SC12 Booth presentation

FUJITSU

shaping tomorrow with you

PRIMEHPC FX10
- Performance evaluations and approach towards the next -

Toshiyuki Shimizu
Next Generation Technical Computing Unit
FUJITSU LIMITED

November, 2012

Outline

FUJITSU

- Introduction of Fujitsu petascale supercomputers and key technologies of massively parallel computers
- K computer and FX10 performance evaluations
  - HPC-ACE: CPU architecture extension
  - VISIMPACT: Hybrid parallel execution support
- Challenges towards Exascale computing
- Summary

Fujitsu HPC from workplace to top-end

FUJITSU

PRIMEHPC FX10
Supercomputers

PRIMERGY x86 Clusters

Celsius
workstations

Customers of large-scale HPC systems

FUJITSU

Customer	Type	Peak
RIKEN (Kobe AICS)	**The K computer**	**11.28 PF**
National Computational Infrastructure, Australia (*)	**x86 Cluster (CX400), FX10**	**1.2 PF**
The University of Tokyo	**FX10**	**1.13 PF**
Central Weather Bureau of Taiwan (*)	**FX10**	**> 1 PF**
Kyushu University	**x86 Cluster (CX400), FX10**	**691 TF**
HPC Wales, UK	x86 Cluster	>300 TF
Japan Atomic Energy Agency	x86 Cluster, FX1, SMP	214 TF
Institute for Molecular Science	x86 Cluster (RX300), FX10	>168 TF
Institute for Molecular Science (*)	**x86 Cluster (CX400)**	>136 TF
Japan Aerospace Exploration Agency	FX1, SMP	>135 TF
RIKEN (Wako Lab. RICC)	x86 Cluster (RX200)	108 TF
Institute for Solid State Physics of the Univ. of Tokyo (*)	FX10	90 TF
NAGOYA University	x86 Cluster (HX600), FX1, SMP	60 TF
A*STAR, Singapore	x86 Cluster (BX900)	> 45 TF
A Manufacturer	x86 Cluster	>250 TF

(*) To be operated, Type definitions: FX10=PRIMEHPC FX10, x86 Cluster=Clusters based on PRIMERGY x86 server, SMP= SPARC Enterprise SMP server

Massively parallel computing products

FUJITSU

- Single CPU / node architecture for multicore CPU
  - FX1(4 cores) → K(8 cores) → FX10(16 cores)
  - High memory bandwidth balanced w/ CPU performance
- Key technologies for massively parallel are developed and inherited

FX1

CY2008～

40GF, 4-core CPU

K computer

CY2010～

128GF, 8-core CPU

PRIMEHPC FX10

CY2012～

236.5GF, 16-core CPU

Key technologies for massively parallel

FUJITSU

- Describe two technologies and show evaluation results of their effects by using real applications
- HPC-ACE: CPU architecture extension
  - Number of floating register extension
  - Floating-point reciprocal approximation instructions
  - Conditional execution (move, store, and mask generation) instructions
- VISIMPACT: Hybrid parallel execution model (process & thread) support
  - Automatic parallelization compiler
  - Hardware inter-core barrier and shared L2 cache

FUJITSU

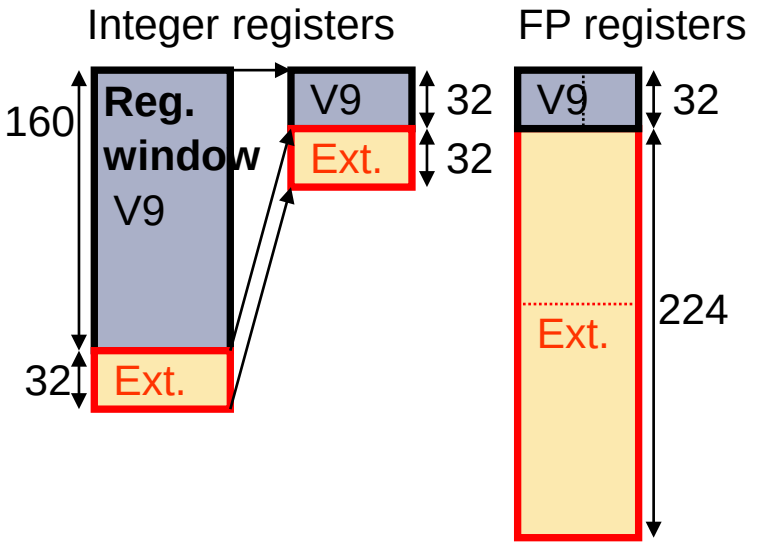
HPC-ACE extension and its effect

- # of register extension
- Conditional execution instructions
- Floating-point reciprocal approximations

