes: Xeon-EP, Xeon Phi, NVIDIA© GPUs, other architectures…
- **Scales up to tens of thousands of nodes**
  - Large building blocks to facilitate scaling
  - Large systems with DLC: 250-64k nodes
- **Embedding the fastest interconnects**
  - Multiple Interconnects: BXI, InfiniBand
  - Optimized interconnect topology for large basic cell / DLC (288 nodes)
  - Fully non-blocking within cell
- **Lowest TCO / Ultra-energy efficient**
  - Enhanced DLC – up to 40°C for inlet water and ~100% DLC

Bull sequana compute blade: X1110

Intel Xeon-EP (Broadwell-EP)

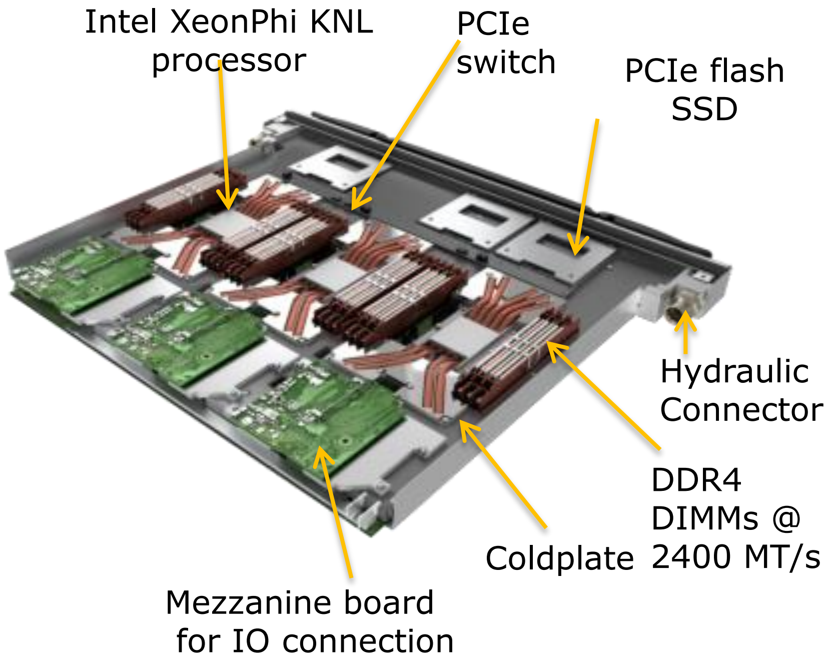


Solid State Disks
DDR4 DIMMs @2400M/s with heat spreaders
Interconnect Mezzanine cards
Cold plate
Hydraulic connectors
3 Compute Nodes bi sockets
2 sockets for Intel Xeon Broadwell-EP processor

- **1U form factor**
- **Direct liquid cooling**
- **3 compute nodes per blade with:**
  - 2 Intel Xeon Broadwell-EP
  - 8 DDR4 DIMM slots
  - 1 I/C mezzanine board (BXI or EDR)
  - Optional 1 x 2.5” 7 mm SATA SSD

Bull sequana compute blade: X1210

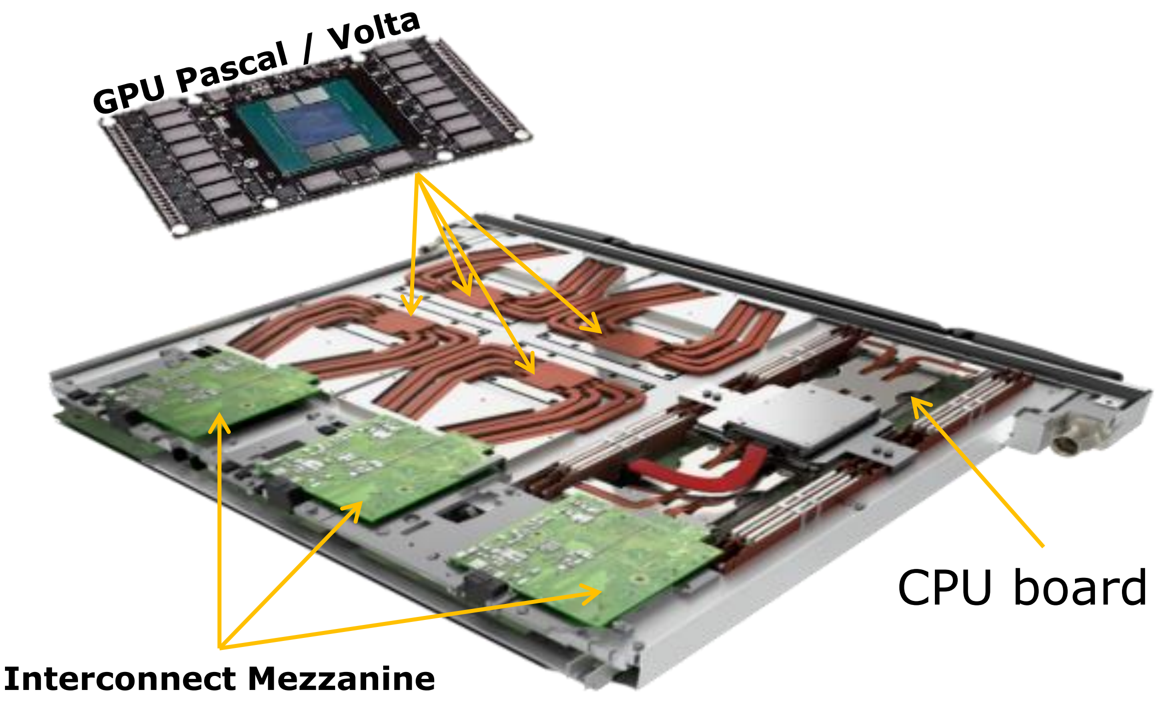
Intel Xeon Phi – Knights Landing (KNL)

