Briefing:
High Performance Computing System Division

Taisuke Boku, Division Leader
Center for Computational Sciences /
Graduate School of Systems and Information Science

Organization

- High Performance Computing System Division
  - System Architecture Research Group
    - Taisuke Boku (Professor)
    - Daisuke Takahashi (Associate Professor)
  - Grid Computing Research Group
    - Mitsuhisa Sato (Professor)
    - Osamu Tatebe (Associate Professor)
  - PD
    - Toshihiro Hanawa
  - Since system architecture and Grid computing related very closely, all faculty members are working in borderless manner

Research field

- HPC system research & development
  - HPC system architecture
    - HPC processor architecture including memory hierarchy
    - Scalable & high-performance interconnection network
    - Total system design and solution
  - System software
    - Compiler
    - HPC math. Library
    - Network drivers for PC cluster
    - Model and language for HPC
  - Grid computing
    - Grid RPC system
    - Distributed file system for Data Grid

Collaboration with application fields

- Advising application people from system viewpoint
  - Large scale parallelization (MPI programming, performance bottleneck checking, ...)
  - Performance tuning for processor architecture (FP acceleration, memory hierarchy, ...)
  - Numerical solutions (development of math. library)
  - System design (file system, new project support, ...)
- System operation support
  - PACS-CS
  - FIRST

Education

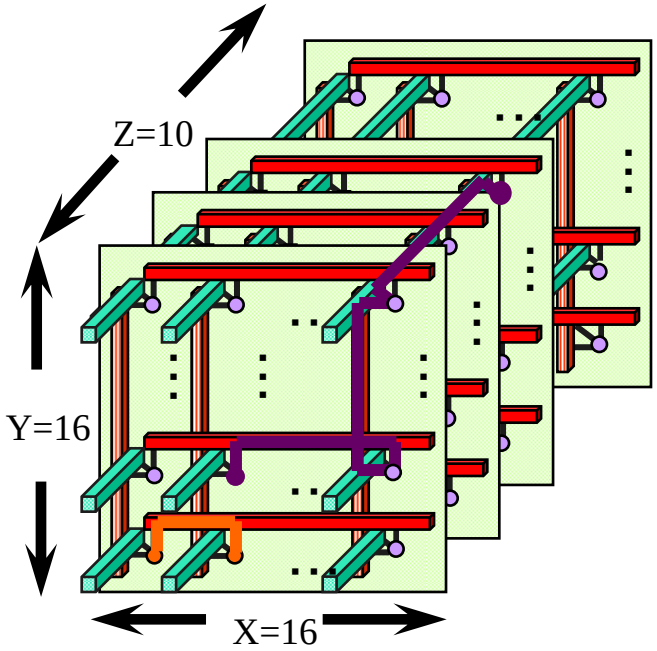
- 4 professors share a collaborative laboratory named “High Performance Computing System Lab.” in Department of Computer Science, Graduate School of Systems and Information Engineering
- Students
  - Doctoral Course: 5
  - Master Course: 7
  - Undergraduate: 4

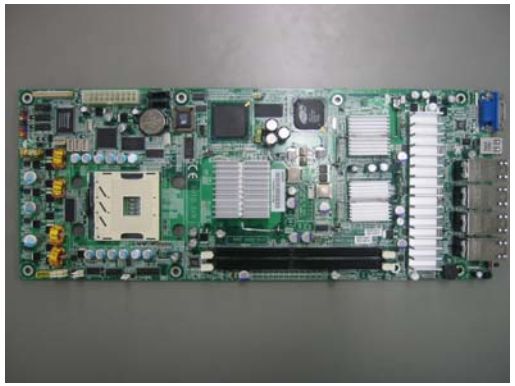
Research topics in HPCS Division

- HPC System Architecture
  - HPC processor architecture and memory system: SCIMA
  - Large scale parallel processing network: HXB
  - Power-aware computing: Optimized DVFS on parallel processing
  - Large scale cluster computing: VFREC-Net, RI2N
- Software
  - OpenMP compiler: Omni OpenMP
  - New language model for large scale parallel processing: OpenMPD
  - High performance and scalable math. Library: FFT, orthogonalization
- HPC Grid
  - Grid RPC: OmniRPC
  - Data Grid on distributed file system: Gfarm
  - Grid interoperability

Topic: PACS-CS development

- Total system design
- 3D-HXB/Ethernet network driver development
- Special communication library for low-level high-performance communication
- After installation, mainly supporting the performance tuning, parallel code development and code porting
  - Lattice QCD
  - RS-DFT
  - Tree-Puzzle
  - WRF





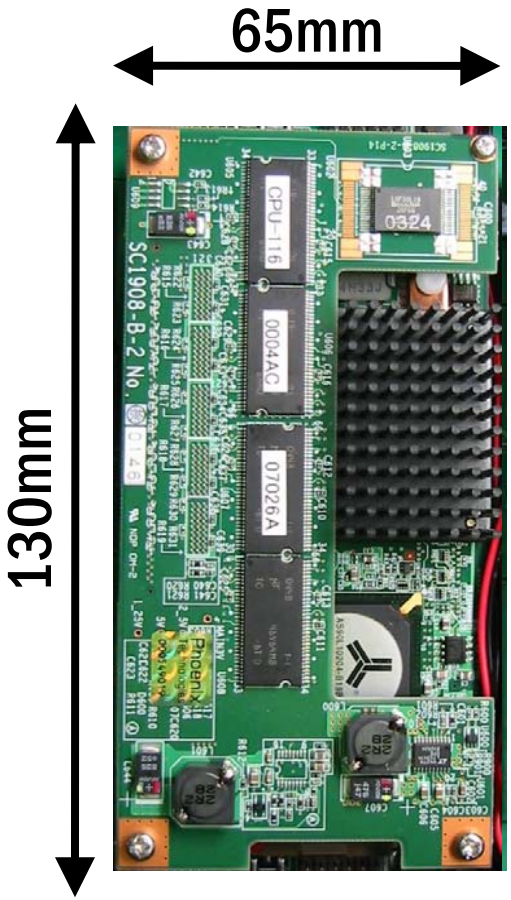
Topic: FIRST development

- Basic concept design of HMCS (Heterogeneous Multi-Computer System)
- Conceptual design of Blade-GRAPE solution
- Cluster configuration
- Gfarm installation and operation for shared file system

