Oakforest-PACS (OFP) : the largest scale many-core supercomputer in Japan

Taisuke Boku
Center for Computational Sciences, University of Tsukuba

(with courtesy by Toshihiro Hanawa
Information Technology Center, the University of Tokyo)

Outline

- Supercomputer deployment plan in Japan
- What is JCAHPC ?
- Supercomputer procurement in JCAHPC
- System specification
- Pre-evaluation by KNC (not KNL)
- Conclusions

Deployment plan of 9 supercomputing center (Oct. 2015 + latest)

Fiscal Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Hokkaido		HITACHI SR16000/M1 (172TF, 22TB) Cloud System BS2000 (44TF, 14TB) Data Science Cloud / Storage HA8000 / WOS7000 (10TF, 1.96PB)			3.2 PF (UCC + CFL/M) 0.96MW 0.3 PF (Cloud) 0.36MW					30 PF (UCC + CFL-M) 2MW		
Tohoku		NEC SX-9他 (60TF)			SX-ACE(707TF,160TB, 655TB/s) LX406e(31TF), Storage(4PB), 3D Vis, 2MW				~30PF, ~30PB/s Mem BW (CFL-D/CFL-M) ~3MW			
Tsukuba		HA-PACS (1166 TF) COMA (PACS-IX) (1001 TF)			PACS-X 10PF (TPF) 2MW							
Tokyo		Fujitsu FX10 (1PFlops, 150TiB, 408 TB/s), Hitachi SR16K/M1 (54.9 TF, 10.9 TiB, 28.7 TB/s)			Oakforest-PACS 25 PF (UCC + TPF) 4.2 MW	Reedbush 1.8～1.9 PF 0.7 MW		50+ PF (FAC) 3.5MW	100+ PF 4.5MW (UCC + TPF)	200+ PF (FAC) 6.5MW		
Tokyo Tech.		TSUBAME 2.5 (5.7 PF, 110+ TB, 1160 TB/s), 1.4MW			TSUBAME 2.5 (3～4 PF, extended)	TSUBAME 3.0 (20 PF, 4~6PB/s) 2.0MW (3.5, 40PF at 2018 if upgradable)			TSUBAME 4.0 (100+ PF, >10PB/s, ~2.0MW)			
Nagoya		FX10(90TF) CX400(470TF) SGI UV2000 (24TF, 20TiB)			Fujitsu FX100 (2.9PF, 81 TiB) Fujitsu CX400 (774TF, 71TiB) 2MW in total		50+ PF (FAC/UCC + CFL-M) up to 4MW			100+ PF (FAC/UCC+CFL-M) up to 4MW		
Kyoto		Cray: XE6 + GB8K + XC30 (983TF) Cray XC30 (584TF)			7-8 PF(FAC/TPF + UCC) 1.5 MW			50-100+ PF (FAC/TPF + UCC) 1.8-2.4 MW				
Osaka		NEC SX-ACE (423TF) NEC Express5800 (22.4TF)			0.7-1PF (UCC)		3.2PB/s, 5-10Pflop/s, 1.0-1.5MW (CFL-M)			25.6 PB/s, 50-100Pflop/s, 1.5-2.0MW		
Kyushu		HA8000 (712TF, 242 TB) SR16000 (8.2TF, 6TB) FX10 (272.4TF, 36 TB) CX400 (966.2 TF, 183TB) 2.0MW			15-20 PF (UCC/TPF) 2.6MW FX10 (90.8TFLOPS)				100-150 PF (FAC/TPF + UCC/TPF) 3MW			

Power consumption indicates maximum of power supply (includes cooling facility)

T2K Open Supercomputer Systems

Open Supercomputer T2K
Open Supercomputer Alliance

- Same timing of procurement for next generation supercomputers in three universities
- Academic leadership for computational science/engineering in research/education/grid-use on same platform

- *Open* hardware architecture with commodity devices & technologies.
- *Open* software stack with open-source middleware & tools.
- *Open* to user's needs not only in FP & HPC field but also INT world.

