科	学	技	術	計	算	分	科	会		選	出

科学技術計算分科会 2013 年度会合 より

エクサスケールで顕在化する Power Wall 問題 ～現状と今後の打開策～

近藤　正章

(電気通信大学)

サイエンティフィック・システム研究会　　2013 年 10 月 23 日
科学技術計算分科会 2013 年度会合

エクサスケールで顕在化する Power Wall 問題
～現状と今後の打開策～

近藤 正章
電気通信大学大学院情報システム学研究科

［アブストラクト］
エクサスケール HPC システムでは、消費電力がその設計や実効性能を制約する最大の要因の一つと考えられており、各構成要素の省電力化のみならず、システ ムレベルでの電力制御技術の開発が重要な課題となっている。本講演では、プロセッサやメモリ、ネットワークデバイス等の省電力化に関する要素技術動向について述べるとともに、現在取り組んでいる電力制約下において電力配分を最適化するための電力マネージメントフレームワークの研究について紹介する。

［キーワード］
エクサスケールシステム、ハイパフォーマンスコンピューティング、Power Wall、
省電力化、電力マネージメント

１．はじめに

エクサスケール高性能計算システムの開発においては、消費電力が最も重要な制約条件の 1 つであると認識されている。2013 年 10 月現在において Top500 リスト中 3 位にランクされるスーパーコンピュータの BlueGene/Q は、8MW 程度の消費電力で 17 ペタフロップスの実効性能を達成している。これに対し、 2020 年前後の稼働を目指すエクサスケール高性能計算機システムでは､20～30MW と現在の電力の 3 倍程度の消費電力で BlueGene/Q の 50 倍超の処理速度向上が要求されることになる。現状のトレンドを外挿すると、(比較的楽観的に見積もったとしても）実効で 1 エクサフロップスを達成するためには 50MW 近い消費電力が必要となる。そのため、消費電力当りの性能(MFlops/Watt)を向上させることが急務であり、この課題の解決無くしてはエクサスケール・スーパーコンピュータは実現し得ない。

この電力効率の改善に向けては、DVFS、Power-gating、不揮発性メモリなど、既存技術の活用をさらに進めるのはもちろんのこと、3 次元積層技術、FinFET、FDSOI、SOTB、Hybrid Memory Cube、シリコン・フォトニクスといった最新の電力効率に優れるデバイスの利用を

検討することが必要である。また、ハードウェア、およびソフトウェアからの電力マネージメント技術も重要になると考えられる。今後、これらの各種要素技術のコストや技術成熟度を検証しつつ、システムを開発を進めていくことが求められる。

２．電力マネージメントフレームワークの研究開発

上述のように、電力マネージメント技術はエクサスケールシステムの実現に向け、重要な技術の一つである。ポストペタスケール時代のアプリケーションは、超大規模システムにスケールさせる上で、システムへの要求は多様化すると予想されるため、供給電力や熱設計電力制約の中でハードウェア資源を投入し、運用時のピーク消費電力が制約を超えないことを保証する従来の設計思想では、アプリケーションを今後の大規模システムに対してスケールさせることは難しい。そこで、今後の HPC システムのあるべき姿として、従来のように利用可能な全ハードウェア資源を使い切るのではなく、ピーク消費電力が電力制約を超過することを積極的に許し(Over-provisioning)、ハードウェアが持つ電力性能ノブを適応的に制御することで限られた電力資源を計算・記憶・通信という要素に適応的に配分しつつ、実効電力を制約以下に抑えるシステム形態を提唱し、そのための電力マネージメントフレームワークの研究開発を富士通株式会社、九州大学、東京大学と連携して実施している。

現在は、電力観測・制御のための各種ミドルウェアの開発や、使用可能電力に制約が存在する中で各アプリケーションの性能を最大化するための電力制御最適化手法を開発しており、初期結果として適応的な電力配分により 1.4～2 倍超の性能向上を達成できることがわかっている。

３．おわりに

エクサスケールシステムでは、消費電力がシステム設計や実効性能を制約する最大の要因であり、電力効率を意識したシステムデザインが特に重要である。今後、限られた電力資源を真に有効利用できるシステムの実現を目指し、ハードウェアのみならず、関連するシステムソフトウェアの開発を、国内だけではなく国際的にも連携して進めて行くことが重要であると考える。

エクサスケールで顕在化するPower Wall問題
～現状と今後の打開策～

電気通信 大学大学院情報システム学研究科
近藤 正章

エクサスケールへの壁

- エクサスケールシステムへの課題
  - 信頼性の壁ーエクサスケールHPLの実行には1週間近くを要する
  - 消費電力の壁ー10倍超の電力効率の向上が必要
  - 低B/F、小メモリ容量、・・・
- EXTREMETECHより
  - “Supercomputing director bets $2,000 that we won’t have exascale computing by 2020…One of the biggest problems standing in our way is power.”

[出展] http://www.extremetech.com/computing/155941-supercomputing-director-bets-2000-that-we-wont-have-exascale-computing-by-2020

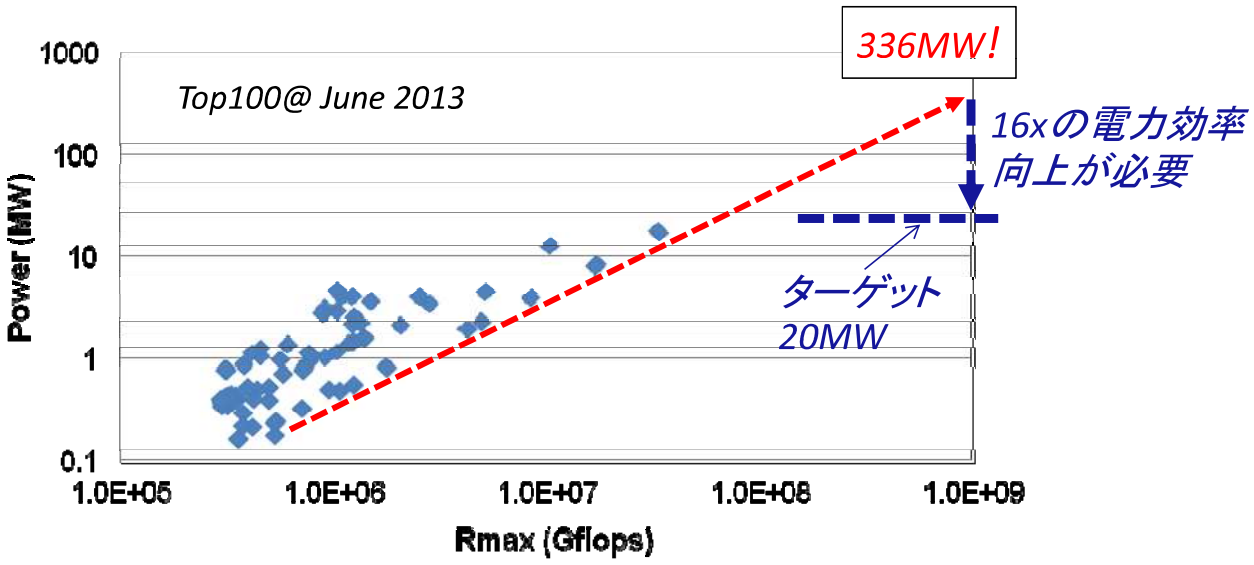
Power Wall問題を解決しない限りエクサスケール
システムの実現は難しい

講演内容

- 高性能計算機の消費電力トレンド
- エクサスケールに向けたPower Wallの検証
- Power Wall打開に向けた要素技術
- 電力マネージメントフレームワークの研究紹介

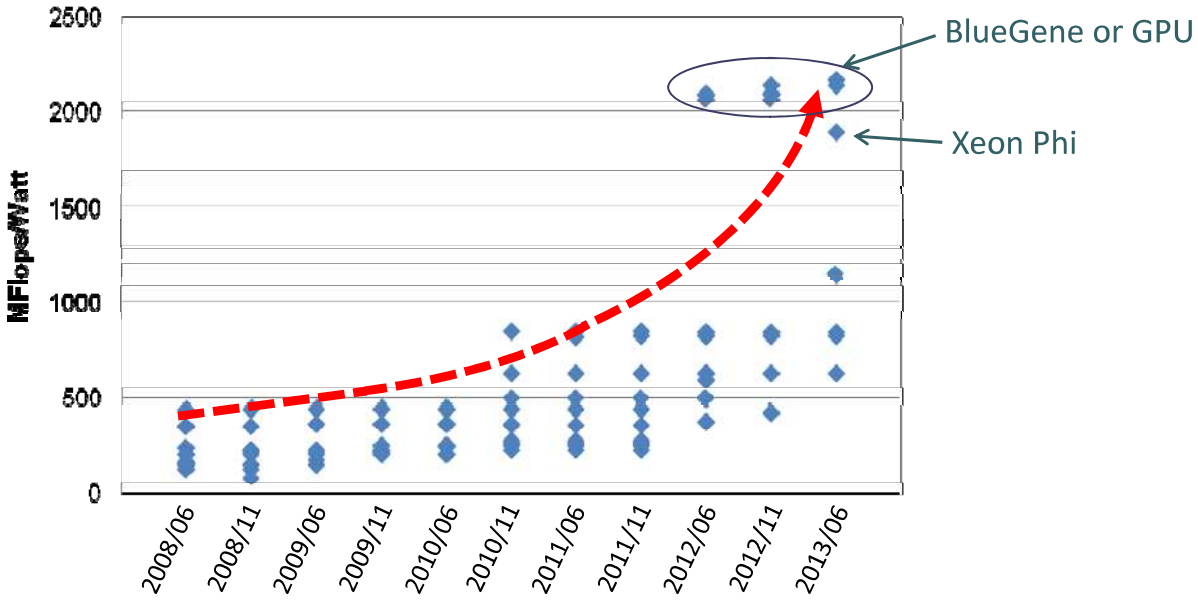
現在のスパコンの電力効率

- Top100スパコンのLinpack性能と消費電力
  - Top100中で最も電力効率が良いスパコン: 2972MFlops/Watt
- 20MWでExaFlops実現には16倍の電力効率向上が必要



システムレベルの電力あたり性能(MFlops/Watt)

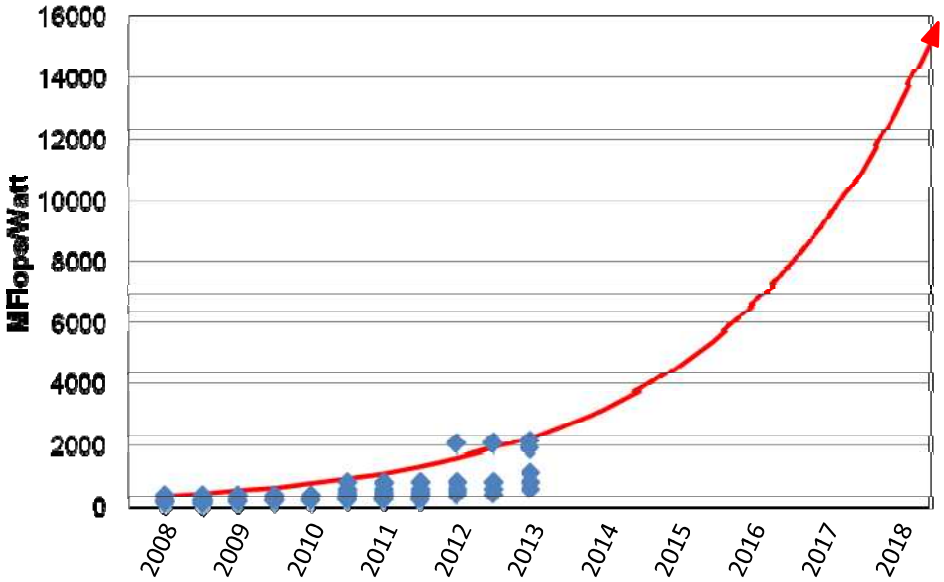
- Top10スパコンのMFlops/Watt
  - 2年(1プロセス世代)でおよそ2倍の向上
  - 今後は電力効率向上ペースは鈍化すると予想されている



システムレベルの電力あたり性能の将来トレンド

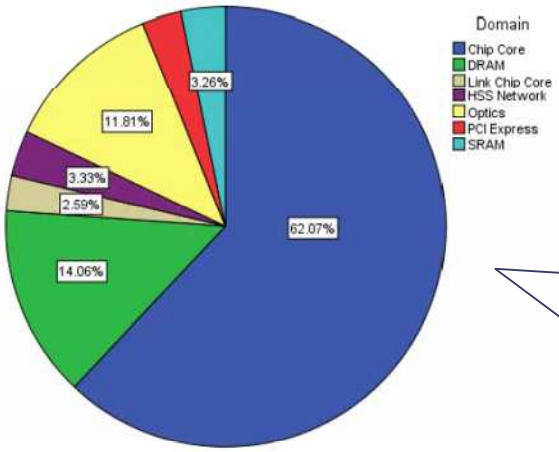
- 2年で2倍のFlops/Watt向上が続くとしても・・・
  - 2018年では10～20GFlops/Watt → エクサシステムで50MW

→ 電力効率向上のためのさらなる技術開発が必須



現在のシステムの消費電力の内訳

- BlueGene/Qの消費電力の内訳



[出展] S. Wallace, et. al. , "Measuring Power Consumption on IBM Blue Gene/Q". Proc. 9th HPPAC, 2013.

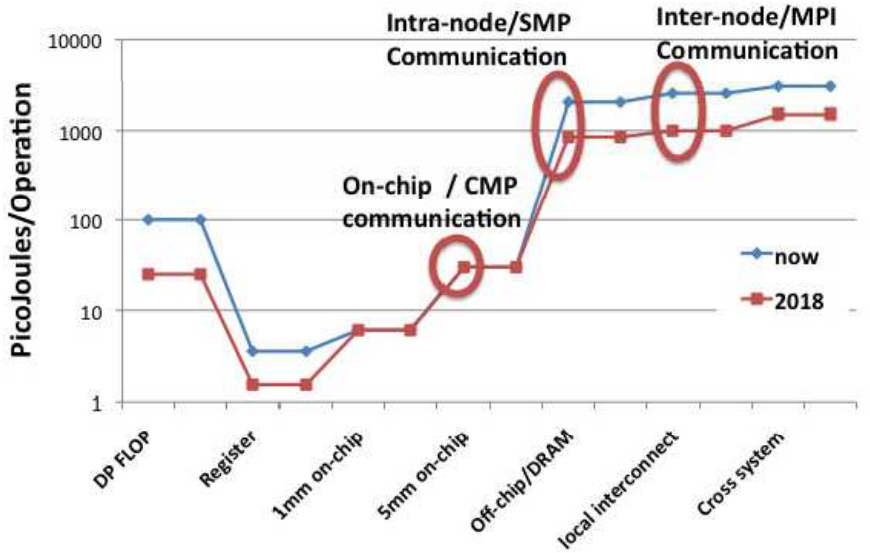
- BGQのenvironment database を通じて取得
- プロセッサ+SRAMの電力: 65%
- DRAM電力：約14%
- ネットワーク電力：約21%

Fig. 10. Pie chart showing relative percentages of total power usage consumed by each of the 7 power domains. Intense network activity largely contributing to optics percentage.

- BlueGeneでもプロセッサ・コア以外の消費電力は大きい

データ移動の電力コスト

- オフチップデータ移動の電力コストは計算よりも高い
  - 今後はメモリやインターコネクトの電力削減の重要性が増加
  - データ移動を抑えるソフトウェア技術(局所性の活用)も重要に



[出展]: Rick Stevens, et. al. "Scientific Grand Challenges: Architectures and Technology for Extreme Scale Computing", DoE, Dec. 2009.

講演内容

- 高性能計算機の消費電力トレンド
- エクサスケールに向けたPower Wallの検証
- Power Wall打開に向けた要素技術
- 電力マネージメントフレームワークの研究紹介

エクサに向けたPower Wallの検証

- エクサシステムに向けPower Wallは大きな課題
- システムの構成要素毎に性能・消費電力を検証
  - プロセッサ(浮動小数点演算)
  - メモリ(DRAM)
  - 通信(インターコネクト)
- 以下の資料を基に最新動向を踏まえてアップデート
  - ITRS roadmap, http://www.itrs.net/
  - P. Kogge, et. al., “ExaScale Computing Study, Technology Challenges in Achieving Exascale Systems”, IPTO Technical report TR-2008-13, DARPA Sep. 2008.
  - R. Stevens, et. al. “Scientific Grand Challenges: Architectures and Technology for Extreme Scale Computing”, Technical report, ASCR Scientific Grand Challenges Workshop Series, Dec. 2009.
  - 石川他, “計算科学研究ロードマップ白書”, 2012年3月.

