Oakforest-PACS: Japan's Fastest Intel Xeon Phi Supercomputer and its Applications

Taisuke Boku

Vice Director, JCAHPC &

Deputy Director, Center for Computational Sciences

University of Tsukuba

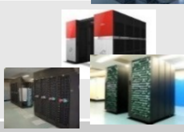
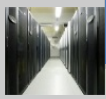
(with courtesy of JCAHPC members)

JCAHPC

- Joint Center for Advanced HPC
- A virtual organization with U. Tsukuba and U. Tokyo
  - for joint procurement on Japan's largest university supercomputer
  - for joint operation of the system
  - to provide the largest resource for HPCI (HPC Infrastructure) program under government
- Two universities contribute for all budget to procure and operate the machine
  - Tokyo : Tsukuba ratio = 2 : 1
- Official operation of the system starts on April 2017 under the name of "Oakforest-PACS"

Deployment plan of 9 supercomputing center（Feb. 2017）

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

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 12 PF, UCC/TPF 2.0MW				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)		5 PF(FAC/TPF) 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			
	K Computer (10PF, 13MW)					????		Post-K Computer (??. ??)				

Oakforest-PACS (OFP)

U. Tokyo convention U. Tsukuba convention

⇒ Don't call it just "Oakforest" !
"OFP" is much better

- 25 PFLOPS peak
- 8208 KNL CPUs
- FBB Fat-Tree by OmniPath
- HPL 13.55 PFLOPS
  #1 in Japan (acting)
  #6→#9
- HPCG #3→#6
- Green500 #6→#22

- Full operation started Dec. 2016
- Official Program started on April 2017

TOP500 list on Nov. 2017 (#50)

#	Machine	Architecture	Country	Rmax (TFLOPS)	Rpeak (TFLOPS)	MFLOPS/W
1	TaihuLight, NSCW	MPP (Sunway, SW26010)	China	93,014.6	125,435.9	6051.3
2	Tianhe-2 (MilkyWay-2), NSCG	Cluster (NUDT, CPU + KNC)	China	33,862.7	54,902.4	1901.5
3	Piz Daint, CSCS	MPP (Cray, XC50: CPU + GPU)	Switzerland	19,590.0	25,326.3	10398.0
4	Gyoukou, JAMSTEC	MPP (Exascaler, PEZY-SC2)	Japan	19,125.8	28,192.0	14167.3
5	Titan, ORNL	MPP (Cray, XK7: CPU + GPU)	United States	17,590.0	27,112.5	2142.8
6	Sequoia, LLNL	MPP (IBM, BlueGene/Q)	United States	17,173.2	20,132.7	2176.6
7	Trinity, NNSA/ LABNL/SNL	MPP (Cray, XC40: MIC)	United States	14,137.3	43,902.6	3667.8
8	Cori, NERSC-LBNL	MPP (Cray, XC40: KNL)	United States	14,014.7	27,880.7	3556.7
9	Oakforest-PACS, JCAHPC	Cluster (Fujitsu, KNL)	Japan	13,554.6	25,004.9	4985.1
10	K Computer, RIKEN AICS	MPP (Fujitsu)	Japan	10,510.0	11,280.4	830.2

HPCG on Nov. 2016

Rank	Site	Computer	Cores	Rmax Pflops	HPCG Pflops	HPCG /HPL	% of Peak
1	RIKEN Advanced Institute for Computational Science	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	705,024	10.5	0.603	5.7%	5.3%
2	NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.8	0.580	1.7%	1.1%
3	Joint Center for Advanced High Performance Computing Japan	Oakforest-PACS – PRIMERGY CX600 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel OmniPath, Fujitsu	557,056	24.9	0.385	2.8%	2.8%
4	National Supercomputing Center in Wuxi, China	Sunway TaihuLight – Sunway MPP, SW26010 260C 1.45GHz, Sunway, NRCPC	10,649,600	93.0	0.3712	0.4%	0.3%
5	DOE/SC/LBNL/NERSC USA	Cori – XC40, Intel Xe Cray Aries, Cray					
6	DOE/NNSA/LLNL USA	Sequoia – IBM BlueG 16C 1.6GHz, 5D Toru					
7	DOE/SC/Oak Ridge Nat Lab	Titan - Cray XK7 , Op 2.200GHz, Cray Gem NVIDIA K20x					
8	DOE/NNSA/LANL/SNL	Trinity - Cray XC40, I custom, Cray					
9	NASA / Mountain View	Pleiades - SGI ICE X, 2680v2, E5-2680v3, FDR, HPE/SGI					
10	DOE/SC/Argonne National Laboratory	Mira - BlueGene/Q, 1.60GHz, 5D Torus, I					

Specification of Oakforest-PACS

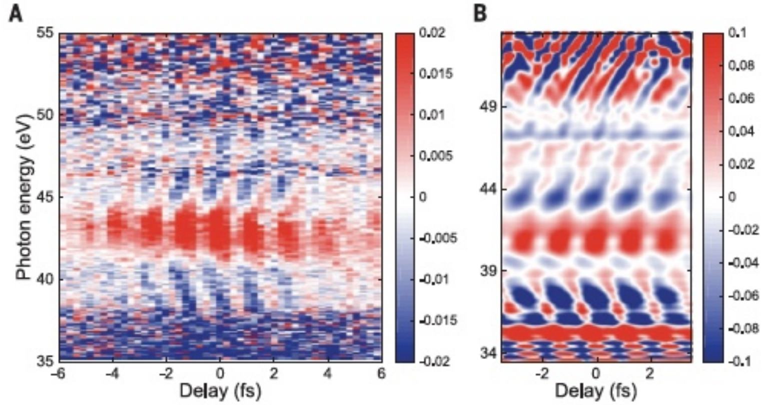
Total peak performance			25 PFLOPS
Total number of compute nodes			8,208
Compute node	Product		Fujitsu Next-generation PRIMERGY server for HPC (under development)
	Processor		Intel® Xeon Phi™（Knights Landing） Xeon Phi 7250 (1.4GHz TDP) with 68 cores
	Memory	High BW	16 GB, > 400 GB/sec (MCDRAM, effective 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 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)

