JCAHPCにおける
国内最大PCクラスタの導入と運用に向けて

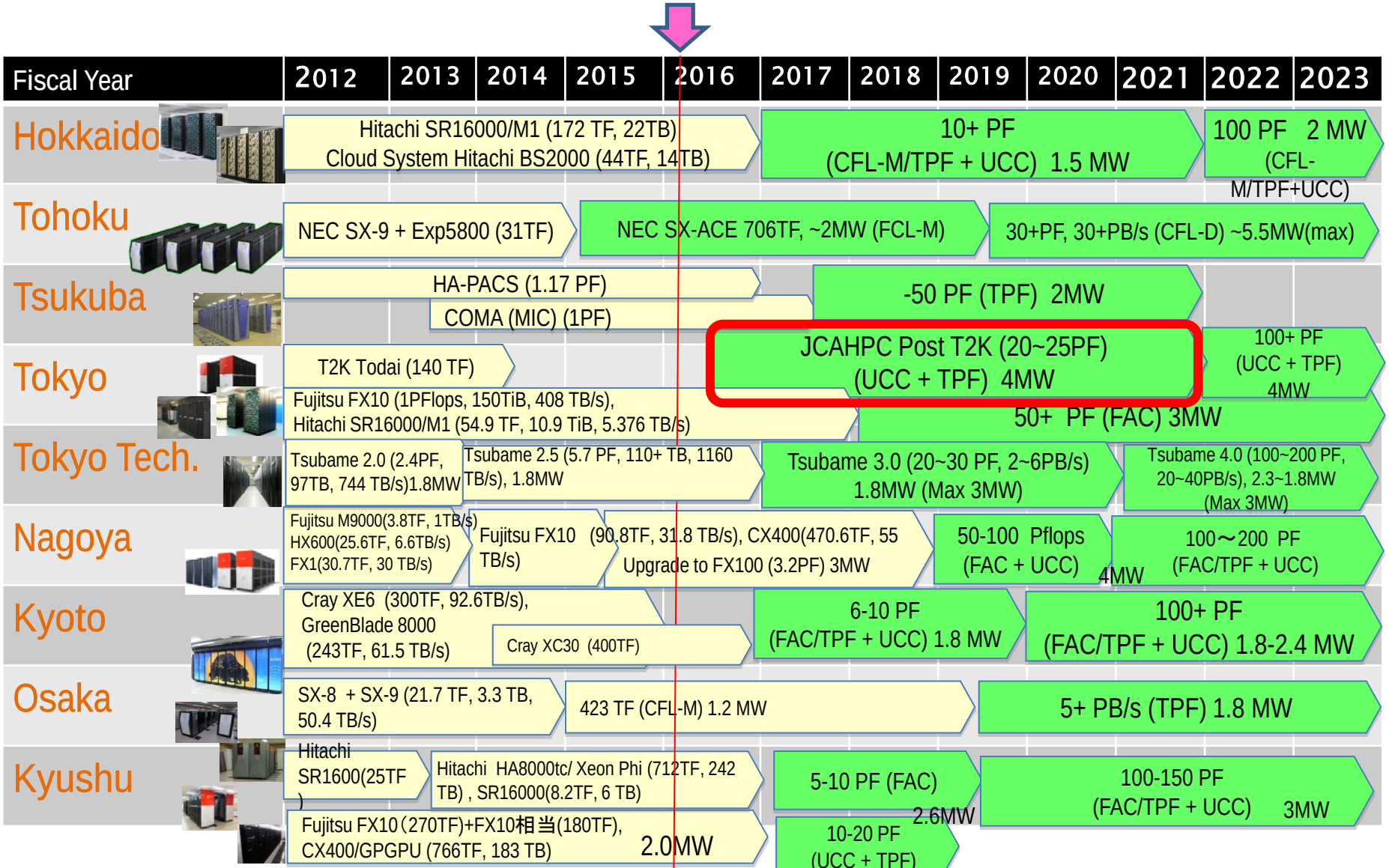
朴　泰祐
筑波大学計算科学研究センター
http://www.hpcs.cs.tsukuba.ac.jp/~taisuke/

アウトライン

- 国立大学におけるスパコン設置状況・計画
- JCAHPCの発足経緯と現状
- JCAHPCで導入されるスパコンの概要
- メニーコア向けチューニング例（based on KNC）
- まとめ

