Rechnerarchitektur

Sommersemester 2013

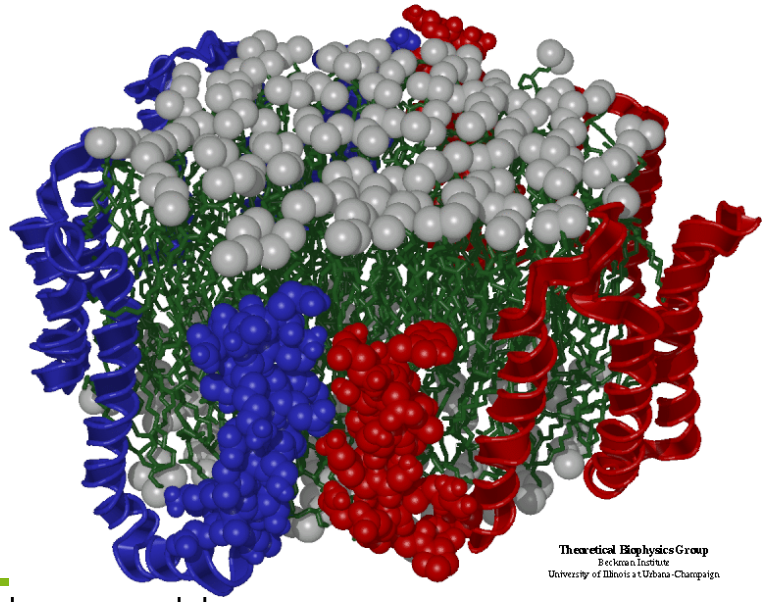
Beispielarchitekturen: IBM BlueGene

Michael Engel
Informatik 12
TU Dortmund

2013/07/11

IBM BlueGene

- Familie von Hochleistungs-rechnern
- Massiv paralleles System
  - Mehrere 10000 Prozessoren
- Entwicklung ab 1999
  - Projektierte Entwicklungs-kosten: US$100.000.000
- Anwendungsgebiet
  - Untersuchung bio-molekularer Probleme
  - Beispiel: Proteinfaltung



IBM BlueGene

- Drei Generationen
  - BlueGene/L, BlueGene/P, BlueGene/Q
- Über einige Jahre hinweg #1 der TOP500 Supercomputer-Liste (Rechner am LLNL)
  - November 2004: BlueGene/P #1 der TOP500
  - 16 Racks zu je 1024 Prozessoren
  - Linpack-Leistung von 70,72 TFLOPS
  - Behielt mit Ausbau Platz #1 für 3,5 Jahre
- Endausbau
  - 104 Racks
  - Linpack-Leistung von 478 TFLOPS
- November 2006: 27 BlueGene/L in TOP500

BlueGene Entwurfsprinzipien

- Tradeoff
  - Prozessorgeschwindigkeit – Energieverbrauch
- Verwendung von PowerPC-Prozessoren
  - Niedrige Taktfrequenzen
  - Verwendung v. Cores aus eingebetteten Systemen
- Leistung pro Chip gering
  - Aber gutes Verhältnis Leistung/Energie
  - Ermöglicht Bau von Systemen mit großer Anzahl an Prozessoren
- Großer Speicherbereich
- Verwendung von Standardcompilern und Message Passing Libraries

BlueGene Entwurfsprinzipien (2)

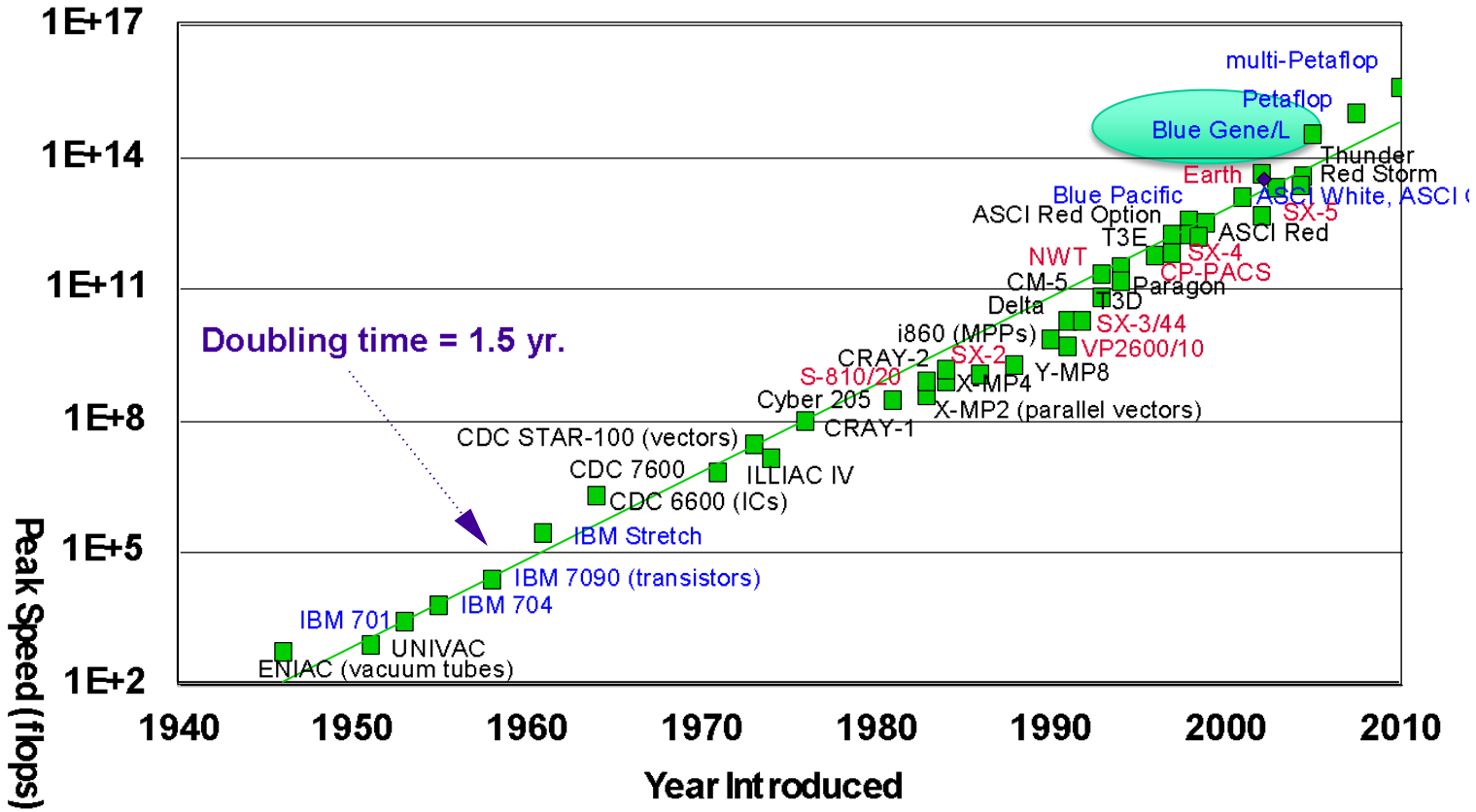
- Networks were chosen with extreme scaling in mind
  - Scale efficiently in both performance and packaging
  - Support very small messages
    - As small as 32 bytes
  - Includes hardware support for collective operations
    - Broadcast, reduction, scan, etc.
- Reliability, Availability and Serviceability critical issue
  - Reliable and usable even at extreme scaling limits
  - 20 fails per 1E9 hrs = 1 node failure every 4.5 weeks
- System Software also important to scaling
  - BG/L designed to efficiently utilize a distributed memory, message-passing programming model
  - MPI is the dominant message-passing model with hardware features added and parameter tuned

BlueGene RAS: Reliability, Availablity, Serviceability

- System designed for RAS from top to bottom
- System issues
  - Redundant bulk supplies, power converters, fans, DRAM bits, cable bits
  - Extensive data logging (voltage, temp, recoverable errors …) for failure forecasting
  - Nearly no single points of failure
- Chip design
  - ECC on all SRAMs
  - Dataflow outside CPUs protected by error detection
  - Access to all state via noninvasive back door
- Low power, simple design leads to higher reliability
- All interconnects have multiple EDAC coverage
  - Virtually zero escape probability for link errors

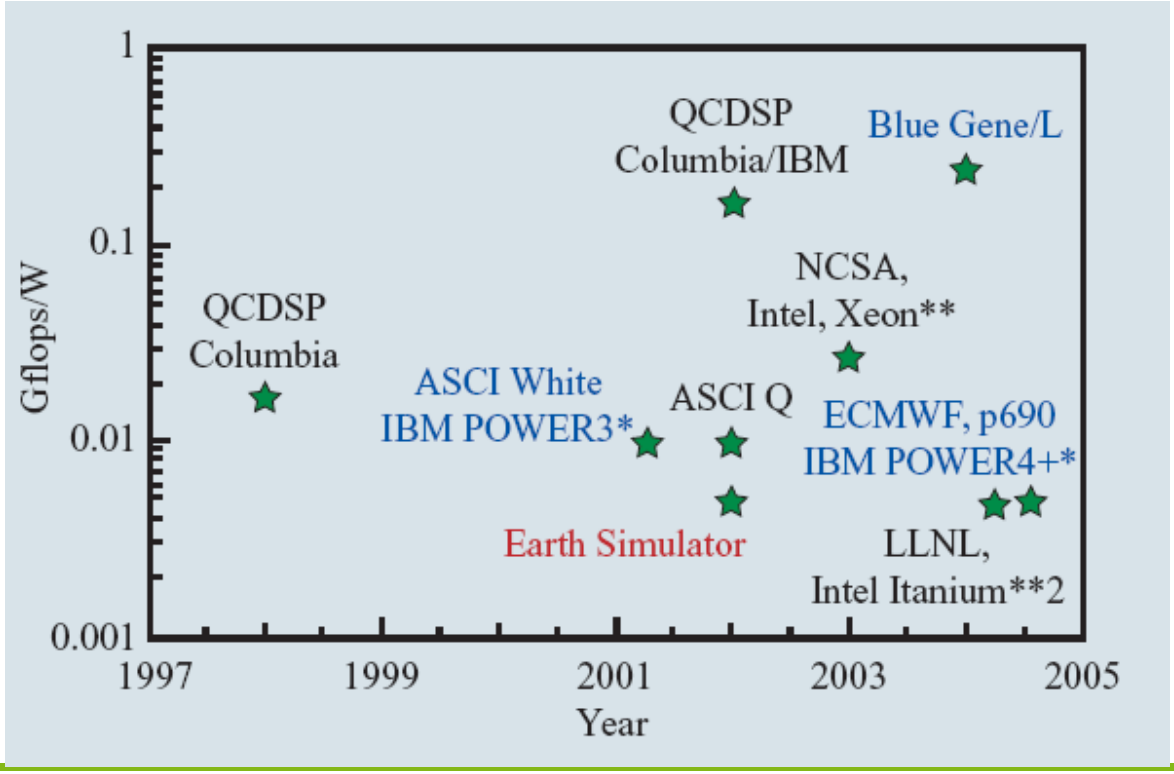
Entwicklung der Supercomputer-Rechenleistung

