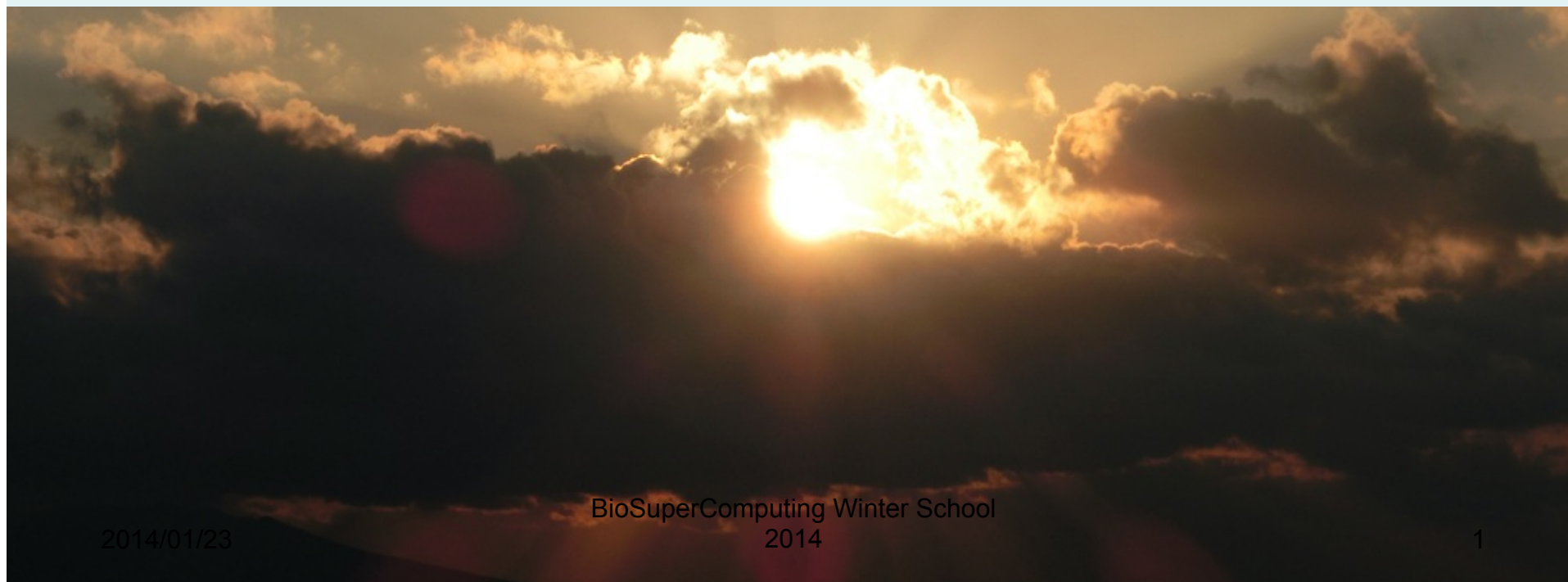


ExaFLOPSへの展望 ～Exaへの道～

バイオスーパーコンピューティング研究会

会長 姫野龍太郎

himeno@riken.jp



京プロジェクトの個人的評価

- ハードウェア
 - 😊10PFLOPs他の目標達成:3.11にも関わらず達成できて良かった
 - 😭目標の電力<10MW、設置面積<3000m²
 - 頑張れば達成できる程度で、いくつかの案のどれでも達成可能。大きな変革が必要な目標にすべきだった
- ソフトウェア
 - 😊実際のアプリで40%を越える実効性能の達成:ナノ, ライフ、天文他
 - 😭Gordon Bell賞を逃した(候補にもならなかった)
 - 😭コンパイラ性能と言語の選択
 - FORTRAN偏重、C++の実効性能の低さ, Java, Ruby, Perl, Python, R, gccなどのサポート
 - 😭商用ソフトがポーティングもサポートされない
- AICS
 - 😊国内に計算科学に関する研究センターが設立
 - 😭予算の名目から研究が歪む
- 京プロジェクトの成否の一つは富士通製の京互換システムが商業的に成功するかどうかで決まる:○/△?

ISLIMに関する個人的評価

- 😊 実際のアプリで40%を超える実効性能の達成: ZZ-EFZI他
- 😊 生命科学に関する学際的なネットワークができた
- 😭 京を利用できる期間が大幅に短縮、せっかくソフトを作ったのに使えなかった
- 😭 開発したソフトの利用が必ずしも進まず、維持・向上も
- 😭 人材の流出

