JÜLICH FORSCHUNGSZENTRUM

Introduction to the massively parallel Blue Gene/Q architecture

D. Pleiter | CODE JAM 2014 | 27 January 2014

Outline

Why going parallel?

- Brief introduction into relevant technology trends

Blue Gene/Q architecture overview

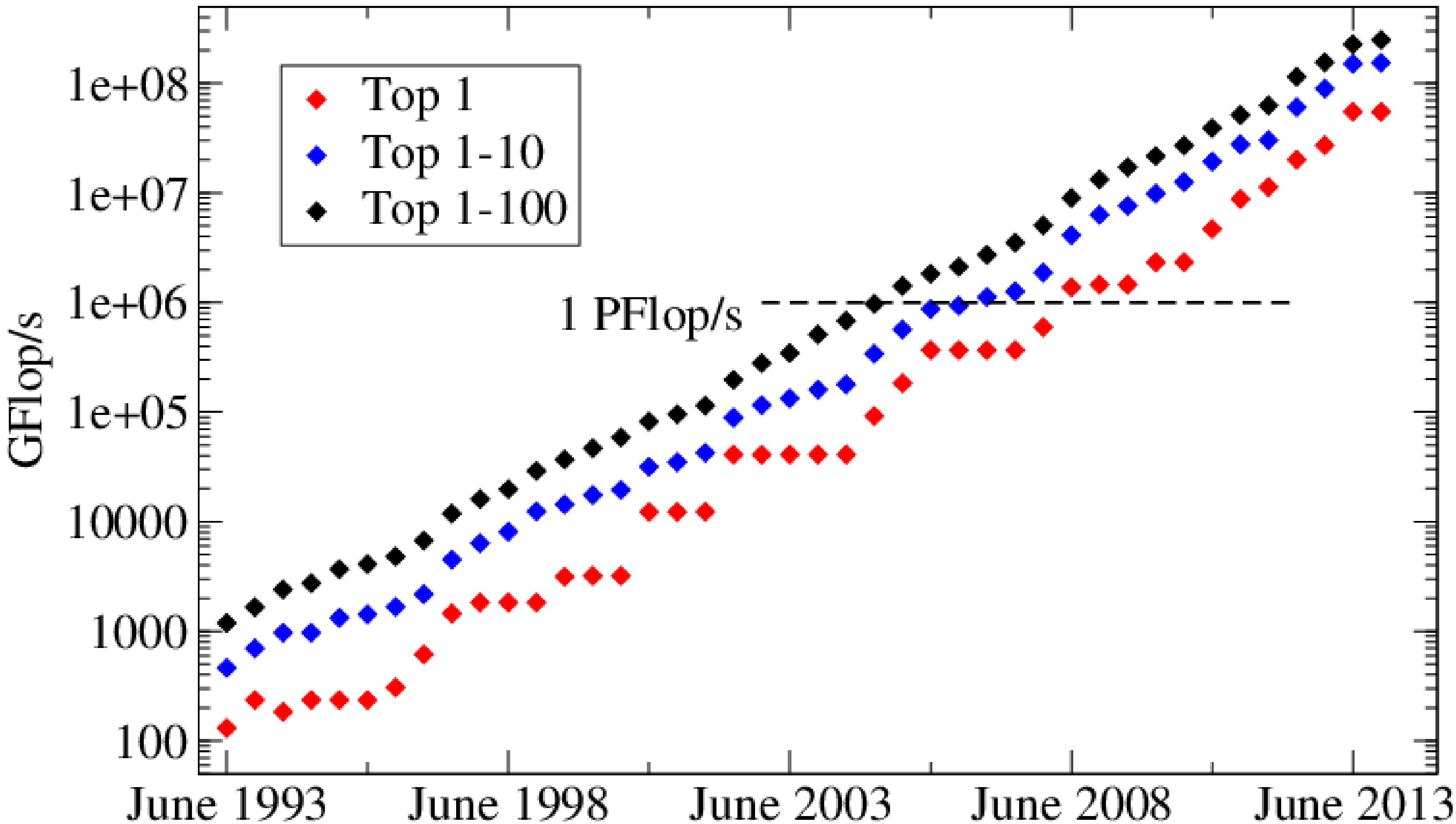
- From core to full system design

Selected features of the Blue Gene/Q architecture

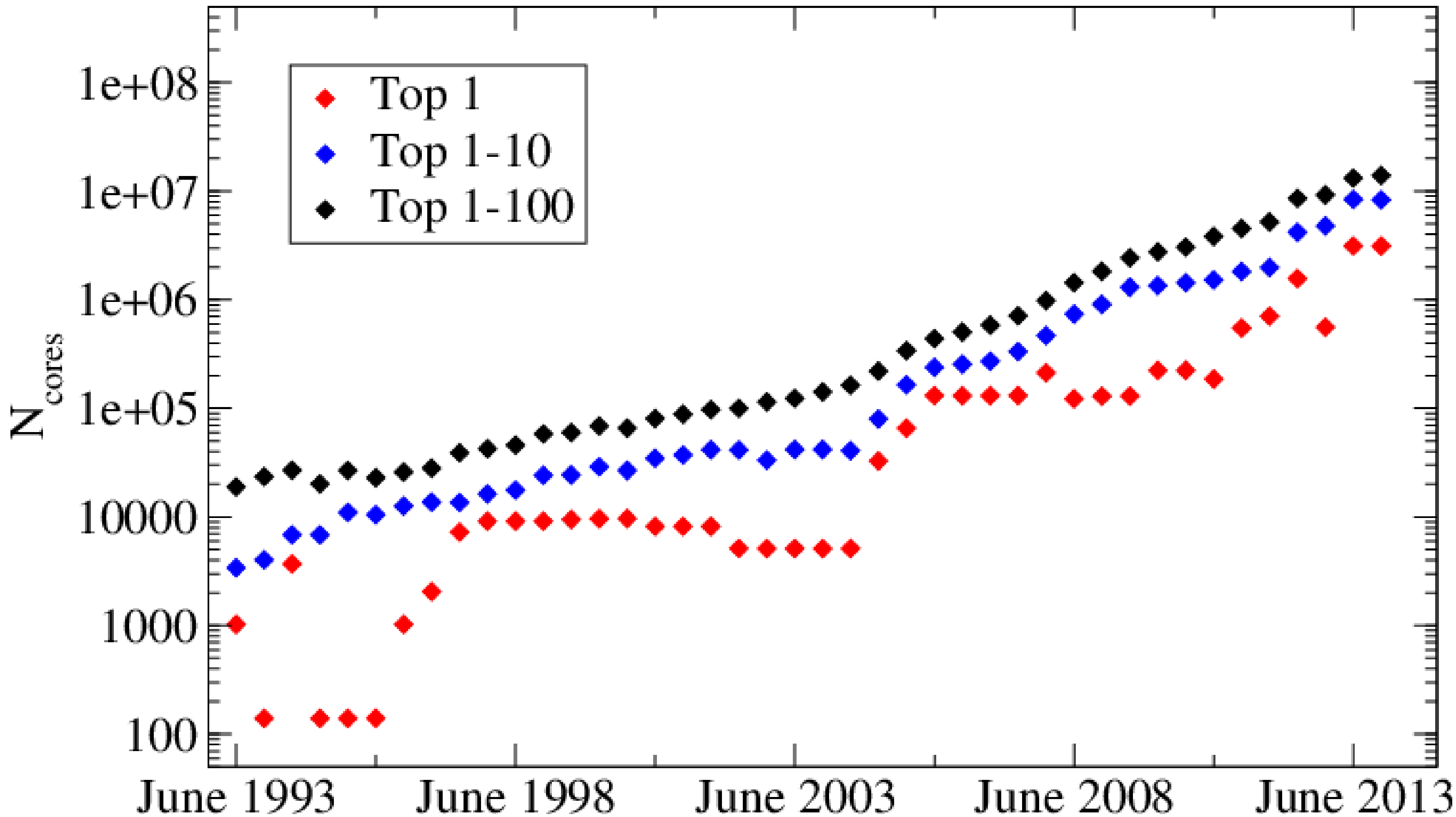
- And how to exploit them

Comparison to other HPC architectures

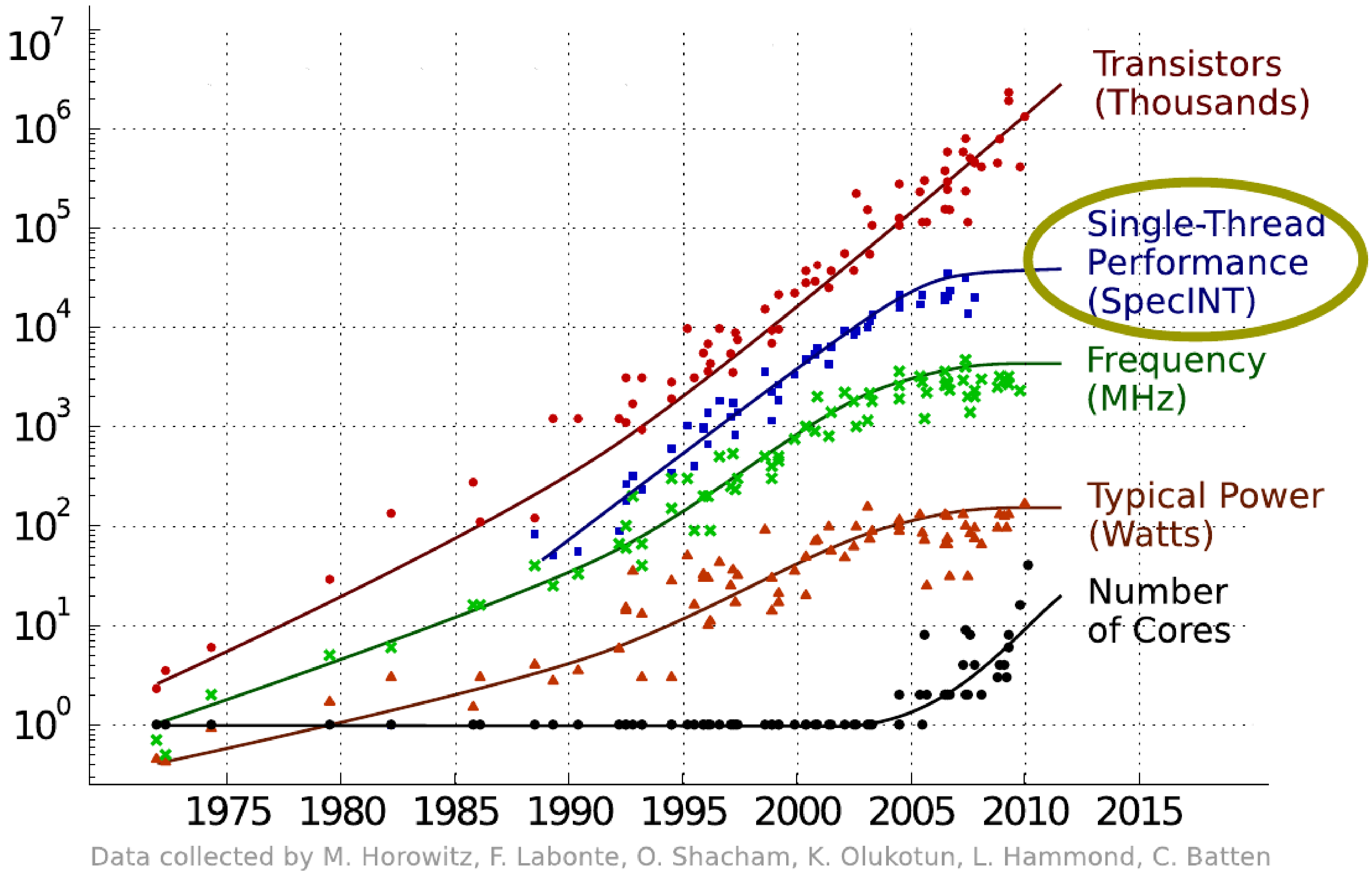
Top500 trends: Peak performance