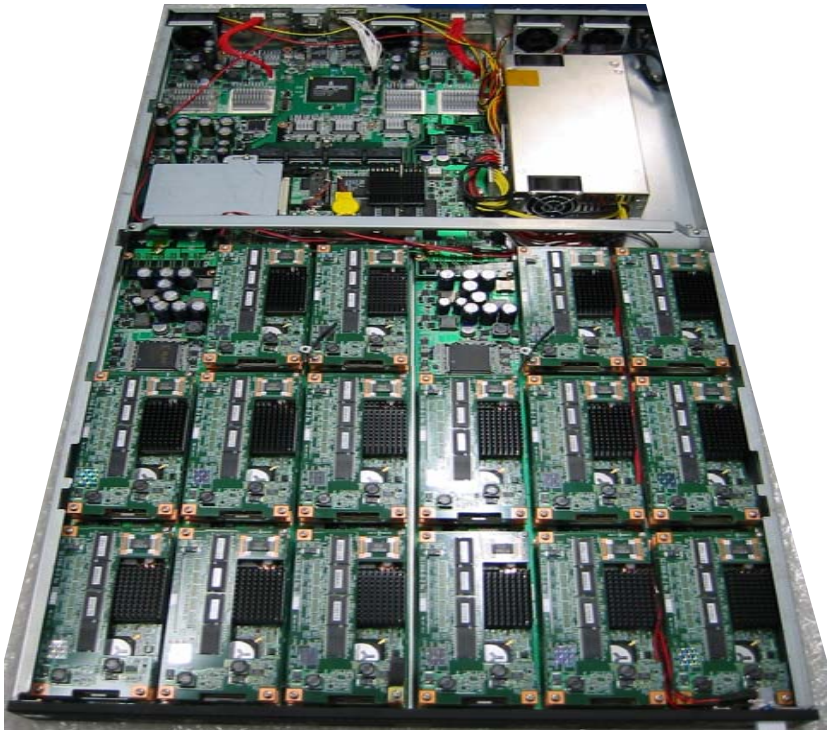
Topic: Power-aware parallel processing

- MegaProto (JST/CREST, finished on Oct. 2006)
  - Prototype R&D for very small and low-power CPU module with embedded technology
  - Using Efficion processor (Transmeta, 3W TDP)
  - 16 of CPU modules + 1 controlling processor module
  - On-board dual-link Gigabit Ethernet + 24 port switch x 2 (with 16 ports of up-link)
  - MegaProto/E prototype with 16 computation nodes and dual-link GbE network in 1U form factor
    ⇒ 32 GFLOPS peak performance (2.7 times faster than dual Xeon SMP solution) with same power consumption

Topic: Power-aware computing (cont'd)