（お断り：本資料における導入システムの仕様は現時点の調達状況に基づくものです。実導入システムでは変更があり得ます。）

国立大学スパコンセンターのシステム設置状況と導入計画



PCCWorkshop2016@仙台

※一部、最新情報でないものがあります

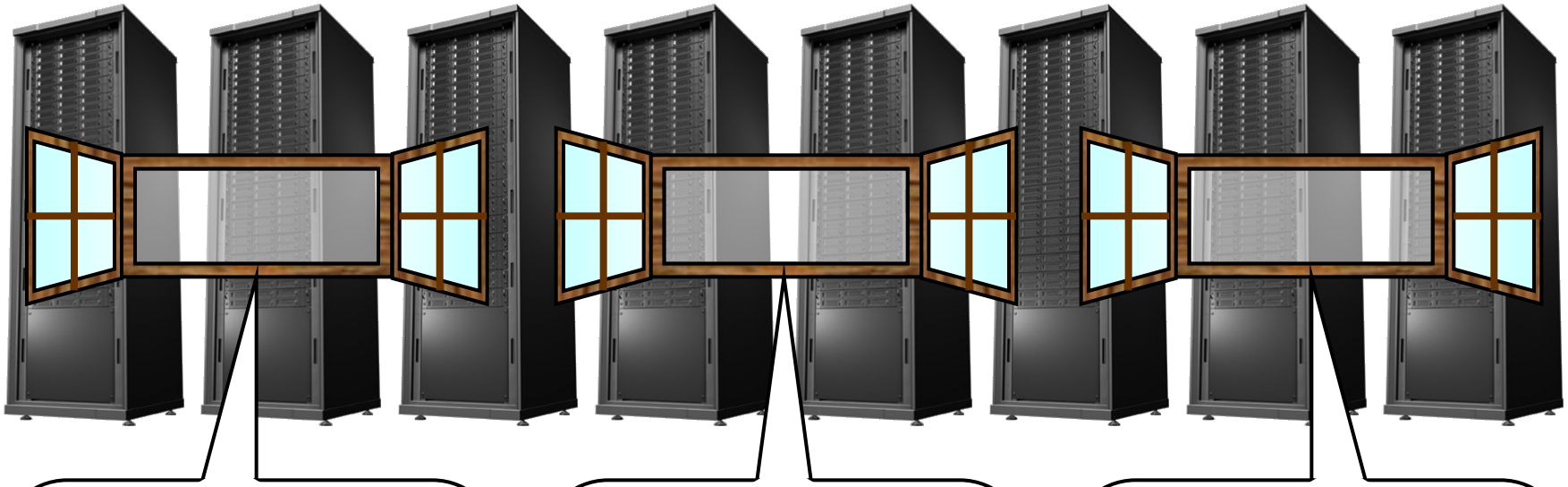
Positioning of infrastructures in Japan (HPCI)

- National Flagship Leading Machine (NFL)
  - K, post-K
- Flagship-Aligned Commercial Machine (FAC)
  - small scaled machine of NFL (or similar system) -> FX10, FX100
- Complimentary Function Leading Machine (CFL-M, CFL-D)
  - special architecture or featured machine for the field not covered by NFL -> Vector
- Upscale Commodity Cluster Machine (UCC)
  - commodity cluster based on conventional technology and commodity market -> Clusters
- Technology Path-Forward Machine (TPF)
  - experimental system toward future technology and next generation HPC system -> original technology

（JCAHPCの前に）T2K Alliance

- T2K Open Supercomputer Alliance
  - 筑波大学
  - 東京大学
  - 京都大学
- 最先端コモディティテクノロジによる大学主導の仕様によるオープンクラスタシステムの導入
  - 3大学で基本仕様を共通化
  - アプリケーション、システムソフトウェアの共有により、システム間の性能可搬性、プラットフォーム共有を実現
  - ベンダー主導のクローズドなシステムからオープンなシステムへ

何がオープンなのか？



Open Hardware Arch.

- **コモディティテクノロジ**
  e.g. x86, IB/Myri-10G
- → **現在のITマーケットで最もcost/performanceの良いもの**
- HPC**向けの特殊ハードは対象としない**

Open Software Stack

- **オープンソース＆標準システムソフトウェア**
  e.g. Linux, MPI, Globus
- → **オープンソースなHPC向けミドルウェア＆ライブラリ**