エクサシステムで想定されるプロセッサ構成

[出展]「計算科学研究ロードマップ白書」より

- レイテンシコア (LC)
  - 高い周波数
  - 深いパイプライン構成
  - アウトオブオーダー実行
  - キャッシュ・プリフェッチ

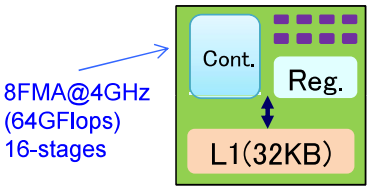
  → シングルスレッド性能に特化

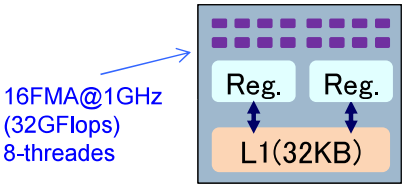
- スループットコア(TC)
  - 低い周波数
  - 浅いパイプライン
  - インオーダ実行
  - マルチスレッドサポート

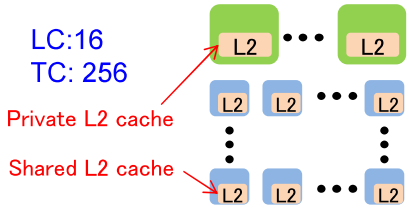
  → 電力性能に特化

- ヘテロ構成
  - LCとTCの融合によるヘテロ構成 (On-chip あるいは Off-chip)
  - プログラミングの複雑化

  →シングル＆マルチスレッド性能の両者に特化







	# cores	FLOPS	Clock speed	LLC
レイテンシコアのみで構成	32	2TFLOPS	4GHz	128MB
スループットコアのみで構成	512	16TFLOPS	1GHz	128MB
ヘテロ構成 (面積比 LC:TC = 1:1)	16L+256T	9TFLOPS	4GHz/1GHz	128MB
(参考) K-computer (58W/CPU)	8	128GFLOPS	2GHz	6MB

仮定：各プロセッサチップの電力は50-200W程度と想定

プロセッサの消費電力

- プロセッサの消費電力効率(GFlops/Watt)の予想
  - レイテンシコア(LC)、スループットコア(TC)、ヘテロ構成それぞれ

	2009年	2011年	2013年	2015年	2017年
プロセス世代	45nm	32nm	22nm	16nm	11nm
[参考]D-FLOP 電力 (pJ/FLOP)	20	10.5	6.0	3.5	2.0
LC電力効率 (GF/W)	0.5 - 2	1 - 4	2 - 16	4 - 30	8 - 40
TC電力効率 (GF/W)	2 - 10	4 - 20	8 - 40	16 - 80	32 - 160
ヘテロ構成電力効率 (GF/W)	—	2 - 10	5- 25	9 - 50	18 - 90

2018年に利用可能なプロセス世代でExaFlopsを実現するにはプロセッサのみで10-20MW程度が必要

メモリ(DRAM)の消費電力

- 各世代のメモリの比較

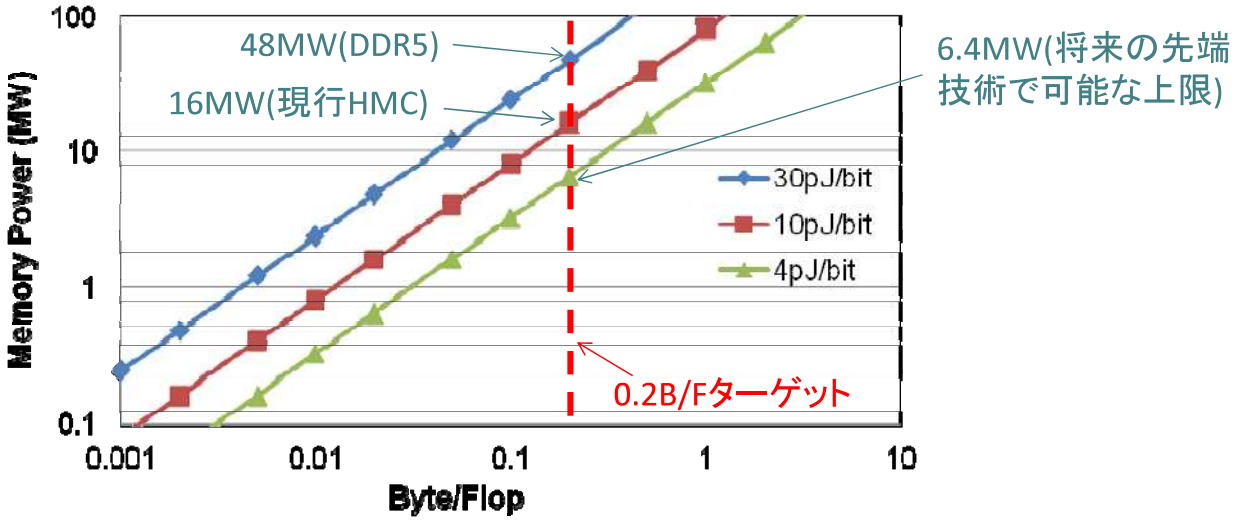
Technology	VDD	IDD	BW GB/s	Power (W)	mw/GB/s	pJ/bit	real pJ/bit
SDRAM PC133 1GB	3.3	1.50	1.06	4.96	4664.97	583.12	762
DDR-333 1GB	2.5	2.19	2.66	5.48	2057.06	257.13	245
DDRII-667 2GB	1.8	2.88	5.34	5.18	971.51	121.44	139
DDR3-1333 2GB	1.5	3.68	10.66	5.52	517.63	64.70	52
DDR4-2667 4GB	1.2	5.50	21.34	6.60	309.34	38.67	39
HMC 4DRAM w/ Logic	1.2	9.23	128.00	11.08	86.53	10.82	13.7

[出展]: J. T. Pawlowski, “Hybrid Memory Cube (HMC)”, Hot Chips23, Aug. 2011.

- 世代が進むにつれてDRAMの電力効率(pJ/bit)は改善
- DRAMモジュールの消費電力は増加傾向
  - 実装密度の向上、バンド幅の向上
- 従来トレンドを外挿すると2018年のDDRテクノロジ (DDR5)では30pJ/bitと予想されている[Stevens2009]

メモリ(DRAM)の消費電力

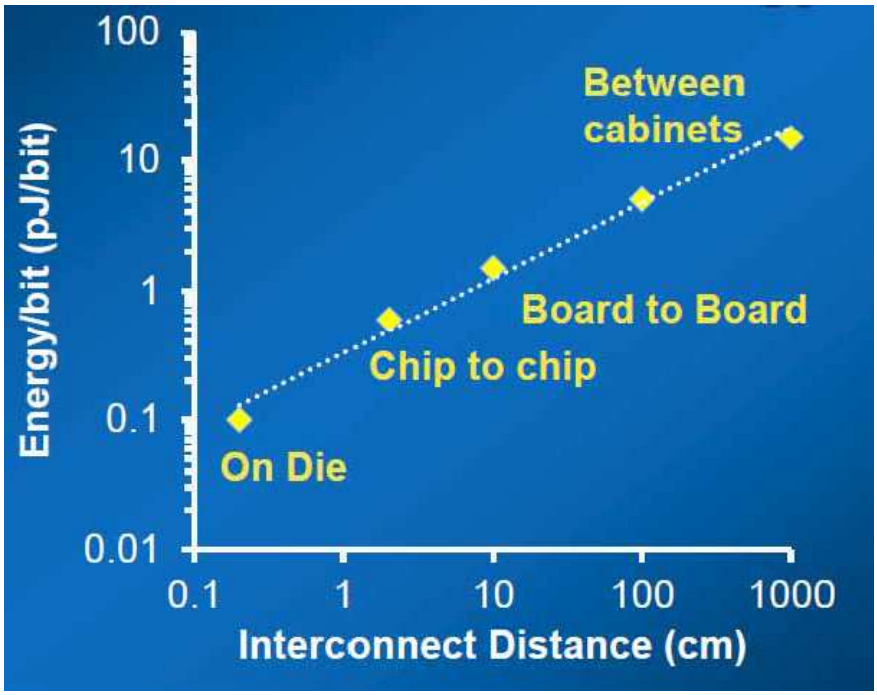
- 要求メモリバンド幅(Byte/Flop)に基づくDRAM電力の推定



- 30pJ/bitでは0.2B/Fの実現に48MWもの電力が必要
- より電力効率の良いメモリ技術の開発が必須
  - 先端テクノロジで7pJ/bit (最小で4pJ/bit) と予想[Stevens2009]

通信の消費電力

- 距離に応じたデータ移動に必要なエネルギー

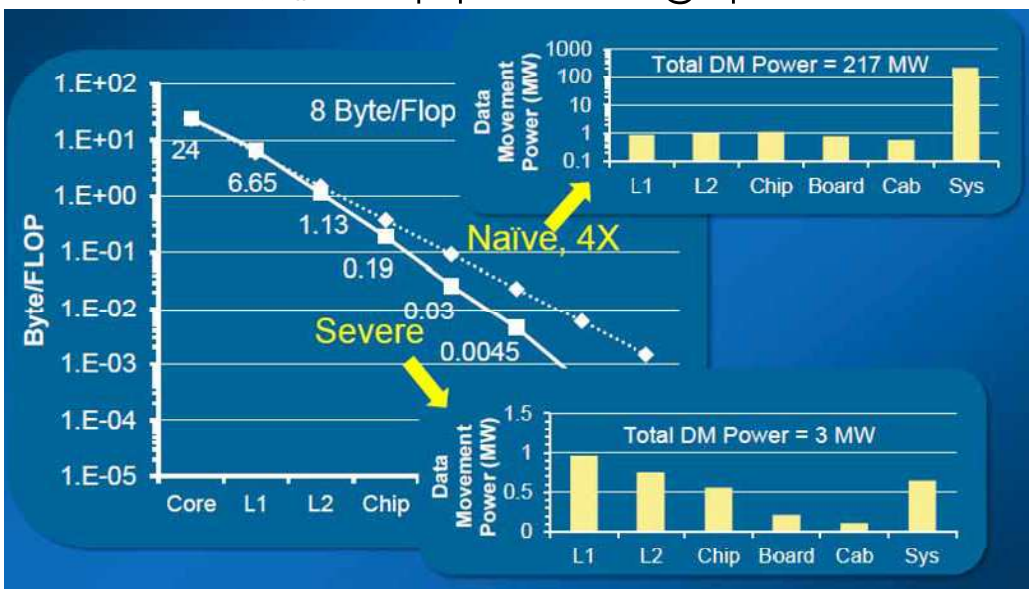


[出展] S. Borkar, “Exascale Computing – a fact or a fiction?”, IPDPS2013 Keynote , May 2013.

通信の消費電力

- エクサシステムにおけるデータ移動にかかる電力

1m以上の通信：40Gbps photonics links@10pJ/bit



[出展] S. Borkar, “Exascale Computing – a fact or a fiction?”, IPDPS2013 Keynote, May 2013.

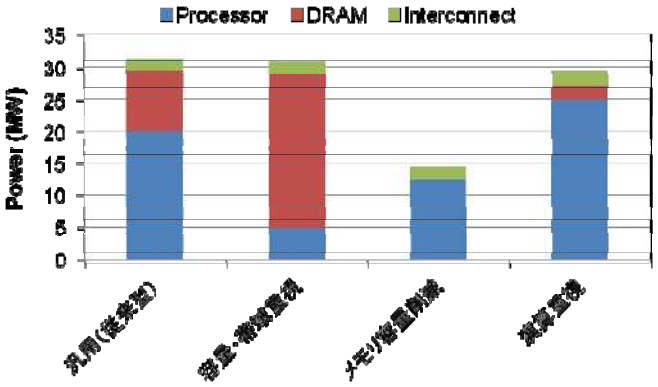
→ 遠距離の通信バンド幅を抑えた設計が必要

エクサスケールアーキテクチャへのマッピング

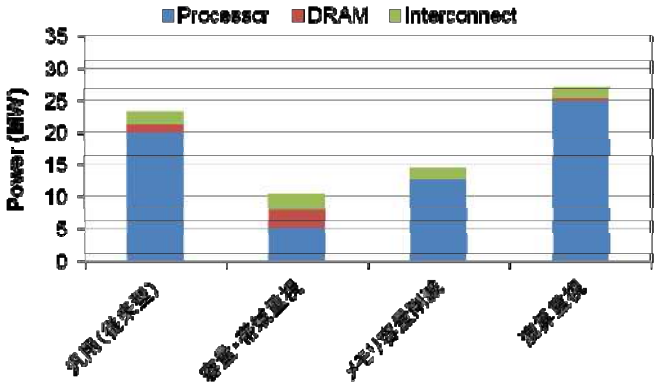
- 計算科学研究ロードマップ白書[石川2012]の各アーキテクチャへのマッピング

	総演算性能 PetaFLOPS	総メモリ帯域 PetaByte/s	総メモリ容量 PetaByte	Byte/Flop	利用コアの仮定
汎用(従来型)	200~400	20~40	20~40	0.1 程度	LC(20GF/W)
容量・帯域重視	50~100	50~100	50~100	1.0 程度	LC(20GF/W)
メモリ容量削減	500~1000	250~500	0.1~0.2	0.5 程度	ヘテロ(45GF/W)
演算重視	1000~2000	5~10	5~10	0.005 程度	TC(80GF/W)

DDR5 メモリ (30pJ/bit)



先端メモリ (4pJ/bit)



※Interconnectは2MWで一定と仮定

講演内容

- 高性能計算機の消費電力トレンド
- エクサスケールに向けたPower Wallの検証
- Power Wall打開に向けた要素技術
- 電力マネージメントフレームワークの研究紹介

電力効率を向上させる技術候補

- プロセッサ
  - (既存技術の延長) DVFS、Power-gating、Clock-gating
  - 3次元積層技術、FinFET、トライゲート、FDSOI、SOTB、・・・
  - Low (Near-Threshold) Voltage Computing
  - SIMD幅拡大、アクセラレータの効率的利用
- メモリ
  - 3次元積層技術(Hybrid Memory Cube、 Wide I/O、HBM)
  - 不揮発性メモリの利用
- インターコネクト
  - **シリコン・フォトニクス**
  - 動作モード制御(リンク幅/リンク速度のスケーリング)
- システムレベル電力マネージメント
  - Power-capping、 電力性能比を最適化するアルゴリズム/ライブラリ
  - 電力モニタリング／制御インタフェースの提供

Low (Near-Threshold) Voltage Computing

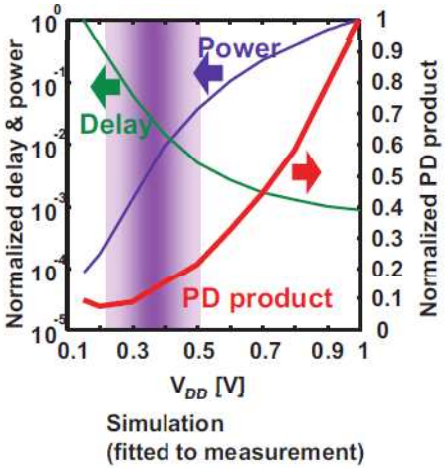
- 電源電圧のトレンドおよび

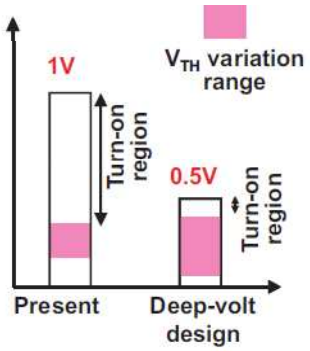
[出展] T. Sakurai, "Pitfalls in deep-volt logic design". ISSCC'11 Forum: Ultra-Low Voltage VLSIs for Energy-Efficient Systems, 2011.

非公開資料

- 低電圧動作の利点と課題

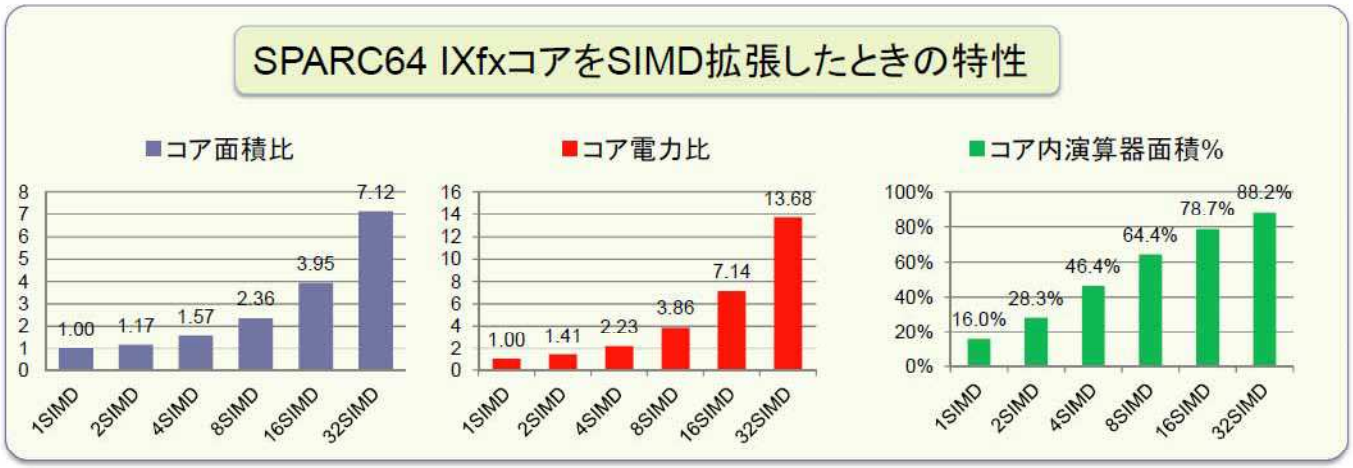
[出展] T. Sakurai, "Designing Ultra-Low Voltage logic". Proc. ISLPED'11, pp57-58, Aug. 2011.





