



Blue Gene/L in Jülich

K. Wolkersdorfer@fz-juelich.de



JuBL (January 2006)

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



Air Flow

Low Power
Consumption

Small
Footprint

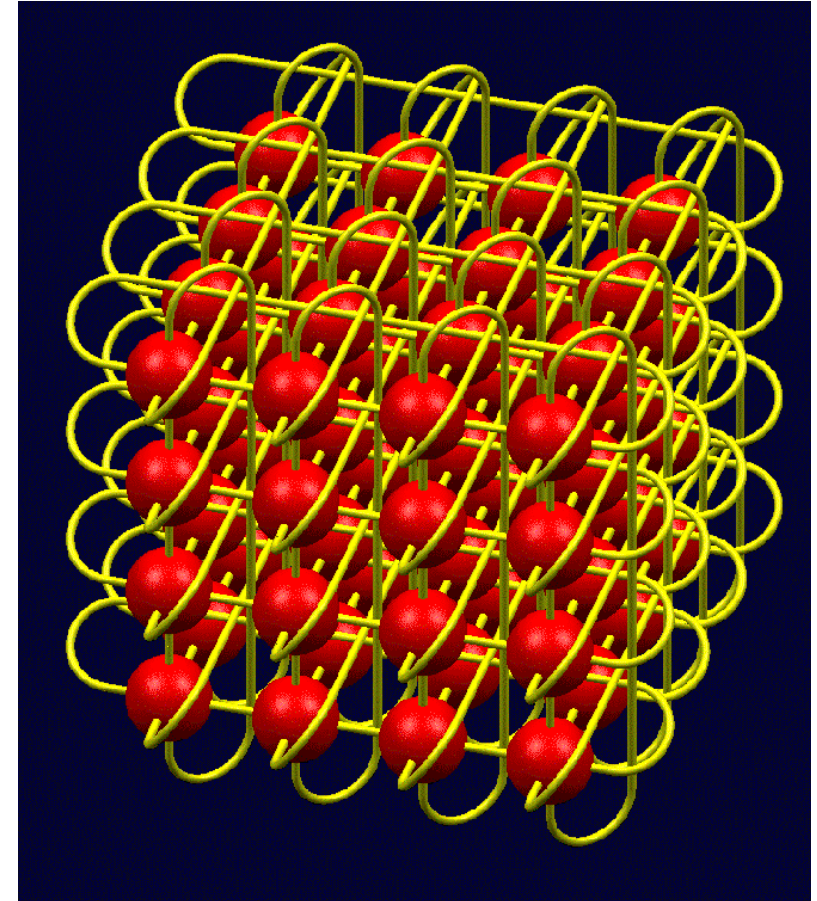
16384 Prozessoren
45.8 Teraflop/s
4.1 Terabytes



- **December 2004**
International Blue Gene Workshop in Jülich
 - **January 2005**
Procurement for a Blue Gene/L Testsystem (1 Frame)
 - **July - September 2005**
Assembly, Installation and Integration with JUMP cluster,
First Users achieved 20% of Peak Performance
-
- **November 2005**
Agreement with IBM to extend to 8 - Rack System
 - **December 2005**
HGF Covenant, Contract Conclusion, Delivery
 - **January 2006**
Assembly, Installation, Integration, LINPACK achieved **36,49 TFlops**
 - **February 2006**
Produktion and Inauguration



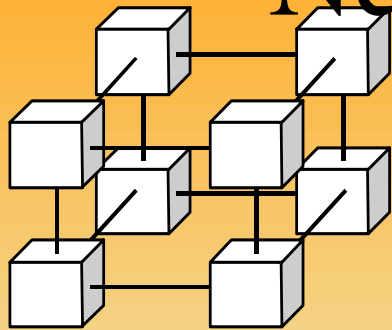
- **A large number of nodes (65,536)**
 - *Low-power nodes for density*
 - *High floating-point performance*
 - *System-on-a-chip technology*
- **Nodes interconnected as 64x32x32 three-dimensional torus**
 - *Full routing in hardware*
 - *Auxiliary networks for I/O and global operations*
- **Applications consist of multiple processes with MPI**
 - *Strictly one process/node*





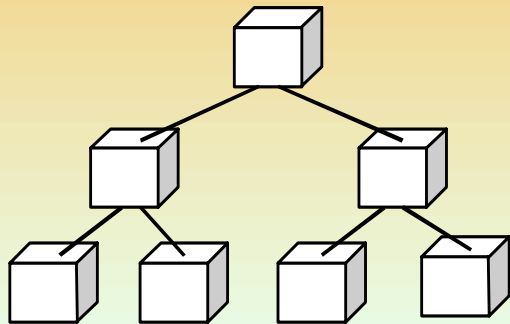
Networks

3-Dimensional Torus



- 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)
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 67TB/s total bandwidth

Global Collective Network (Tree)



- One-to-all broadcast functionality
- Reduction operations functionality
- 2.8 Gb/s of bandwidth per link; Latency of tree traversal 2.5 μ s
- ~23TB/s total binary tree bandwidth (64k machine)
- Interconnects all compute and I/O nodes (1024)

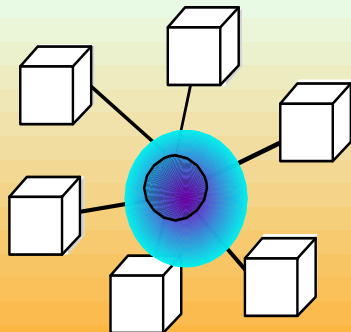
Low Latency Global Barrier and Interrupt

