FUJITSU

shaping tomorrow with you

Maximise Application Performance

Copyright 2013 FUJITSU LIMITED

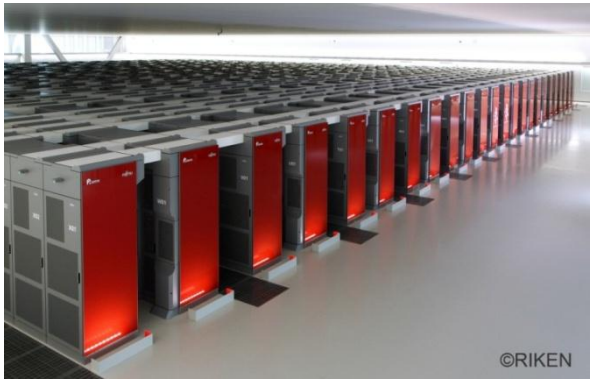
The K computer - Developed by RIKEN & Fujitsu

FUJITSU

System Overview

- Super-large-scale system – combining 88,128 processors
- Combination of advanced technology developed by Fujitsu
- Development was completed in June 2012.

K computer



The K computer in Kobe, Hyogo

System	Theoretical calculation speed : 11.28 petaflops(1) LINPACK performance : 10.51petaflops Processors: 88,128 Total memory: 1.26 petabytes
CPU	SPARC64™ VIIIfx (8 cores, 128 gigaflops)
Interconnect	6-dimensional mesh/torus topology (Tofu)

(1) FLOPS: the number of FLoating Operations Per Second. 1 petaflops is 10^{15} (1 quadrillion), 1 gigaflops is 10^{9} (1 billion) calculations per second

Copyright 2013 FUJITSU LIMITED

K computer and **PRIMEHPC FX10**

FUJITSU

- Fujitsu supercomputer w/ enhanced technology introduced for K computer

	K computer	PRIMEHPC FX10	Note
CPU	SPARC64 VIIIfx	SPARC64 IXfx	SPARC V9 + HPC-ACE
Peak perf.	128 GFLOPS	236.5 GFLOPS	
# of cores	8	16	
Memory	16GB	32GB/64GB	2GB/core~
BW	64GB/s	85GB/s	
Interconnect	6D mesh/torus	←	Tofu interconnect
System size	X x Y x 17	X x Y x 9	Z=0 is I/O node
link BW	5GB/s x bidirectional	←	

Copyright 2013 FUJITSU LIMITED

Software stack for Fujitsu supercomputers

FUJITSU

Applications

HPC Portal / System Management Portal

Technical Computing Suite

System Management

- System management
- System control
- System monitoring
- System operation support

Job Management

- Job manager
- Job scheduler
- Resource management
- Parallel

High Performance File System *FEFS*

- Lustre based high performance distributed file system
- High scalability, high reliability and availability

Automatic parallelisation compiler

- Fortran
- C
- C++

Tools and math. libraries

- Programming support tools
- Mathematical libraries

Parallel languages and libraries

- OpenMP
- MPI
- XPFortran

Linux based OS (enhanced for FX10)

PRIMEHPC FX10

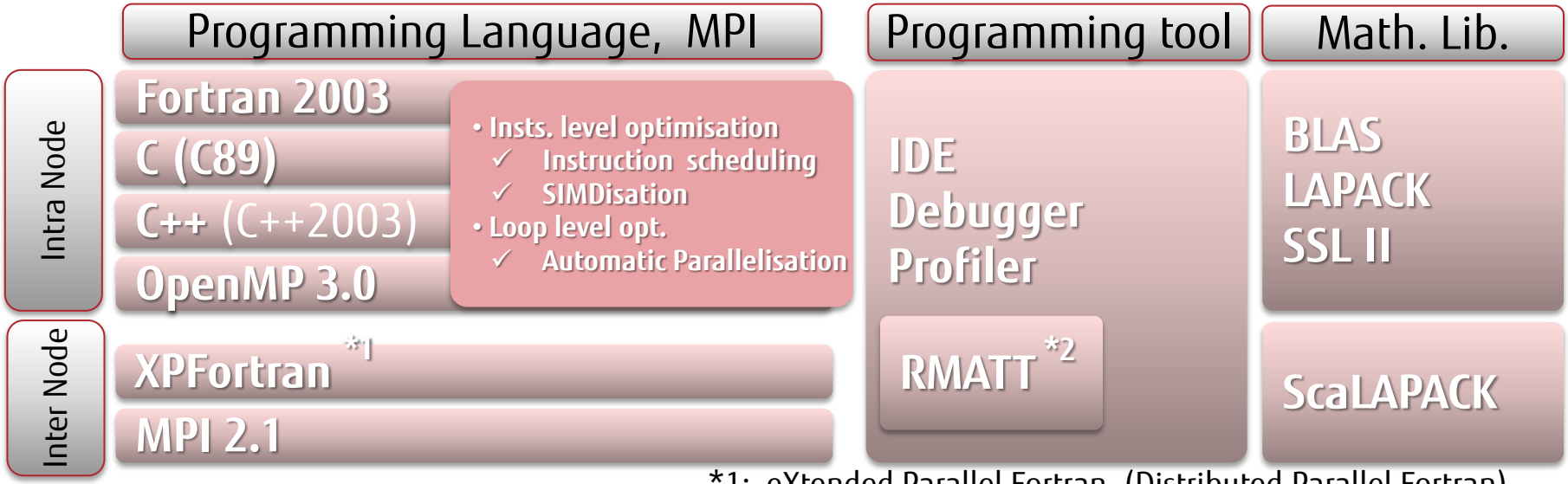
*Technical Computing Suite is also available for x86 clusters

