JCAHPC

新しい科学を開拓する
Oakforest-PACS

中島 研吾
東京大学情報基盤センター
最先端共同HPC基盤施設
（研究開発部門長）

2017年10月26日
サイエンティフィックシステム研究会
科学技術計算分科会2017年度会合
拡がるHPC ～新たな鼓動～

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

筑波大学
計算科学研究センター
Center for Computational Sciences

Oakforest-PACS （OFP）

- 2016年12月1日稼働開始
- 8,208 Intel Xeon/Phi (KNL)，ピーク性能25PFLOPS
  - 富士通が構築
- **TOP 500　7位（国内1位），HPCG 5位（国内2位）（2017年6月）**
- **最先端共同HPC 基盤施設(JCAHPC: Joint Center for Advanced High Performance Computing)**
  - 筑波大学計算科学研究センター
  - 東京大学情報基盤センター
    - 東京大学柏キャンパスの東京大学情報基盤センター内に、両機関の教職員が中心となって設計するスーパーコンピュータシステムを設置し，最先端の大規模高性能計算基盤を構築・運営するための組織
  - http://jcahpc.jp

JCAHPC

東京大学 THE UNIVERSITY OF TOKYO　筑波大学 University of Tsukuba

JCAHPC

最先端共同HPC基盤施設 JCAHPC

- Joint Center for Advanced High Performance Computing (http://jcahpc.jp)
- 平成25年3月、筑波大学と東京大学は「計算科学・工学及びその推進のための計算機科学・工学の発展に資するための連携・協力推進に関する協定」を締結
- 本協定の下、筑波大学計算科学研究センターと東京大学情報基盤センターが JCAHPC を設置

東京大学 THE UNIVERSITY OF TOKYO　筑波大学 University of Tsukuba

平成25年のプレスリリース

平成 25 年 7 月 22 日

国立大学法人 東京大学情報基盤センター
国立大学法人 筑波大学計算科学研究センター
最先端共同 HPC 基盤施設

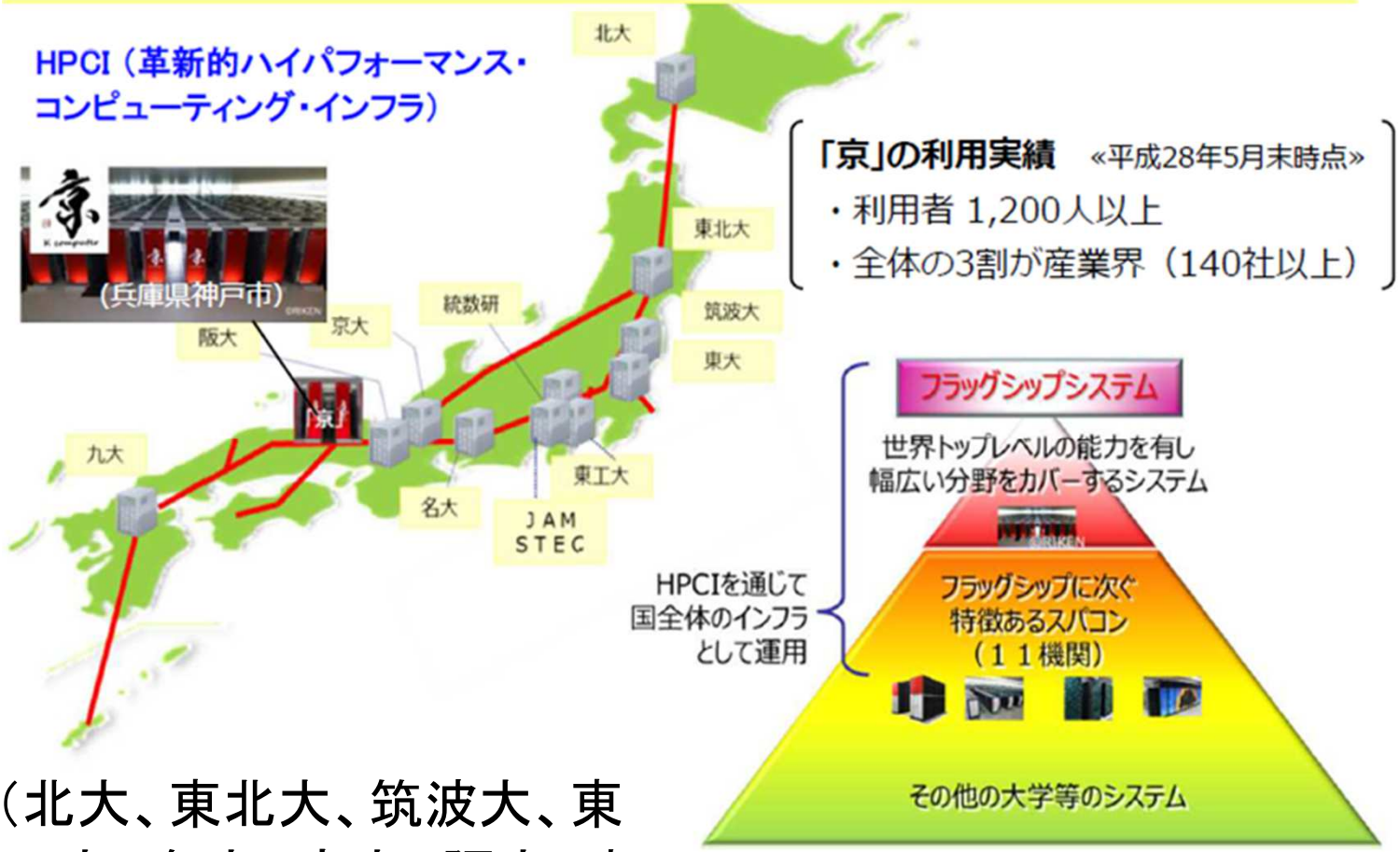
最先端共同 HPC 基盤施設の活動を開始

筑波大学と東京大学によるスーパーコンピュータ共同開発、共同運営・管理

2センターのミッション

- 筑波大学計算科学研究センターのミッション:
  - 計算機科学と計算科学の協働:学際的な高性能計算機開発
    → PACSシリーズの開発:CP-PACS@1996 TOP1
  - 先端学際科学共同研究拠点: 最先端の計算科学研究推進
  - これからの計算科学に必要な学際性を持つ人材を育成
- 東京大学情報基盤センターのミッション:
  - 学際大規模情報基盤共同利用・共同研究拠点(8大学の情報基盤センター群からなるネットワーク型)の中核拠点:
    大規模情報基盤を活用し学際研究を発展
  - HPCI資源提供機関:最先端スパコンの共同設計開発及び運用、Capability資源および共用ストレージ資源の提供
  - 人材育成:計算科学の新機軸を創造できる人材の育成

HPCI: High Performance Computing Infrastructure
日本全体におけるスパコンインフラ



- 9大学（北大、東北大、筑波大、東大、東工大、名大、京大、阪大、九大）の情報基盤センター
- 海洋開発研究機構、統数研＋SINET

Plan of 9 Tier-2 Supercomputer Centers (May 2017)

JCAHPC

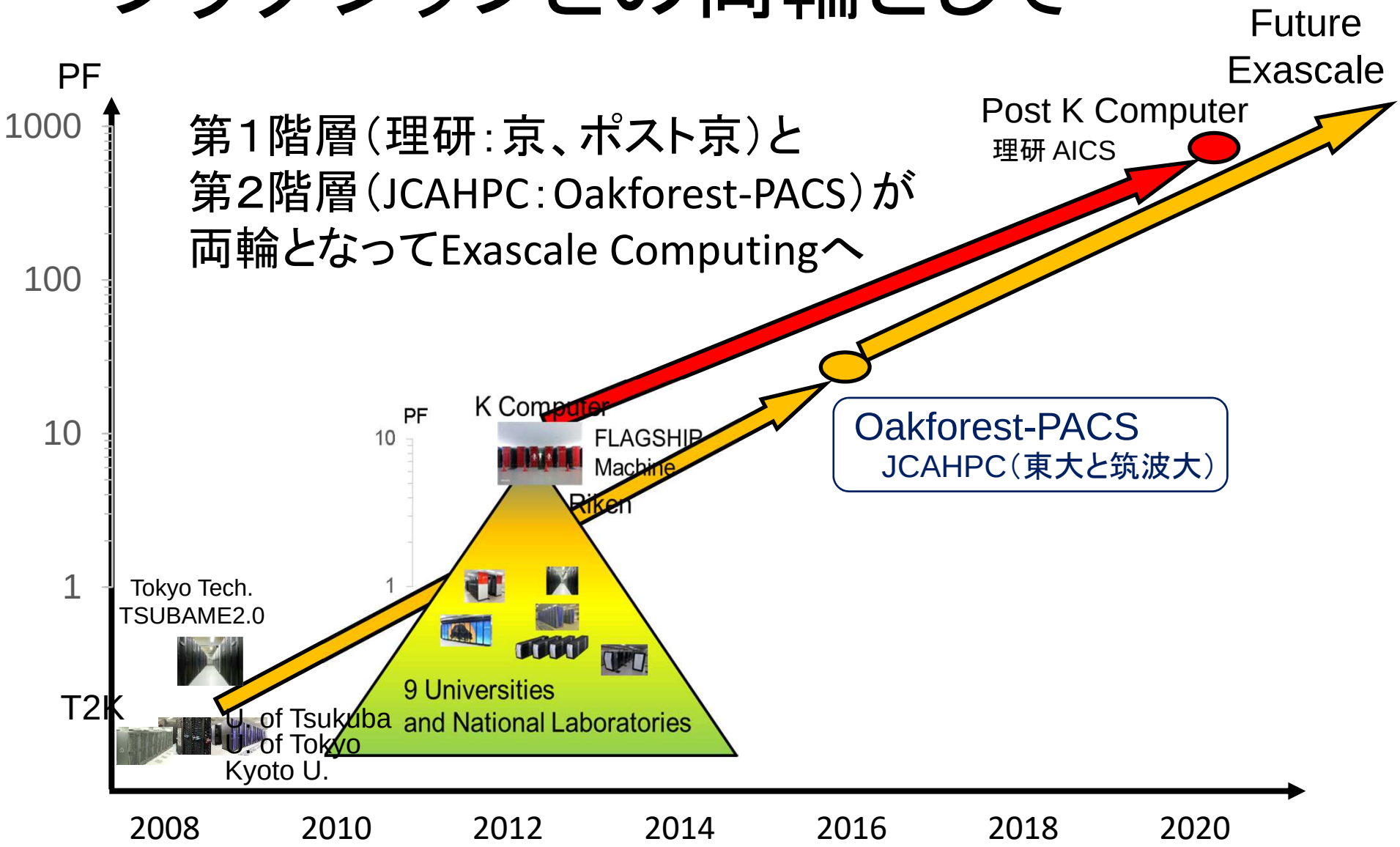
Fiscal Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Hokkaido	HITACHI SR16000/M1 (172TF, 22TB) Cloud System BS2000 (44TF, 14TB)			Data Science Cloud / Storage HA8000 / WOS7000 (10TF, 1.96PB)	3.2 PF (UCC + CFL/M) 0.96MW		0.3 PF (Cloud) 0.36MW			30 PF (UCC + CFL-M) 2MW		
Tohoku	NEC SX-9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)			PACS-X 10PF (TPF) 2MW								
	COMA (PACS-IX) (1001 TF)											
Tokyo	Fujitsu FX10 (1PFlops, 150TiB, 408 TB/s)			Oakforest-PACS 25 PF (UCC + TPF) 3.2 MW				100+ PF 4.5MW (UCC + TPF)				
	Hitachi SR16K/M1 (54.9 TF, 10.9 TiB, 28.7 TB/s)			Reedbush 1.93 PF 0.7 MW	Reedbush-L 1.40 PF 0.18 MW		BDEC: 50+ PF (FAC) 3.5MW				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)			Fujitsu FX100 (2.9PF, 81 TiB)								
	CX400(470TF)			Fujitsu CX400 (774TF, 71TiB)								
	SGI UV2000 (24TF, 20TiB)			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)					7-8 PF(FAC/TPF + UCC) 1.5 MW		50-100+ PF (FAC/TPF + UCC) 1.8-2.4 MW				
	Cray XC30 (584TF)											
Osaka	NEC SX-ACE (423TF)	NEC Express5800 (22.4TF)				3.2PB/s, 5-10Pflop/s, 1.0-1.5MW (CFL-M)					25.6 PB/s, 50-100Pflop/s,1.5-2.0MW	
					0.7-1PF (UCC)							
Kyushu	HA8000 (712TF, 242 TB) SR16000 (8.2TF, 6TB)			2.0MW		15-20 PF (UCC/TPF) 2.6MW				100-150 PF (FAC/TPF + UCC/TPF) 3MW		
	FX10 (272.4TF, 36 TB) CX400 (966.2 TF, 183TB)					FX10 (90.8TFLOPS)						

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

東京大学情報基盤センター INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

筑波大学 計算科学研究センター Center for Computational Sciences

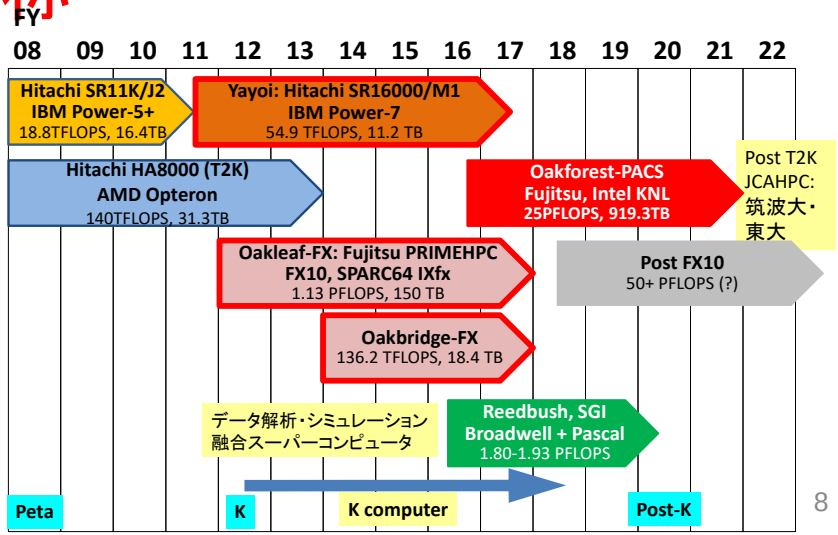
フラグシップとの両輪として



Oakforest-PACSの由来

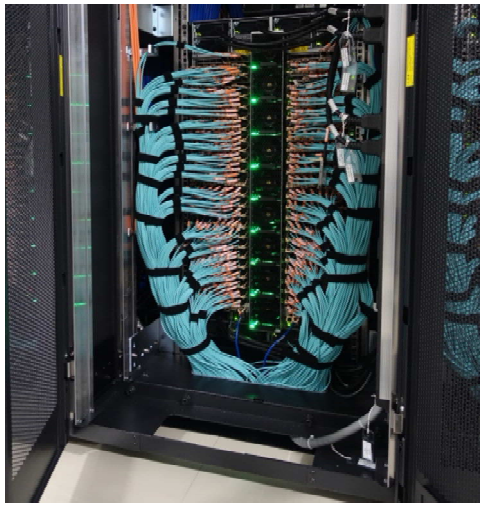
- Oakforest：東大
  - 2012年4月　東大情報基盤センター柏拠点開所
  - Oakleaf-FX（Fujitsu PRIMEHPC FX10，1+PF）運用開始
  - 1F：Oakleaf-XXX，2F：Oakforest-YYY
- PACS：筑波大
  - PACS： Parallel Advanced System for Computational Science
  - 9世代にわたって継承
- 両センターの思いの詰まった名称
  - 必ず「オークフォレストパックス」と呼んでください（OFPでも可）。

Year	System	Performance
1978	PACS-9 (PACS I)	7 KFLOPS
1980	PACS-32 (PACS II)	500 KFLOPS
1983	PAX-128 (PACS III)	4 MFLOPS
1984	PAX-32J (PACS IV)	3 MFLOPS
1989	QCDPAX (PACS V)	14 GFLOPS
1996	CP-PACS (PACS VI)	614 GFLOPS
2006	PACS-CS (PACS VII)	14.3 TFLOPS
2012	HA-PACS (PACS VIII)	1.166 PFLOPS
2014	COMA (PACS IX)	1.001 PFLOPS

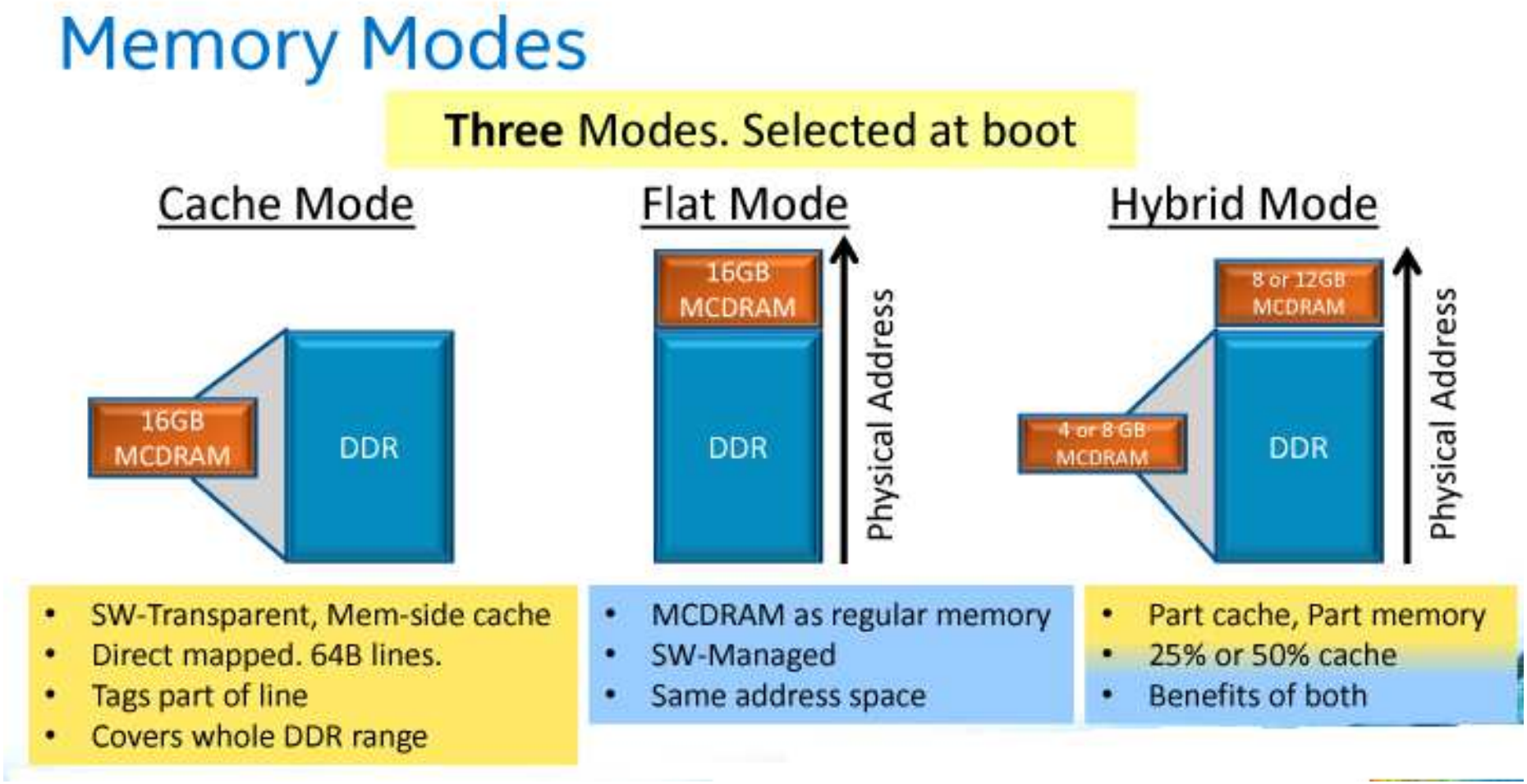


Oakforest-PACSの特徴

- 計算ノード
  - 1ノード 68コア，3TFLOPS×8,208ノード＝ 25 PFLOPS
  - MCDRAM（3D積層・高速，16GB）＋DDR4（低速，96GB）
    - 多様なメモリモード，クラスタリングモードの選択可能
- ノード間通信
  - フルバイセクションバンド幅を持つFat-Treeネットワーク
  - 全系運用時のアプリケーション性能に効果を発揮，多ジョブ運用
  - Intel Omni-Path Architecture
- ファイルI/O
  - 並列ファイルシステム
  - 高速ファイルキャッシュシステム（DDN IME）
    - 計算科学・ビッグデータ解析・機械学習にも貢献
- 消費電力
  - Green 500でも世界6位（2016年11月）
  - Linpack：4,986 MFLOPS/W（OFP），830 MFLOPS/W（京）