プロセッサ・コア内のSIMD幅拡大

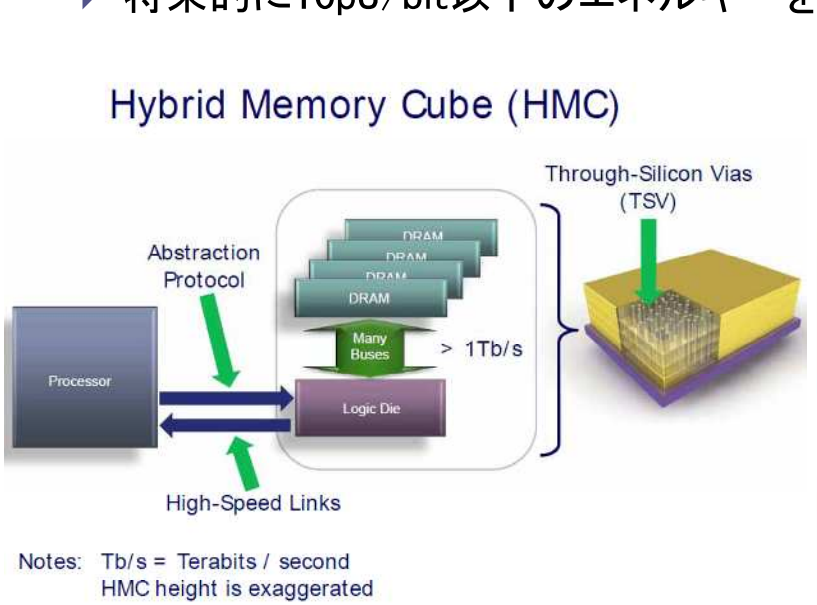
- SIMD幅を拡大することで電力性能効率が向上
- 多くのプロセッサがSIMD幅を増加させる傾向
  - e.g.) Intel SSE(128bit) → AVX(256bit) → AVX-512(512bit)
- 高い実効性能を出すためのソフトウェア環境が重要に



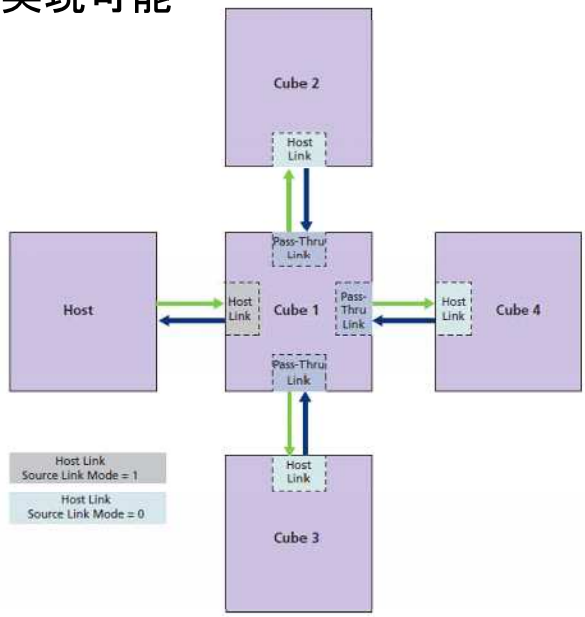
[出展] 追永, "FXシリーズの今後の取り組みについて", SS研HPCフォーラム2013, 2013年8月.

Hybrid Memory Cube (HMC)

- DRAMダイとロジックダイを3次元積層
- プロセッサ・HMC間、HMCモジュール間を高速なシリアルリンクで接続
- 将来的に10pJ/bit以下のエネルギーを実現可能



[出展]: J. T. Pawlowski, "Hybrid Memory Cube (HMC)", Hot Chips23, Aug. 2011.



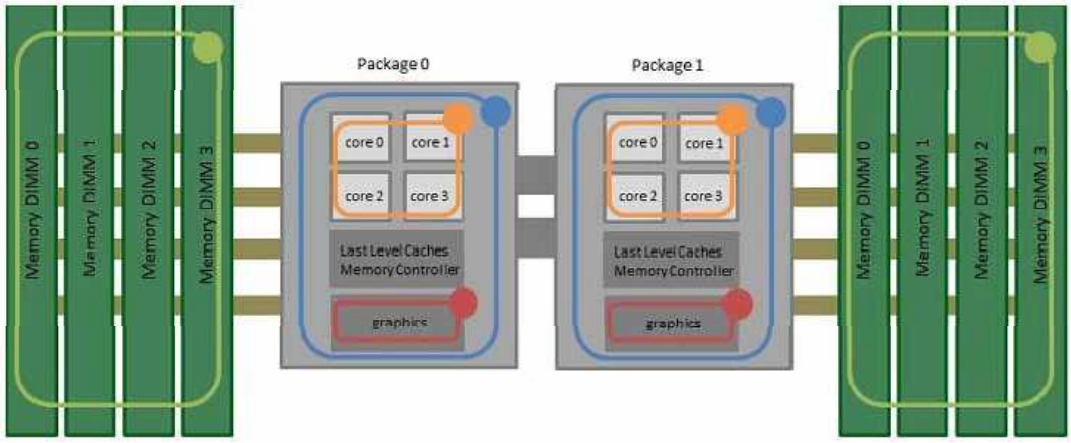
[出展] Hybrid Memory Cube Consortium, "Hybrid Memory Cube Specification 1.0", 2013.

電力マネージメント技術

- エクサスケールでは電力マネージメントが重要
  - ハードウェア／ソフトウェアの電力管理
  - 各構成要素、各システム階層でのきめ細かな電力制御
  - Power-cappingのもとでの Over-provisioningも有望
- 重要技術項目
  - 電力消費状況の(リアルタイム)モニタリング
    - 電力観測インタフェースを備えるシステムが普及しつつある
    - e.g.) BlueGene/P、BlueGene/Q、Intel Sandy Bridge (RAPL)
  - 電力性能比を最適化するアルゴリズム/ライブラリ
  - 電力制御インタフェースの標準化
  - 電力制御用*Knob*の適切なモデリングと最適化制御

電力観測・制御インタフェースの例: RAPL

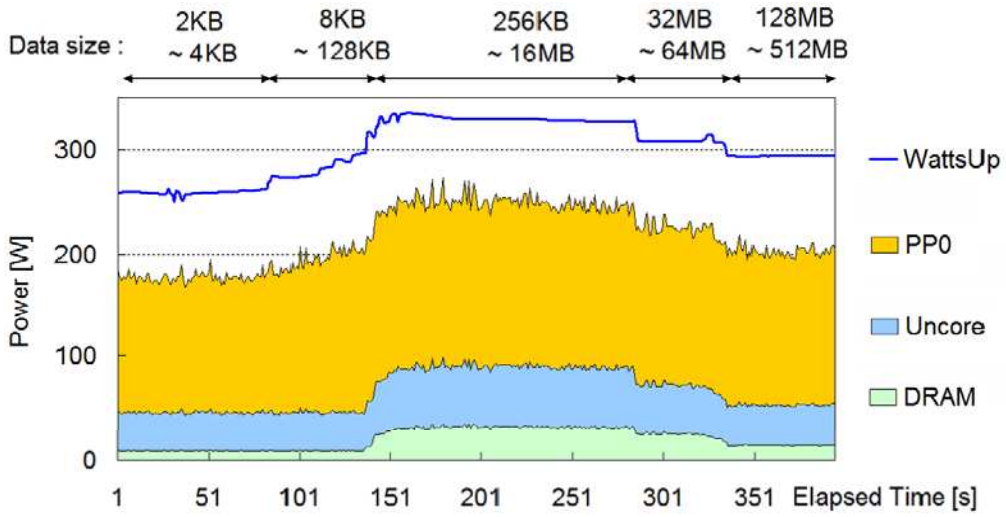
- RAPL (Running Average Power Limit)インタフェース
  - Intel Sandy Bridgeマイクロアーキテクチャより搭載
  - パフォーマンスカウンタや温度等の情報を基に消費電力の見積り・制御
  - MSRを介して消費電力の取得や電力上限設定が可能
  - ドメイン毎に電力を計測



[出展]: Intel Power Governor, http://software.intel.com/en-us/articles/intel-power-governor

RAPLによる電力モニタリングの例 [カオ2013]

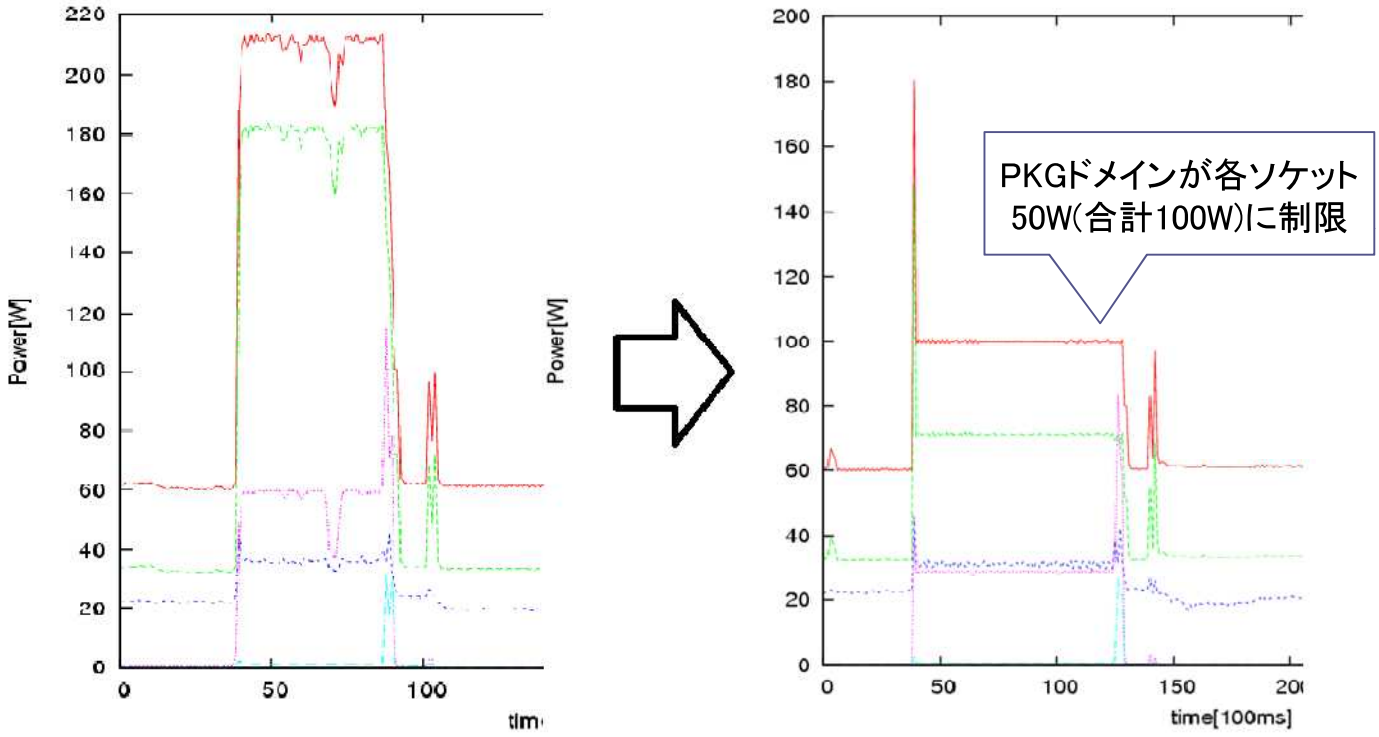
- ストリームアクセスプログラム
  - 16コア(MPI並列)、配列サイズ:2KB～2GB



- RAPLと外部電力計(WattsUp)の消費電力傾向は非常に良く一致
- メモリアクセス頻度の違いにより消費電力が大きく異なる
  - L2サイズ(256KB)以上:電力増加 → キャッシュアクセス増 + DRAMアクセス増
  - L3サイズ(20MB)以上:電力減少 → 単位時間あたりのアクセス頻度減少

消費電力制約の設定例:CPU

- パッケージ電力(PKG)を50Wに設定



講演内容

- 高性能計算機の消費電力トレンド
- エクサスケールに向けたPower Wallの検証
- Power Wall打開に向けた要素技術
- 電力マネージメントフレームワークの研究紹介

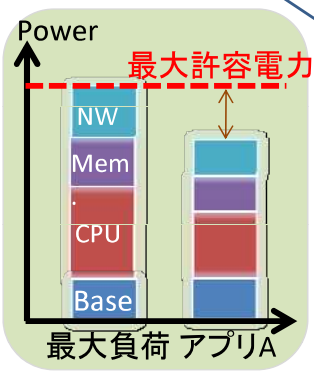
電力マネージメントフレームワークの研究紹介

- JST CREST「ポストペタスケールシステムのための電力マネージメントフレームワークの開発」
  - 共同研究機関：富士通、九州大学、東京大学、電気通信大学
- 研究背景
  - ポストペタ時代のシステムは消費電力が最大の設計制約
  - アプリケーションのシステムへの要求の多様化
    - 計算・記憶・通信の各要素への要求が異なる
    - 電力がシステム制約となる状況下では各要素へ投入するハードウェア資源は制限せざるを得ない

運用時のピーク電力が制約を超えないことを保証する
worst case設計ではシステムをスケールさせることは難しい

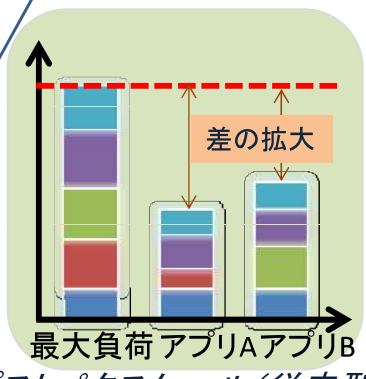
最大許容電力と実効消費電力

- 最大許容電力と実効電力の差の拡大
  - システムのヘテロジニアス構成化(特定アプリ専用・得意な構成要素の導入)
    - 設計制約が集積度ではなくなるため、HW資源を使い尽くすという設計からアプリにとって必要な資源を投入するという設計思想への転換
  - 局所性の利用(高性能化としての必須技術)
    - 一部要素への負荷集中の拡大
  - 省電力技術の積極的利用
    - DVFS・Power-Gatingなど普及、負荷比例電力成分の増加

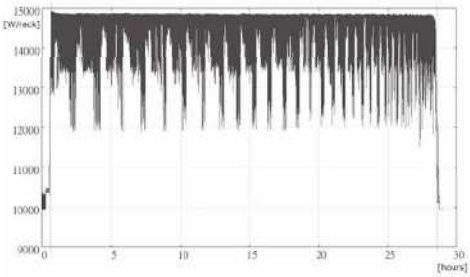


ペタスケール

・ヘテロ化
・局所性利用
・省電力技術



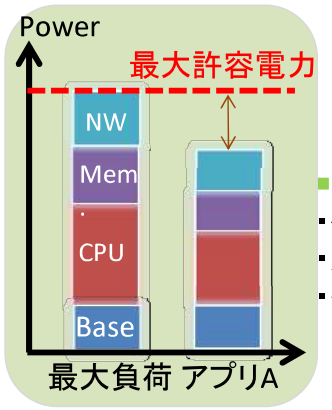
ポストペタスケール(従来型)



参考データ:京コンピュータにおけるLinpack実行時の電力 [Miyazaki2012]

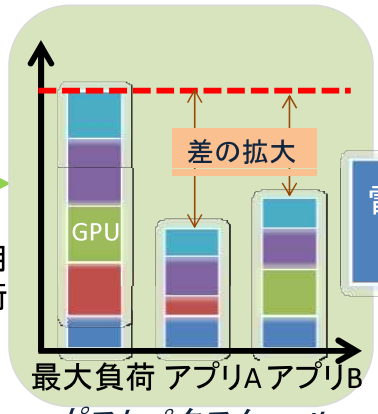
ポストペタスケールシステムのあるべき姿

- ハードウェア資源の有効利用から電力資源の有効利用へのパラダイムシフト



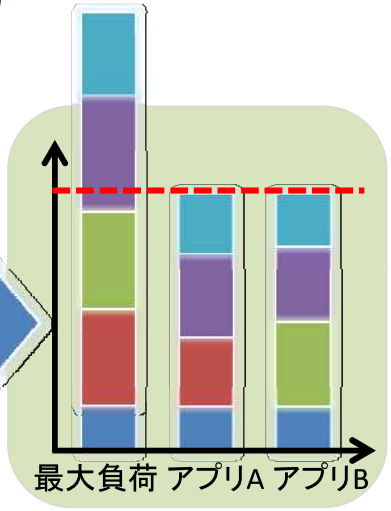
ペタスケール

・ヘテロ化
・局所性利用
・省電力技術



ポストペタスケール(従来型)