ともかく、やって良かった！！

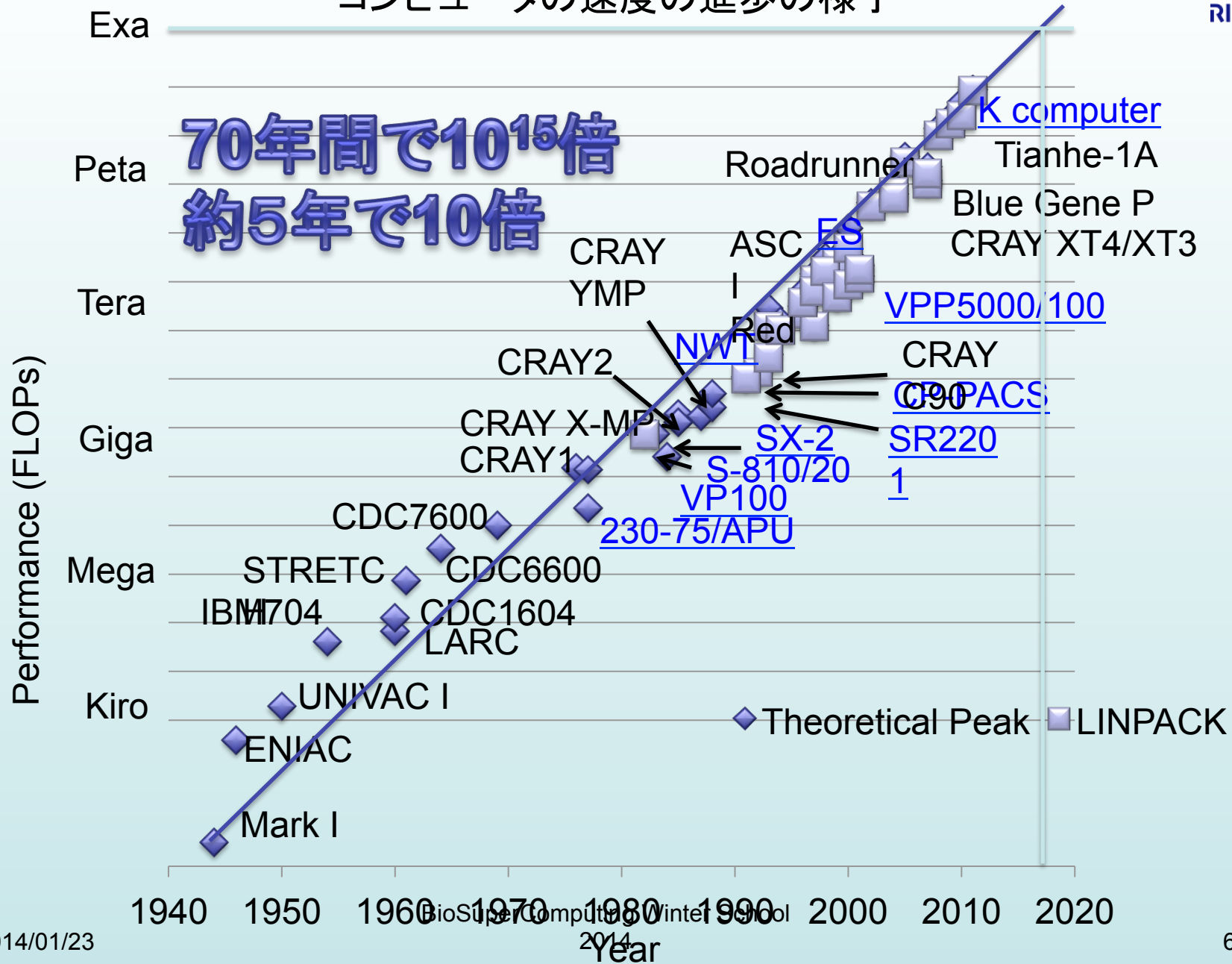


この反省に立って、次のExaに向けて

1. ハードウェアは速くなり続けられるか
2. アプリ側で高速化のために考えないといけないこと
3. ソフト開発の仕組みで考えないといけないこと

1. コンピュータは速くなり続けるか

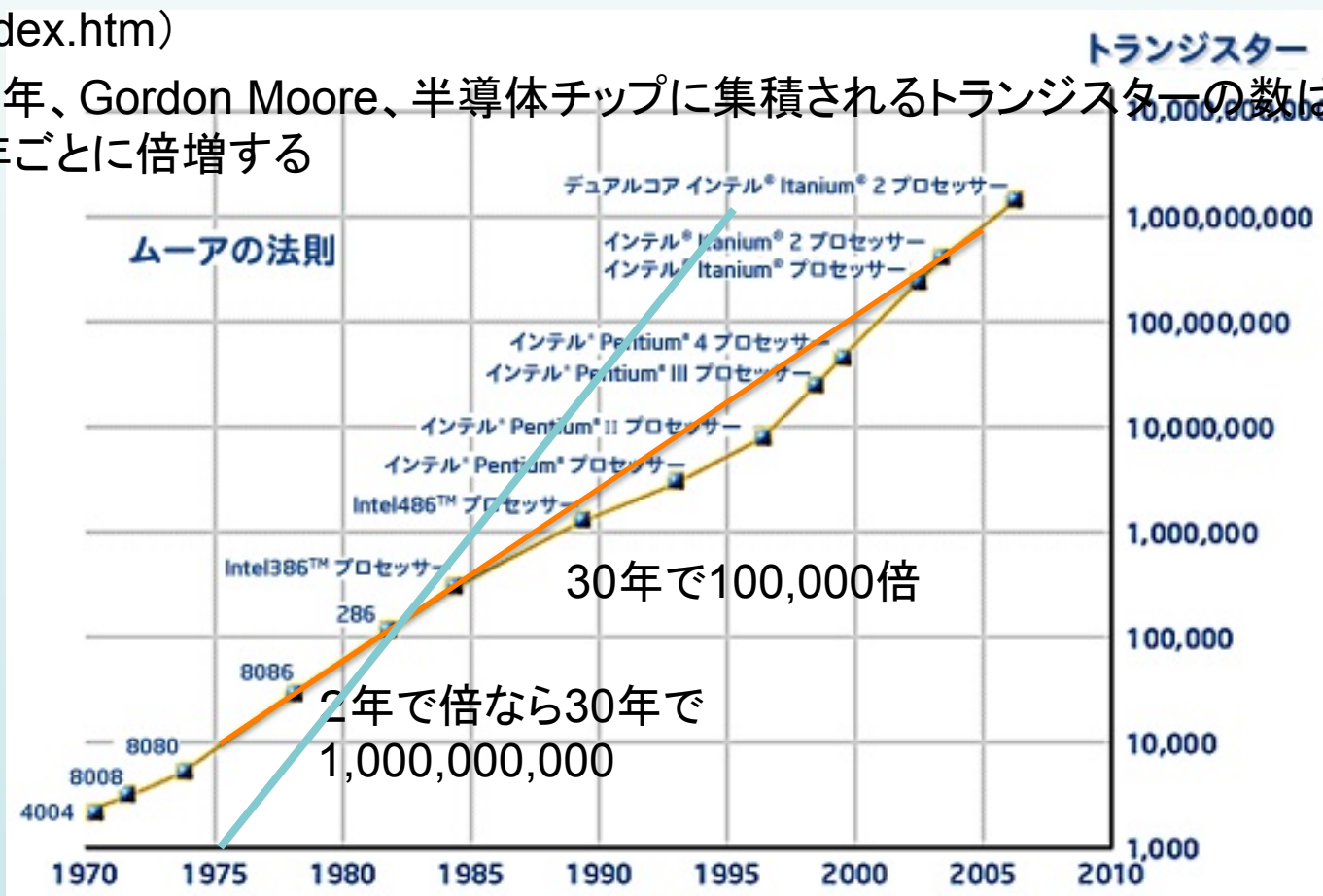
コンピュータの速度の進歩の様子



LSIの進歩

ムーアの法則(出典: <http://www.intel.co.jp/jp/technology/mooreslaw/index.htm>)

1965年、Gordon Moore、半導体チップに集積されるトランジスターの数は約2年ごとに倍増する



コンピュータ性能の向上のトレンドの変化

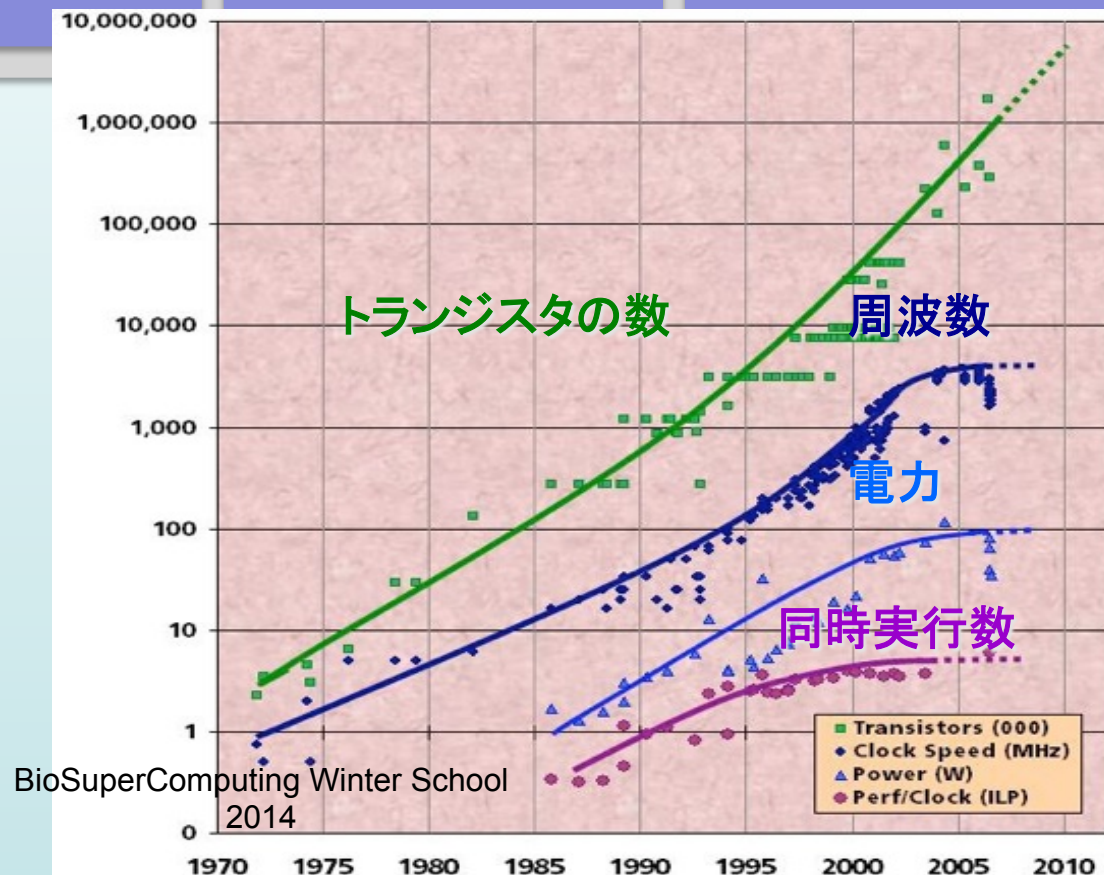
線幅を狭くした結果、漏れ電流が増え、消費電力が下がらなかった

発熱は冷却の限界に
• >100W. 発熱密度はホットプレート並み

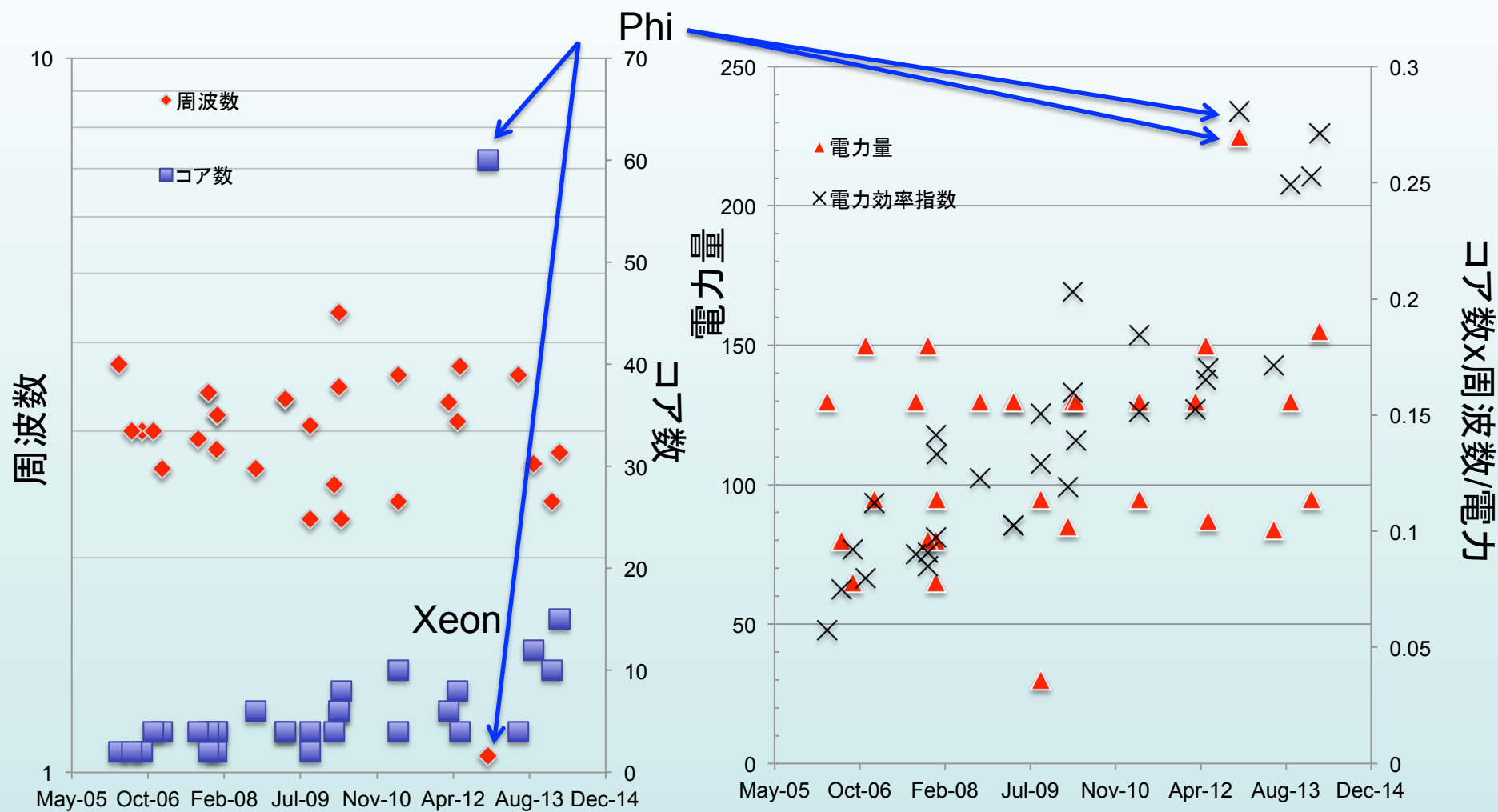
その結果、周波数は増やせなくなっている

性能向上はプロセッサのコア数の増加で対応
1→2→4→8→?