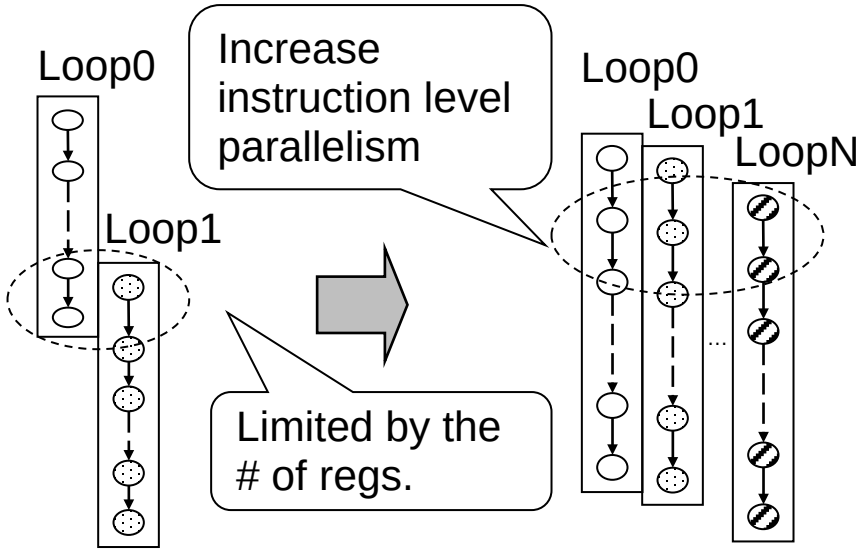
HPC-ACE: # of register extension

- Enhancement of SPARC V9 specification
  - # of integer regs from 32 to 64
  - # of DP FP regs from 32 to 256

    The extended registers can be used as same as standard registers with any instructions



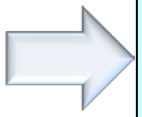
- Increasing # of register enables to explorer more parallelism
  - Large loops can be software pipelined



HPC-ACE: Conditional execution instructions

FUJITSU

```
subroutine sub(a, b, c, x, n)
 real*8  a(n), b(n), c(n), x(n)
 do i = 1 , n
   if ( x(i) .gt. 0.0 ) then
      a(i) = b(i) * c(i)
    else
      a(i) = b(i) - c(i)
   endif
 enddo
end
```



Conditional execution instructions

```
   :
.L100:
   :  calculations of b(i)*c(i) and b(i)-c(i)
fcmpgted,s      %f32,%f34,%f32
fselmovd,s      %f42,%f40,%f32,%f42
std,s           %f42,[%o5+%xg1]
add             %xg1,16,%xg1
bne,pt          %icc, .L100
nop
   :
```

- Conditional branch is eliminated
- SIMDization and software pipelining can be widely applied

HPC-ACE: Floating-point reciprocal approx.

FUJITSU

```
real(kind=8),dimension(1024) :: a, b, c
do i=1,1024
  a(i) = b(i) / sqrt(c(i))
end do
```

No reciprocal approximations

```
.LL1:
    ldd       [%o2 + %g2], %f2
    ldd       [%o1 + %g2], %f0
    subcc     %g3, 1, %g3
    add       %g2, 8, %g1
    fsqrtd    %f2, %f2
    fdivd     %f0, %f2, %f0
    std       %f0, [%o0 + %g2]
    mov       %g1, %g2
    bne,pt    %icc, .LL1
    nop
```