電力制約適応型

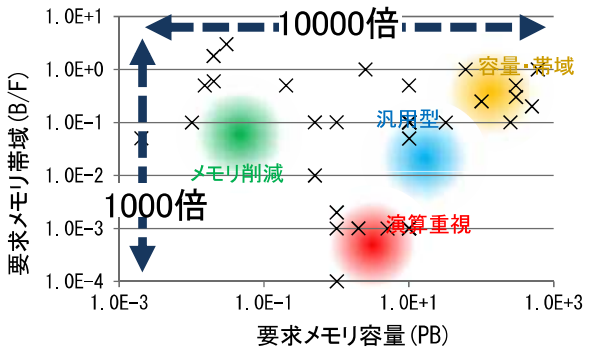


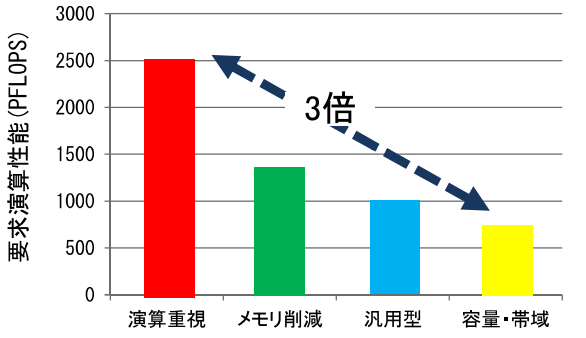
ポストペタスケール(電力制約適応型)

電力制約適応型システム

- 最大負荷時電力が電力制約を超過することを積極的に許容
- 電力性能ノブを自動制御することで実効電力を制約以下に抑制
- 電力資源を計算・記憶・通信へ適応的に配分することで実効性能向上へ

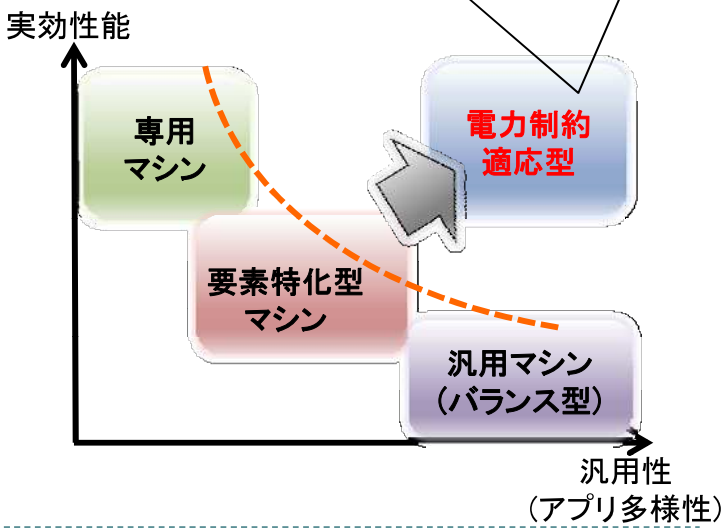
電力制約適応型システムのねらい





アプリケーションのシステム要求
（参考：今後のHPCI技術開発に関する報告書）

- 電力資源の適応的配分による実効性能向上
  - 演算効率10%の場合、10倍の性能向上の可能性あり！
- 多様なアプリに対応可能なシステムの実現
- ☹電力性能ノブ最適化の負担増

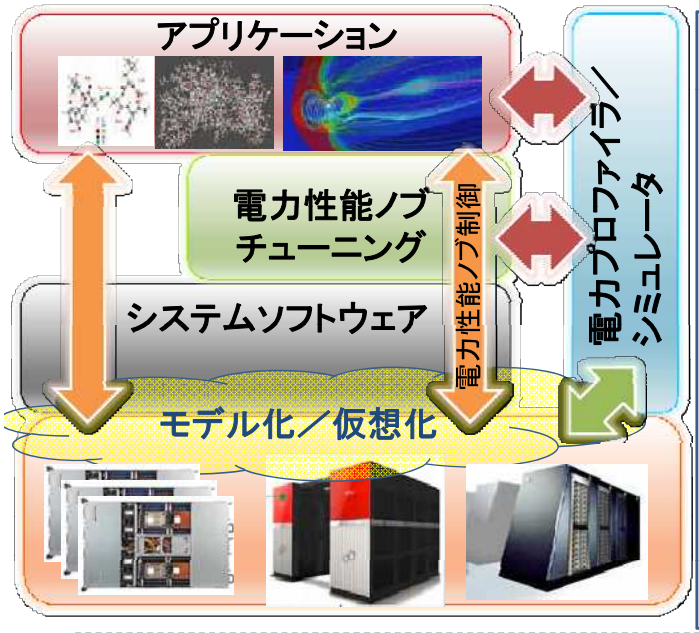


主要チャレンジと研究目的

電力制約適応型システムに向けた主要チャレンジ

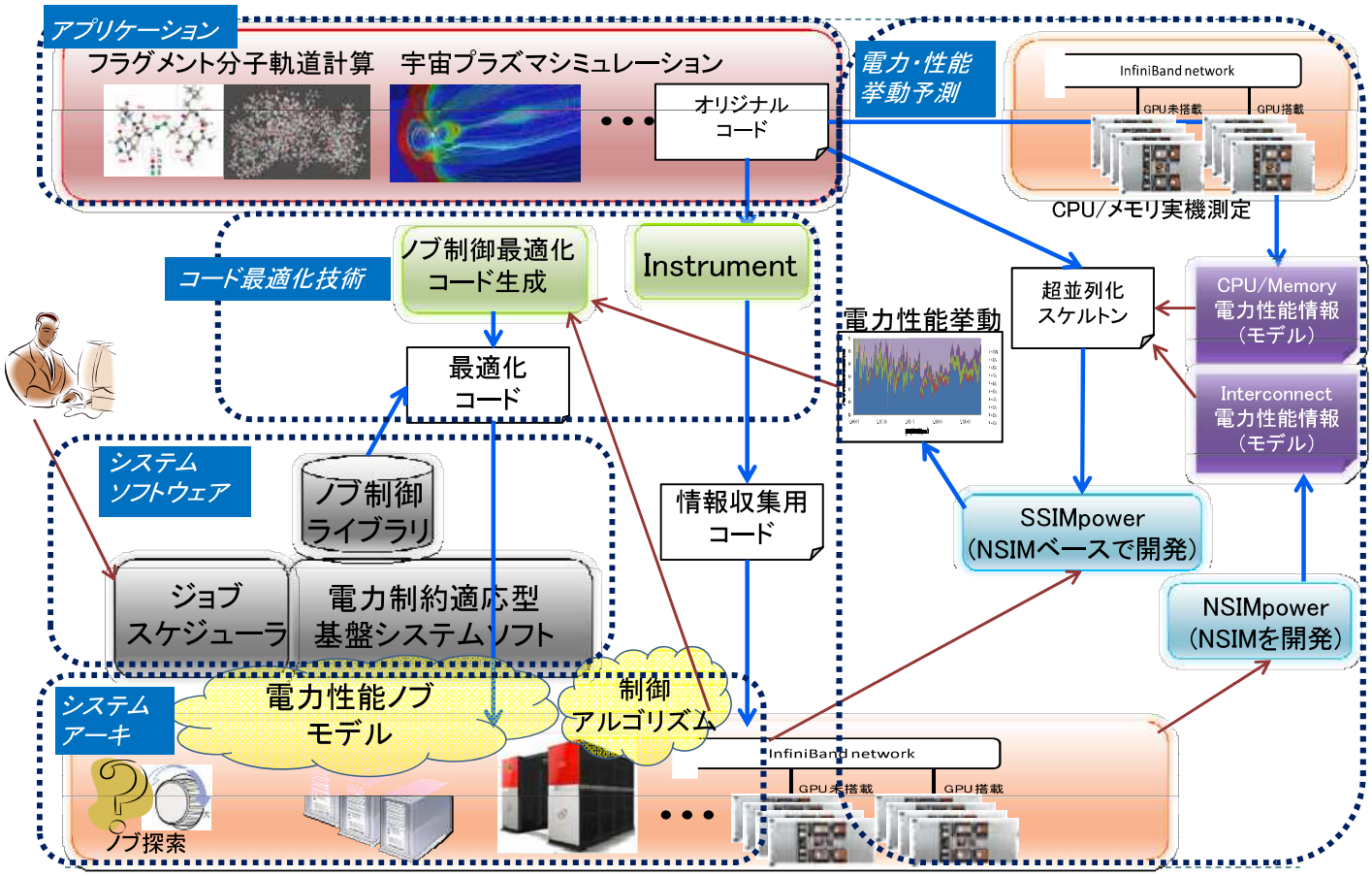
1. 電力性能ノブ制御最適化のユーザ負担の低減・隠蔽
2. 電力制約適応型システムの制御と電力性能ノブ仮想化

→ 性能／電力最適化をアプリケーション最適化として統合できるフレームワーク



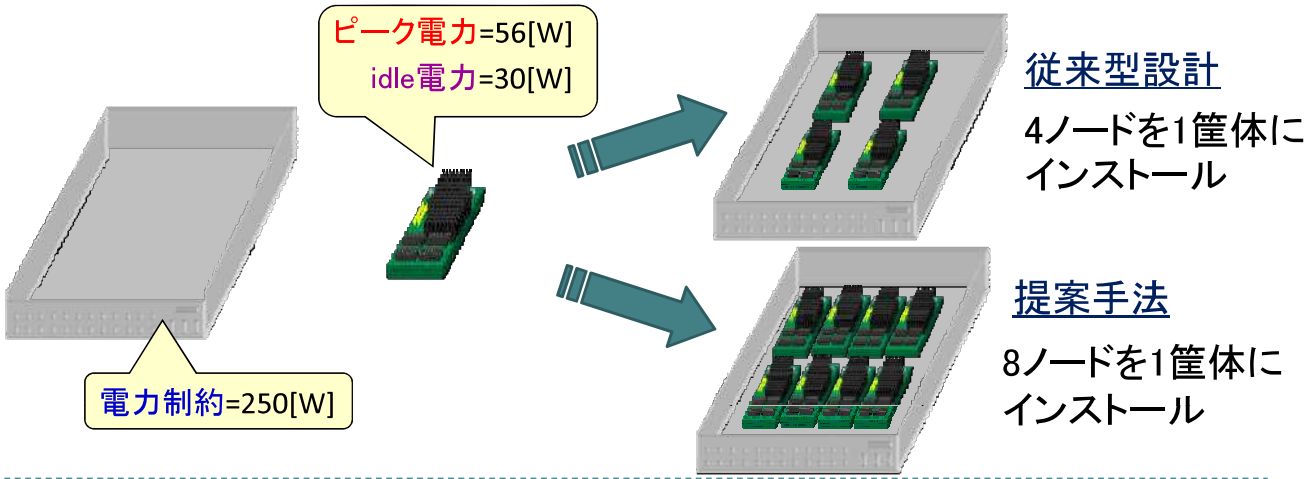
- **電力性能ノブ最適化による性能向上**
  - 自動・半自動での電力性能ノブ制御最適化
  - 大規模アプリの電力挙動予測
- **電力制約適応型システムの制御**
  - HWの詳細をアプリから隠蔽しつつノブの制御、電力資源の管理が可能なシステムソフトウェア
  - 動的変動に対応するジョブ管理技術
- **電力性能ノブ仮想化**
  - 新しいHWノブに対応し、その詳細を隠蔽可能な電力性能ノブモデリング
  - 空間・時間粒度を考慮した最適化アルゴリズム

電力マネージメントフレームワークの全体像



ノード数とCPU周波数の電力バランス最適化

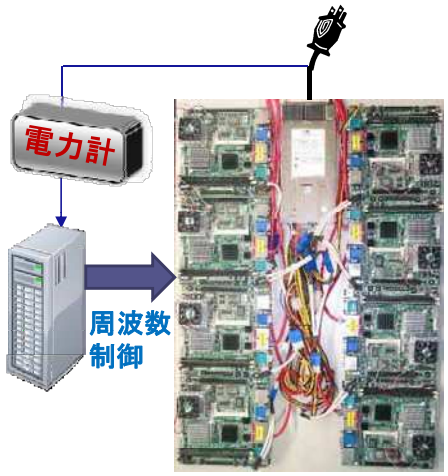
- 計算機クラスタ上での適応的電力制御 [kondo2007]
  - 電力制約適応型システムをコンセプトとするプロトタイプ
  - 目的：電力(発熱)制約内で電力資源を有効利用することで性能向上
  - 戦略：制約を超えるノードを1筐体にインストール、クラスタ構成(ノード数・CPU周波数)をアプリに応じて最適化



アプリに応じた最適化の方法

- 使用ノード数最適化
  - 1筐体を用いてノード数を変化させてプロファイリング(周波数は動的制御)
  - 最も高いMIPS値が得られるノード数の構成を採用
- 周波数の最適化
  - 実行時の電力(Peff.)をモニタリングしつつ電力制約内に収まるように周波数(電源電圧)を調整
  - アルゴリズム
    1. Peff. を一定周期でモニタリング
    2. 周波数を1段階上げる
    3. Peff.が電力制約を超えたら周波数を1段階下げる
    4. 電力制約を超過した場合、一定時間周波数を上げるのを抑制

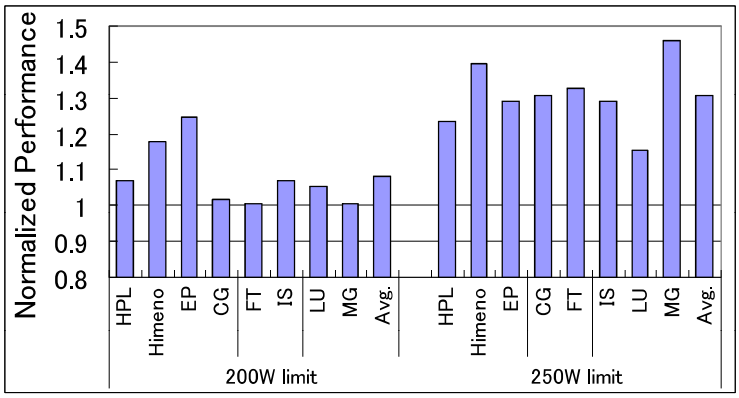
    (周波数は筐体内の全ノードで同一)



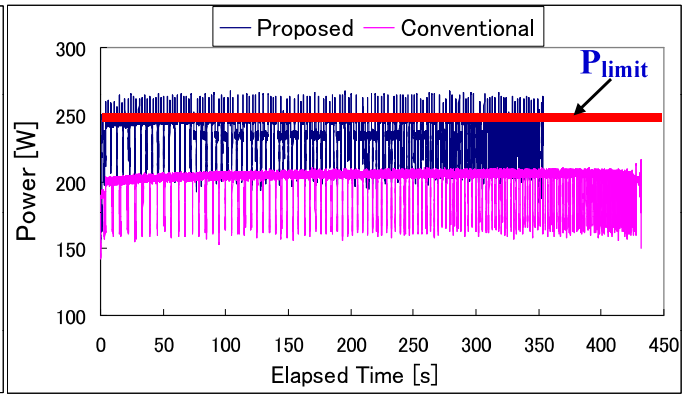
評価結果

- システム構成・仮定
  - CPU: Pentium M 760 (周波数2.0～ 0.8GHz)、メモリ: DDR SDRAM 1GB
  - 8ノード／筐体 (ピーク消費電力 450[W])
  - 電力制約:200[W]および250[W]

従来型システムに対する性能



消費電力(HPL)