- Round trip latency 1.3 μ s

Control Network: Boot, monitoring and diagnostics

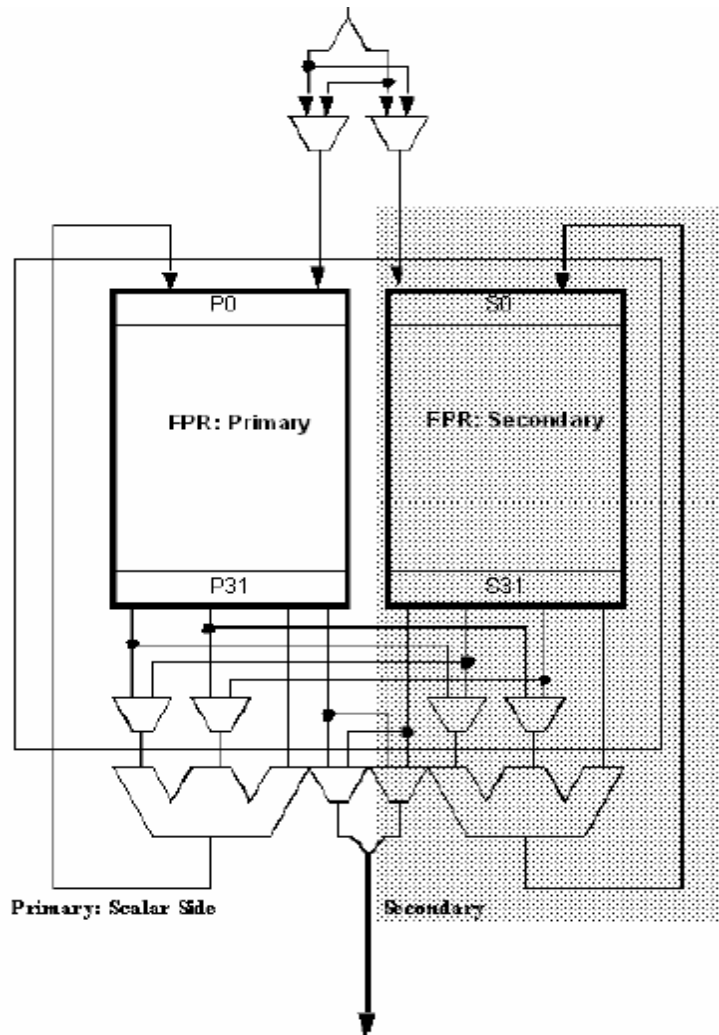
Ethernet

- Incorporated into every node ASIC
- Active in the I/O nodes (1:64)
- All external comm. (file I/O, control, user interaction, etc.)



Dual FPU Architecture (Hummer-2)

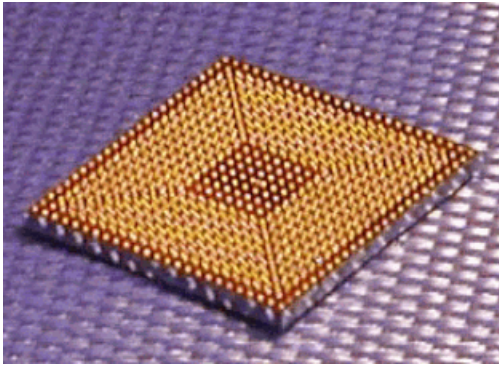
Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



- **440 PowerPC (32 bit)**
- **SIMD instructions over both register files**
 - *FMA operations over double precision data*
 - *More general operations available with cross and replicated operands*
 - Useful for complex arithmetic, matrix multiply, FFT
- **Parallel (quadword) loads/stores**
 - *Fastest way to transfer data between processors and memory*
 - *Data needs to be 16-byte aligned*
 - *Load/store with swap order available*
 - Useful for matrix transpose