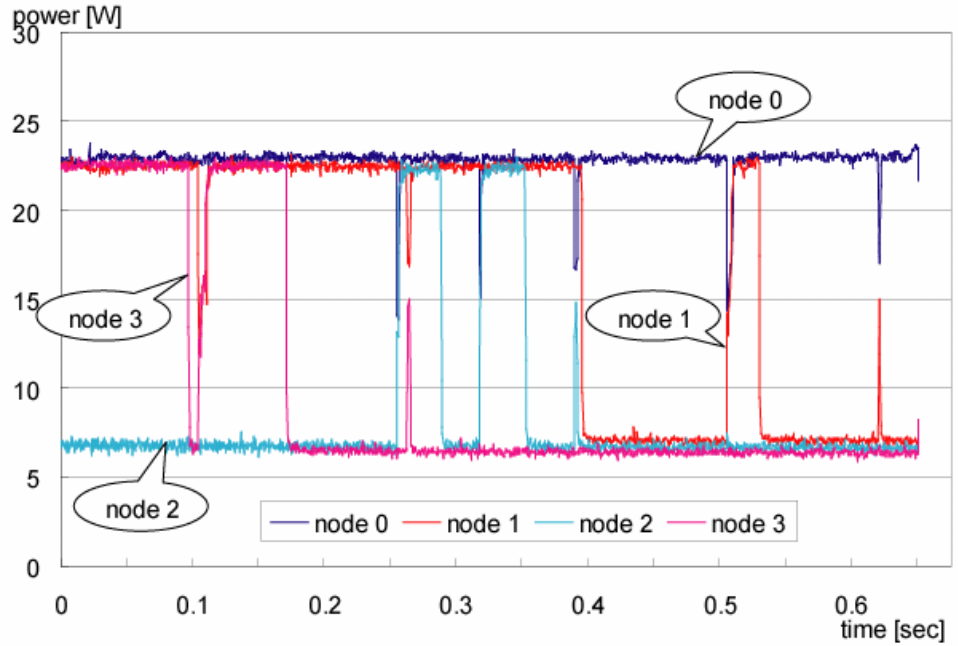
CPU module of MegaProto/E



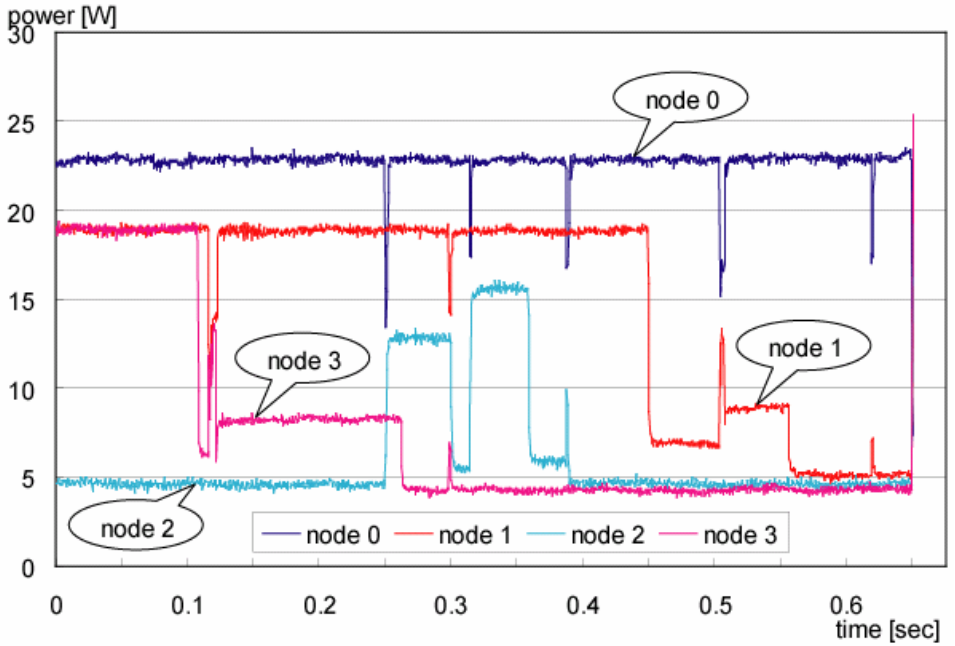
Cluster Unit of MegaProto/E with 16 CPUs

Power-aware computing (cont'd)

- Power optimization on parallel program
  - DVFS: Dynamic Voltage and Frequency Scaling
  - Changing CPU "gear" for optimal one utilizing "slack-time" of parallel execution with imbalanced parallel application



Program executed at standard gear