Open to User's Needs

- Floating Point**ユーザだけでなく，**
- Integer**ユーザ（大規模データ処理等）を含めた幅広いユーザを対象に**

何が共通仕様か？

- 共通する仕様
  - Hardware
    - shared memory node of 16+ x86 cores and 32+GB ECC memory with 40+GB/sec (aggr.)
      Fat Node Architecture for wide variety of applications
    - bundle (even #) of inter-node links of 5+GB/sec (aggr.)
      High bandwidth to support very high performance of computation node
    - on-node 250+GB RAID-1 disk (optional) and IPMI2.0
  - Software
    - Red Hat or SuSE Linux
    - Fortran, C and C++ with OpenMP and auto-parallelizer
    - Java with JIT compiler
    - MPI of 4+GB/sec and 8.5-μsec RT latency
    - BLAS, LAPACK and ScaLAPACK
  - Benchmarks (性能数値自体は各大学により異なる)
    - SPEC CPU2006, SPEC OMP2001, HPC Challenge (part)
    - our own for memory, MPI and storage performance

T2K Open Supercomputer Alliance

- 元々は次期スパコン調達における共通仕様策定と運用連携が目的
- 学際的な計算機科学・計算科学の展開を目指し、研究・教育・グリッド運用等における連携活動へ

- *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

T2K時代の日本のTOP-4スパコン

TOP500 2008/06→2008/11

Machine	Site	Vendor	Rpeak (GF)	Rmax (GF)	#rank
T2K-Todai	Univ. Tokyo	Hitachi	113050	82984	16→27
T2K-Tsukuba	Univ. Tsukuba	Appro	92000	76460	20→32
TSUBAME	Tokyo Inst. Tech.	Sun	109728 161816	67700 77480	24→29
T2K-Kyodai	Kyoto Univ.	Fujitsu	61235	50510	34→51

- 現在、大学の計算センターのマシンがTOP-4を占めている
- 4台中3台が “T2K Open Supercomputer Alliance” のマシン
- T2K システムは全て quad-core Opteron (Barcelona) と quad-rail SAS (Myrinet10G or Infiniband)を利用
- 東工大TSUBAMEは dual-core Opteron ＋アクセラレータ(ClearSpeed + GT200)

T2KからPost-T2Kへ

- T2K Allianceは3大学のスパコン調達時期が一致し、研究コミュニティとしてもタイトな関係を築くことができた

- T2Kシステムの後、各大学の調達は時期が異なり、目的もそれぞれ独立化
  - 京大：4年リース周期
  - 筑波大：アクセラレータ重視等
  - 東大：FX10の導入等

- その後、筑波大・東大で再度、よりタイトな形でのスパコン連携運用の機運が生じた
  ⇒ post-T2K （ただし京大はいない）

JCAHPC

- Joint Center for Advanced High Performance Computing
- 最先端共同HPC基盤施設
  （http://jcahpc.jp）
- post-T2K Allianceとして、よりタイトな形
  - メインとなるスパコンリソースを「仕様の統一化」から「共有マシン」へ
  - 両大学のpost-T2Kスパコン予算を持ち寄り、共同調達形式で単一のシステムを導入
  - これをスムーズに運用管理するため、両大学による共同施設を仮想設置 ⇒ JCAHPC

JCAHPC沿革

- 2013年3月「最先端共同HPC基盤施設の設置及び運営に関する協定」を両大学で締結
  - 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
- 2013年4月　JCAHPC発足
  - 初代　施設長：佐藤三久（筑波大）　副施設長：石川裕（東大）
  - 現　　施設長：中村宏（東大）　　副施設長：梅村雅之（筑波大）
- 2013年7月　両大学独立に資料招請を開始
  - この時点ではまだ共同調達の形が確立していなかった
    ⇒ その後、意見招請フェーズからは共同
  - 最先端テクノロジであることに配慮し、ベンダーに十分なテスト・検討期間を与えるため資料招請期間を1年以上の長期に設置
- 複数大学の共同調達によるスパコン共同設置は国内初の試み！

JCAHPCシステムの特徴

- T2Kの精神を引き継ぎ、コモディティテクノロジによるオープンシステムの導入
  - 超並列PCクラスタ
  - 最先端のHPC向けプロセッサ
  - 使い易く効率の良い相互結合網
  - 大規模共有ファイルシステム
- 両大学による共同調達
  - 予算的に(京を除き)国内最大規模
  - システム規模も国内最大
  - 幅広いユーザ層を支援するためアクセラレータを導入しない
    ⇒ 絶対的ピーク性能追求よりも使い易さと一定の高性能
- single systemの強み
  - 通常運用では相互の予算に按分されたリソース共有
  - 特別運用(例:Gordon Bell Challenge)では全システム占有利用も可能
  - 大規模システム調達によるスケールメリット