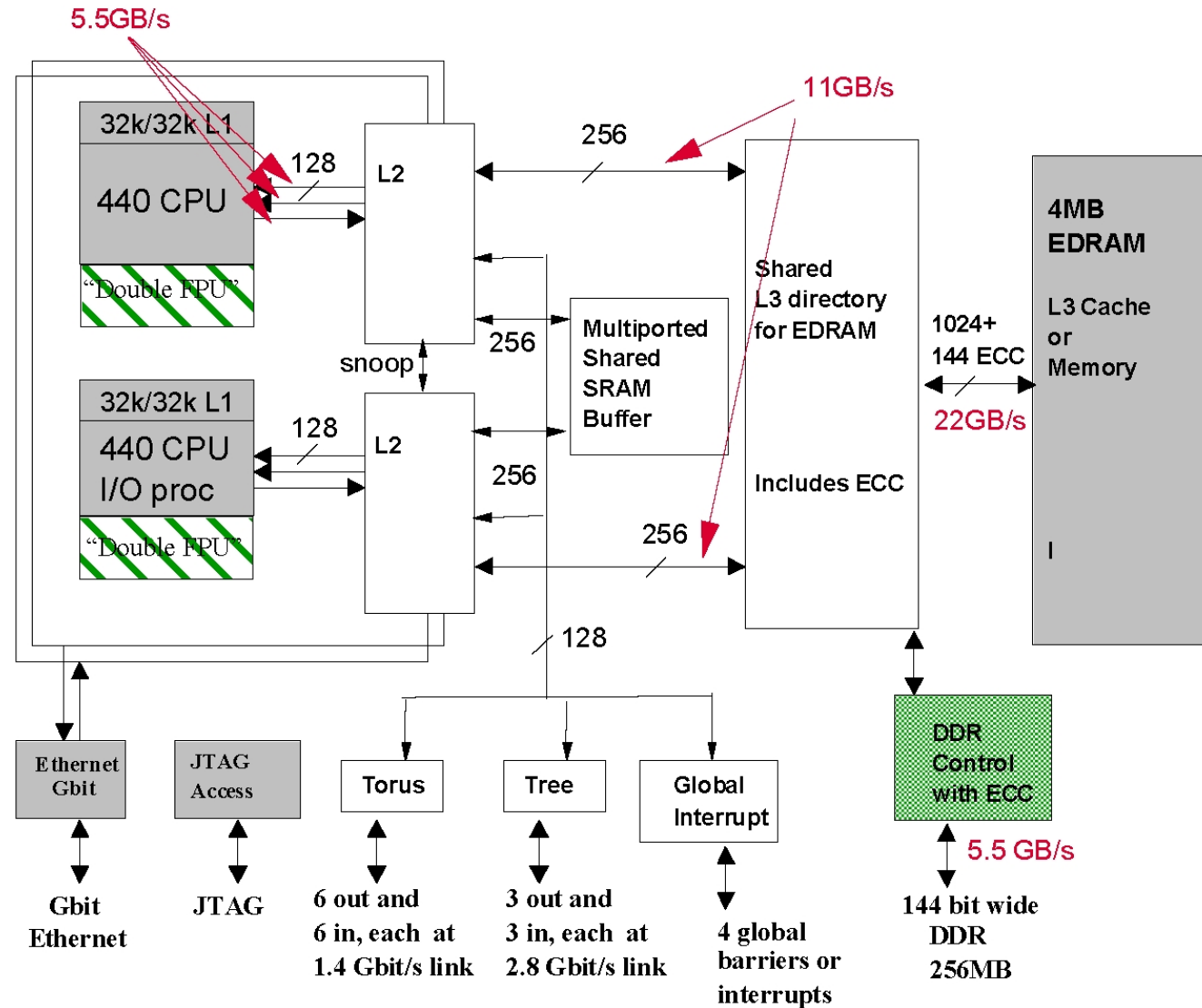
Dual Core BlueGene Processor

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



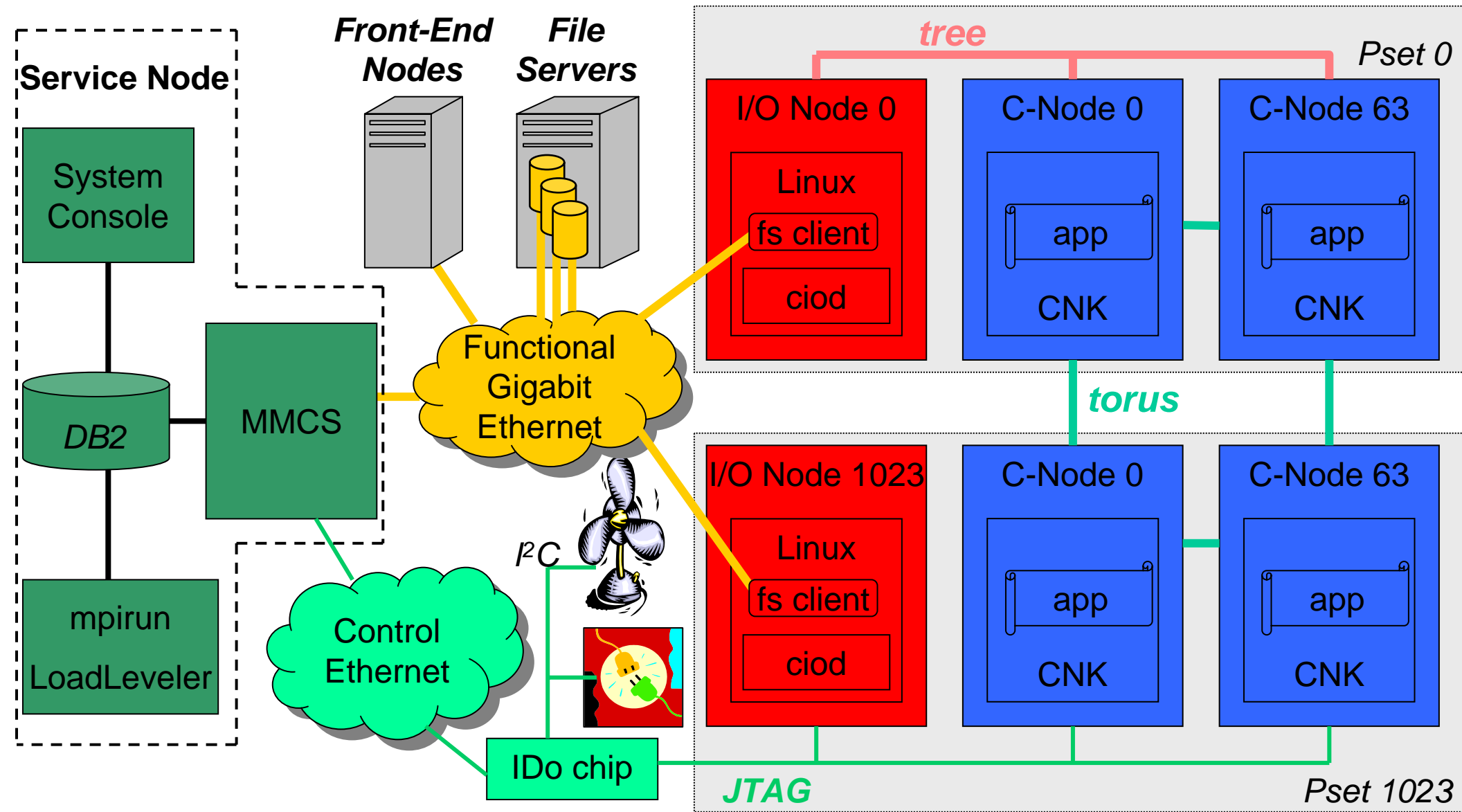
Fused Multiply/Add
= 2 FLOPs/cycle
Double FP pipeline
= 4 FLOPs/(cycle * FPU)
2 FPUs
= 8 FLOPs/cycle
700 MHz
= 5.6 GFLOPs/chip