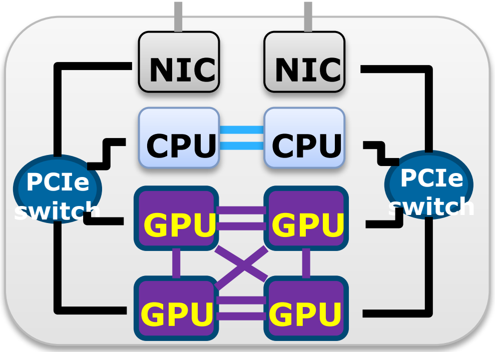


- **1U form factor**
- **Direct liquid cooling**
- **3 compute nodes per blade with:**
  - 1 Intel Xeon Phi KNL processor
  - 6 DDR4 DIMM slots
  - 1 I/C mezzanine board (BXI or EDR)
  - 1 optional board with a PCIe switch
  - Optional disks
    - 1 x 2.5” SATA SSD
    - Or 1x 2.5” PCIe flash SSD, connected to an embedded PCIe switch

Bull sequana accelerator blade: X1125

Nvidia Pascal & Volta



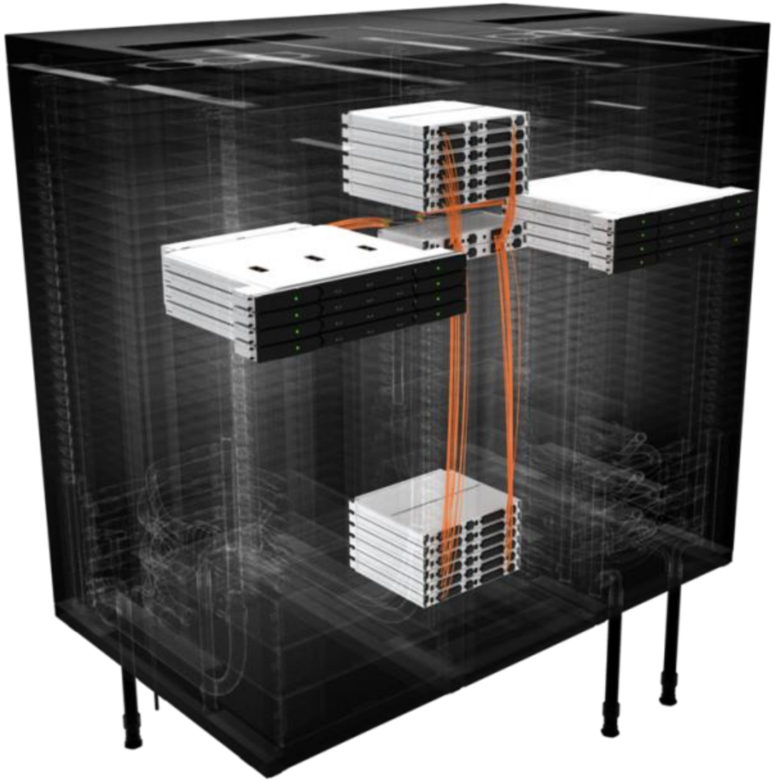


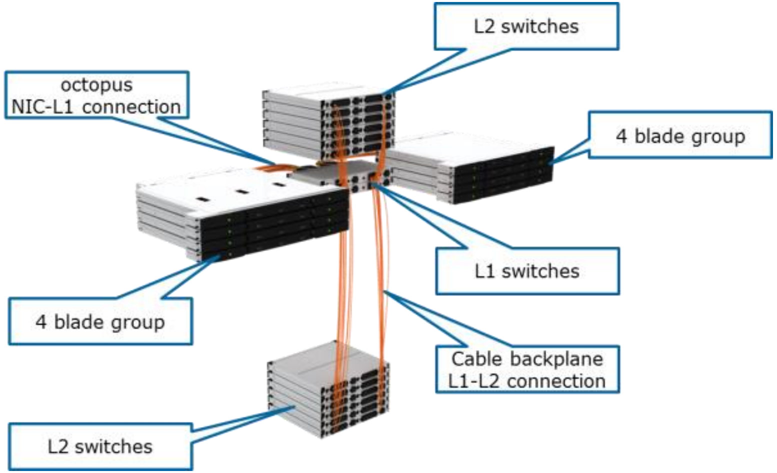
- **1U form factor**
- **Direct liquid cooling**
- **1 compute node per blade with:**
  - 1 CPU Board with 2 sockets
  - 1 GPU board supports up to 4 GPUs
  - 2 I/C mezzanine boards (BXI or EDR)
  - Optional 1 x 2.5” 7 mm SATA SSD

BXI – Interconnect overview

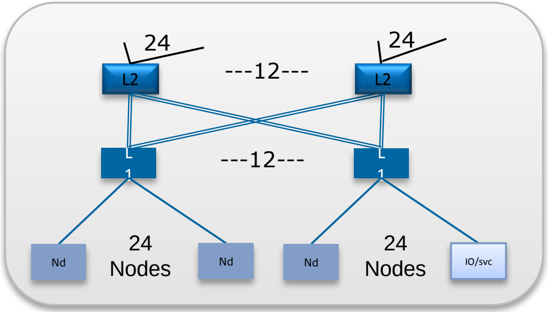
- **BXI 1st generation of Bull Exascale Interconnect**
  - Hardware acceleration → sustained performance under heavy load,
  - High Bandwidth, low latency, high message rate at scale.
- **BXI full acceleration in hardware for HPC applications**
  - Based on Portals 4, a rich low level network API for message passing.
  - HW support for:
    - MPI and PGAS communications over Portals 4 (send/recv, RDMA),
    - High performance collective operations.
- **BXI highly scalable, efficient and reliable**
  - Exascale scalability → 64k nodes,
  - Adaptive Routing,
  - Quality of Service (QoS),
  - End-to-end error checking + link level CRC + ASIC ECC.