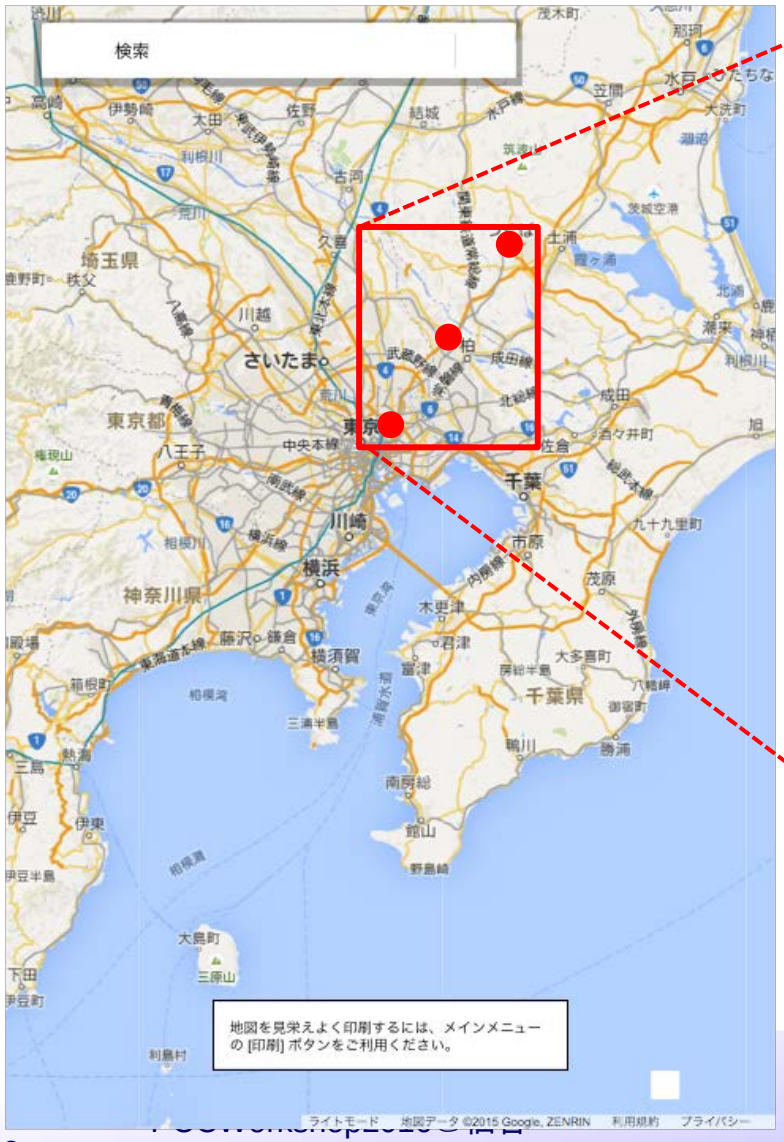
JCAHPCシステムの特徴(続き)

- 計算ノード
  - メニーコアアーキテクチャ／テクノロジによる汎用コアベースの超高性能計算ノード
  - アクセラレータなし、OpenMP+MPIをベースとするコーディング
    ⇒ 従来システムからの連続性
- 相互結合網
  - 100Gbpsクラスタの超高速汎用ネットワーク
  - Full-Bisection BandwidthをサポートするFat-Tree構成
  - 計算ノードと共有ファイルシステムをフラットに収容
    ⇒ flat構造によるスケジューリングの柔軟性と single system image の維持
- 共有ファイルシステム
  - 全計算ノードからフラットに見えるクラスタファイルシステム
  - SSD等によるファイルキャッシュシステム(加点)

主な仕様（仕様書より）

項目	要求仕様
総ピーク演算性能	20～25PFLOPS
プロセッサ	メニーコアアーキテクチャ, X86-64互換
相互結合網リンク	> 100Gbps
相互結合網トポロジ	Fat-Tree (full-bisection B/W)
ノード当たりメモリ	> 96GiB (low speed) + > 16GiB (high speed)
ノード当たりメモリB/W	> 115GB/s (low speed) + > 850GB/s (high speed)
共有ファイルシステム容量	16～26PB
共有ファイルシステムB/W	265-380 GB/s ～ 1500 GB/s
冷却方式	提案に含む