従来型のシステムに比べ最大で1.5倍程度の性能向上

構成要素間の電力バジェット配分最適化

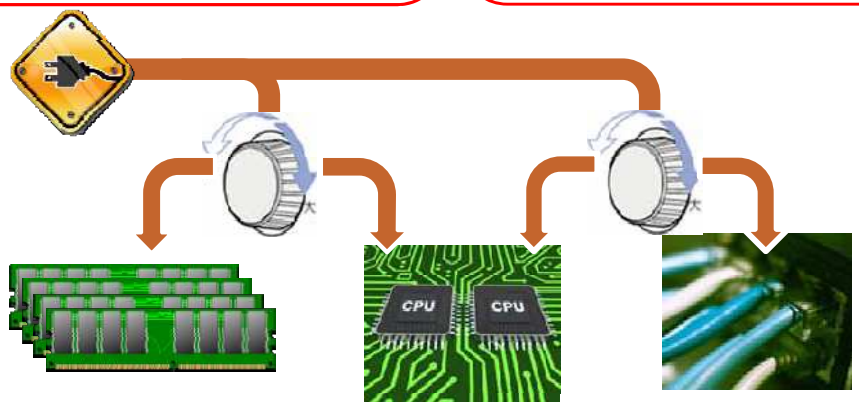
- 電力資源の適応的配分に向けた電力制御の事例

プロセッサ-DRAM間

- RAPLの電力制約機能を利用
- まずは単一ノードに着目
- 電力バジェット（総量）とその配分量を様々に変えた際の実行時間を計測

プロセッサ-ネットワーク間

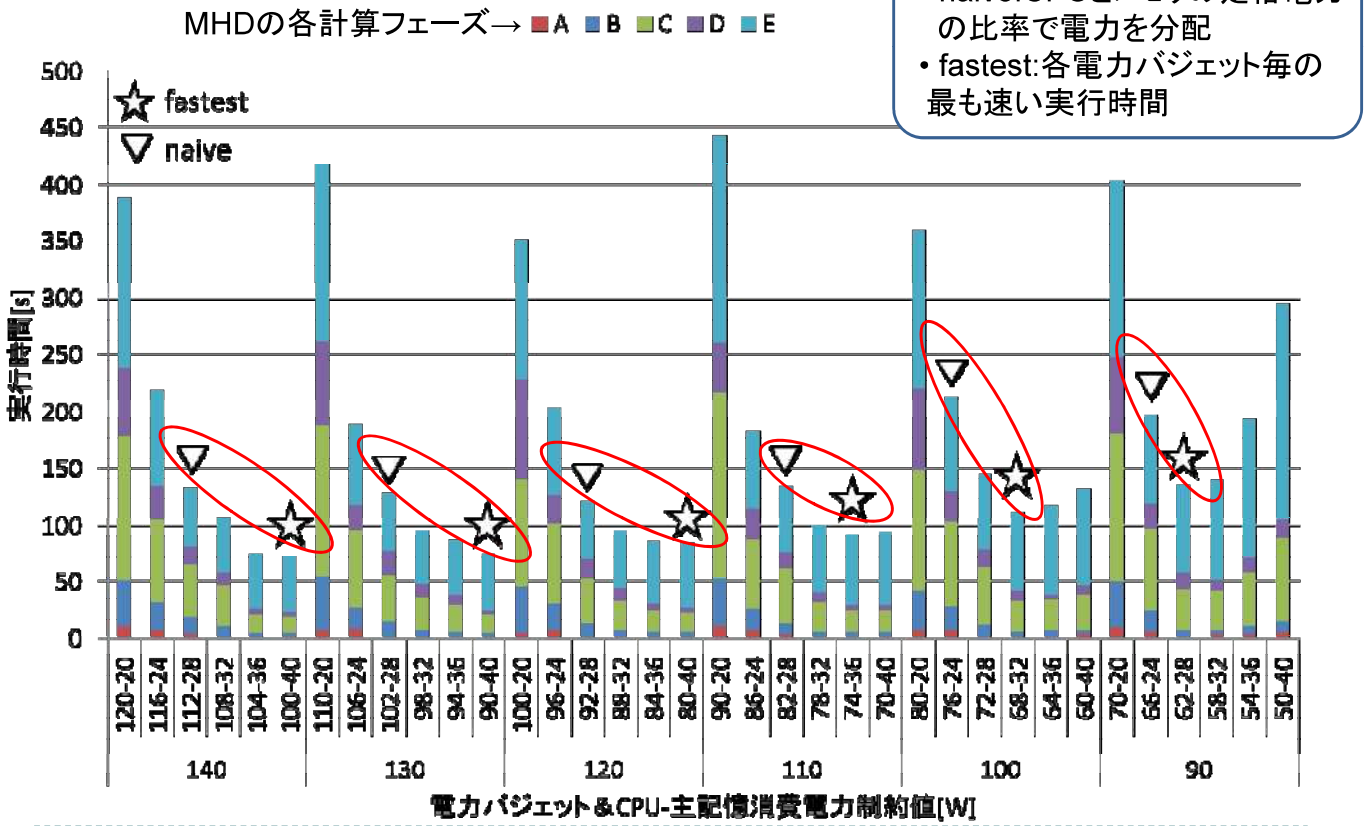
- EEE(Energy Efficient Ethernet)で生まれた余剰電力を各コアに配分
- 各ノードのコア周波数はクリティカリティに応じて電力を再分配
- プロファイリングにより最適化



DRAM・プロセッサ　　プロセッサ・ネットワーク

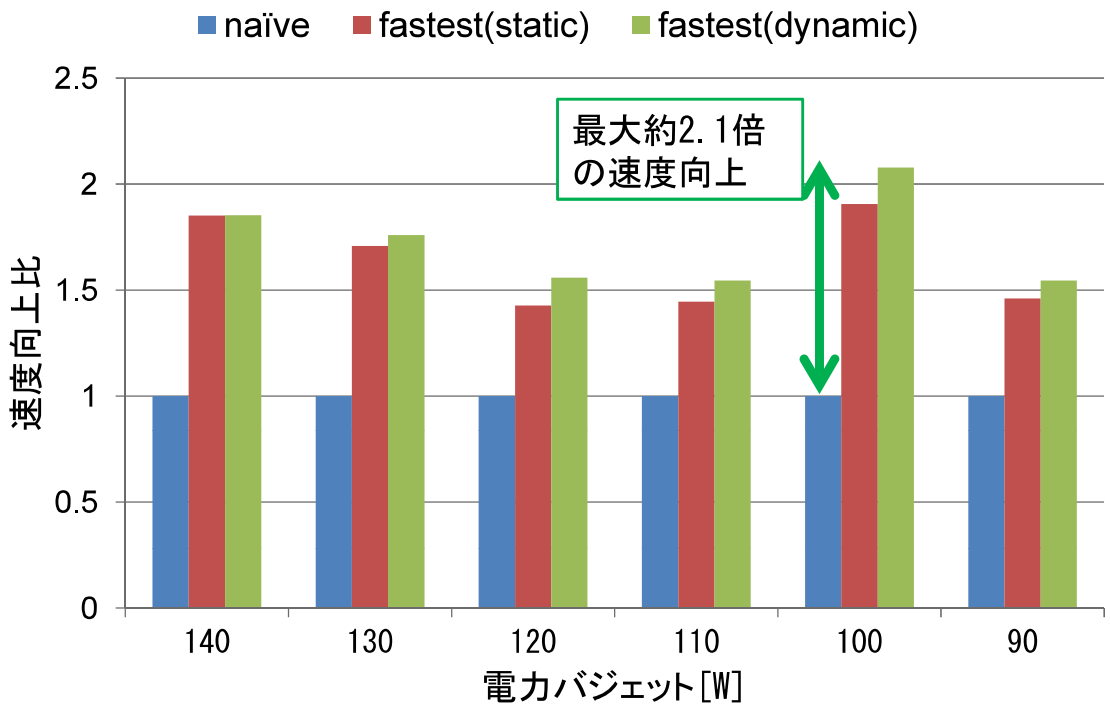
プロセッサ-DRAM間の電力配分実験結果(1) [吉田2013]

- naive:CPUとメモリの定格電力の比率で電力を分配
- fastest:各電力バジェット毎の最も速い実行時間



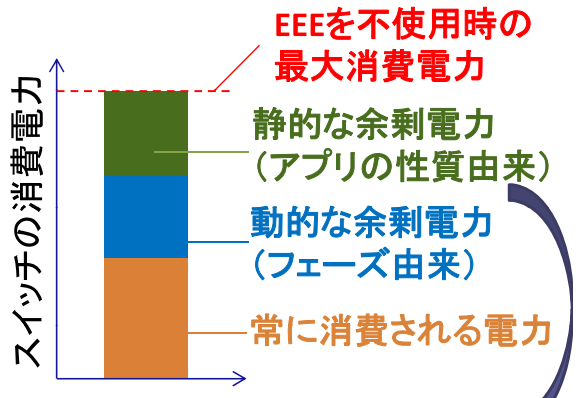
プロセッサ-DRAM間の電力配分実験結果(2) [吉田2013]

- naïveに対する電力配分最適時による速度向上

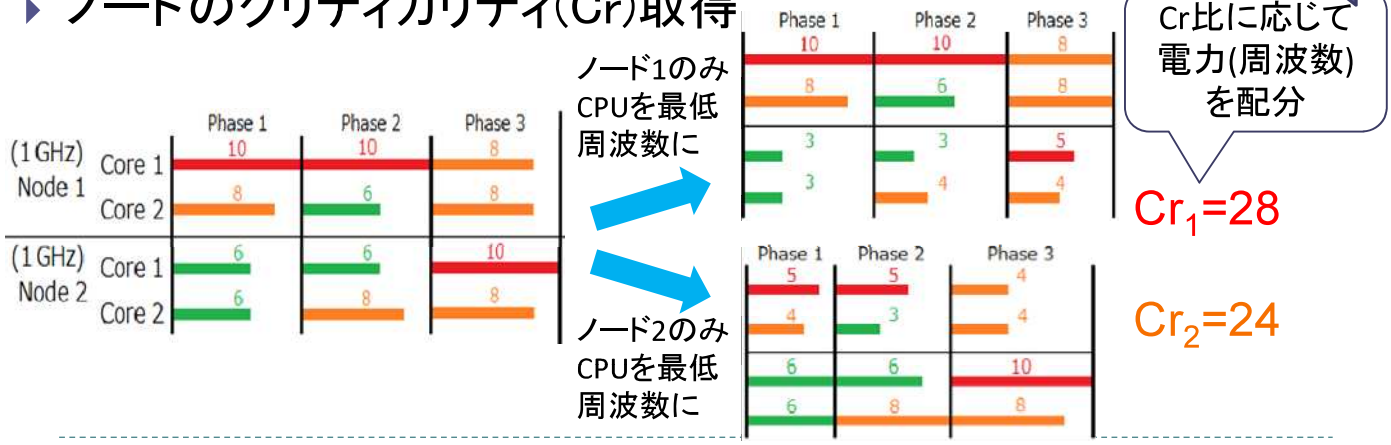


プロセッサ-ネットワーク間の電力分配法[會田2013]

- Energy Efficient Ethernet (EEE)
  - インタフェース上にデータが流れない期間にリンクを省電力モードへ移行
  - タイムアウトで省電力モード
  - オンデマンドで復帰



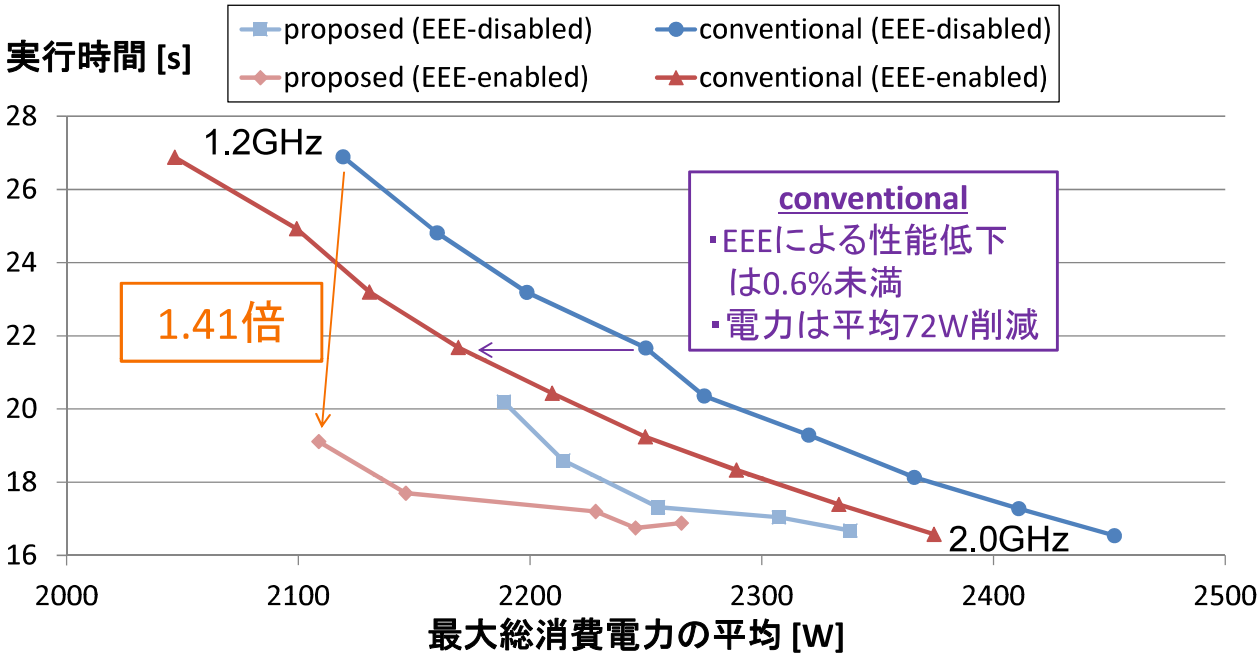
- ノードのクリティカリティ(Cr)取得



プロセッサ-ネットワーク間の電力配分実験結果

- ネットワークの静的な余剰電力を利用することで1.4倍の性能向上を実現

ノード: Dell PowerEdge r620 x 16
スイッチ: Dell PowerConnect 8132 x 3
- 10GbE: 24ポート, EEE対応)
- Fat-Tree構成



まとめ

- エクサスケールに向けたPower Wall問題
  - 大規模システムの電力トレンドと将来予測
  - 2020年前後に20MWでExaFlopsの実現は簡単ではない
- 今後の打開策
  - 電力効率を向上するための種々の要素技術
  - 電力を最重要資源と考えた電力マネージメント
- 電力マネージメントフレームワークの研究紹介
  - 電力制約適応型システムの必要性
  - 適切な電力分配で性能は大きく向上する可能性あり
- 電力効率の向上に向けた継続的な技術開発が必須

参考文献

- [Wallance2013] S. Wallace, V. Vishwanath, S. Coghlan, Z. Lan, M. Papka, “Measuring Power Consumption on IBM Blue Gene/Q”, Proc. 9th HPPAC, May 2013.
- [Stevens 2008] R. Stevens, et. al. “Scientific Grand Challenges: Architectures and Technology for Extreme Scale Computing”, Technical report, ASCR Scientific Grand Challenges Workshop Series, Dec. 2009.
- [Kogge2008] P. Kogge, et. al., “ExaScale Computing Study, Technology Challenges in Achieving Exascale Systems”, IPTO tech. report TR-2008-13, DARPA, Sep. 2008.
- [石川2012] 石川他, “計算科学研究ロードマップ白書”, 2012年3月.
- [Pawlowski 2011] J. T. Pawlowski, “Hybrid Memory Cube (HMC)”, Hot Chips23, Aug. 2011.
- [Borkar2013] S. Borkar, “Exascale Computing – a fact or a fiction?”, IPDPS2013 Keynote, May 2013.
- [Sakurai2011] T. Sakurai, “Pitfalls in deep-volt logic design”. ISSCC’11 Forum: Ultra-Low Voltage VLSIs for Energy-Efficient Systems, Feb. 2011.
- [Sakurai2011-2] T. Sakurai, “Designing Ultra-Low Voltage logic”. Proc. ISLPED’11, pp57-58, Aug. 2011.
- [HMC2013] Hybrid Memory Cube Consortium, “Hybrid Memory Cube Specification 1.0”, 2013.

参考文献

- [追永] 追永, “FXシリーズの今後の取り組みについて”, SS研HPCフォーラム2013, 2013年8月.
- [Intel2012] Intel Power Governor, http://software.intel.com/en-us/articles/intel-power-governor, Jul. 2012
- [カオ2013] カオ, 和田, 近藤, 本多, “RAPLインタフェースを用いたHPCシステムの消費電力モデリングと電力評価”, 情報処理学会HPC研究会, 2013年10月.
- [Miyazaki2013] H. Miyazaki,“K Computer: 8.162 PetaFLOPS Massively Parallel Scalar Supercomputer Built with Over 548k Cores”, ISSCC’12, Feb. 2012.
- [Kondo2007] M. Kondo, Y. Ikeda, and H. Nakamura,“A High Performance Cluster System Design by Adaptive Power Control”, HPPAC2007, Mar. 2007.
- [吉田2013] 吉田, 佐々木, 深沢, 稲富, 上田, 井上, 青柳, “CPUと主記憶への電力バジェット配分を考慮したHPCアプリケーションの性能評価”, 情報処理学会HPC研究会, 2013年10月.
- [會田2013] 會田, 三輪, 中村, “電力制約下におけるCPUとネットワークの電力制御協調手法, SWoPP2013, 2013年7月.