Supercomputer Peak Performance

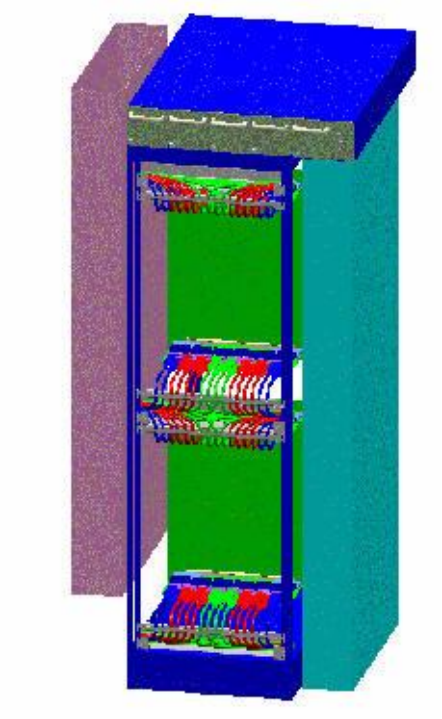


BlueGene: Energieeffizienz

- 360 TFLOPS benötigen mit konventionellen Prozessoren 20 MW Leitung
- Ziel: 2–10x bessere Energieffizienz

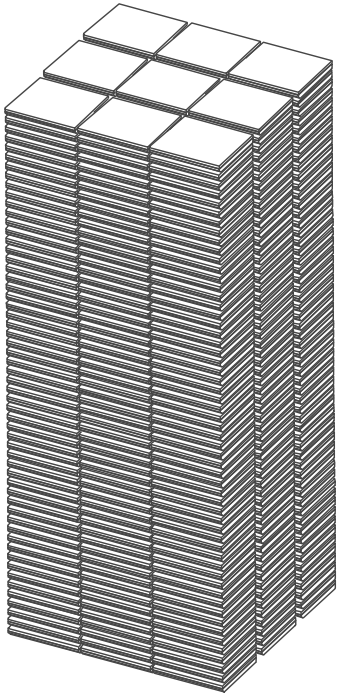


Energieeffizienz: Vergleich



BG/L
2048 processors

20.1 kW



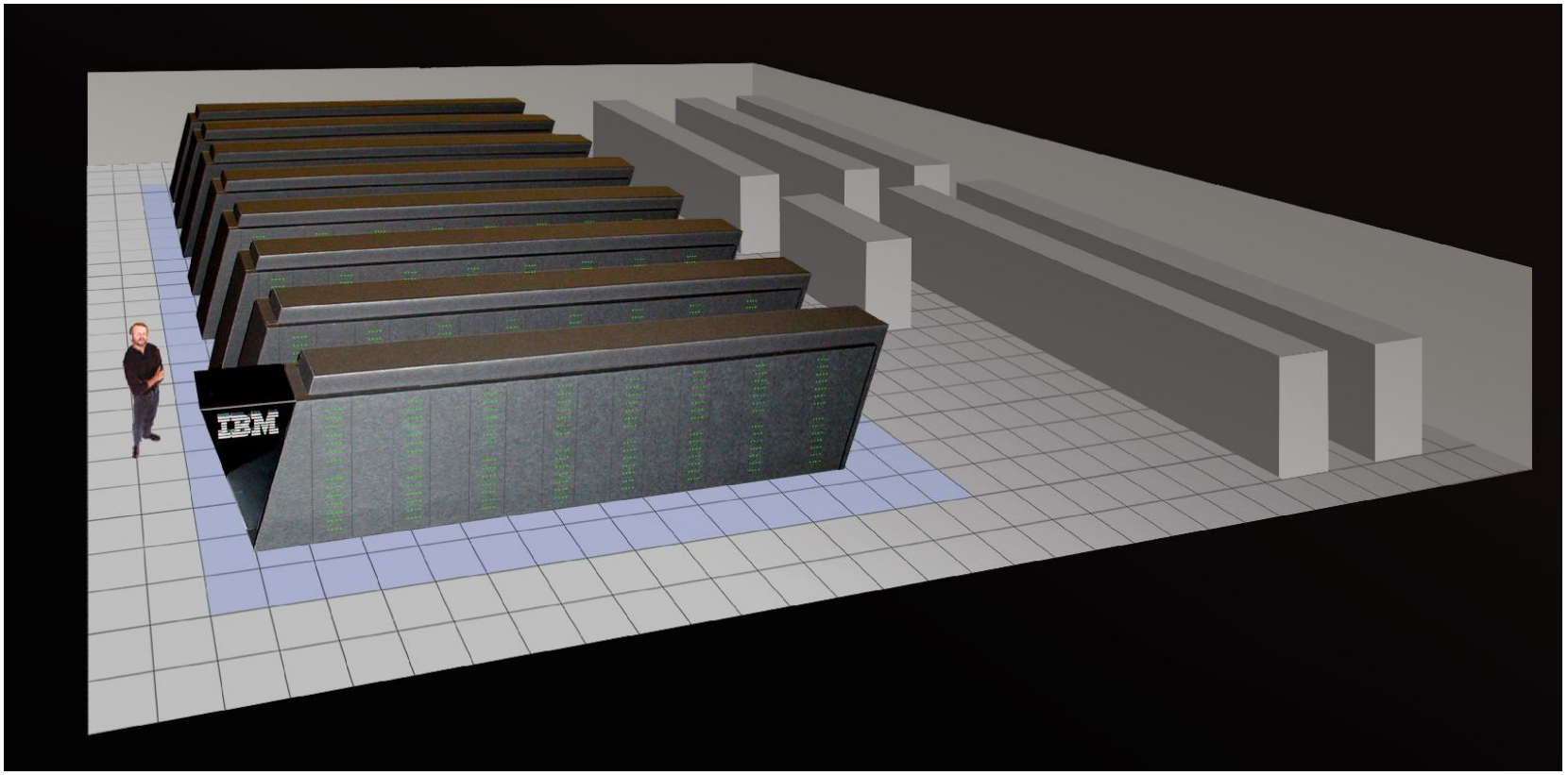
450 Thinkpads

20.3 kW

BlueGene/L: Rechenknoten

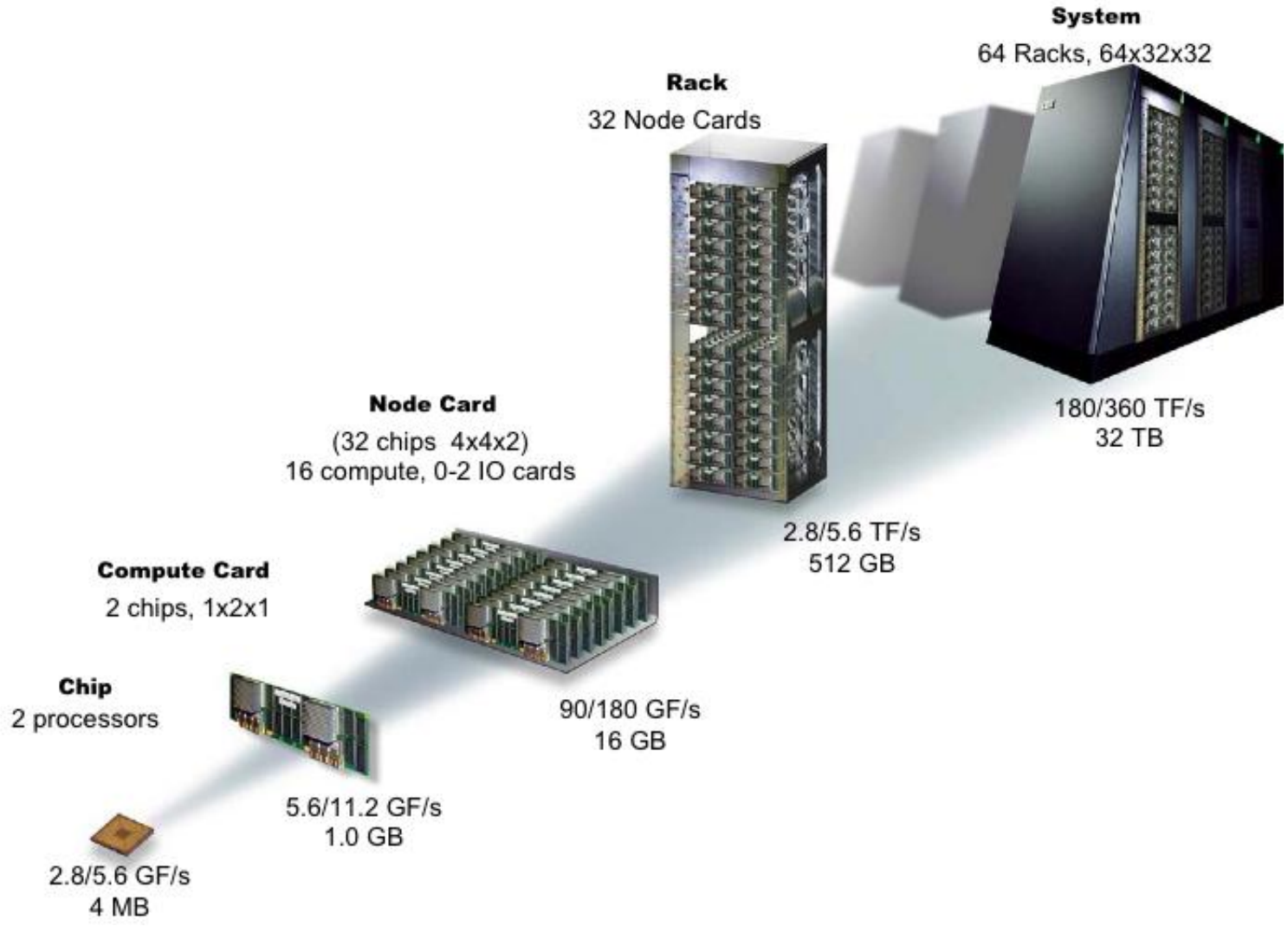
- Zwei Prozessoren pro Node, zwei Modi:
  - Eine CPU für Rechnen, eine für Kommunikation
  - Beide CPUs teilen sich die Aufgaben (“virtual node”-Modus)
  - Nicht Cache-kohärent zueinander
- System-on-Chip
  - Alle Komponenten auf einem Chip integriert
  - …bis auf 512 MB externes DRAM
- Große Anzahl von Knoten: in Schritten von 1.024 bis zu mindestens 65.536