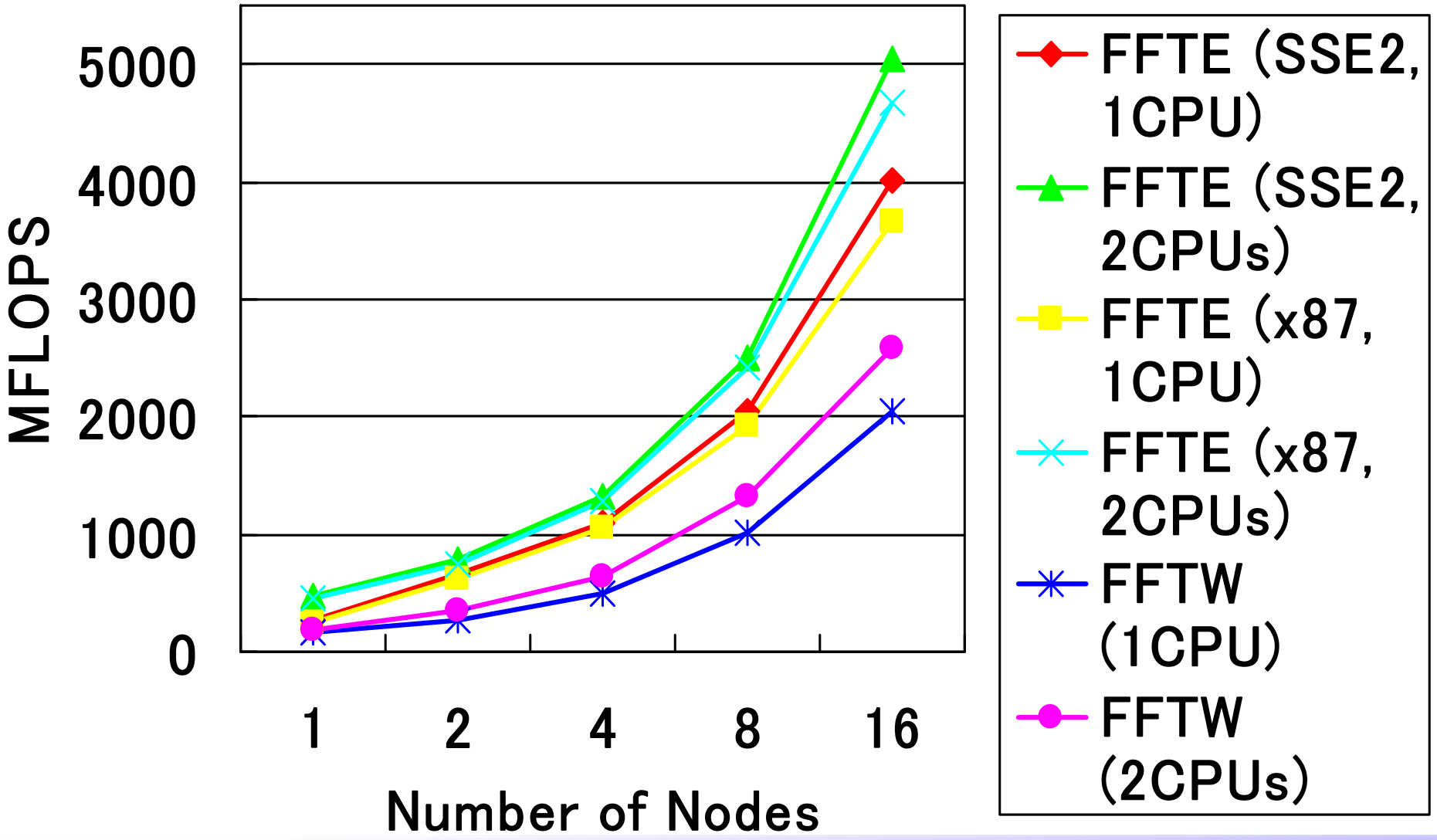


Power optimized program

Topic: HPC math. libraries

- FFT-E (FFT East)
  - Included in HPCC benchmark suite
  - Multi-platform highly optimized FFT library
  - Well-tuned with cache awareness for large scale PC cluster
  - Combining multicolumn FFT and data transposing to reduce cache miss-hit on each node
  - Better performance than well-known FFT-W library