Kyoto Univ.
416 nodes (61.2TF) / 13TB
Linpack Result:
Rpeak = 61.2TF (416 nodes)
Rmax = 50.5TF

Univ. Tokyo
952 nodes (140.1TF) / 31TB
Linpack Result:
Rpeak = 113.1TF (512+256 nodes)
Rmax = 83.0TF

Univ. Tsukuba
648 nodes (95.4TF) / 20TB
Linpack Result:
Rpeak = 92.0TF (625 nodes)
Rmax = 76.5TF

From T2K to Post-T2K

- Effect of T2K Alliance
  - Three supercomputers are introduced at the same time, sharing wide knowledge for system construction and commodity technology, followed by academic research collaboration among these players
- After T2K, three universities had different time of new system procurement
  - Kyotop U.: four year period of procurement
  - U. Tsukuba: accelerated computing
  - U. Tokyo: T2K + Fujitsu FX10 and other systems
- Post-T2K (with two ”Ts”)
  - in 2013, U. Tsukuba and U. Tokyo collaborated again for new supercomputer procurement in ***much more tight framework***

JCAHPC

- Joint Center for Advanced High Performance Computing (http://jcahpc.jp)
  （最先端共同HPC基盤施設）
- Very tight collaboration for “post-T2K” with two universities
  - For main supercomputer resources, *uniform specification* to ***single shared system***
  - Each university is financially responsible to introduce the machine and its operation
    -> unified procurement toward single system with ***largest scale in Japan***
  - To manage everything smoothly, a joint organization was established
    -> JCAHPC

History of JCAHPC

- March 2013: U. Tsukuba and U. Tokyo exchanged agreement for "JCAHPC establishment and operation"
  - Center for Computational Sciences, University of Tsukuba and Information Technology Center, University of Tokyo
- April 2013: JCAHPC started
  - 1st period director: Mitsuhisa Sato (Tsukuba), vice director: Yutaka Ishikawa (Tokyo)
  - 2nd period (2016~) director: Hiroshi Nakamura (Tokyo), vice director: Masayuki Umemura (Tsukuba)
- July 2013: RFI for procurement
  - at this time, the joint procurement style was not fixed
    -> then a single system procurement was decided
  - to give enough time for very advanced technology for processor, network, memory, etc., more than 1 year of period was taken to fix the specification
- It is the first trial to introduce a shared single supercomputer system by multiple national universities in Japan !

Procurement Policies of JCAHPC

JCAHPC

- based on the spirit of T2K, introducing open advanced technology
  - massively parallel PC cluster
  - advanced processor for HPC
  - easy to use and efficient interconnection
  - large scale shared file system flatly shared by all nodes
- joint procurement by two universities
  - the largest class of budget as national universities' supercomputer in Japan
  - the largest system scale as PC cluster in Japan
  - no accelerator to support wide variety of users and application fields
    -> not chasing absolute peak performance and inheriting traditional application codes (basically)
- goodness of single system
  - scale-merit by merging budget -> largest in Japan
  - ultra large scale single job execution at special occasion such as "Gordon Bell Prize Challenge"

⇒ Oakforest-PACS (OFP)