BlueGene/L: Schema Gesamtsystem

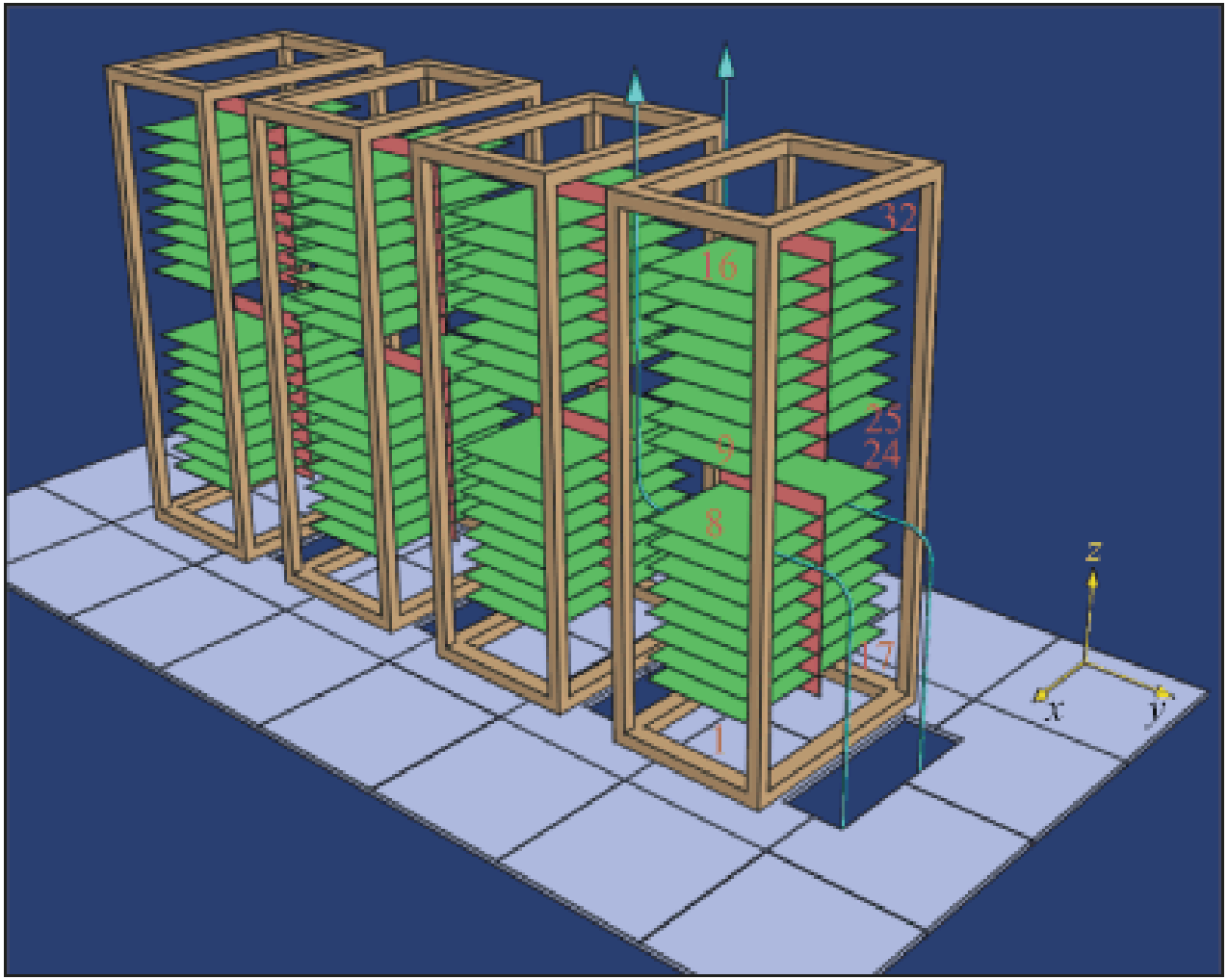


BlueGene/L

BlueGene/L: physikalischer Aufbau

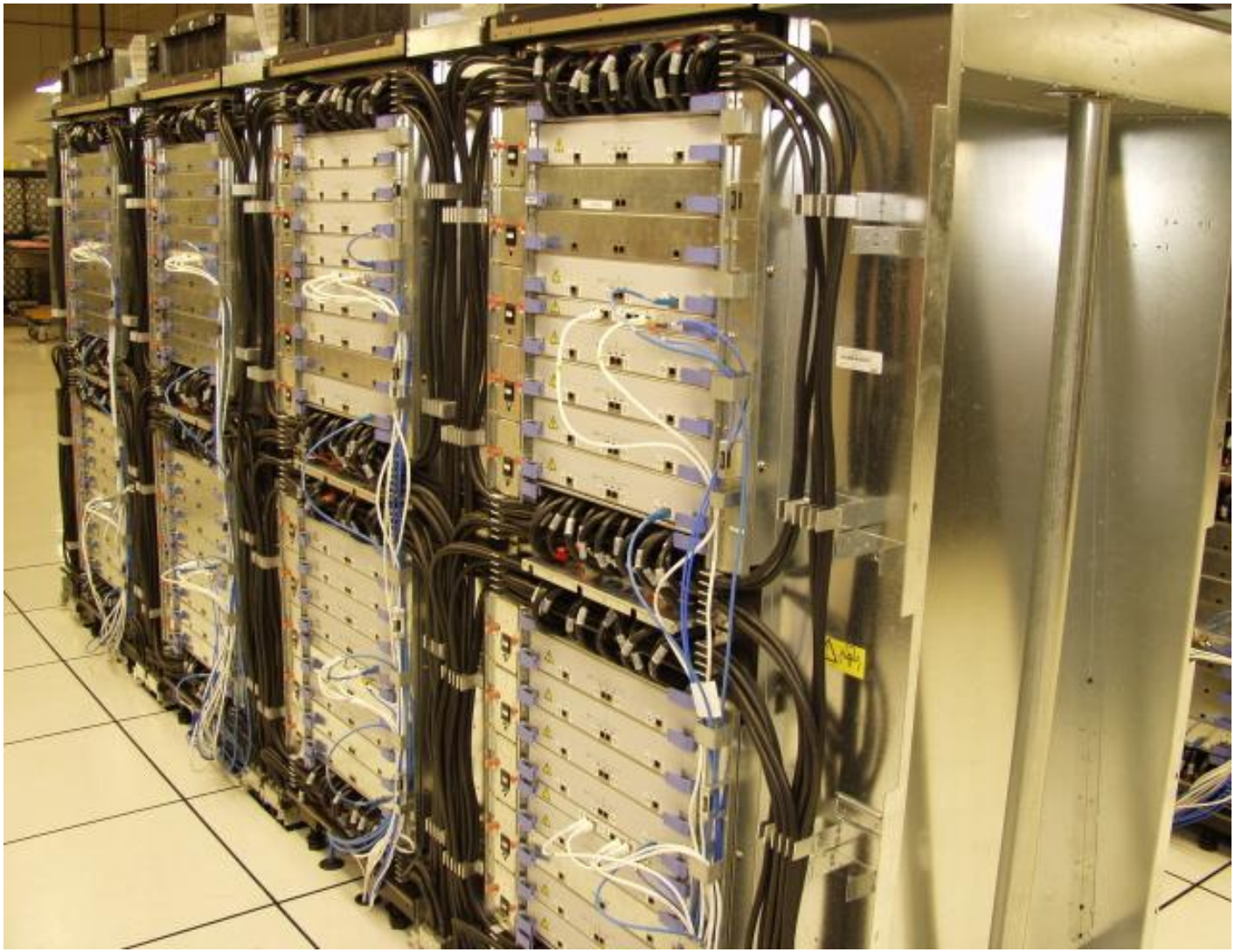


BlueGene/L: Midplanes und Racks



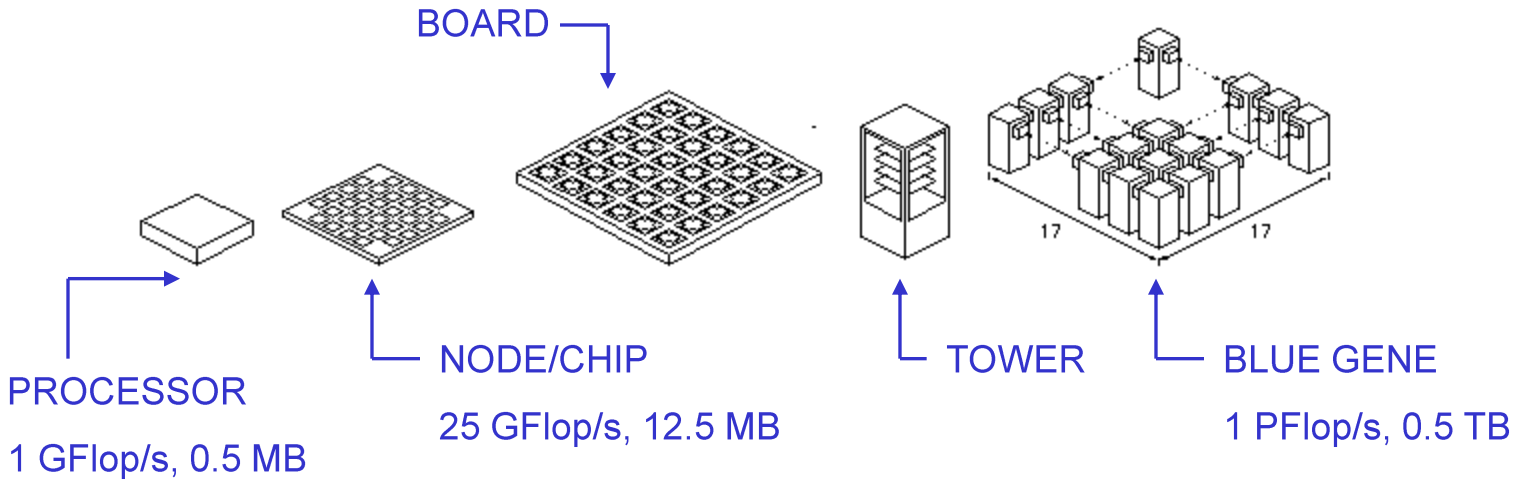
BlueGene/L

BlueGene/L



Systemarchitektur

- Skalierbarkeit (aus http://www.research.ibm.com/bluegene/)
  - "5 Schritte zum TeraFLOP"



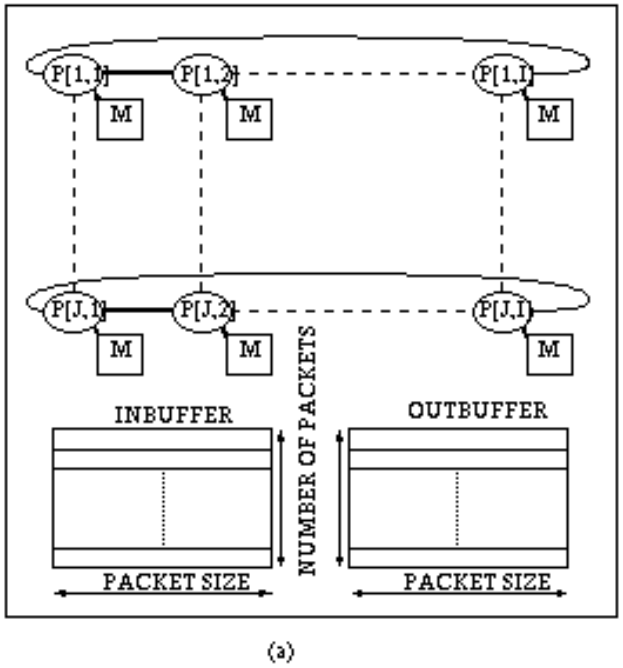
- Struktur
  - 34x34x36-"Würfel" auf shared memory nodes mit jeweils 25 Prozessoren

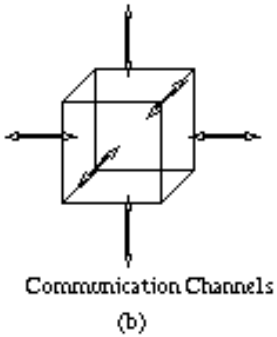