科	学	技	術	計	算	分	科	会		選	出

科学技術計算分科会 2013 年度会合 より

これで我々のアプリケーション プログラムは速くなるか？ -マルチコアクラスタ性能 WG 成果報告-

高木 亮治

(宇宙航空研究開発機構)

サイエンティフィック・システム研究会　　2013 年 10 月 23 日
計算科学技術分科会 2013 年度会合

これで我々のアプリケーションプログラムは速くなるか？

－マルチコアクラスタ性能 WG 成果報告－

高木　亮治

宇宙航空研究開発機構 宇宙科学研究所

[アブストラクト]
マルチコアクラスタ性能WGでは、会員が開発した各分野のアプリケーションプログラムを対象に、FX1、次世代スーパーコンピュータ「京」、PRIMEHPC FX10 上での性能分析・評価およびチューニングを実施することで、マルチコアクラスタマシンに向けた並列プログラミングモデルや高速化手法の検討およびノウハウの共有を行った。ここでは本 WG 活動を通じて得られた実践的な成果について報告する。はたして我々のアプリケーションプログラムを速くすることができただろうか？

[キーワード]
マルチコアクラスタ、アプリケーション、性能評価、高速化

1. はじめに

次世代スーパーコンピュータ「京」（以下、単に「京」）は、ノード内マルチコアおよび大規模ノード数を特徴とした並列計算機であり 2012 年に本格稼働を開始した。このようなマルチコアかつ大規模ノード数といった特徴を有する並列計算機を有効活用するためのノウハウは「京」開発当時は十分ではなく、そのため既存のスーパーコンピュータユーザがすみやかに「京」に移行し、「京」の潜在能力を有効活用するためのノウハウを蓄積することを目的に、2010 年 12 月にマルチコアクラスタ性能 WG が設立された。ここでは本 WG の活動概要とその成果について紹介する。

2. WG の活動概要

WG は 2010 年 12 月から 2013 年 5 月までの 2.5 年間に全 10 回の会合を持った。WG では会員が開発/保有する各分野のアプリケーションプログラム（流体解析が多く、他に構造解析、プラズマ解析など）を対象に、まずはノードアーキテクチャが「京」に似た FX1 での性能分析や評価を行った。その後稼働した「京」、更に WG 活動の後半で利用可能となった PRIMEHPC FX10（以下、単に FX10）を用いて性能分析・評価およびチューニングを実践することで、マルチコアクラスタマシンに向けた並列プログラミングスタイルや高速化手法の検討およびノウハウの共有を行った。また、PA ツールや会員から提供されたツールなど性能分析ツールの紹介や利用方法、ノウハウの共有も実施した。WG の活動を通じて得られた成果である各アプリケーションプログラムの性能測定結果及びその評価、高速化チューニングの各種ノウハウなどを実践的事例として成果報告書「実践、アプリ高速化に向けて」（約 180 ページ）にまとめた。また「PRIMEHPC FX10 チューニングチュートリアル」（約 300 ページ）にも反映された。

3. 活動の総括

WG で扱った会員のアプリケーションプログラムとしては流体、構造など連続体系のものが多く、そ

のためメモリアクセス性能を向上させるチューニングが主となった。その中で特にプリフェッチの有効利用に関する議論を行った。プリフェッチの有効活用ではハードウェアプリフェッチ（HWPF）とソフトウェアプリフェッチ（SWPF）の使い分けがポイントとなるが、その判断や効果が結局のところ試行錯誤である場合が多く、一般的なユーザーが取り組む手法としてはハードルが高いものと感じられた。この様に性能評価および高速化チューニングにおいて、性能評価や分析のためのデータを取得するツールが色々と整備されてきており、以前に比べて性能評価や分析が効率よくできる様になった。しかしながら、ユーザーの何故こうなるのか？どうすれば良いのか？の質問に的確に答えられないのが現状である。これはユーザーだけではなく専門家？でさえ試行錯誤を行っているためであり、この様な現状は、はなはだ憂慮すべきことであると感じた。同じ実装でもデータサイズが異なっただけで正反対のチューニングを実施するような状況では、WG の求める汎用化、共有化は非常に困難となり、ややもすると個別ケースに関する各論になりがちであった。もっとメタな議論が必要とされるが、そのためには、ハードウェア開発者、コンパイラ開発者、ユーザーの三者間での密接な議論が必要（所謂 Co-design）と痛感した。不十分であったかもしれないがそういった場を提供できた本 WG の活動は有意義であり今後も継続していくことが望まれる。特に、今後想定されるより高いハードル（さらに使いにくいマシンへの対応）に向けては Co-design と呼ばれるより密な議論が必要と思われ、そういった活動をどう実現していくかが喫緊の課題である。

4. おわりに

マルチコアクラスタ性能 WG では FX1 や「京」、FX10 などマルチコアクラスタを対象として会員のアプリケーションプログラムの評価および高速化チューニングを実施し、その成果やノウハウの共有を試みた。しかしながらアプリケーションプログラムの開発者にとって自分のプログラムの高速化チューニングはますます困難な状況になっていることを痛感したのではないかと思われる。今後のエクサスケール計算機では、更に危機的状況になると予想されており、ハードウェアおよびシステム、コンパイラなどの開発者の支援の下、ユーザーが主体的に性能評価・高速化チューニングを実施する本 WG の様な活動は貴重であり、今後も継続されることが期待される。

本講演のタイトルである「これで我々のアプリケーションプログラムは速くなる（なった）か？」であるが、（少なくとも WG に参加すれば）速くなったと回答したい。WG に参加することで、計算機の専門家、先達ユーザーから貴重なアドバイスをもらえ、それらが高速化の役に立ったと思う。皆さん是非 WG に参加しよう！！

最後になるが、貴重な時間を割いて 2.5 年間の WG 活動に参加していただいた会員の皆さま、富士通の担当者、至らないまとめ役を全面的にバックアップして WG の円滑な運営に尽力していただいた SS 研事務局のみなさまに感謝する。また、WG では以下の環境を利用した。改めてここに感謝の意を表する。

スーパーコンピュータ「京」試験利用枠、一般公募枠
宇宙航空研究開発機構 JSS
日本原子力研究開発機構 BX900, FX1
国際核融合エネルギー研究センター Helios
名古屋大学 FX1
東京大学 FX10
九州大学 CX400, FX10
理化学研究所 RICC
富士通社内機

これで我々のアプリケーションプログラムは速くなるか？

—マルチコアクラスタ性能WG成果報告—

マルチコアクラスタ性能WG まとめ役　高木亮治
宇宙航空研究開発機構

内容

- WGの概要
- アプリケーションチューニングの実践例
  - 会員のいくつかの例より
- 活動の総括
  - かなり個人的
- まとめ

WG概要

WG設置の背景

- WG立ち上げ：2010年
- FX1から次世代スーパーコンピュータ「京」（、FX10）への流れ
  - ノード内マルチコアの大規模クラスタ
    - FX1：4コア、「京」：8コア、FX10：16コア
  - 「京」の本格稼働：2012年
  - 「京」の利用促進に向けたプログラムの性能評価と高速化手法の検討
- 「京」を始めとしたマルチコアクラスタマシンに向けた、
  - 並列プログラミングモデル
  - 性能評価ツールの利用法、分析手法
  - 高速化チューニング

  に関するノウハウの共有を目指した。

活動期間とメンバー

- 活動期間：2010年12月～2013年5月（2.5年）
- メンバー：

	氏名	所属		氏名	所属
担当幹事	石井 克哉	名古屋大学	推進委員	福島 正雄	富士通（株）
推進委員	高木 亮治	宇宙航空研究開発機構		青木 正樹	富士通（株）
	井戸村 泰宏	日本原子力研究開発機構		山中 栄次	富士通（株）
	梅田 隆行	名古屋大学		三吉 郁夫	富士通（株）
	荻野 正雄	名古屋大学		三輪 英樹	富士通（株）
	坂下 雅秀	宇宙航空研究開発機構		内藤 俊也	富士通（株）
	佐藤 幸紀	北陸先端科学技術大学院大学		錦 龍生	富士通（株）
	柴村 英智	九州先端科学技術研究所		瀧 康太郎	富士通（株）
	野田 茂穂	理化学研究所		千葉 修一	富士通（株）
	姫野 龍太郎	理化学研究所	オブザーバー	森重 博司	富士通（株）
	堀之内 成明	（株）豊田中央研究所	オブザーバー	市川 真一	富士通（株）
	南 一生	理化学研究所			

活動内容の概略

- 全10回の会合（2.5年間）
- 活動内容
  - 情報提供
    - 次世代スーパーコンピュータ「京」
    - 性能解析ツール：PAツール、会員ツール
    - 各種チュートリアル
  - 会員アプリの測定報告
    - 性能測定、チューニング
- 成果
  - 成果報告書（約180ページ）
  - PRIMEHPC FX10チューニングチュートリアル（約300ページ）


マルチコアクラスタ性能WG成果報告書

「「実践、アプリ」高速化に向けて」

（WG活動期間2010年12月-2013年5月）

2013年5月10日

サイエンティフィック・システム研究会

マルチコアクラスタ性能WG

成果報告書(1/2)

- アプリケーションの測定評価(会員から)
  - 3次元FEM構造解析コード:ADVENTURE
  - 3次元非圧縮性流体計算プログラム:COSMOS
  - 3次元電磁界コード:FDTD3
  - 核融合プラズマ5次元格子コード:GT5D
  - 圧縮性流体解析プログラム:UPACS
  - 宇宙プラズマ5次元ブラソフコード:Vlasov5
  - 流体構造連成解析アプリ:ZZ-EFSI
  - 超音波集束シミュレータ:ZZ-HIFU
  - 非構造格子CFDソルバ:JTAS

成果報告書(2/2)

- 性能評価ツール(会員から)
  - インターコネクトシミュレータ:NSIM
  - 実行駆動型アプリ解析ツール:Exana
- 共通事項(富士通から)
  - H/WとS/Wプリフェッチの仕様
  - キャッシュミス数/ミス率
  - FMA命令化

別冊 PRIMEHPC FX10 チューニングチュートリアル

- 第1章 プログラミング言語処理系概略
- 第2章 PAイベント偏
- 第3章 Fortran偏
- 第4章 C/C++偏
- 第5章 チューニングツール偏
- 第6章 ノード内チューニング偏
- 第7章 MPIおよびノード間チューニング偏

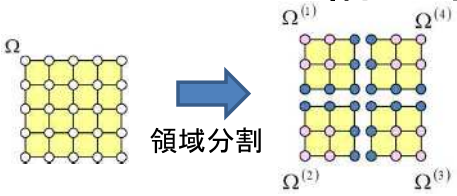
会員のいくつかの例より

アプリケーションチューニングの実践例

ADVENTURE（荻野、名大）

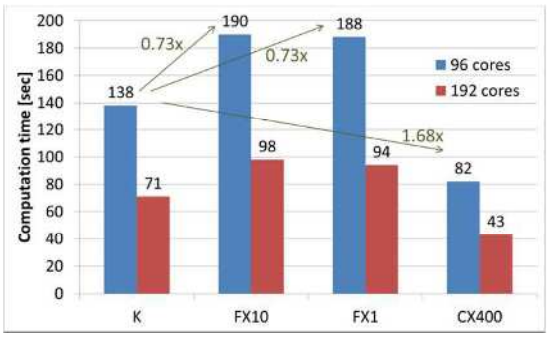
- FEMによる弾塑性解析
  - 非構造格子、疎行列の反転
- FX1, FX10, CX400, 京で評価
  - メモリバンド幅ネック
  - 直接法→反復法でメモリバンド幅ネックを緩和
  - 現時点では直接法が速いが、コア数が増え、コア当たりのメモリバンド幅が低下した場合は反復法が有利

3次元FEM構造解析コードADVENTURE



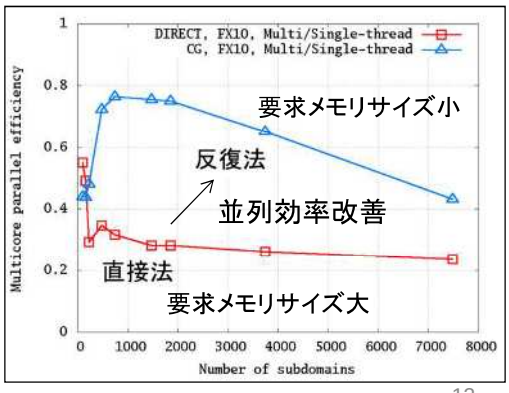

```
#pragma omp parallel for schedule(dynamic,1)
for i=1,...,N_j        領域方向ループ
  s^(i)n = -A_IB^(i) R_B^(i) p^n
  t^(i)n = A_II^(i)-1 s^(i)n
  q^(i)n = A_BB^(i) R_B^(i) p^n + A_IB^(i)T t^(i)n      領域FEM
#pragma omp critical
  q_j^n = q_j^n + R_B^(i)T q^(i)n
endfor
q^n = q^n + q_j^n
```

領域分割法における高コスト部分



同一コア数における性能比較

- 領域分割法は, 粗粒度の並列性を持つ
- 一方, 各スレッドが高いB/Fを要求する
- 同一コア数で比較するとシステムのB/F値に従った性能差が見られた
- 領域FEMの要求メモリサイズが小さくなる実装を行い, CPU内並列効率の改善が得られた



マルチコアCPU内の並列効率

COSMOS（堀之内、豊田中研）

- 非定常非圧縮性乱流計算プログラム
  - LES, 構造格子（物体適合、重合）, 陰解法（行列反転, SOR）
- RX600, FX1, FX10, 京で評価
  - マルチカラー化（8色）、コンパイラオプション
  - 反復解法レベルでのアルゴリズムの見直し
  - マルチコアに特化した配列構造の利用？

チューニング対象とした計算の概要

- LES(*)による非圧縮性流れの計算
  - 基礎方程式: 連続の式，運動方程式（Navier-Stokes 方程式）
    ↓ SMAC法に準じた陰解法（時間積分: Crank-Nicolson法）
    運動方程式，圧力Poisson方程式に対する連立一次方程式に帰着
    （計算時間全体の8割程度を費やす）
  - 格子体系: 構造格子を組み合わせた重合格子
    ↓ 各部分格子ごとに離散化
    係数行列: 規則的（非対称）スパース行列
- 連立一次方程式 $Ax=b$ の解法: SOR(**)法
  - $A=L+D+U$ とした時，下三角行列 L にかかる要素は常に最新の値を参照；
    $(D+\omega L)\boldsymbol{x}^{(n+1)}=\{(1-\omega)D-\omega U\}\boldsymbol{x}^{(n)}+\omega\boldsymbol{b}$
    → 回帰参照が発生
  - マルチカラーオーダリングによる並列化
    ⇒ これの高速化がターゲット

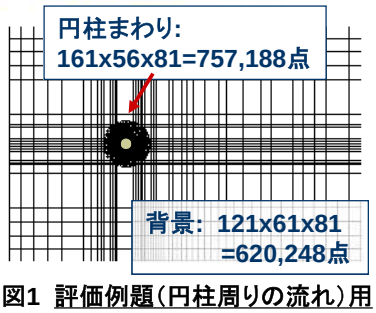


図1 評価例題（円柱周りの流れ）用重合格子

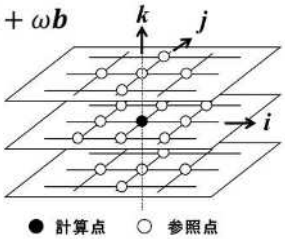


図2 格子点の参照関係

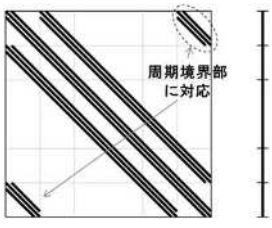


図3 係数行列のイメージ

(*) Large Eddy Simulation
(**) Successive Over-Relaxation

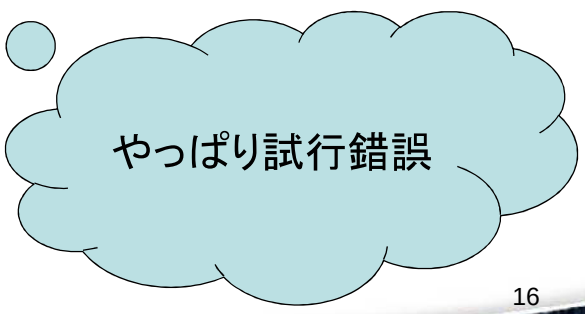
チューニング結果のまとめ

- オーダリングの修正による高速化（*on FX1 4core*）
  - オリジナル：3次元の格子点を1次元化した配列に入れて，最小の色数となる7色でオーダリング → 1次元ループのストライドアクセス
  - 改良版：3次元（*i,j,k*）各方向ごとに2色化した8色でオーダリングし，かつ，各色ごとのループに分ける．（配列のとり方は変えていない）
    ⇨ **プログラム全体で8%の実効速度向上（ｷｬｯｼｭｱｸｾｽ待ち削減）**
- コンパイルオプションによるチューニング
  （上記改良版に対して，*on FX10 16core*）
  - 圧力Poisson方程式から得られる連立一次方程式の計算：
    ソフトウェアパイプラインによる命令スケジューリング
    ⇨ **該当ルーチンで11%の実効速度向上（浮動小数点演算待ち，整数演算待ち削減 ⇒ ﾒﾓﾘｽﾙｰﾌﾟｯﾄ改善）**
  - 運動方程式から得られる連立一次方程式の計算：
    ソフトウェアプリフェッチと，ストライドアクセスオプション指定
    ⇨ **該当ルーチンで11%の実効速度向上**
    （浮動小数点ﾛｰﾄﾞﾒﾓﾘｱｸｾｽ待ち，L2ﾐｽﾃﾞﾏﾝﾄﾞ率削減
    ⇒ ﾒﾓﾘｽﾙｰﾌﾟｯﾄ改善）

FDTD3（梅田、名大）

- 3次元の電磁場解析
  - 構造格子
- FX1, FX10, 京でノード性能の評価
  - メモリバンド幅ネック
    - 配列インデックスの違い（ベクトル型、スカラー型）
    - キャッシュの再利用具合はアルゴリズムに依存
  - ループ分割か融合か？
  - 配列の融合の是非？

A(i,j,k,n) or
A(n,i,j,k) or
A1(i,j,k), A2(i,j,k),…, An(i,j,k)



GT5D（井戸村、JAEA）

- 第一原理プラズマ乱流コード
  - 5次元位相空間（3次元空間×2次元速度空間）
    - 3次元流体に比べて100^2倍自由度が大
- BX900, Helios, 京, FX1, FX10で評価
  - バンド幅ネック
  - ルーフラインモデルによる性能予測と実測の比較
  - 通信マスク手法の適用
    - 袖通信：15%削減（京）、41%削減（Helios）
    - 60万コアまで99.99989%の並列化効率を達成（24,576コアから589,824コアまでのストロングスケーリングでの評価）
- 課題：大規模並列I/O、可視化