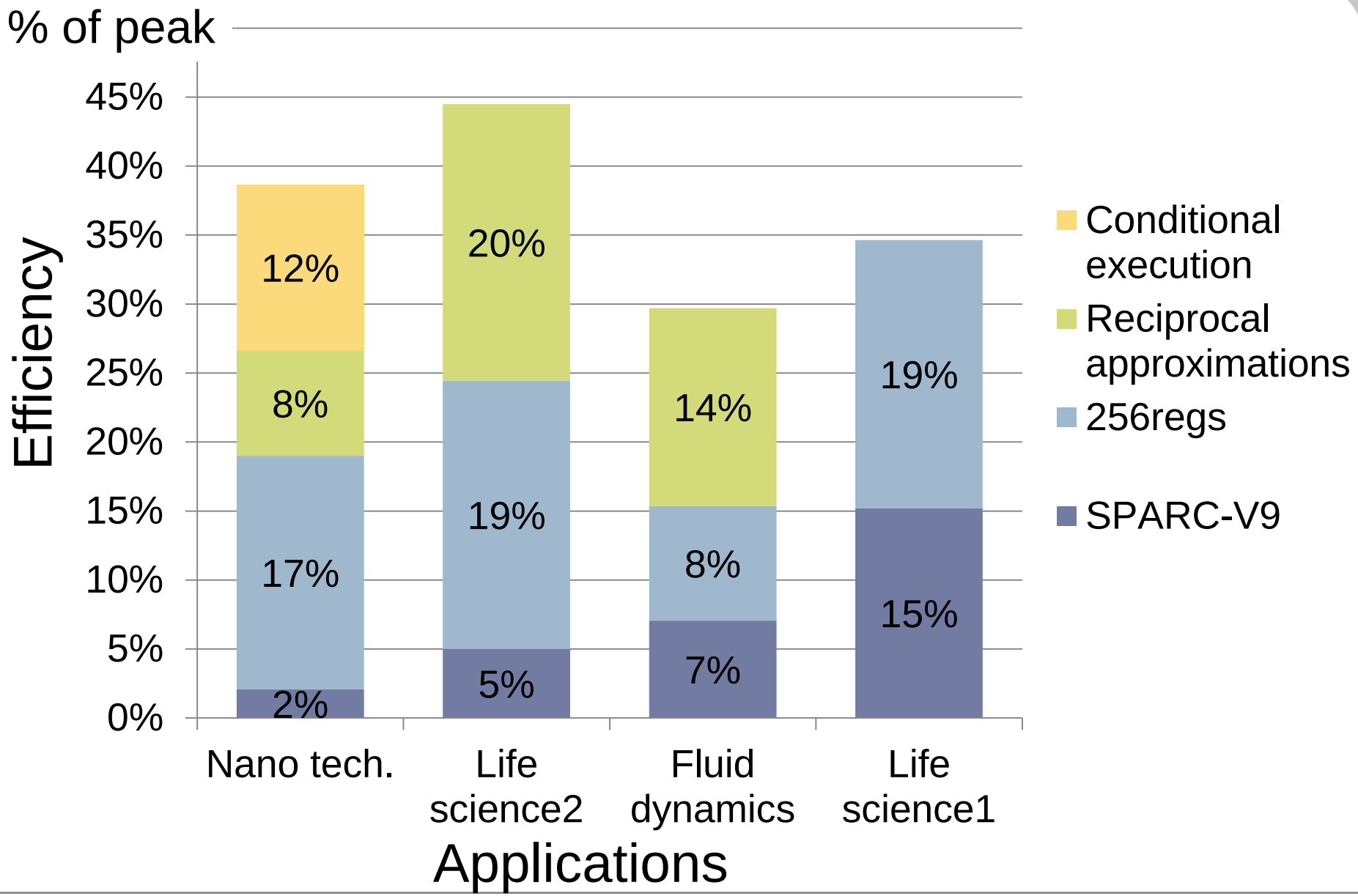
Reciprocal approximation instructions

```
.LL1:
    ldd       [%o2 + %g2], %f2
    ldd       [%o1 + %g2], %f0
    subcc     %g3, 1, %g3
    add       %g2, 8, %g1
    fmuld     %f8, %f2, %f10
    frsqrtad  %f2, %f2
    fmuld     %f10, %f2, %f14
    fnmsubd   %f2, %f14, %f8, %f14
    fmaddd    %f2, %f14, %f2, %f2
    fmuld     %f10, %f2, %f12
    fnmsubd   %f2, %f12, %f8, %f12
    fmaddd    %f2, %f12, %f2, %f2
    fmuld     %f10, %f2, %f10
    fnmsubd   %f2, %f10, %f8, %f6
    fcmpeqd   %f10, %f16, %f18
    fmaddd    %f2, %f6, %f2, %f2
    fselmovd  %f4, %f2, %f18, %f2
    fmuld     %f0, %f2, %f0
    std       %f0, [%o0 + %g2]
    mov       %g1, %g2
    bne,pt    %icc, .LL1
    nop
```