Rechenknoten

- 25 Prozessoren
- 200 Recheneinheiten

- Input/Output-Buffer
  - 32x128 Byte

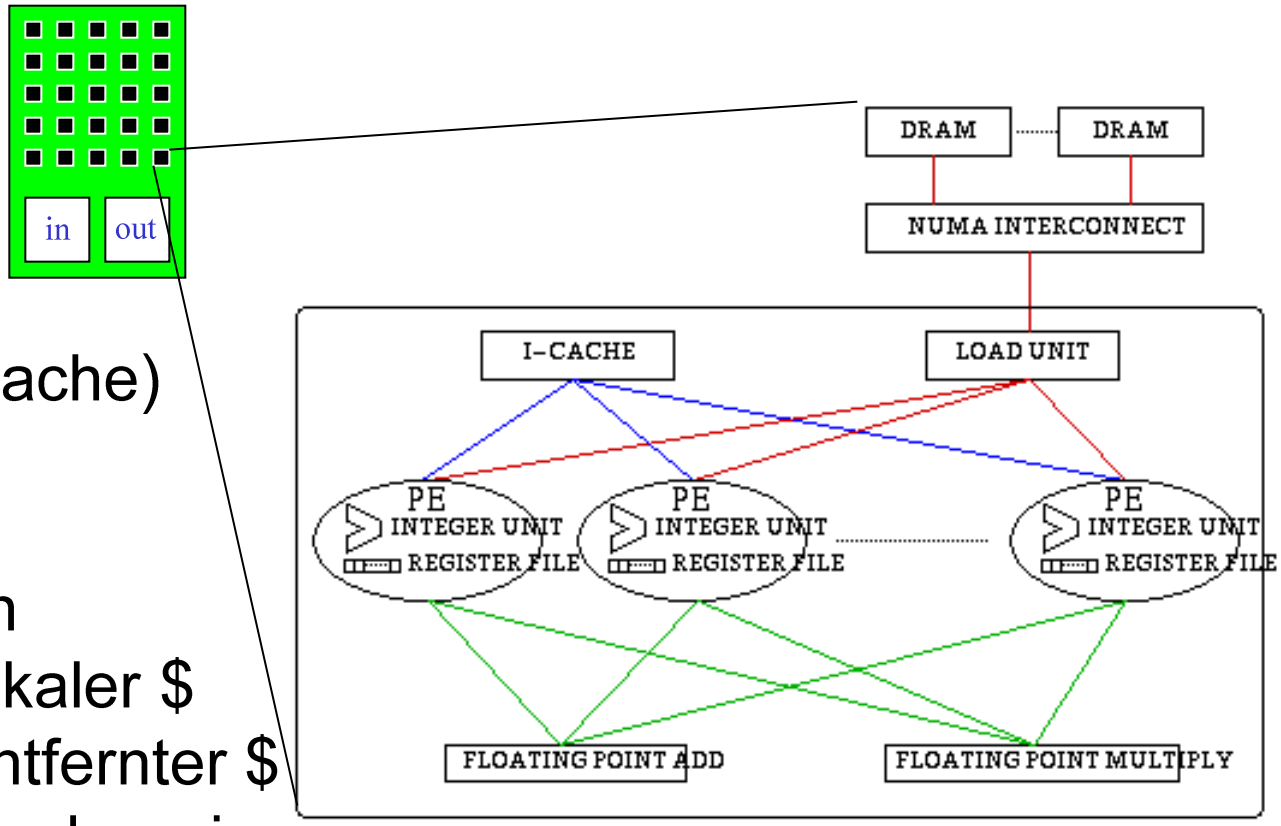
- Netzwerk
  - Duplex-Verbindung zu je 6 Nachbarn
  - 16 Bit @ 500 MHz = 1 GB/s
- Latenzen
  - 5 Zyklen/hop, 75 Zyklen pro Durchlauf





Prozessoren

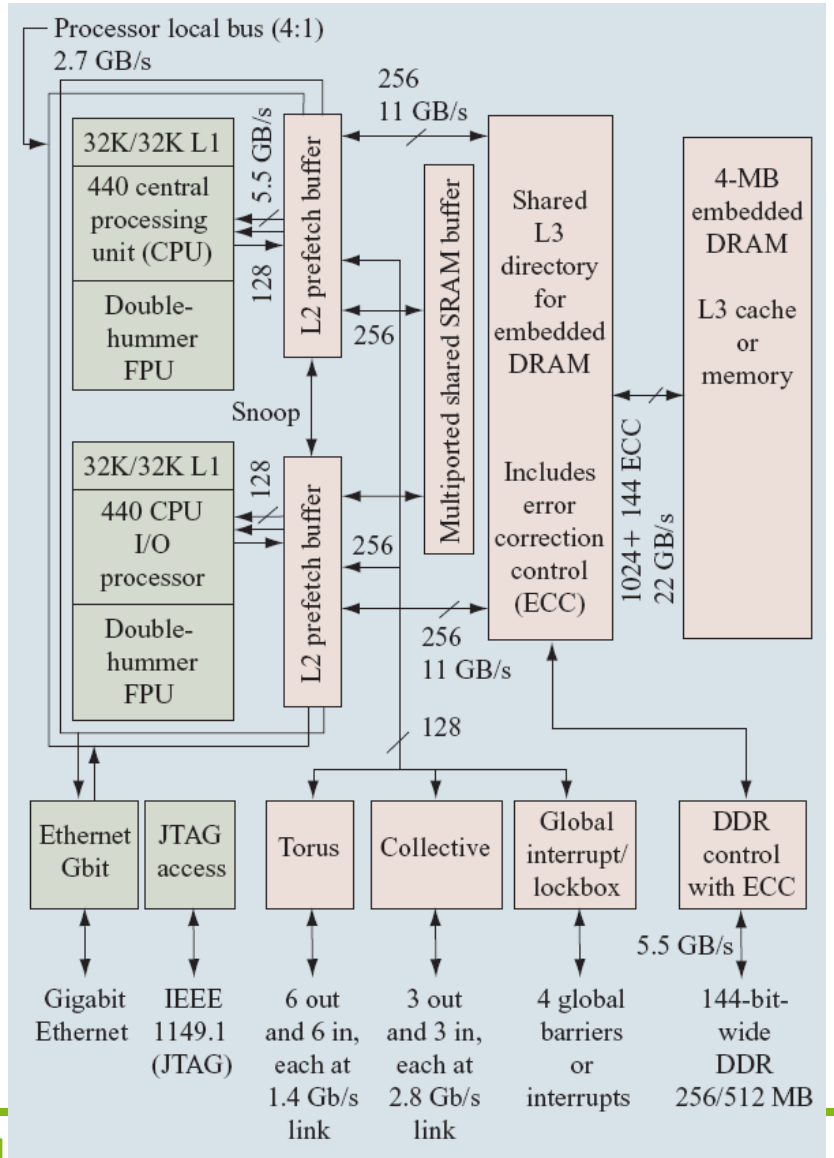
- 700 MHz
- Cache an Speicher (memory side cache) löst Kohärenz-probleme
- Zugriffslatenzen
  - 10 Zyklen lokaler $
  - 20 Zyklen entfernter $
  - 10 Zyklen Cache miss
- 8 Integer-Einheiten teilen sich 2 FP-Einheiten