核融合プラズマ5次元格子コードGT5Dの測定評価

■ 概要

核融合プラズマ5次元格子コードGT5Dの並列化率向上を目的として、演算と通信を同時処理する通信マスク手法を開発し、10万コア以上のストロングスケーリングを実現

■ 通信マスク手法

①MPIライブラリ（富士通、インテル）における*RendezVouz*プロトコルの問題により演算中に非同期通信が機能しない原因を解明

②この問題を回避する2つの手法を開発

B.MPI_Testによる*RendezVouz*プロトコル促進

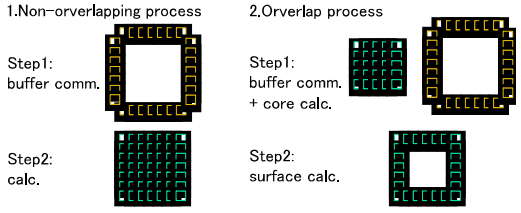
C.OpenMPによる通信スレッドの実装

③手法B、CをGT5Dにおける差分演算の袖領域通信、さらに、手法Cをデータ転置の集団通信に適用し有効性を確認

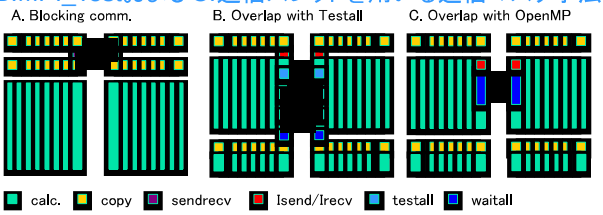
→並列化率～99.9999%を達成し、京60万コアを用いてBX900の約35倍の高速計算を実現

[Idomura et al., Int. J. HPC Appl. 2013]

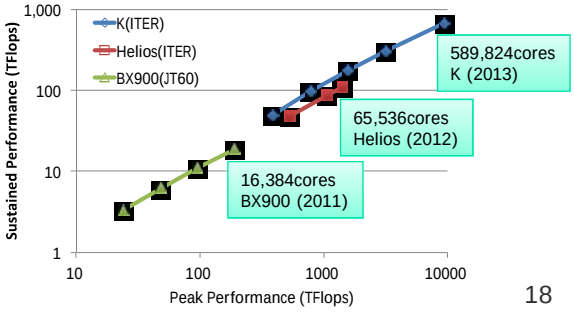
通信と演算の同時処理によるオーバーヘッド削減



B.MPI_TestおよびC.通信スレッドを用いる通信マスク手法



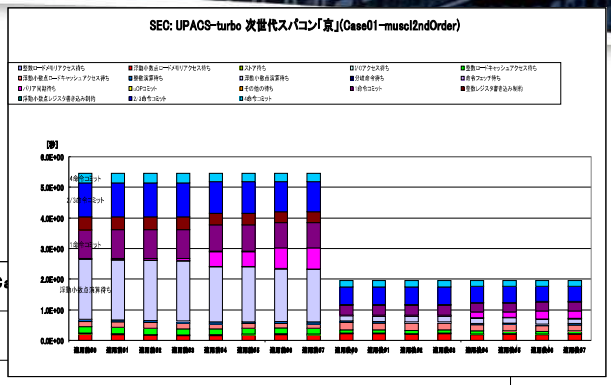
京およびHeliosにおけるGT5Dのストロングスケーリング

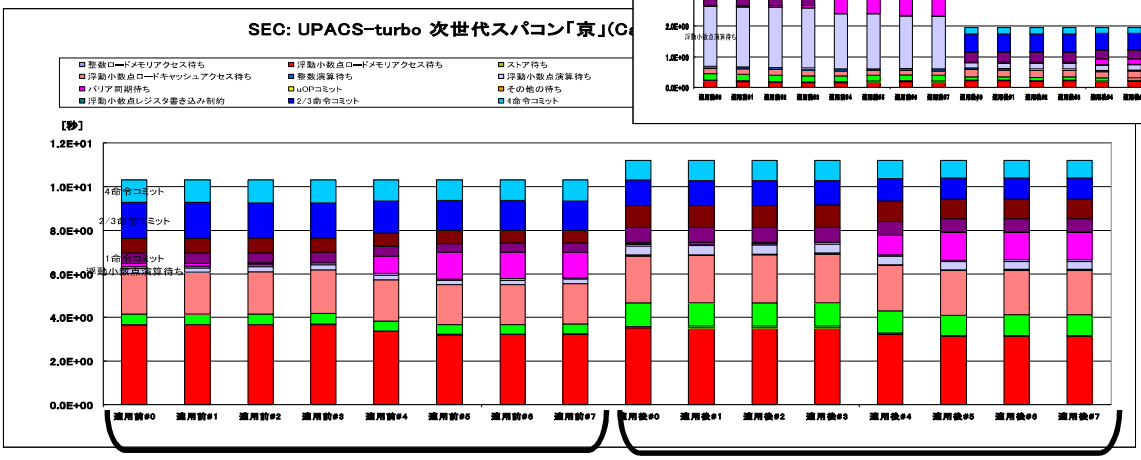


UPACS（高木、JAXA）

- 3次元圧縮性流体解析プログラム
  - 構造格子（マルチブロック, 重合）
- FX1, 京で評価
  - スレッド並列の促進、キャッシュチューニング、SIMD化
    - Allocate/deallocate：アリーナ開放の抑止
  - SIMD化を促進してもメモリバンド幅ネックの場合は速度向上なし
    - チューニングの指針として何を見るべきか？

- SIMD化率とFLOPSの関係
  - SIMD化を促進しても性能が向上せず。





SIMD化促進前　　SIMD化促進後

PAデータ 測定区間	実行時間 (sec)	浮動小数点 演算ピーク比	MFLOPS	MIPS	浮動小数点 演算数	SIMD演算命令率 (/対象演算命令数)
適用前#0	10.31	9.45%	1512	1929	1.56E+10	0.00%
適用後#0	11.21	8.69%	1391	1707	1.56E+10	98.85%

UPACSカーネル（高木、JAXA）

- UPACSのカーネル部分（対流項、時間積分:陽解法）
  - 従来のベクトル型ループ
    - 空間スイープが多い
  - 局所性を意識したループ
- JSS, Intel CPUで評価
  - キャッシュミス率の低減
  - 低B/Fでの性能向上が期待

データ&ループ構造

データ：Q(i,j,k,n), I,j,k：空間、n：物理量

ループA：

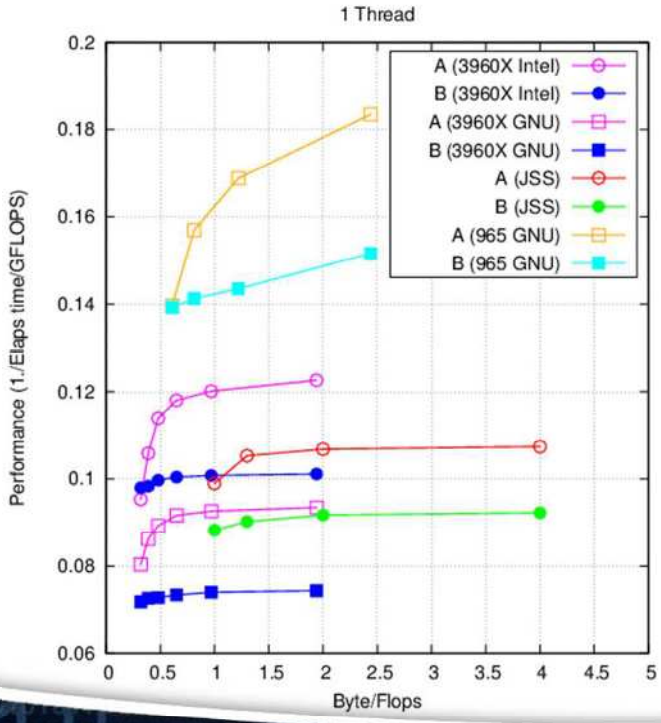
空間ループ

```
do dir=1,3
  do k=1,kmax, do j=1,jmax, do i=1,imax
      MUSCLの計算
  enddo, enddo, enddo
  do k=1,kmax, do j=1,jmax, do i=1,imax
      FLUXの計算
  enddo, enddo, enddo
  do k=1,kmax, do j=1,jmax, do i=1,imax
      RHS(ΔQ)の計算
  enddo, enddo, enddo
enddo
do k=1,kmax, do j=1,jmax, do i=1,imax
      時間積分
enddo, enddo, enddo
```

ループB：

```
do k=1,kmax, do j=1,jmax, i=1,imax
    do ndir=1,3
        MUSCLの計算
        FLUXの計算
        RHS(ΔQ)の計算
        境界での処理(MUSCL, FLUX, ΔQ)
    enddo
enddo, enddo, enddo
do k=1,kmax, do j=1,jmax, i=1,imax
      時間積分
enddo, enddo, enddo
```

ループAとBの比較（1スレッド）



- 仮想的にメモリバンド幅を変化させた。
  - スレッド数は1で固定
  - ブロックサイズは80
- 縦軸は理論ピーク性能あたりの性能（経過時間の逆数）
- Byte/FLOPは理論性能
- B/Fが悪化すると、ループAは急激に性能が悪化する。

ZZ-EFSI（野田、理研）

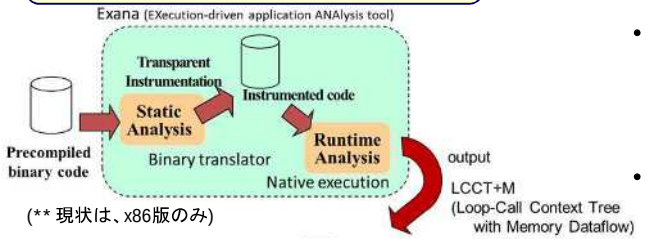
- 流体構造連成解析
  - ボクセル格子, WENO
  - 既存のチューニングではなく新たに設計（京の性能を出す）
    - 計算アルゴリズムの選択
- RICC@理研, 京で評価
  - 高い実行性能
    - ノード：46.4%、12,288ノード：43.2%
  - **優秀なスタッフの理詰め、でも最後は試行錯誤**

性能評価ツール

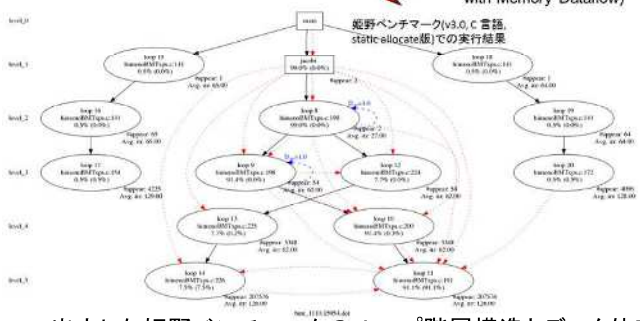
- NSIM（柴村、九州先端）
  - インターコネクトシミュレータ
  - 残念ながらユーザーの利用はなかった。
- Exana（佐藤、北陸先端）
  - プロファイラー
    - ホットスポット、ループ階層構造とそれらの間の並列性の検出
  - 残念ながらユーザーの利用はなかった。
    - チューニングのノウハウや事例に基づき機能要件を検討した。
    - 並列性はループだけでない。
    - キャッシュの挙動の考慮、デバッガとの連携、部分解析、オーバーヘッド

実行駆動型アプリケーション解析ツール　Exana

本ツールは動的バイナリ変換によりコード実行時にループおよびデータフロー情報を抽出



(** 現状は、x86版のみ)



出力した姫野ベンチマークのループ階層構造とデータ依存

丸いノードがループ
四角のノードは関数
実線はコントロールフロー
　親子関係はネストで子ノードは内部ループ
点線はデータ依存

本WGでの議論による知見

- 流体と構造の連成解析のようなマルチフィジクスでは並列性はループだけとは限らないため本手法により関数とループのコンテクストによりコードを俯瞰することは有益
- チューニングへの応用のためには、ループ階層構造をキャッシュの挙動を如何に結びつけるかが鍵

出力したOpenMXのループ階層構造

- データ依存プロファイルをなしとすると解析オーバーヘッドは大幅に小
  - データ依存解析ありで50倍、なしで3倍程度のオーバーヘッド
  - データ依存なしでもプログラムのコードを俯瞰する手段としてはOK
  - ループ階層構造をキャッシュの挙動と結び付ける解析が望まれる
  - 本ツールでキャッシュ性能を推測できるかということは検討が必要

かなり個人的

活動の総括

議論を通じて得られた知見

- メモリバンド幅ネックのアプリケーションが多く、そこのチューニングが主。
- メモリのスループットを上げるために今回はプリフェッチ（PF）に注目
  - 通常はHWPFを使うが、場合によってはSWPFを使った方が良い場合がある。
    - 何時SWPFを使うか？
    - 「今でしょ！！」という簡単明瞭な基準がなくて「ケースバイケース」（←個人的には悪夢の言葉！！）

議論を通じて得られた知見

- HWPFとSWPFの仕組み
- どういう場合はどちらを使うか？
  - ケースバイケースだが、いくつかの事例はまとめた。
  - SWPFを使うとき：
    - 連続アクセスだが、途中でアクセスが飛ぶ
      - 無駄なアクセスをしない
    - 翻訳時オプションと最適化指示子の利用法
      - 最適化指示子が確実
- ベンダーはコンパイラにお任せあれと言うが…。

議論を通じて得られた成果

- チューニング支援機能としてコンパイラへの改善要求
  - FMA命令化のメッセージ等を出力する。
  - 現状の問題点
    - FMA命令化はコーディングスタイルに依存しない → ユーザーは操作できない。
    - FMA命令化したかどうかはアセンブラを見るしかない。

チューニングの現状（かなり個人的）

- やっていること:
  - ある程度見通しをつけたら、まずは試してみる！
  - どんどん試す、ひたすら試す！！！
- 「試行錯誤の世界」
  - 微かにある理詰めも最後は「ケースバイケース」で粉砕される。
- やっぱり一般ユーザーの手に負えないレベル！
  - 専門家にまかせるしかない？
  - 専門家でも試行錯誤！！
    - WGでも何故何故攻撃の繰り返し

チューニングの現状（かなり個人的）

- アーキテクチャが複雑化してチューニングが大変
  - ベクトル型CPUは単純で使いやすかった。
- 何が問題か？
  - B/Fを要求するアプリケーションに対して計算機のB/Fが低下 → アプリケーション側で工夫が必要。
  - 性能評価データの分析も困難。
  - チューニングも含めて、プログラムをどう書けば良いかわからない。
    - 私だけかもしれないが、どうもそうでもない…。
- 何故そうなったか？
  - アーキテクチャ（H/W, コンパイラ）と実際の動きを正確に理解できていない。

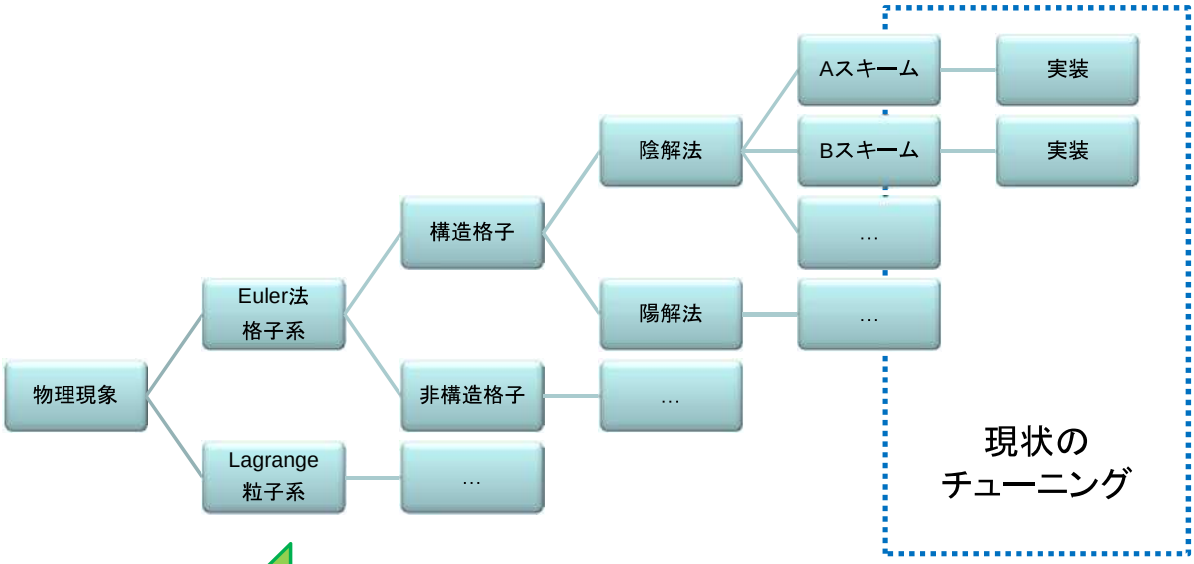
チューニングの現状（かなり個人的）

- ユーザーはH/W、コンパイラについてどんだけ勉強すれば良いのでしょうか？
  - 「試行錯誤」、「ケースバイケース」では先に進めない！！
- そもそもこんなスーパーな人はいない？
  - ユーザーの何故何故攻撃に答えられる。
  - 理詰めで分析・チューニングができる。
  - ケースバイケースの判断基準が明確。
  - 試行錯誤しなくてもどっちが良いか判断できる。

