FUJITSU

shaping tomorrow with you

Programming for Fujitsu Supercomputers

Koh Hotta

The Next Generation Technical Computing

Fujitsu Limited

FUJITSU

To Programmers who are busy on their own research, Fujitsu provides environments for

Parallel Programming

Tuning Parallel Programs

System Software Stack

FUJITSU

User/ISV Applications

HPC Portal / System Management Portal

Technical Computing Suite

System operations management

- System configuration management
- System control
- System monitoring
- System installation & operation

High-performance file system

- Lustre-based distributed file system
- High scalability
- IO bandwidth guarantee
- High reliability & availability

Job operations management

- Job manager
- Job scheduler
- Resource management
- Parallel execution environment

VISIMPACT™

- Shared L2 cache on a chip
- Hardware intra-processor synchronization

Compilers

- Hybrid parallel programming
- Sector cache support
- SIMD / Register file extensions

Support Tools

- IDE
- Profiler & Tuning tools
- Interactive debugger

MPI Library

- Scalability of High-Func.
- Barrier Comm.

Linux-based enhanced Operating System | **Red Hat Enterprise Linux**

Super Computer: PRIMEHPC FX10 | **PC cluster: PRIMERGY**

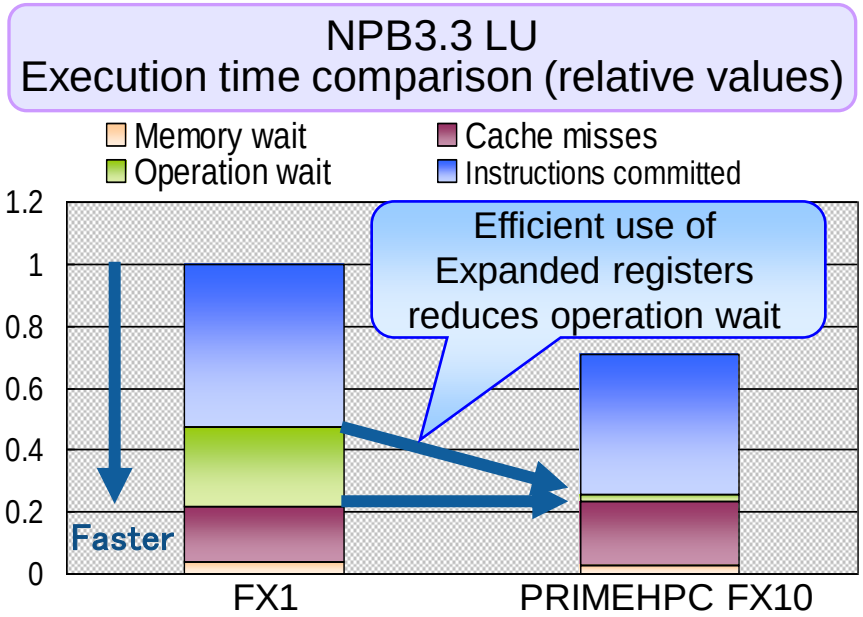
HPC-ACE architecture & Compiler

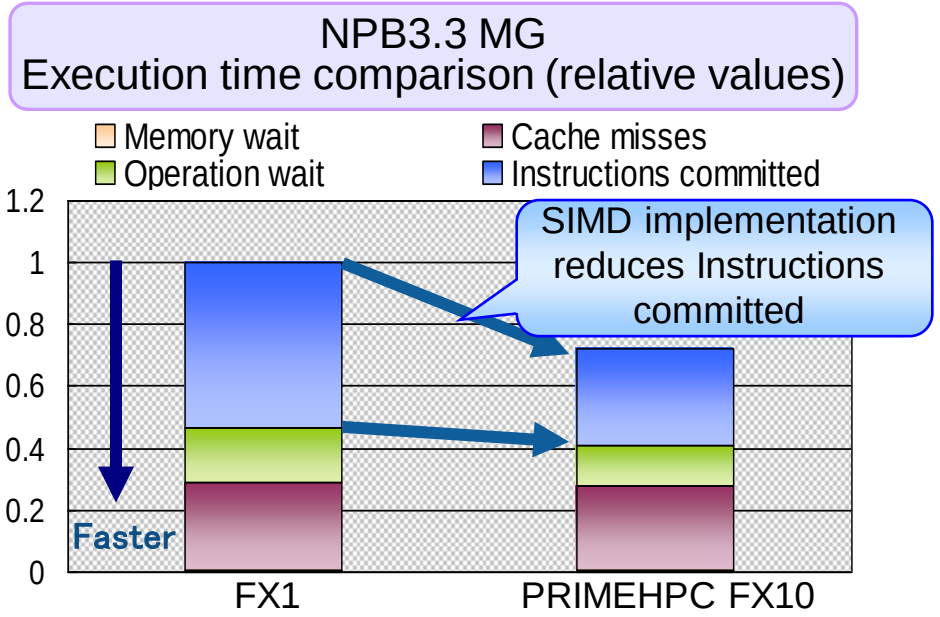
FUJITSU

- Expanded Registers
  - reduces the wait-time of instructions
- SIMD instructions
  - reduces the number of instructions
- Sector Cache
  - improves cache efficiency

Compiler Gives Life to HPC-ACE

- NPB3.3 LU: operation wait-time dramatically reduced
- NPB3.3 MG: number of instruction is halved





Parallel Programming for Busy Researchers

FUJITSU

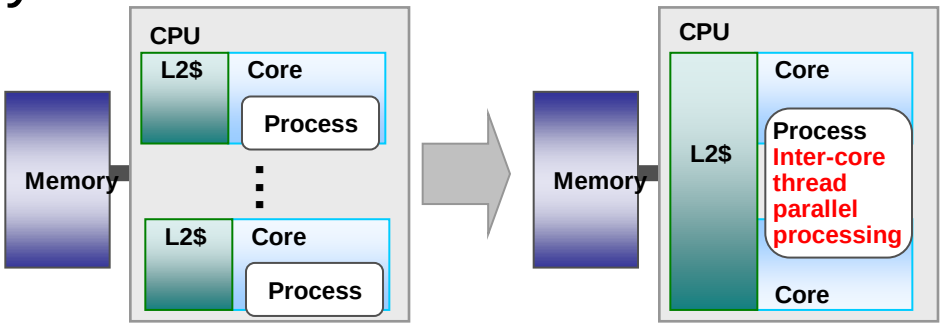
- Large # of Parallelism for Large Scale Systems
  - Large # processes need Large Memory & Overhead
    - **hybrid thread-process programming to reduce number of processes**
  - ***Hybrid parallel programming is annoying for programmers***
- Even for multi-threading, the more coarse grain, the better.
  - Procedure level or outer loop parallelism is desired
  - Little opportunity for such coarse grain parallelism
  - ***System support for "fine grain" parallelism is required***
- **VISIMPACT solves these problems**

VISIMPACT™
(Virtual Single Processor by Integrated Multi-core Parallel Architecture)

FUJITSU

- Mechanism that treats multiple cores as one high-speed CPU through automatic parallelization
  - Just program and compile, and enjoy high-speed