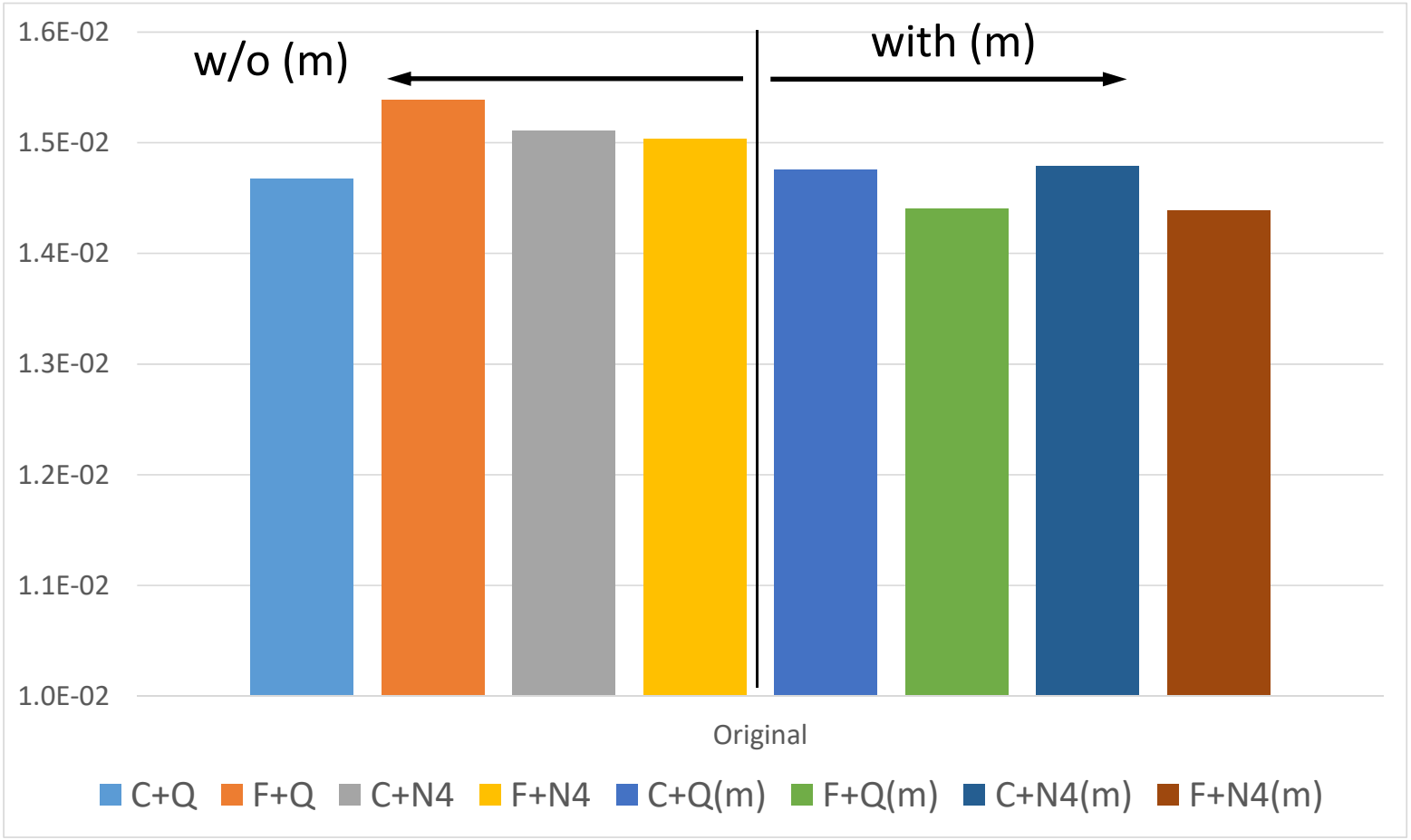


Memory Modes for KNL



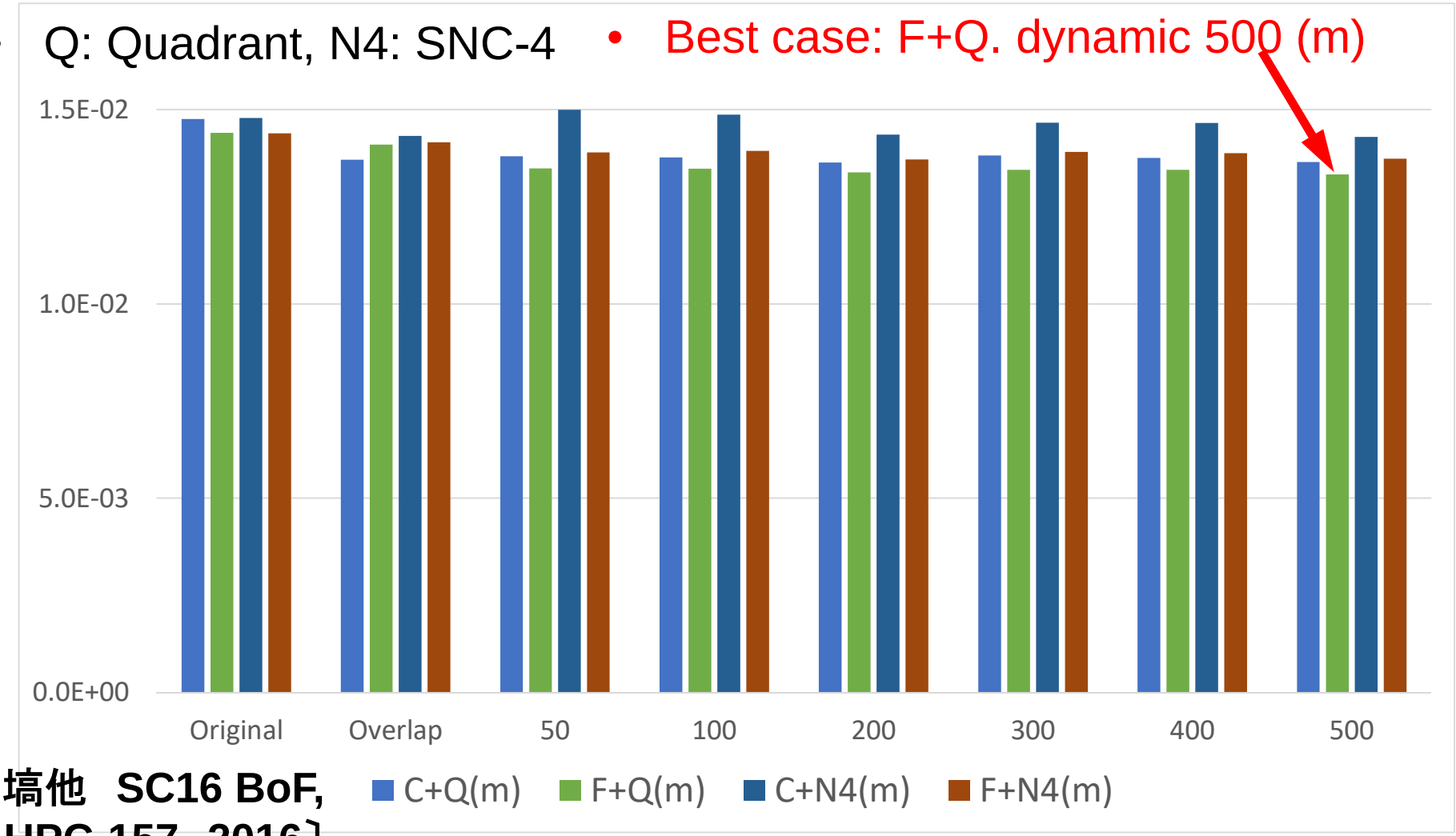
Results (OFP: 32 nodes) 〔塙他 SC16 BoF, HPC-157，2016〕 Parallel FEM (GeoFEM), HB 16x4

- C: Cache, F: Flat
- Q: Quadrant, N4: SNC-4
- (m) : core binding to avoid “busy” core
- (m) case are generally better



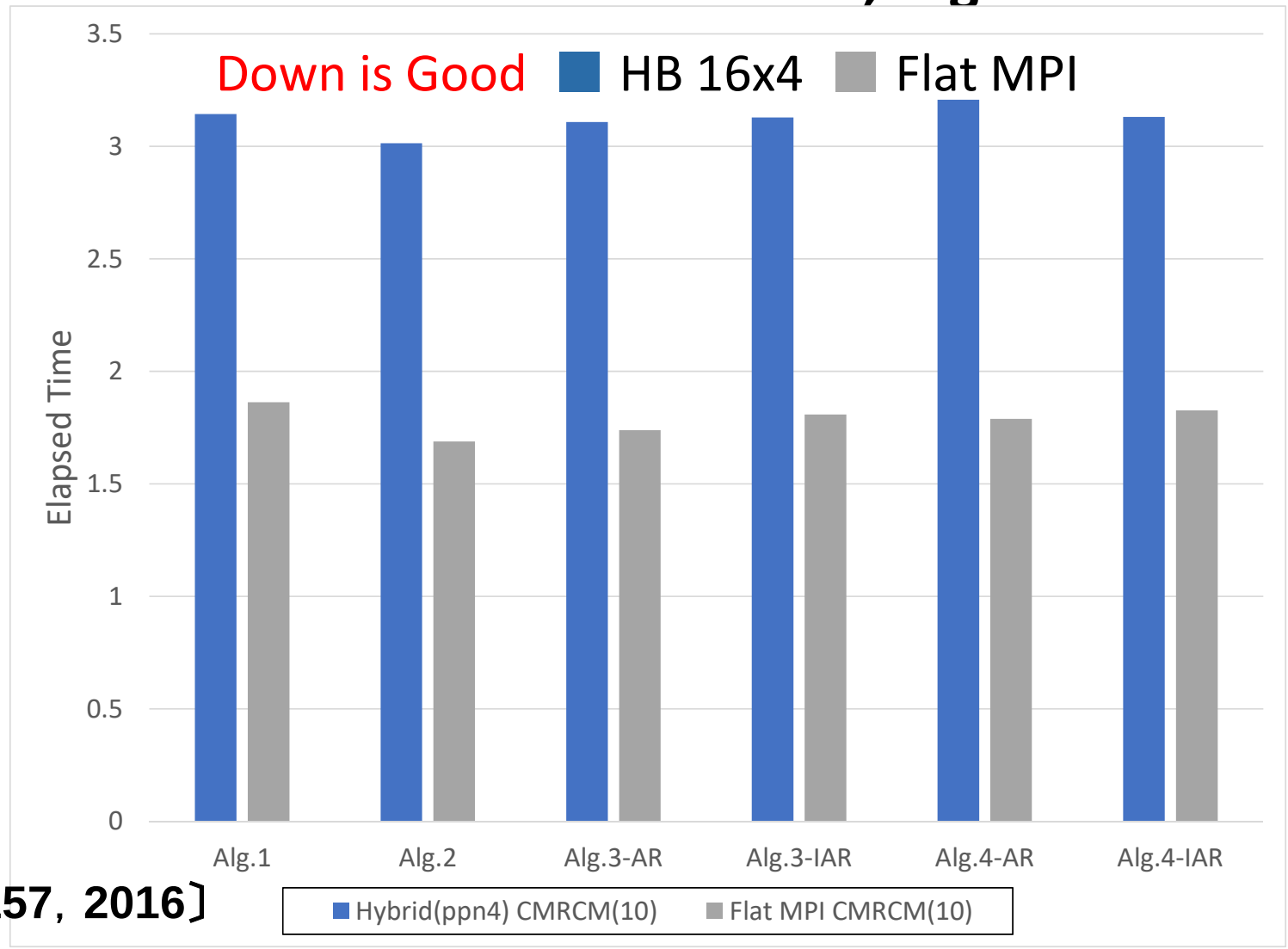
Results (OFP: 32 nodes): (m) Parallel FEM, HB 16x4 (16t x 4p)

- C: Cache, F: Flat
- Q: Quadrant, N4: SNC-4
- (m) : core binding to avoid “busy” core
- Best case: F+Q. dynamic 500 (m)



〔塙他 SC16 BoF, HPC-157, 2016〕

Comparisons at 8,192 cores
OFP: 128 nodes, Parallel FEM
Flat MPI is much faster than HB 16x4, Alg.2 is the best



〔塙他, HPC-157，2016〕

Oakforest-PACS の仕様

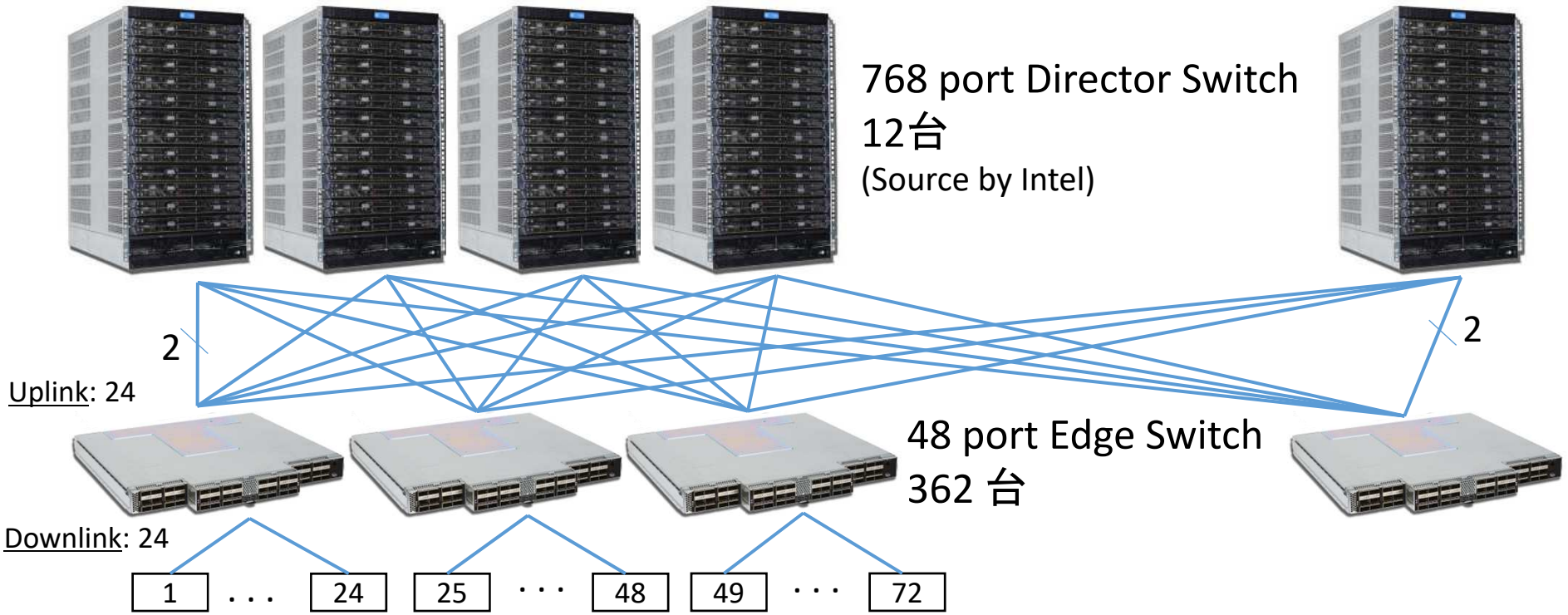
総ピーク演算性能			25 PFLOPS
ノード数			8,208
計算ノード	Product		富士通 PRIMERGY CX600 M1 (2U) + CX1640 M1 x 8node
	プロセッサ		Intel® Xeon Phi™ 7250 （開発コード: Knights Landing） 68 コア、1.4 GHz
	メモリ	高バンド幅	16 GB, MCDRAM, 実効 490 GB/sec
		低バンド幅	96 GB, DDR4-2400, ピーク 115.2 GB/sec
相互結合網	Product		Intel® Omni-Path Architecture
	リンク速度		100 Gbps
	トポロジ		フルバイセクションバンド幅Fat-tree網

計算ノードとシャーシ

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

Chassis with 8 nodes, 2U size

Intel® Omni-Path Architecture を用いた フルバイセクションバンド幅Fat-tree網



コストはかかるがフルバイセクションバンド幅を維持
- システム全系使用時にも高い並列性能を実現
- 柔軟な運用：ジョブに対する計算ノード割り当ての自由度が高い

Oakforest-PACS の仕様（続き）

並列ファイルシステム	Type	Lustre File System
	総容量	26.2 PB
	Product	DataDirect Networks SFA14KE
	総バンド幅	500 GB/sec
高速ファイルキャッシュシステム	Type	Burst Buffer, Infinite Memory Engine (by DDN)
	総容量	940 TB (NVMe SSD, パリティを含む)
	Product	DataDirect Networks IME14K
	総バンド幅	1,560 GB/sec
総消費電力		4.2MW（冷却を含む）
総ラック数		102

各種ベンチマーク

- TOP 500（Linpack，HPL（High Performance Linpack））
  - 連立一次方程式ソルバー（直接法），計算速度（FLOPS値）
  - 規則的な密行列：連続メモリアクセス
  - 計算性能
- HPCG
  - 連立一次方程式ソルバー（反復法），計算速度（FLOPS値）
  - 有限要素法から得られる疎行列（ゼロが多い）
    - 不連続メモリアクセス
    - 実アプリケーションに近い
  - メモリアクセス性能，通信性能
- Green 500
  - HPL（TOP500）実行時の消費電力量から求めたFLOPS/W値

49th TOP500 List (June, 2017)

	Site	Computer/Year Vendor	Cores	R_{max} (TFLOPS)	R_{peak} (TFLOPS)	Power (kW)	
1	National Supercomputing Center in Wuxi, China	**Sunway TaihuLight** , Sunway MPP, Sunway SW26010 260C 1.45GHz, 2016 NRCPC	10,649,600	93,015 (= 93.0 PF)	125,436	15,371	
2	National Supercomputing Center in Tianjin, China	**Tianhe-2**, Intel Xeon E5-2692, TH Express-2, Xeon Phi, 2013 NUDT	3,120,000	33,863 (= 33.9 PF)	54,902	17,808	
3	Swiss Natl. Supercomputer Center, Switzerland	**Piz Daint** Cray XC30/NVIDIA P100, 2013 Cray	361,760	19,590	33,863	2,272	
4	Oak Ridge National Laboratory, USA	**Titan** Cray XK7/NVIDIA K20x, 2012 Cray	560,640	17,590	27,113	8,209	
5	Lawrence Livermore National Laboratory, USA	**Sequoia** BlueGene/Q, 2011 IBM	1,572,864	17,173	20,133	7,890	
6	DOE/SC/LBNL/NERSC USA	**Cori**, Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Cray Aries, 2016 Cray	632,400	14,015	27,881	3,939	
7	Joint Center for Advanced High Performance Computing, Japan	**Oakforest-PACS**, PRIMERGY CX600 M1, Intel Xeon Phi Processor 7250 68C 1.4GHz, Intel Omni-Path, 2016 Fujitsu	557,056	13,555	24,914	2,719	**919 TB**
8	RIKEN AICS, Japan	**K computer**, SPARC64 VIIIfx , 2011 Fujitsu	705,024	10,510	11,280	12,660	**1,410TB**
9	Argonne National Laboratory, USA	**Mira** BlueGene/Q, 2012 IBM	786,432	8,587	10,066	3,945	
10	DOE/NNSA/LANL/SNL, USA	**Trinity**, Cray XC40, Xeon E5-2698v3 16C 2.3GHz, 2016 Cray	301,056	8,101	11,079	4,233	

R_{max}: Performance of Linpack (TFLOPS)
R_{peak}: Peak Performance (TFLOPS), Power: kW

http://www.top500.org/

HPCG Ranking (June, 2017)

	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Peak
1	**K computer**	705,024	10.510	8	0.6027	5.3%
2	**Tianhe-2 (MilkyWay-2)**	3,120,000	33.863	2	0.5801	1.1%
3	**Sunway TaihuLight**	10,649,600	93.015	1	0.4808	0.4%
4	**Piz Daint**	361,760	19.590	3	0.4767	1.9%
5	**Oakforest-PACS**	557,056	13.555	7	0.3855	1.5%
6	**Cori**	632,400	13.832	6	0.3554	1.3%
7	**Sequoia**	1,572,864	17.173	5	0.3304	1.6%
8	**Titan**	560,640	17.590	4	0.3223	1.2%
9	**Trinity**	301,056	8.101	10	0.1826	1.6%
10	**Pleiades** – NASA/SGI	243,008	5.952	15	0.1752	2.5%

http://www.hpcg-benchmark.org/

Green 500 Ranking (November, 2016)

	Site	Computer	CPU	HPL Rmax (Pflop/s)	TOP500 Rank	Power (MW)	GFLOPS/W
1	NVIDIA Corporation	DGX SATURNV	NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100	3.307	28	0.350	9.462
2	Swiss National Supercomputing Centre (CSCS)	Piz Daint	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100	9.779	8	1.312	7.454
3	RIKEN ACCS	Shoubu	ZettaScaler-1.6 etc.	1.001	116	0.150	6.674
4	National SC Center in Wuxi	Sunway TaihuLight	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	93.01	1	15.37	6.051
5	SFB/TR55 at Fujitsu Tech. Solutions GmbH	QPACE3	PRIMERGY CX1640 M1, Intel Xeon Phi 7210 64C 1.3GHz, Intel Omni-Path	0.447	375	0.077	5.806
6	JCAHPC	Oakforest-PACS	PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	1.355	6	2.719	4.986
7	DOE/SC/Argonne National Lab.	Theta	Cray XC40, Intel Xeon Phi 7230 64C 1.3GHz, Aries interconnect	5.096	18	1.087	4.688
8	Stanford Research Computing Center	XStream	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	0.781	162	0.190	4.112
9	ACCMS, Kyoto University	Camphor 2	Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect	3.057	33	0.748	4.087
10	Jefferson Natl. Accel. Facility	SciPhi XVI	KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path	0.426	397	0.111	3.837

http://www.top500.org/

Green 500 Ranking (June, 2017)

	Site	Computer	CPU	HPL Rmax (Pflop/s)	TOP500 Rank	Power (MW)	GFLOPS/W
1	Tokyo Tech.	TSUBAME3.0	SGI ICE XA, IP139-SXM2, Xeon E5-2680v4, NVIDIA Tesla P100 SXM2, HPE	1,998.0	61	142	14.110
2	Yahoo Japan	kukai	ZettaScaler-1.6, Xeon E5-2650Lv4,, NVIDIA Tesla P100 , Exascalar	460.7	465	33	14.046
3	AIST, Japan	AIST AI Cloud	NEC 4U-8GPU Server, Xeon E5-2630Lv4, NVIDIA Tesla P100 SXM2 , NEC	961.0	148	76	12.681
4	CAIP, RIKEN, JAPAN	RAIDEN GPU subsystem -	NVIDIA DGX-1, Xeon E5-2698v4, NVIDIA Tesla P100 , Fujitsu	635.1	305	60	10.603
5	Univ. Cambridge, UK	Wilkes-2 -	Dell C4130, Xeon E5-2650v4, NVIDIA Tesla P100 , Dell	1,193.0	100	114	10.428
6	Swiss Natl. SC. Center (CSCS)	Piz Daint	Cray XC50, Xeon E5-2690v3, NVIDIA Tesla P100 , Cray Inc.	19,590.0	3	2,272	10.398
7	JAMSTEC, Japan	Gyoukou,	ZettaScaler-2.0 HPC system, Xeon D-1571, PEZY-SC2 , ExaScalar	1,677.1	69	164	10.226
8	Inst. for Env. Studies, Japan	GOSAT-2 (RCF2)	SGI Rackable C1104-GP1, Xeon E5-2650v4, NVIDIA Tesla P100 , NSSOL/HPE	770.4	220	79	9.797
9	Facebook, USA	Penguin Relion	Xeon E5-2698v4/E5-2650v4, NVIDIA Tesla P100 , Acer Group	3,307.0	31	350	9.462
10	NVIDIA, USA	DGX Saturn V	Xeon E5-2698v4, NVIDIA Tesla P100 , Nvidia	3,307.0	32	350	9.462
11	ITC, U.Tokyo, Japan	Reedbush-H	SGI Rackable C1102-GP8, Xeon E5-2695v4, NVIDIA Tesla P100 SXM2 , HPE	802.4	203	94	8.575

http://www.top500.org/

Now operating 4 (or 6)systems !!