- 並列で処理できないプログラムは速くならない
- 何を速くするかが設計の肝



Intelのサーバー用プロセッサ: 二つの方向



プロセッサのトレンド

- トランジスター数の増加は依然継続
- プロセッサの設計は二つの方向
 - Many core方向
 - 設計を単純化してコア数を増加。
 - 消費電力を下げるため、平均よりも周波数は低い
 - 実行時間短縮方向
 - 周波数は3-4GHz
 - コア数は徐々に増加、複雑な機構で演算効率を向上

Top10の現状

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Super Computer Center in Guangzhou China	Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2 Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	Titan - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,848	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,585.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect, NVIDIA K20x Cray Inc.	115,934	6,271.0	7,788.9	2,325
7	Texas Advanced Computing Center/Univ. of Texas United States	Stampede - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi Dell	462,462	5,168.1	8,520.1	4,510
8	Forschungszentrum Juelich (FZJ) Germany	JUQUEEN - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301
9	DOE/NNSA/LLNL United States	Vulcan - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.3	5,033.2	1,972
10	Leibniz Rechenzentrum Germany	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147,456	2,897.0	3,185.1	3,423

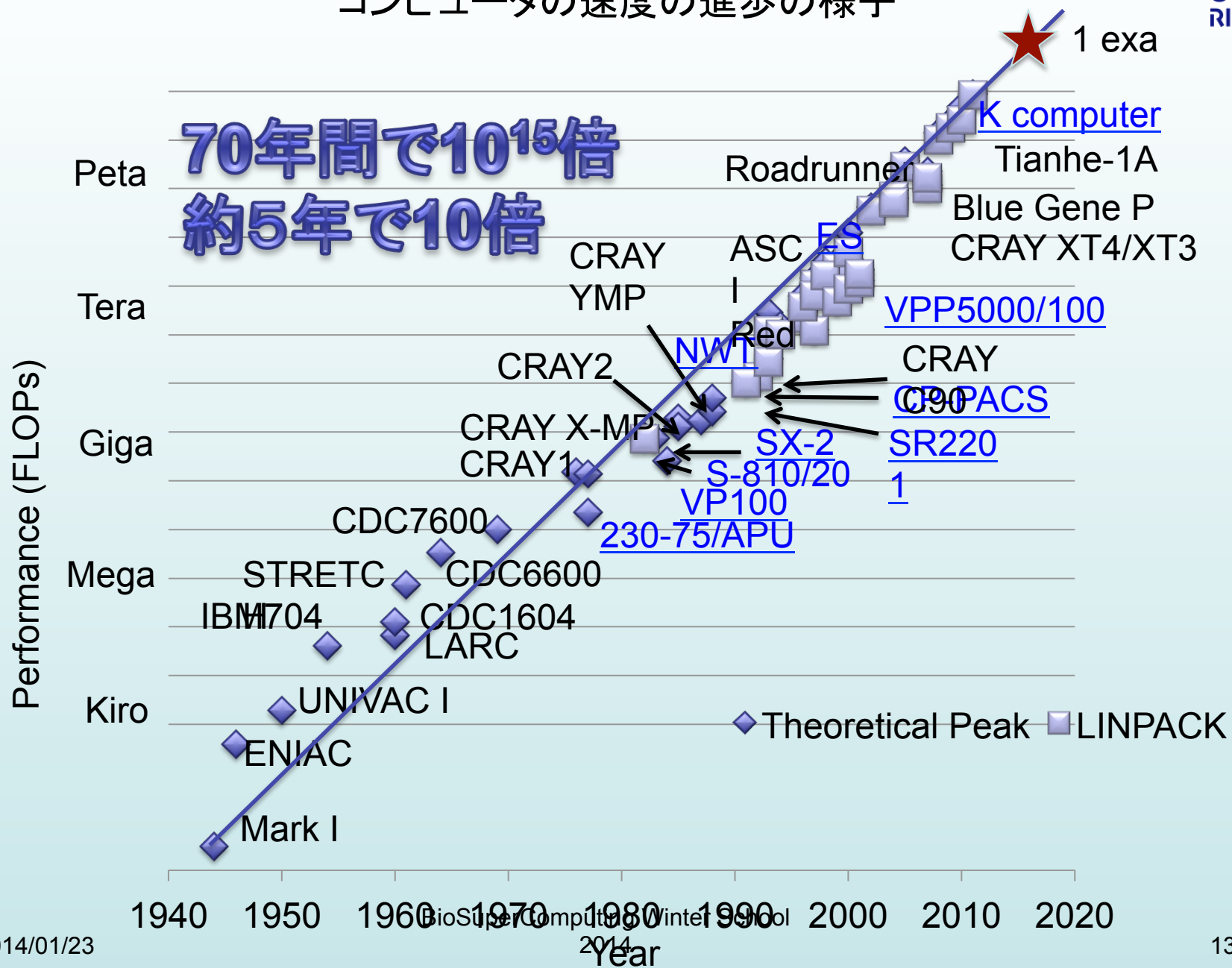
アクセラレータ

Top 10 of Green500

全てアクセラレータ付き

Green500 Rank	MFLOPSW	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR NVIDIA K20x	922.54
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14 NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR Nvidia K20m	71.01

コンピュータの速度の進歩の様子



そのとき、それぞれの性能は？

単純に技術進歩だけで実現しようとする、
電力は100MW

System attributes	2010	“2015”		“2018”		Difference 2010 & 2018
System peak	2 Pflop/s	200 Pflop/s		1 Eflop/sec		O(1000)
Power	6 MW	15 MW		~20 MW		
System memory	0.3 PB	5 PB		32-64 PB		O(100)
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF	O(10) – O(100)
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec	O(100)
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)	O(100) – O(1000)
Total Concurrency	225,000	O(10 ⁸)		O(10 ⁹)		O(10,000)
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec		O(100)
MTTI	days	O(1day)		O(1 day)		- O(10)