Copyright 2013 FUJITSU LIMITED

Language system overview

FUJITSU

- Fortran C/C++ Compiler
- Programming model (OpenMP, MPI, XPFortran)
- Instruction level /Loop level optimisation using HPC-ACE
- Debugging and Tuning tools for massively parallel supercomputer



*1: eXtended Parallel Fortran (Distributed Parallel Fortran)
*2: Rank Map Automatic Tuning Tool

Copyright 2013 FUJITSU LIMITED

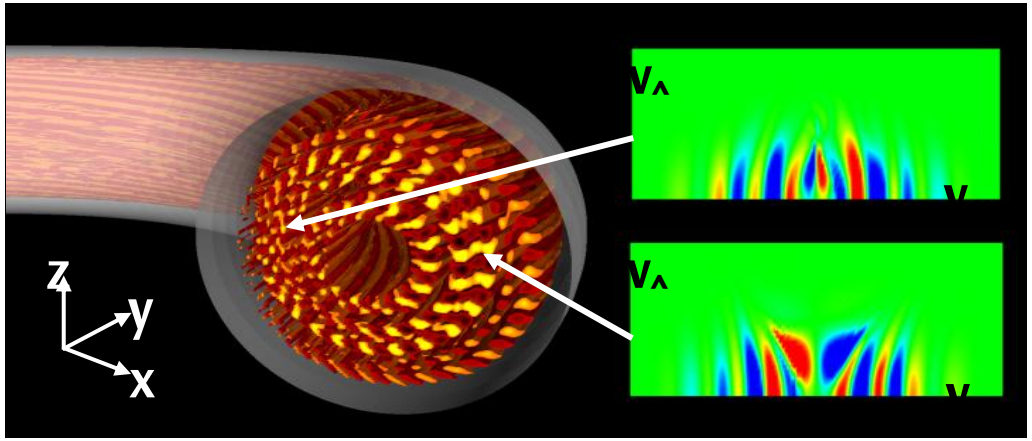
Application Results

FUJITSU

Simulation of turbulent fusion plasma

- Gyrokinetic Toroidal 5D Eulerian Code GT5D

  [Idomura et al., Comput. Phys. Commun (2008); Nuclear Fusion (2009)]



- Prediction of reactor performance limited by plasma turbulent transport
- Describe dynamics of fuel plasma distribution in 5D phase space
- Resolve from machine size ~1m to ion gyro-radius ~1mm

Copyright 2013 FUJITSU LIMITED

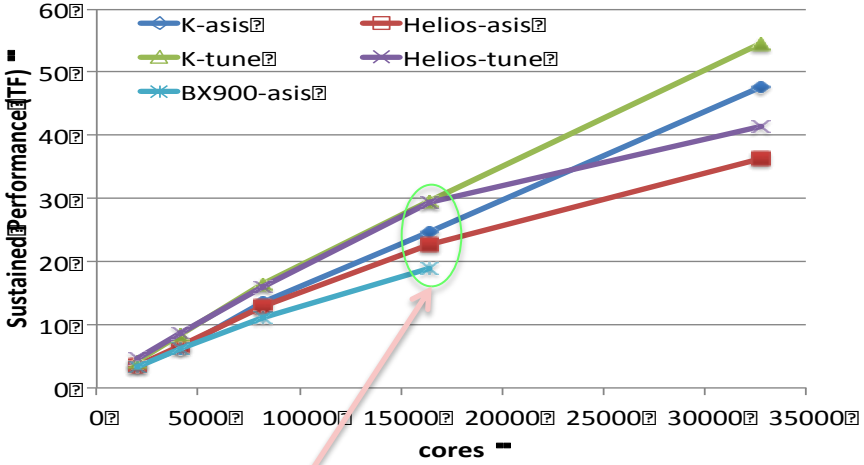
Strong Scheduling of GT5D code (JT-60 scale)