- Oakleaf-FX (Fujitsu PRIMEHPC FX10)
  - 1.135 PF, Commercial Version of K, Apr.2012 – Mar.2018

- Oakbridge-FX (Fujitsu PRIMEHPC FX10)
  - 136.2 TF, for long-time use (up to 168 hr), Apr.2014 – Mar.2018
- Reedbush (HPE, Intel BDW + NVIDIA P100 (Pascal))
  - Integrated Supercomputer System for Data Analyses & Scientific Simulations
    - Jul.2016-Jun.2020

  - Our first GPU System (Mar.2017), DDN IME (Burst Buffer)
  - **Reedbush-U: CPU only, 420 nodes, 508 TF (Jul.2016)**
  - **Reedbush-H: 120 nodes, 2 GPUs/node: 1.42 PF (Mar.2017)**
  - **Reedbush-L: 64 nodes, 4 GPUs/node: 1.43 PF (Oct.2017)**
- Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))
  - JCAHPC (U.Tsukuba & U.Tokyo)
  - 25 PF, #7 in 49th TOP 500 (June.2017) (#1 in Japan)
  - Omni-Path Architecture, DDN IME (Burst Buffer)

JPY (=Watt)/GFLOPS Rate
Smaller is better (efficient)

System	JPY/GFLOPS
Oakleaf/Oakbridge-FX (Fujitsu) (Fujitsu PRIMEHPC FX10)	125
Reedbush-U (SGI) (Intel BDW)	62.0
Reedbush-H (SGI) (Intel BDW+NVIDIA P100)	17.1
Oakforest-PACS (Fujitsu) (Intel Xeon Phi/Knights Landing)	16.5

運 用

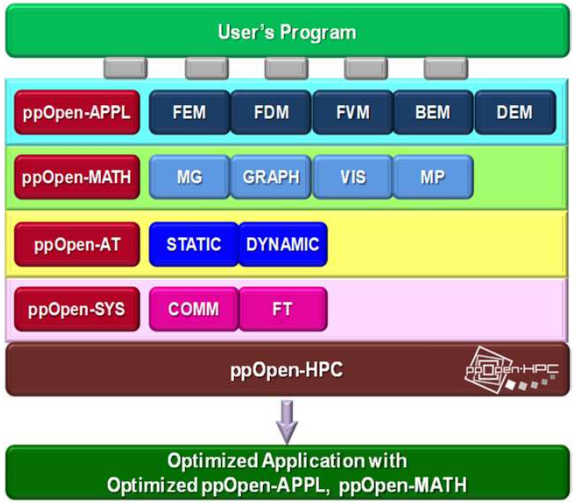
- 計算資源は全系を共用（パーティション分けはしない）
  - ✓全8,208ノード（25PF）を常に全系で運用できるようにしておき，国内最大の計算資源を有効に活用する
  - ✓Flat/Quadrant，Cache/Quadrantで半々
- 様々な利用形態
  - ✓各大学独自の利用コース
    - グループ利用（最大2,048ノード）
    - パーソナル利用（東大のみ）
    - 企業利用（東大のみ）
    - 国際共同利用（グループ利用の一部）
  - ✓HPCI
    - 全資源の20%を「JCAHPC」として拠出，企業利用可能
  - ✓JHPCN（学際大規模情報基盤共同利用共同研究拠点）
    - 全資源の5%程度：企業共同研究，国際共同研究も含む（東大のみ）
  - ✓教育（講義，講習会）
  - ✓大規模HPCチャレンジ：全ノード占有：公募中（非ユーザも可）

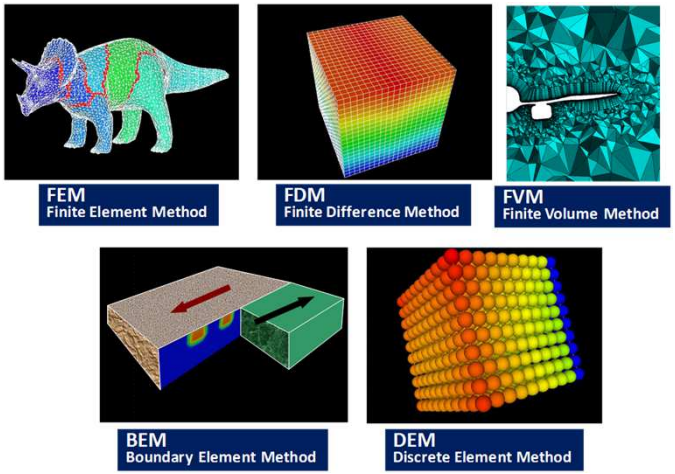
筑波大学 計算科学研究センター	COMA(PACS-IX) ▸資源提供元	Xeon +Phi(KNC)	計算ノード：90ノード(約230TFLOPS) 1,800コア （＋メニーコアアーキテクチャ 10,980コア） 資源量：756,000ノード時間積 ストレージ：300TB
最先端共同HPC基盤施設 (JCAHPC)※	Oakforest-PACS ▸資源提供元	Xeon Phi(KNL)	計算ノード：1,600ノード(4,872 TFLOPS) 108,800コア 資源量：13,824,000ノード時間積 ストレージ：3,000TB
東京大学 情報基盤センター	スーパーコンピュータ FX10 ▸資源提供元	SPARC	計算ノード：1,500ノード (354.78TFLOPS) 24,000コア 資源量：12,960,000ノード時間積 ストレージ：500TB
	Reedbush-U ▸資源提供元	Xeon	計算ノード：60ノード(72.58TFLOPS) 2,160コア 資源量：518,400ノード時間積 ストレージ：60TB
	Reedbush-H ▸資源提供元	Xeon +Tesla	計算ノード：18ノード (193.11-212.73TFLOPS) コア数 648 + GPU 36枚 資源量：155,520ノード時間積 ストレージ：18TB

Oakforest-PACS のソフトウェア

- OS: Red Hat Enterprise Linux (ログインノード), CentOS および McKernel (計算ノード, 切替可能)
  - McKernel: 理研AICSで開発中のメニーコア向けOS
    - Linuxに比べ軽量, ユーザプログラムに与える影響なし
    - ポスト京にも搭載される予定。
- コンパイラ: GCC, Intel Compiler, XcalableMP
  - XcalableMP: 理研AICS・筑波大で開発中の並列プログラミング言語
- ライブラリ・アプリケーション: オープンソースソフトウェア
  - ppOpen-HPC, OpenFOAM, ABINIT-MP, PHASE system, FrontFlow/blue, LAPACK, ScaLAPACK, PETSc, METIS, SuperLU etc.

McKernel





ppOpen-HPC: Overview

- Application framework with automatic tuning (AT)
  - ✓ “pp” : post-peta-scale
- Five-year project (FY.2011-2015) (since April 2011)
  - ✓ P.I.: Kengo Nakajima (ITC, The University of Tokyo)
  - ✓ Part of “Development of System Software Technologies for Post-Peta Scale High Performance Computing” funded by JST/CREST (Supervisor: Prof. Mitsuhisa Sato, RIKEN AICS)
- Target: Oakforest-PACS (Original Schedule: FY.2015)
  - ✓ could be extended to various types of platforms
- Team with 7 institutes, >50 people (5 PDs) from various fields: Co-Design
- Open Source Software
  - ✓ http://ppopenhpc.cc.u-tokyo.ac.jp/
  - ✓ English Documents, MIT License

独立行政法人
科学技術振興機構
Japan Science and Technology Agency

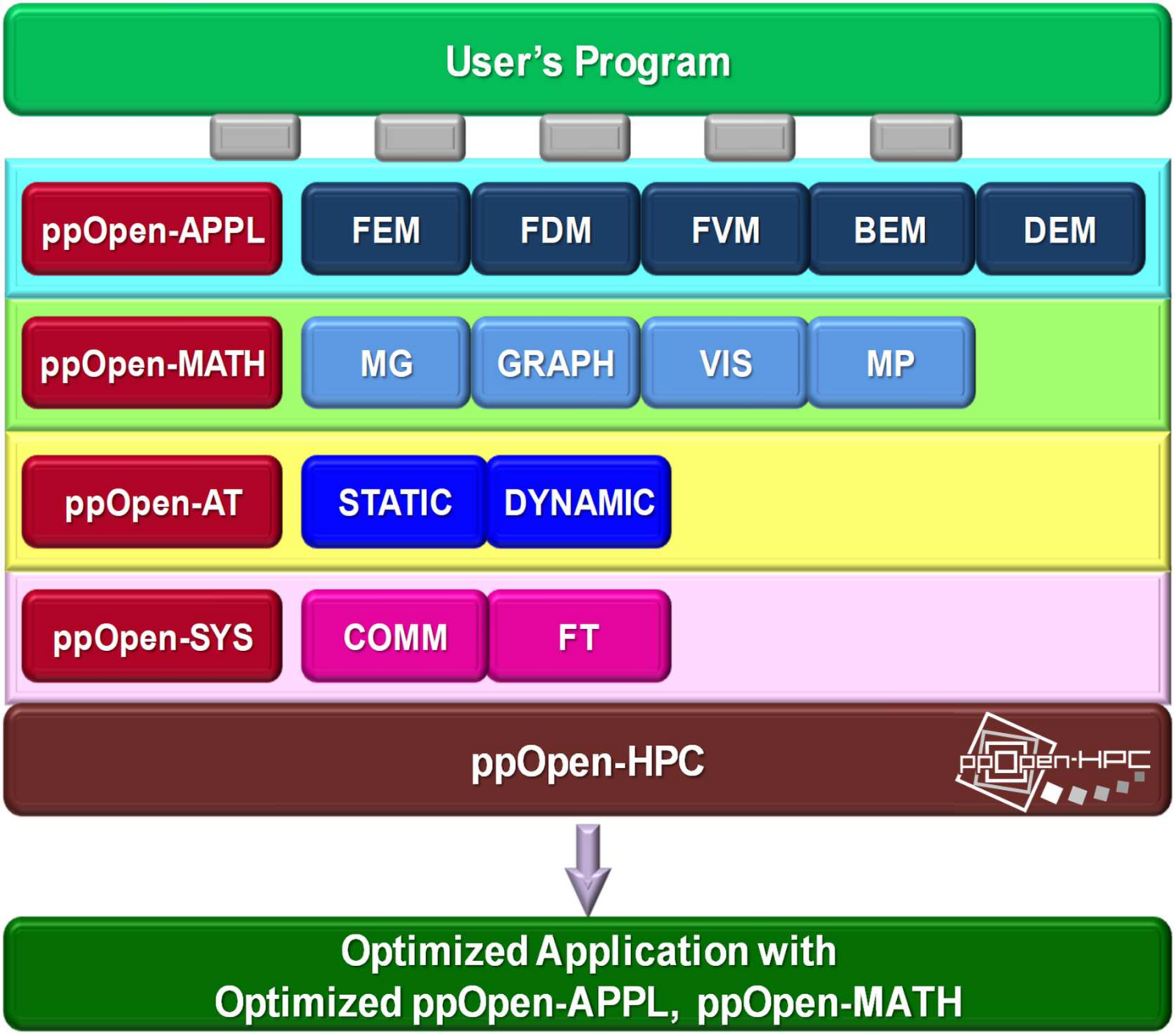
CREST
戦略的創造研究推進事業
Core Research for Evolutionary Science and Technology

Framework Appl. Dev.

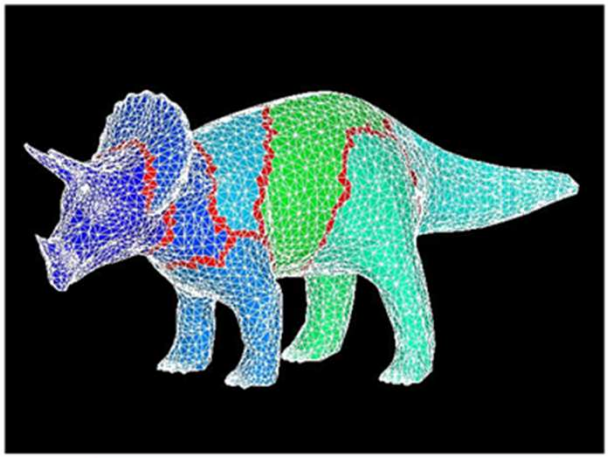
Math Libraries

Automatic Tuning (AT)

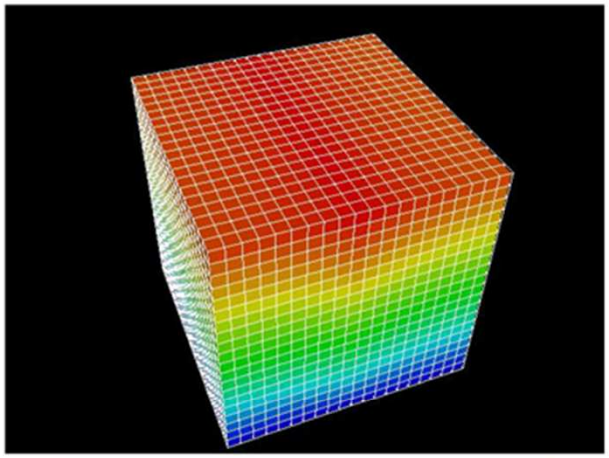
System Software



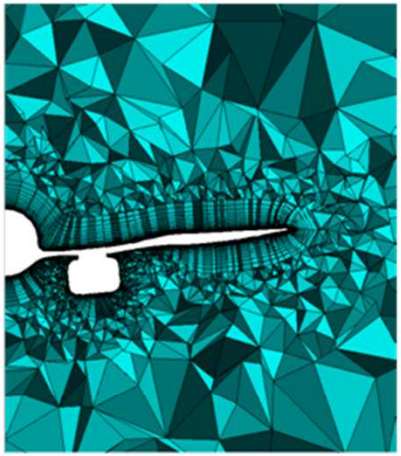
ppOpen-HPC covers …



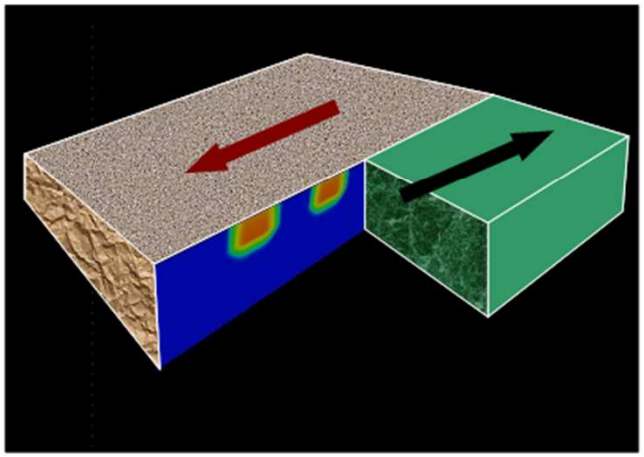
FEM
Finite Element Method



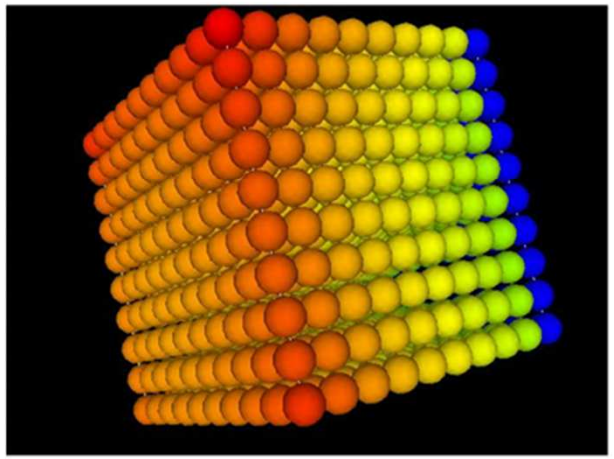
FDM
Finite Difference Method



FVM
Finite Volume Method



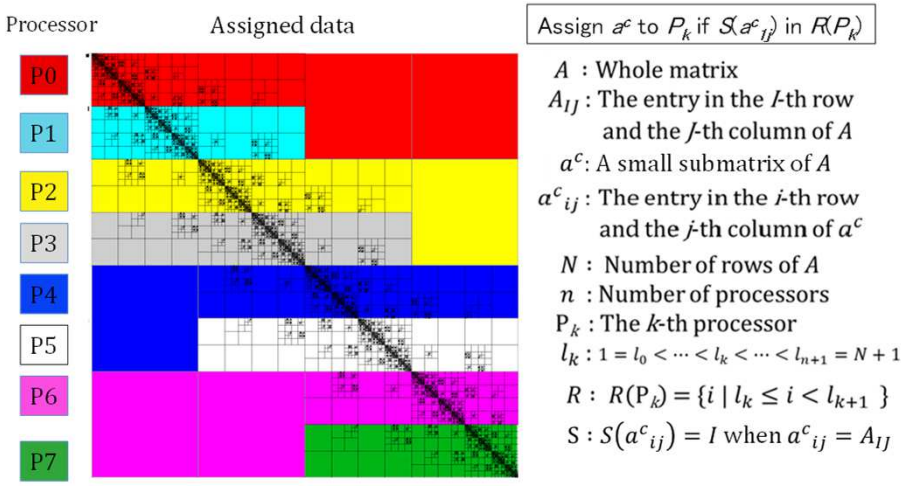
BEM
Boundary Element Method

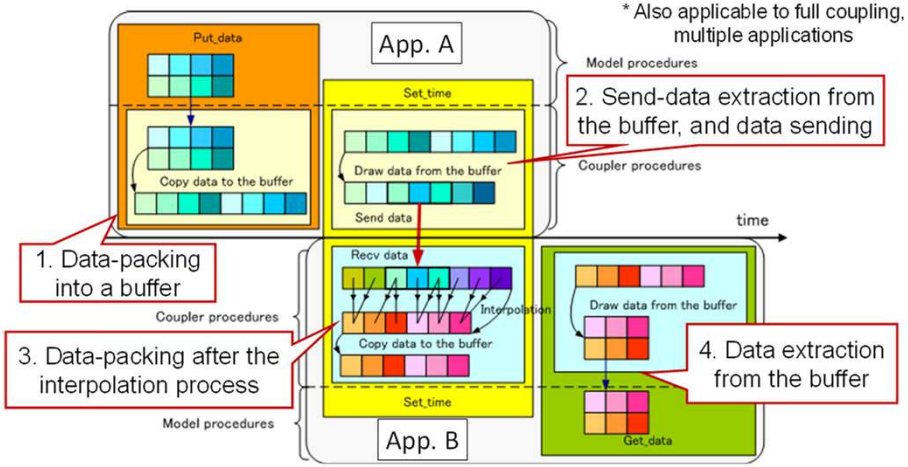


DEM
Discrete Element Method

Featured Developments

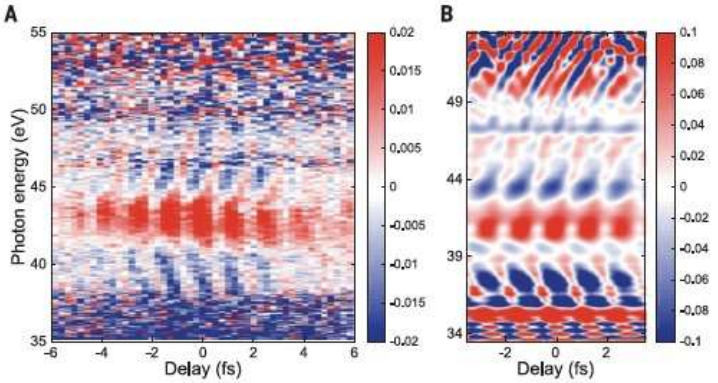
- **ppOpen-AT: AT Language for Loop Optimization**
- HACApK library for H-matrix comp. in ppOpen-APPL/BEM (OpenMP/MPI Hybrid Version)
  - First Open Source Library by OpenMP/MPI Hybrid
- ppOpen-MATH/MP (Coupler for Multiphysics Simulations, Loose Coupling of FEM & FDM)
- Sparse Linear Solvers