- Limited overlapping of instructions

- Divide and Sqrt are pipelined
- Software pipelining can also be applied

HPC-ACE improves application efficiency



Performance efficiency of real applications

FUJITSU

- Measured by petascale K computer and PRIMEHPC FX10
- K computer & FX10 runs broad applications efficiently

Applications	# of Cores	Efficiency	IPC	System (Peak)
A*	98,304	30%	1.7	K(1.57PF)
B*	98,304	41%	1.5	
C*	98,304	32%	1.3	
D*	98,304	27%	1.0	
E*	98,304	12%	0.6	
F*	98,304	52%	1.5	
G*	98,304	9%	0.7	
H*	20,480	3%	0.6	K(0.33PF)
I*	32,768	23%	0.9	K(0.52PF)
J*	98,304	40%	1.5	K(1.57PF)
K	6,144	30%	1.6	FX10(0.1PF)

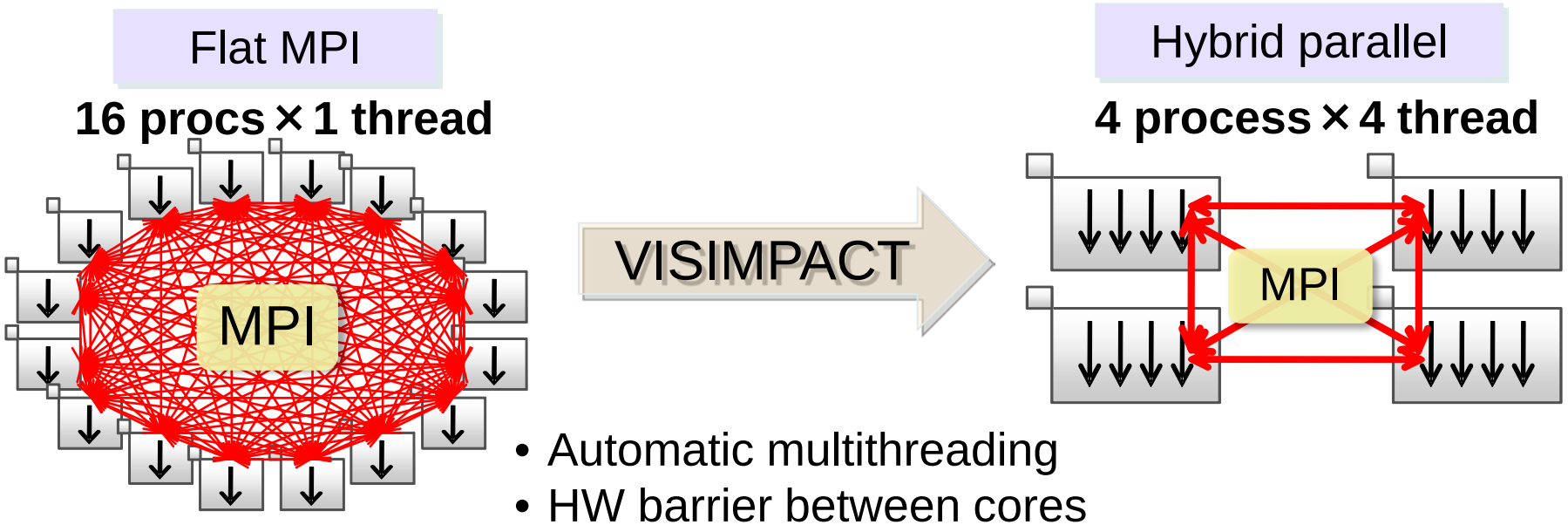
* Measured by K computer: Results are from trial use & not the final.