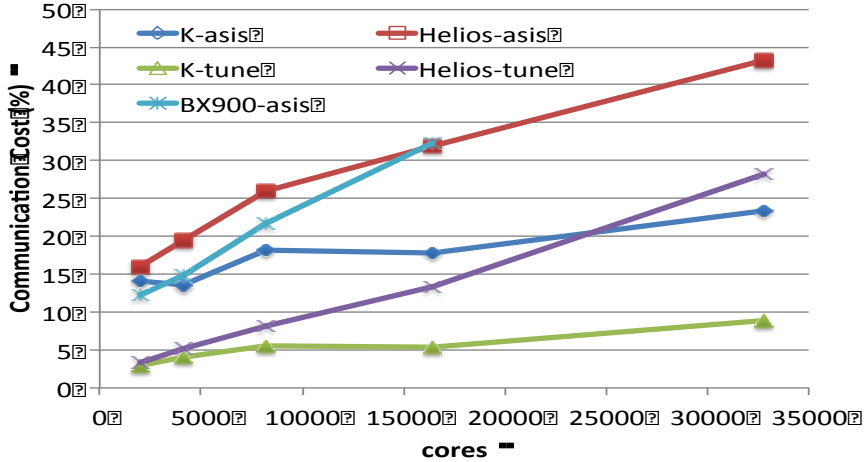
FUJITSU

Sustained performance (Tflops)

Communication cost ratio (%)

- JT-60U Scale : $(N_R, N_z, N_Z, N_{vl}, N_m) = (240, 64, 240, 128, 32) \sim 1.5 \times 10^{10}$





K 25TF(9.5%)→30TF(11.3%)
Helios 23TF(6.4%)→29TF(8.3%)
BX900(Fujitsu Blade Server) 19TF(9.8%)

※Data of K computer is provisional

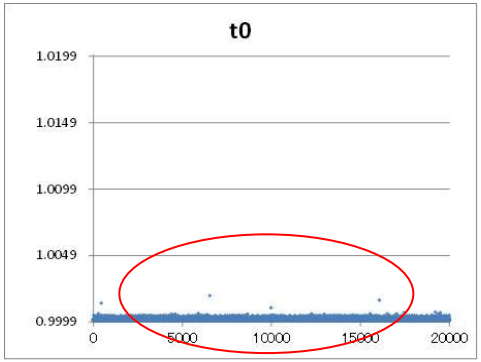
[Y. Idomura et al., Int. J. HPC Appl. in press] Presents by Y. Idomura in The Japan Atomic Energy Agency

Copyright 2013 FUJITSU LIMITED

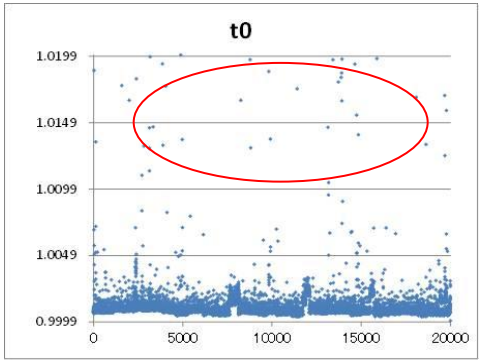
Reduce system jitter

- Optimised system software (incl. OS/FEFS) to minimise Operating System activity

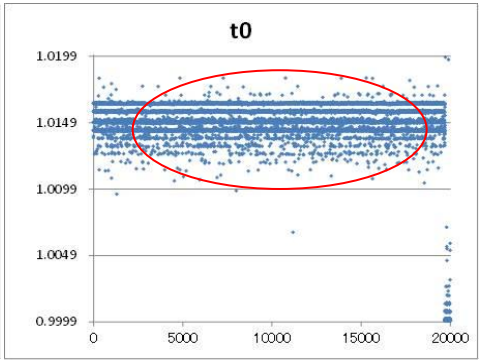
<FX10 + TCSuite>



<IA Cluster w/ TCSuite>

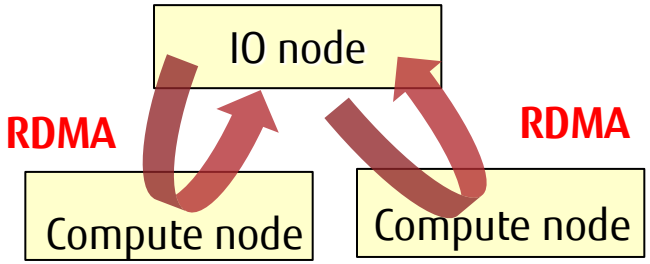


<IA Cluster w/o TCSuite>



e.g) Minimise Operating System jitter with RDMA of Tofu

a. Node / service health check
b. System information monitor (remote sadc)
c. Job information monitor (CPU time/used memory)