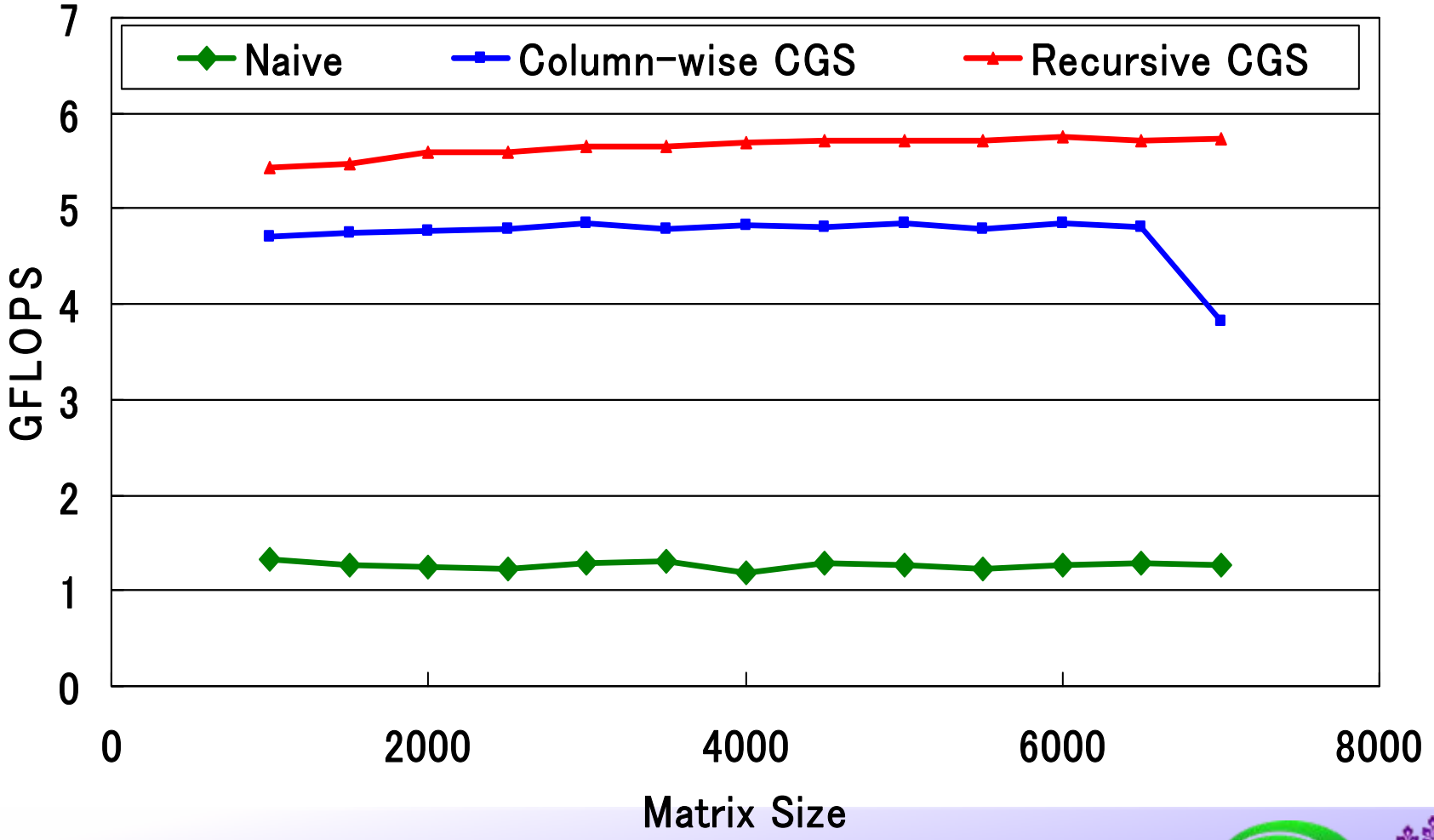
Performance of parallel 3-D FFTs
(dual Xeon PC SMP cluster, N1xN2xN3 = 2^24xP)



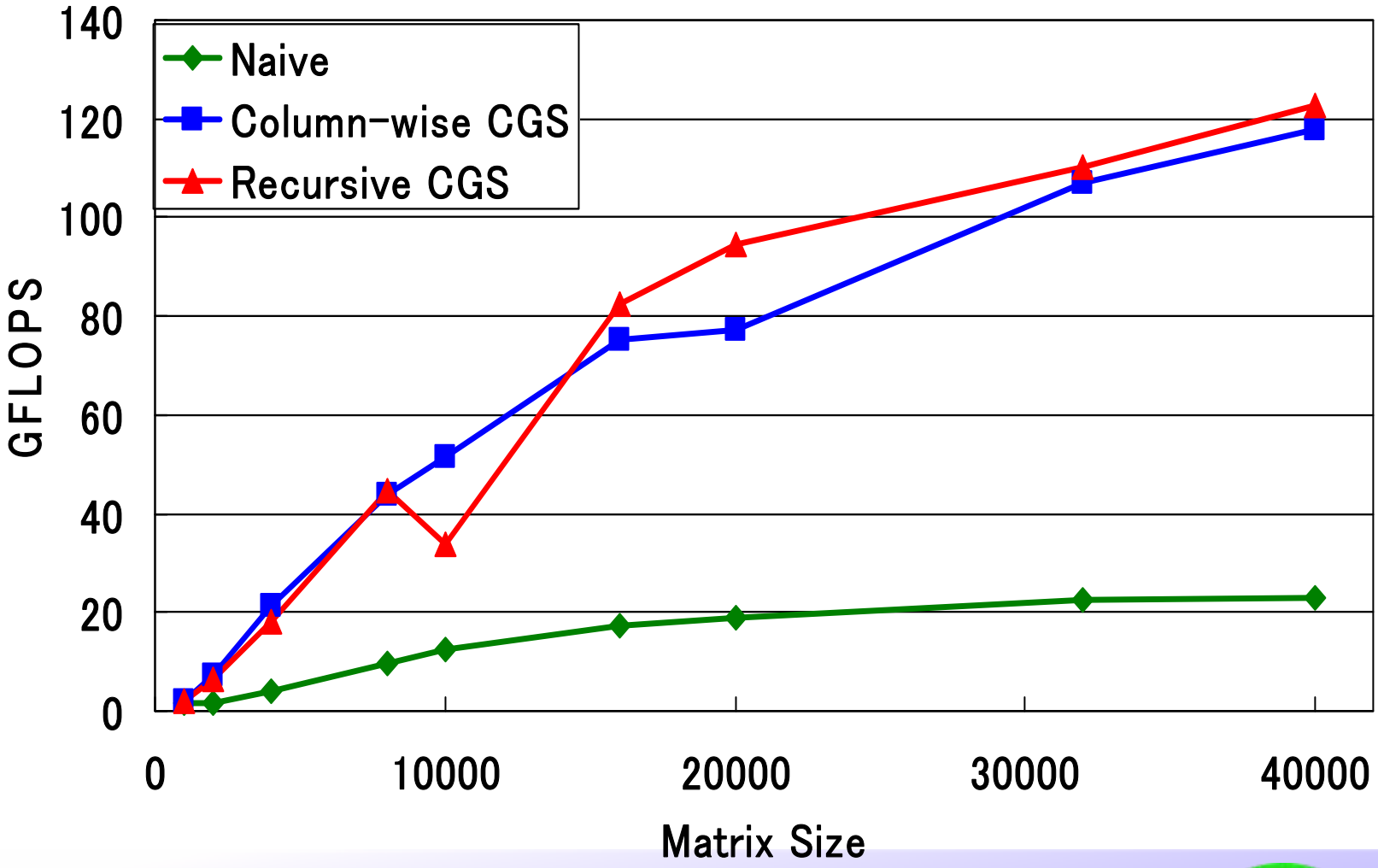
Topic: math. library (cont'd)