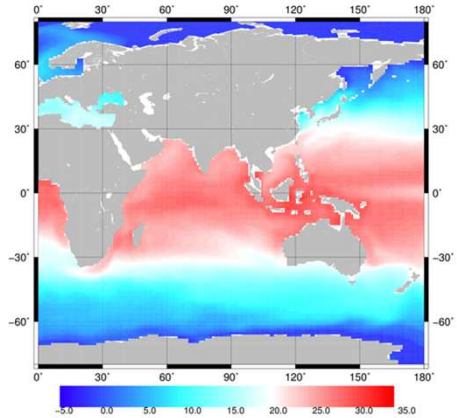


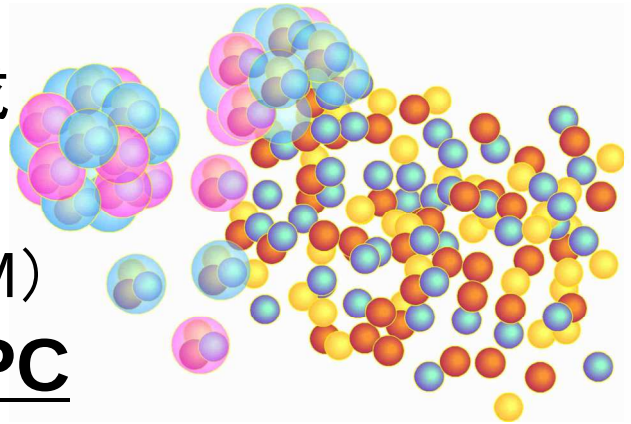


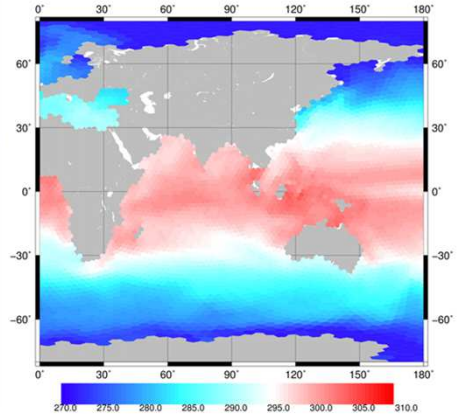
Oakforest-PACS：代表的アプリケーション

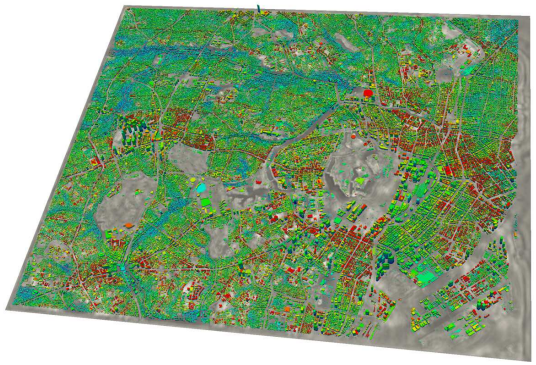
- **SALMON/ARTE**
  - 電子動力学
- **Lattice QCD**
  - 格子量子色力学
- **NICAM-COCO**
  - 全地球大気・海洋連成
- **GHYDRA**
  - 三次元地盤震動（FEM）
- **Seism3D/OpenSWPC**
  - 三次元広域波動伝搬（FDM）

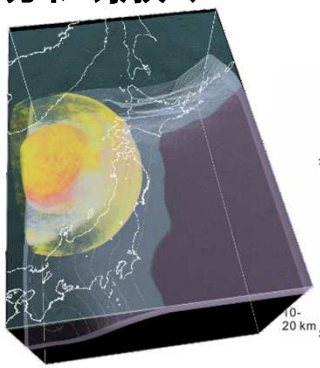


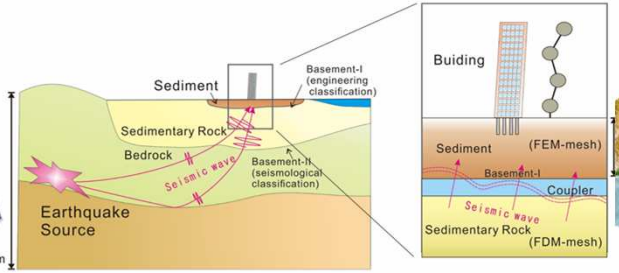


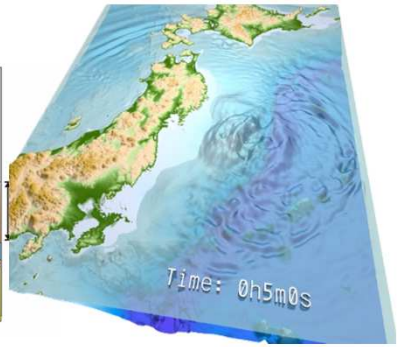




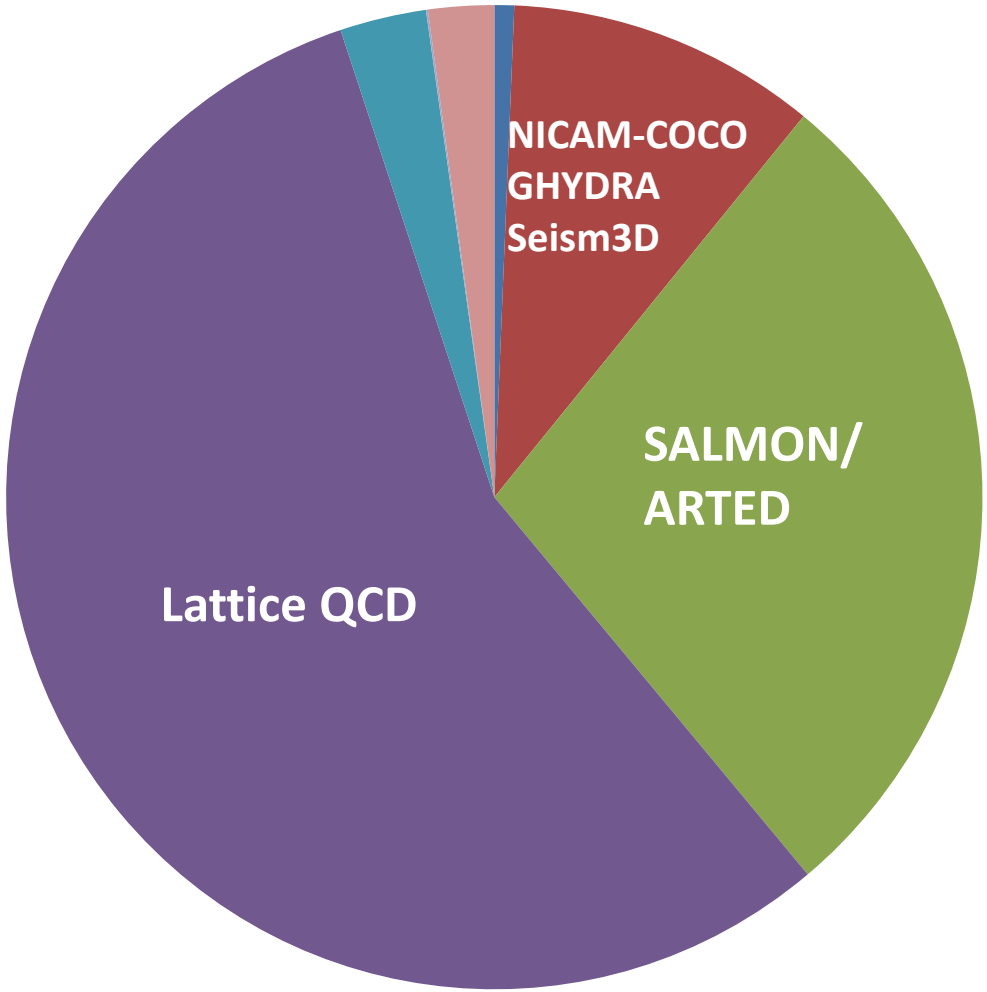






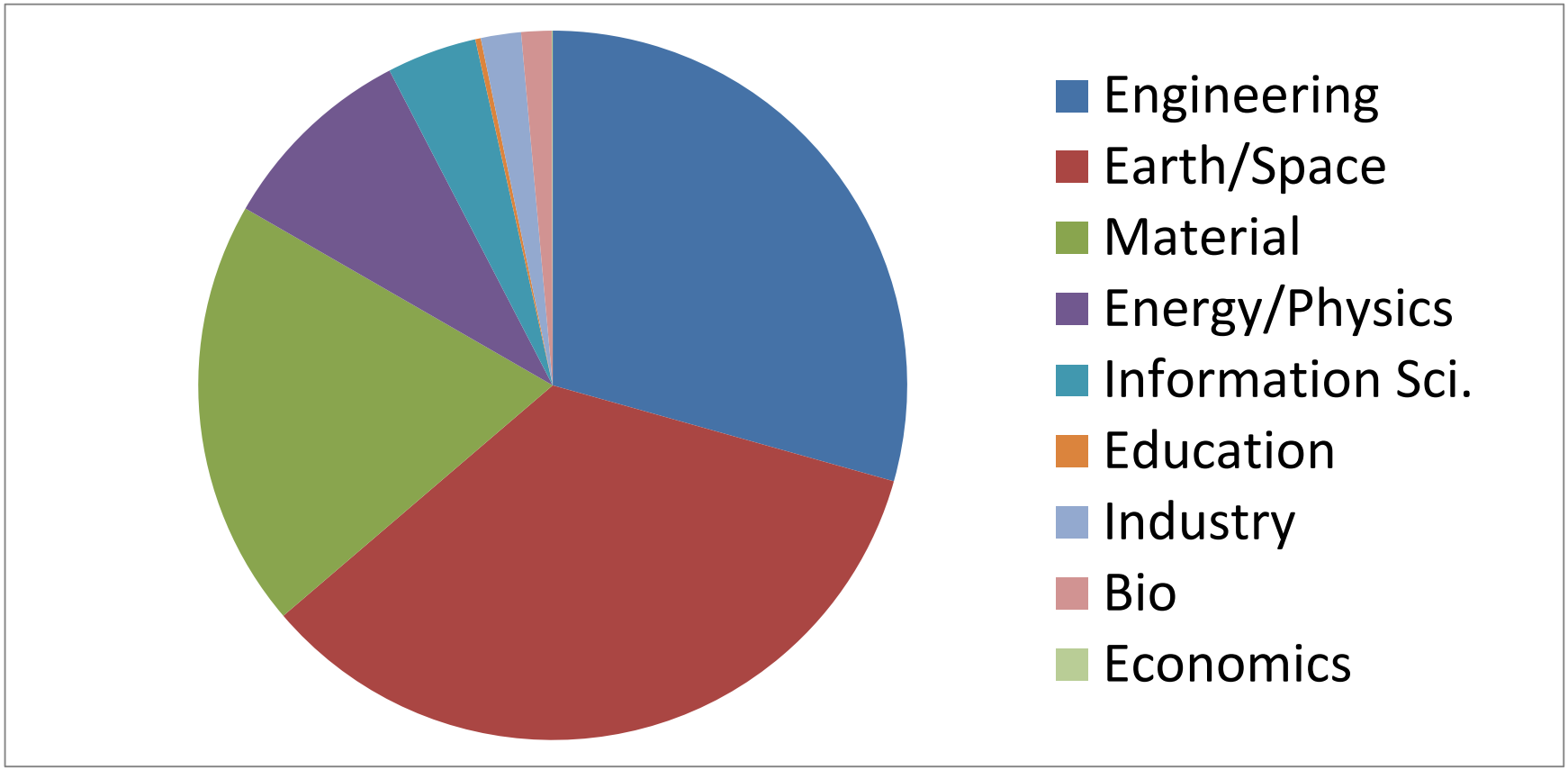


Research Area based on CPU Hours Oakforest-PACS in FY.2017 (2017.4~2017.9M)



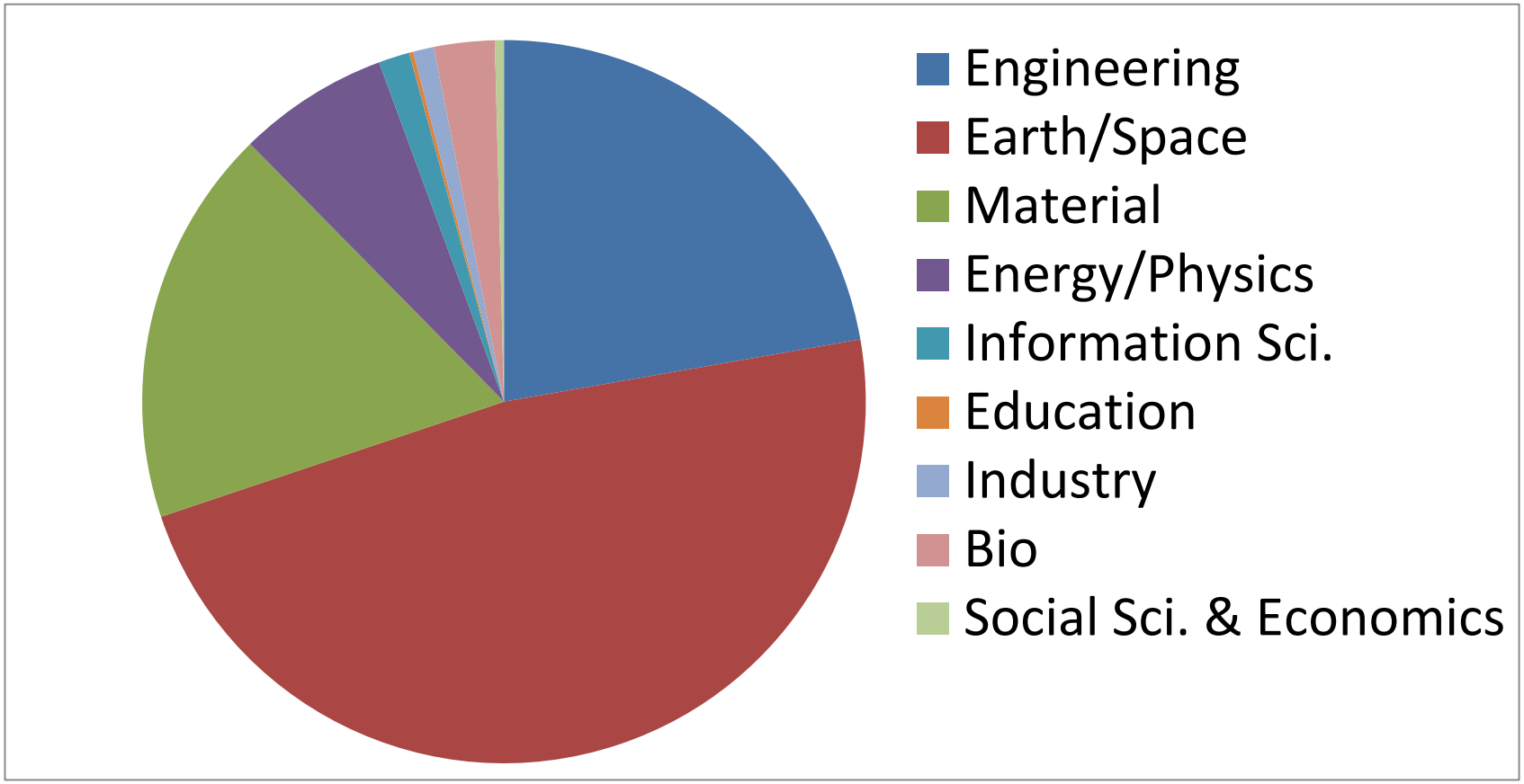
- Engineering
- Earth/Space
- Material
- Energy/Physics
- Information Sci.
- Education
- Industry
- Bio
- Social Sci. & Economics
- Data

Research Area based on CPU Hours FX10 in FY.2015 (2015.4~2016.3E)



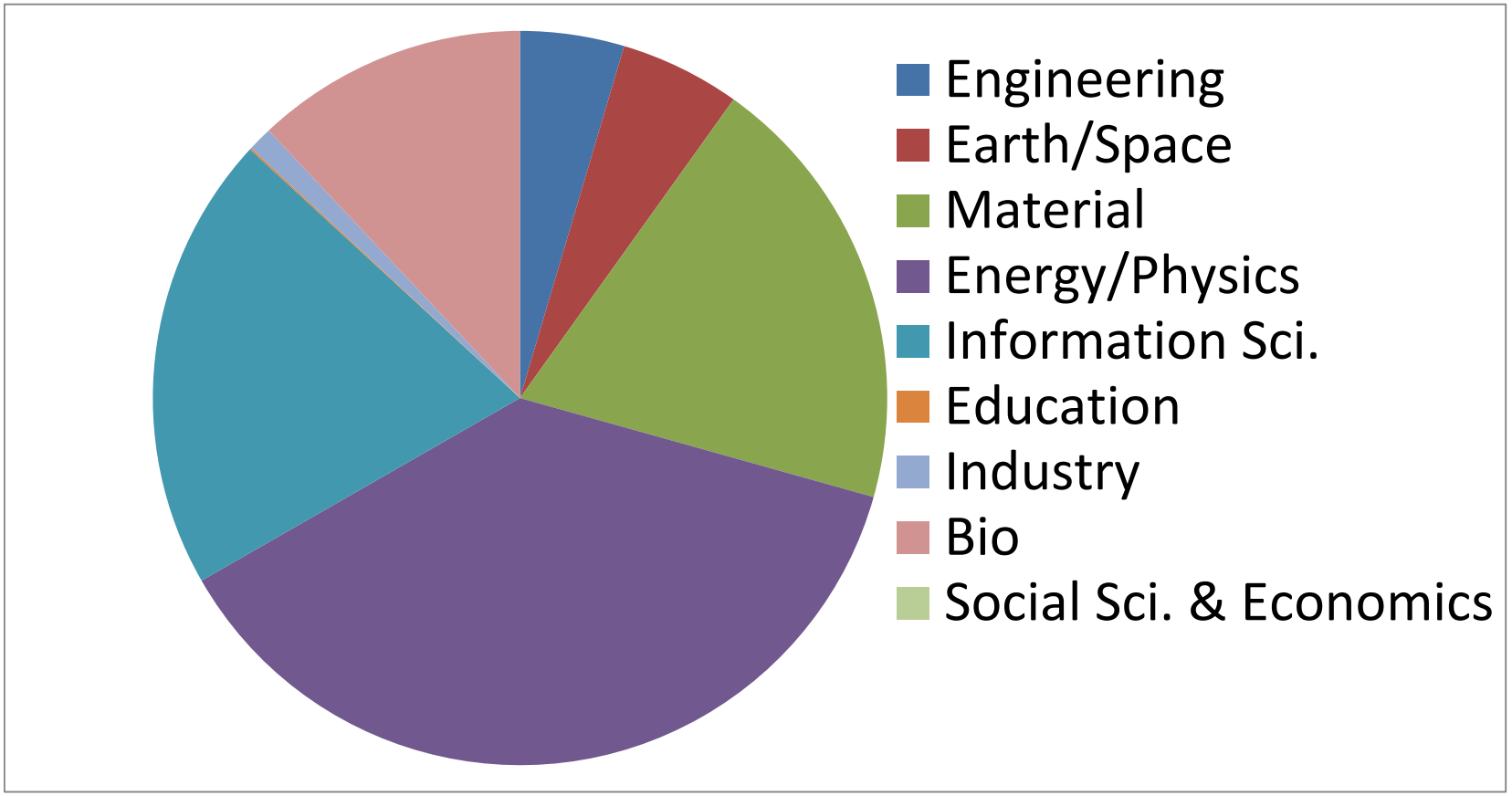
Oakleaf-FX + Oakbridge-FX

Research Area based on CPU Hours FX10 in FY.2016 (2016.4~2017.3E)



Oakleaf-FX + Oakbridge-FX

Research Area based on CPU Hours Reedbush-U in FY.2016 (2016.7~2017.3E)



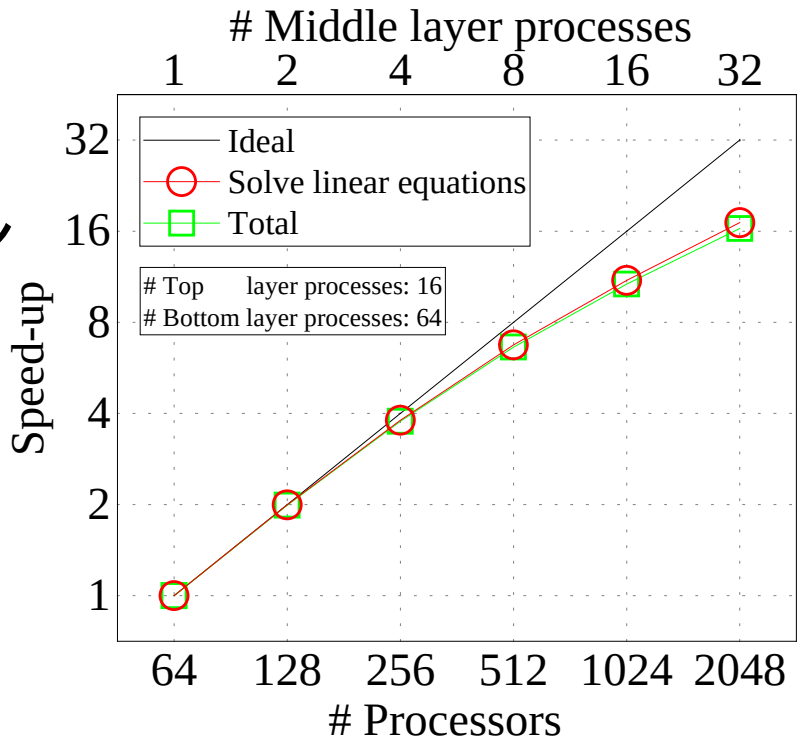
第1回OFP利活用報告会（5th JCAHPCセミナー）
2017年10月12日（木）13:30-18:00

- **13:30 - 13:45　中島研吾（JCAHPC/東京大学情報基盤センター）**
  - 開会　システム概要
- **13:45 - 14:15　斉藤圭亮（東京大学）**
  - 光合成蛋白質内の水分子のシミュレーション
- **14:15 - 14:45　佐々木勝一（東北大学）**
  - 物理点格子QCDシミュレーションによる核子構造の研究
- **14:45 - 15:15　廣川祐太（筑波大学）**
  - 電子動力学シミュレーションARTEDのOakforest-PACSでの全系性能評価
- **15:30 - 16:00　井戸村泰宏（日本原子力研究開発機構）**
  - Oakforest-PACSにおける大規模CFDコードの計算技術開発
- **16:00 - 16:30　八代　尚（理化学研究所計算科学研究機構）**
  - 『超』並列時代の気象・気候シミュレーション：メニーコアは怖くないぞ！
- **16:30 - 17:00　渡邊啓正（HPCシステムズ株式会社）**
  - Oakforest-PACS上でのFMO計算プログラムABINIT-MPの性能評価
- **17:10 - 17:40　講演者，センター教員**
  - 質疑応答
- **17:40 - 17:45　朴泰祐（JCAHPC/筑波大学計算科学研究センター）**
  - 閉会（終了後希望者のみでOakforest-PACSを見学）