FUJITSU

VISIMPACT and its effect

- Hybrid execution model support
- Scalability and efficiency
- Load imbalance

- Hybrid parallel execution is preferable for
  - Good scalability (reduce communications by reducing the # of processes)
  - Larger usable memory per process
  - Efficient use of memory (smaller work area)
- But…
  - Programing is harder due to two level of parallelization



- Automatic multithreading
- HW barrier between cores
- Shared L2 cache

Evaluations with practical meteorological apps.

FUJITSU

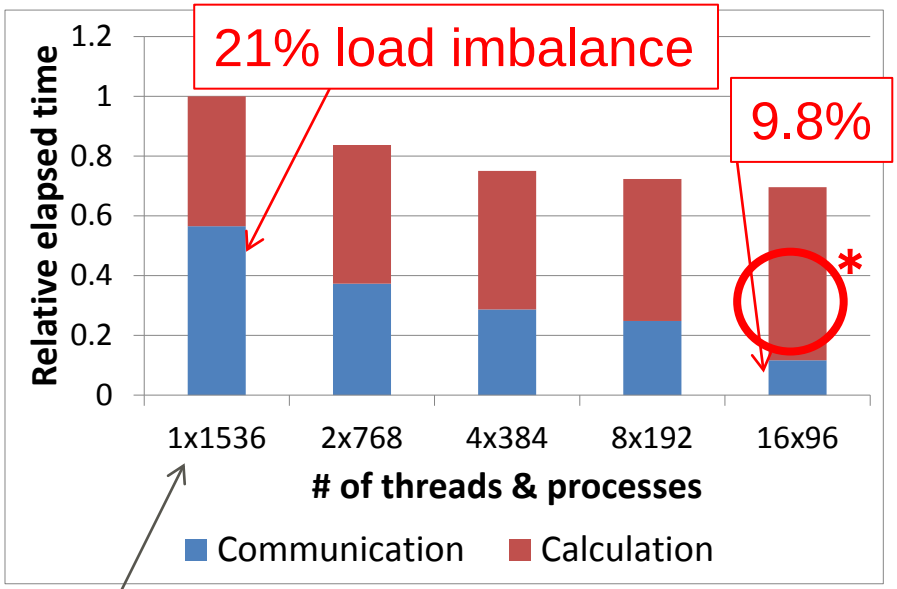
- WRF and COSMO were evaluated
  - Operational weather forecasting codes
  - Regional models
- Different computational characteristics
  - WRF uses “IO Quilting” mechanism
  - COSMO is memory intensive
  - WRF is thread-parallelized by OpenMP, but COSMO is not