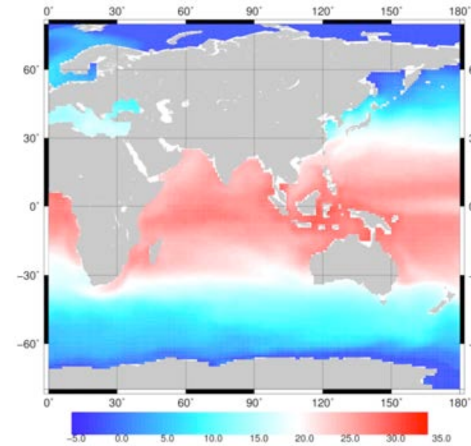
Specification of Oakforest-PACS (I/O)

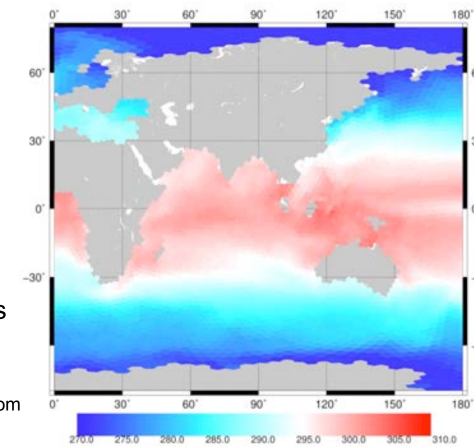
Parallel File System	Type		Lustre File System
	Total Capacity		26.2 PB
	Meta data	Product	DataDirect Networks MDS server + SFA7700X
		# of MDS	4 servers x 3 set
		MDT	7.7 TB (SAS SSD) x 3 set
	Object storage	Product	DataDirect Networks SFA14KE
		# of OSS (Nodes)	10 (20)
		Aggregate BW	~500 GB/sec
Fast 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

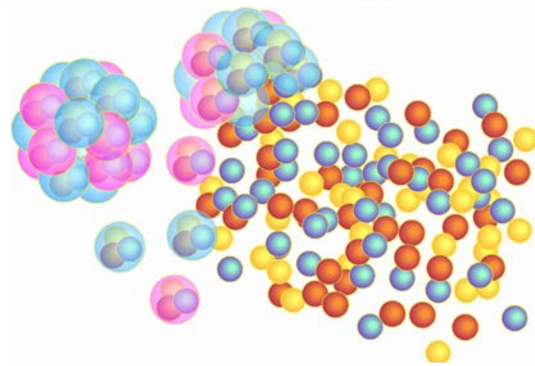
Large Scal Applications on Oakforest-PACS

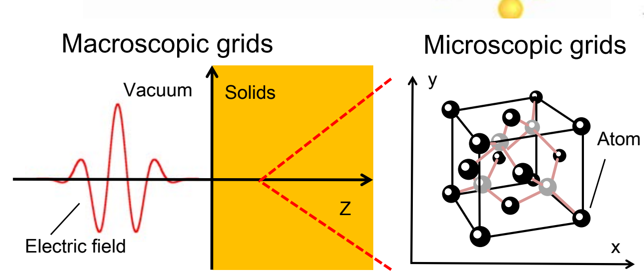
- **ARTED (SALMON)**
  - Electron Dynamics
- **Lattice QCD**
  - Quantum Chrono Dynamics
- **NICAM & COCO**
  - Atmosphere & Ocean Coupling
- **GHYDRA**
  - Earthquake Simulations
- **Seism3D**
  - Seismic Wave Propagation