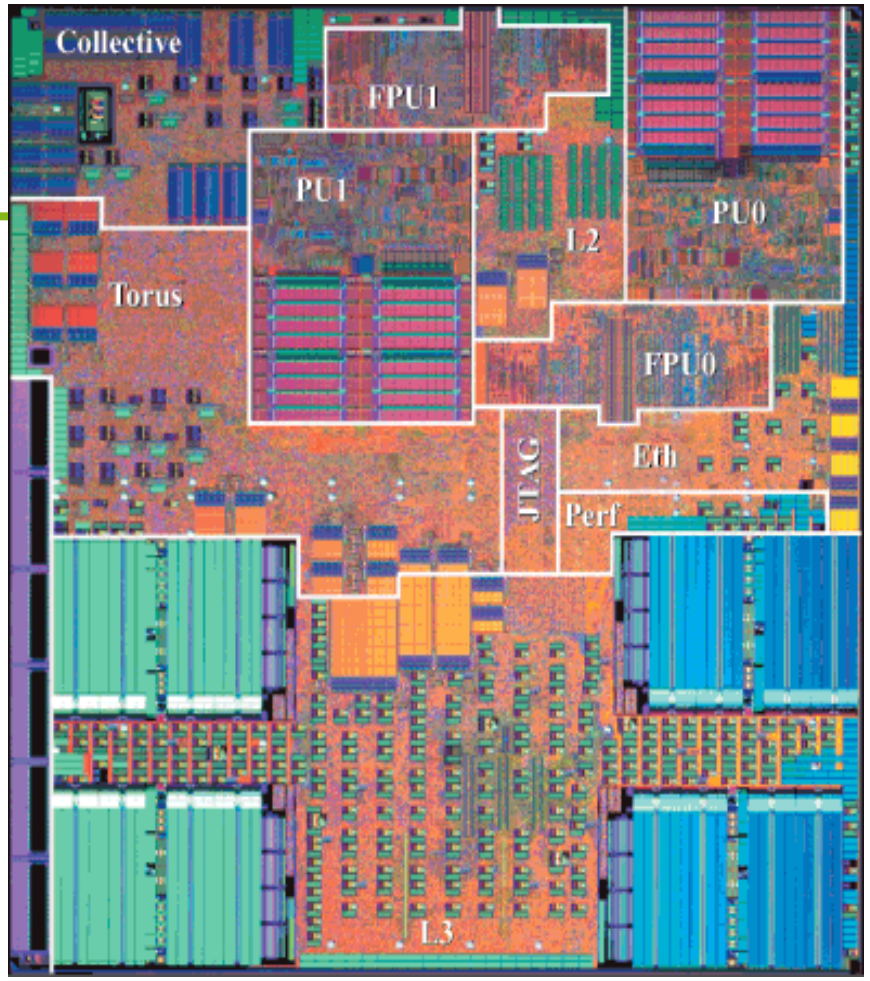


Prozessoren

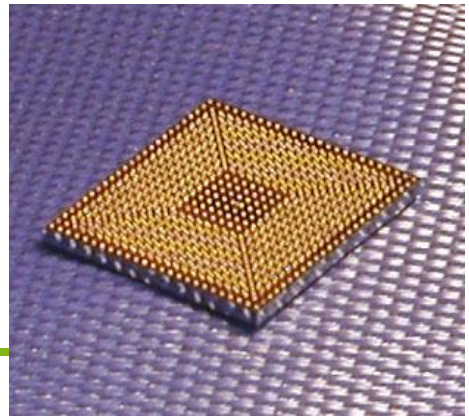
- PowerPC 440
  - Mit speziellen erweiterten FPUs
- 700 MHz bei 1W Leistungsaufnahme durch komplexe Energiesparmechanismen
- Typischer superskalarer Prozessor
  - Pipelined Mikroarchitektur
  - dual instruction fetch, decode
  - out of order issue, out of order dispatch, out of order execution and out of order completion

Compute ASIC



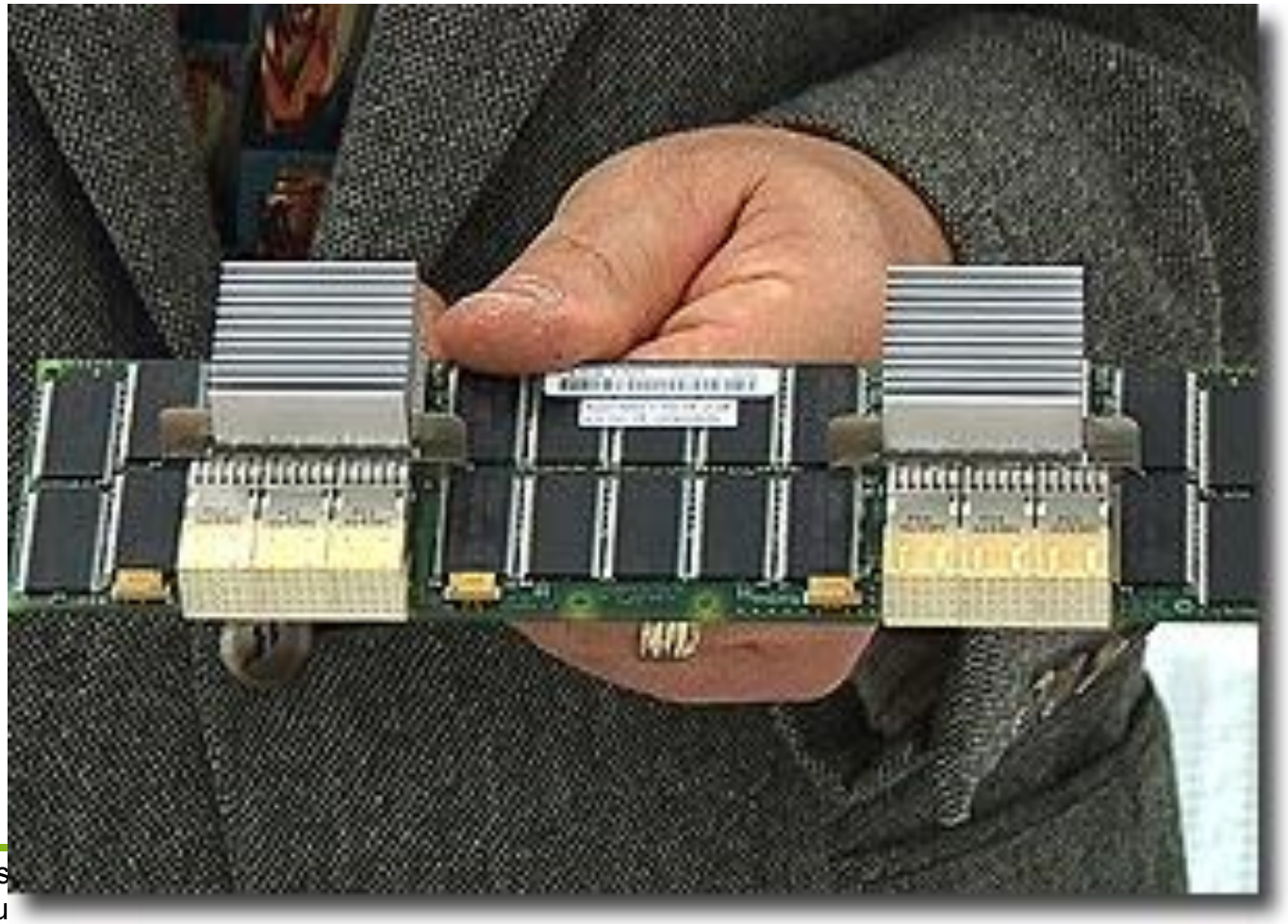


- **IBM CU-11, 0.13 µm**
- **11 x 11 mm die size**
- **25 x 32 mm CBGA**
- **474 pins, 328 signal**
- **1.5/2.5 Volt**