設置場所：東大柏キャンパス（情報基盤センター内）



筑波大

東大
柏キャンパス

東大
本郷キャンパス

調達スケジュール

- 2013/7 資料招請
- 2015/1 仕様書原案(意見招請)
- 2016/1 仕様書、入札公告
- 2016/3/30 入札締め切り
- 2016/4/20 開札
- 2016/10/1 第一次システム運用開始
  (フルシステムの5%以上)
- 2016/12/1 フルシステム運用開始
- 2017/4 HPCIを含む本格運用開始(予定)
- 2022/3 システム運用終了(予定)

システム運用イメージ

- 通常運用
  - ベースラインとして筑波大と東大で予算に応じたノード時間積のリソースを按分
  - 特定の買い上げパーティションを除き、ノード固定の「資産分配」は行わず、柔軟なスケジューリングを行う
  - HPCIの他、各大学固有の運用プログラムがあり、これらはそれぞれのノード時間積内で収容
- 特別運用
  - 超大規模期間限定運用
    ⇒ 国内最大規模の計算実行プログラム、Gordon Bell Challenge等の特別な機会向け
- 省電力運用
  - 夏期節電期間等では power capping を行い一定数のノードを休止（ダイナミック）

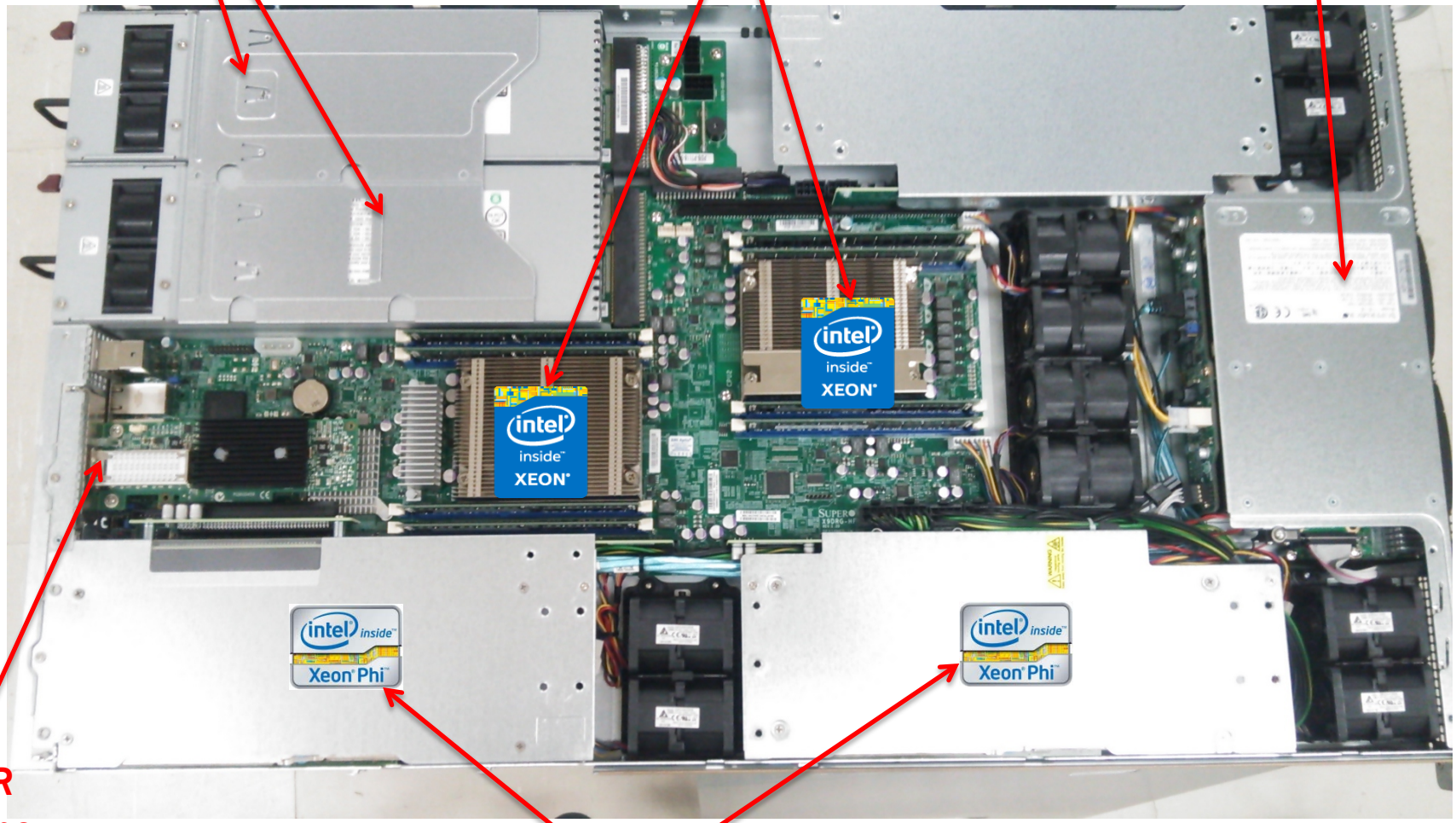
メニーコアシステム予備評価

- 現在利用可能な商用・汎用メニーコアプロセッサとして、Intel Xeon Phi (KNC)を用いたクラスタを両大学で運用中
  - 筑波大：COMA (PACS-IV), 393 nodes, 786 Xeon Phi
  - 東大：64 nodes, 64 Xeon Phi
- 筑波大COMAはHPCI、学際共同利用等の通常プログラムにおいて2015/4より一般運用
- メニーコアプロセッサ固有の特性に応じたアプリケーションチューニング

COMA (PACS-IX)

- Cray社 CS300 ベース
- Intel Xeon Phi (KNC: Knights Corner)を全面採用
- 393ノード（2 Xeon E5-2670v2 +
  2 Xeon Phi 7110P)
- Mellanox IniniBand FDR, Fat Tree
- 2015/10時点でXeon Phi搭載クラスタとして日本最大