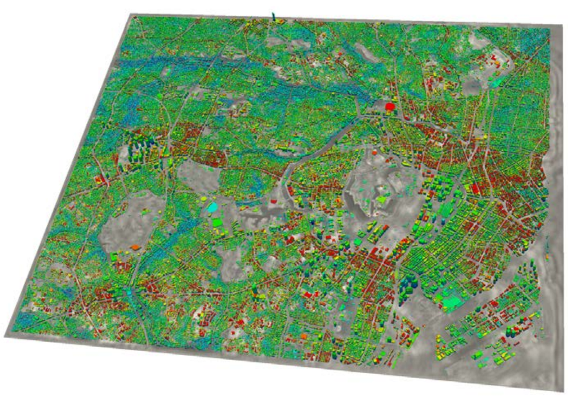


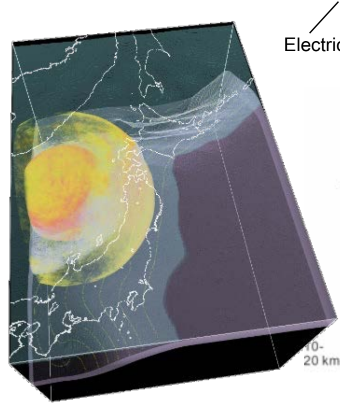


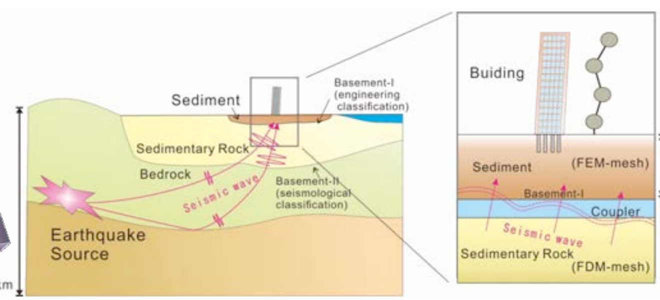


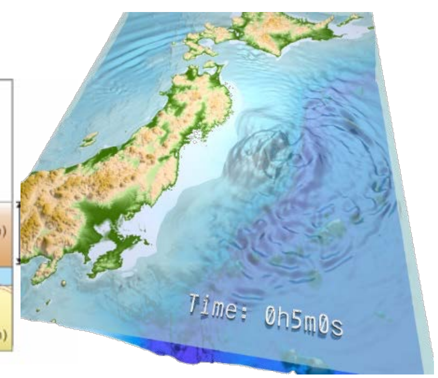






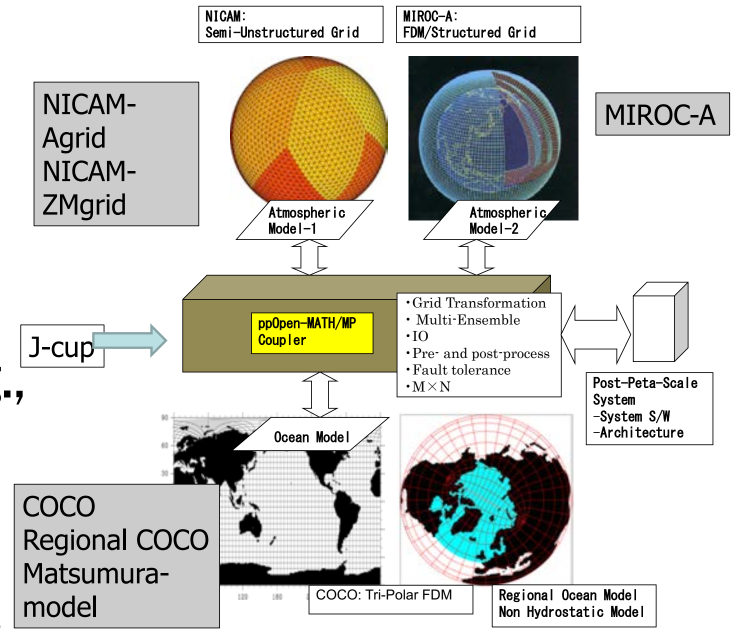


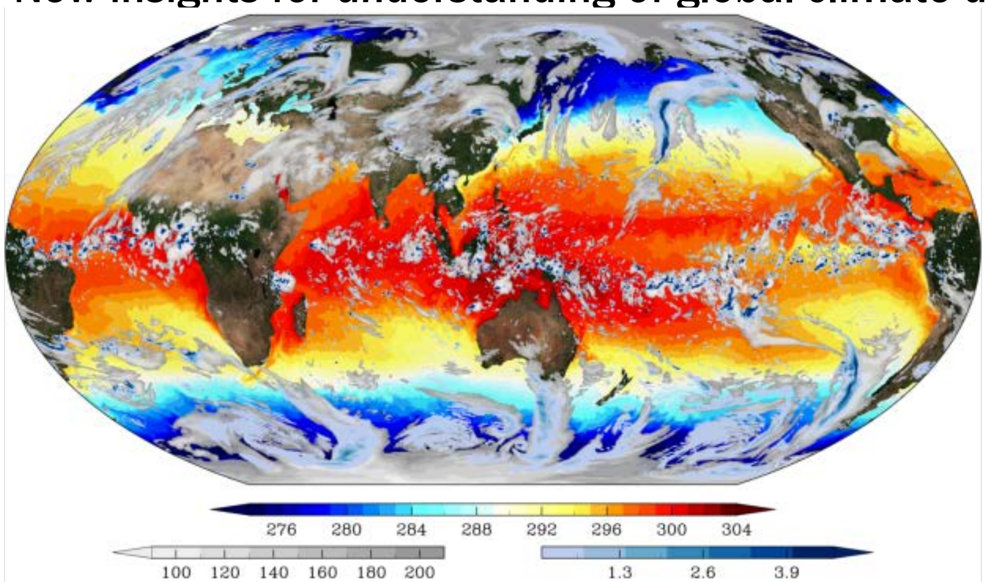


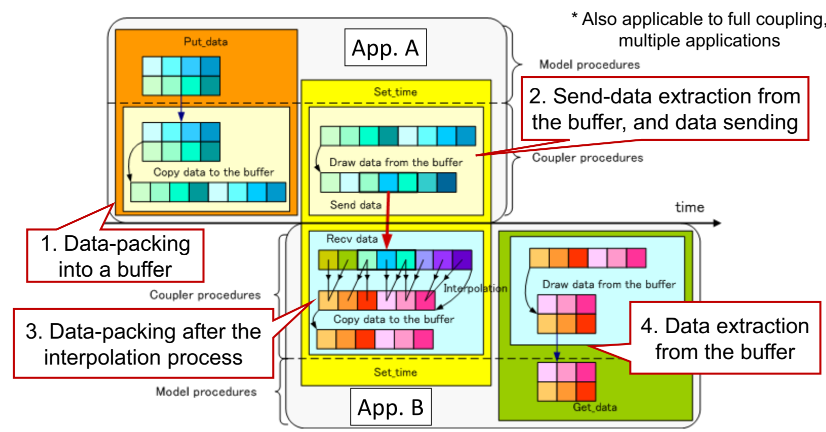


Atmosphere-Ocean Coupling
on OFP by NICAM/COCO/ppOpen-MATH/MP

- High-resolution global atmosphere-ocean coupled simulation by NICAM and COCO (Ocean Simulation) through ppOpen-MATH/MP on the K computer is achieved.
  - ppOpen-MATH/MP is a coupling software for the models employing various discretization method.
- An O(km)-mesh NICAM-COCO coupled simulation is planned on the Oakforest-PACS system (3.5km-0.10deg., 5+B Meshes).
  - A big challenge for optimization of the codes on new Intel Xeon Phi processor
  - New insights for understanding of global climate dynamics



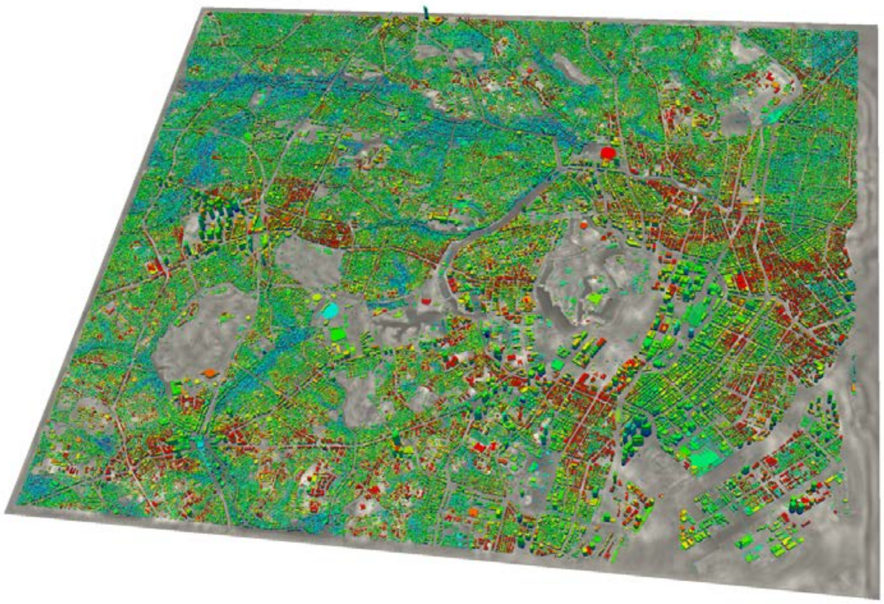




[C/O M. Satoh (AORI/UTokyo)@SC16]