Compute Card

- 2 Compute ASICs plus RAM



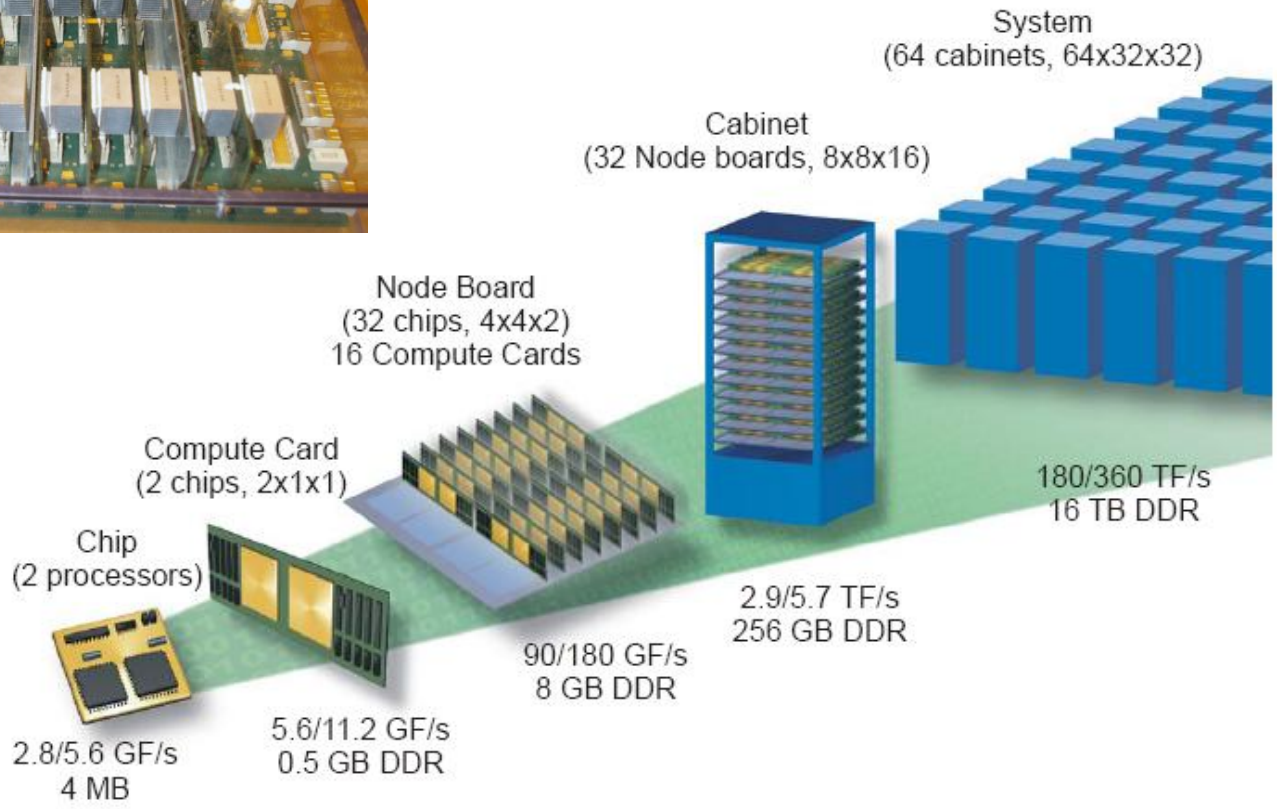
Node Card

- “Mainboard” für 16 Compute Cards

Speichersystem

- BG/L only supports distributed memory paradigm.
- No need for efficient support for cache coherence on each node.
  - Coherence enforced by software if needed.
- Two cores operate in two modes:
- Communication coprocessor mode
  - Need coherence, managed in system level libraries
- Virtual node mode
  - Memory is physical partitioned (not shared).

BlueGene/L: Leistungsdaten

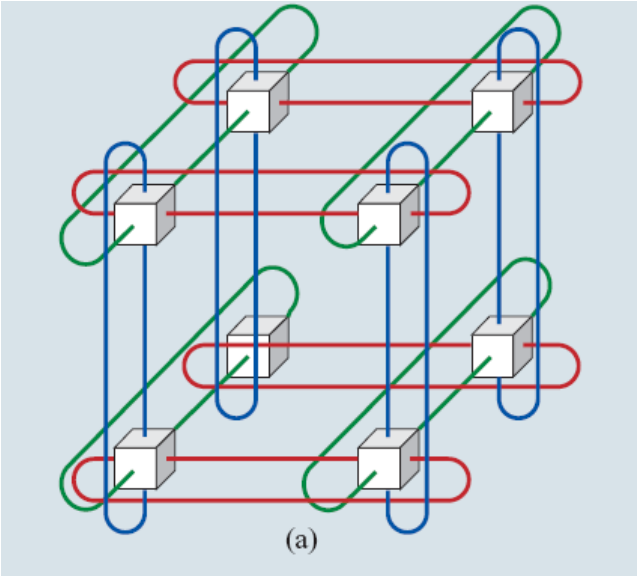


Netzwerk

- Five networks
  - 100 Mbps Ethernet control network for diagnostics, debugging, and some other things
  - 1000 Mbps Ethernet for I/O
  - Three high-band width, low-latency networks for data transmission and synchronization.
  - 3-D torus network for point-to-point communication
  - Collective network for global operations
  - Barrier network
- All network logic is integrated in the BG/L node ASIC
  - Memory mapped interfaces from user space

3D-Torus-Netzwerk

- Support p2p communication
- Link bandwidth 1.4Gb/s
- 6 bidirectional link per node (1.2GB/s).
- 64x32x32 torus: diameter 32+16+16=64 hops, worst case hardware latency 6.4us.
- Cut-through routing
- Adaptive routing



3D-Torus-Netzwerk (2)

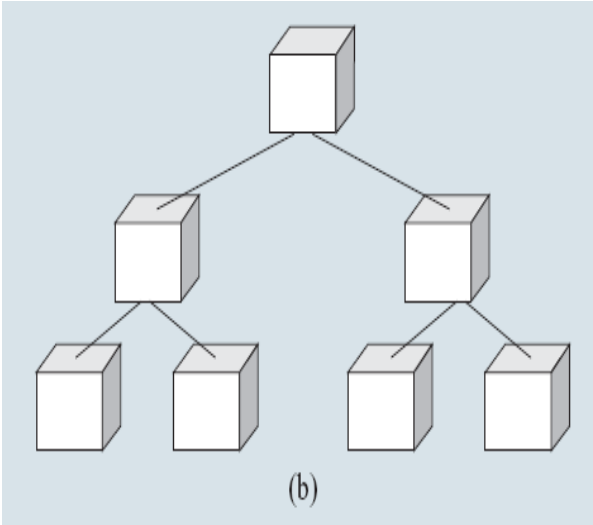
- Main network, for point-to-point communication
- High-speed, high-bandwidth
- Interconnects all compute nodes (65,536)
- Virtual cut-through hardware routing
- 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- 1 µs latency between nearest neighbors, 5 µs to the farthest
- 4 µs latency for one hop with MPI, 10 µs to the farthest
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 68TB/s total bandwidth

3D-Torus-Netzwerk (3)

- 3 dimensional: 64 x 32 x 32
  - Each compute node connected to its six neighbors: x+, x-, y+, y-, z+, z-
  - Compute card is 1x2x1
  - Node card is 4x4x2
  - 16 compute cards in 4x2x2 arrangement
  - Midplane is 8x8x8
  - 16 node cards in 2x2x4 arrangement
- Communication path
  - Each node can send and receive at 1.05GB/s.
  - Supports cut-through routing, along with both deterministic and adaptive routing.
  - Variable-sized packets of 32,64,96…256 bytes
  - Guarantees reliable delivery

Collective Network

- Binary tree topology, static routing
- Link bandwidth: 2.8Gb/s
- Maximum hardware latency: 5us
- With arithmetic and logical hardware: can perform integer operation on the data
  - Efficient support for reduce, scan, global sum, and broadcast operations
  - Floating point operation can be done with 2 passes.



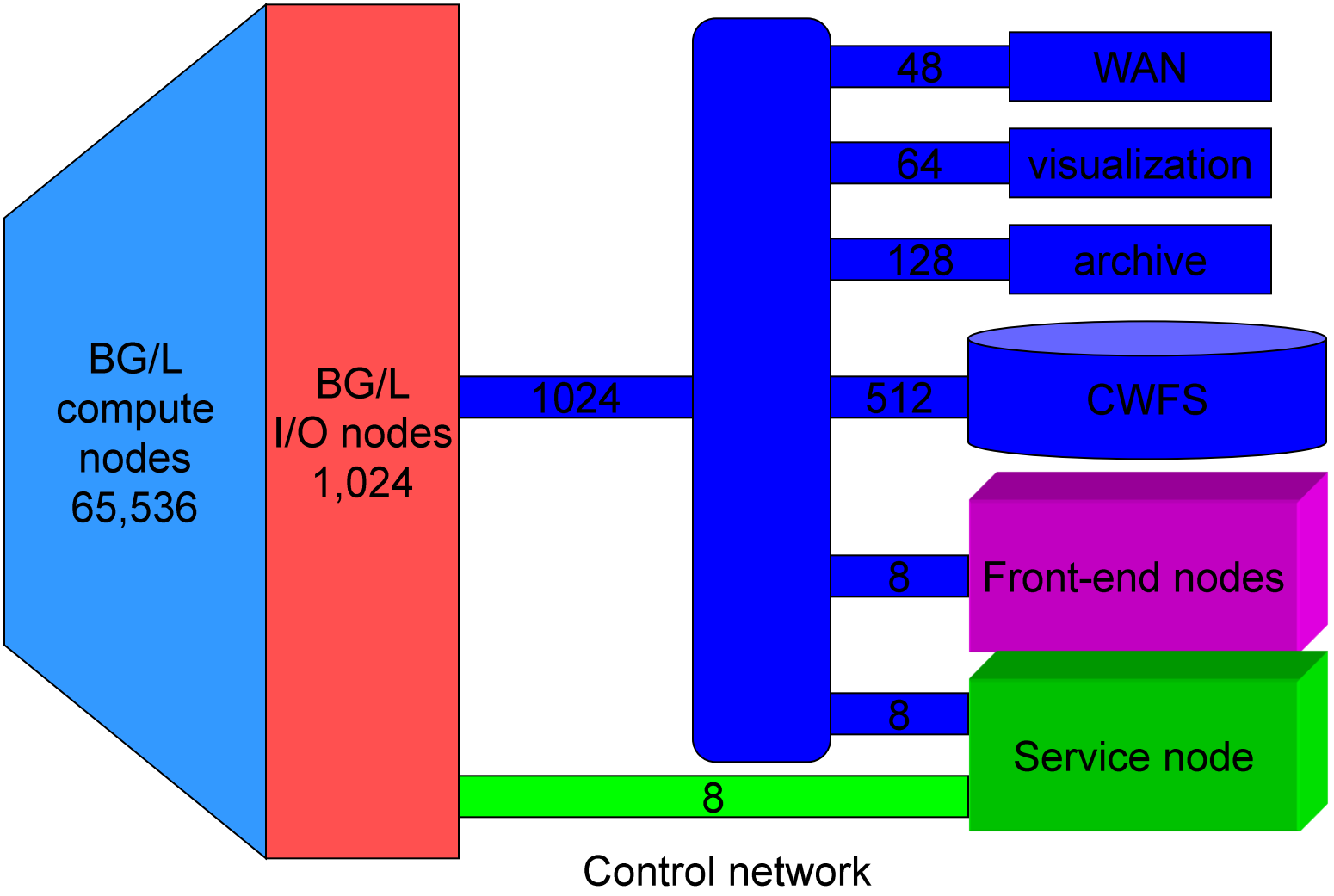
Collective Network (2)

- One-to-all broadcast functionality
- Reduction operations functionality
- MPI collective ops in hardware
- Fixed-size 256 byte packets
- 2.8 Gb/s of bandwidth per link
- Latency of one way tree traversal 2.5 µs
- ~23TB/s total binary tree bandwidth (64k machine)
- Interconnects all compute and I/O nodes (1024)
- Also guarantees reliable delivery

Barriere-Netzwerk

- Hardware support for global synchronization.
- 1,5 µs for barrier on 64K nodes.