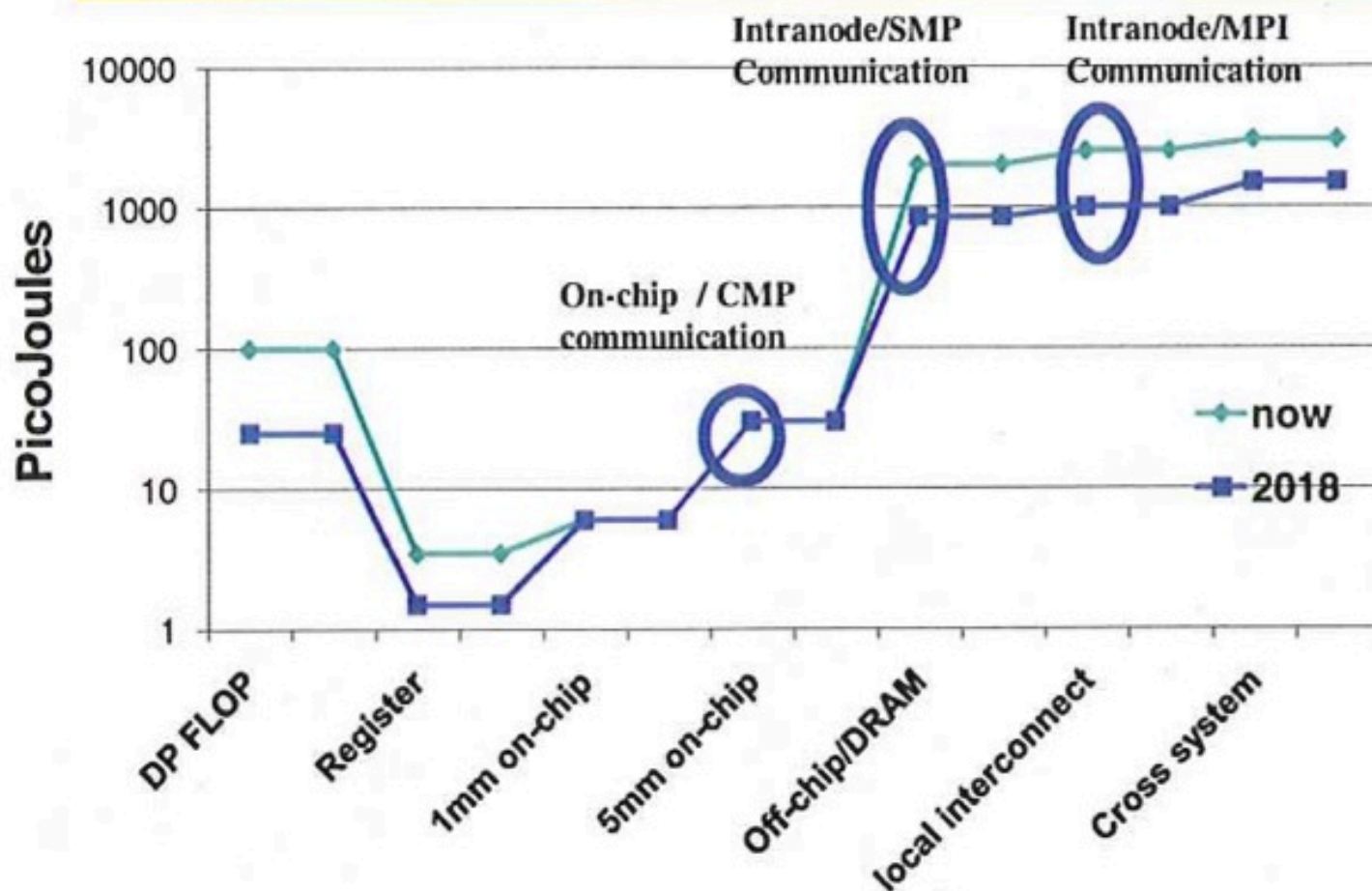
BioSuperComputing Winter School



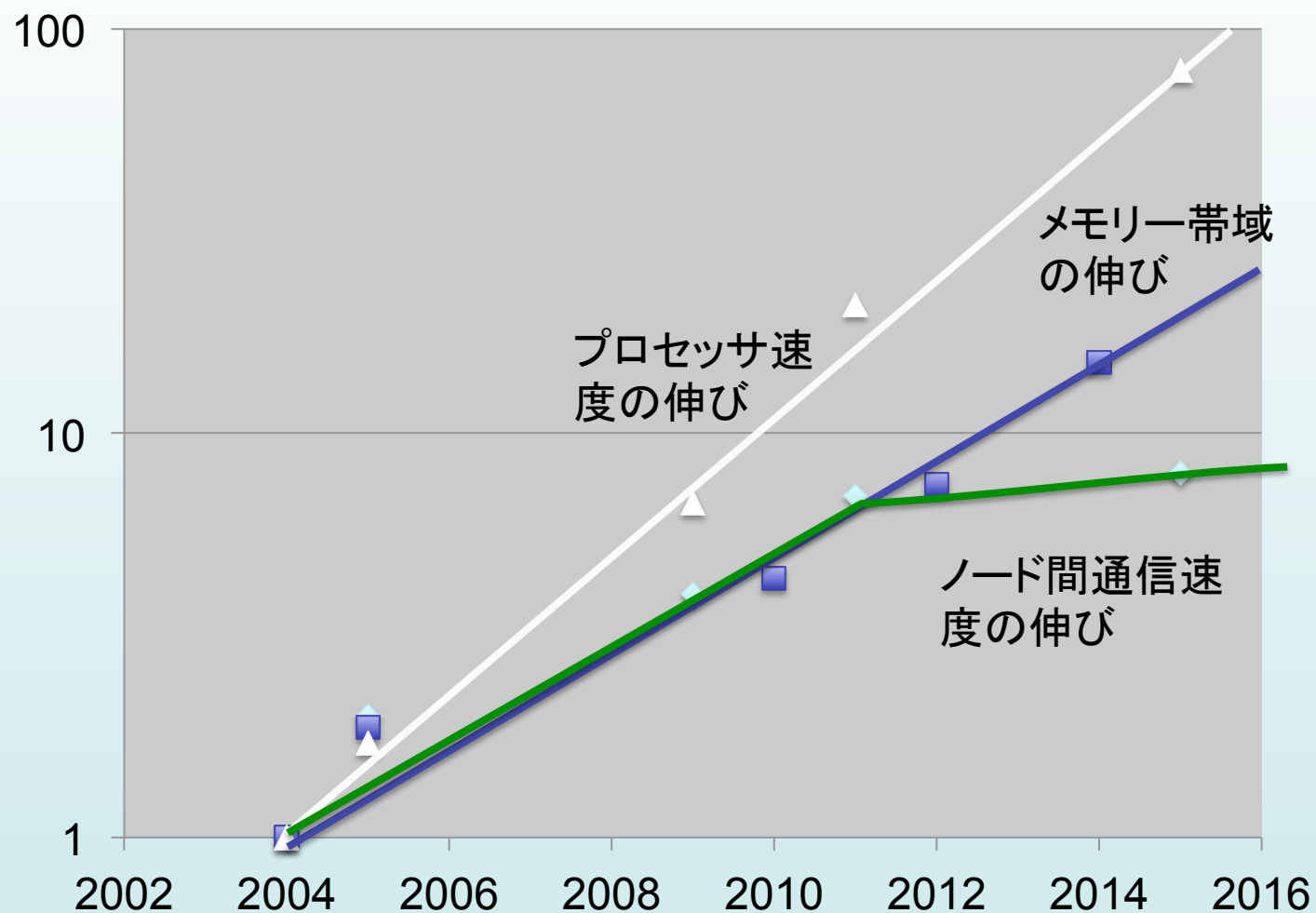
U.S. DEPARTMENT OF
ENERGY

Office of Science

Why Exascale is Hard



今後のコンピュータの要素性能のトレンド



システム設計の基本的考え方

○将来動向

- 高い性能電力比と不規則非構造データ処理もできるCPUとして、汎用コアと演算加速コアが統合されていく。しかし、どのように統合されていくかは今後の研究成果に依存
- ノード単体の演算性能向上に比べ搭載可能メモリ容量はさほど増えない

○開発に関する考え方

- システム設計

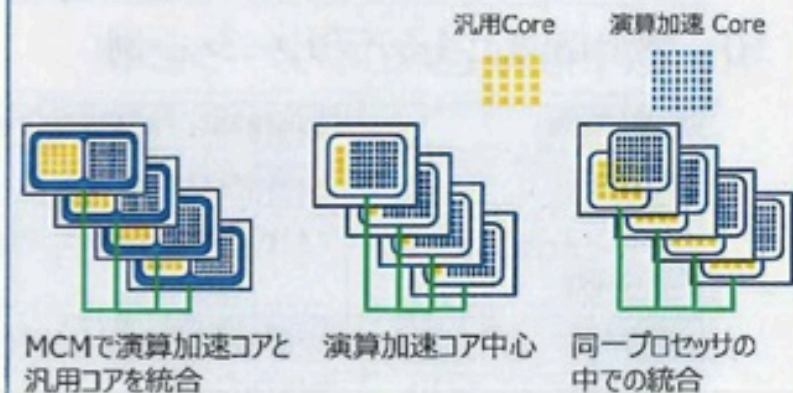
汎用部（汎用コア）と演算加速部（演算加速コア）を有するマシン

- 汎用コアでないと性能がだせないアプリケーション、演算加速コアで性能だせるアプリケーション、両コアを使って性能がだせるアプリケーションを棲み分け、全電力時間積削減
- 将来にわたって有効な統一プログラミングモデル、ライブラリ、フレームワークを提供
- 早期成果創出のためにキラーアプリケーションの同時開発を通してシステム設計に反映

	汎用コア (Server)	演算加速コア (GPU, SIMD)
データ構造	不規則・非構造	規則構造
性能電力比	低 Ex. 4.27GF/W, 22nm	高 Ex. 7GF/W, 28nm
メモリ階層の現状	キャッシュと主メモリ	GPU側とホスト側メモリ

将来マシンイメージ群

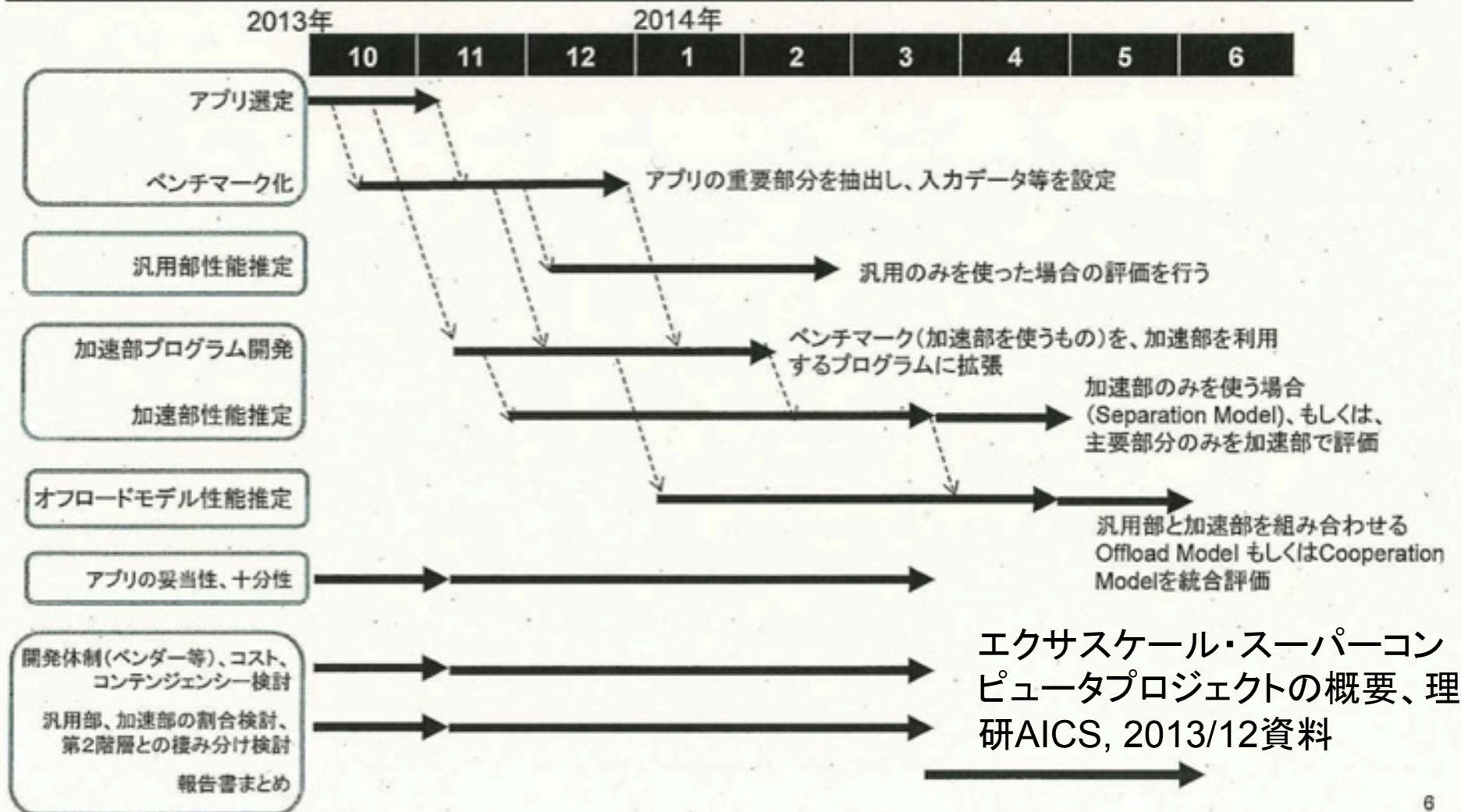
- 汎用コア、演算加速コア、ネットワークが統合されていく
- 統合方法および搭載メモリ容量とメモリバンド幅は、実装技術によって変わっていく
- 3D実装可能メモリ容量以上が必要な場合、低速メモリが導入される





フラグシップシステム概念設計性能推定等工程表

方針：社会的課題を含む課題を、計算科学ロードマップ白書(第2版)から選定。ベンチマーク化し、汎用部および加速部を用いた評価を進める。ファイルI/Oに関しては2014年度実施。

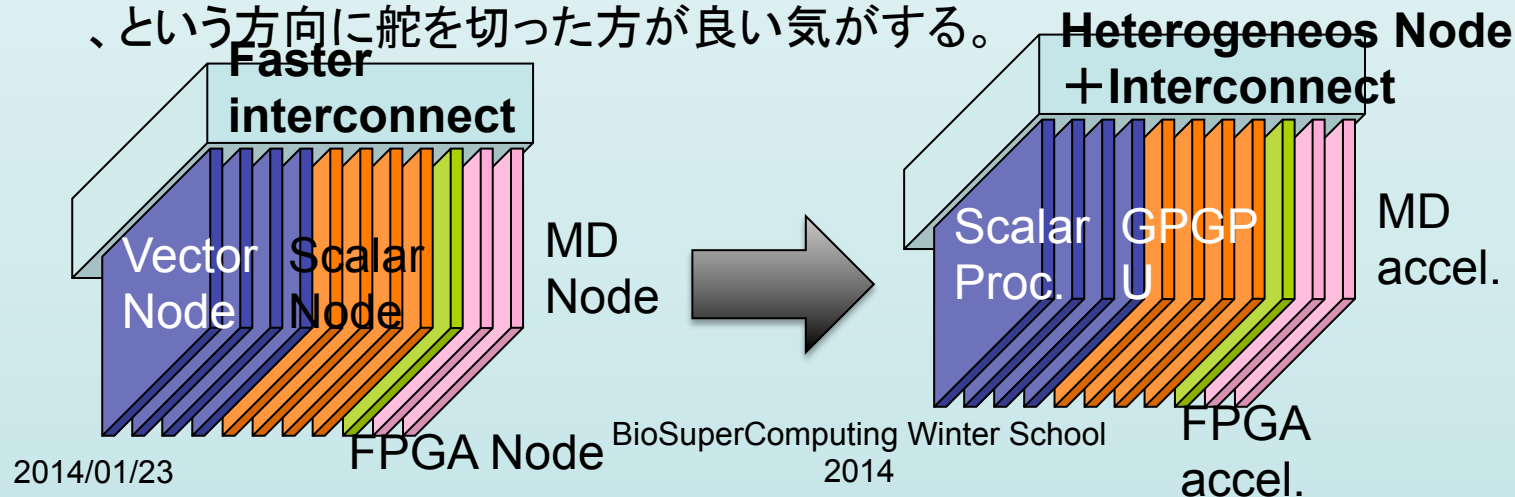


電力が最大の問題となるExaでは

- アクセラレータの採用は合理的
- ただ考えなければならないことが
 - プログラミングの負荷が高まる。問題を解くまでの道のりが長くなる。
 - アクセラレータを使うと、全てのソフトが速くなる訳ではない。合うソフトのみ。
 - アクセラレータ用の言語をどうするか、まだ流動的。OpenCL?
 - プログラムはそのアクセラレータに特化して開発が必要
 - アプリ開発用にアクセラレータが開発者に早めに供給されるか
 - PCやPCクラスタ用も同時に提供しないと普及しない
 - 国策として日本で独自に開発するとすると
 - 継続して開発し続ける事が必要: でないとソフトが移植されない
 - NVIDIAやIntelと戦い続け、1～2年毎に新製品を供給できるか
 - 誰がそれを保証できるか
 - 2年で性能的には陳腐化、4年で交換が必要か?
 - 商業ベースに乗るか?