- Cache-aware orthogonalization library
  - Especially required for RS-DFT (Real Space Density Function Theory) on large number of atoms
  - On classical Gram-Shcmidt Orthogonalization does not work well on cache architecture
  - Modifying the algorithm to fit Level-2 or Level-3 BLAS

Performance on Xeon 3GHz (1CPU)

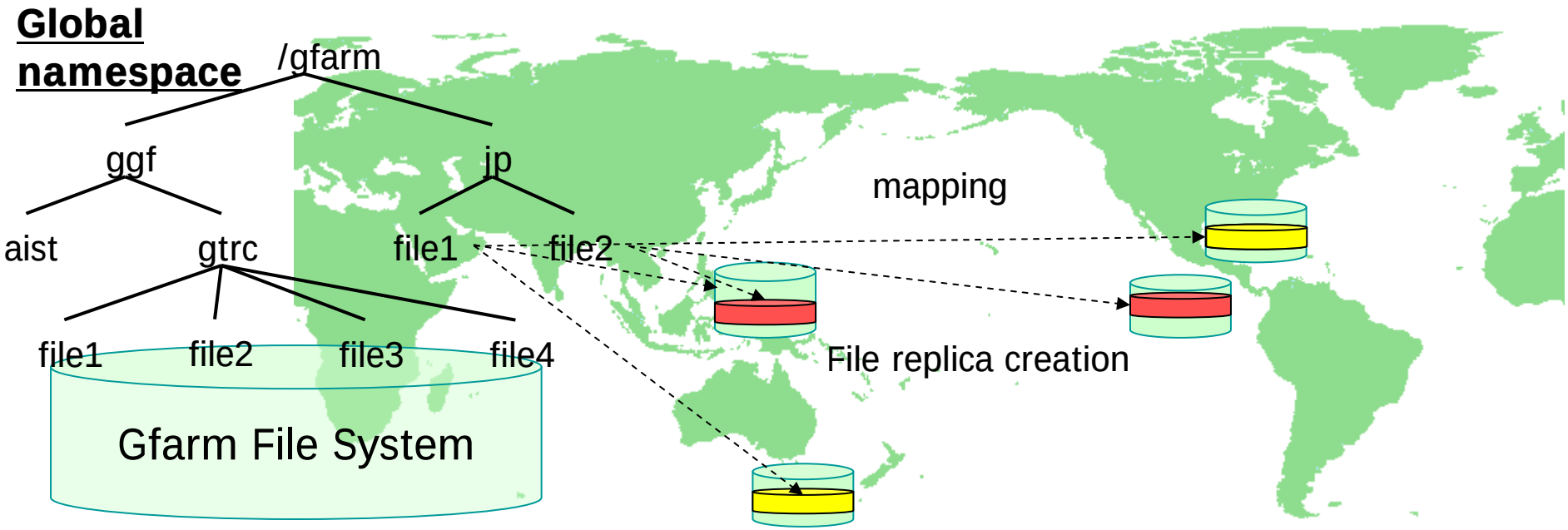


Performance on 32 node 3GHz Xeon PC Cluster



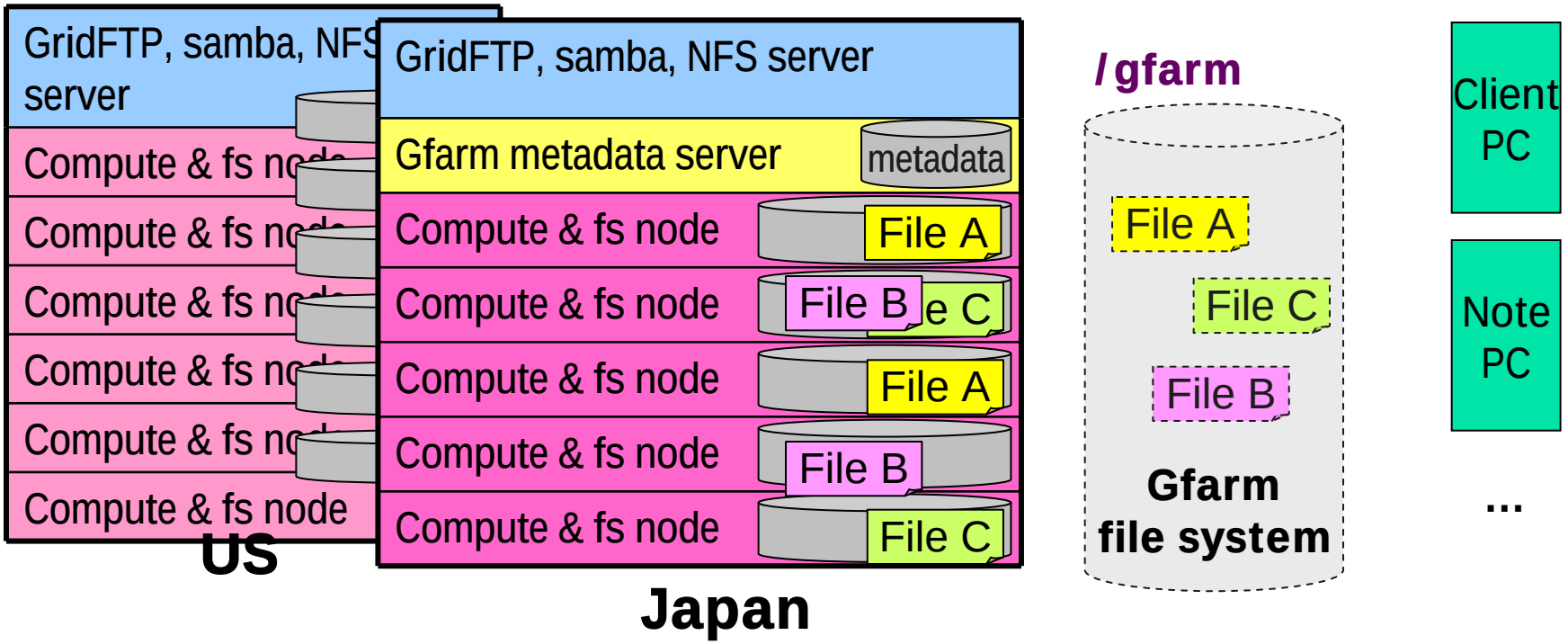
Topic: Gfarm distributed file system for Data Grid

- **Commodity-based distributed file system** that federates storage of each site
- It can be **mounted** from all cluster nodes and clients
- It provides **scalable I/O performance** wrt the number of parallel processes and users
- It supports fault tolerance and avoids access concentration by automatic replica selection



Topic: Gfarm (cont'd)

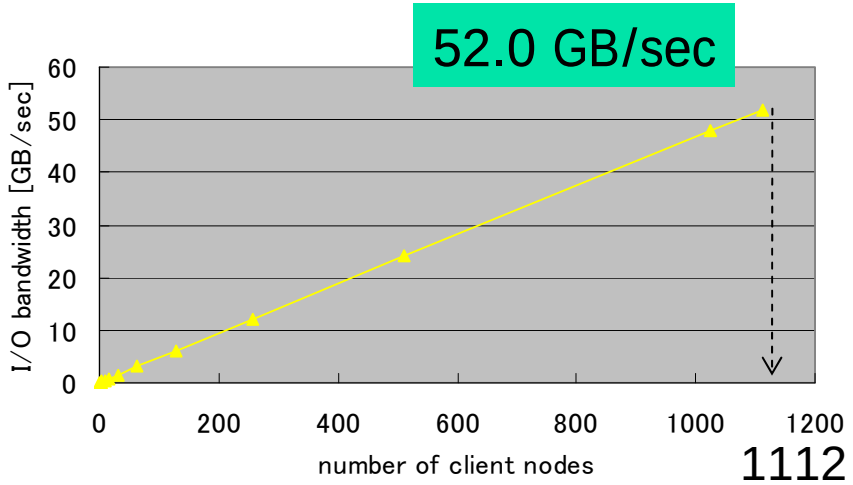
- Files can be shared among all nodes and clients
- Physically, it may be replicated and stored on any file system node
- Applications can access it regardless of its location
- File system nodes can be distributed



Gfarm for particle physics data analysis

- **O. Tatebe et al, "High Performance Data Analysis for Particle Physics using the Gfarm File System", SC06 HPC Storage Challenge, Winner – Large Systems, 2006**