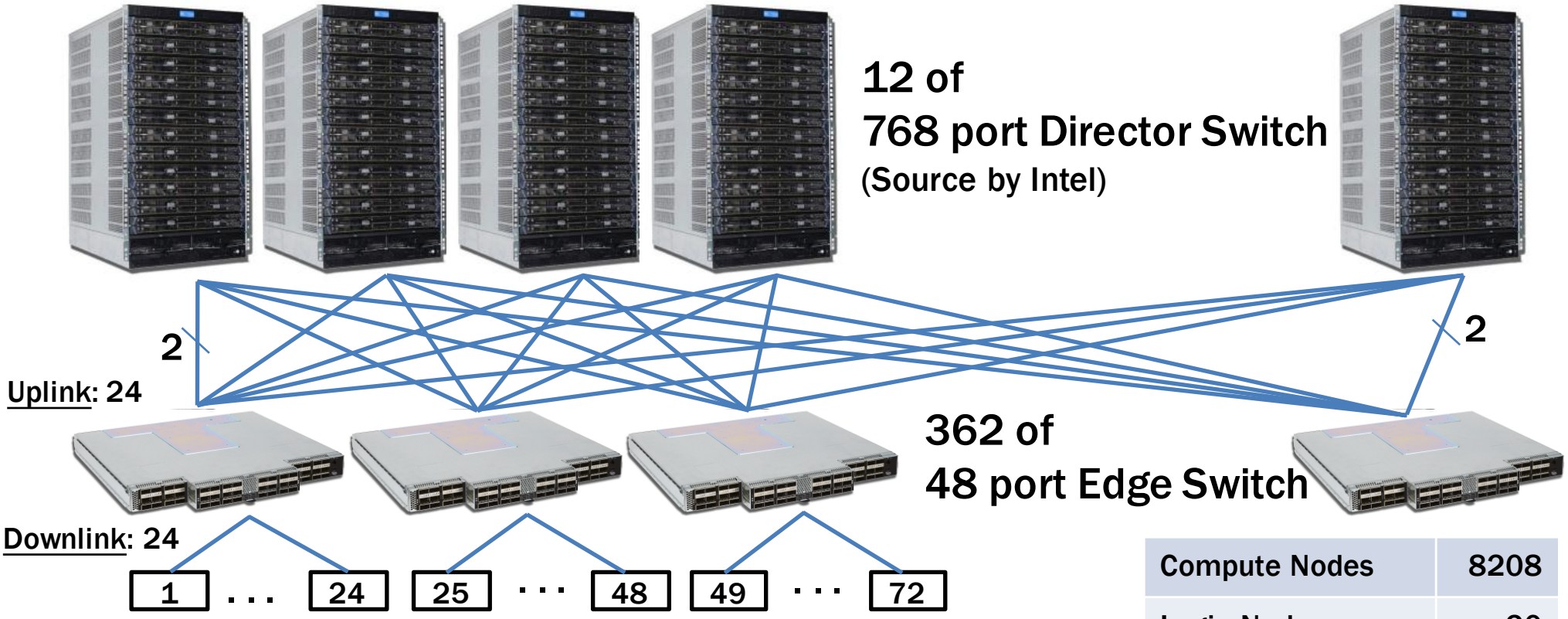
Specification of Oakforest-PACS

Total peak performance			25 PFLOPS
Total number of compute nodes			8,208
Compute node	Product		Fujitsu Next-generation PRIMERGY server for HPC (under development)
	Processor		Next-generation of Intel® Xeon Phi™ (Code name: Knights Landing), >60 cores
	Memory	High BW	16 GB, > 400 GB/sec (MCDRAM, effective rate)
		Low BW	96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate)
Inter-connect	Product		Intel® Omni-Path Architecture
	Link speed		100 Gbps
	Topology		Fat-tree with (completely) full-bisection bandwidth
Login node	Product		Fujitsu PRIMERGY RX2530 M2 server
	# of servers		20
	Processor		Intel Xeon E5-2690v4 (2.6 GHz 14 core x 2 socket)
	Memory		256 GB, 153 GB/sec (DDR4-2400 x 4ch x 2 socket)

Specification of Oakforest-PACS (I/O)

Parallel File System	Type		Lustre File System
	Total Capacity		26.2 PB
	Meta data	Product	DataDirect Networks MDS server + SFA7700X
		# of MDS	4 servers x 3 set
		MDT	7.7 TB (SAS SSD) x 3 set
	Object storage	Product	DataDirect Networks SFA14KE
		# of OSS (Nodes)	10 (20)
		Aggregate BW	500 GB/sec
Fast File Cache System	Type		Burst Buffer, Infinite Memory Engine (by DDN)
	Total capacity		940 TB (NVMe SSD, including parity data by erasure coding)
	Product		DataDirect Networks IME14K
	# of servers (Nodes)		25 (50)
	Aggregate BW		1,560 GB/sec

Full bisection bandwidth Fat-tree by Intel® Omni-Path Architecture



Compute Nodes	8208
Login Nodes	20
Parallel FS	64
IME	300
Mgmt, etc.	8
Total	8600

Firstly, to reduce switches&cables, we considered :
- All the nodes into subgroups are connected with FBB Fat-tree
- Subgroups are connected with each other with >20% of FBB

But, HW quantity is not so different from globally FBB, and globally FBB is preferredfor flexible job management.