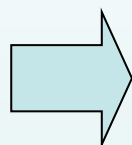
個人的には

- アクセラレータを独自開発し、商業的に展開する事は無謀
 - 商業的に成功しない限り、一時的な消え行くあだ花になる
 - 応用範囲を絞り込んだ専用機は有望:Anton, MD-GRAPE4
- 分野毎の専用機を簡単に開発できる仕組み作り(プラットフォームやキット)を開発する事の方が有意義
- 分野のアプリに特化して簡単に開発できる専用アクセラレータと、それを予算に合わせてスケーリングできるノード・インターコネクトを開発するのが得策では？
- 科学技術計算に特化したスーパーコンピュータより、人と会話でき、人の問いに答えられたり、人に代わって乗り物を操縦したり、病気の診断や手術ができた、という方向に舵を切った方が良い気がする。



トレンド

科学技術計算に
特化した専用設
計のコンピュータ

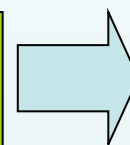


PCベース
のスパコン

組み込みチップ
ベースのスパコン

汎用チップ+ゲー
ムチップ

汎用チップ増強型



PC+アクセラ
レータ

組み込みチップ
増強型スパコン

汎用チップ増強型

電力とコスト
なら

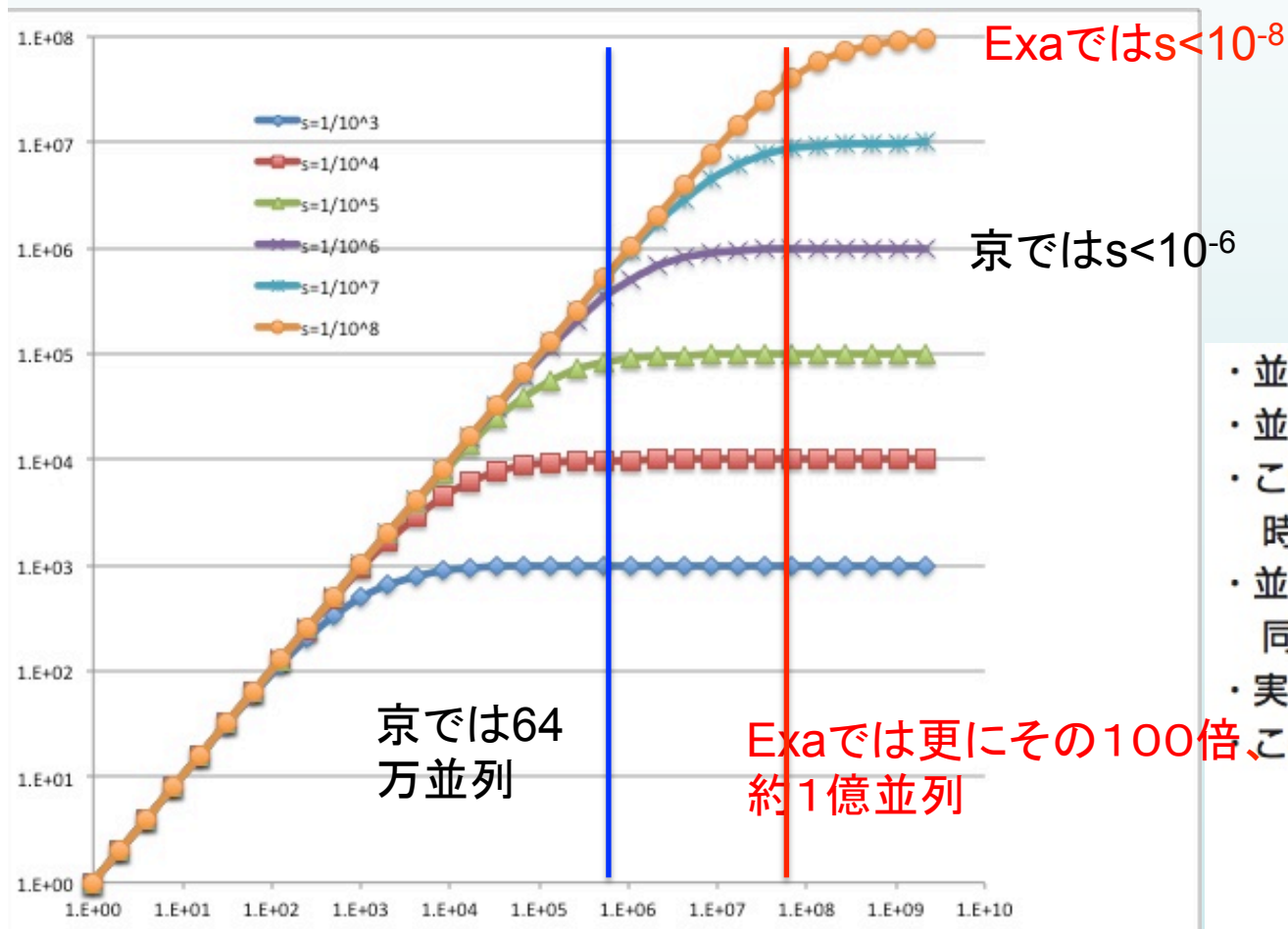
組み込みチップ
増強型スパコン

携帯チップ増強
型スパコン？



アプリ側で考えなければならない事

アムダールの法則から見たExaの難しさ



- ・ 並列化できない部分の比を s とすると
- ・ 並列化できる部分は $(1 - s)$
- ・ この部分は n 個のプロセッサで $1/n$ の時間で実行可能
- ・ 並列化できない部分はいくつ使おうと同じ時間がかかる
- ・ 実行時間は $s + (1 - s)/n$ この逆数が速度向上比