- File Server: DDN 1.5PB (RAID6+Lustre)
- 1.001 PFLOPS
  (HPL: 746 TFLOPS)
  June '14 TOP500 #51
- HPL効率 74.7%

COMA (PACS-IX) 計算ノード (Cray 1U 1027GR)



ARTED: 電子動力学シミュレーションコードにおける Xeon Phi向け性能チューニング　（by 廣川祐太@筑波大）

- 第一原理計算に基づく電子動力学計算コード
  - 筑波大学計算科学研究センターの in-house code
  - 電子の波動関数のハミルトニアン計算が支配的
    - 特に**25点ステンシル計算が計算時間の大半**を占める
  - Fortran90 で実装，メインターゲットは京コンピュータ

- 波動関数は倍精度複素数で下記のパラメータで表される
  - (NZ, NK, NB, NL)
  - NZ　　: マクロ格子点
  - NK　　: Bloch Wave Number k
  - NB　　: Wave Band
  - NL　　: 3次元空間格子 (NLx, NLy, NLz)

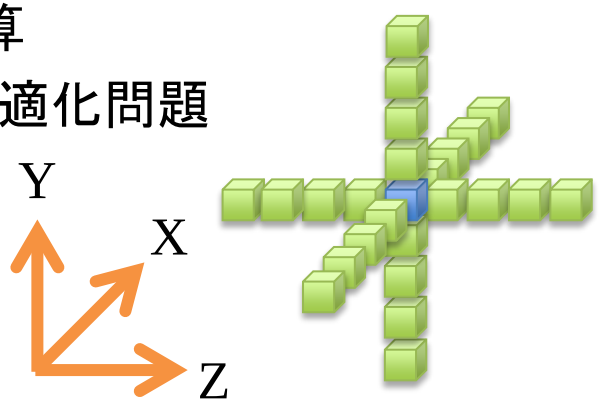
波数空間

実空間

ARTED の計算

- 計算領域は波数空間を MPI + OpenMP で並列分散
  - 波数空間のサイズが実空間よりも非常に大きい
  - 実空間は L2 キャッシュに載る程度に小さい
  - 波数空間の分割によって袖領域の交換が不要
    - 通信時間がボトルネックとならない

- 周期境界条件の 25 点ステンシル計算が支配的
  - 158 FLOP / Point
  - OpenMP **1スレッドで1個の空間格子**を計算
  - シングルスレッドでのステンシル計算の最適化問題



ステンシル計算コード（オリジナル）

```
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
```

間接参照配列：近傍点のインデックスを保存

書き込んだ値は使用しない

長さ4の複素数ベクトル演算となり，**512-bit SIMD** 命令1個で計算できる

自動ベクトル化 (Compiler Vec.)

```
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)
```