Earthquake Simulations
Prof. Ichimura (ERI, U.Tokyo)

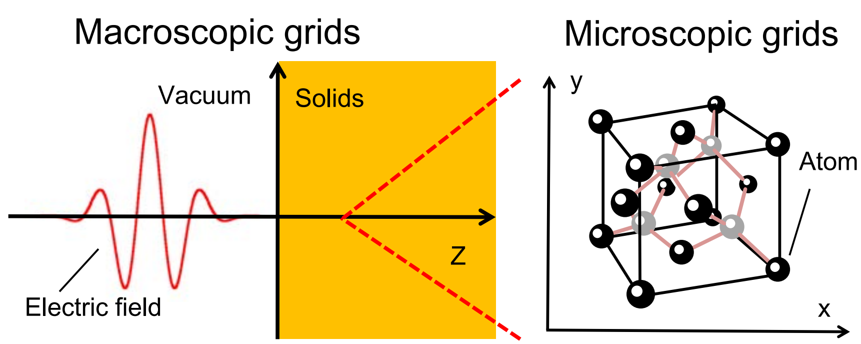
- GOJIRA/GAMERA
  - ✓ FEM with Tetrahedral Elements (2nd Order)
  - ✓ Nonlinear/Linear, Dynamic/Static Solid Mechanics
  - ✓ Mixed Precision, EBE-based Multigrid
  - ✓ SC14, SC15: Gordon Bell Finalist
  - ✓ SC16: Best Poster
- GHYDRA
  - ✓ Time-Parallel Algorithm
  - ✓ Oakforest-PACS (on-going)



Simulation example: Earthquake simulation of 10.25 km x 9.25 km area of central Tokyo using full K computer. Response of 328 thousand buildings are evaluated using three-dimensional ground data and building data. Analyzed using a 133 billion degrees-of-freedom nonlinear finite-element model.

Xeon Phi tuning on ARTED (with Y. Hirokawa under collaboration with Prof. K. Yabana, CCS) → SALMON now

- ARTED – Ab-initio Real-Time Electron Dynamics simulator
- Multi-scale simulator based on RTRSDFT 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

supported by JST-CREST and Post-K important field development program (field-7)

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
```

Original code just compiled on KNC with "-O3" option → less than 5% of peak!

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
```

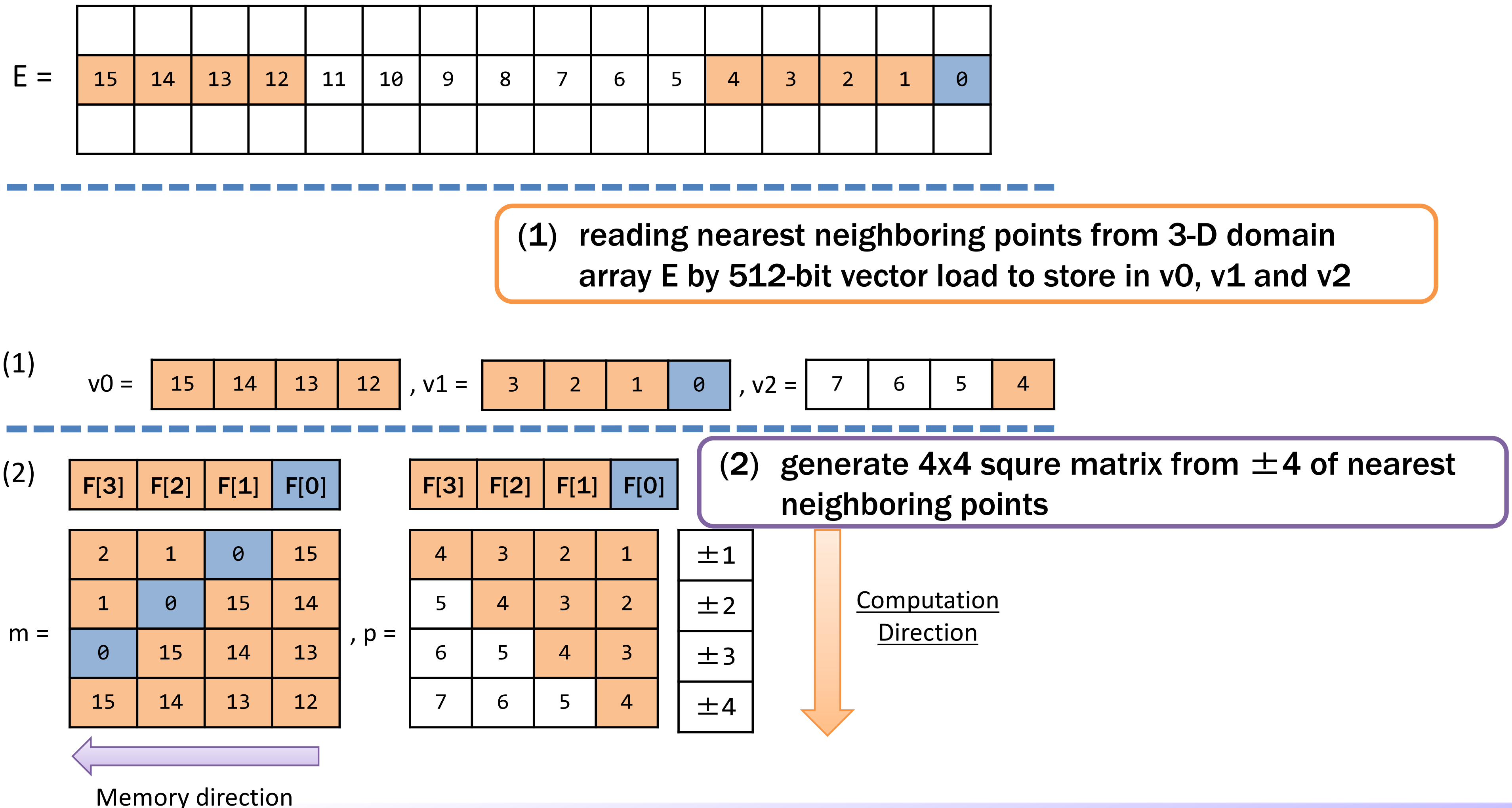
direct index calculation

```
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
```