Advantage of hybrid parallelization on FX10

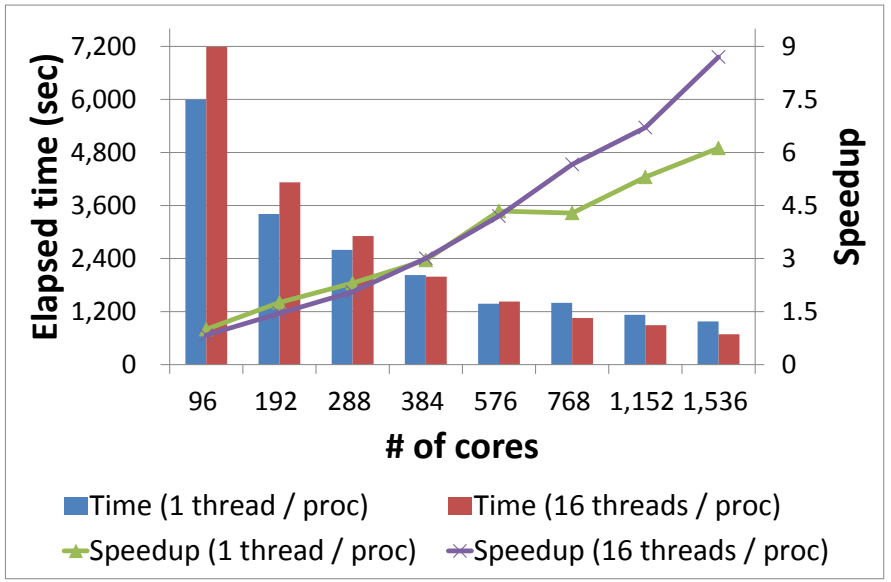
FUJITSU

WRF V3.3.1

Reduction of Communication Cost



Difference of Node Scalability



(WRF V3.3.1, 1200x700x45 grids, by courtesy of CWB)

- Flat-MPI (1x1536) spends much time for communication
- The communication includes load imbalance between processes
  - Load imbalance ratio† of Flat-MPI is 21%, while 9.8% for 16x96
  - * Load imbalance between threads in a process is in the calculation time
- Node scalability of hybrid execution (16 threads/proc) is better

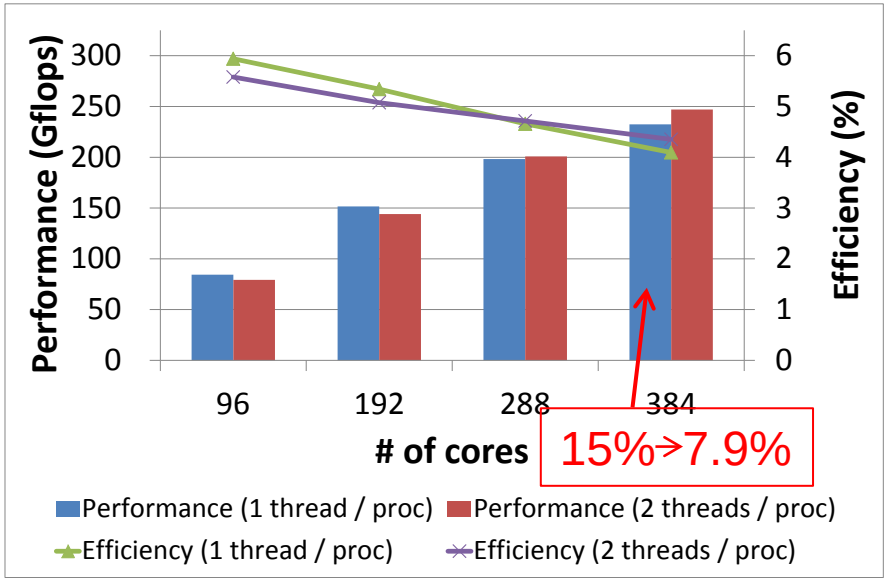
†:*LoadImbalanceRatio*＝(*MaxCalculationTime*÷*AvgCalculationTime*)－1

Automatic thread-parallelization on FX10

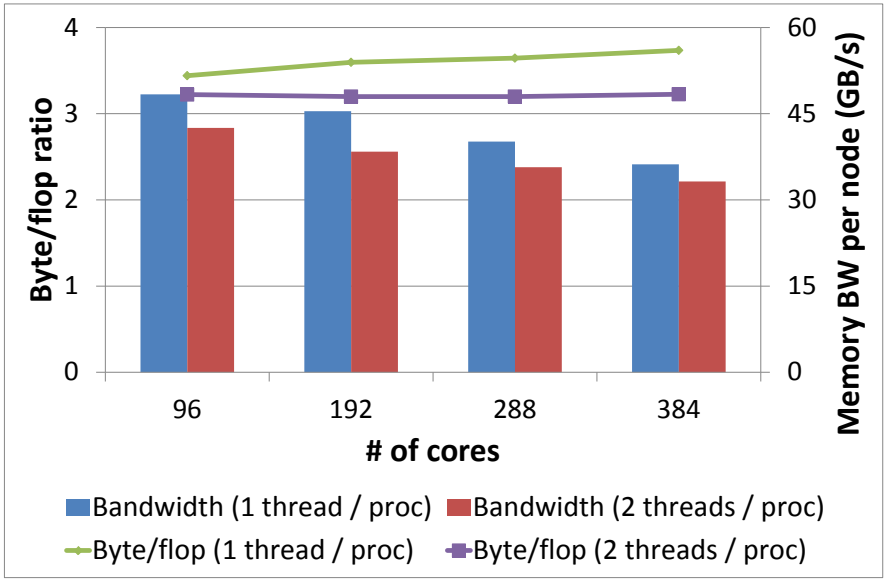
FUJITSU

COSMO-DE(RAPS 5.0)