Bull sequana X1000 – Embedded interconnect





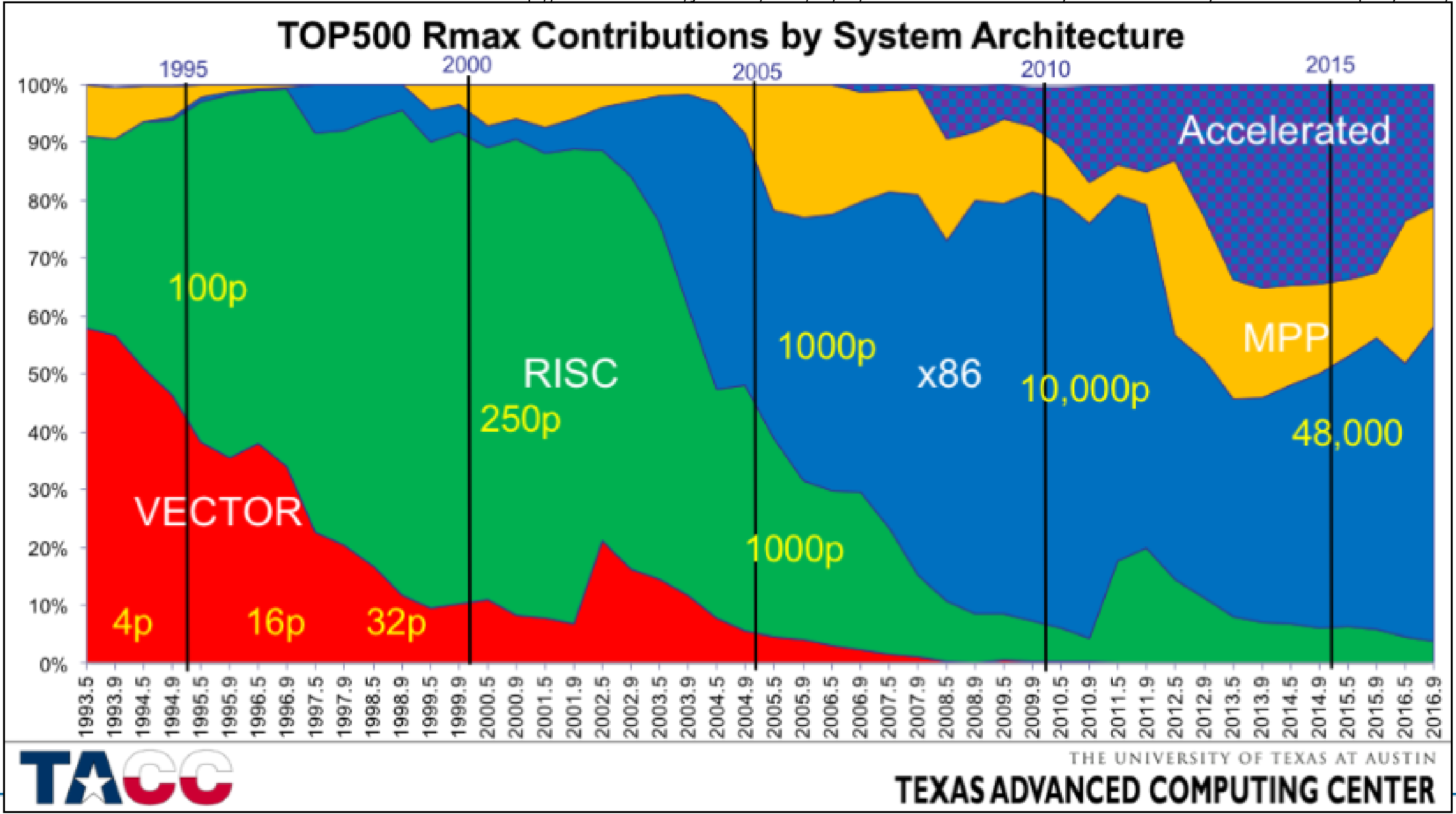
Fat-Tree – 2 Levels – 288 nodes



Fast Interconnect layout

Accelerated computing

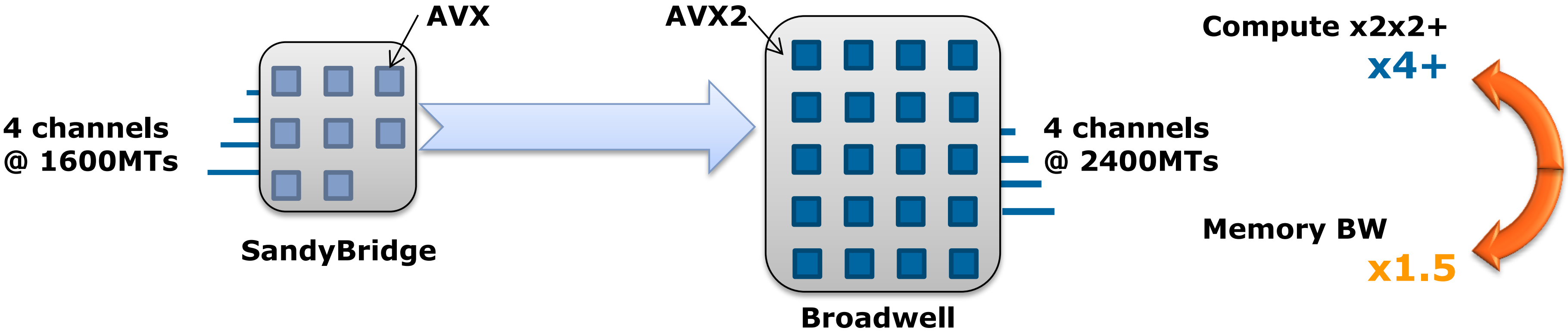
http://sites.utexas.edu/jdm4372/2016/11/22/sc16-invited-talk-memory-bandwidth-and-system-balance-in-hpc-systems/