Facility of Oakforest-PACS system

JCAHPC

Power consumption			4.2 MW (including cooling)
# of racks			102
Cooling system	Compute Node	Type	Warm-water cooling Direct cooling (CPU) Rear door cooling (except CPU)
		Facility	Cooling tower & Chiller
	Others	Type	Air cooling
		Facility	PAC

Software of Oakforest-PACS

JCAHPC

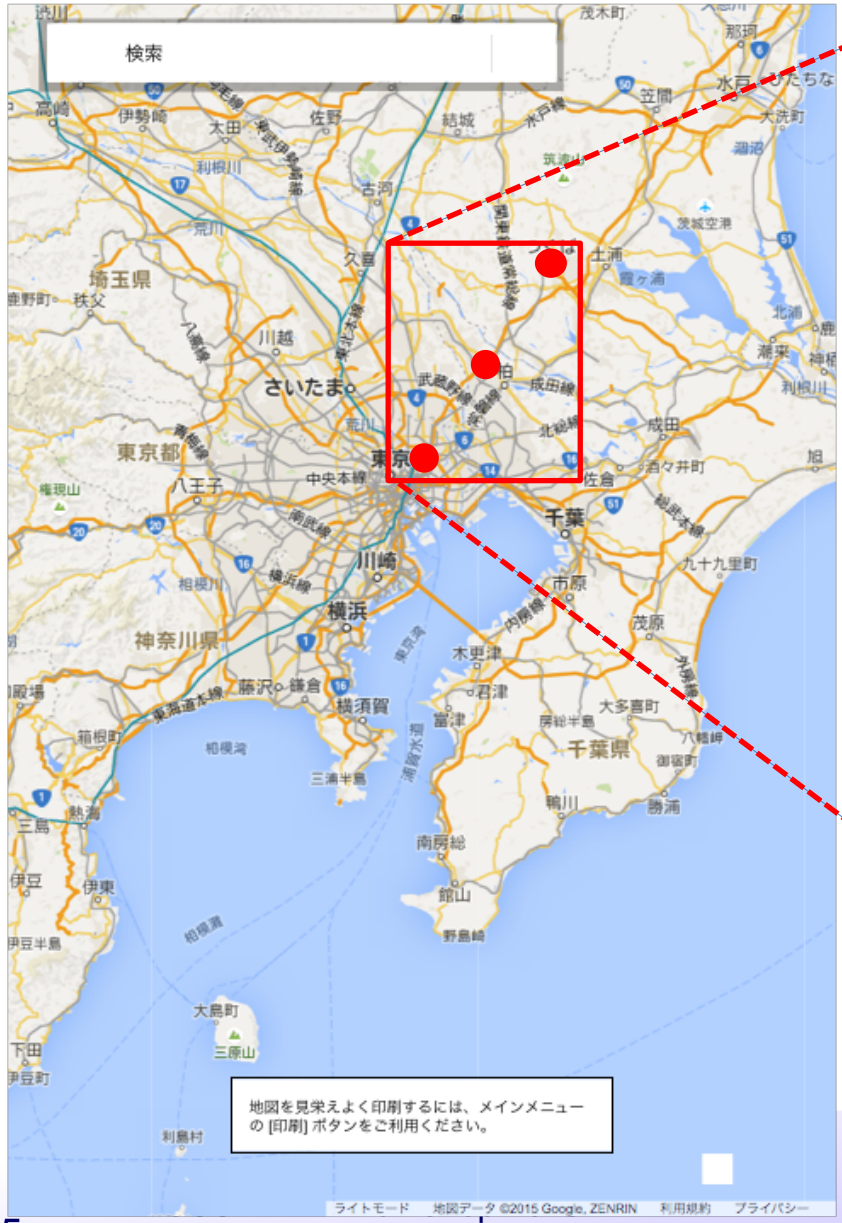
	Compute node	Login node
OS	CentOS 7, McKernel	Red Hat Enterprise Linux 7
Compiler	gcc, Intel compiler (C, C++, Fortran)	
MPI	Intel MPI, MVAPICH2	
Library	Intel MKL LAPACK, FFTW, SuperLU, PETSc, METIS, Scotch, ScaLAPACK, GNU Scientific Library, NetCDF, Parallel netCDF, Xabclib, ppOpen-HPC, ppOpen-AT, MassiveThreads	
Application	mpijava, XcalableMP, OpenFOAM, ABINIT-MP, PHASE system, FrontFlow/blue, FrontISTR, REVOCAP, OpenMX, xTAPP, AkaiKKR, MODYLAS, ALPS, feram, GROMACS, BLAST, R packages, Bioconductor, BioPerl, BioRuby	
Distributed FS		Globus Toolkit, Gfarm
Job Scheduler	Fujitsu Technical Computing Suite	
Debugger	Allinea DDT	
Profiler	Intel VTune Amplifier, Trace Analyzer & Collector	