Top500 trends: Number of "cores"



Single processor performance limits



Parallel data processing

Many nodes

- Nodes with one or more processors connected via a high-performance network
- Typically no shared memory

Multiple processors and cores per node

- Multiple processing units within a single die, multi-chip modules or different sockets connected via a fast interconnect bus
- Typically shared memory and cache coherent

Multiple hardware threads per core

- Hardware support for running multiple sequences of instructions concurrently on the same core

SIMD instructions

- Single instructions processing multiple data streams

Data storage and transport

Memory and network performance

- Fast enough to feed processing pipelines
- Performance = bandwidth, start-up latency

Memory capacity

- Large enough to hold the full problem

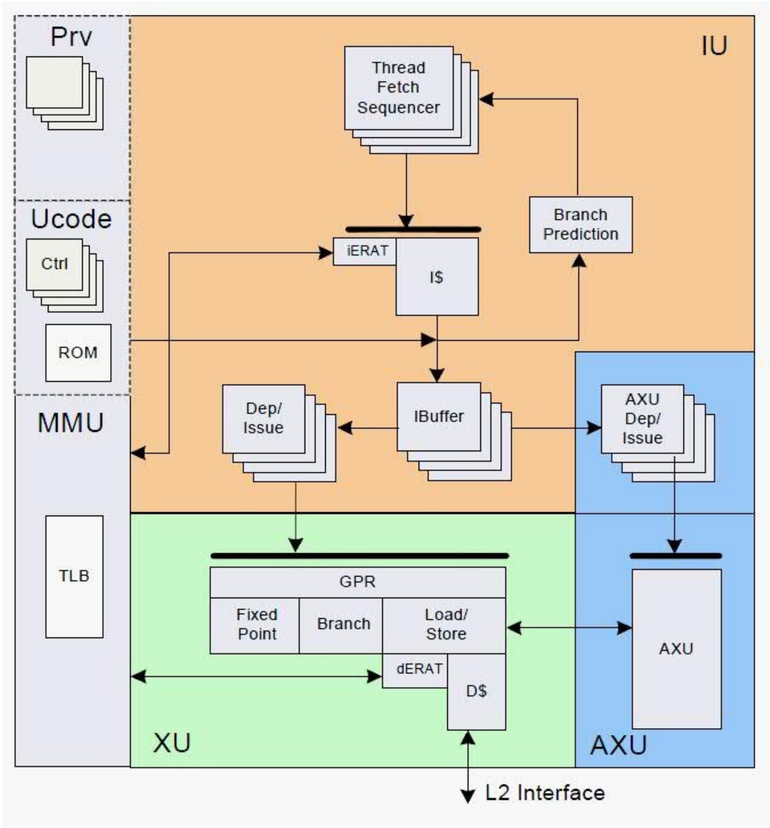
Memory hierarchies

- Different types of memory technologies with different capacity and performance properties
- Memory with
  - Large capacity but low performance
  - Small capacity but high performance