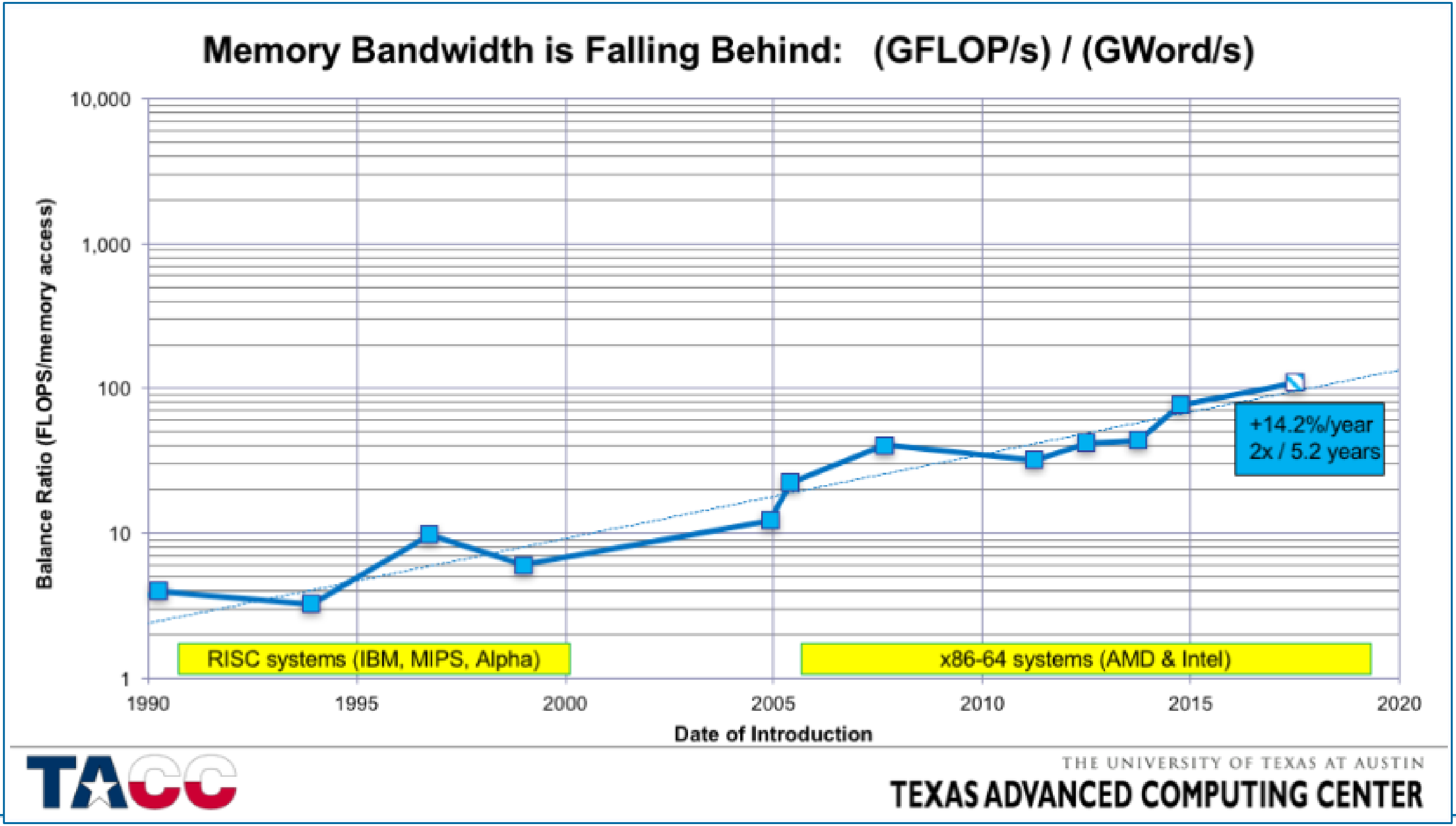
TOP500 Rmax Contributions by System Architecture

SC16
John McCalpin

From Petaflops …
to moreFlops but fewmoreBytes



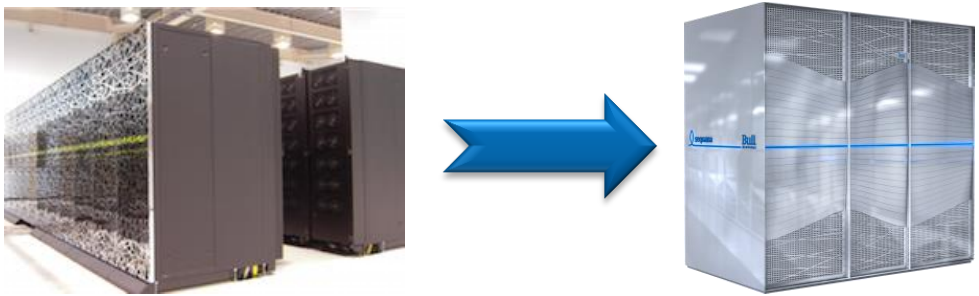
the bandwidth!