- You need not think about hybrid

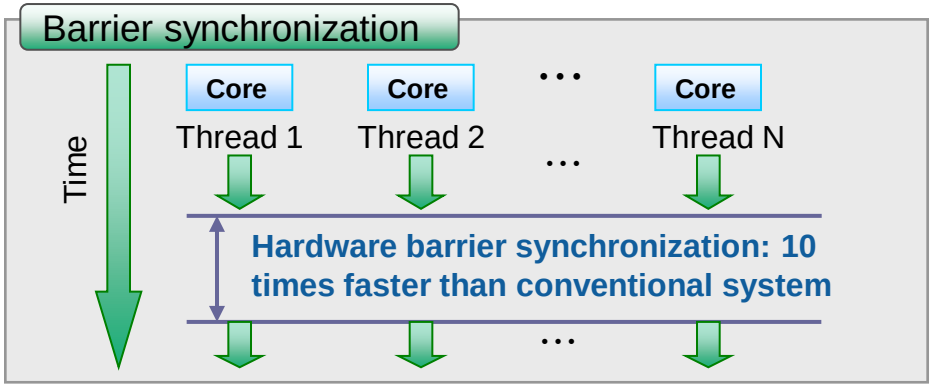


VISIMPACT™

(Virtual Single Processor by Integrated Multi-core Parallel Architecture)

FUJITSU

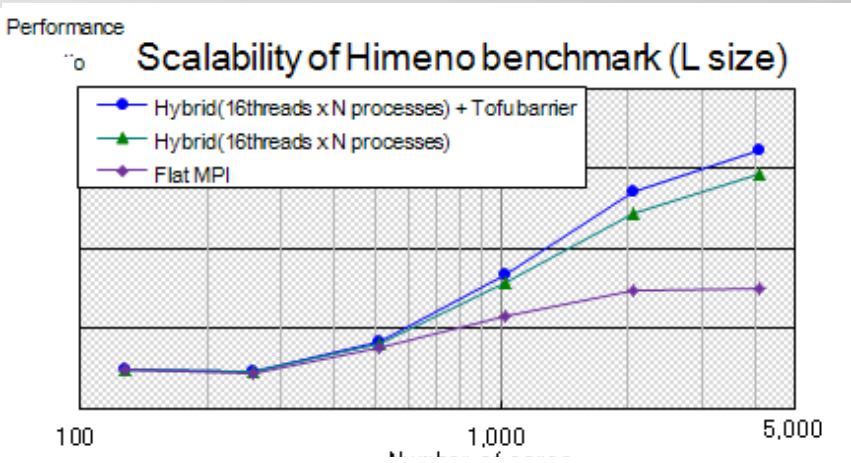
- Technologies to make multiple cores to single high speed CPU
  - Shared L2 cache memory to avoid false sharing
  - Inter-core hardware barrier facilities to reduce overhead of thread synchronization
- Thread parallel programs use these technologies

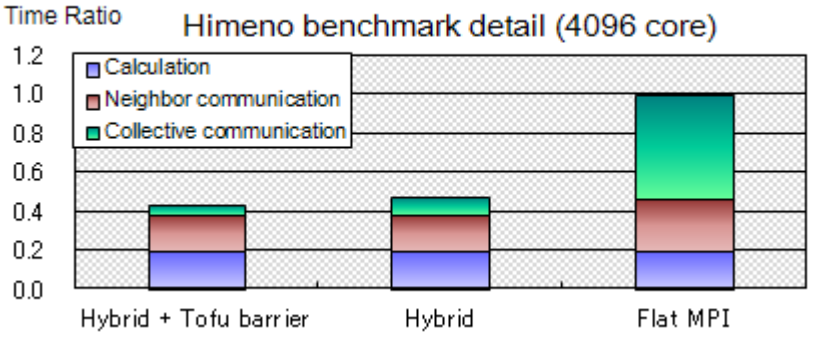


thread & process Hybrid-Parallel Programming

FUJITSU

- thread parallel in a chip
  - Auto-parallel
    or explicit parallel description using OpenMP™ API
  - VISIMPACT enables low overhead thread execution
- process parallel beyond CPU
  - MPI programming
  - Collective communication is tuned using Tofu barrier facility