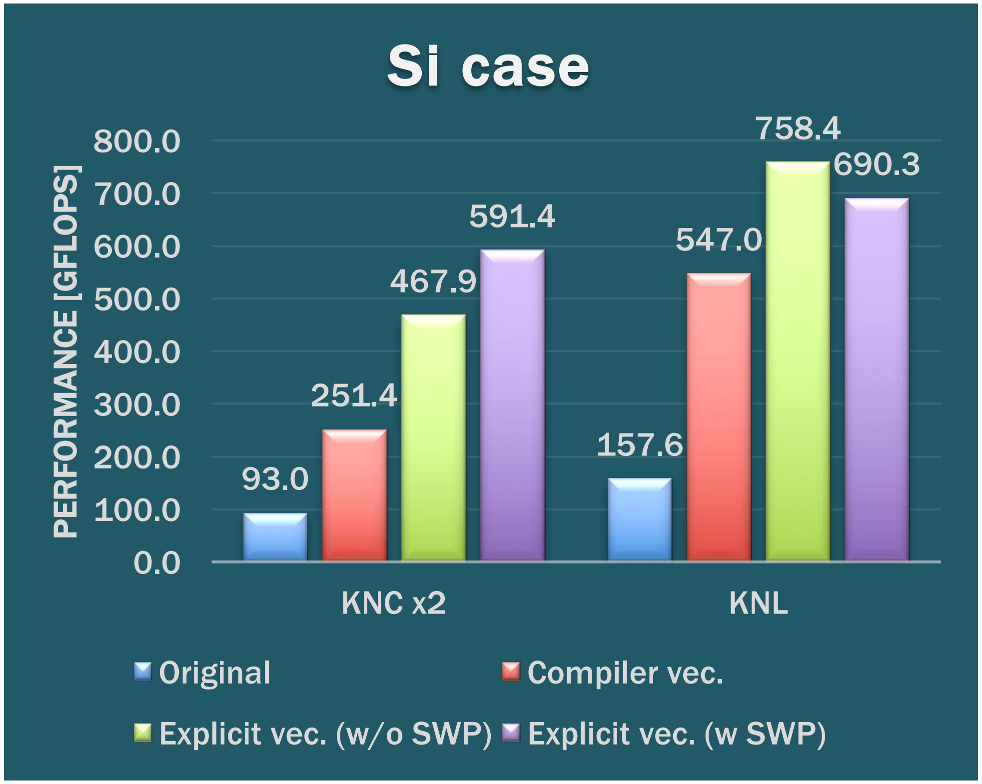
non-temporal write without cache

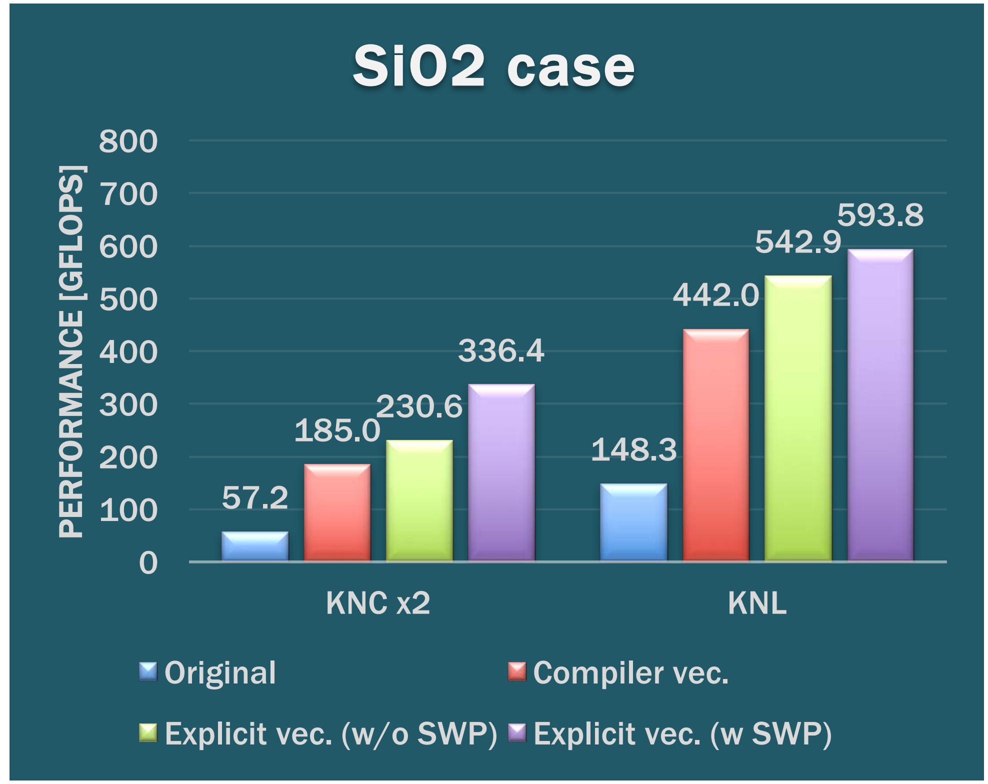
reordering according to memory access sequence

Hand vectorization – unit-stride memory access optimization (how to utilize AVX-512 SIMD load and operation)