Blue Gene/Q Architecture

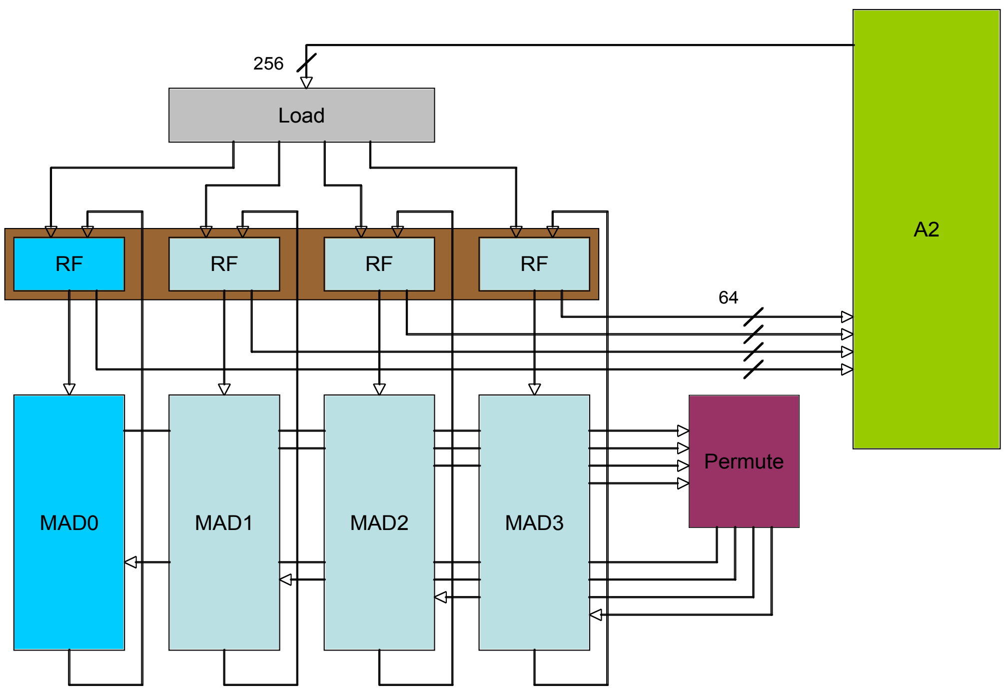
Processor core

- Embedded processor core
- 64-bit Power ISA
- 4-way Simultaneously Multi-Threaded (SMT)
- In-order dispatch, execution, completion
- Execution unit (XU)
  - 32x4x64-bit general purpose registers
  - Dynamic branch prediction
- Auxiliary execution unit (AXU)
  - BG/Q: Quad-FPU



Quad Floating-Point Processing Unit

- ISA extension: Quad Processing eXtension (QPX)
- 4 64-bit pipelines, SIMD processing
  - Vector operands
- 32x4x256 bit registers
- Multiply-Add Dataflow
  - Can process in each cycle, e.g., ± [(A * B) ± C]
- Peak performance 4 FMA / cycle → 12.8 GFlops at 1.6 GHz