$$\frac{1}{s + \frac{1 - s}{n}}$$

速く計算するために使う側で考えること

計算を分析

多くの計算を同時にできるように考える

数を増やす → 複雑な計算の方が有利

ばらつきをなくす

通信を減らす

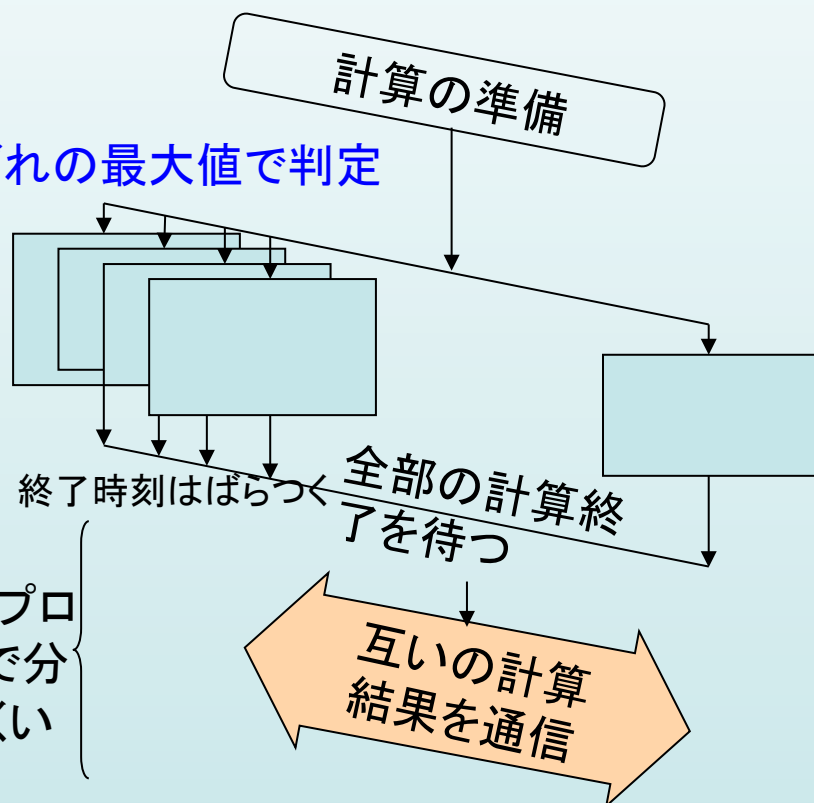
例えば誤差の合計 → 誤差のそれぞれの最大値で判定

プロセッサ数に比例して高速化

高速化できない部分

プロセッサ
毎に分担で
きる部分

複数のプロ
セッサで分
担しにくい
部分

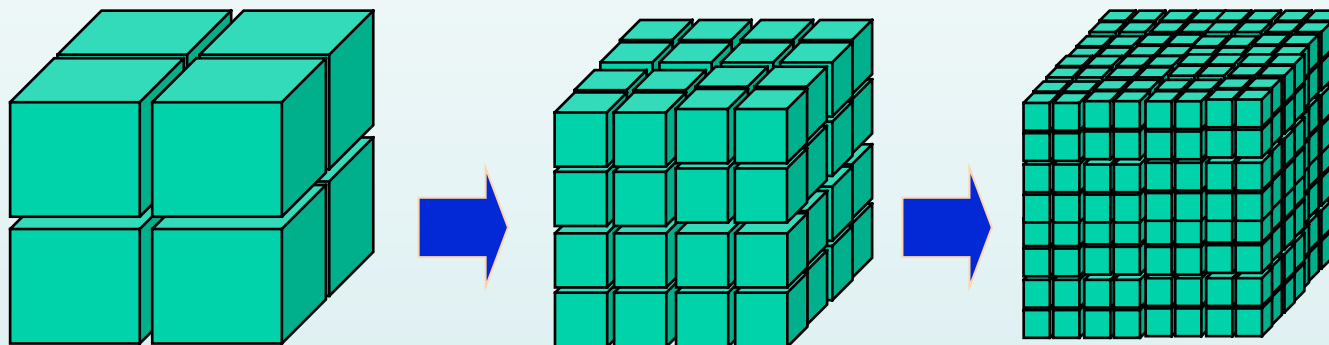


空間分割

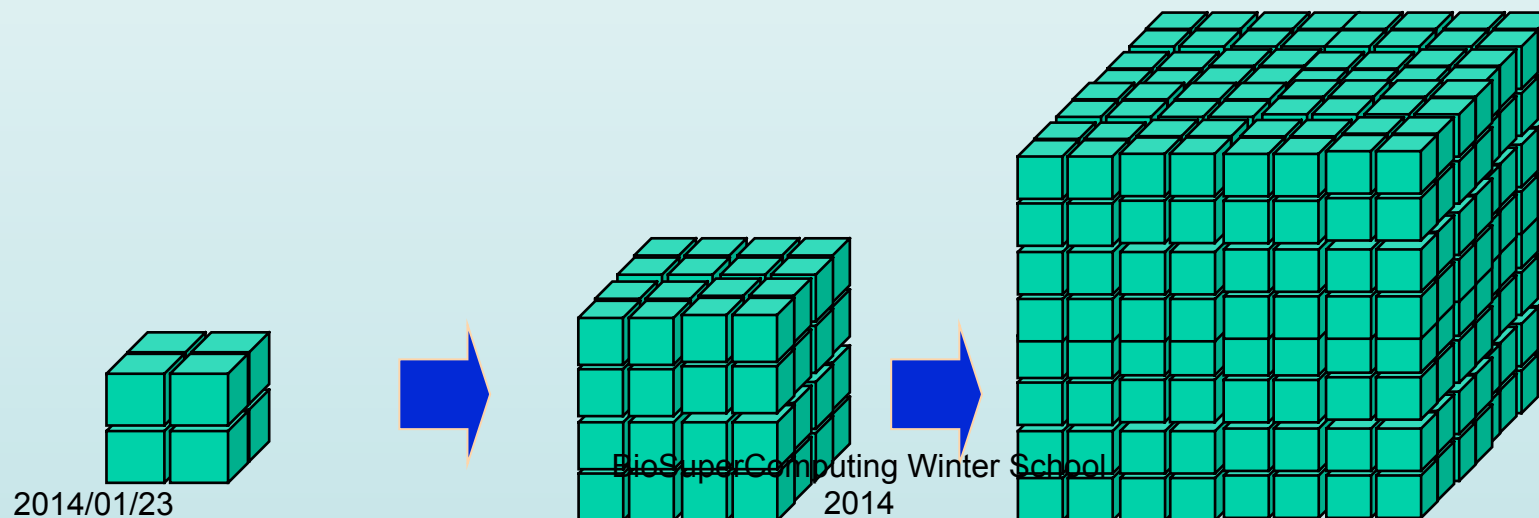
計算量は体積に比例

通信は表面積に比例

問題規模一定で分割を増やすと、一つあたりの通信が相対的に増える



CPU数に比例して計算規模を大きくすると、同じ比率を維持



2014/01/23

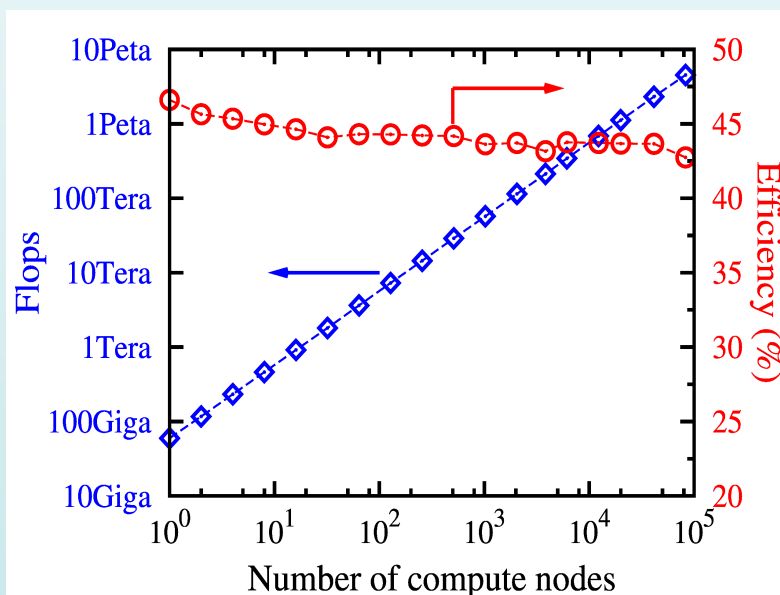
BioSuperComputing Winter School

2014

25

通信性能がプロセッサ性能に比べて速くならないことから(1)

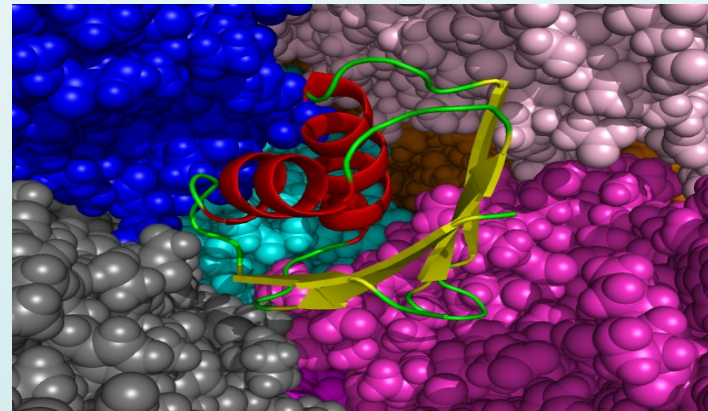
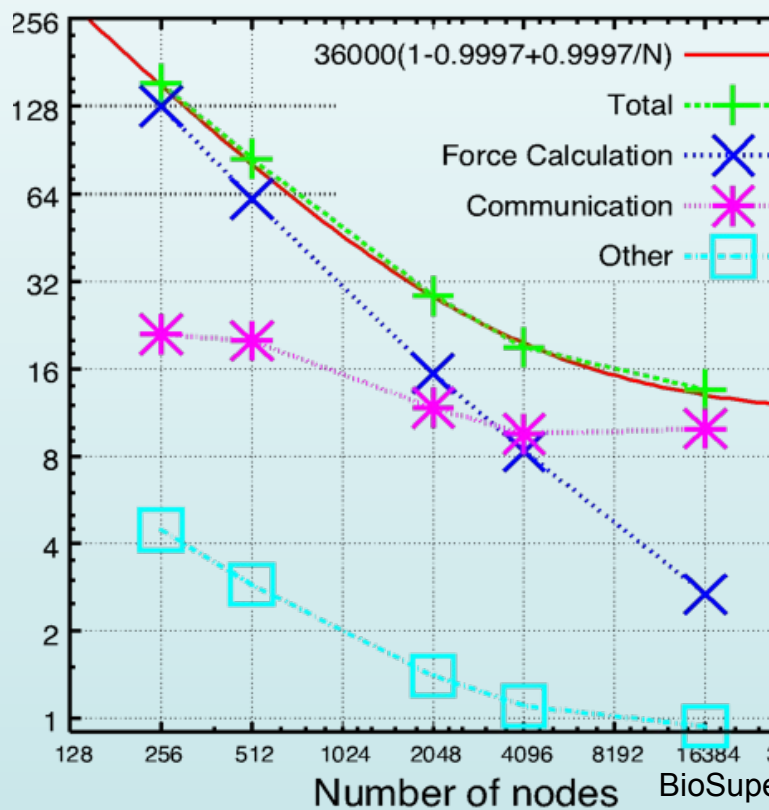
- 問題規模を大きくしないと性能が出ない
 - 京(10PetaFLOPs)でも全体を使った流体計算は1週間から1ヶ月かかる
 - 現実的には京の1/10程度を使って計算
 - 京の計算効率を維持しようとする、計算メッシュを100倍(三方向4.3倍)に
 - その結果、時間方向の安定性から時間方向には1/4.3倍
 - 結果的に、計算量は430倍に。必然的に計算時間は今より4.3倍に伸びる!!



ZZ-EFSIの場合の
実測

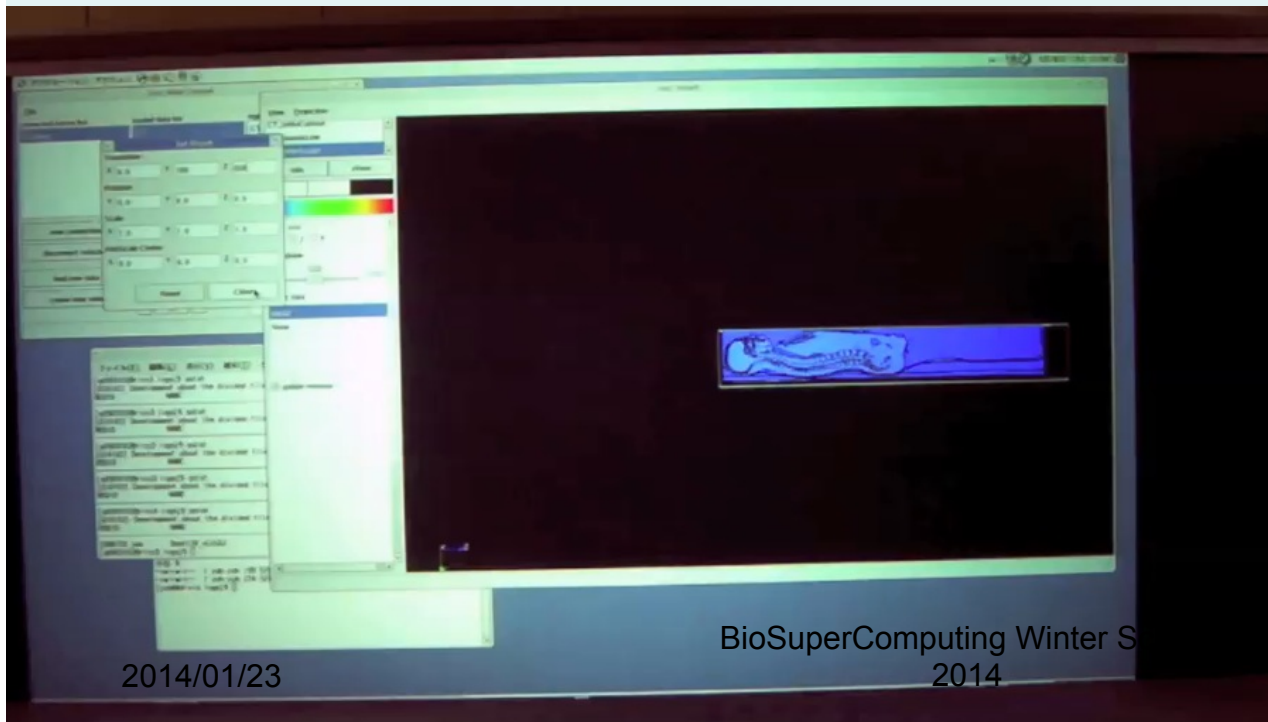
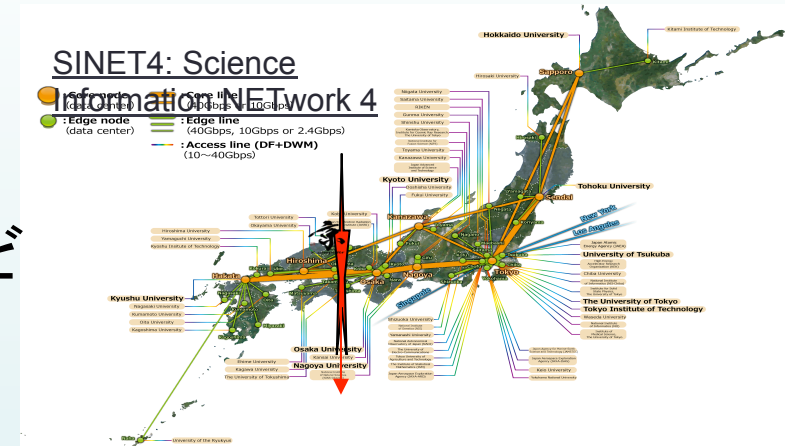
通信性能がプロセッサ性能に比べて速くならないことから(2)