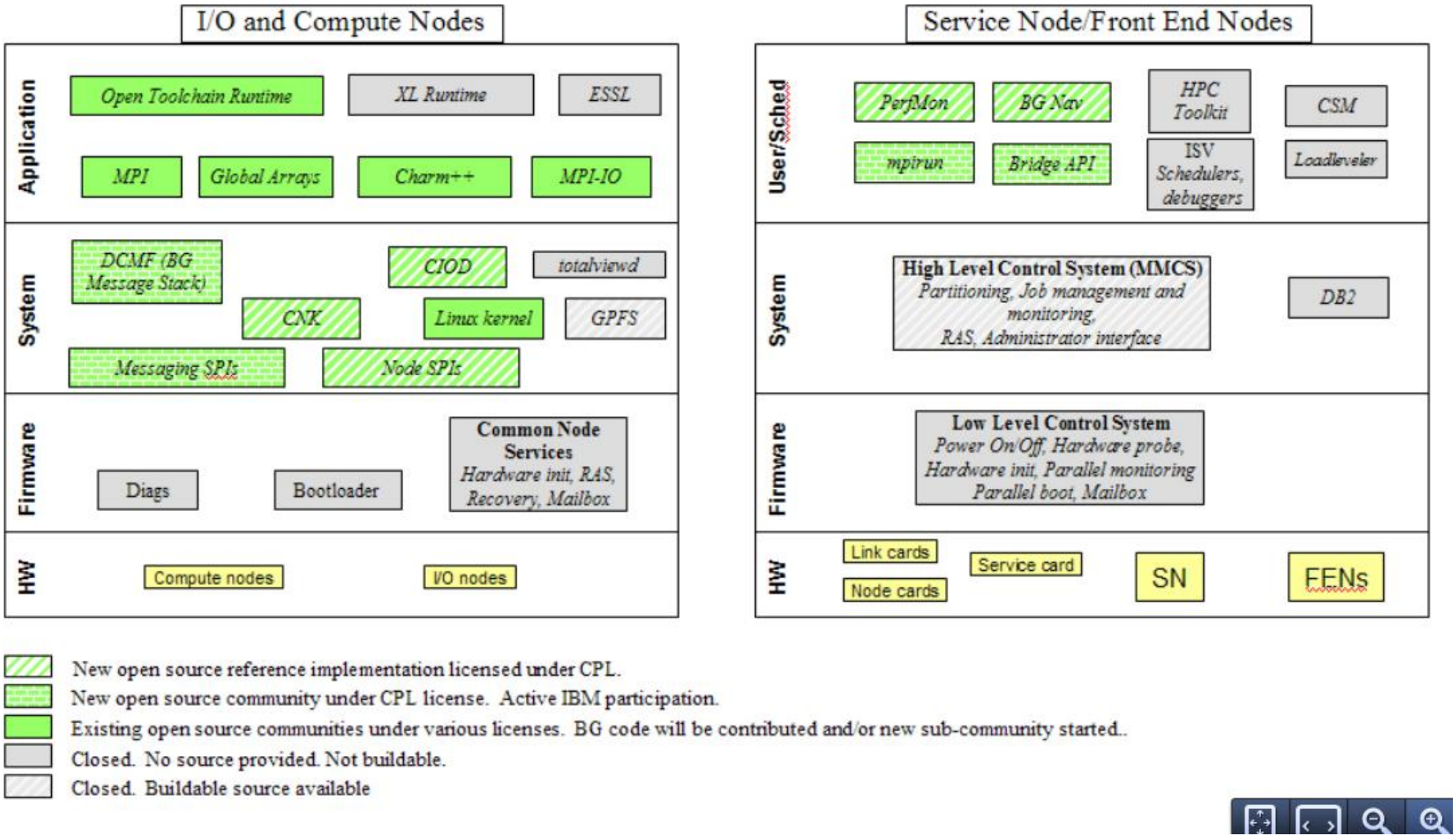
BG/L-System am LLNL



Systemsoftware

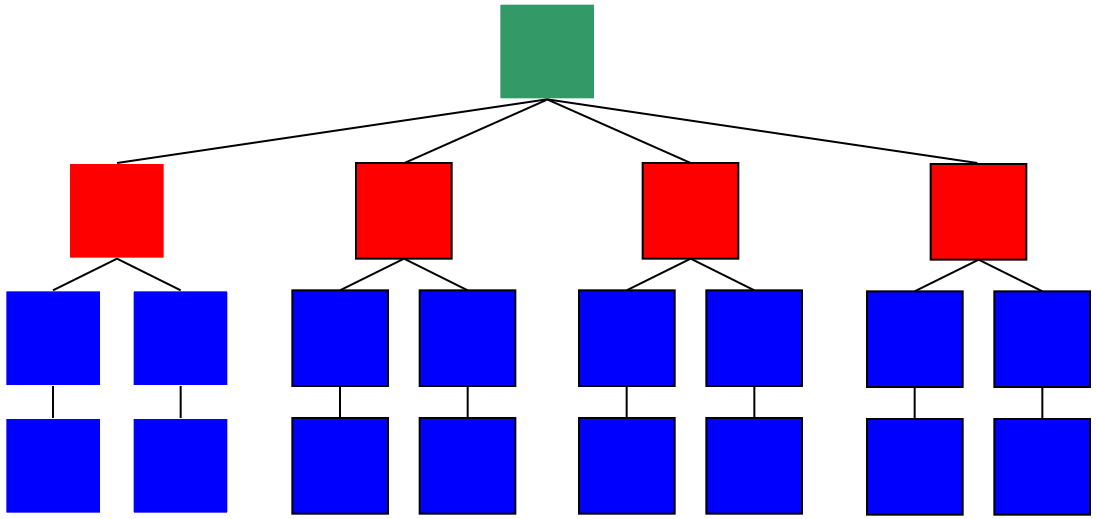
- Operating system - Linux
- Compilers - IBM XL C, C++, Fortran95
- Communication - MPI, TCP/IP
- Parallel File System - GPFS, NFS support
- System Management - extensions to CSM
- Job scheduling - based on LoadLeveler
- Math libraries - ESSL

Systemsoftware



Hierarchische Softwareorganisation

- Compute nodes dedicated to running user application, and almost nothing else – simple compute node kernel (CNK)
- I/O nodes run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination
- Service node performs system management services (e.g., heart beating, monitoring errors) – transparent to application software



Systemsoftware: Entwurfsprinzipien

- Simplicity
  - Space-sharing
  - Single-threaded
  - No demand paging
- Familiarity
  - MPI (MPICH2)
  - IBM XL Compilers for PowerPC
- Struktur
  - Front-end nodes are commodity systems running Linux
  - I/O nodes run a customized Linux kernel
  - Compute nodes use an extremely lightweight custom kernel
  - Service node is a single multiprocessor machine running a custom OS

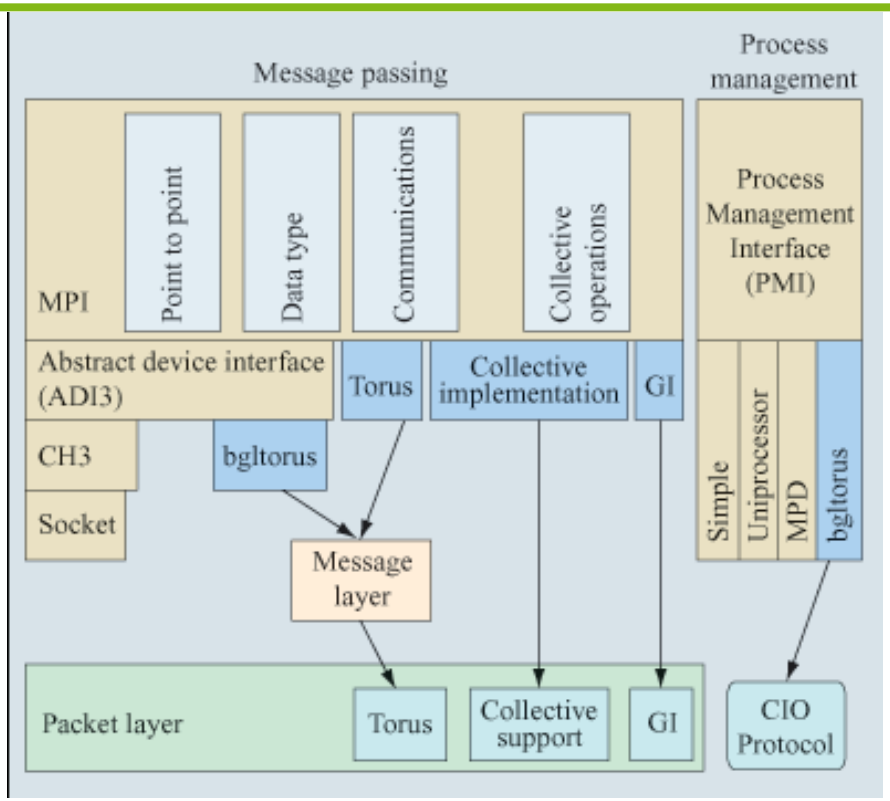
Compute Node Kernel (CNK)