Copyright 2013 FUJITSU LIMITED

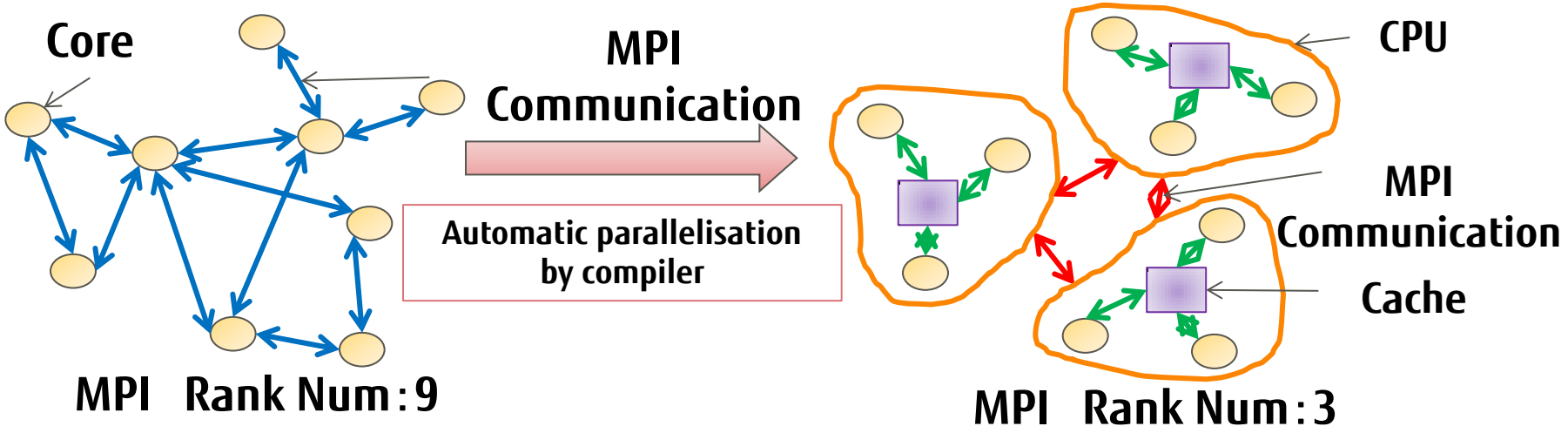
VISIMPACT – Hybrid parallelisation –

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

MPI program is automatically compiled to hybrid parallelisation (process and multithread) and executed

⇒ Communication overhead and memory usage reduced by MPI rank number decrease

⇒ Synchronisation overhead reduced by hardware barrier synchronisation between CPU cores



Inheriting and advancing vectorisation technology, corresponding to the Many core era

Copyright 2013 FUJITSU LIMITED

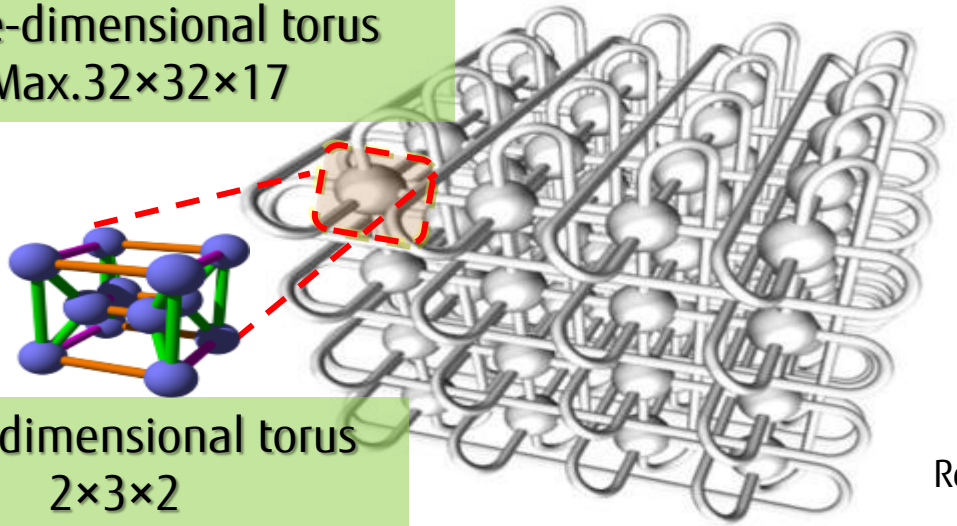
Network topology of Tofu interconnect

FUJITSU

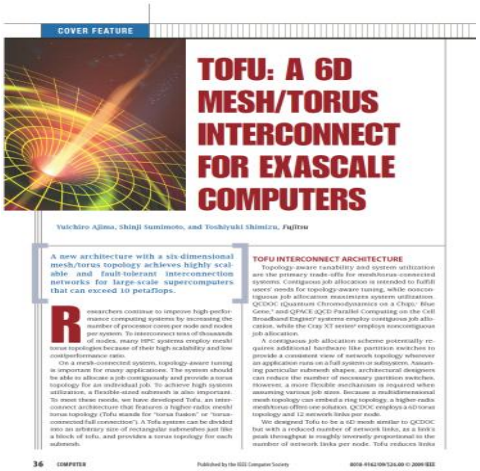
Six-dimensional mesh/torus direct connection network (MAX.32x32x17x2x3x2)

- Highly scalable compared to three-dimensional torus
- High operability and reliability
- Average number of hops, bisection bandwidth is improved with additional dimensions

Three-dimensional torus
Max.32×32×17



Three-dimensional torus
2×3×2


COVER FEATURE

TOFU: A 6D MESH/TORUS INTERCONNECT FOR EXASCALE COMPUTERS

Yuichiro Ajima, Shinji Sumimoto, and Toshiyuki Shimizu, Fujitsu

A new architecture with a six-dimensional mesh/torus topology achieves highly scalable and fault-tolerant interconnection networks for large-scale supercomputers that can exceed 10 petaflops.