- 問題規模を大きくしないと性能が出ない
 - MDの場合も同様、原子数を増やして演算量を増やさないと、通信の比率が上がり、効率が下がる。原子数を増やすと当然、計算時間が伸びる
 - 創薬等では原子数がそれほど大きくないところで、長時間計算する方が重要



もう一つの問題

- 結果を見る: インタラクティブ・動画
- 広域での転送が必要
- 長距離のネットワーク帯域はそれほど増えていない。



2014/01/23

28

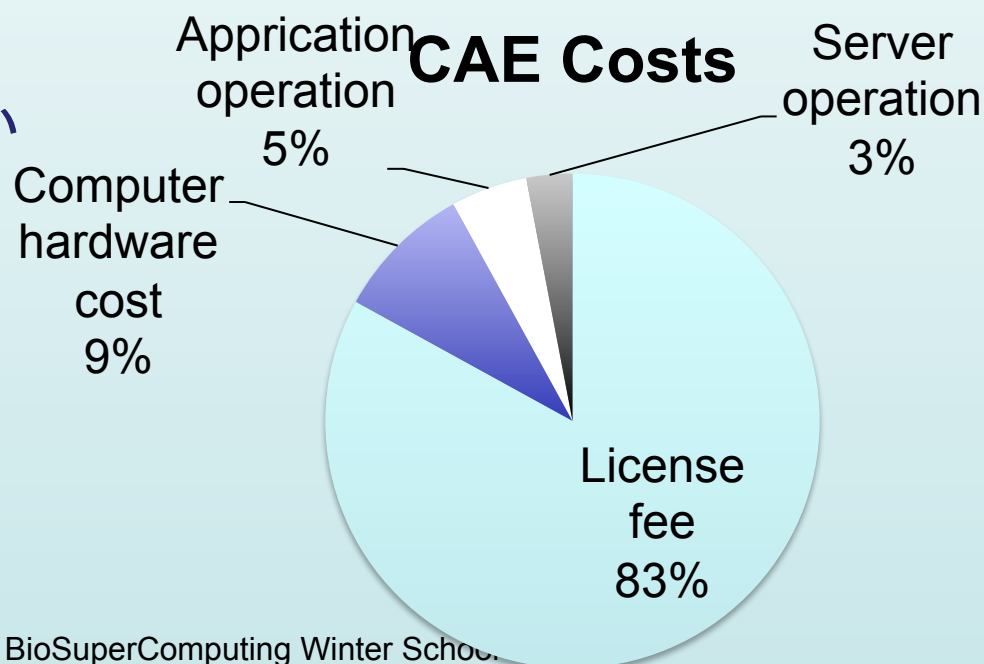
- 問題規模を大きくしないと性能が出ない
 - 京(10PetaFLOPs)でも全体を使った流体計算は1週間から1ヶ月かかる
 - 現実的には京の1/10程度を使って計算
 - 京の計算効率を維持しようとする、計算メッシュを100倍(三方向4.3倍)に
 - その結果、時間方向の安定性から時間方向には1/4.3倍
 - 結果的に、計算量は430倍に。必然的に計算時間は今より4.3倍に伸びる!!
- コア数は増えるけどコアあたりのメモリーバンド幅は増えない
 - 演算数の少ないアルゴリズム vs. **演算数の多い複雑なアルゴリズム**
 - 反復回数の多いアルゴリズム vs. **反復回数の少ないアルゴリズム**
- 通信のバンド幅は増えるが、コアあたりで見た通信バンド幅は増えない
 - ローカルに処理できた方が有利
 - 誤差の総量で反復を判定 → 局所的な誤差の最大値で判定
- アクセラレータ
 - 電力効率からアクセラレータの利用は今後も続く
 - プログラミングの観点からは複雑化
 - ライブラリーや分野別パッケージの共同開発と利用が必要

- 時間進行: 陰的か陽的か
 - 時間ステップ回数が少ない陰的解法がこれまで有利
 - > 通信がネックになる可能性
 - ローカルに決まる陽的が有利
- 単純な計算から、複雑な計算へ
 - 形状の複雑さより、計算アルゴリズムの複雑化の方が計算効率上有利
 - 単純な解析から、化学反応を伴うようなマルチ・フジックスへ
 - 単一のスケールの解析から、ミクロからマクロまでを解析するマルチスケールへ
- 時間方向の並列化
 - 種々の問題では統計量だけが問題、それなら時系列で追いかける必要はないはず。異なる初期状態からの統計処理で、処理できるはず
 - この方向の研究が必要

アプリの開発の仕組みで考えなければいけないこと

CAEに関する状況

- CAE抜きに設計をすることはできない状況
 - コンピュータ速度向上で、精度良く予測が可能で設計開発のコストと期間の短縮が可能
 - 一方で、寡占化が進んだアプリのライセンス料がCAEコストの大きな比率を占めるようになった
 - コンピュータの高性能化の恩恵は多くのアプリで受けられない



市販ソフトの問題

- ライセンス料が高い
 - 使えば使うほど高くなる？
- 計算機の性能向上に追隨していない
 - 計算機性能は上がるのに、同じ問題の計算時間は短縮されない
 - モデル規模を大きくした場合、使うノード数を増やしても計算時間が短縮されない(典型的には10ノード以上で速度向上は頭打ち)

大学や国立研究機関発のソフト

- 多くのソフトがこれまで開発された
 - 流体解析コード、大規模構造解析コード
 - 企業で使われているか？
- 使われていない理由
 - 設計:使われるように作っていない
 - 設計段階からの使う側の参加
 - 信頼性:使い続けられるか
 - 開発プロジェクトが終了すると開発者がいなくなる
 - 改良や機能追加ができるか
 - 計算機の進歩に追隨した性能向上が可能か

フリーソフトの問題

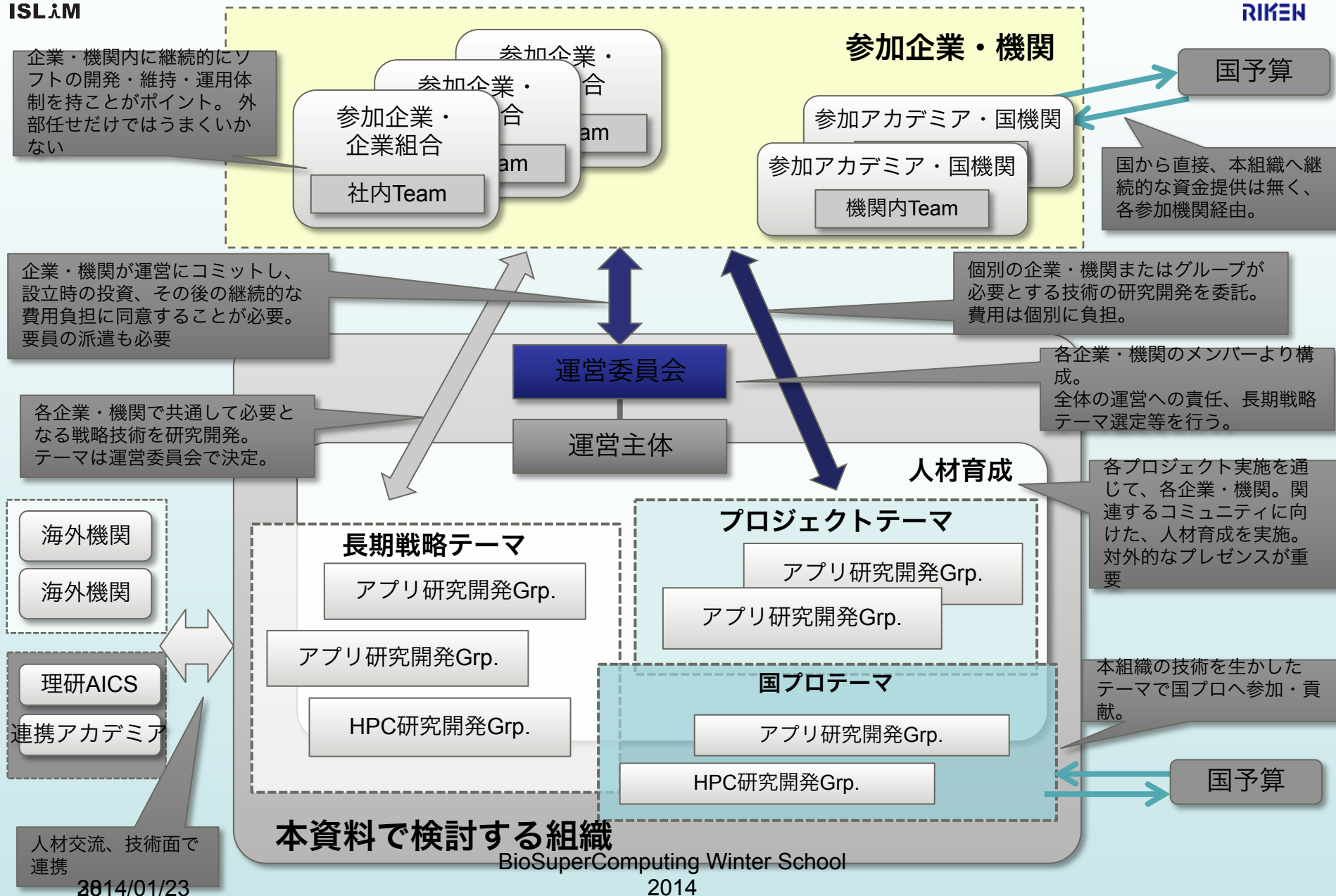
- 信頼して使い続けることができるか
 - 機能追加、性能向上は可能か
 - 一般にはできる。でも、誰が責任を持つか。
 - 現実的には現在主流にはなっていない
 - 今後ライセンス料の状況が変わらないと、フリーソフトへの移行が起こる可能性もありえる