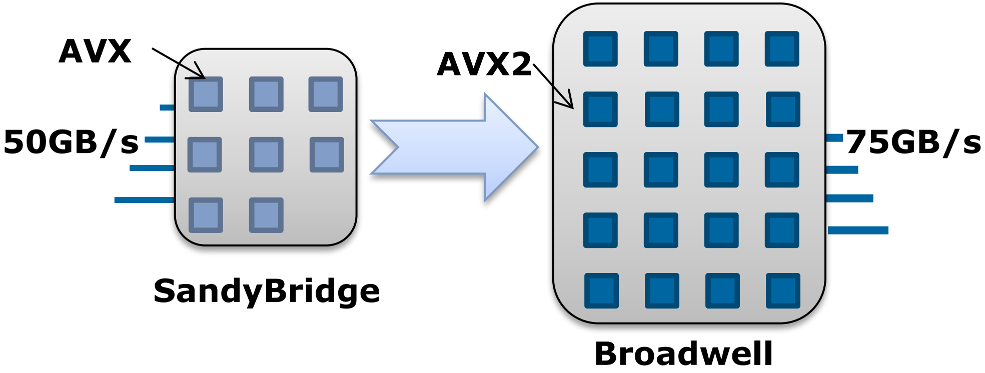


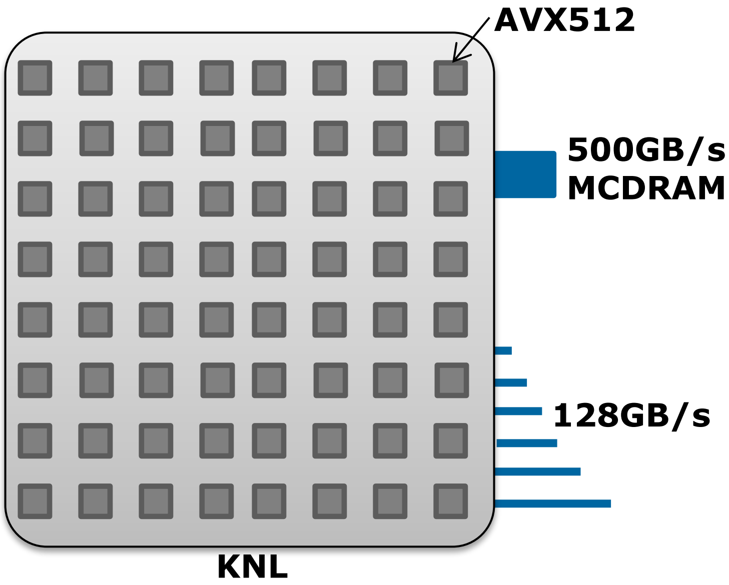
Memory Bandwidth is Falling Behind: (GFLOP/s) / (GWord/s)