cache ~43 cycles latency
DRAMs ~90 cycles latency



Blue Gene System Architecture

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft

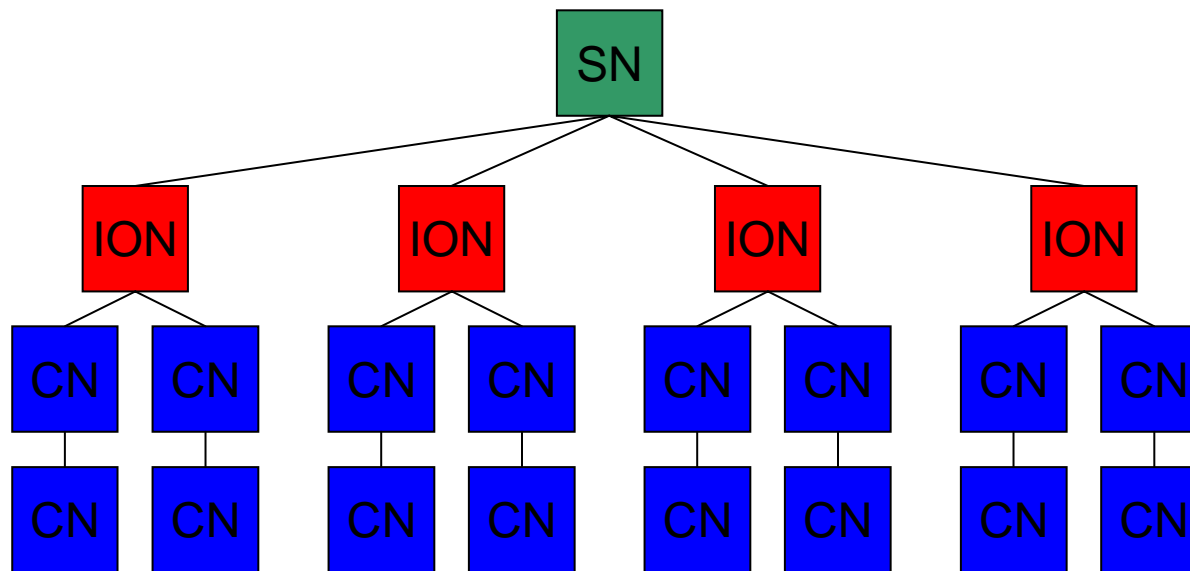


BG/L Hierarchical Node Organization

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



- **Compute nodes (CNs)** dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)
- **I/O nodes (IONs)** run Linux and provide a more complete range of OS services – files, sockets, process management, debugging
- **Service node (SN)** performs system management services (e.g., heart beating, monitoring errors) – transparent to application
 - the SN is a pSeries server running SLES9, not a BG node