TOFU INTERCONNECT ARCHITECTURE

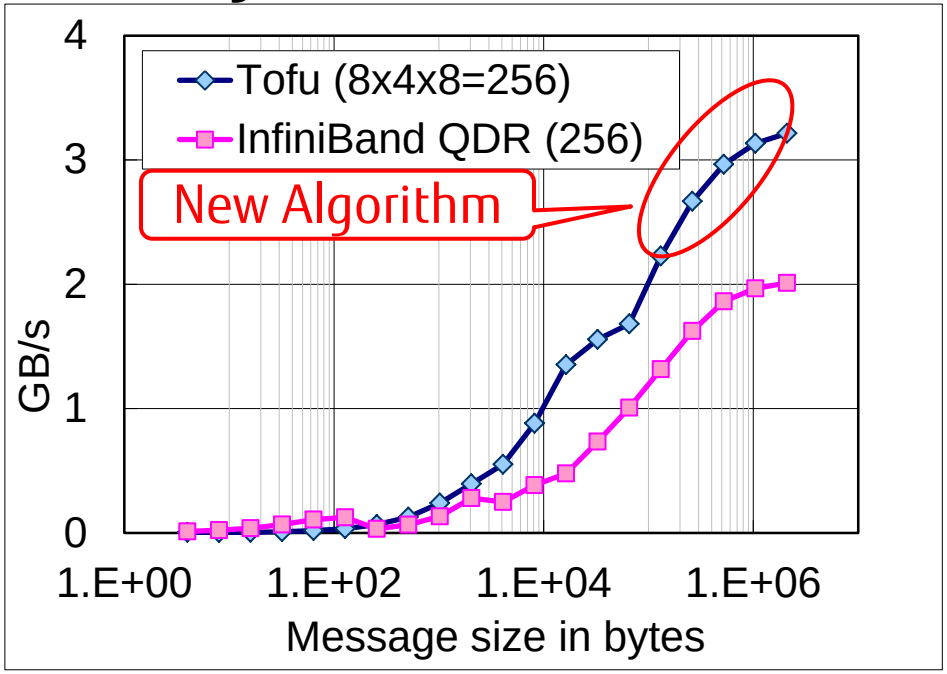
Research Paper of Tofu interconnect technology published in IEEE computer(2009)

Copyright 2013 FUJITSU LIMITED

Optimised alltoall communication of Tofu interconnect

FUJITSU

- Usage of uniform link is important for alltoall communication performance efficiency
- Development of a new algorithm to take advantage of Tofu
- Provide optimised library
- Surpassed Fat-Tree in measurement of 256 nodes (InfiniBand QDR)
  - Ease of porting applications