3次元配列に変換

```
#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
```

インデックスを直接計算

```
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
```

キャッシュを経由しない書き込みを指示

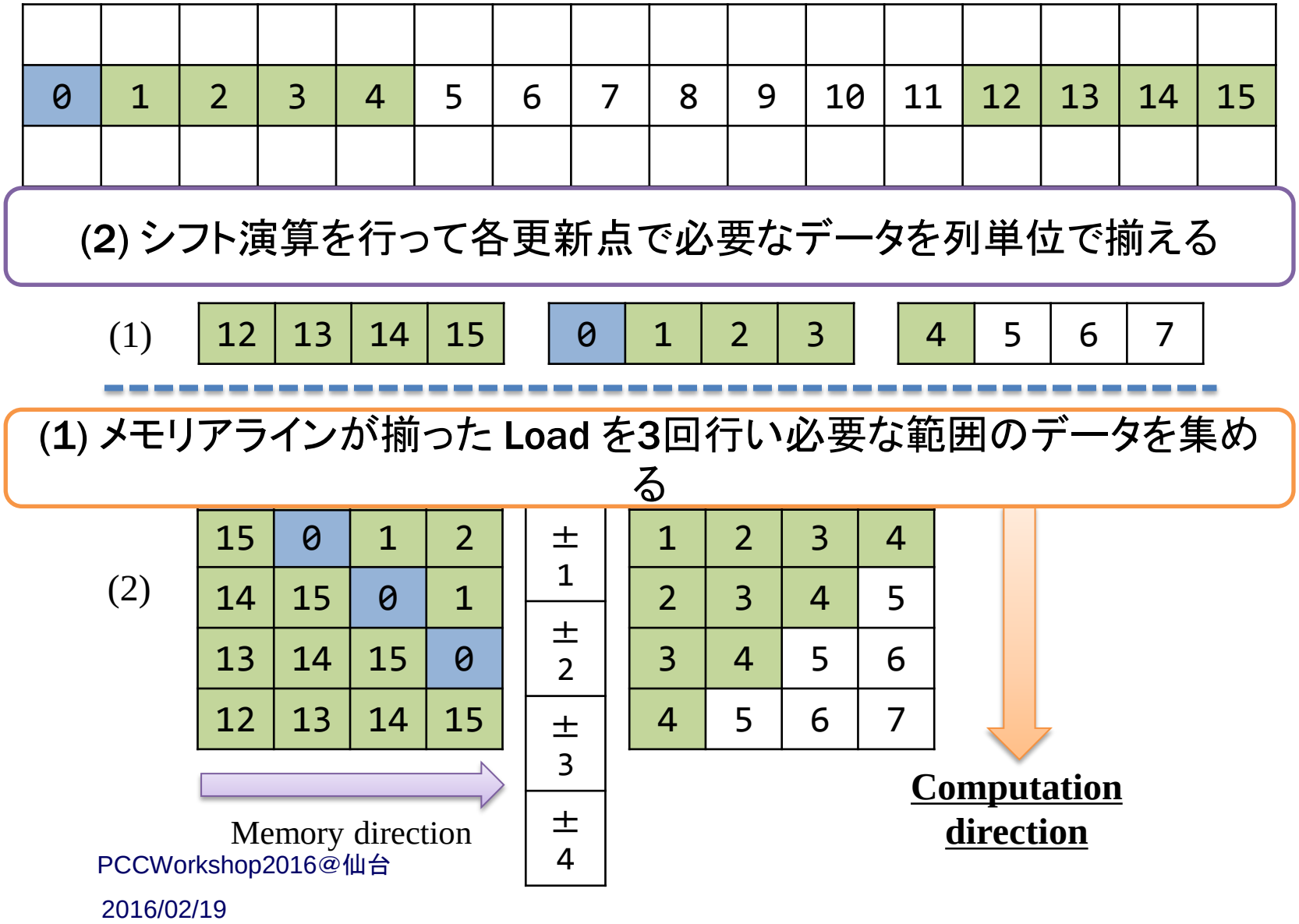
メモリ上連続な領域から計算

手動ベクトル化 (Explicit Vec.)

- 下記の問題点が考えられる
  1. 複素数積の最適化 (省略)
     - Xeon Phi は複素数積用の命令が未実装
     - 定数積のため展開して計算
  2. 連続方向のメモリアクセス最適化
     - 必ずアラインがずれたメモリアクセスが発生