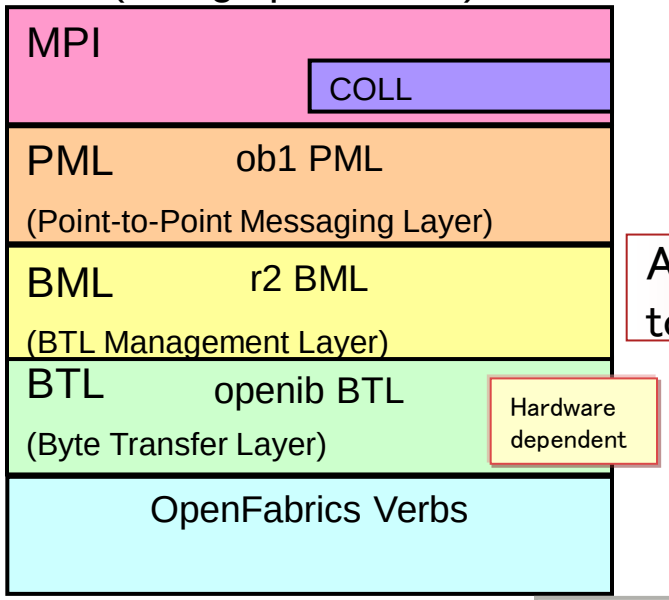




MPI Software stack

FUJITSU

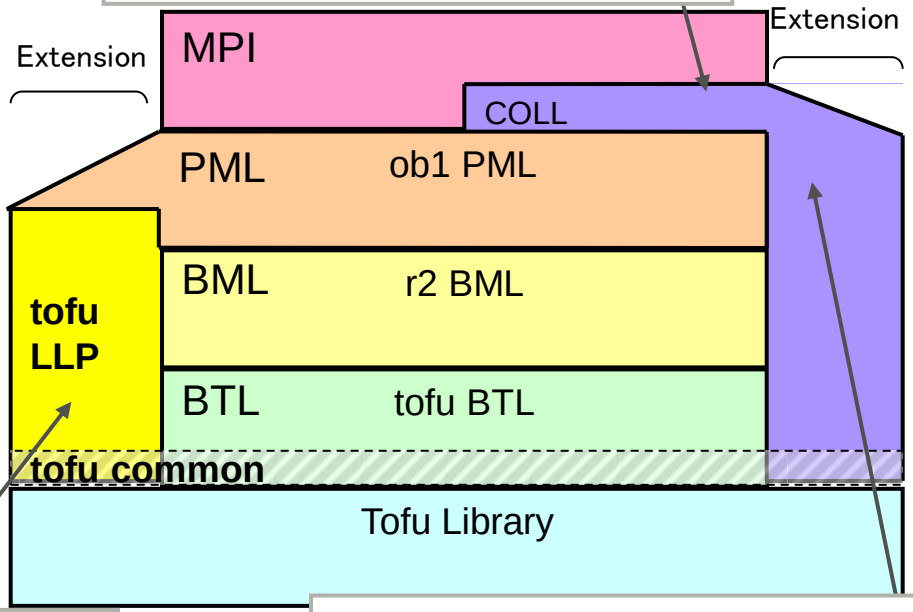
Original Open MPI Software Stack
(Using openib BTL)



Hardware dependent

Adapting to tofu

Supported special Bcast・Allgather・Alltoall・Allreduce for Tofu



LLP (Low Latency Path)