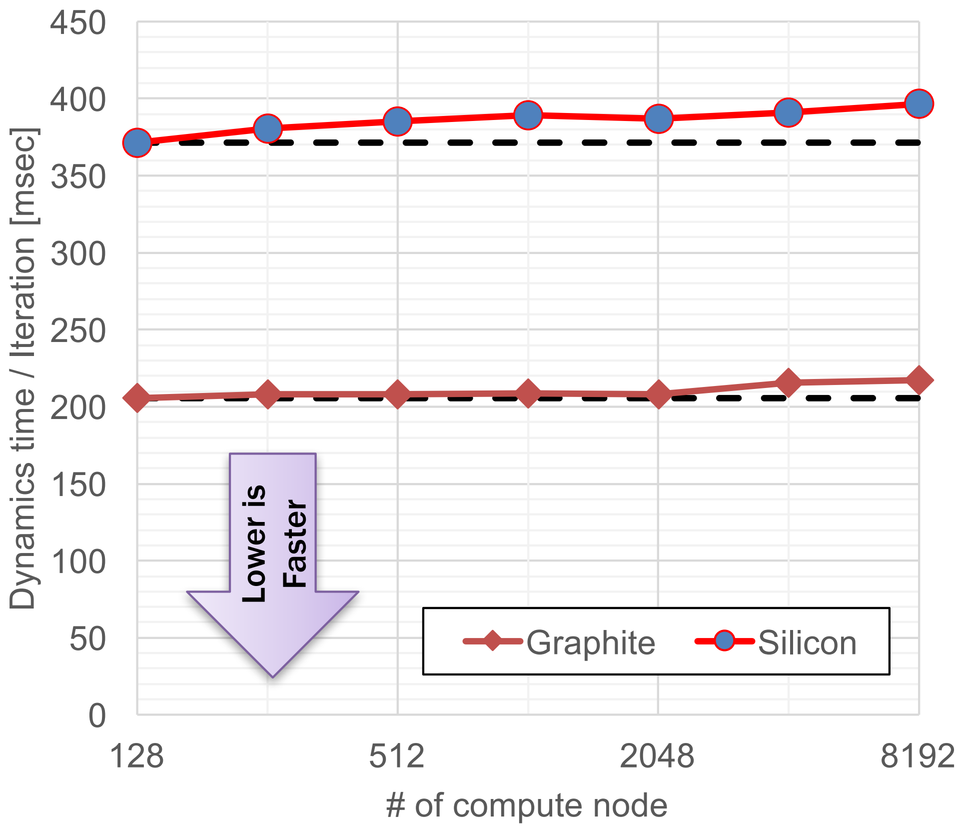
Stencil computation (3D) performance

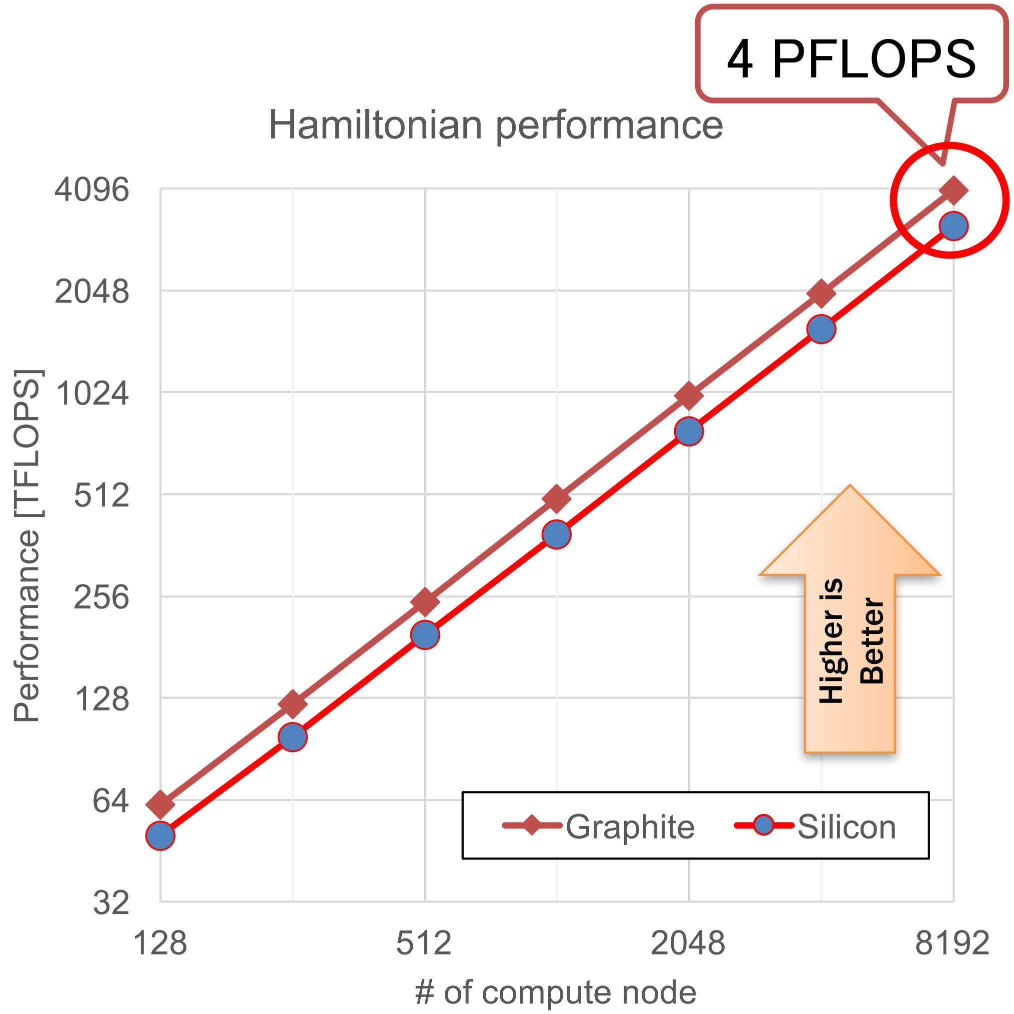




>2x faster than KNC (at maximum) -> up to 25% of theoretical peak of KNL

Weak scaling on OFP full system





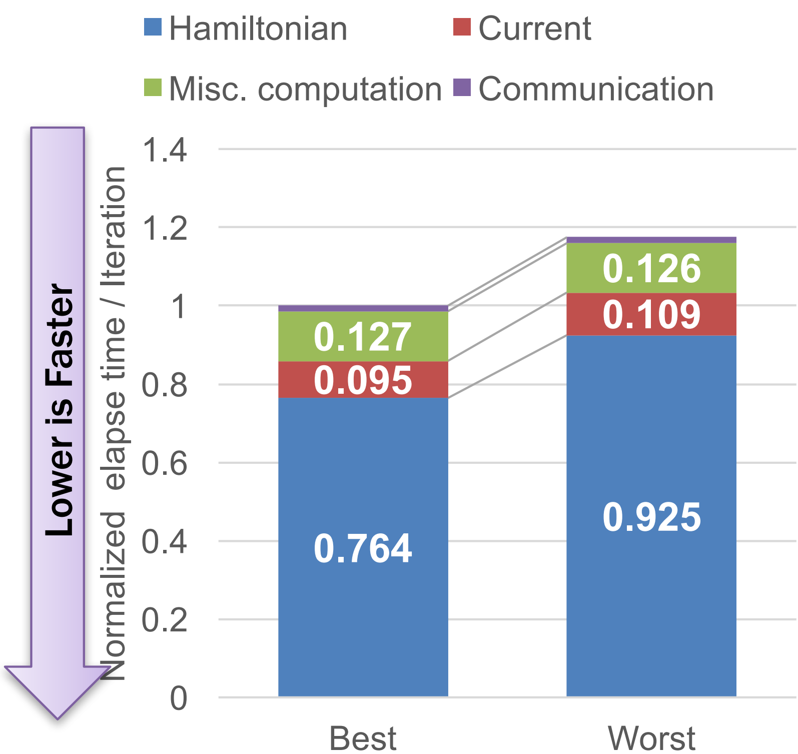
KNL vs GPU for ARTED (3D stencil part)

Si case	GFLOPS	vs. Peak perf.
Xeon E5-2670v2 x2 (IVB)	232.1	58.0%
Xeon Phi 7110P x2 (KNC)	592.3	27.6%
OFP: Xeon Phi 7250 (KNL)	758.0	24.8%
Tesla K40 x2 (Kepler)	476.0	33.3%
Tesla P100 (Pascal)	788.2	14.9%

	Peak performance (DP)	Actual memory bandwidth	Actual B/F
Xeon Phi 7110P (KNC)	1074 GFLOPS	177.1 GB/s	0.16
Xeon Phi 7250 (KNL)	2998 GFLOPS	456.2 GB/s	0.15
Tesla K40 (Kepler)	1430 GFLOPS	180.5 GB/s	0.13
Tesla P100 (Pascal)	5300 GFLOPS	514.8 GB/s	0.10

Update will be presented as ISC2018 at Frankfurt

Performance variant between nodes



normalized to best case

- **most of time is consumed for Hamiltonian calculation**
  - **not including communication time**
  - **domain size is equal for all nodes**
- **root cause of strong scaling saturation**
  - **performance gap exists on any materials**
- **Non-algorithmic load-imbalancing**
  - **dynamic clock adjustment (DVFS) on turbo boost is applied individually on all processors**
  - **it is observed on under same condition of nodes**
  - **on KNL, more sensitive than Xeon**
  - **serious performance degradation on synchronized large scale system**

Lattice QCD

- Xeon Phi tuning under IPCC (Intel Parallel Computing Center) program at CCS, U. Tsukuba
  - PI: T. Boku
    members: K. Ishikawa (Hiroshima U.), M. Umemura, Y. Kuramashi