各アプリ最適化

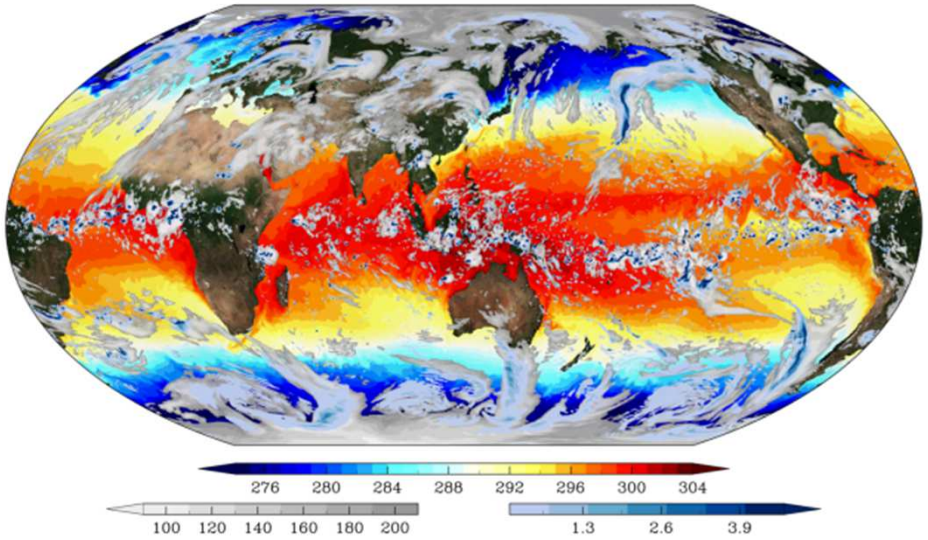
- SALMON，Lattice QCD，GAMERA/GHYDRA
  - OFP全系での稼働実績あり，ピーク性能の12-16%を達成済

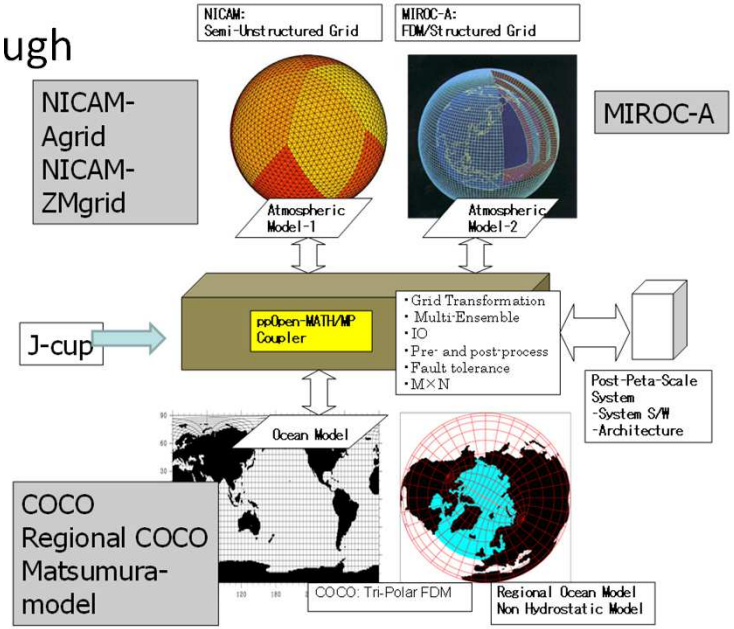
- z-Pares
  - 筑波大学　櫻井鉄也教授
  - 積分型疎行列固有値解析エンジン
  - 1万原子系の複素バンド構造計算
  - 3300万次元の非線形固有値問題
  - OFPの2048ノードまでを使用，高いスケーラビリティを確認
  - SC17のPaperに採択

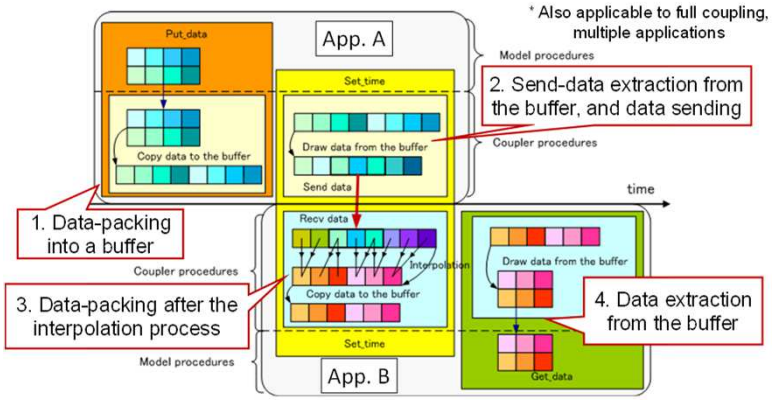


Atmosphere-Ocean Coupling
on OFP by NICAM/COCO/ppOpen-MATH/MP

- 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]

NICAM-COCO連成

東京大学(東大)は9月20日、雲の生成・消滅を詳細に計算できる全球大気モデルNICAMに、全球気候モデルMIROCの海洋部分を連結させた気象-海洋モデル「NICAM-COCO(NICOCO)」を開発し、熱帯域を東進する巨大な雲群マッデン＝ジュリアン振動(MJ0)と、東太平洋の海面温度が通常より高くなるエルニーニョ現象との相互作用の再現を可能にしたと発表した。

NICAM-COCOのカプラー：
ppOpen-MATH/MG

- 現状：全球大気14 km・全球海洋0.25-1.00°メッシュ（京）
- OFP：大気3.5km・海洋0.10°（従来の20倍超，50億メッシュ）

2017.9.20
日経(夕)

「スーパーエルニーニョ」終息

巨大雲 引き金に

東大など解明

東京大学大気海洋研究所と理化学研究所、海洋研究開発機構は、1997～98年に発生して世界に異常気象をもたらした「スーパーエルニーニョ」現象が急に終息した原因を突き止めた。熱帯の巨大雲が引き金になっていた。台風の発生予測などの精度向上にもつながる成果という。

エルニーニョは太平洋の東側の海面温度が平年より高くなる現象。特に強まった状態はスーパーエルニーニョと呼ばれ、97～98年のほか2015～16年にも発生した。世界各地で豪雨など異常気象の一因になったと考えられている。

研究チームは雲の発生や動きを計算する手法を改良し、海洋との相互作用を精度良く予測する手法を開発した。理研のスーパーコンピューター「京」を使った数値計算で、赤道付近に巨大な雲ができて東に進む「マッデン・ジュリアン振動」と呼ぶ現象がスーパーエルニーニョ終息の引き金になったと突き止めた。

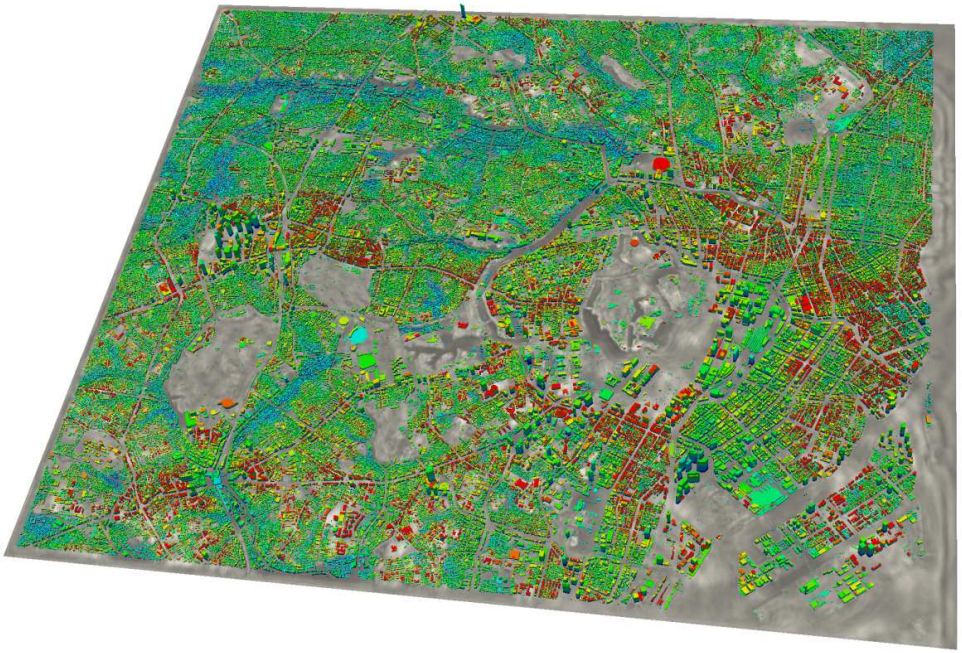
東西方向の数千キロに広がる巨大な雲から吹き出す東風が強まって海面付近の温かい海水が動かされ、入れ替わりに冷たい海水が下から上がってくるようになったという。

2017年9月20日 日経

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.

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

- GAMERA
  - multi-**G**rid method **A**daptive conjugate gradient method, **M**ulti-precision arithmetic, **E**lement-by-element method, p**R**edictor with **A**dams-bashworth method

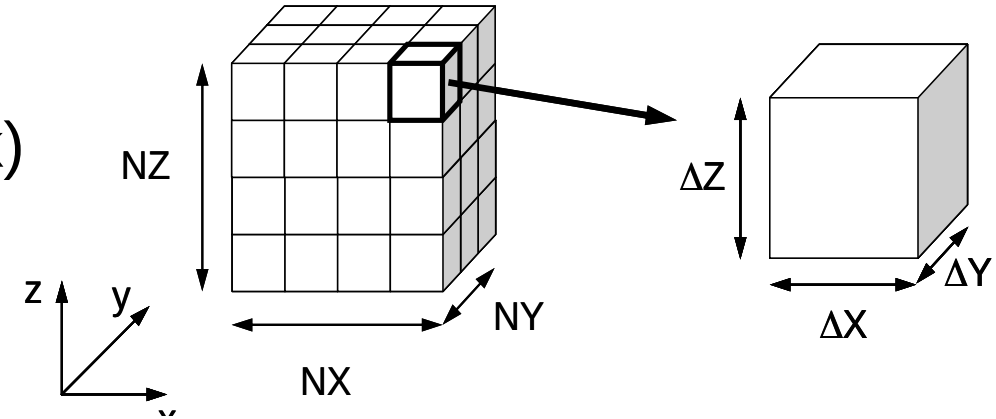
- GHYDRA
  - Great-HYDRA
  - **HY**bird tempo-spatial-arithmetic multi-gri**D** solve**R** with concentrated comput**A**tion

GAMERA

- **Poisson 3D** **ICCG for FVM (7-pt)**
- **GeoFEM/Cube** **Matrix Assembly for FEM**
- **GeoFEM/Heat** **CG with SELL-C-σ**
- **Dynamic Loop Scheduling of Parallel FEM [ICPP 2017]**

Poisson3D-OMP

- Finite Volume Method, Poisson Equations (128^3 cells)
  - FDM-type mesh (7-pt. Stencil), Unstructured data structure
  - SPD matrix
  - ICCG: Data Dependency for Incomplete Cholesky Fact.
- Fortran 90 + OpenMP
- Thread Parallelization by OpenMP: Reordering needed
  - CM-RCM + Coalesced/Sequential
- Storage of Matrix
  - CRS, ELL (Ellpack-Itpack)
- Outer Loops
  - Row-Wise, Column-Wise

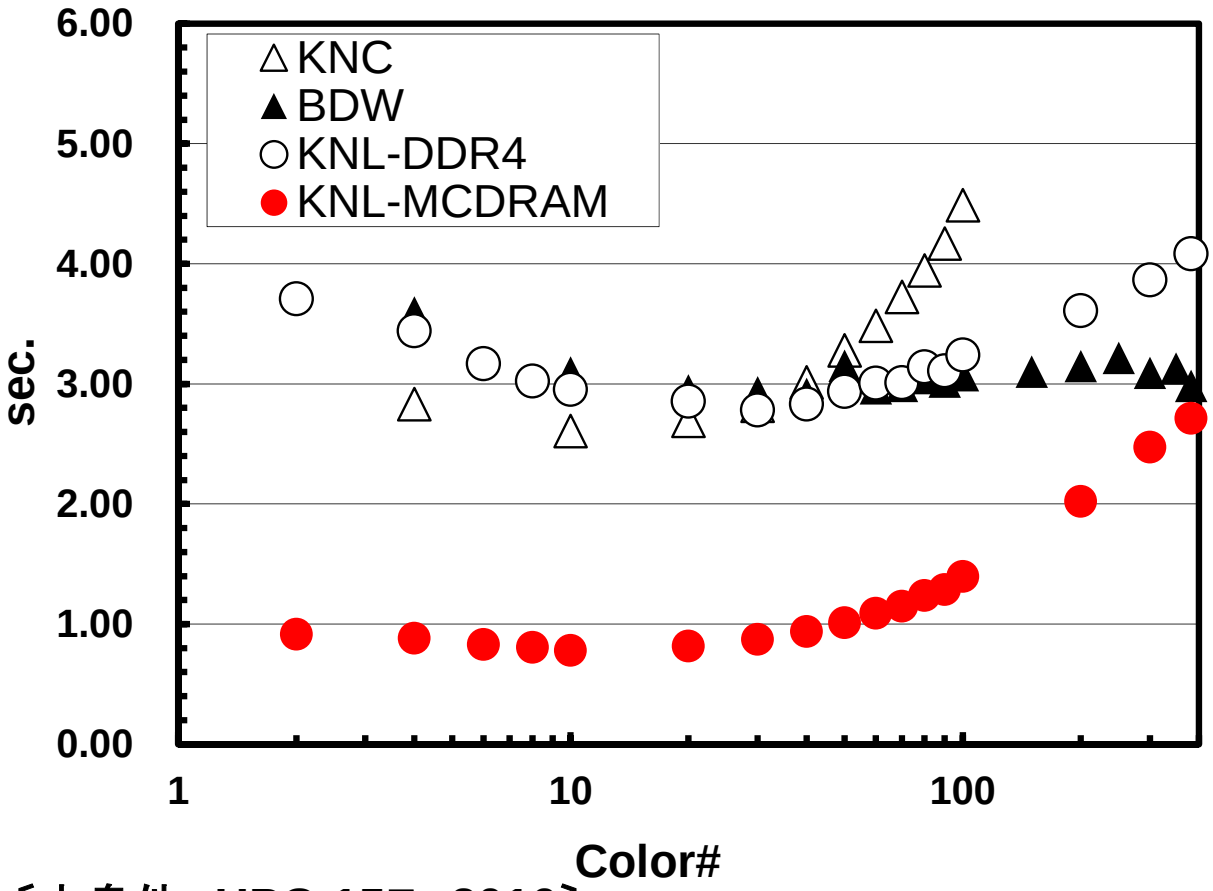


Code Name	KNC	KNL-0	BDW
Architecture	Intel Xeon Phi 5110P (Knights Corner)	Intel Xeon Phi 7210 (Knights Landing)	Intel Xeon E5-2695 v4 (Broadwell-EP)
Frequency (GHz)	1.053	1.30	2.10
Core # (Max Thread #)	60 (240)	64 (256)	18 (18)
Peak Performance (GFLOPS)	1,010.9	2,662.4	604.8
Memory (GB)	8	MCDRAM: 16 DDR4: 96	128
Memory Bandwidth(GB/sec., Stream Triad)	159	MCDRAM: 454 DDR4: 72.5	65.5
Out-of-Order	N	Y	Y
System		OFP-mini	Reedbush-U

Comp. Time for ICCG (Best ELL)

Effects of synchronization overhead are significant on KNL & KNC, if number of colors is larger
Generally, optimum number of color is 10 for KNL/KNC

Down is Good !!



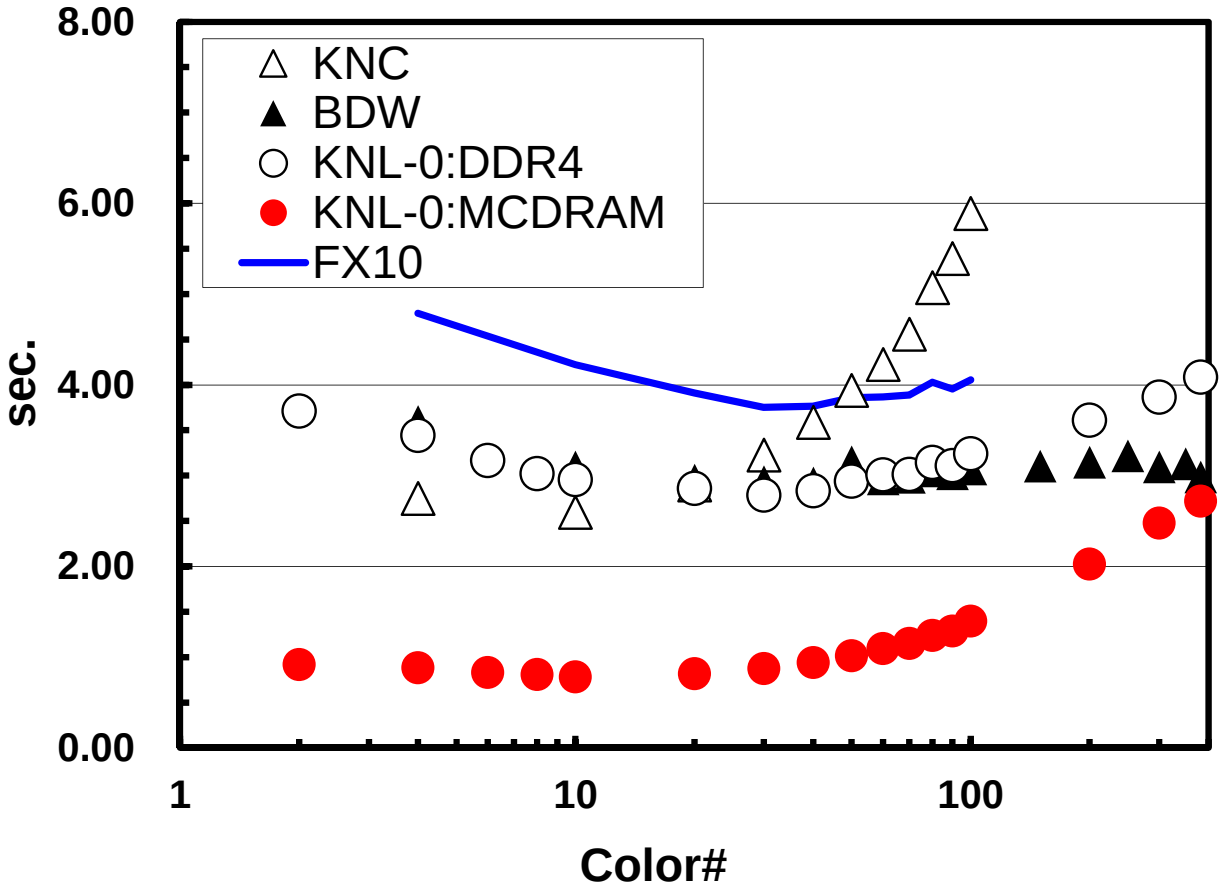
〔中島他 HPC-157，2016〕

- KNC/KNL-MCDRAM比は妥当
- メモリバンド幅を考慮するとKNC，KNL-MCDRAMはこの倍の性能が出ても良い（現在ピークの1%）
- BDW，KNL-DDR4はTriad並にメモリバンド幅使い切り
- 7-ptステンシルなので性能は出にくい

Comp. Time for ICCG (Best ELL)

Effects of synchronization overhead are significant on KNL & KNC, if number of colors is larger
Generally, optimum number of color is 10 for KNL/KNC

Down is Good !!



- FX10は60GB/sec位出ている@20色

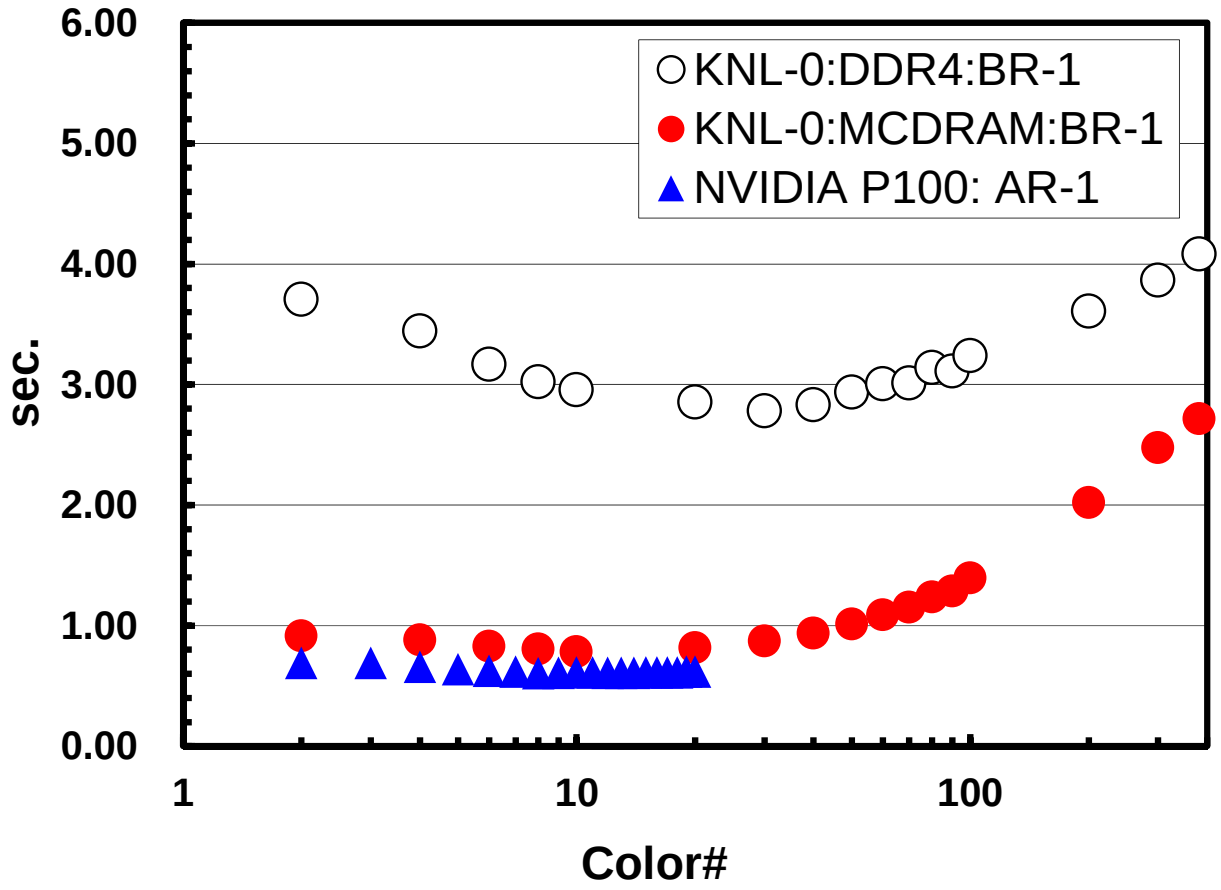
〔中島他 HPC-157，2016〕

Code Name	KNC	KNL-0	P100
Architecture	Intel Xeon Phi 5110P (Knights Corner)	Intel Xeon Phi 7210 (Knights Landing)	NVIDIA Tesla P100 (Pascal)
Frequency (GHz)	1.053	1.30	1.328
Core # (Max Thread #)	60 (240)	64 (256)	1,792
Peak Performance (GFLOPS)	1,010.9	2,662.4	4,759
Memory (GB)	8	MCDRAM: 16 DDR4: 96	16
Memory Bandwidth(GB/sec., Stream Triad)	159	MCDRAM: 454 DDR4: 72.5	530

Comp. Time for ICCG (Best ELL)

Xeon Phi (KNL) and P100 are competitive

Down is Good !!