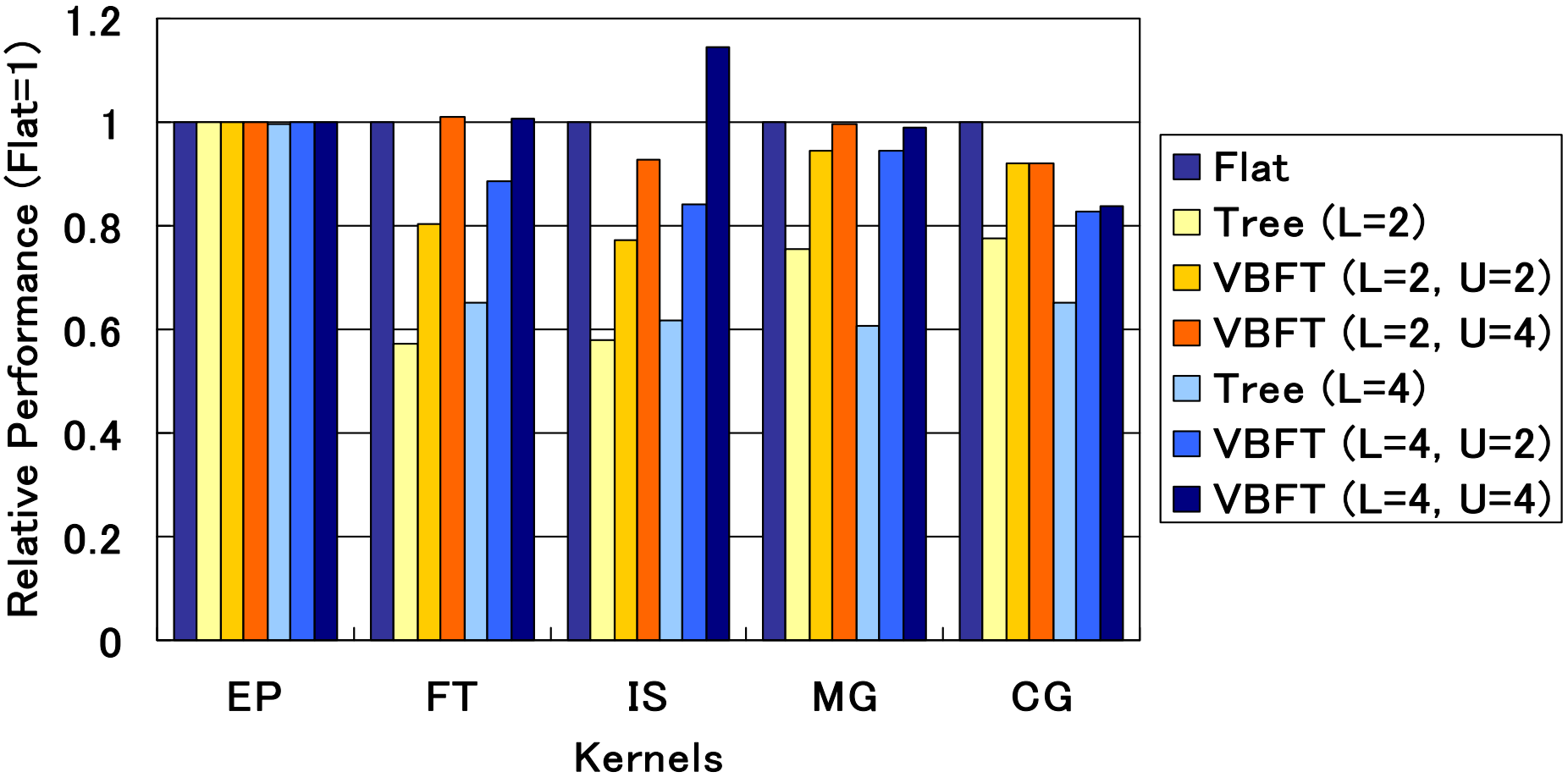
• Construct 26 TB of Gfarm FS using **1112** nodes
• Store all 24.6 TB of Belle experiment data
• 52.0GB/s in parallel read
→ **3,024** times speedup
• 24.0GB/s in skimming process for b → s γ decays using 704 nodes
→ **3 weeks to 30 minutes**



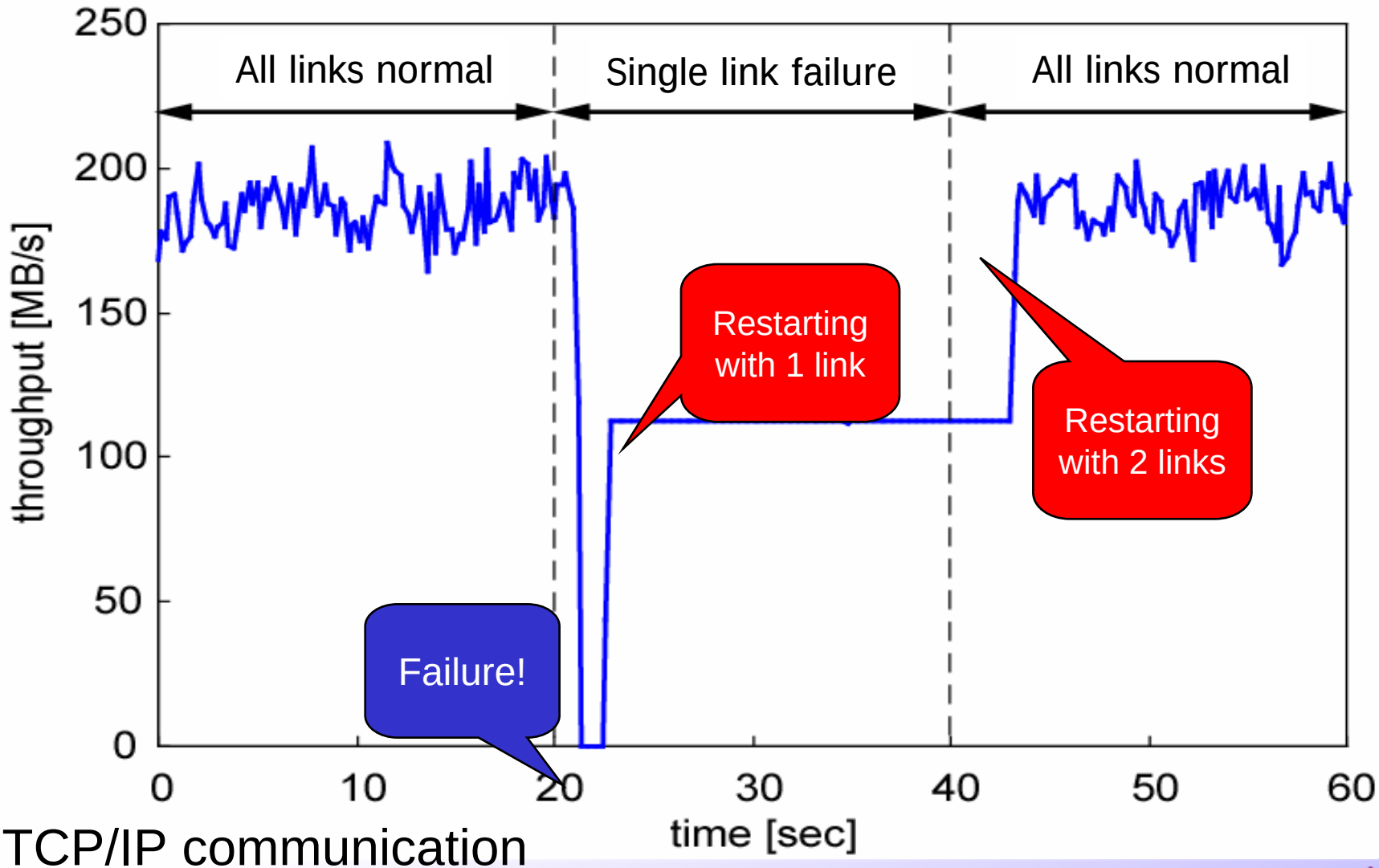
Topic: HPC interconnection for PC cluster