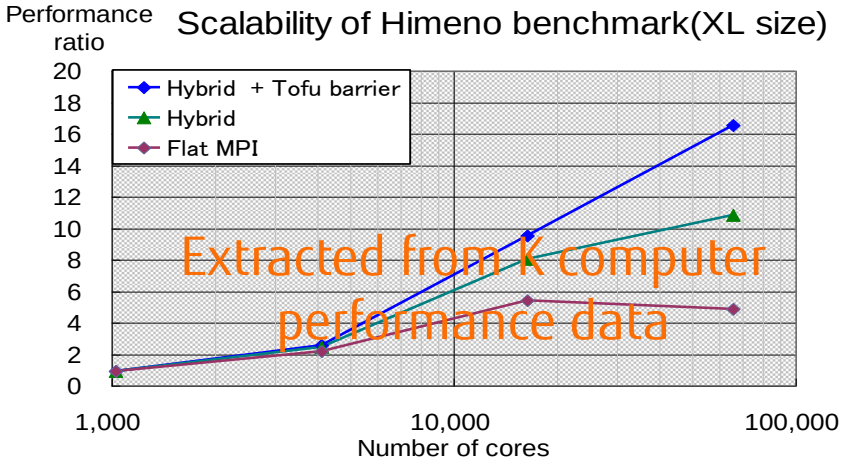
Copyright 2013 FUJITSU LIMITED

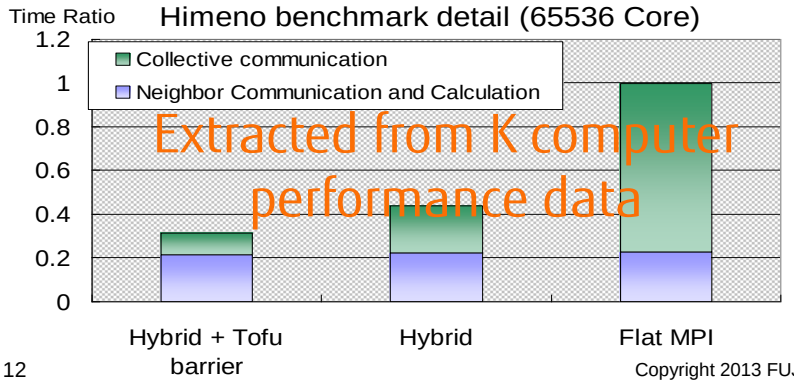
Programming Model for High Scalability

FUJITSU

Hybrid parallelisation by VISIMPACT and MPI library

- VISIMPACT
  - Automated multi-thread parallelisation
  - High performance thread synchronisation using Inter-core hardware barrier synchronisation function
- MPI library
  - High performance collective communications using Tofu barrier synchronisation function



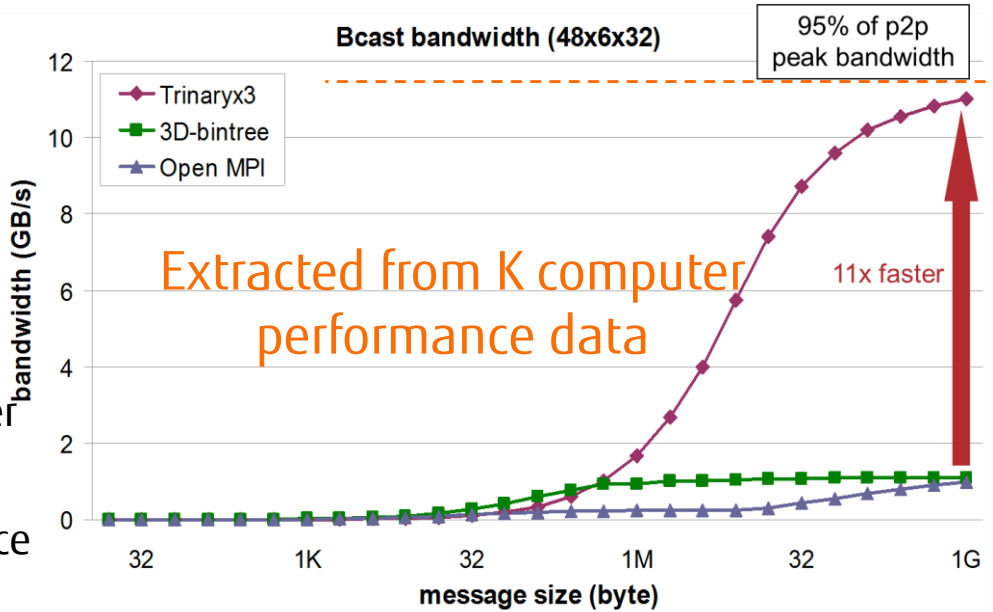


Copyright 2013 FUJITSU LIMITED

Customised MPI Library for High Scalability

FUJITSU

- Point-to-Point communication
  - Use special type of low-latency path that bypasses the software layer
  - Transfer method optimisation according to data length, process location and number of hops
- Collective communication
  - High performance Barrier synchronisation, Allreduce, Bcast and Reduce using Tofu barrier function
  - Scalable Bcast, Allgather, Allgatherv, Allreduce and Alltoall algorithm optimised for Tofu network