- 産官学でソフト開発機関を設立
 - 民間: 参加する企業が出資、技術者を出向
 - 研究機関: 理研等、受け皿となる組織を用意、研究者を出向
 - 大学: 研究者を兼務等で派遣、ポスドクを供給
 - 運営: 出資者で組織する運営
- 産業界主導で必要とする新規解析ソフトを開発
 - 開発は産業界からの要望のあるもののみ
 - 科研費やNEDOなど公的資金による開発にも応募
 - 継続的に開発を行う
 - 開発したソフトの改良、新規プラットフォームへの適用等を行う

既存アプリに対して

- 権利関係を託し、クリーンにできるものに関しては、出資企業のメンバーの承認を得て、対応する
 - 理研で主導したV-CAD, ISLiMの他、大学や研究機関で開発した有益なソフトで民間出資者からの要望のある物を対象
- フリーソフトも同様、民間出資者の要望に応じて対象とする

対応案詳細 - 全体概念図 -



第二案：実用アプリ高速化コンソーシアム構想

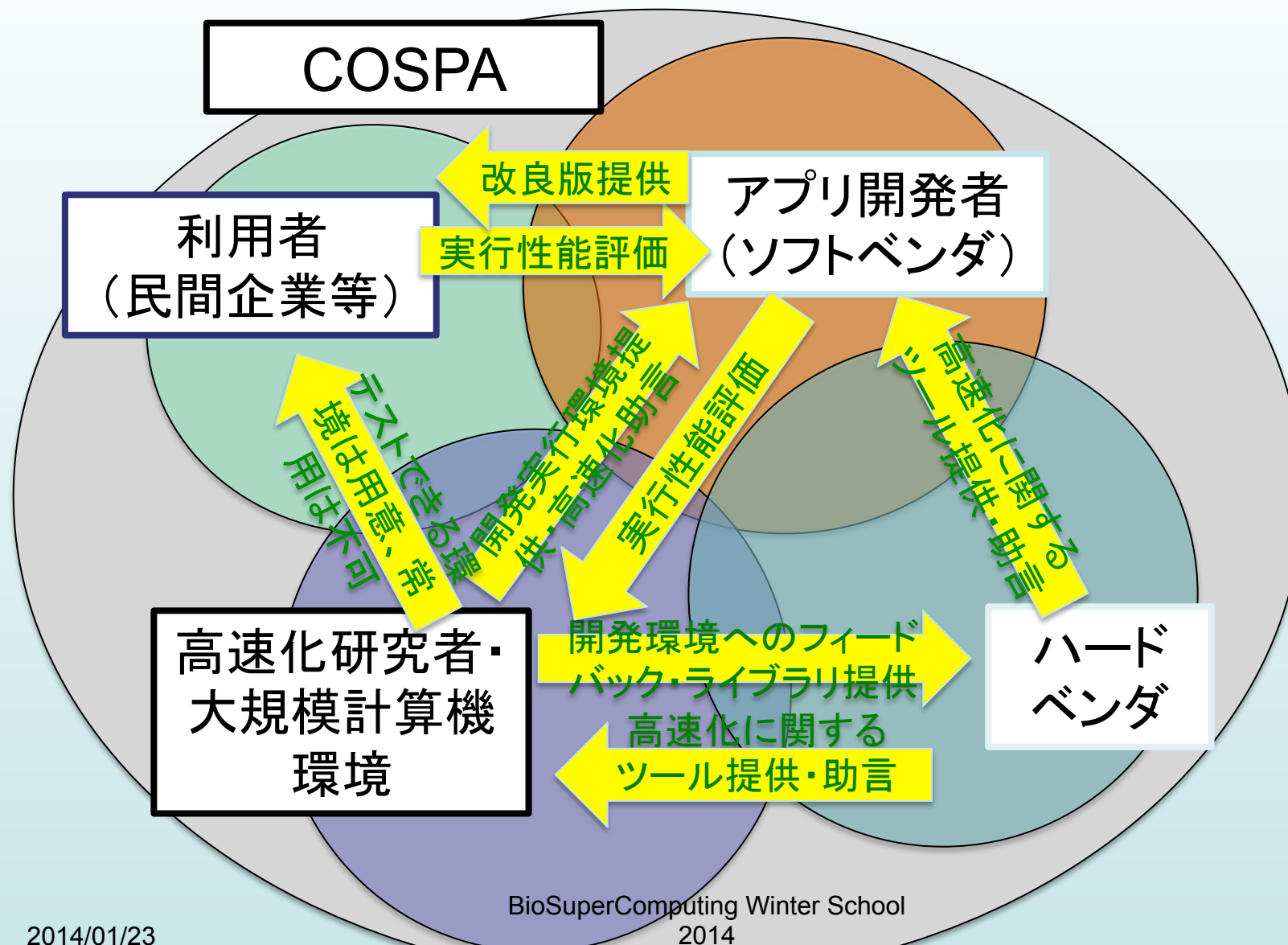
COSPA: COnsortium for Speeding up Practical Applications

- ・ 背景：マルチコア、GPGPU、超並列など、コンピュータの高速化と発展は、それに応じて対応するアプリの開発なくしては、実用的に使っているアプリの高速化に結びつかない。しかし、これらのハードに対応しようとすると、従来のMPIによる書き換えだけではなく、アルゴリズムレベルで見直しをする必要が出てきている。
- ・ 趣旨：このような背景のもと、1)アプリを実際に日常的な道具として使っている利用者、2)アプリの開発者、3)高速化の研究を行っていたり、大規模計算環境を持つ研究機関や大学、4)計算機ハードウェアや開発環境を提供しているベンダの4者が協力して、実用アプリの高速化・大規模化を実現する

従来のアプローチとの違い

- 実際に使われているアプリを対象とするため、高速化したのに使われないということはおこらない(利用者側からの要請で開発を行う)
- それぞれの強みを合わせることで、単独で行う高速化に比べて効果大きい
- 高速化に関するソフト開発で、官学は研究結果、ノウハウの提供を行うが、開発自体はソフトベンダーが行い、著作権はソフトベンダーが保持することを原則とする(個別の共同研究はありえる)

実用アプリ高速化コンソーシアム構想



将来の対象分野

- シミュレーションアプリ分野
 - 各参加機関・企業が戦略的かつ長期的に必要なとする分野
 - 先進的またはニッチ分野で、商用ソフトが参入していない、解析できない、または十分に解析技術が確立されていない分野
 - 具体例
 - 燃料電池(マルチスケール流体解析ほか)
 - 電磁場解析(電気自動車)
 - 材料開発、バイオ・ナノ・材料開発
 - バイオ・メカニクス
 - 気象・気候、防災
 - 社会科学
 - :
- アプリ利用技術分野
 - 各参加機関・企業での利用にとどまらず、適用分野が広い、またはシミュレーションアプリ開発上必須の要素技術
 - 具体例
 - アプリ最適化技術
 - 数学ライブラリ
 - マルチスケール・マルチフィジクス解析技術
 - :

海外で参考となる事例：仏)CERFACS

設立の経緯

- － 1980年代初めに設立された、官民出資の計算科学分野の研究開発・人材育成を行なう組織
 - － 産業界・官側の利用機関の研究開発を加速する事を期待されている
 - － 仏)ツールーズに設置されている
- 株主構成と関連アカデミア機関(☆印が設立当初の株主。出資金の情報無し、下線は公的機関)
 - 株主* : CNES ☆(21%) , EADS France ☆ (9%) , EDF (21%) , Météo-France ☆(21%) , ONERA(9%) , SAFRAN(9%) , TOTAL (9%)
 - * : CNES (the French Space Agency), EADS France(European Aeronautic and Defence Space Company), EDF (Electricité de France), Météo-France (the French meteorological service), ONERA(the French Aerospace Lab), SAFRAN(an international high-technology group), TOTAL (a multinational energy compagn)
 - － 仏の主要なHPC関連のアカデミア機関も連携
- 予算・業務内容
 - － 年間予算：約 8Mユーロ
 - 研究開発・業務テーマは株主により構成される委員会で決定 (カッコ内は予算比率)
 - ・ 各株主が共通に関心を持つ長期的なテーマ(5-10年レンジ) (40%)
 - ・ 各社が個別に費用を払って進めるプロジェクトテーマ (30%)
 - ・ 国・EU予算によるテーマ (20%)
 - 研究開発分野
 - ・ HPC基盤分野：並列処理(数学ライブラリ、ソルバー、ツール等)
 - ・ アプリケーション分野：(気象・気候、CFD、電磁気・音響、航空宇宙)
 - ・ 参加機関・企業向けの人材育成・教育も実施
- 人材
 - － 約150 名、内 研究者が 130名。半分が正規職員で残りは一時的な職員(アカデミア・企業から)
- その他補足
 - － 日本との企業構成の違い：仏は国内の企業が寡占化されており、CERFACSの場合も企業間の競合が起きにくい
 - 失敗例からの教訓：同様な組織がカナダで設立され、CERFACSも一時協力したが5年で頓挫。主な理由は官主導で(95%がパブリックセクターからの出資、5%が企業)、専任職員がいなかった為(大学の先生が一時的にいただけ)。

今後のBioSupercomputingの発展のために

“自分の研究や計算が社会にどう貢献できるか“

社会に役立つものだけが、支持され、援助される



1st Joint Workshop on Computational Science
7-8 Jul 2008



4th Biosupercomputing Symposium
2012.12.3-5 Tokyo International Forum

4th Biosupercomputing Symposium in 2012

Workshop in 2008

2014/01/23

Workshop in 2009



BioSuperComputing Winter School
2014