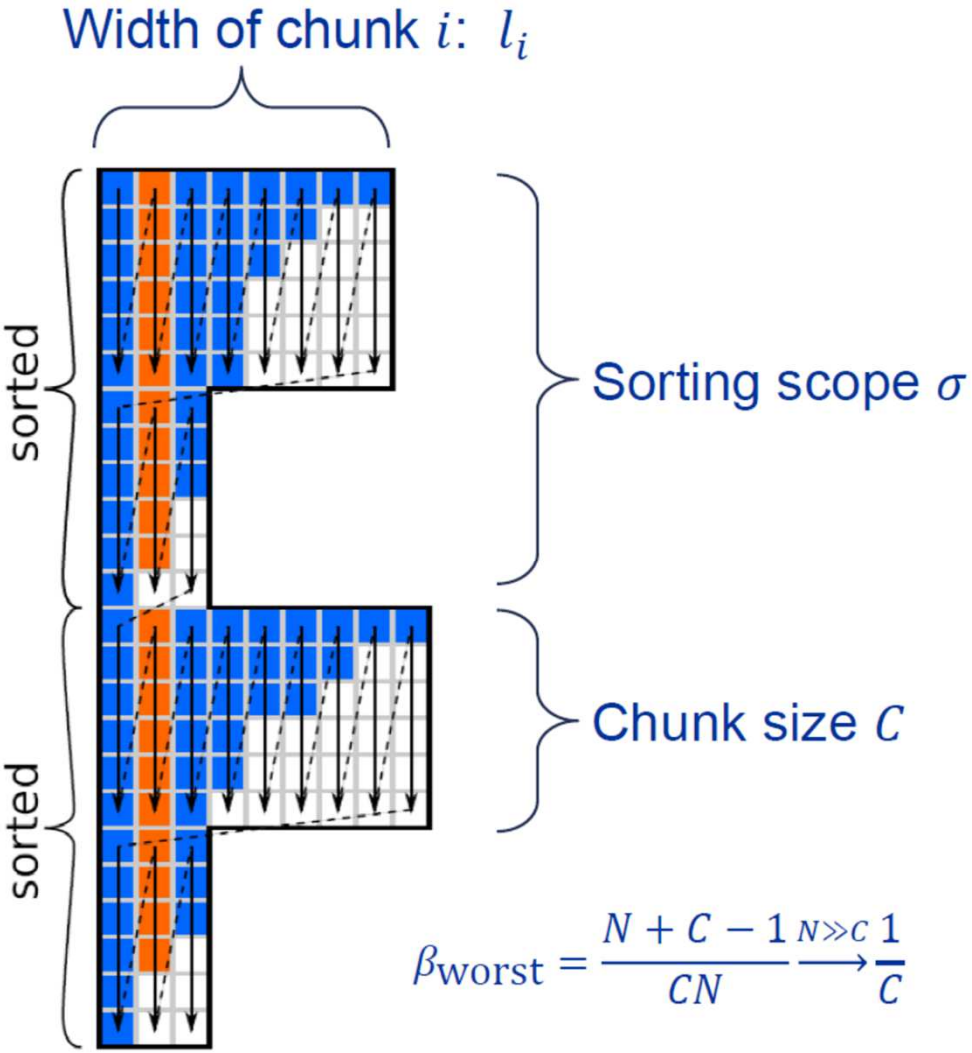


Constructing SELL-C-σ

1. Pick chunk size C (guided by SIMD/T widths)
2. Pick sorting scope σ
3. Sort rows by length within each sorting scope
4. Pad chunks with zeros to make them rectangular
5. Store matrix data in “chunk column major order”

“Chunk occupancy”: fraction of “useful” matrix entries

$$\beta = \frac{N_{nz}}{\sum_{i=0}^{N_c} C \cdot l_i}$$

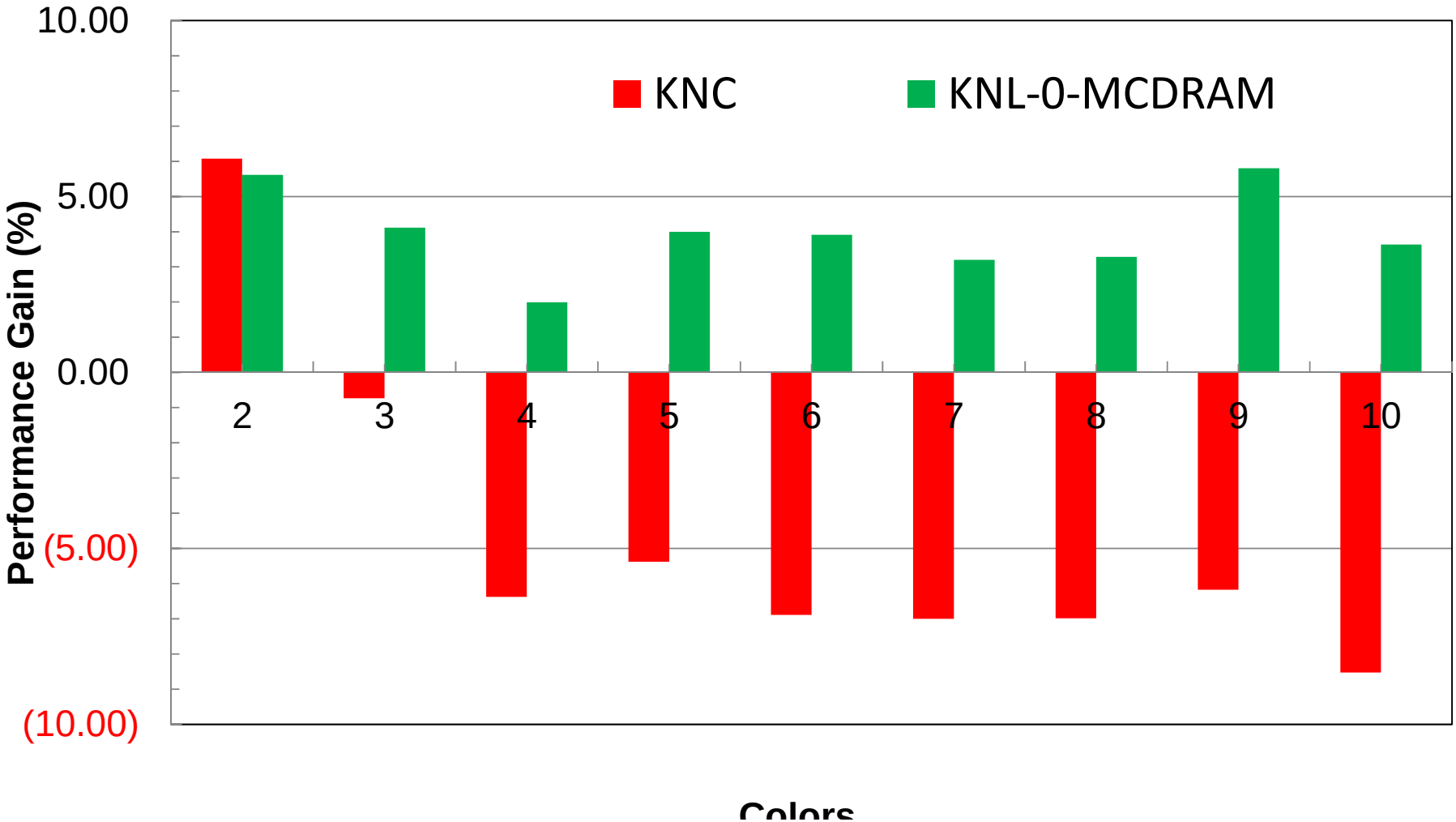


[Kreutzer, Hager, Wellein 2014]

Effects of SELL-C-σ on KNL-0-MC/D

Improvement over Original Best ELL Cases (2-10 colors)

SELL-C-σ is rather slower on KNC



Further Reduction of OpenMP Overhead on KNL

[Hoshino et al. 2017]

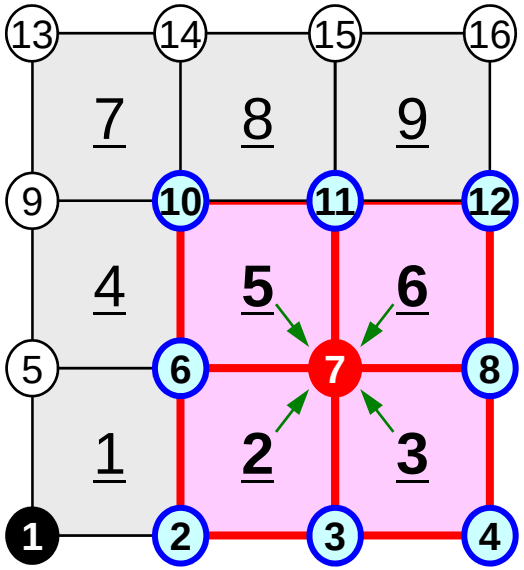
星野他
HPC-158
2017

- NO Wait for SpMV
- Call $OMP PARALLEL only once at Each Iteration
- Remove $OMP DO
- Remove REDUCTION: Dot products are calculated at each thread
- Performance Improvement on SELL-8-1
  - KNL: 16.7 %, KNC: 25.0 %
- Finally, KNL-MCDRAM is 3.6x faster than KNL-DDR4
  - Room for further performance engineering
  - 1% of peak (e.g. HPCG (27-pt stencil): 1.5% of peak)
  - Throughput: 70+%

- **Poisson 3D ICCG for FVM (7-pt)**
- **GeoFEM/Cube Matrix Assembly for FEM**
- **GeoFEM/Heat CG with SELL-C-σ**
- **Dynamic Loop Scheduling of Parallel FEM [ICPP 2017]**
-

Matrix Assembly in FEM

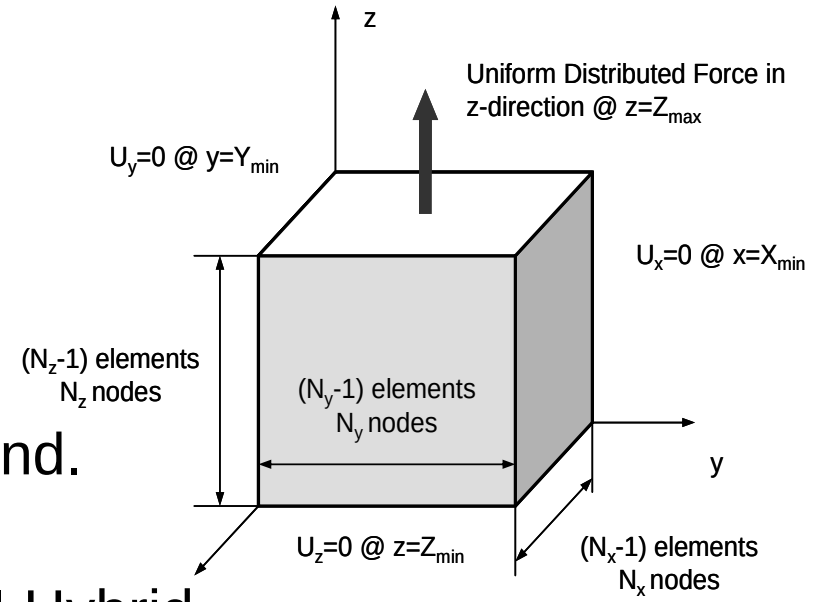
- Integration on each Element: Element Matrix (Dense)
- Accumulation of Elem. Mat: Global Matrix (Sparse)
- Integration/Matrix: Element
- Matrix Computation: Node
  - Components of each Node: Accumulation by Neighboring Elements.
  - Data dependency may happen
  - Coloring needed: Multicoloring
- Hardware
  - Intel Xeon Phi (KNC/KNL)
  - NVIDIA Tesla K40/P100
    - Atomic OP's supported by H/W



[KN, Naruse et al. HPC-152, 2015]

GeoFEM/Cube

- Parallel FEM Code (& Benchmarks)
- 3D-Static-Elastic-Linear (Solid Mechanics)
- Performance of Parallel Preconditioned Iterative Solvers
  - 3D Tri-linear Hexa. Elements
  - SPD matrices: CG solver
  - Fortran90+MPI+OpenMP
  - Distributed Data Structure
  - Reordering for IC/SGS Precond.
    - MC, RCM, CM-RCM
  - MPI, OpenMP, OpenMP/MPI Hybrid
    - only OpenMP case



- Focusing on Matrix Assembling in this Study: 128^3 cubes

Type-A

BLKSIZ: Elem.# in Blocks, NBLK: Block #, icel: Elem. ID

```
!$omp parallel (…)
    do color= 1, COLORtot
!$omp do
    do ip= 1, THREAD_total
      NBLK: calculated by (col_index, color, thread#)
    do ib= 1, NBLK
      do blk= 1, BLKSIZ
        icel: calculated by (col_index, ib, blk)
!$omp simd
        do ie= 1, 8; do je= 1, 8
          <② Address in Sparse Matrix>
        enddo; enddo
      enddo
      do blk= 1, BLKSIZ
        icel: calculated by (col_index, ib, blk)
        <① Jacobian & Grad. Shape Fn' s>
!$omp simd
        do ie= 1, 8; do je= 1, 8
          <③ Element Matrices >
        enddo; enddo
      enddo
      do blk= 1, BLKSIZ
        icel: calculated by (col_index, ib, blk)
!$omp simd
        do ie= 1, 8; do je= 1, 8
          <④ Global Matrices>
        enddo; enddo
      enddo
    enddo
    enddo
    enddo
!$omp end parallel
```

Code Name	KNC	KNL-1	K40	P100
Architecture	Intel Xeon Phi 5110P (Knights Corner)	Intel Xeon Phi 7250 (Knights Landing)	NVIDIA Tesla K40	NVIDIA Tesla P100 (Pascal)
Frequency (GHz)	1.053	1.40	0.745	1.328
Core # (Max Thread #)	60 (240)	68 (272)	2,800	1,792
Peak Performance (GFLOPS)	1,010.9	3,046.4	1,430	4,759
Memory (GB)	8	MCDRAM: 16 DDR4: 96	12	16
Memory Bandwidth(GB/sec., Stream Triad)	159	MCDRAM: 490 DDR4: 84.5	218	530
		Real OFP		

Results (Fastest Case) sec. (GFLOPS)

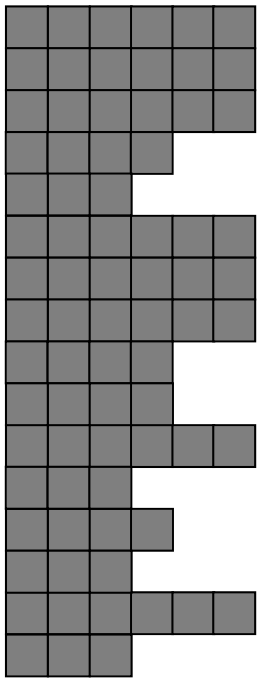
	Coloring + OpenMP/OpenACC	Atomic+ OpenACC	Atomic + CUDA
KNC	no SIMD: 1.339 (48.5) SIMD: 0.690 (94.0)	0.576 (112.7) No Coloring	
KNL	no SIMD: 0.379 (171.1) SIMD: 0.184 (352.5)	0.185 (351.5) No coloring	
K40	0.675 (96.2)	0.507 (128.1)	0.362 (179.4)
P100	0.210 (308.9)	0.190 (341.4)	0.154 (412.2)

Based on〔KN, Naruse et al. HPC-152, 2015〕
Computation on KNL by Dr. Horikoshi (Intel)

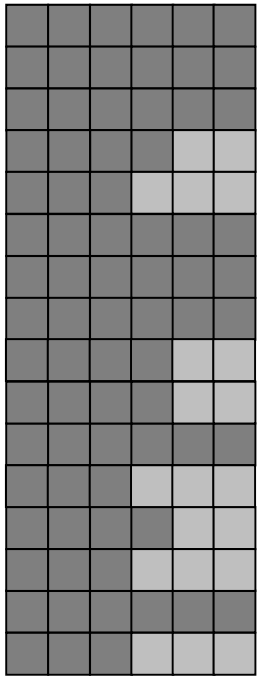
- **Poisson 3D** **ICCG for FVM (7-pt)**
- **GeoFEM/Cube** **Matrix Assembly for FEM**
- **GeoFEM/Heat** **CG with SELL-C-σ**
- **Dynamic Loop Scheduling of Parallel FEM [ICPP 2017]**

Storing Formats for Sparse Matrices

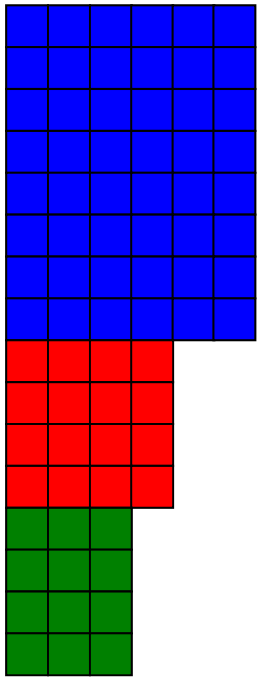
- ELL/Sliced ELL: Excellent Performance by Prefetching
- SELL-C-σ: Good for Vector/SIMD (by FAU/Erlangen)
  - This work is the first example where SELL-C-σ is applied to IC/ILU preconditioning.



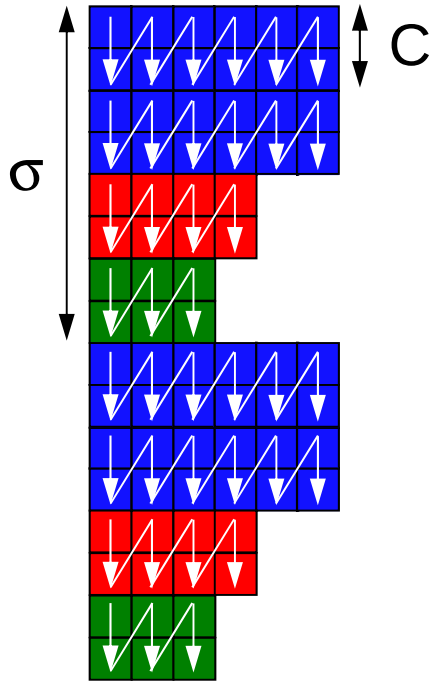
CRS



ELL



Sliced ELL



SELL-C-σ
(SELL-2-8)

SELL-C-σ for SpMV in GeoFEM/Heat

27-pt. Stencil，HPCGと似ている（1.5%@8,192nodes）

- KNC (240 threads)
  - CRS 1.699 sec
  - SELL-128-1 1.063 sec
    - 20GFLOPS
- KNL-1 (192 threads)
  - CRS 0.421 sec
  - SELL-128-1 0.384 sec
    - 55GFLOPS
      - read 350/400 GB/sec.
    - miniFEM (SNL): 42GF
- Generally, difference between CRS & ELL is smaller on KNL (BDW)
  - Out-of-Order

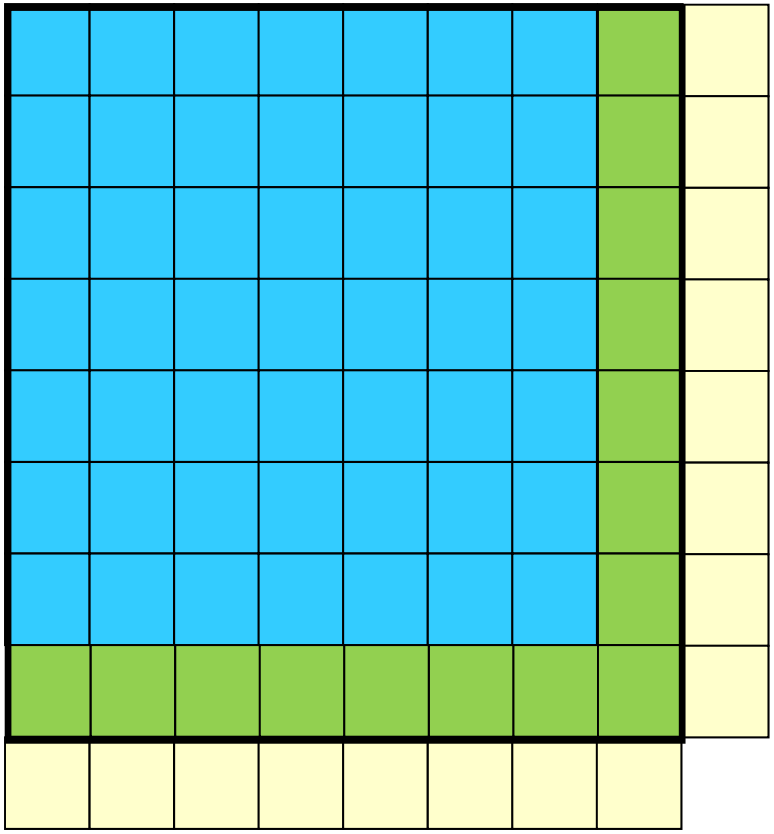
```
!$omp parallel do private (j,k)
      do j= 1, N
        WW(j,Q)= D(j)*WW(j,P)
      enddo

!$omp parallel do private (j,k)
      do j= 1, N
      do k= INL(j-1)+1, INL(j)
        WW(j,Q)= WW(j,Q) + AL(k)*WW(IAL(k),P)
      enddo
      enddo
```

```
!$omp parallel do
      do i= 1, N/2
        WW(2*(i-1)+1,Q)= D(2*(i-1)+1) * WW(2*(i-1)+1,P)
        WW(2*(i-1)+2,Q)= D(2*(i-1)+2) * WW(2*(i-1)+2,P)
      enddo

!$omp parallel do
      do i= 1, Ndiv
      do j= 1, NCOL
!$omp simd
      do k= 1, 128
        WW(128*(i-1)+k,Q)= WW(128*(i-1)+k,Q) +
     &                     AL(sell_CS(i-1)+(j-1)*128+k)          &
     &                 * WW(IAL(sell_CS(i-1)+(j-1)*128+k),P)
      enddo
      enddo
      enddo
```