Special Hardware dependent layer
For Tofu Interconnect

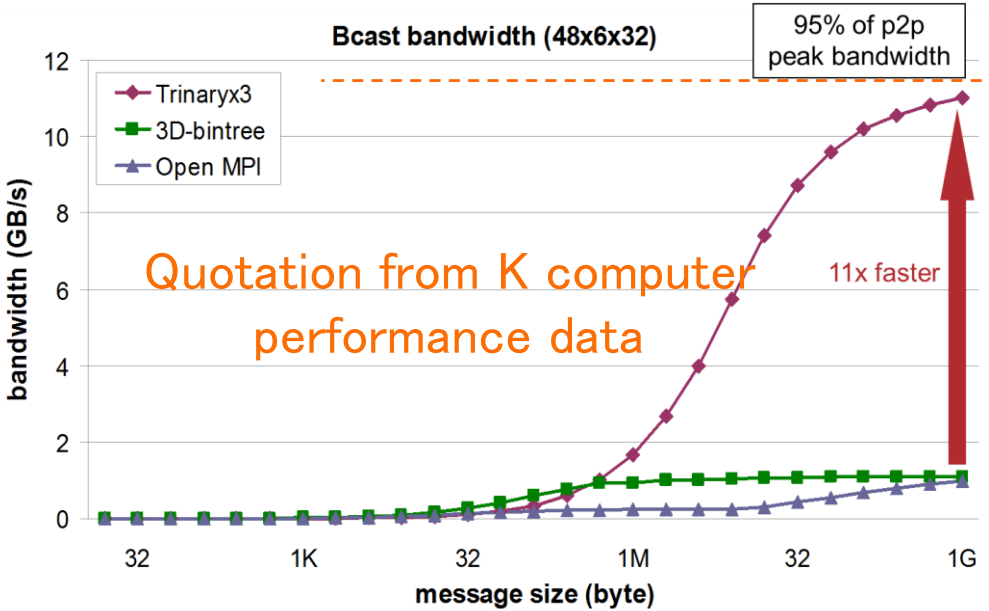
Providing Common Data processing and structures for BTL・LLP・COLL

Rendezvous Protocol Optimization etc

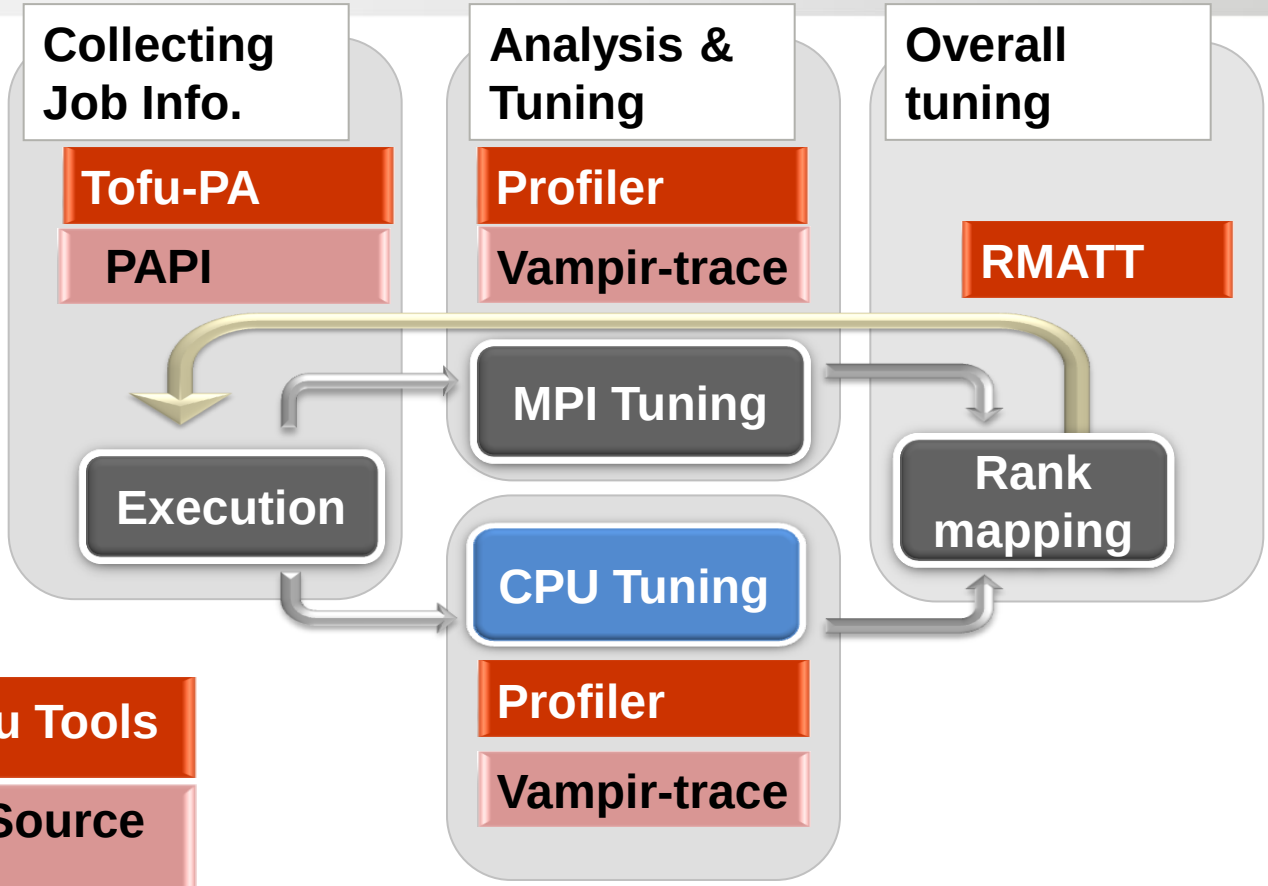
Customized MPI Library for High Scalability