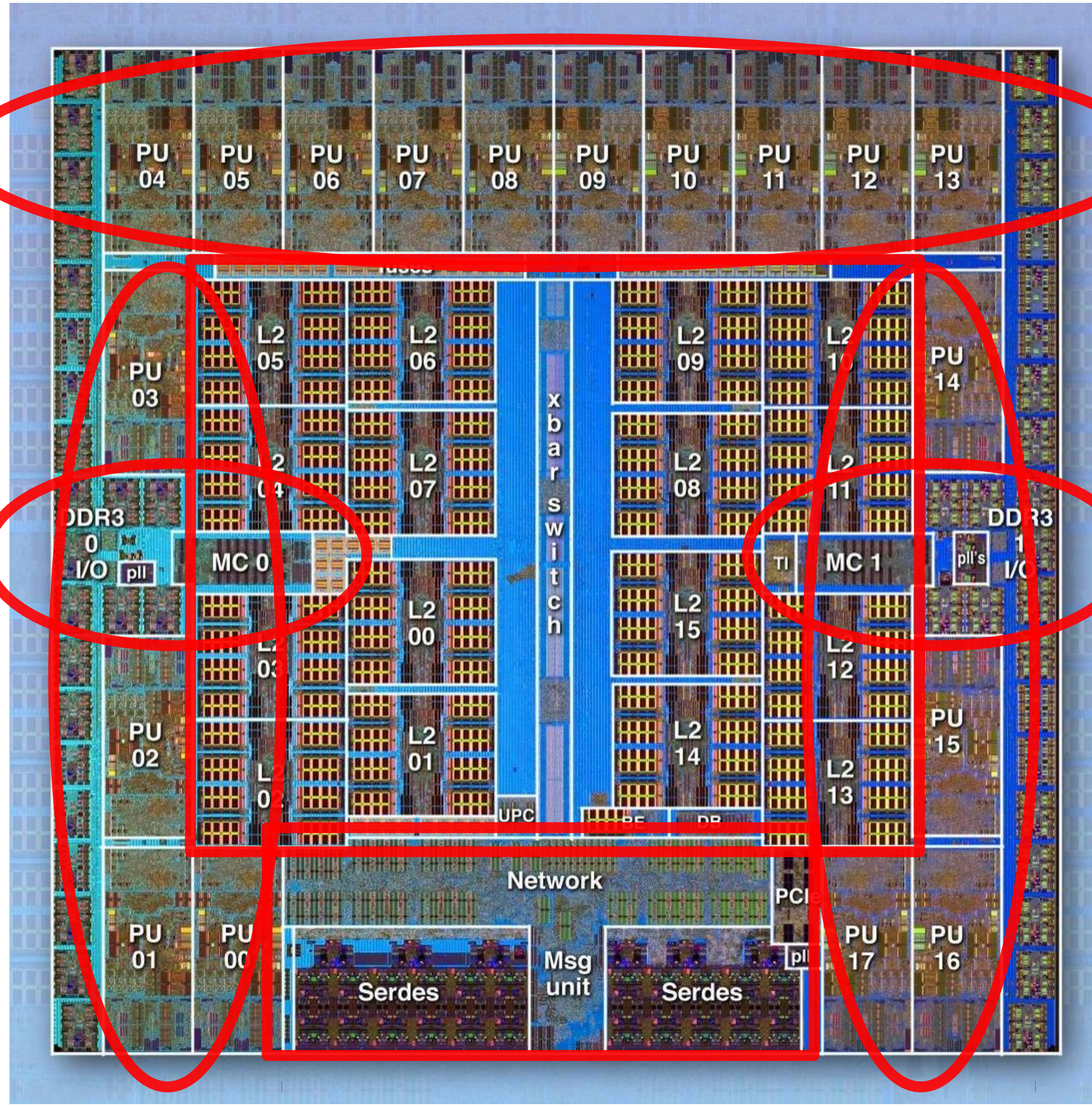


Multi-core processor

L2 cache +
crossbar switch

16+1+1 processing units

2 memory controller



Network

High-speed link technology

- 4 lanes, 4 Gbit/s effective data rate → 2 GByte/link/direction

5-dimensional torus

- High bi-section bandwidth
- Flexible partitioning in low dimensions

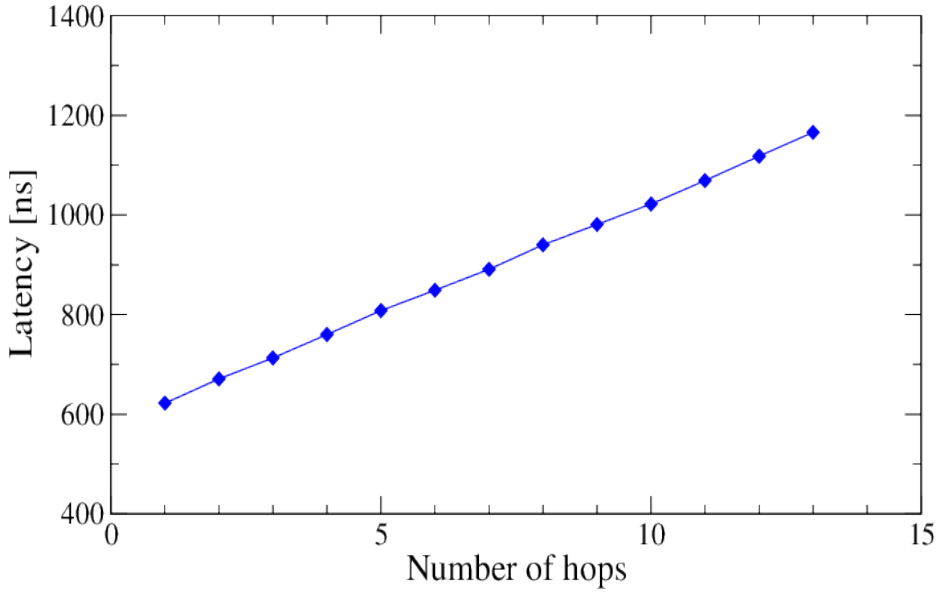
Deterministic and dynamic routing support

- Support of different routing algorithms

Very low latency

- Addition hop in network adds little latency

Ping-pong latency:



[Dong Chen et al., SC'11]

Network: Collectives support

Logical tree network mapped on physical torus

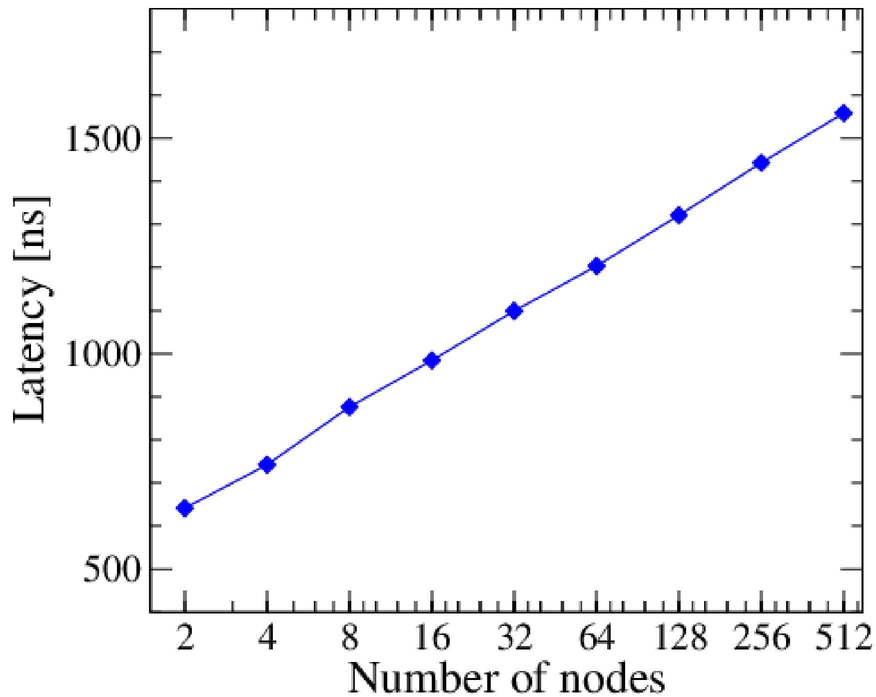
Reduction operations implemented in network unit