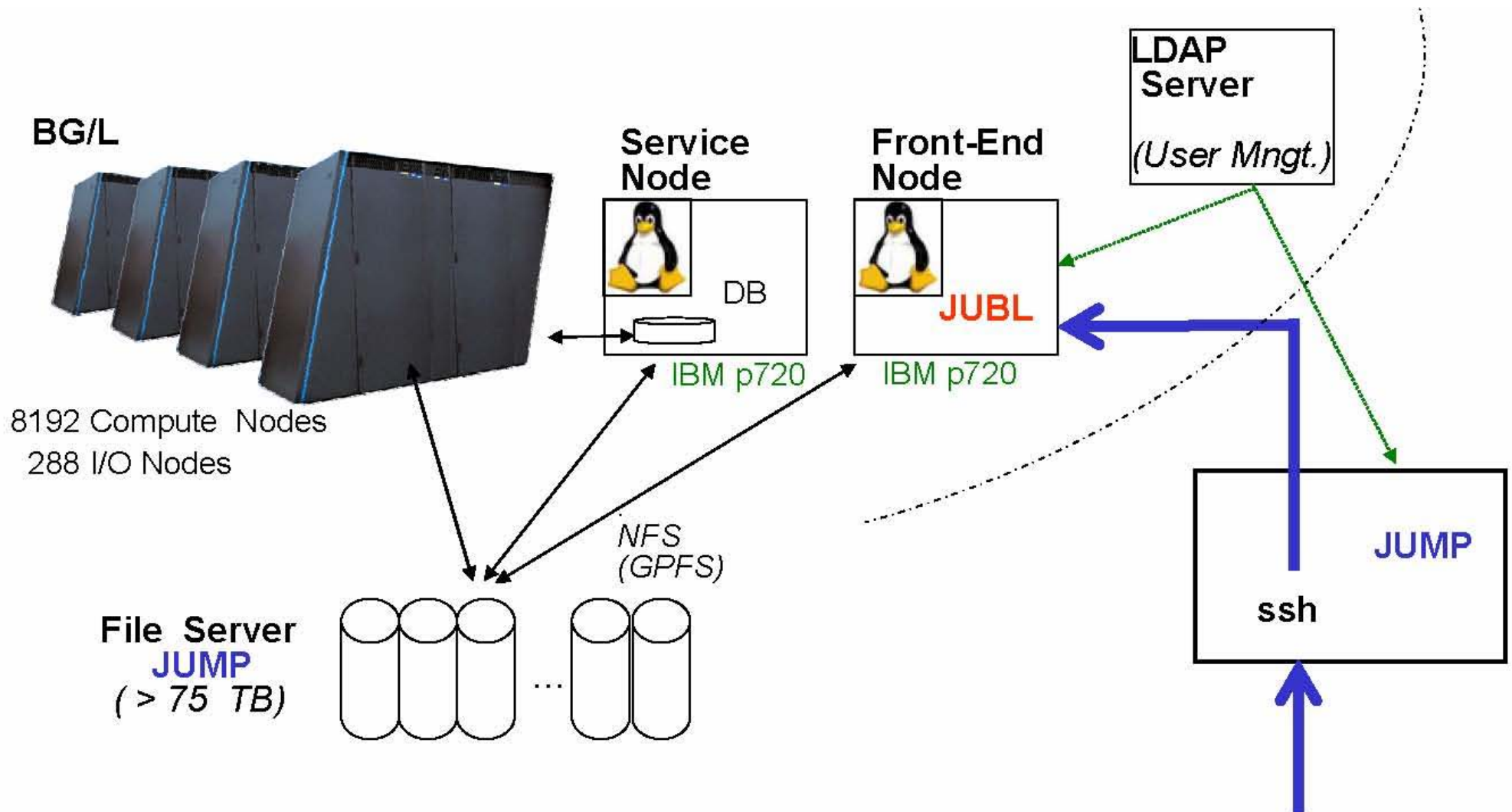




- **Linux (SLES 9) cross-compilation environment**
 - *User accesses system through frontend nodes for compilation, job submission, debugging*
- **Space sharing**
 - *one parallel job per BG/L partition*
 - *partitions are rebooted for each new job*
 - *one process per processor of the compute node*
- **SPMD model**
 - *all compute nodes run the same executable*
- **XL Fortran, XL C/C++, and MPI message passing**
 - *Virtual memory limited to physical memory (512MB/node)*
 - *Libraries statically linked, some syscall limitations*

JUBL integration with JuMP Cluster

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



JuBL Advance Booking and Monitoring

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft

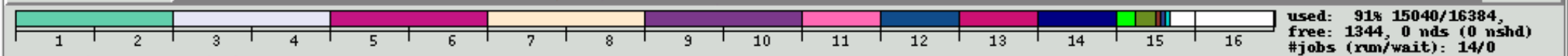


- **Advance Booking System**

- *FZJ Development (Dr. M. Stephan)*
- *Before Availability of LoadLeveler*
- *Standard: One Rack*
- *2 Priorities*

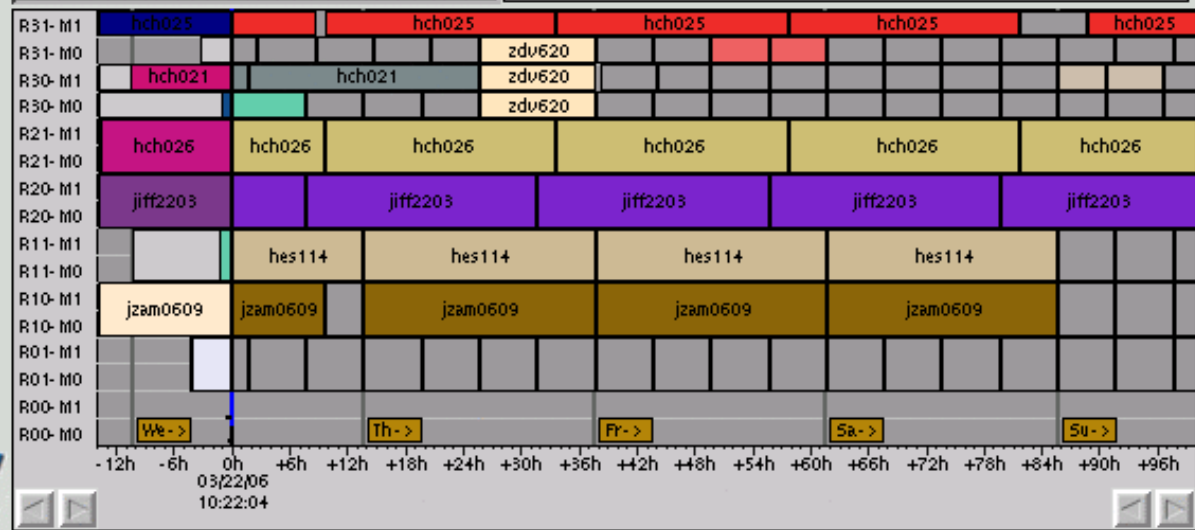
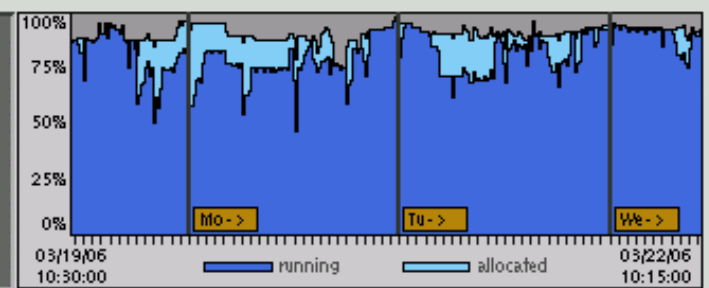
- **Ilview: Monitoring of Usage and Reservation**

- *FZJ Development (W. Frings)*
- *shows: BG/L - Usage, - Queues and - Reservations*
- *uses information from the Service Node database*
- *Web – Interface – Access to XML – file*
- *available for all users*



CPU	Userid	Class	books	nodes	tasks	mode	torus	cpuh	jobid	
2.	2048	hch026	running	32	1024	2048	v	111	13.6	R21
3.	2048	jiff2203	running	32	1024	2048	v	111	16.2	R20
4.	2048	hes114	running	32	1024	1024	c	111	1.3	R11
5.	2048	jzam0609	running	32	1024	2048	v	111	4.2	R01
6.	2048	jzam0609	running	32	1024	2048	v	111	14.3	R10
7.	1024	jzam0608	running	16	512	1024	v	111	0.3	R310
8.	1024	hch025	running	16	512	1024	v	111	15.3	R311
9.	1024	jiff0502	running	16	512	512	c	111	1.2	R300
10.	1024	hch021	running	16	512	1024	v	111	10.7	R301
11.	256	zdv625	running	4	128	128	c	000	0.2	RMP22Mr10074
12.	256	hes114	running	4	128	128	c	000	0.1	RMP22Mr10134
13.	64	zam014	init	1	32	0	v	000	0.8	R001_N1
14.	64	hes114	running	1	32	64	v	000	0.5	RMP22Mr09513
15.	64	hes114	running	1	32	64	v	000	0.3	RMP22Mr10060