FUJITSU

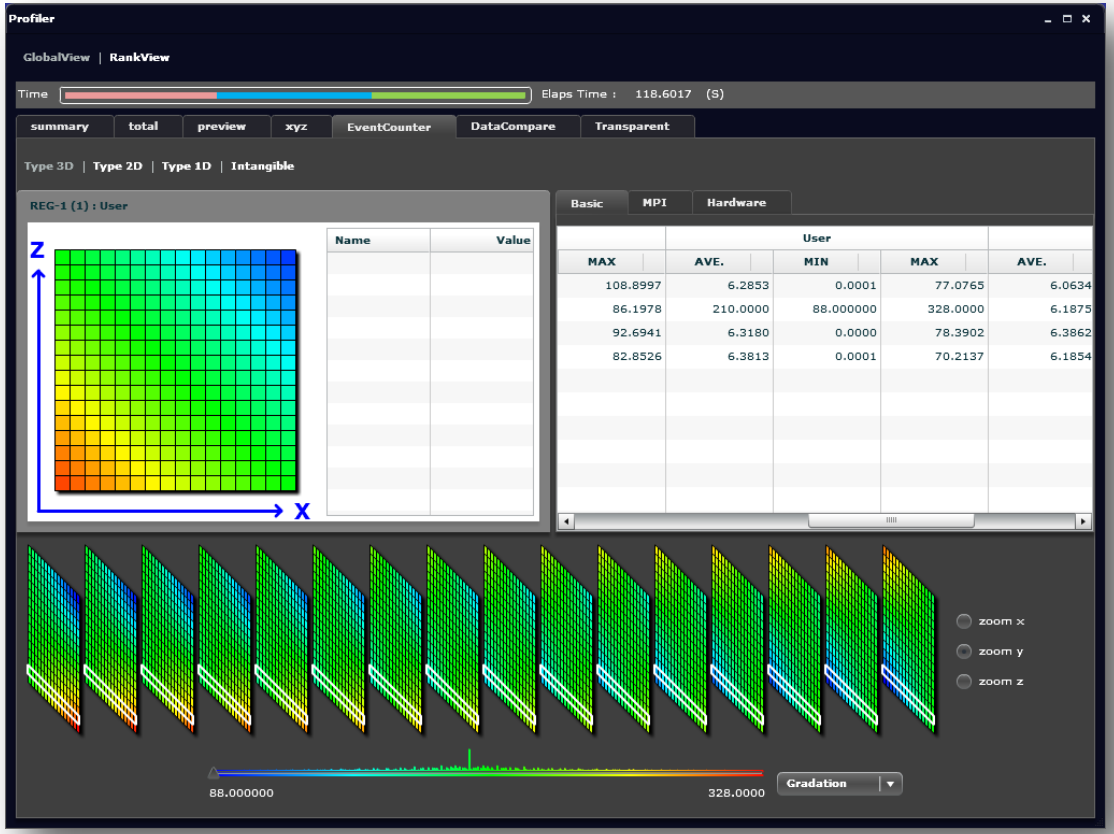
- Point-to-Point communication
  - Two methods for inside communication
  - The transfer method selection by
    - the data length
    - process location
    - number of hops
- Collective communication
  - Barrier, Allreduce, Bcast and Reduce use Tofu-barrier facility
  - Bcast, Allgather, Allgatherv, Allreduce and Alltoall use Tofu-optimized algorithm