SC16
John McCalpin

From Petaflops …
to moreFlops and moreBytes

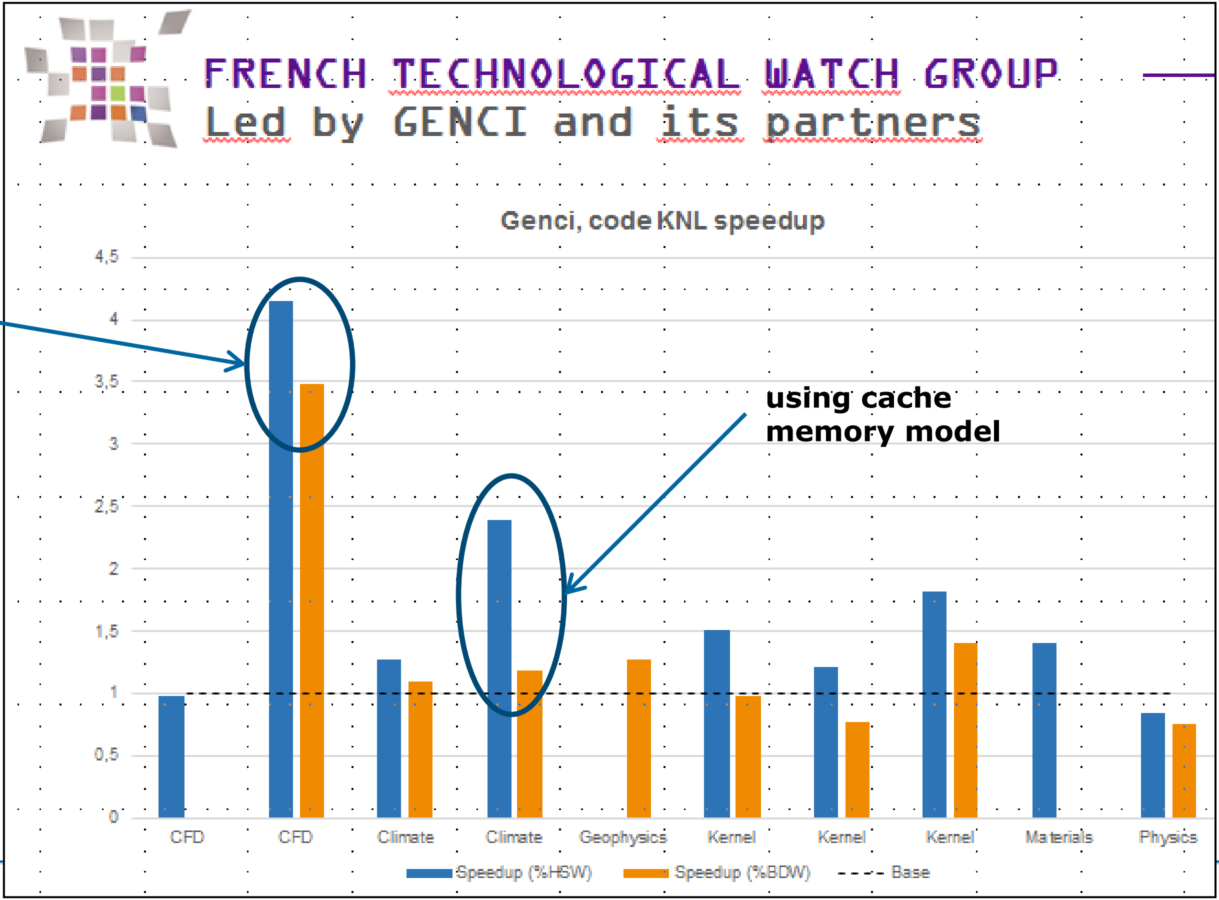






Performance may vary: 1KNL vs 1 node bi-socket

using flat memory model



FRENCH TECHNOLOGICAL WATCH GROUP
Led by GENCI and its partners

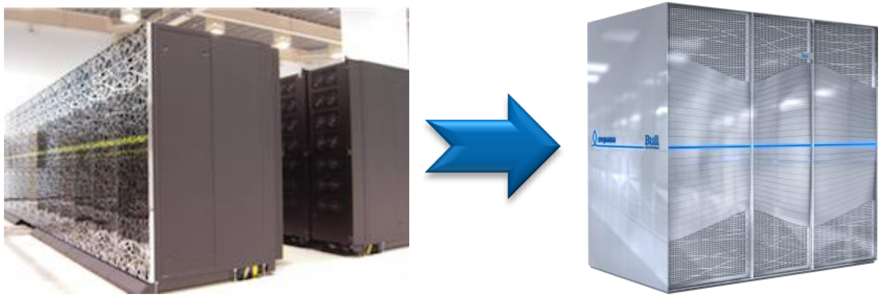
Genci, code KNL speedup