Machine: Blue/Gene Juelich
 Memory: - GB, #cpus: 16384,
 speed=0.7 GHz type=PowerPC,
 #frames: 8, peak=44800 GFLOPS
 type=B6/L
 Date/Time: 03/22/06-10:22:04
 Usable Nodes: 16

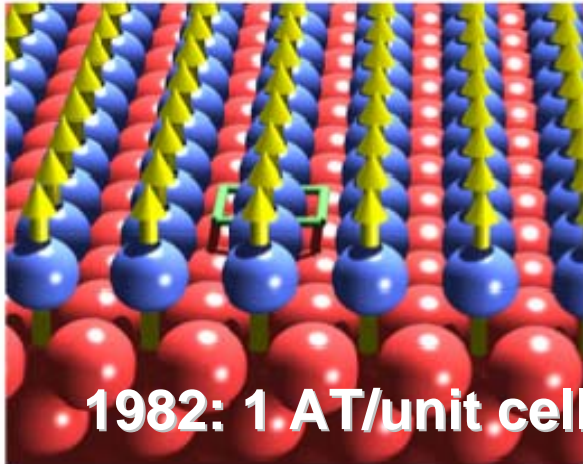




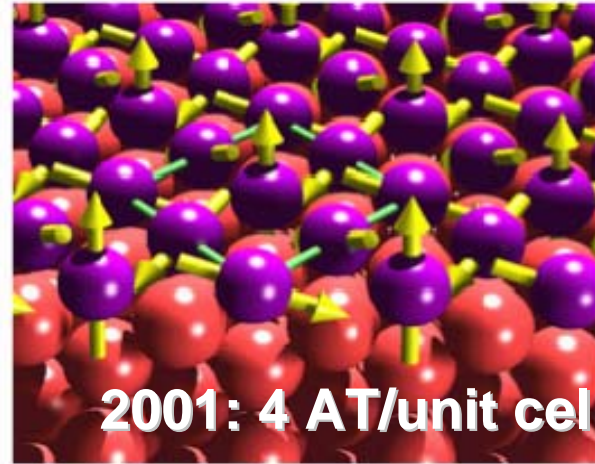
- Summer 2005: Pre-production status
- Sep. 2005: First official release V1R1M0
- Dec. 2005: Update to V1R2M0
- Feb. 2006: Update to V1R2M1
- Sep. 2005: **CTRL/X Kernel (compute node) for QCD**
 - research project together with IBM Yorktown Heights
 - performance improvements by kernel APIs and code reorganization
 - QCD reached 17% peak with standard BG/L - MPI
 - QCD reached 20,4% peak with low-level communication library
 - QCD scaled nearly 100% on 1024 nodes
- Mar 2006: ESSL, GPFS, LoadLeveler available

JuBL Application: Material Science

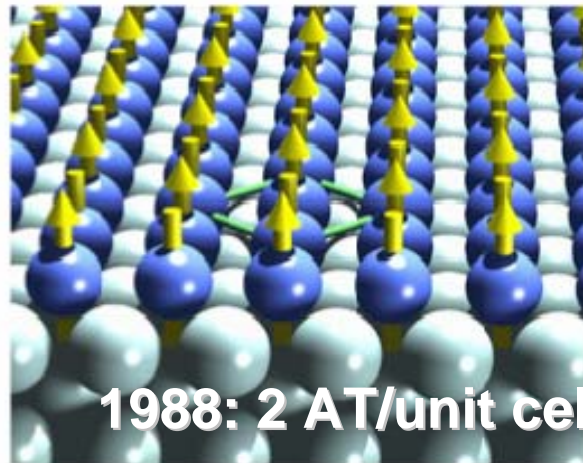
Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



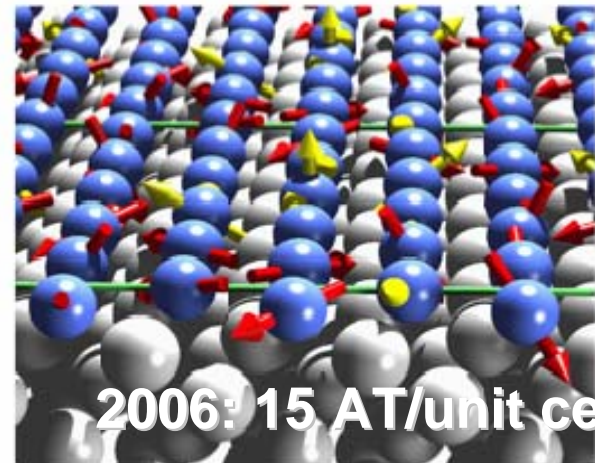
Fe/Cu(001)



Mn/Cu(111)

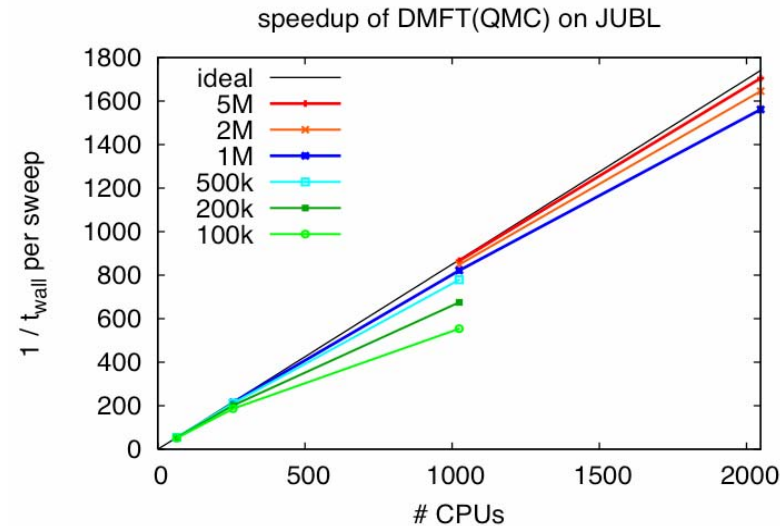


Fe/W(001)