- Operations:
  - floating-point add/min/max
  - fixed-point operations
- Increases per hop point-to-point latency by 12 ns

Broadcast operations

- Can be used for fast synchronisation

Double-precision all-reduce:



[Dong Chen et al., SC'11]

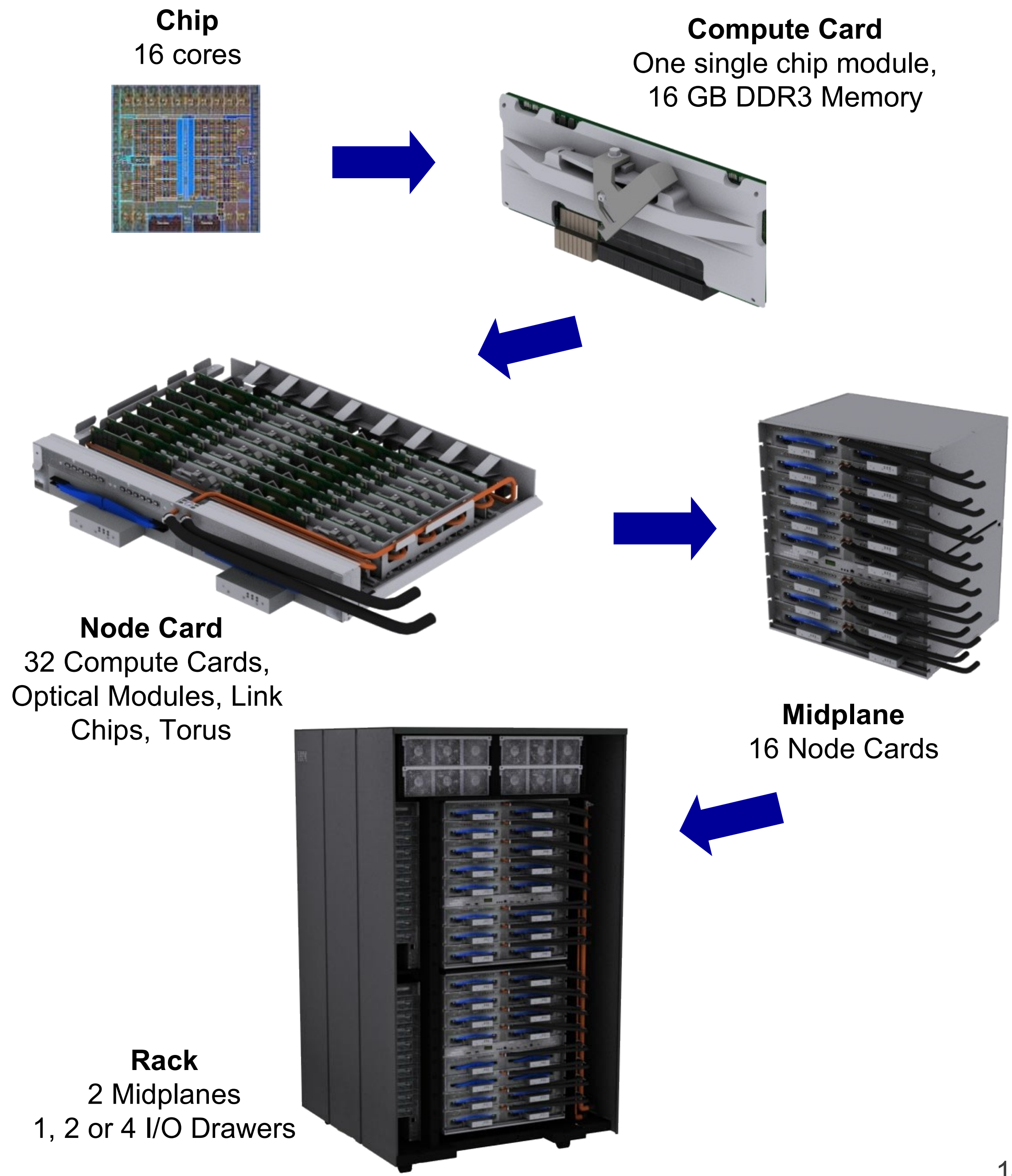
Integration

Rack parameters

- 1,024 nodes = 16,384 cores
- 210 TFlop/s
- Typically: 68 kW (JSC 2012 average)