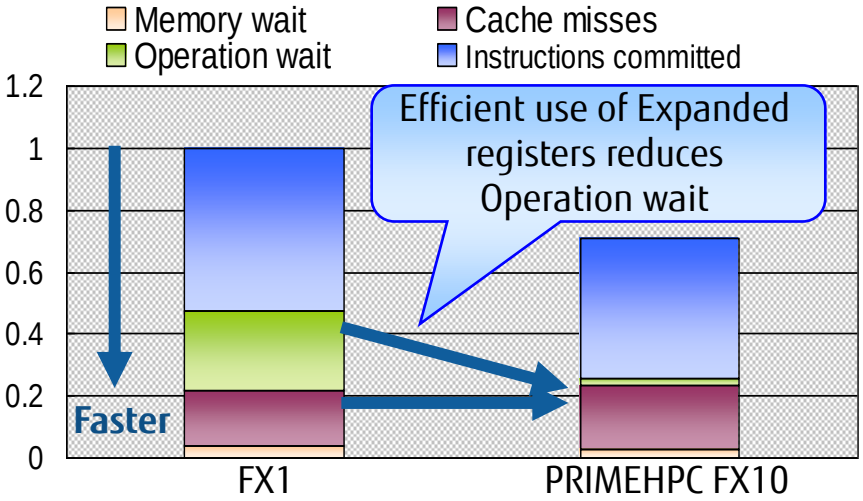


Copyright 2013 FUJITSU LIMITED

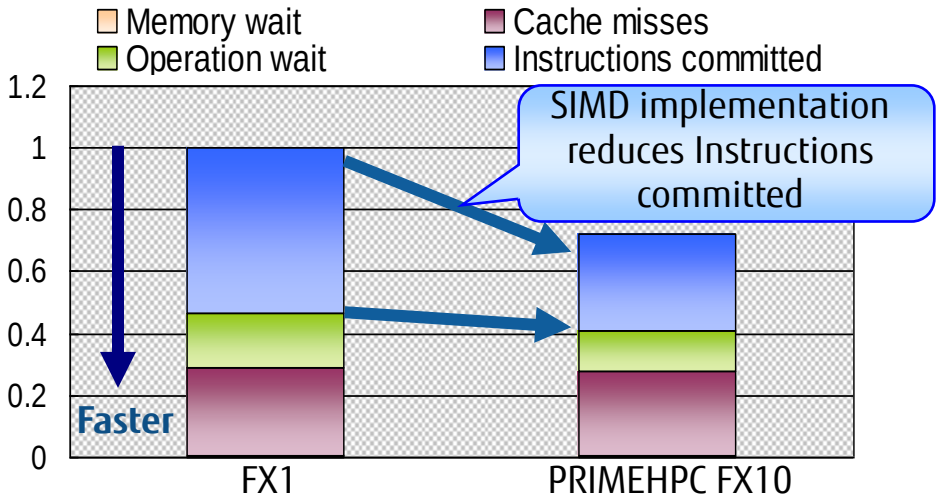
Compiler Optimisation for High Performance

- Instruction-level parallelism with SIMD instructions
- Improvement of computing efficiency using Expanded registers
- Improvement of cache efficiency using Sector cache

NPB3.3 LU
Execution time comparison (relative values)



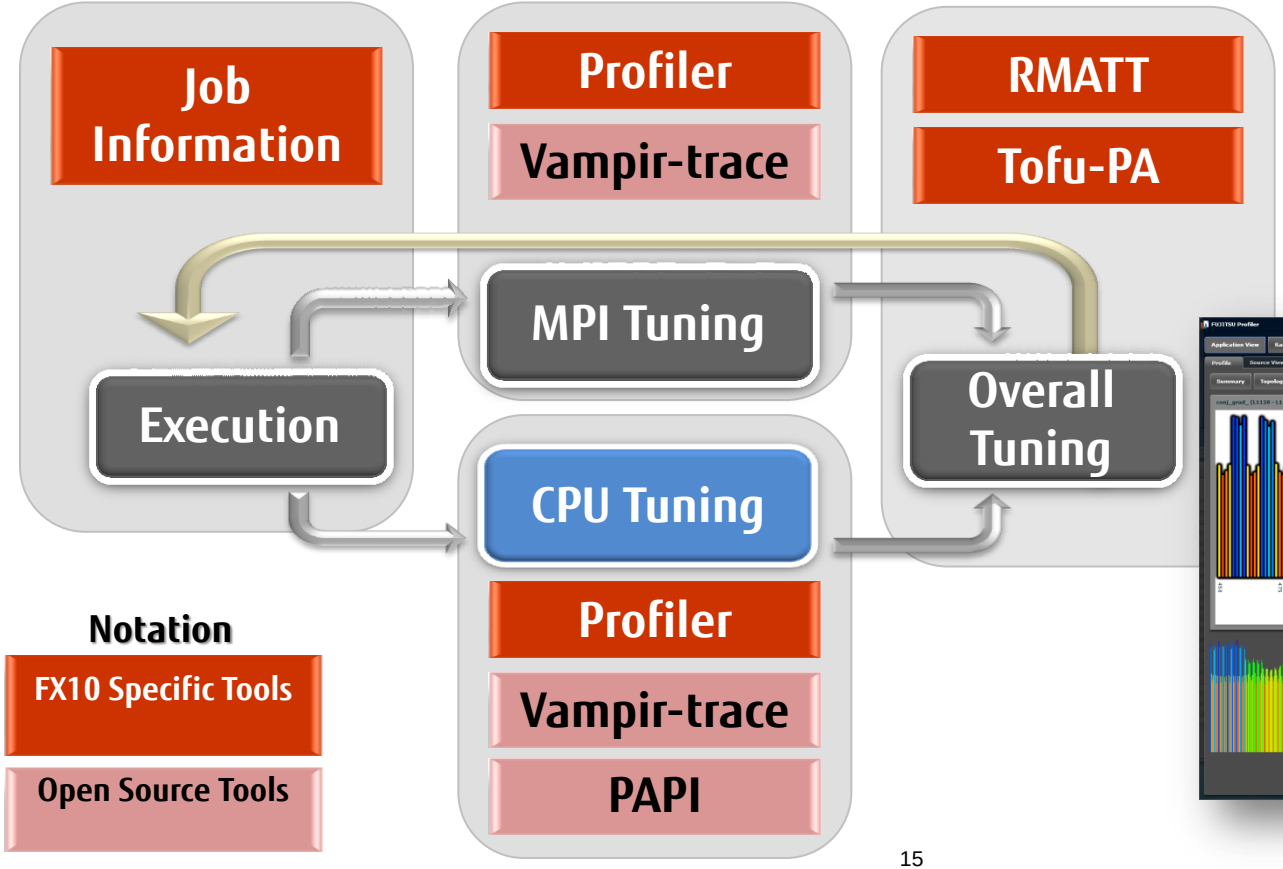
NPB3.3 MG
Execution time comparison (relative values)