〔中島他　HPC-151，2015〕に基づく

- **Poisson 3D ICCG for FVM (7-pt)**
- **GeoFEM/Cube Matrix Assembly for FEM**
- **GeoFEM/Heat CG with SELL-C-σ**
- **Dynamic Loop Scheduling of Parallel FEM [ICPP 2017]**

Comm.-Comp. Overlapping (CC-Overlapping): ***Static***



Pure Internal Meshes

External (HALO) Meshes

Internal Meshes on Boundary's (Boundary Meshes)

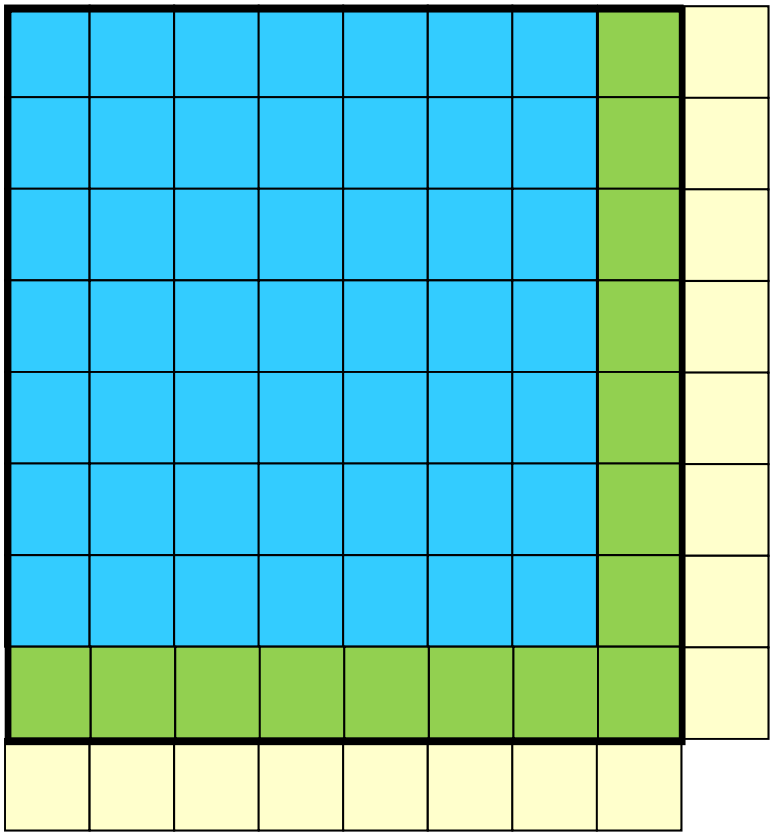
```
call MPI_Isend
call MPI_Irecv
```

```
do i= 1, Ninn
  (calculations)
enddo
call MPI_Waitall
```

```
do i= Ninn+1, Nall
  (calculationas)
enddo
```

Good for Stencil
Not so Effective SpMV

Comm.-Comp. Overlapping + Dynamic Loop Scheduling: *Dynamic*



Pure Internal Meshes

External (HALO) Meshes

Internal Meshes on Boundary's (Boundary Meshes)

```
call MPI_Isend
call MPI_Irecv
call MPI_Waitall
```

Master

```
do i= 1, Ninn
  (calculations)
enddo
```

Dynamic

```
do i= Ninn+1, Nall
  (calculationas)
enddo
```

Static

Dynamic Loop Scheduling

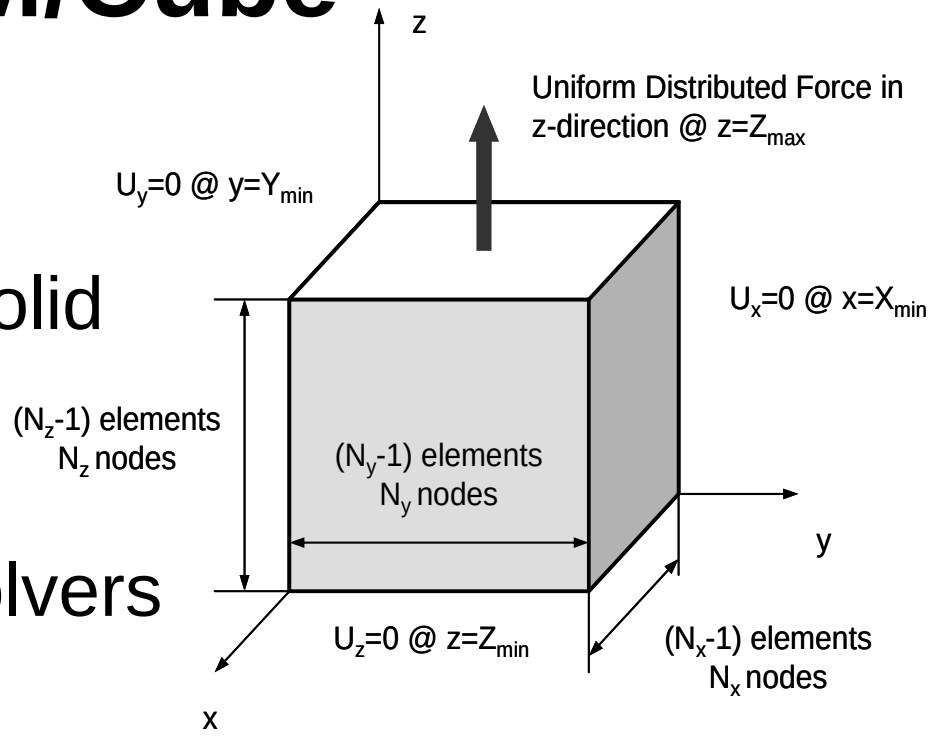
- "dynamic"
- "!$omp master～!$omp end master"

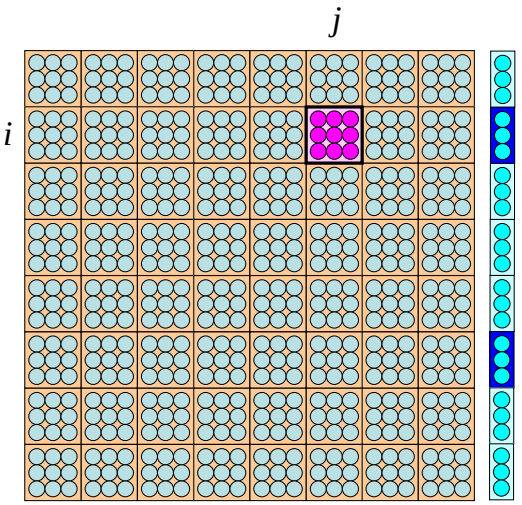
```
!$omp parallel private (neib,j,k,i,X1,X2,X3,WVAL1,WVAL2,WVAL3)
!$omp&         private (istart,inum,ii,ierr)

!$omp master              Communication is done by the master thread (#0)
!C
!C- Send & Recv.
(…)
      call MPI_WAITALL (2*NEIBPETOT, req1, sta1, ierr)
!$omp end master

!C                            The master thread can join computing of internal
!C-- Pure Internal Nodes      nodes after the completion of communication

!$omp do schedule (dynamic,200)    Chunk Size= 200
      do j= 1, Ninn
        (…)
      enddo
!C
!C-- Boundary Nodes           Computing for boundary nodes are by all threads

!$omp do                      default: !$omp do schedule (static)
      do j= Ninn+1, N
        (…)
      enddo

!$omp end parallel
```

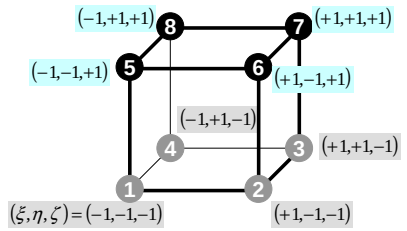
Ina, T., Asahi, Y., Idomura, Y., Development of optimization of stencil calculation on Tera-flops many-core architecture, IPSJ SIG Technical Reports 2015-HPC-152-10, 2015 (in Japanese)

GeoFEM/Cube

- Parallel FEM Code (& Benchmarks)
- 3D-Static-Elastic-Linear (Solid Mechanics)
- Performance of Parallel Preconditioned Iterative Solvers
  - 3D Tri-linear Hexahedral Elements
  - Block Diagonal LU + CG
  - Fortran90+MPI+OpenMP
  - Distributed Data Structure
  - MPI, OpenMP, OpenMP/MPI Hybrid
  - Block CRS Format







$$\begin{bmatrix} a_{i_e j_e 11} & a_{i_e j_e 12} & a_{i_e j_e 13} \\ a_{i_e j_e 21} & a_{i_e j_e 22} & a_{i_e j_e 23} \\ a_{i_e j_e 31} & a_{i_e j_e 32} & a_{i_e j_e 33} \end{bmatrix} (i_e, j_e = 1 \ldots 8)$$

Code Name	KNL	BDW	FX10
Architecture	Intel Xeon Phi 7250 (Knights Landing)	Intel Xeon E5-2695 v4 (Broadwell-EP)	SPARC IX fx
Frequency (GHz)	1.40	2.10	1.848
Core # (Max Thread #)	68 (272)	18 (18)	16 (16)
Peak Performance (GFLOPS)	3,046.4	604.8	236.5
Memory (GB)	MCDRAM: 16 DDR4: 96	128	32
Memory Bandwidth(GB/sec., Stream Triad)	MCDRAM: 490 DDR4: 80.1	65.5	64.7
Out-of-Order	Y	Y	N
System	Oakforest-PACS	Reedbush-U	Oakleaf-FX

Code Name	KNL	BDW	FX10
Architecture	Intel Xeon Phi 7250 (Knights Landing)	Intel Xeon E5-2695 v4 (Broadwell-EP)	SPARC IX fx
Frequency (GHz)	1.40	2.10	1.848
Core # (Max Thread #)	68 (272)	18 (18)	16 (16)
Peak Performance (GFLOPS)/core	**44.8**	**33.6**	**14.8**
Memory Bandwidth(GB/sec., Stream Triad)/core	**MCDRAM: 7.21 DDR4: 1.24**	**3.64**	**4.04**
Out-of-Order	Y	Y	N
Network	**Omni-Path Architecture**	**Mellanox EDR Infiniband**	**Tofu 6D Torus**

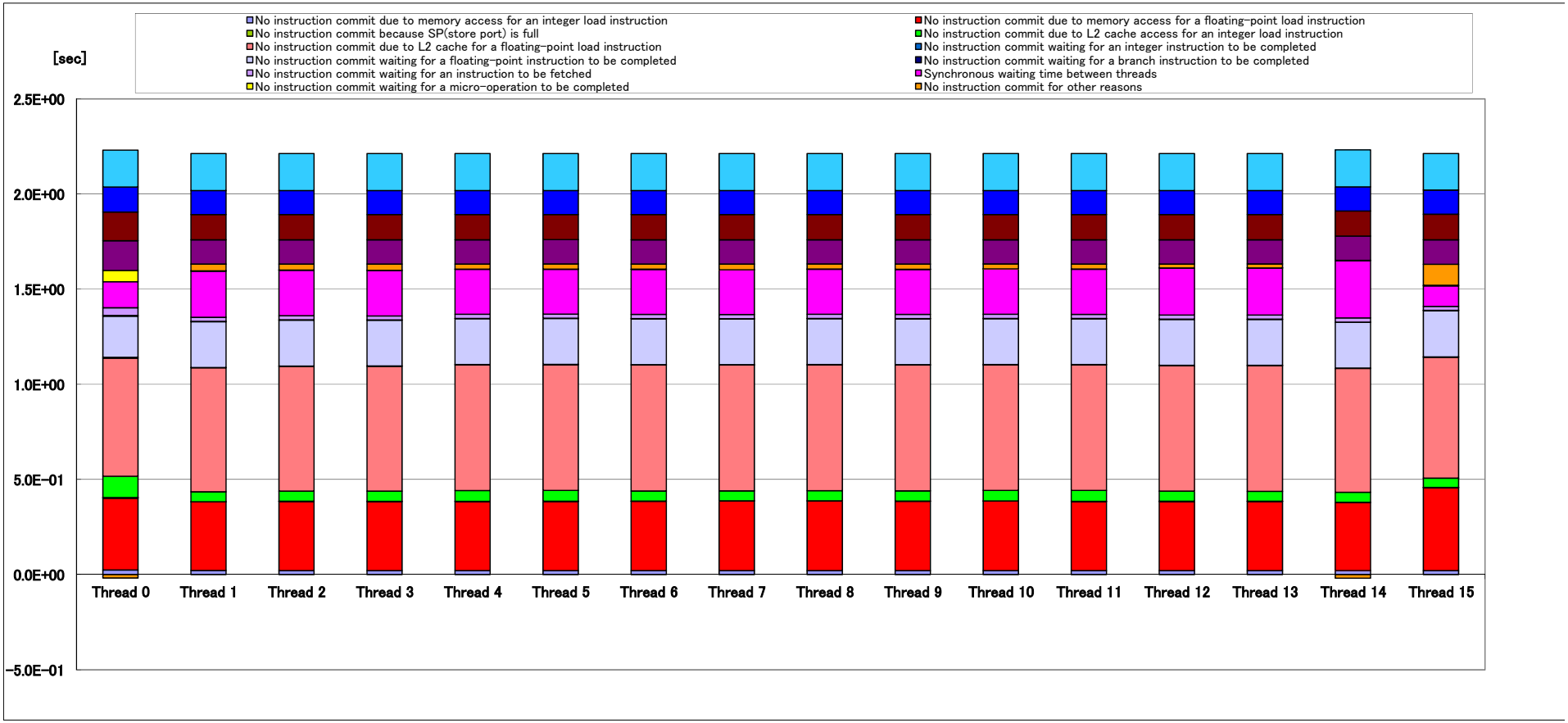
FX10: 240 nodes, 368,640,000 DOF

HB 16x1, Performance Analysis by Fujitsu's Profiler (single node)

	Original	Static	Dynamic: Chunk Size=100	Dynamic: Chunk Size=500
GFLOPS/node	12.59	13.33	14.47	13.82
Memory Throughput (GB/sec)	61.61	64.86	69.44	68.07
L2 Throughput (GB/sec)	71.13	75.03	84.15	79.03
sec./(100 iterations)	2.21	2.10	1.93	2.00
Synchronous waiting time between threads (sec) Averaged	.229	.122	.073	.061
L2 waiting for FP Load (sec) Averaged	.655	.657	.540	.614

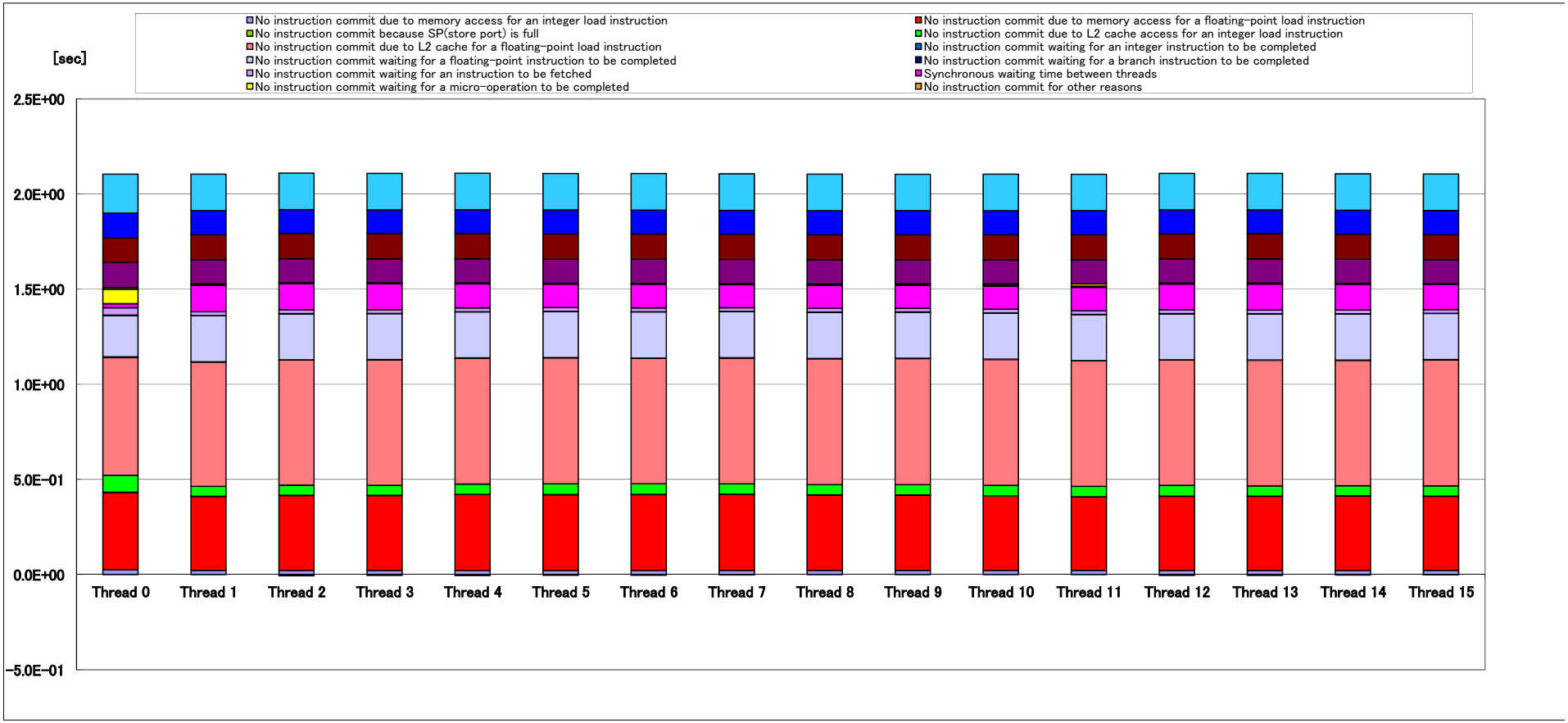
3,840 cores, PA Profiler
FX10: 240 nodes, 368,640,000 DOF
“Original”: 2.21 sec.

■ Synchronization Waiting, ■ L2 Load



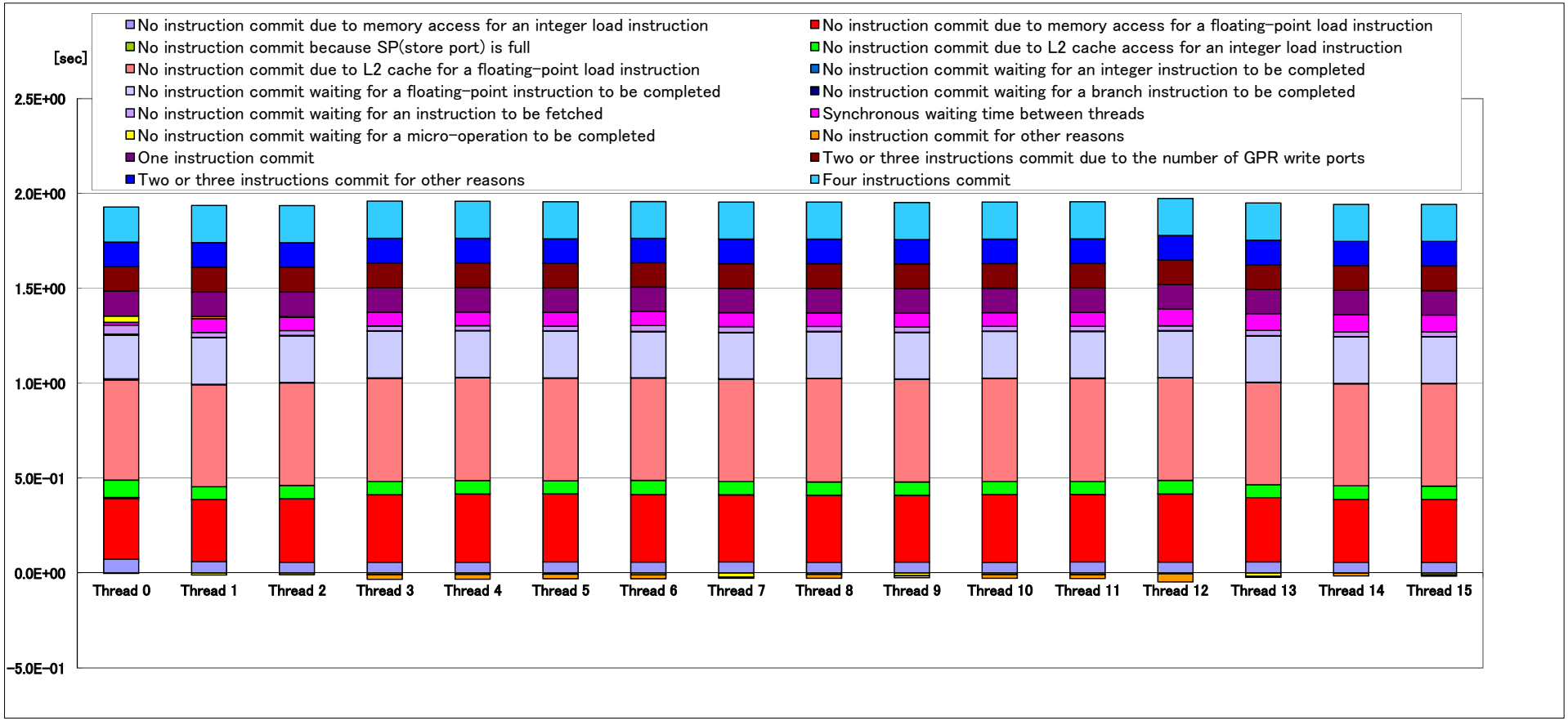
3,840 cores, PA Profiler
FX10: 240 nodes, 368,640,000 DOF
“Static”: 2.10 sec.