Features

- Direct liquid cooling
- Network cabling:
  - Copper inside midplane
  - Otherwise optical



JUQUEEN and others

Blue Gene/Q installation at JSC

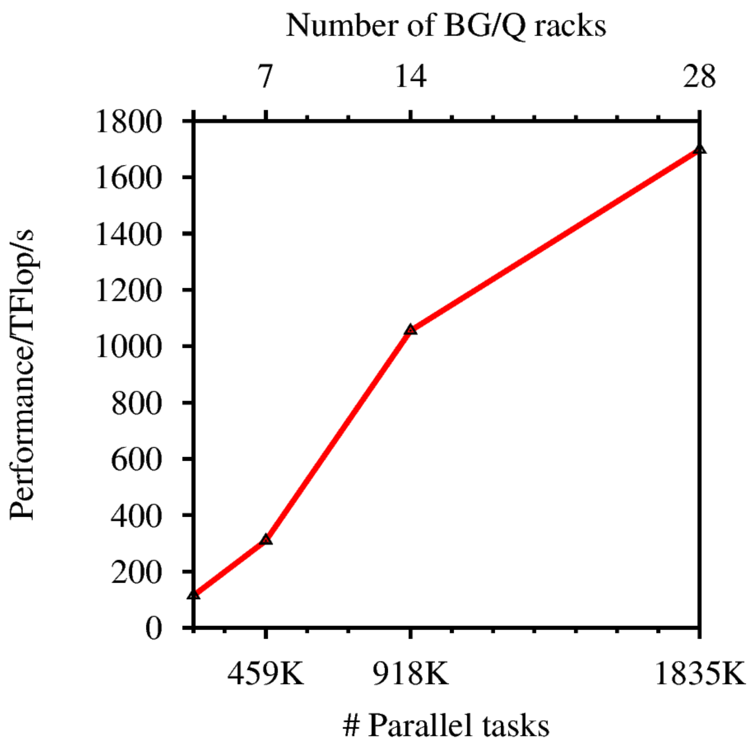
- 28 racks → 5.9 PFlop/s peak performance (DP)
- 448 TiByte main memory
- Installed in 2012

Other installations
(Top500 rank as of 11/2013)

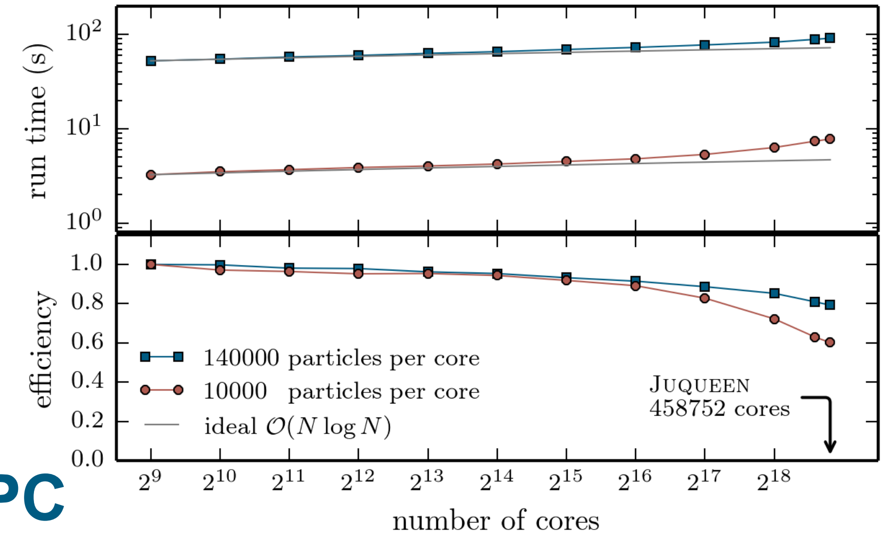
- Sequoia (LLNL, #3)
- Mira (ANL, #5)
- Vulcan (LLNL, #9)
- Fermi (CINECA, #15)
- Blue Joule (STFC, #23)
- Dirac (EPCC, #27)
- ...

Scaling examples: High-Q Club

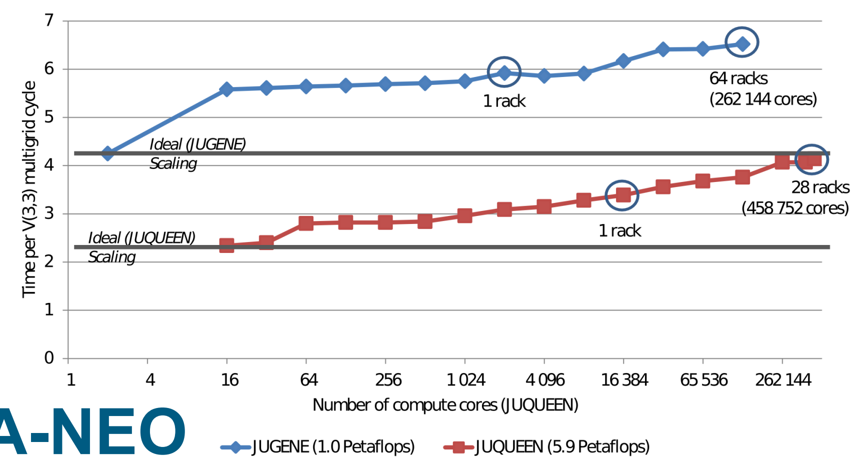
http://www.fz-juelich.de/ias/jsc/EN/Expertise/High-Q-Club/_node.html