(pre) Photo of computation node

Chassis with 4 nodes, 2U size

Computation node (Fujitsu next generation PRIMERGY)
with single chip Intel Xeon Phi (Knights Landing, 3+TFLOPS)
and Intel Omni-Path Architecture card (100Gbps)

Machine location: Kashiwa Campus of U. Tokyo



Schedule

- 2013/7 RFI
- 2015/1 RFC
- 2016/1 RFP
- 2016/3/30 Proposal deadline
- 2016/4/20 Bid opening
- 2016/10/1 1st step system operation (more than 5% of full system)
- 2016/12/1 2nd step, full system operation
- 2017/4 National open use starts including HPCI
- 2022/3 System shutdown (planned)

System operation outline

- Regular operation
  - both universities share the CPU time based on the budget ratio
  - not split the system hardware, but split the “CPU time” for flexible operation (except several specially dedicated partitions)
  - single system entry for HPCI program, and other own program by each university is performed under “CPU time” sharing
- Special operation
  - massively large scale operation (limited period)
    -> effectively using the largest class resource in Japan for special occasion (ex. Gordon Bell Challenge)
- Power saving operation
  - power capping feature for energy saving scheduling feature reacts to power saving requirement (ex. summer time)

Pre-evaluation environment

- U. Tsukuba and U. Tokyo are running own systems based on current Intel Xeon Phi (KNC) for performance evaluation and tuning
  - Tsukuba: COMA (PACS-IV), 393 nodes, 786 Xeon Phi
  - Tokyo: KNSoaring, 64 nodes, 64 Xeon Phi

- COMA at U. Tsukuba is also dedicated to HPCI and other programs since April 2015.

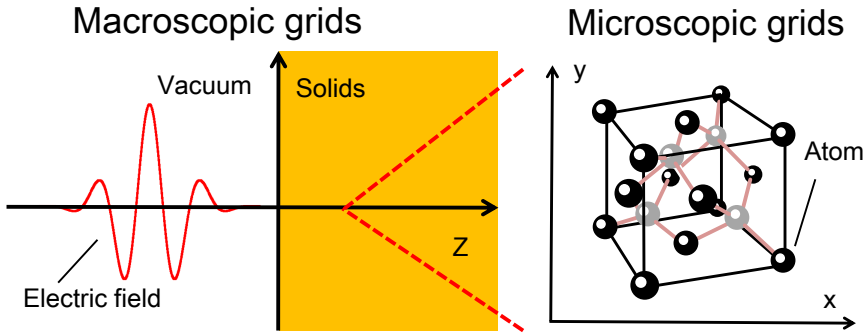
  $\Rightarrow$ we will introduce KNL-based small system soon

COMA (PACS-IX) – KNC cluster

- Cray CS300 Cluster
- Intel Xeon Phi (KNC: Knights Corner)
- 393 nodes（2 Xeon E5-2670v2 + 2 Xeon Phi 7110P)
- Mellanox IniniBand FDR, Fat Tree
- Largest Xeon Phi Cluster in Japan, as at Oct. 2015
- File Server: DDN 1.5PB (RAID6+Lustre)
- 1.001 PFLOPS (HPL: 746 TFLOPS) June ’14 TOP500 #51

Xeon Phi tuning on ARTED (with Yuta Hirokawa under collaboration with Prof. Kazuhiro Yabana, CCS)

- ARTED – Ab-initio Real-Time Electron Dynamics simulator
- Multi-scale simulator based on RTRSDFT (Real-Time Real-Space Density Functional Theory) developed in CCS, U. Tsukuba to be used for Electron Dynamics Simulation
  - RSDFT : basic status of electron (no movement of electron)
  - RTRSDFT : electron state under external force
- In RTRSDFT, RSDFT is used for ground state
  - RSDFT : large scale simulation with 1000～10000 atoms (ex. K-Computer)
  - RTRSDFT : calculate a number of unit-cells with 10 ~ 100 atoms