Fe/Ir(111)

- **GMR**: Gigant Magnetoresistance today in every harddisk !
- **CMR**: Colossal Magnetoresistance is now computed on **JuBL**!
- → Harddisk of the Future



Prof. Stefan Blügel et al. (FZJ)

Dr. Eva Pavarini, Dr. Erik Koch

JuBL: Laser - Plasma - Simulation

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



Petawatt - Laser

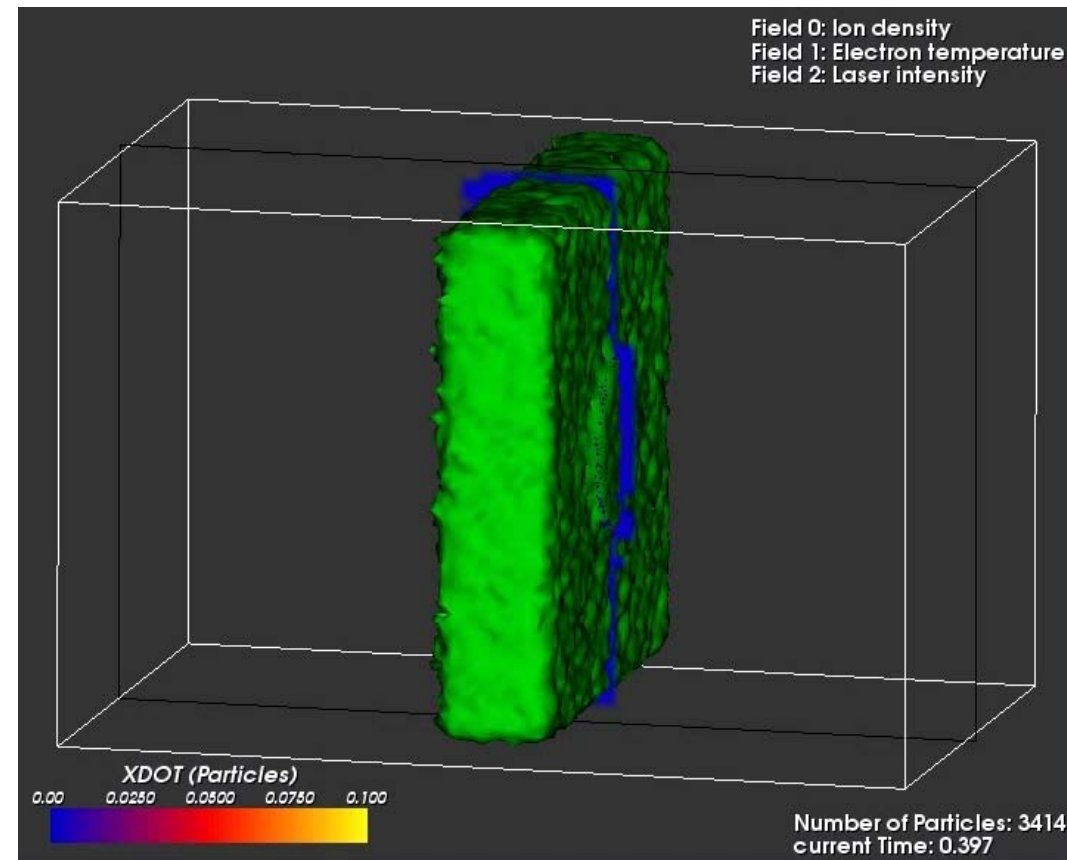
10 Billion Degree
hot Electrons

Applications

Tumor Treatment

Fusion

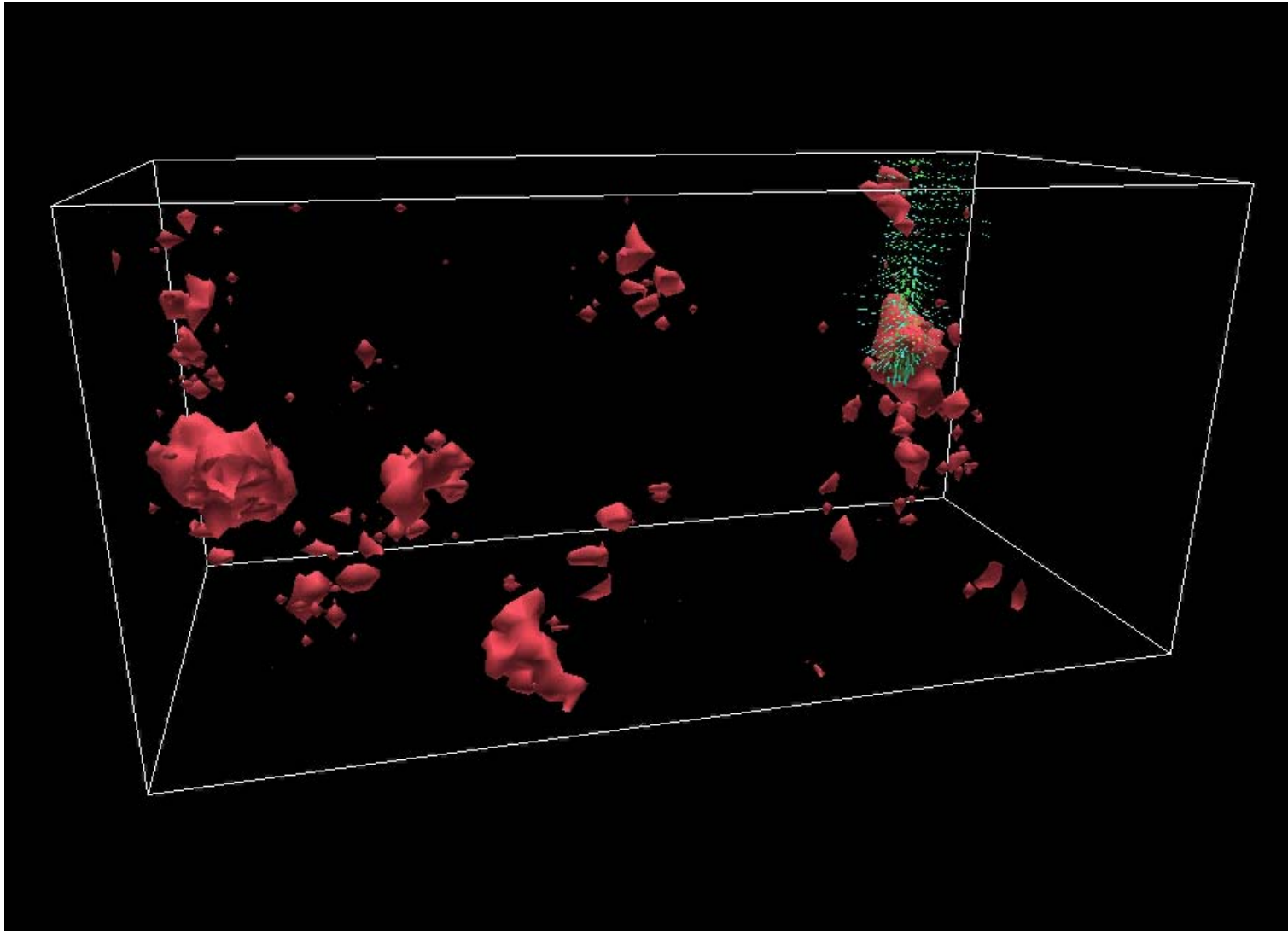
Particle Accelerator



Prof. Paul Gibbon, FZJ

JuBL: Groundwater Contamination

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft



Prof. Harry Vereecken, FZJ



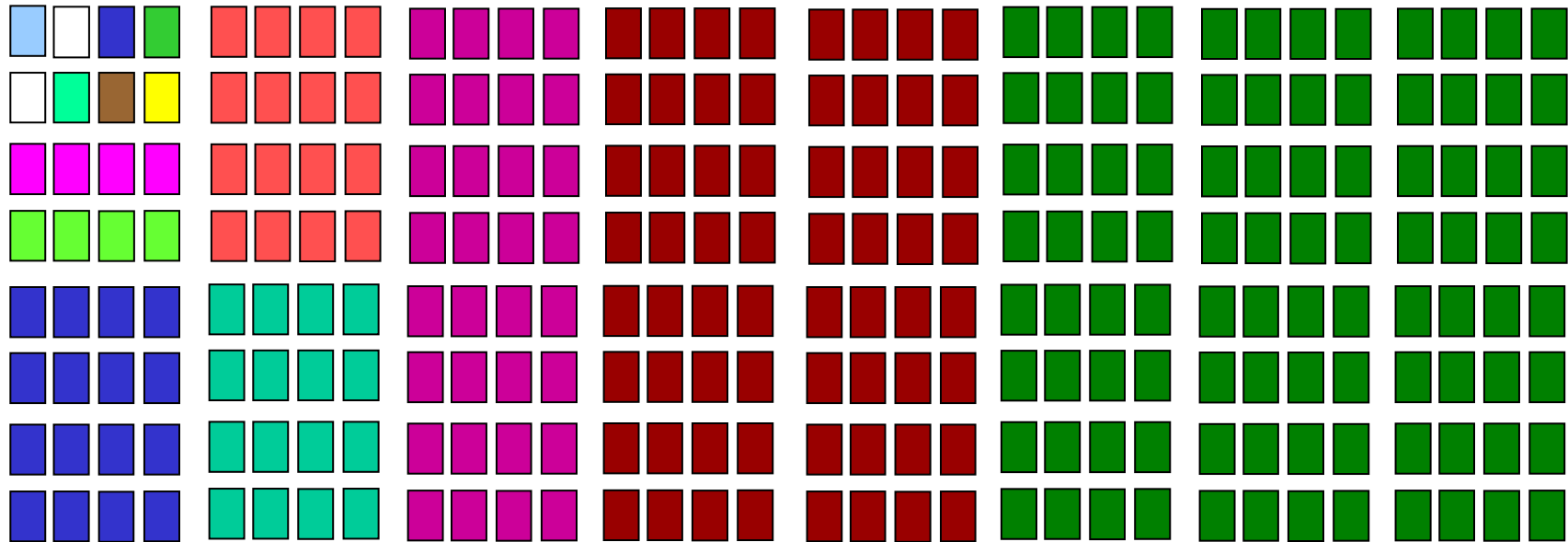
Variety of Applications:

- Material Science, CPMD, DMFT, FZJ-IFF
- Biophysics, SMMP, Protein Folding, NIC-Research Group
- Quantum Chemistry, VASP, Univ. Duisburg
- Elementary Particle Physics, DESY-Zeuthen, Wuppertal, ZAM
- Laser-Plasma-Interaction, PEPC, ZAM
- Astrophysics, nbody6++, Heidelberg

Variety of Methods:

- Molecular Dynamics
- (Quantum-) Monte Carlo
- DFT, CFD

JuBL Partitions



1-2 racks with small blocks (32, 128, 512 nodes)

Other racks used for applications with 1024 or more nodes

Production with big applications only!



- *Blue Gene/L is balanced regarding processor-, memory- and communication- performance*
- *Applications should be balanced and scale via MPI*
- *Easy porting from p690+ cluster*
- *Memory restrictions not as serious as expected*
- *Some applications sensitive to mapping 3D torus*
- *More high scaling applications as expected*
- *Very short installation time (8 racks in 10 days)*
- *Great demand, users already see long waiting times*
- *New application areas*

- *Price / Performance !*