  - Intel: L. Meadows, M. D'Mello, M. Troute, R. Vemuri
- CCS-QCD benchmark
  - has been developed in CCS for more than 10 years
  - selected as one of key programs for Post-K project, next national flagship supercomputer in Japan
  - we need many-core ready version of code and OFP is the largest target
- Performance of CCS-QCD
  - mainly bottlenecked by memory bandwidth for stencil computing
  - key: how to reduce the communication overhead

CCS-QCD on OFP (before tuning)

- We measured the performance of the CCS-QCD using the full system of Oakforest-PACS system
- Weak scaling test (1000 nodes -> 8000 nodes)

Lattice size	# of nodes	**SOLVE**	**MULT**(w/o comm.)	**MULT**(w/ comm.)	**YCOPY**	**MPI_Allreduce**
$200^3 \times 800$	1000	90.109 [s]	50.874 [s]	51.002 [s]	0.031 [s]	14.627 [s]
$400 \times 200^2 \times 800$	2000	92.340	51.067	51.187	0.025	16.531
$400^2 \times 200 \times 800$	4000	94.481	50.991	51.145	0.057	18.887
$400^3 \times 800$	8000	98.794	50.420	50.543	0.031	23.554
Lattice size	# of nodes	**SOLVE** [GFLOPS/node]	**MULT**(w/o comm.) [GFLOPS/node]	**MULT**(w/ comm.) [GFLOPS/node]	-	-
$200^3 \times 800$	1000	356.2	561.6	560.2		
$400 \times 200^2 \times 800$	2000	347.6	559.5	558.1		
$400^2 \times 200 \times 800$	4000	339.7	560.3	558.6		
$400^3 \times 800$	8000	324.9	566.6	565.2		

- Communication overhead was almost hidden in MULT.
- MPI_Allreduce was the bottleneck for the good performance and scaling.
- MULT performance reaches 560GF/node [=280GF/MPI].
- SOLVE(single precision solver) reaches 325-356 GF.
- Using 16000 MPI procs on 8000 nodes, the total performance reached 2.6 PFLOPS sustained and was ~ 10% of the peak 26PF of OFP.