Application Tuning Cycle and Tools



Program Profile (Event Counter Example)



- 3-D job example
  - Display 4096 procs in 16 x 16 x 16 cells
  - Cells painted in colors according to the proc status (e.g. CPU time)
  - Cut a slice of jobs along x-, y-, or z-axis to view

Rank Mapping Optimization (RMATT)

FUJITSU

Network Construction
Communication Pattern (Communication processing contents between Rank)

input

RMATT

output

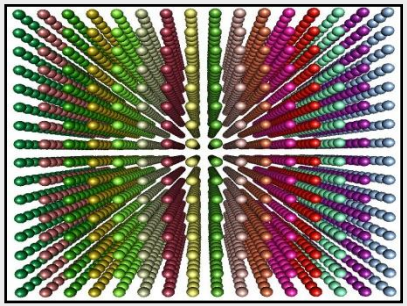
Optimized Rank Map
Reduce number of hop and congestion

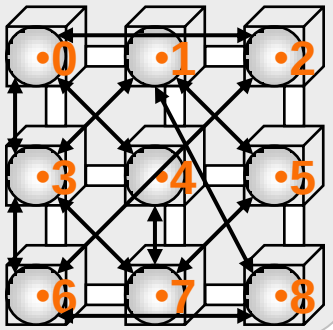
apply

Apply MPI_Allgather Communication Processing Performance

- Rank number : 4096 rank
- Network construction : 16x16x16 node (4096)

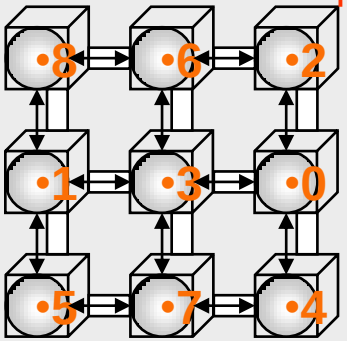
x,y,z order mapping 22.3ms

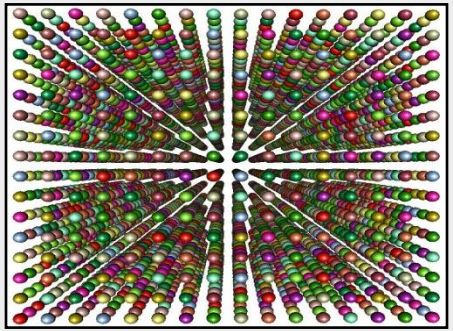


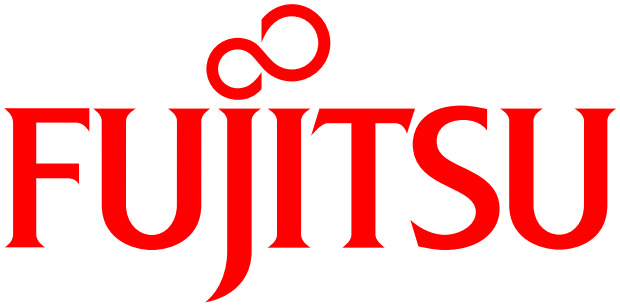


Remapping used RMATT 5.5ms

4 times performance Up







shaping tomorrow with you