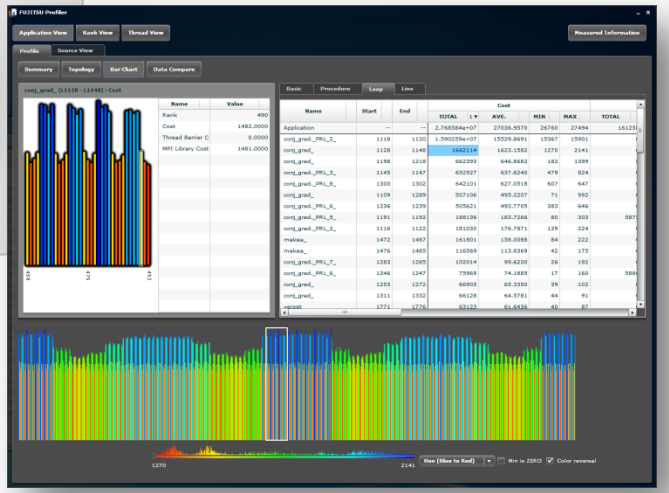
Copyright 2013 FUJITSU LIMITED

Application Tuning Cycle and Tools

FUJITSU



Profiler snapshot



Copyright 2013 FUJITSU LIMITED

Rank Mapping Optimisation (RMATT)

FUJITSU

Network Configuration
Communication Pattern (Communication processing contents between Rank)

input

RMATT

output

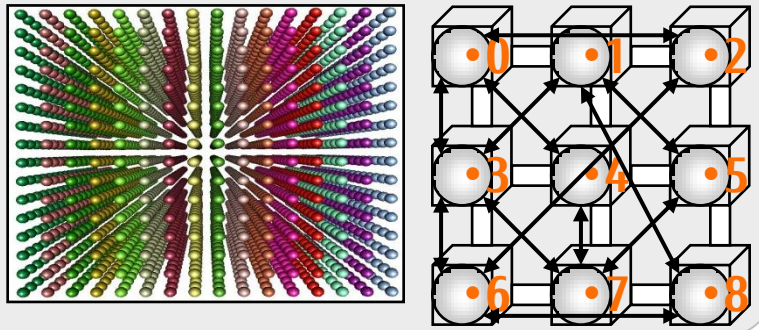
Optimised Rank Map
Reduce number of hops and congestion

apply

Apply MPI_Allgather Communication Processing Performance

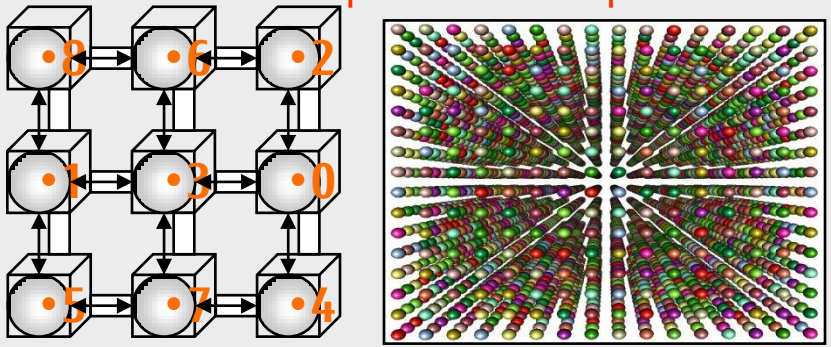
- Rank number : 4096 rank
- Network Configuration : 16x16x16 node (4096)

x,y,z order mapping 22.3ms

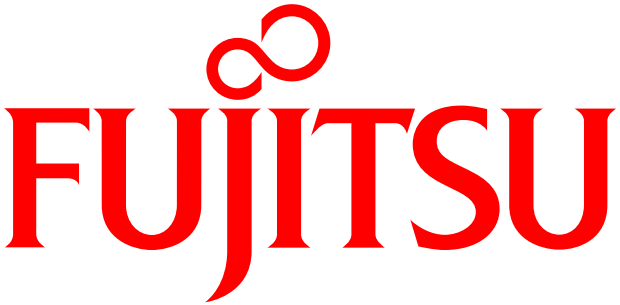


Remapping used RMATT 5.5ms

4 times performance Up



Copyright 2013 FUJITSU LIMITED



shaping tomorrow with you