Sustained Performance & Efficiency



Effective B/F Ratio & Memory BW



(COSMO RAPS 5.0, 421x461x50 grids)

- VISIMPACT parallelizes with threads and improves performance when cores are increased
- For 384 cores, load imbalance ratio† was mitigated from 15% to 7.9%
- Thread parallelization reduces byte/flop and utilizes memory BW

†:*LoadImbalanceRatio* = (*MaxCalculationTime* ÷ *AvgCalculationTime*) − 1

Summary of VISIMPACT (hybrid execution)

FUJITSU

- Real applications run on the hybrid parallel mode using VISIMPACT show better performance than Flat MPI
  - By eliminating memory copy
  - By reducing load imbalance between processes
- Hybrid parallel execution model is less memory usage
  - Intrinsic and indispensable features for many core environment
- Reducing a load imbalance between threads in a process should be studied

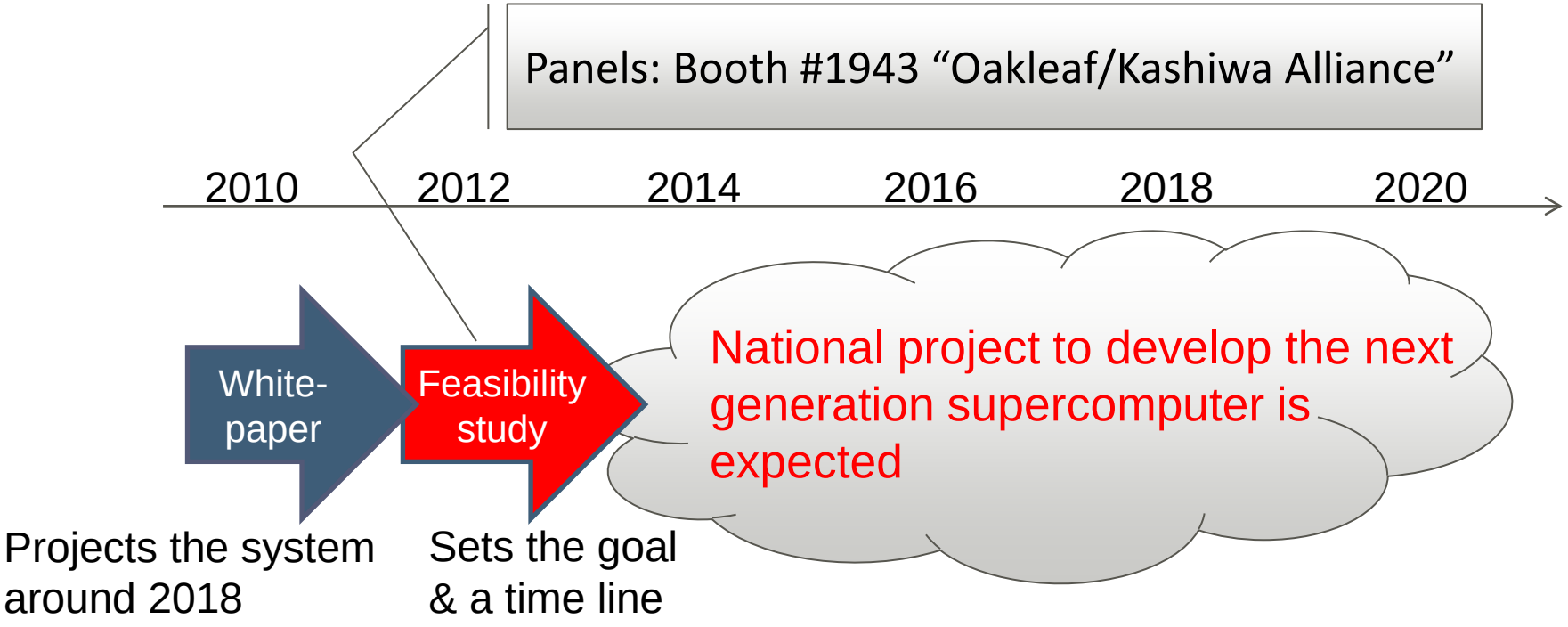