dynQCD



PEPC



TERRA-NEO

Blue Gene/Q Features

SIMD instructions

SIMD = Singe Instruction Multiple Data

- Single instruction is applied to multiple data streams
- E.g., 4-way multiply-add:

$$\begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \end{pmatrix} \leftarrow \begin{pmatrix} a_0 + b_0 \cdot c_0 \\ a_1 + b_1 \cdot c_1 \\ a_2 + b_2 \cdot c_2 \\ a_3 + b_3 \cdot c_3 \end{pmatrix}$$

Vector data type: `VECTOR(REAL(8))`, `vector4double`

Selected instrinsics

- Load/store instructions: `vec_ld(offset, double*)`
- Multiply-add: `d = vec_madd(a, b, c)`

Simultaneous Multi-Threading (SMT)

Hardware support for up to 4 threads

- Need >1 thread to fully fill instruction pipelines

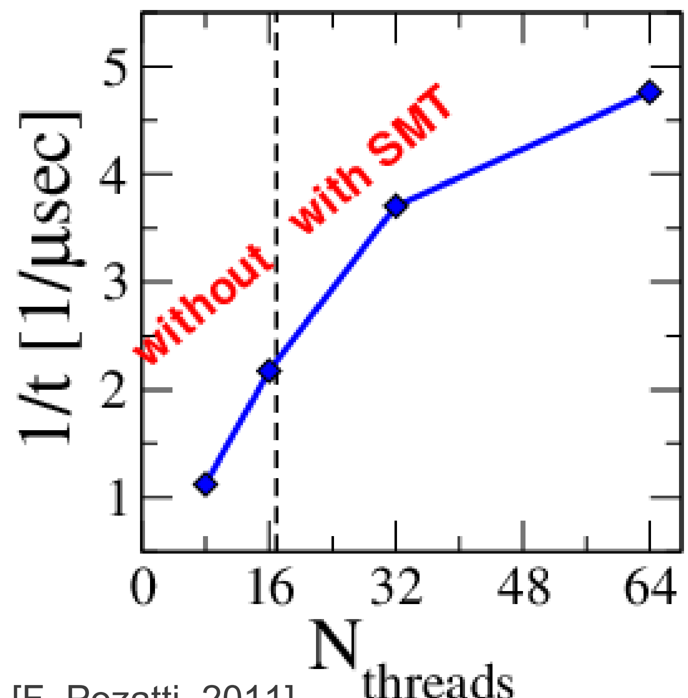
Supported programming models

- OpenMP
- POSIX threads

Hybrid parallelization

- n_p processes per node
- n_t threads per process
- Keep $n_p * n_t \leq 64$

LBM D2Q37, n_p=1



[F. Pozatti, 2011]

Other features

Memory pre-fetchers

- Different pre-fetchers available
  - Stream pre-fetcher
  - List-based pre-fetcher

L2 atomics

- Useful, e.g., if large number of threads update shared data structures

Transactional Memory and Speculative Execution

- Typical computational science applications are too regular to benefit

Other Architectures

Hardware architecture components

Processors

- Only multi-core processors
- >80% of Top500 systems are based on Intel Xeon

Accelerators

- GPU or Xeon Phi

Memory technologies

- DDR3 for processors
- GDDR5 for accelerators

Network

- Link technologies: Ethernet, Infiniband or proprietary
- Topology: fat tree, torus, dragonfly

Comparison of leading Top500 systems

	Sequoia	K Computer	Piz Daint
System architecture	Blue Gene/Q	K Computer	XC30
Vendor	IBM	Fujitsu	CRAY
Top500 (11/2013)	#3	#4	#6
Processor type	Blue Gene/Q A2	Sparc64 VIIIfx	Xeon E5-2650
N_{core}	1 572 864	663 552	42 176
Accelerator type	–	–	K20 GPU
N_{acc}	–	–	5 272
Network topology	5d torus	3d toroidal	dragonfly
Link bandwidth B_{link} [GByte/s]	2	5	4.7-5.25
Floating-point peak B_{fp} [PFlop/s]	20.1	10.6	7.8
Memory capacity C_{mem} [PiByte]	1.5	1.3	0.2
Power [MWatt]	7.9	12.7	2.3

JÜLICH FORSCHUNGSZENTRUM

Summary

High-end HPC architectures continue to become more parallel

- Parallelism at different levels

Blue Gene/Q architecture

- Highly parallel node architecture with 128 Flops/cycle
- Low-latency, high-bandwidth network
  → Crucial for scalability

Performance relevant hardware features

- 4-way SIMD instructions
- Simultaneous Multi-Threading
  → Need for hybrid parallelization