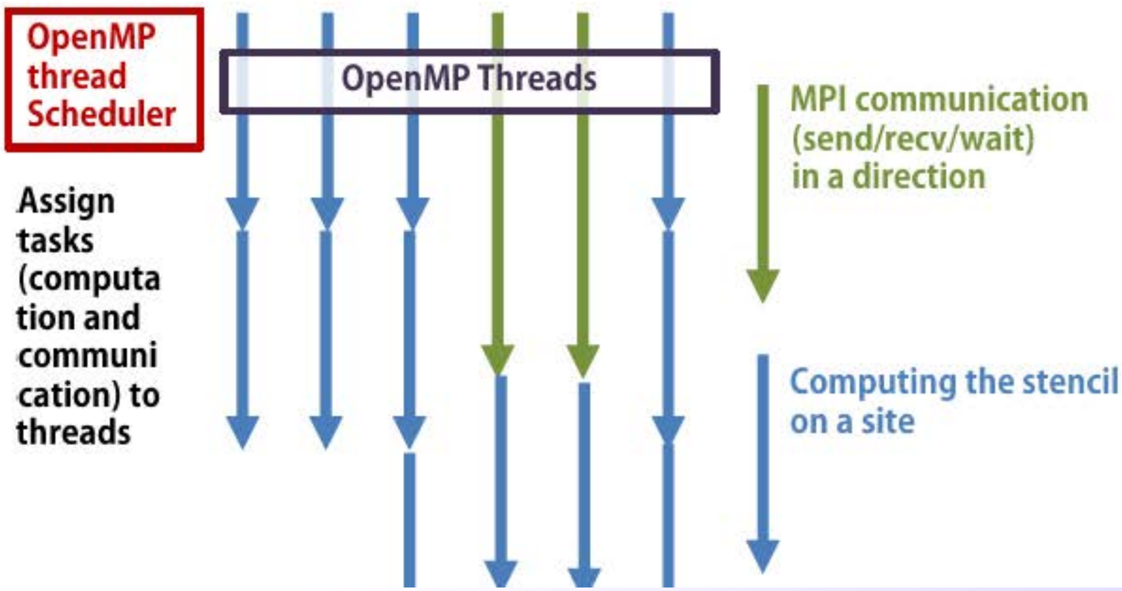
CCS-QCD with multiple endpoints (Multi-EP)

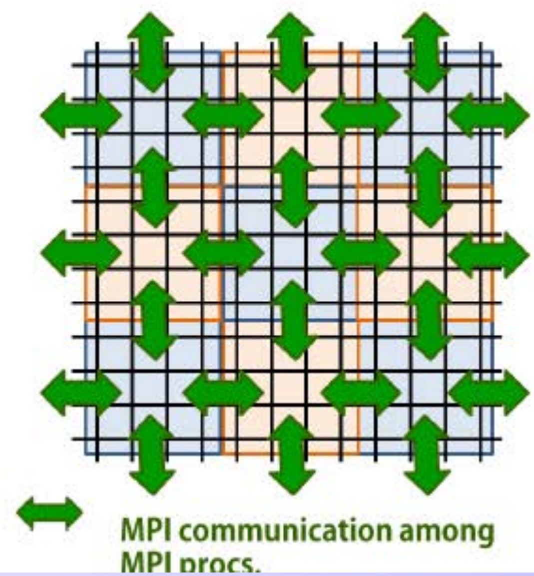
- **More improvements achieved in FY2017**
  - Test the Multiple Endpoints facility of the 2019 Technical Preview release of Intel MPI library.
    - No MPI offloading
    - MPI_THREAD_MULTIPLE enabled.
    - Computation threads and MPI threads using thread scheduler.
    - Split the COMMUNICATOR (MPI_COMM_WORLD) to several communicators for each threads handing MPI-communication.

- L. Meadows and K.-I.I., "OpenMP Tasking and MPI in a Lattice QCD Benchmark", In: Scaling OpenMP for Exascale Performance and Portability. IWOMP 2017. Lecture Notes in Computer Science, vol 10468. Springer, Cham, (2017) 77-91.
- L. Meadows, K.-I.I., T. Boku, M. Horikoshi, "Multiple Endpoints for Improved MPI Performance on a Lattice QCD Code", Proceedings of Workshops of HPC Asia, HPC Asia '18, (2018) 67--70.

CCS-QCD with multiple endpoints (Multi-EP)

- **More improvements achieved in FY2017**
  - Test the Multiple Endpoints facility of the 2019 Technical Preview release of Intel MPI library.
    - No MPI offloading
    - MPI_THREAD_MULTIPLE enabled.
    - Computation threads and MPI threads using thread scheduler.
    - Split the COMMUNICATOR (MPI_COMM_WORLD) to several communicators for each threads handing MPI-communication.





CCS-QCD with multiple endpoints (Multi-EP)

- More improvements achieved in FY2017
  - Performance comparison :
    Previous implementation vs New Multi-EP version
  - Communication Timing Comparison
  - 16^3 x 64 Lattice
    - Prev. ver.: 4x2x2x1 MPI (16MPI), 2 MPI/node, 32 Threads/MPI
    - Multi-EP ver.: 2x2x2x1 MPI(8MPI), 1 MPI/node, 64 Threads/MPI

Version	Num. of Threads for Comm.	Solve [sec]	Mult [sec]	Mult_PRE [sec]	Mult_IN [sec]	Mult_PST [sec]	COMM [sec]
Previous	NA	13.8	5.94	0.881	2.14	2.90	5.64
Multi-EP	1	16.2	4.08	0.427	2.26	1.40	10.4
	2	12.5	4.04	0.429	2.16	1.45	6.37
	4	10.3	4.23	0.435	2.28	1.51	4.45
	6	8.42	4.14	0.450	2.02	1.67	2.51
	8	8.46	4.01	0.443	1.94	1.63	2.53

Summary

- **JCAHPC** is a joint resource center for advanced HPC by **U. Tokyo and U. Tsukuba** as the first case in Japan
- **Full system scale applications** including fundamental physics, global science, disaster simulation, material science, etc. are under development with extreme scale and getting new results
- Two **JCAHPC** universities lead the advanced performance tuning on many scientific codes
- **ARTED** (in SALMON) is an application with the highest sustained performance of Oakforect-PACS
- **CCS-QCD** optimization leads to Post-K supercomputer key program development