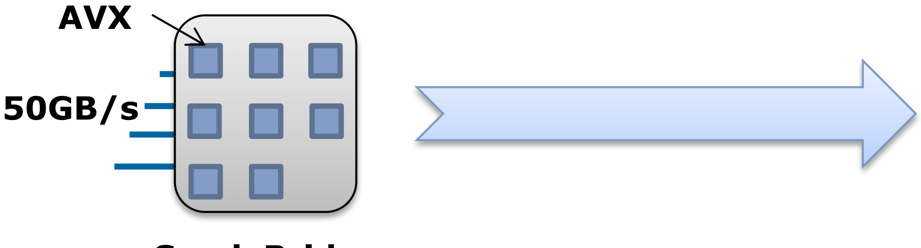
using cache memory model

SC16
Atos / GENCI

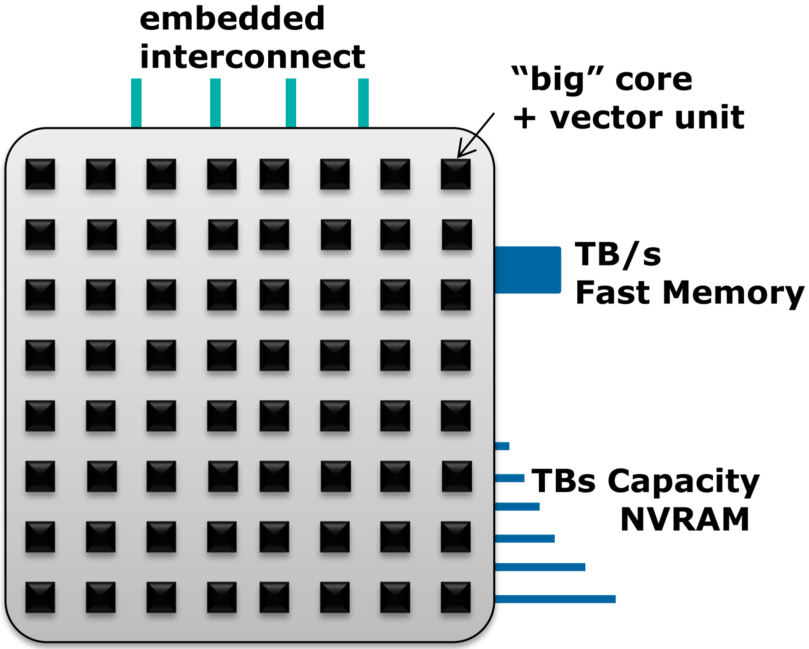
From Petascale …

… to Exascale





SandyBridge

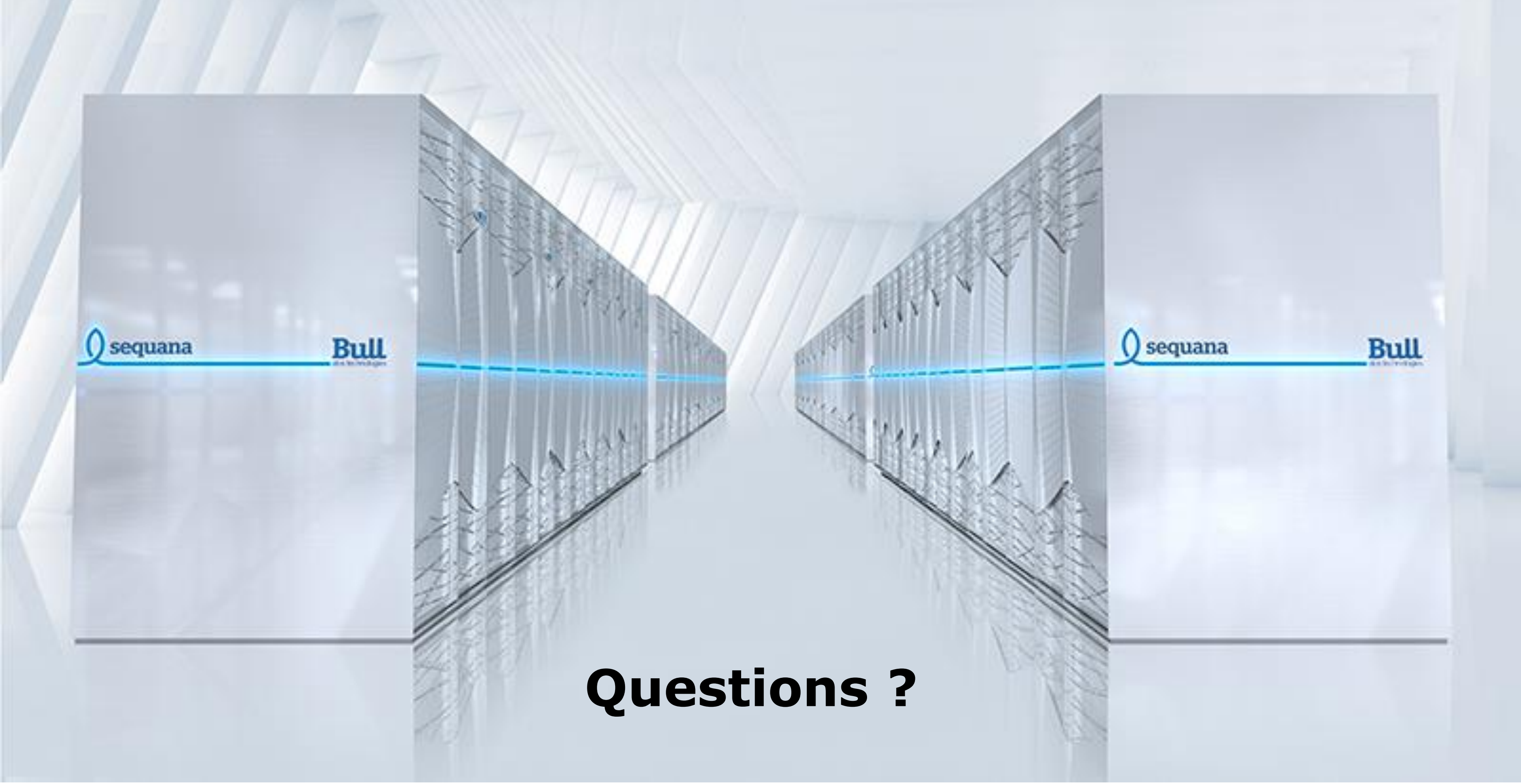


possible Exascale Processor

Accelerating towards Exascale

Exascale system architecture:

- Large nodes → fewer nodes
- Powerful processing units : “big” cores + vector unit
- Fast memory to feed processing units: larger BW, reduced latency?
- Large capacity & performant data access
  - NVMe interfaced devices
  - fast NVRAM: disk capacity, ~DRAM performance
- embedded interconnect: low latency, high message rate, high BW



Questions ?