Synchronization Waiting, L2 Load



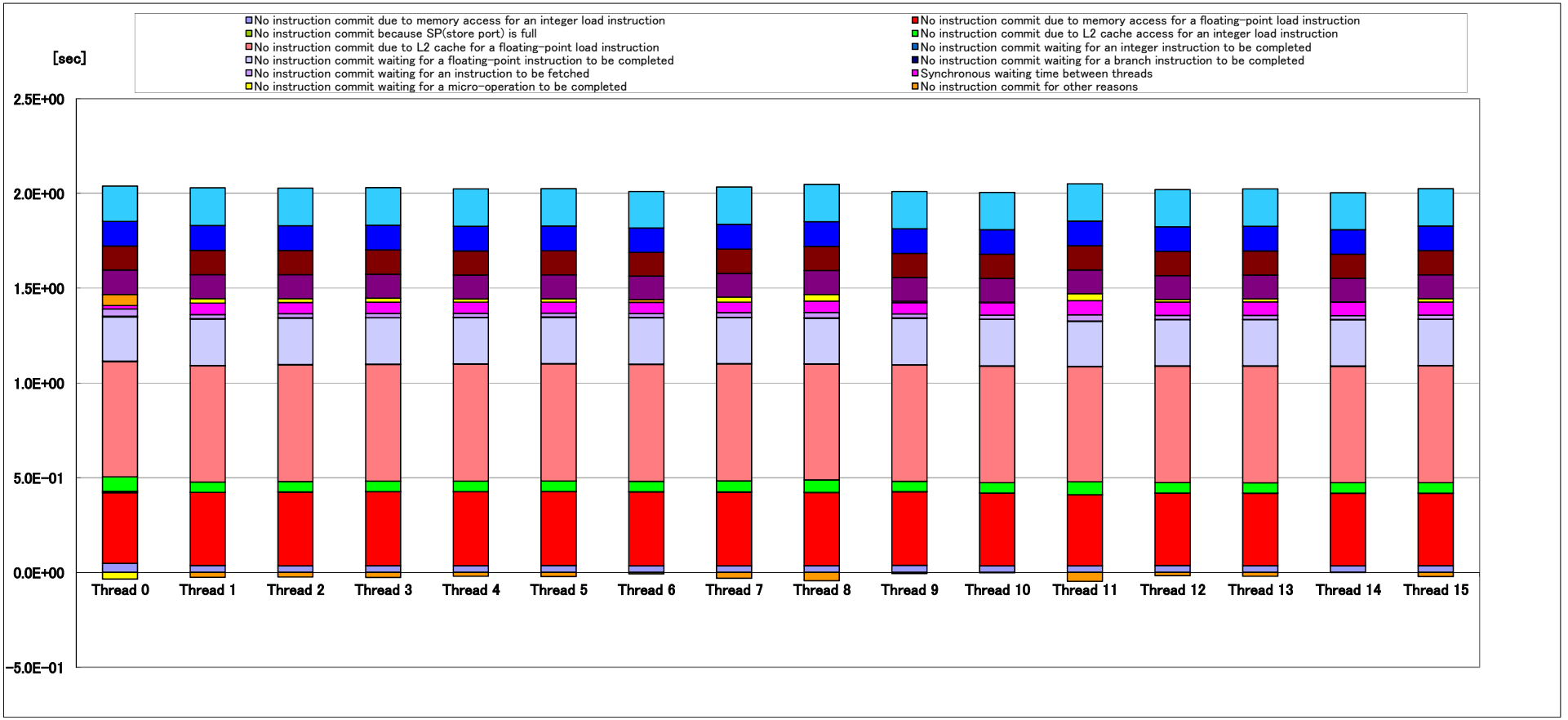
3,840 cores, PA Profiler
FX10: 240 nodes, 368,640,000 DOF
“Dynamic, Csz=100”: 1.93 sec.

■ Synchronization Waiting, ■ L2 Load



3,840 cores, PA Profiler
FX10: 240 nodes, 368,640,000 DOF
“Dynamic, Csz=500”: 2.00 sec.

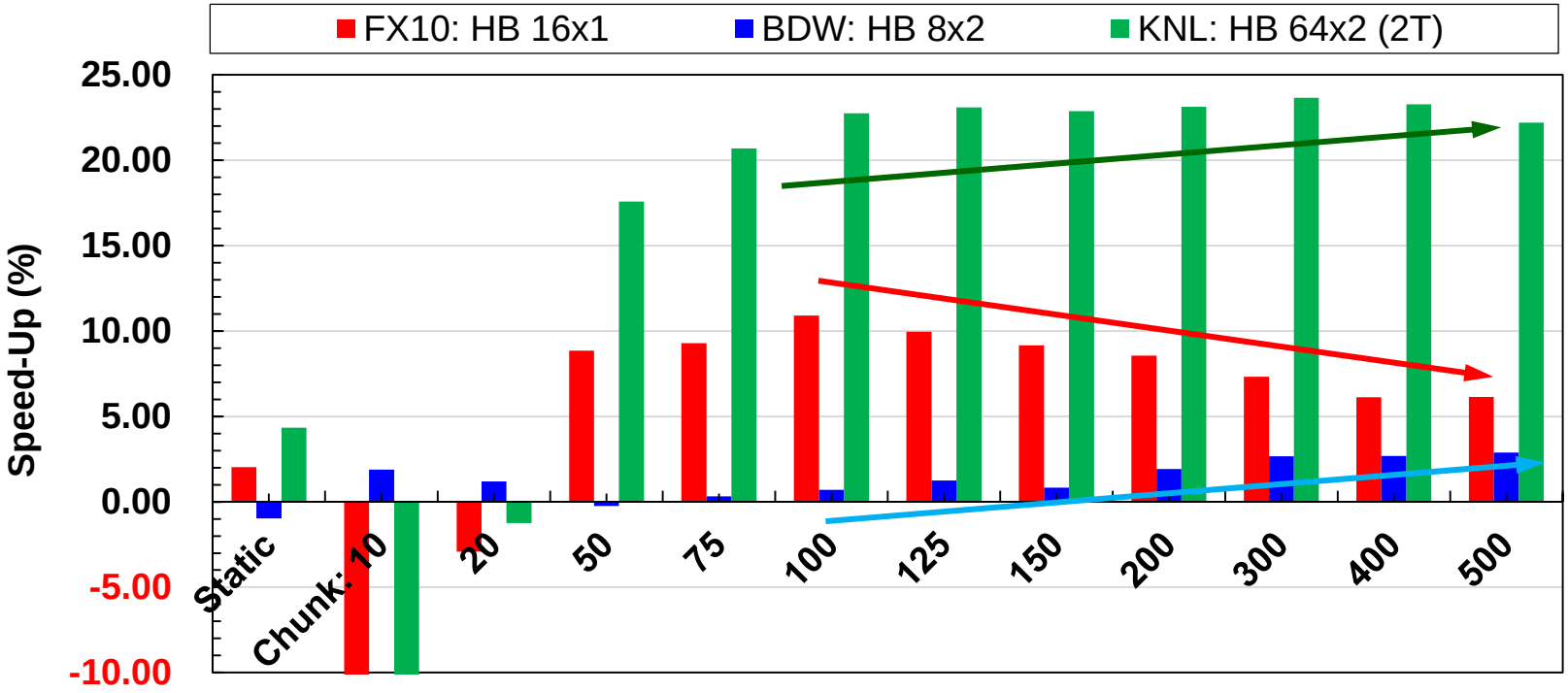
■ Synchronization Waiting, ■ L2 Load



Preliminary Results: Best Cases

3,840 cores, 368,640,000 DOF

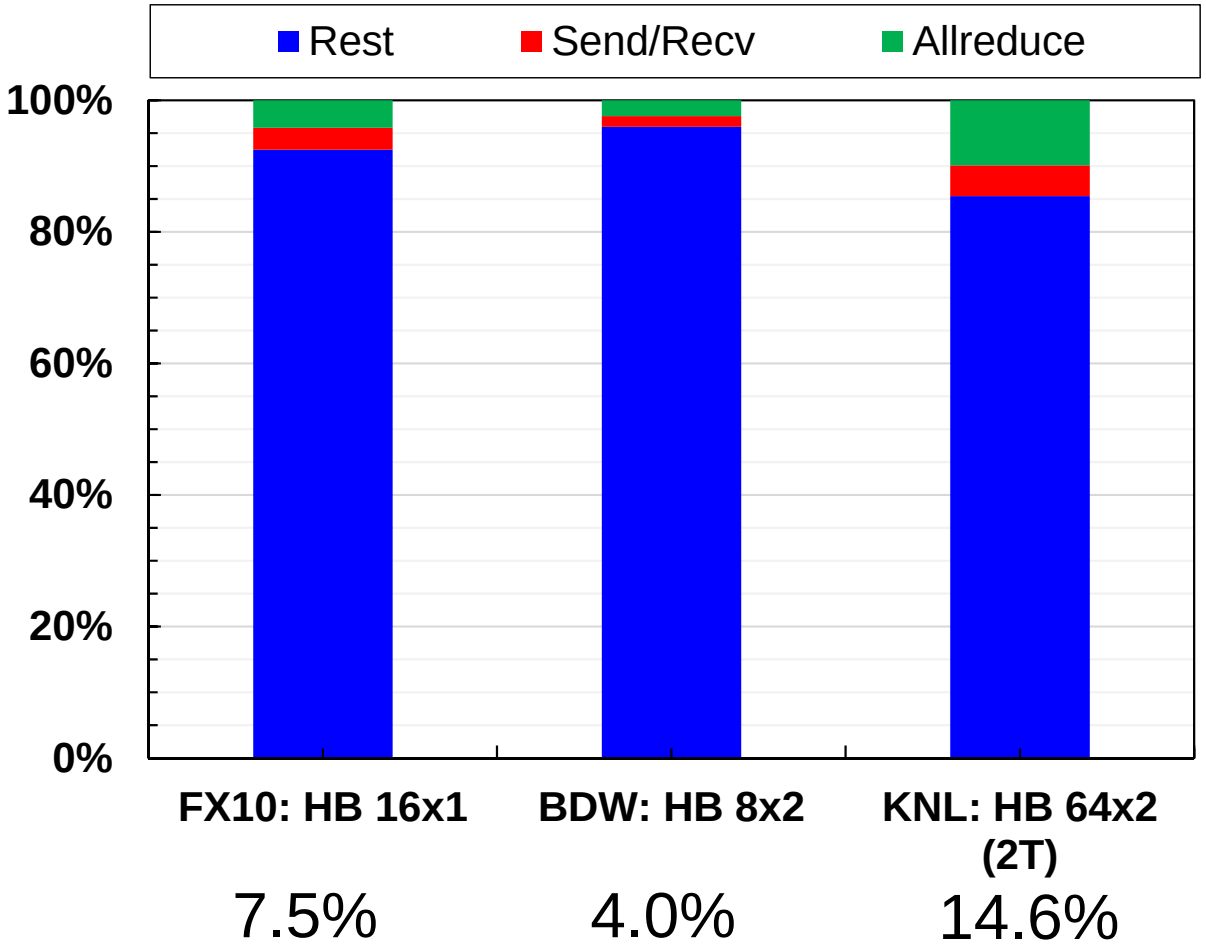
Improvement of CG from Original Cases



Preliminary Results: Original Cases

3,840 cores, 368,640,000 DOF

Communication Overhead by Collective/Point-to-Point Communications



Features

	Effect of Dynamic Scheduling	Optimum Chunk Size	Notes
FX10	Medium	100	Memory Throughput
BDW	Small	500+	Low Comm. Overhead Small number of threads.
KNL	Large	300-500	Effects are significant for HB 64x2, 128x1, where loss of performance by communications on master thread is rather smaller.

Strong Scaling

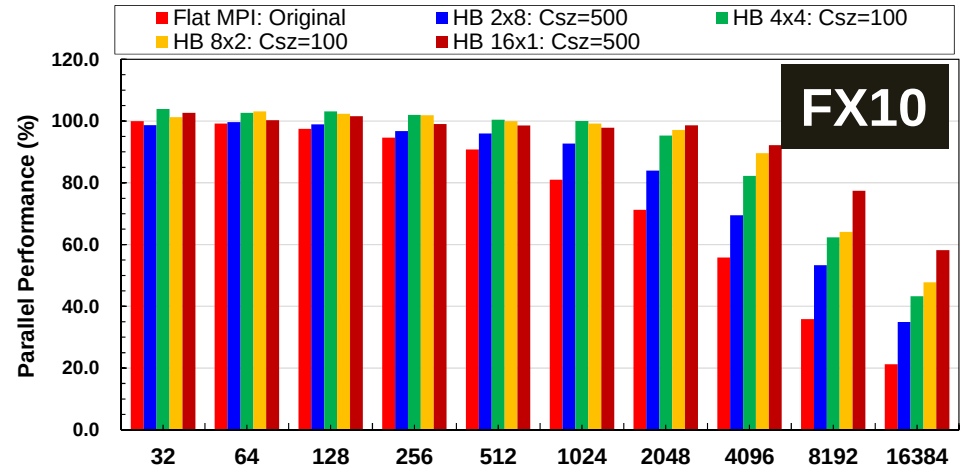
Parallel Performance (%)

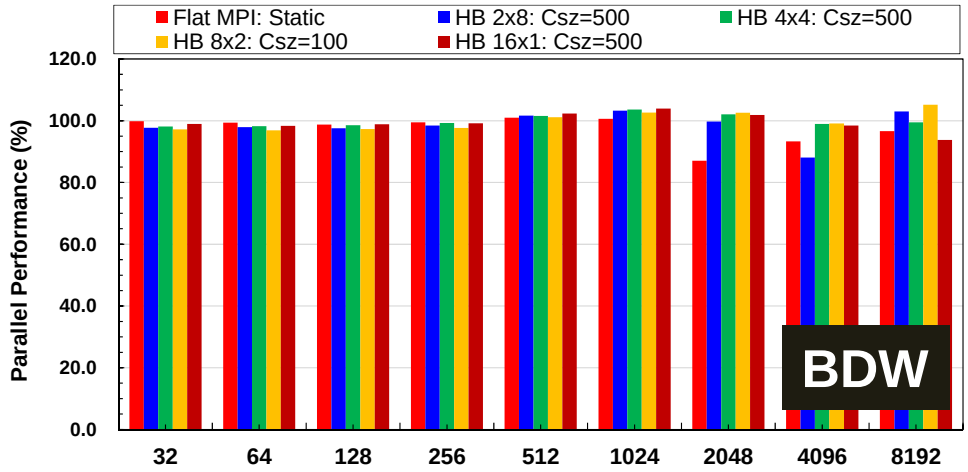
BEST case for each HB MxN

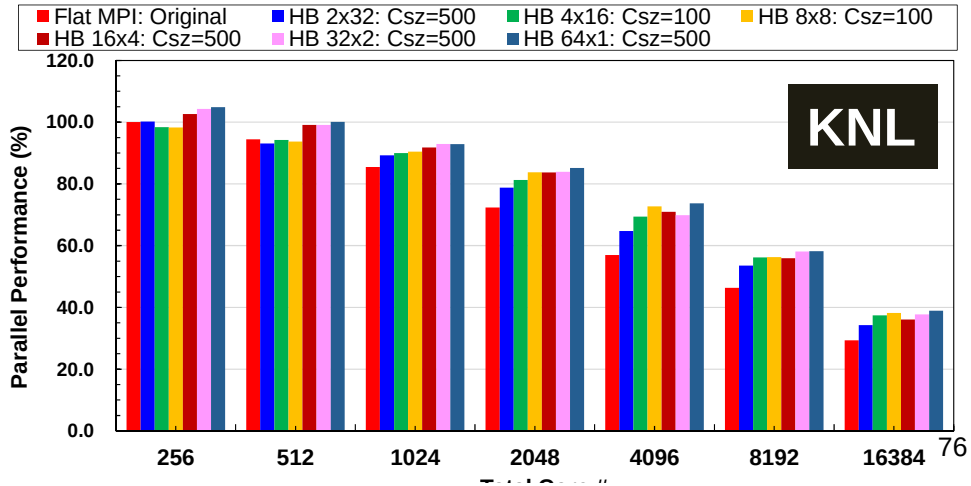
50,331,648 DOF

Computation Time of Flat MPI at Min.# cores: 100%

This difference between BDW and KNL might be because difference of performance between Infiniband EDR and Omni-Path Architecture.







Strong Scaling

Parallel Performance (%)

Effect of Dynamic Loop Scheduling

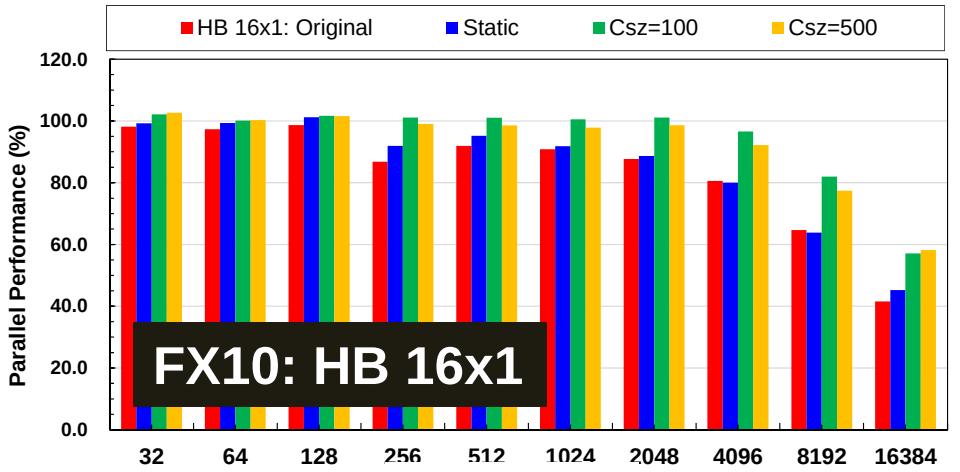
50,331,648 DOF

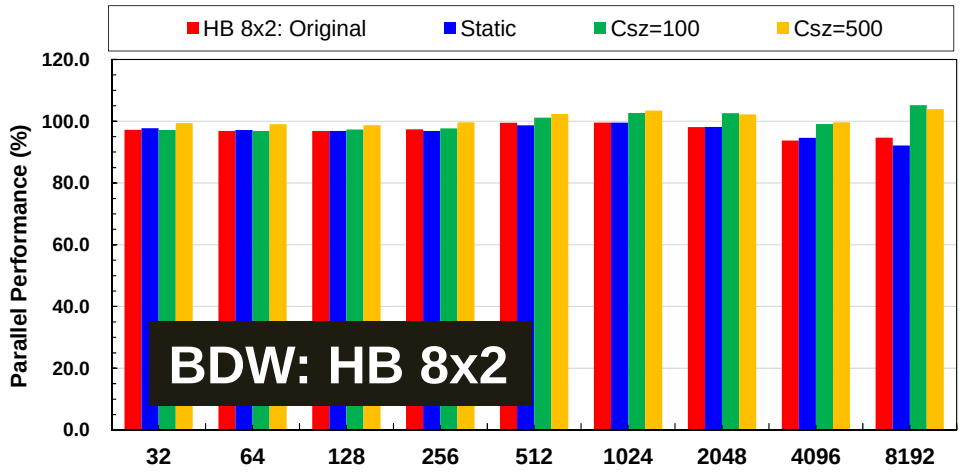
Computation Time of Flat MPI at Min.# cores: 100%

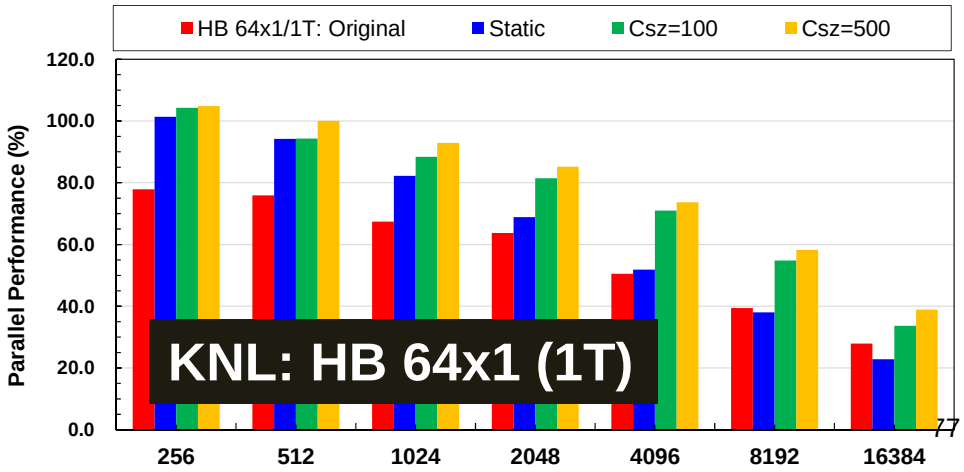
Effect of Dynamic Loop Scheduling with more than 8,192 cores

- FX10: 20%-40%
- BDW: 6%-10%
- KNL with HB 8×8 (1T): 20%-30%
- KNL with HB 64×1 (1T): 40%-50%

Idomura







まとめ

- Oakforest-PACSの概要
- Oakforest-PACSを使用したアプリケーション事例・最適化事例
- 今後
  - 更なる最適化
  - SC18 Gordon Bellへのチャレンジ
  - Deep Learning等への適用
    - SC17　Technical Paper, Cori/LBNLを使用した事例, IntelCaffe
    - Deep Learning at 15PF: Supervised and Semi-Supervised Classification for Scientific Data

おまけ

SIAM Conference on Parallel Processing for Scientific Computing (PP18)

Thanks for Your Contributions !!

March 7-10, 2018
Waseda University, Tokyo, Japan

Record Breaking 126 MS Proposals, 100+ CP

700+ Participants ? Please book your Hotel ASAP !

- Organizing Committee Co-Chairs's
  - Satoshi Matsuoka (Tokyo Institute of Technology, Japan)
  - Kengo Nakajima (The University of Tokyo, Japan)
  - Olaf Schenk (Universita della Svizzera italiana, Switzerland)
- Contact
  - Kengo Nakajima, nakajima(at)cc.u-tokyo.ac.jp
  - http://www.siam.org/meetings/pp18/
  - http://siampp18.jsiam.org/

siam