- High-performance, scalable and reliable commodity network for inexpensive PC clusters: VFREC-Net & RI2N
  - Utilizing multi-link Gigabit Ethernet
  - High-performance: aggregated bandwidth of multiple links
  - Scalability: VLAN-based Fat-Tree is enable even with inexpensive Layer-2 Ethernet switch
  - Reliability: Multiple links work as redundant connection on the failure of switches and links
  - Everything is implemented as a special network driver

NAS Parallel Benchmark with Xeon cluster with 32 nodes

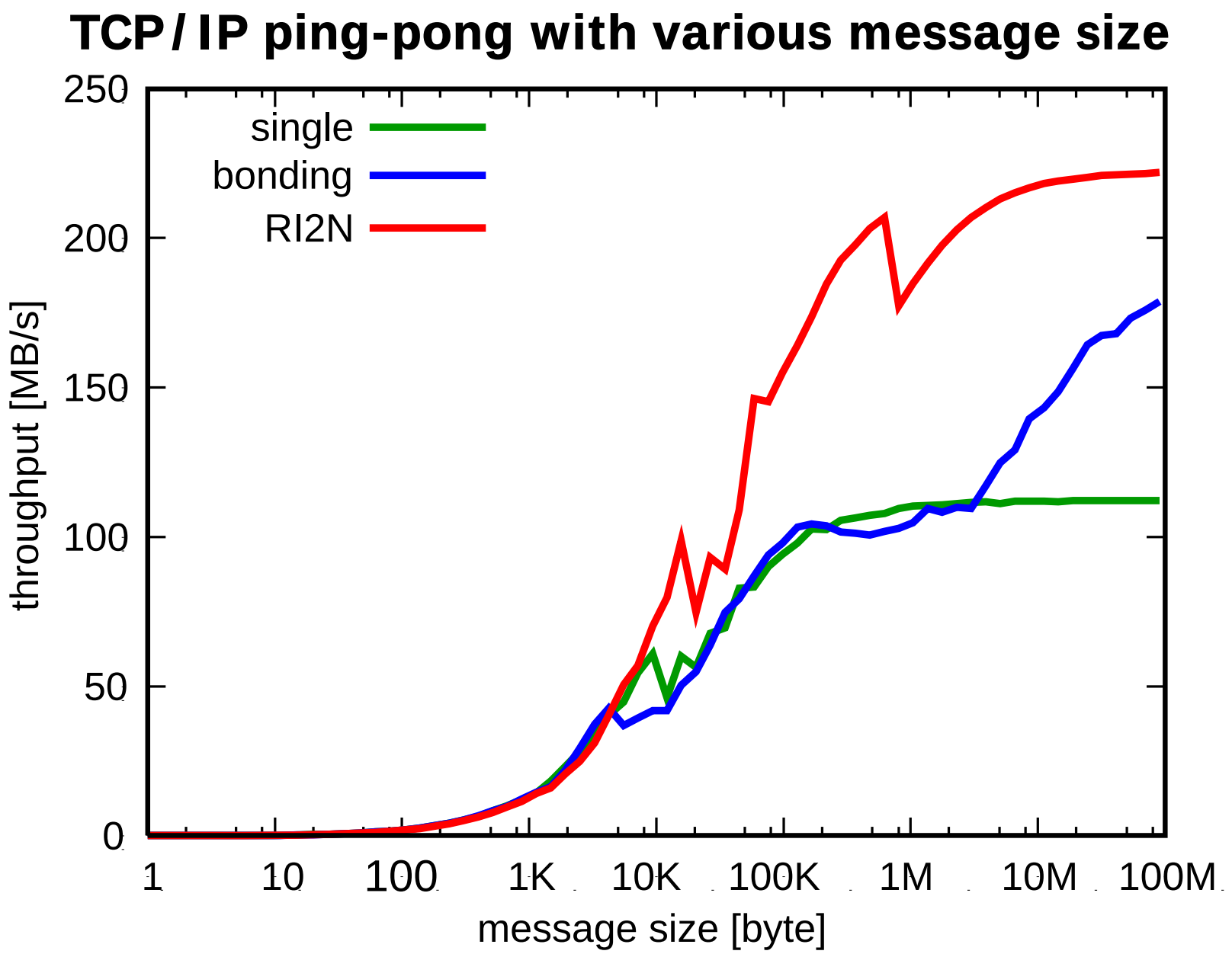


Throughput on failure



TCP/IP communication

Comparison with Linux “channel bonding”



Major Research Collaboration

- Large scale cluster computing and Grid
  - T2K Alliance (U. Tokyo & Kyoto U.)
  - AIST
- Low-power & High-performance processor architecture
  - U. Tokyo
- Next Generation Supercomputer System
  - RIKEN
- HPC Grid computing
  - INRIA
  - PRAGMA (Asia-Pacific Grid middleware community)
- HPC Cluster network system
  - AIST
  - NII

Activities outside CCS

- Contribution to RIKEN's Next Generation Supercomputer Development Project
  - Sato and Boku are the members of system architecture working group as visiting researchers of RIKEN
  - All members will contribute the performance tuning and evaluation on large scale QCD, RS-DFT and FFT under research contract with RIKEN
- Social works
  - All members have been playing important roles on HPC society such as symposium and workshop organization, PC chairs, etc.
  - IPSJ SIGHPC Chairs: Sato (1998-2001) & Boku (2006-)
  - Recognized as very active research group in HPC community in Japan, Asia and the world