- 本研究では，空間格子点サイズに制限を設ける
  - NLz (メモリ上連続方向) のサイズを4の倍数に固定
  - ベクトル長で割り切れるように

連続方向のメモリアクセス最適化

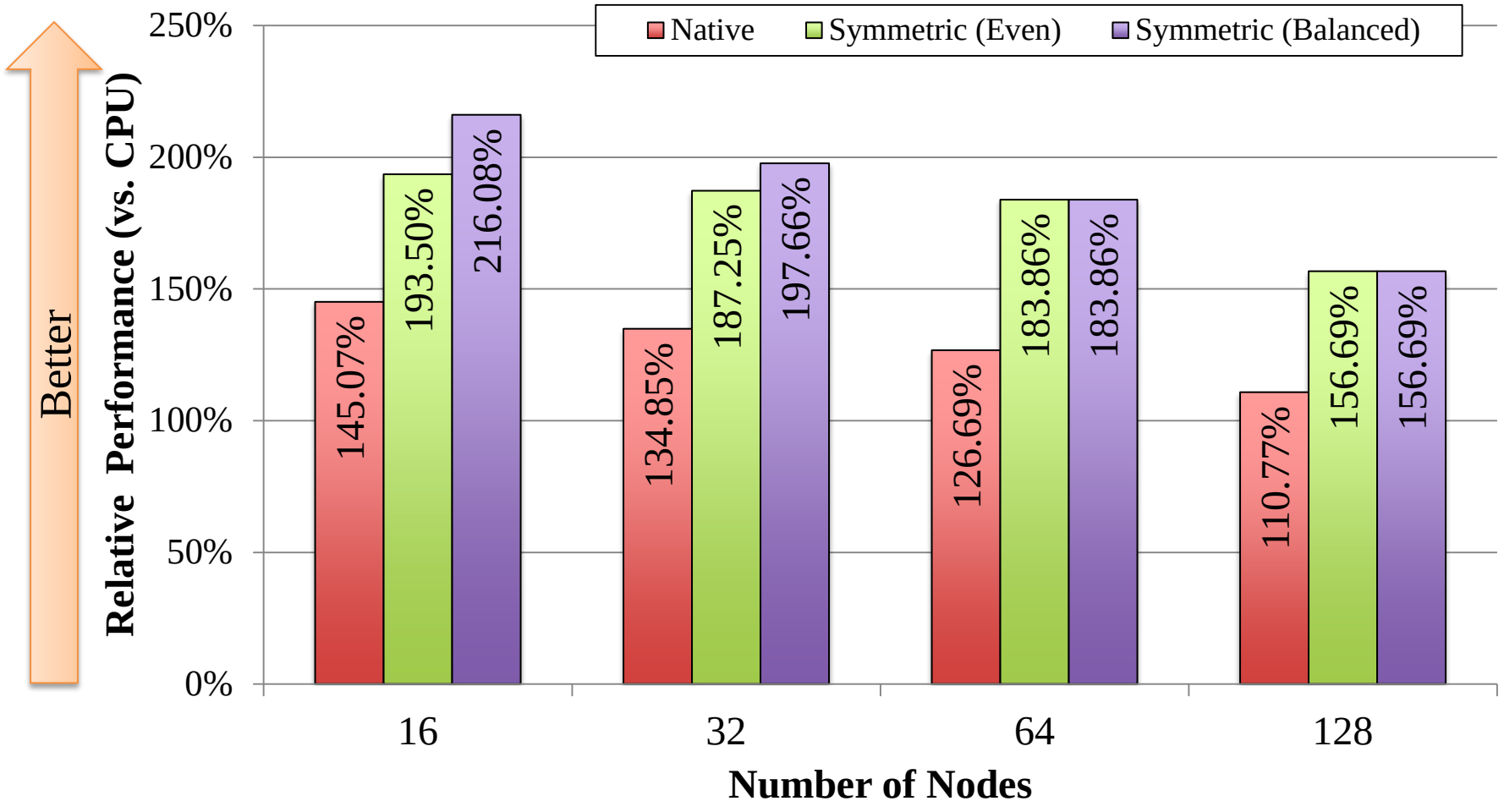


ステンシル計算性能

- (NK, NB, NL) = (8^3, 16, 16^3) とし 1 プロセスで計算
  - Xeon Phi では高い並列性が必要となる
  - ここでは NK を大きく取り，並列性を高くする ($8^3 \times 16$ 並列)

	Type	GFLOPS	ピーク性能比
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 %

全コードのStrong Scaling 性能評価 (CPU との相対性能)



N 台の CPU ノードでの実行性能 ≦ N/2 台での Symmetric 実行性能

まとめ

- 筑波大学と東京大学の共同運用によるJCAHPCにおいて、最大25PFLOPSピーク性能のメニーコア型大規模クラスタを2016年度下半期から運用予定
- 国内初の2大学の共同調達・共同運用によるスケールメリットを活かした大規模システム導入
- メニーコアプロセッサの利用・チューニングは今度のトレンド⇒ポスト京にもつながる高性能計算技術
- 本システムは国内最大規模の汎用スーパーコンピュータとして、今後様々な局面で重要な役割を果たしていく予定