RSDFT : Real-Space Density Functional Theory
RTRSDFT : Real-Time RSDFT

Computation domain and amount

- Parameters for wave function expression
  - k-points (NK), band-number (NB), 3-D lattice points (NL)
  - valuables are in double precision complex with matrix of (NK, NB, NL)
  - for stencil computation, size NL of calculation is performed NKxNB times
- Parameters used in this research (two models)
  - SiO_2 : (4^3, 48, 36000 = (20, 36, 50)) -> not enough large
  - Si : (24^3, 32, 4096 = (16, 16, 16)) -> larger parallelism on thread
- NK is parallelized by MPI, then NKxNB is parallelized in OpenMP
  - domain of each process: (NK/NP, NB, NL)
    (NP = number of processes)
  - space domain is not decomposed to minimize MPI communication

Stencil code (original)

```
integer, intent(in) :: IDX(-4:4,NL),IDY(-4:4,NL),IDZ(-4:4,NL)

! NL = NLx*NLy*NLz
do i=0,NL-1
   ! x-computation
   v(1)=Cx(1)*(E(IDX(1,i))+E(IDX(-1,i))) ...
   w(1)=Dx(1)*(E(IDX(1,i))-E(IDX(-1,i))) ...

   ! y-computation
   v(2)=Cy(1)*(E(IDY(1,i))+E(IDY(-1,i))) ...
   w(2)=Dy(1)*(E(IDY(1,i))-E(IDY(-1,i))) ...

   ! z-computation
   v(3)=Cz(1)*(E(IDZ(1,i))+E(IDZ(-1,i))) ...
   w(3)=Dz(1)*(E(IDZ(1,i))-E(IDZ(-1,i))) ...

   ! update
   F(i) = B(i)*E(i) + A*E(i) - 0.5d0*(v(1)+v(2)+v(3)) - zI*(w(1)+w(2)+w(3))
end do
```

indirect index array:
keeping nearest neighbor index

write-only in the loop

vector length=4, for DP-complex vector calculation-> 512-bit AVX fittable

For automatic vectorization

```
real(8),    intent(in)   :: B(0:NLz-1,0:NLy-1,0:NLx-1)
complex(8),intent(in)    :: E(0:NLz-1,0:NLy-1,0:NLx-1)
complex(8),intent(out)   :: F(0:NLz-1,0:NLy-1,0:NLx-1)
```

convet to 3D array

```
#define IDX(dt) iz,iy,iand(ix+(dt)+NLx,NLx-1)
#define IDY(dt) iz,iand(iy+(dt)+NLy,NLy-1),ix
#define IDZ(dt) iand(iz+(dt)+NLz,NLz-1),iy,ix
```

direct index calculation

```
do ix=0,NLx-1
do iy=0,NLy-1
!dir$ vector nontemporal(F)
do iz=0,NLz-1
   v=0; w=0
   ! z-computation
   v=v+Cz(1)*(E(IDZ(1))+E(IDZ(-1)))  ...
   w=w+Dz(1)*(E(IDZ(1))-E(IDZ(-1)))  ...
   ! y-computation
   ! x-computation
   F(iz,iy,ix)  = B(iz,iy,ix)*E(iz,iy,ix)  &
   &            + A           *E(iz,iy,ix)  &
   &            - 0.5d0*v  - zI*w
end do
end do
end do
```