FUJITSU

Challenges towards Exascale computing

Challenges toward exascale computing

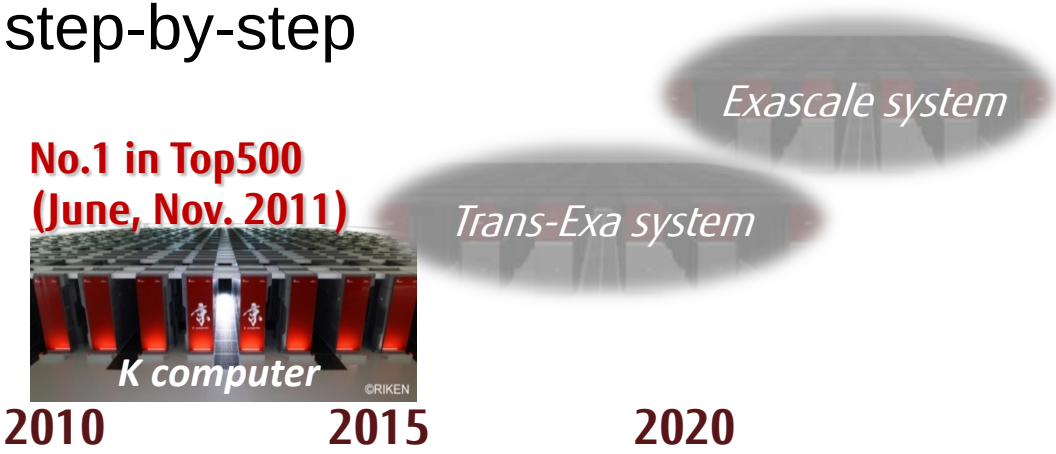
FUJITSU

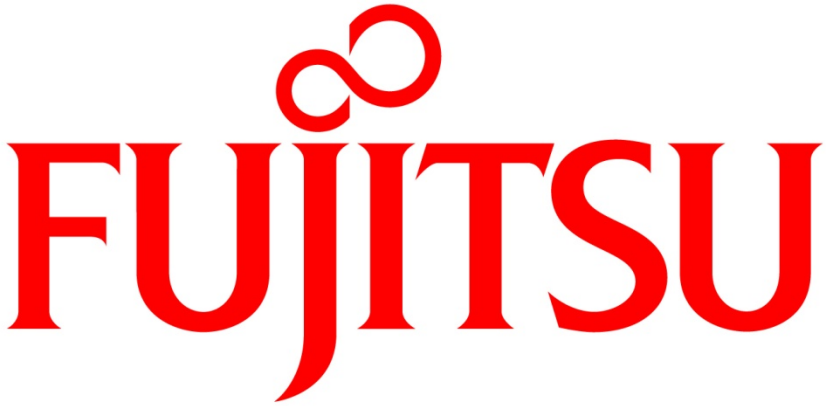
- Fujitsu is developing a 100 Petaflops capable system as a midterm goal
- Participates two consecutive national projects for exascale
  - Fujitsu contributed to develop the whitepaper
    - "Report on Strategic Direction/Development of HPC in Japan"
  - Fujitsu has started two-year feasibility study to set the goal & schedule since July 2012



Summary

- K computer & FX10 employ massively parallel technologies
  - HPC-ACE
  - VISIMPACT
- Evaluation running real applications
  - Good performance efficiency
  - Good scalability and smaller load imbalance
- Research and develop massively parallel architecture and approach toward exascale step-by-step





shaping tomorrow with you