- Single user, dual-threaded
- Flat address space, no paging
- Physical resources are memory-mapped
- Provides standard POSIX functionality (mostly)
- Two execution modes:
  - Virtual node mode
  - Coprocessor mode
- ca. 5000 Zeilen C++

Compute Node Kernel (2)

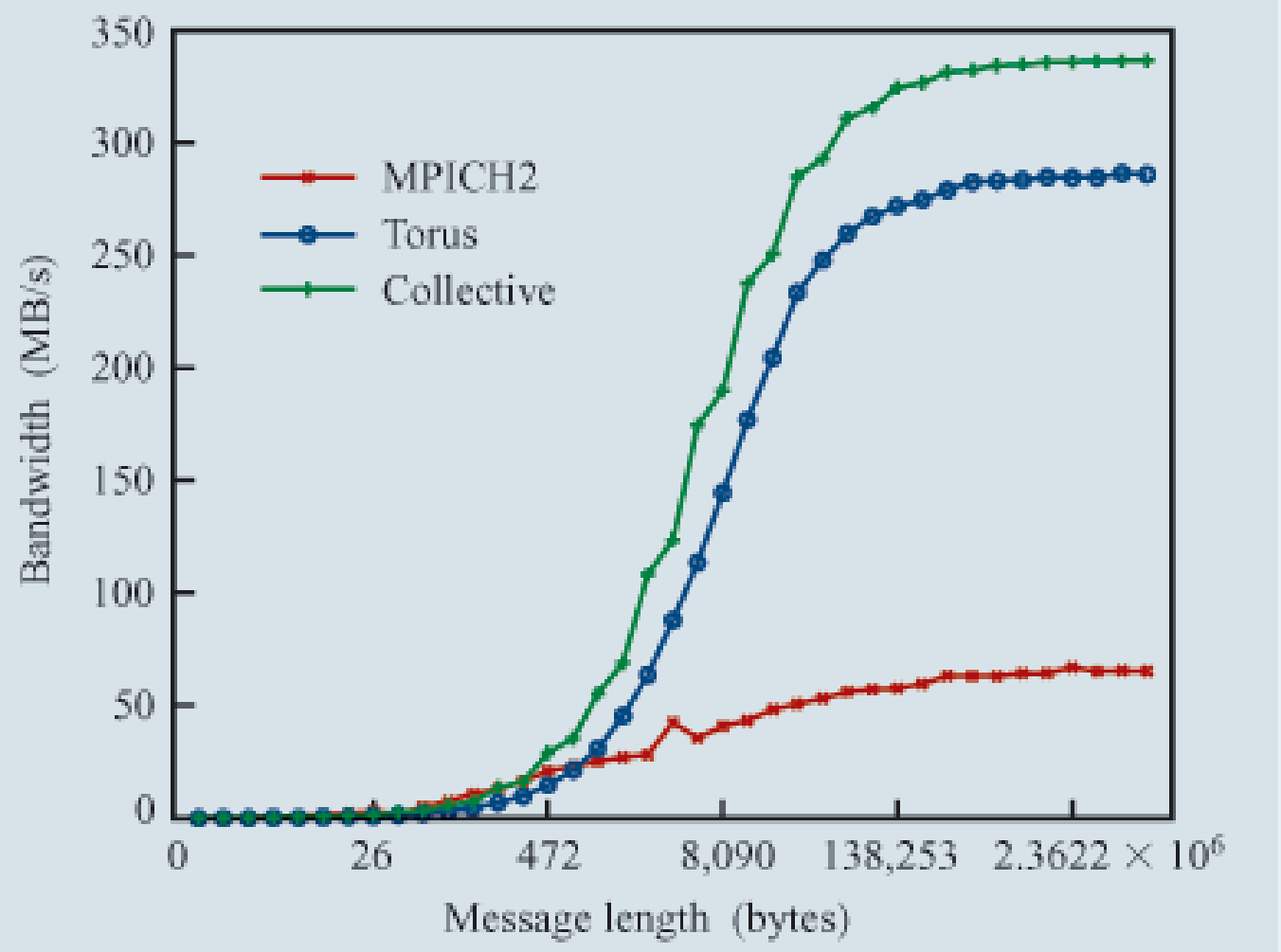
- Physical memory is statically mapped
- CNK neither needs nor provides scheduling or context switching
  - at each point it runs a single application for a single user
- By not allowing virtual memory or multi-tasking, the design of CNK aimed to devote as many cycles as possible to application processing
- CNK does not even implement file I/O on the compute node
  - delegates that to dedicated I/O nodes running Linux: INK (I/O node kernel)

MPI Software-Architektur

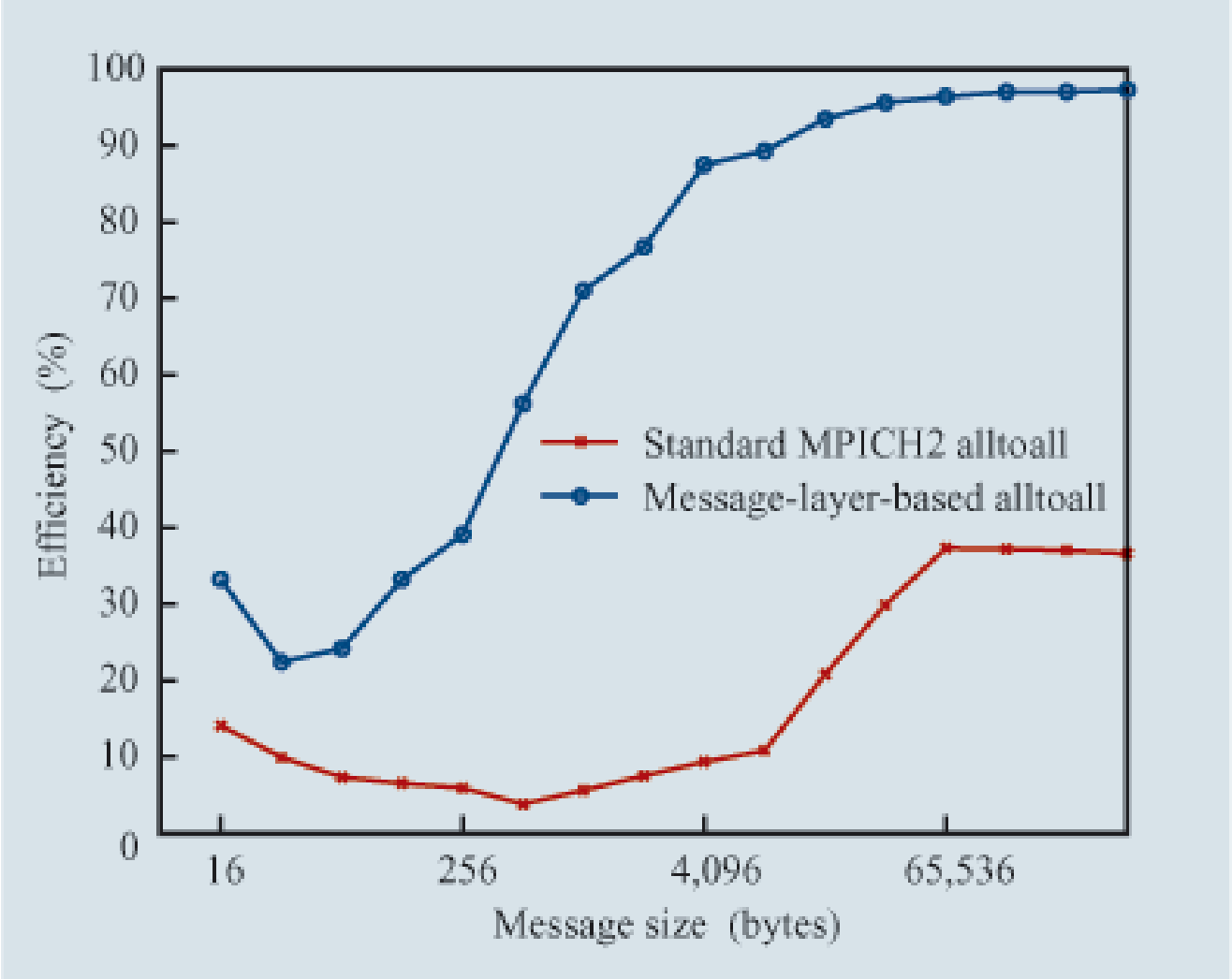


GI = Global Interrupt
CIO = Control and I/O Protocol
CH3 = Primary device distributed with MPICH2 communication
MPD = Multipurpose Daemon

Performance: MPI_Bcast



Performance: MPI_Alltoall



BlueGene/P

- Leistungssteigerung…
- Each Blue Gene/P Compute chip contains four PowerPC 450 processor cores, running at 850 MHz
- The cores are cache coherent and the chip can operate as a 4-way symmetric multiprocessor (SMP)
- The memory subsystem on the chip consists of small private L2 caches, a central shared 8 MB L3 cache, and dual DDR2 memory controllers.
- Compute card contains a Blue Gene/P chip with 2 or 4 GB DRAM, comprising a "compute node“
  - A single compute node has a peak performance of 13.6 GFLOPS
- Compute Node Linux statt CNK

BlueGene/Q

- 20 Petaflops in 2012
- The Blue Gene/Q Compute chip is an 18 core chip
- The 64-bit PowerPC A2 processor cores are 4-way simultaneously multithreaded, and run at 1.6 GHz
- Each processor core has a SIMD Quad-vector double precision floating point unit (IBM QPX)
- 19x19 mm² (359.5 mm²),1.47 x10^9 transistors
- 16 Processor cores are used for computing
- 17th core for operating system assist functions such as interrupts, asynchronous I/O, MPI pacing and RAS
- 18th core used as a redundant spare to increase yield
- Processor cores are linked by a crossbar switch to a 32 MB eDRAM L2 cache, operating at half core speed