non-temporal write without cache

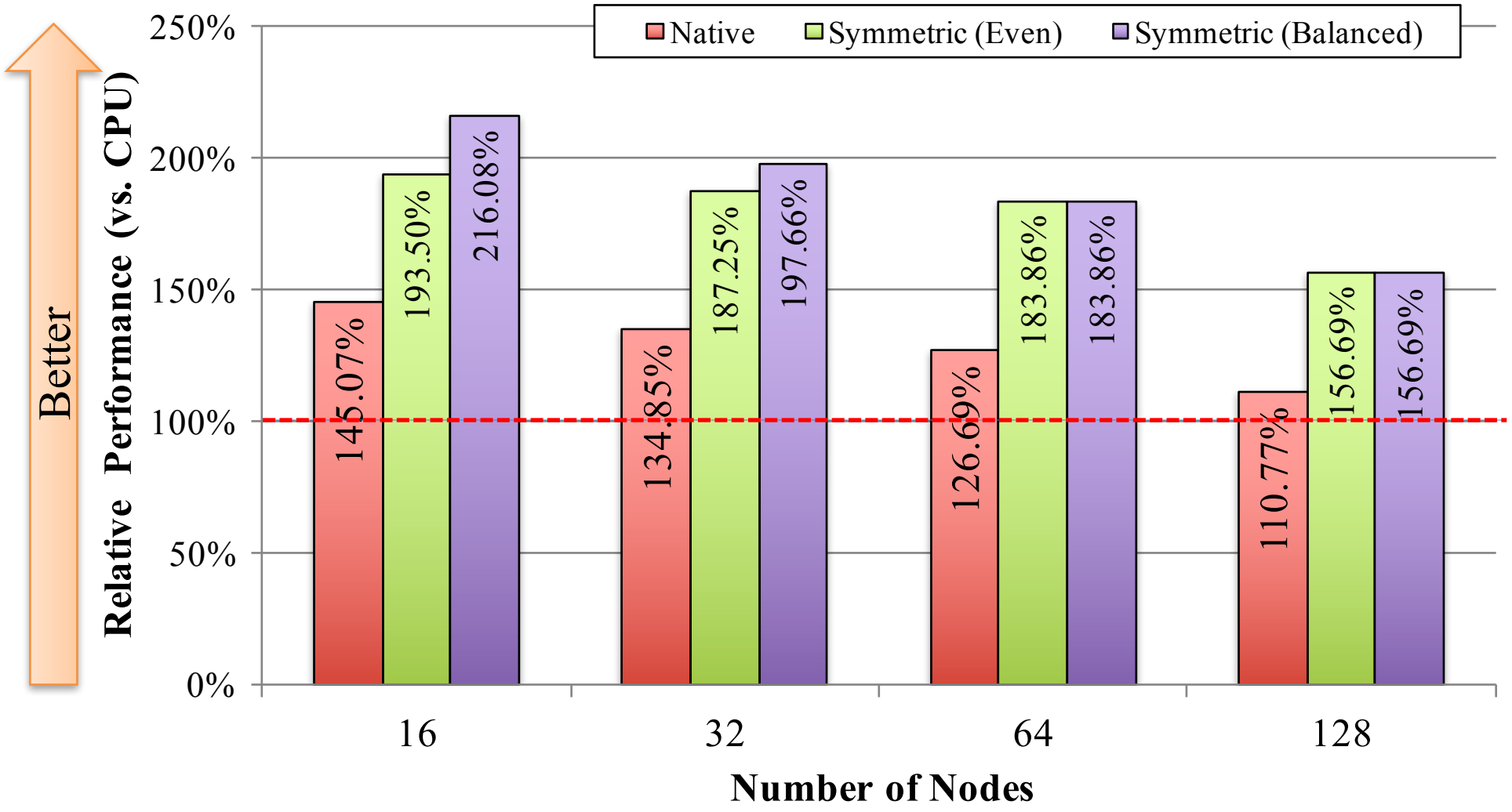
reordering according to memory access sequence

Stencil computation performance

- (NK, NB, NL) = (8^3, 16, 16^3) , single process
- 2x performance of Ivy-Bridge Xeon

	Type	GFLOPS	ratio to peak (%)
Xeon Phi 7110P	Original	29.0	2.70
	Compiler Vec.	132.2	12.30
	Explicit Vec.	212.2	19.75
Ivy-Bridge E5-2670v2	Original	53.7	26.85
	Compiler Vec.	102.7	51.35
	Explicit Vec.	106.9	53.45

Relative perf. to CPU on entire code (strong scaling)



Comparison with other systems

Entire code execution performance of ARTED

3x in KNL -> 300 GFLOPS ?

Processor	Vectorization	Performance [GFLOPS]	ratio to peak [%]
KNC	compiler	64.4	6.0
	explicit	103.4	9.6
IvyBridge	compiler	54.7	27.4
	explicit	55.9	27.9
Haswell	compiler	83.2	21.7
	explicit	70.7	18.4
Sparc64 VIIIfx	compiler	13.0	10.1

K Computer

exec. efficiency is comparable with K Computer

Summary

- JCAHPC is a joint resource center for advanced HPC by U. Tokyo and U. Tsukuba
- Oakforest-PACS with 25 PFLOPS peak performance with Intel Xeon Phi (Knights Landing) and Omni-Path Architecture as interconnect
- small part will start operation on Oct. 2016, and the full system will be available on Dec. 2016
- Under JCAHPC, both universities perform multiple resource sharing programs including HPCI
- JCAHPC is not just an organization to manage the resource but also a basic community for advanced HPC research