WGの限界

- 個別アプリケーションのチューニングに終始
  - 「試行錯誤」と「ケースバイケース」
- メタな情報としては汎用化、共有化できるが、具体化したとたんに個別ケースになりがち。
- もっとメタと個別をつなぐ部分が必要。
  - ある意味Co-Designだと思うが、どう実現すれば良いか？

Co-Design（ex. 流体解析）



Co-Designではどこまで上流に行けるか？
上流に行けば行くほど共有化、汎用化が可能では？

まとめ

- マルチコアクラスタ性能WGでは、FX1,「京」, FX10などのマルチコアクラスタを対象としたアプリケーションプログラムの性能評価および高速化チューニングを実施した。
  - 報告書、FX10向けチューニングチュートリアル
- 今後のエクサスケール計算機開発に向けて、H/Wおよびシステム開発者、コンパイラ開発者の支援の下、ユーザーが主体的に性能評価・高速化チューニングを実施する本WG的活動が継続されることを期待する。

まとめ

- これで我々のアプリケーションプログラムは速くなる（なった）か？
- 回答：（少なくともWGに参加すれば）速くなった。
  - WGでは専門家（H/W、コンパイラ）に分析をしてもらえる。
    - 進むべき道も教えてもらえる？
  - 先達ユーザーからも教えてもらえる。
  - 継続することが必要。
- 皆さんWGに参加しましょう！！

今後

- チューニングも必要だがCo-Designを目指した活動が必要！！
  - WGでやれるか？（SS研でできるか？）
  - もう少し柔軟な活動形態が適切かも。
- SS研ならではのWG活動が求められている。
  - HPCI, HPCIコンソーシアム
  - 京、AICS、…

謝辞

- 2.5年間のWG活動に参加していただいた会員の皆様、富士通担当者様、SS研事務局の皆様に感謝の意を表します。
- WGでは以下の環境を利用させていただきました。改めてここに感謝の意を表します。
  - スーパーコンピュータ「京」試験利用枠、一般公募枠
  - 宇宙航空研究開発機構 JSS
  - 日本原子力研究開発機構 BX900, FX1
  - 国際核融合エネルギー研究センター Helios
  - 名古屋大学 FX1
  - 東京大学 FX10
  - 九州大学 CX400, FX10
  - 理化学研究所 RICC
  - 富士通社内機

科	学	技	術	計	算	分	科	会		選	出

HPC フォーラム 2013 より

Why we need Exascale, and why we won't get there by 2020

Horst D. Simon
(Lawrence Berkeley National Laboratory)

BERKELEY LAB

Why we need Exascale and why we won't get there by 2020

Horst Simon
Lawrence Berkeley National Laboratory

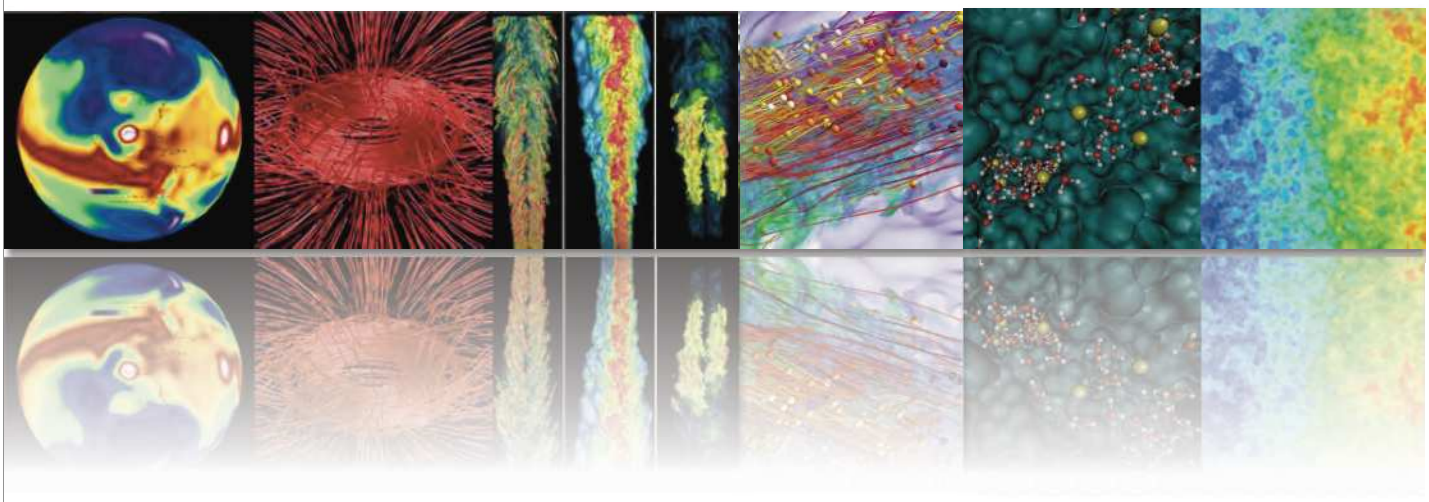
August 27, 2013

Overview

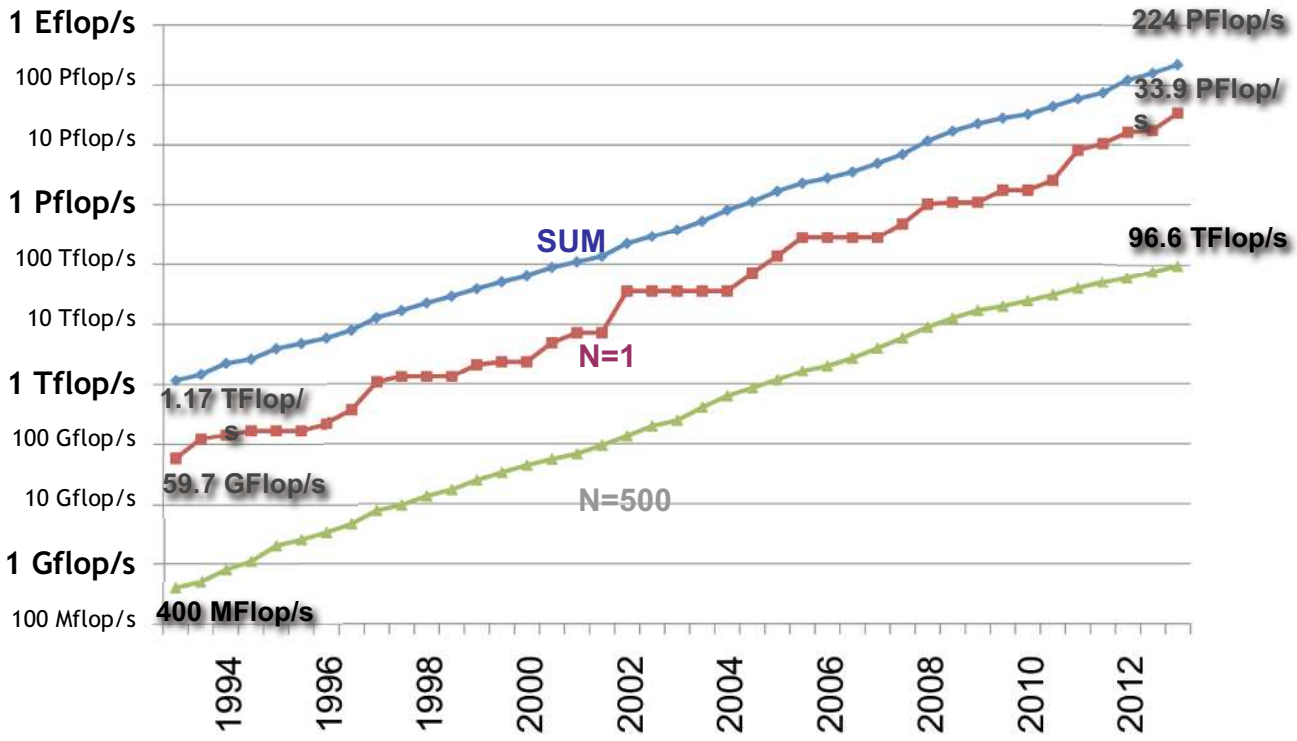
- **Current state of HPC: petaflops firmly established**
- **Why we won't get to exaflops (in 2020)**
- **The case for exascale**
  - **Science (old)**
  - **Science (new)**
  - **Technology and economic competitiveness**
  - **Discovery**

Thank you to the DOE Hexlab group, the TOP500 team, John Shalf, Kathy Yelick and many others

Current State of HPC: petaflops firmly established



Performance Development



Source: TOP500 June 2013

Top 10 Systems in June 2013

#	Site	Manufacturer	Computer	Country	Cores	Rmax [Pflops]	Power [MW]
1	National University of Defense Technology	NUDT	Tianhe-2 NUDT TH-IVB-FEP, Xeon 12C 2.2GHz, IntelXeon Phi	China	3,120,000	33.9	17.8
2	Oak Ridge National Laboratory	Cray	Titan Cray XK7, Opteron 16C 2.2GHz, Gemini, NVIDIA K20x	USA	560,640	17.6	8.21
3	Lawrence Livermore National Laboratory	IBM	Sequoia BlueGene/Q, Power BQC 16C 1.6GHz, Custom	USA	1,572,864	17.2	7.89
4	RIKEN Advanced Institute for Computational Science	Fujitsu	K Computer SPARC64 VIIIfx 2.0GHz, Tofu Interconnect	Japan	795,024	10.5	12.7
5	Argonne National Laboratory	IBM	Mira BlueGene/Q, Power BQC 16C 1.6GHz, Custom	USA	786,432	8.59	3.95
6	Texas Advanced Computing Center/UT	Dell	Stampede PowerEdge C8220, Xeon E5 8C 2.7GHz, Intel Xeon Phi	USA	462,462	5.17	4.51
7	Forschungszentrum Juelich (FZJ)	IBM	JuQUEEN BlueGene/Q, Power BQC 16C 1.6GHz, Custom	Germany	458,752	5.01	2.30
8	Lawrence Livermore National Laboratory	IBM	Vulcan BlueGene/Q, Power BQC 16C 1.6GHz, Custom	USA	393,216	4.29	1.97
9	Leibniz Rechenzentrum	IBM	SuperMUC iDataPlex DX360M4, Xeon E5 8C 2.7GHz, Infiniband FDR	Germany	147,456	2.90	3.52
10	National SuperComputer Center in Tianjin	NUDT	Tianhe-1A NUDT TH MPP, Xeon 6C, NVidia, FT-1000 8C	China	186,368	2.57	4.04

Technology Paths to Exascale

Leading Technology Paths (Swim Lanes):

Multicore — Maintain complex cores, and replicate (x86, SPARC, Power7) [#4 and 9]

Manycore / Embedded — Use many simpler, low power cores from embedded (BlueGene) [#3, 5, 7, and 8]

GPU / Accelerator — Use highly specialized processors from gaming/graphics market space (NVidia Fermi, Cell, Intel Phi (MIC)) [# 1, 2, 6, and 8]

Tianhe-2 (TH-2) at NUDT, China

Summary of the Tianhe-2 (TH-2) Milkyway 2

Model	TH-IVB-FEP
Nodes	16,000
Processor	Intel Xeon IvyBridge E5-2692
Speed	2.200 GHz
Sockets per Node:	2
Cores per Socket:	12
Coprocessors:	Intel Xeon Phi 31S1P
Coprocessors per Node:	3
Cores per Coprocessor:	57
Coprocessors total	48,000
Operating System:	Kylin Linux
Primary Interconnect:	Proprietary high-speed interconnecting network (TH Express-2)
Peak Power (MW):	17.8
Size of Power Measurement (Cores)	3,120,000
Memory per Node (GB)	64

Summary of all components

CPU Cores	384,000
Accelerators/CP	48,000
Accelerator/CP Cores	
Memory	1,024,000 GB

Summary of the Tainhe-2 (TH-2) or Milkyway-2	
Items	Configuration
Processors	32,000 Intel Xeon CPU's + 48,000 Xeon Phi's (+ 4096 FT-1500 CPU's frontend) Peak Performance 54.9 PFlop/s (just Intel parts)
Interconnect	Proprietary high-speed interconnection network, TH Express-2
Memory	1 PB
Storage	Global Shared parallel storage system, 12.4 PB
Cabinets	125 + 13 + 24 = 162 compute/communication/storage cabinets
Power	17.8 MW
Cooling	Closed air cooling system

BERKELEY LAB

Sequoia at Lawrence Livermore National Laboratory (Manycore/embedded)

- **Fully deployed in June 2012**
- **16.32475 petaflops Rmax and 20.13266 petaflops Rpeak**
- **7.89 MW Power consumption**
- **IBM Blue Gene/Q**
- **98,304 compute nodes**
- **1.6 million processor cores**
- **1.6 PB of memory**
- **8 or 16 core Power Architecture processors built on a 45 nm fabrication method**
- **Lustre Parallel File System**

BERKELEY LAB

K Computer at RIKEN Advanced Institute for Computational Science (Multicore)

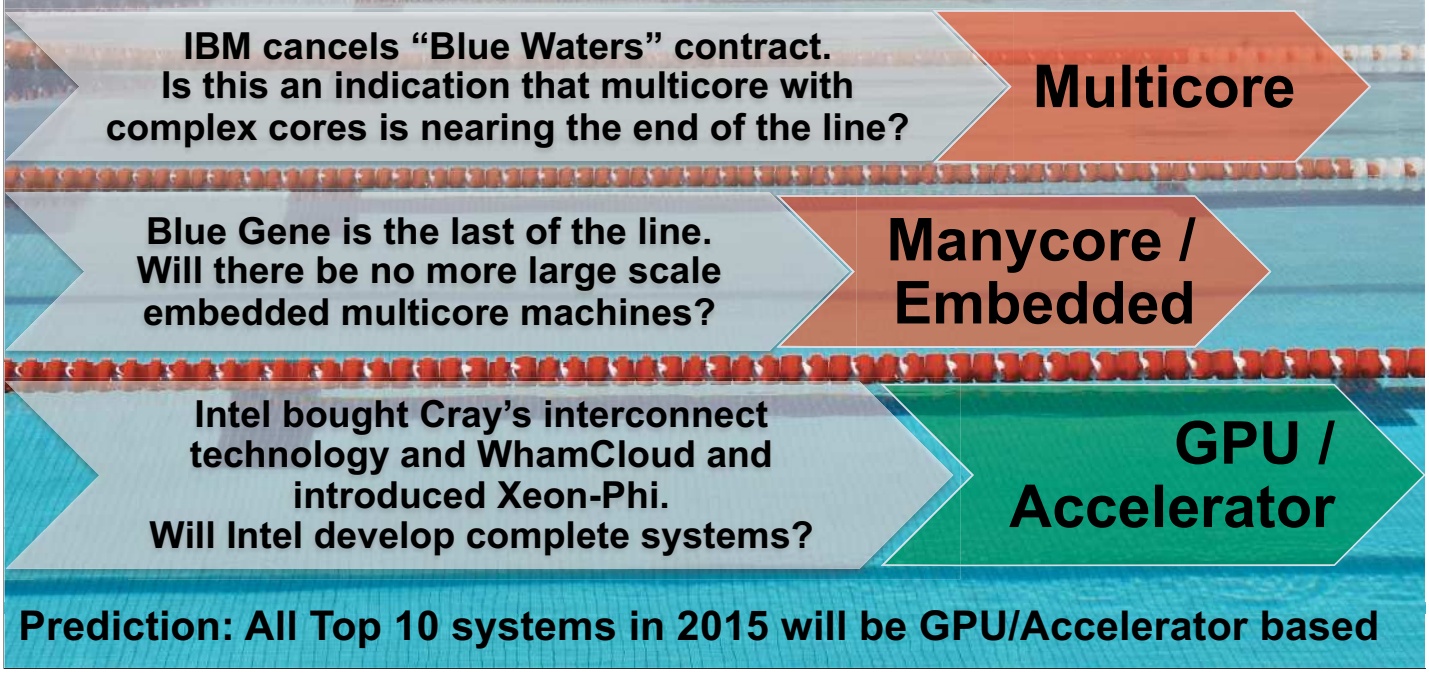
- #3 on the TOP500 list
- Distributed memory architecture
- Manufactured by Fujitsu
- Performance: 10.51 PFLOPS Rmax and 11.2804 PFLOPS Rpeak
- 12.65989 MW Power Consumption, 824.6 GFLOPS/Kwatt
- Annual running costs - $10 million
- 864 cabinets:
  - 88,128 8-core SPARC64 VIIIfx processors @ 2.0 GHz
  - 705,024 cores
  - 96 computing nodes + 6 I/O nodes per cabinet
- Water cooling system minimizes failure rate and power consumption

Choice of Swim Lanes

- **Swim Lane choices – currently all three are equally represented**
- **Risk in Swim Lane switch for large facilities**
  - **Select too soon:** Users cannot follow
  - **Select too late**: Fall behind performance curve
  - **Select incorrectly:** Subject users to multiple disruptive technology change

Events of 2011/2012 and the Big Questions for the Next Three Years until 2015

Leading Technology Paths (Swim Lanes):

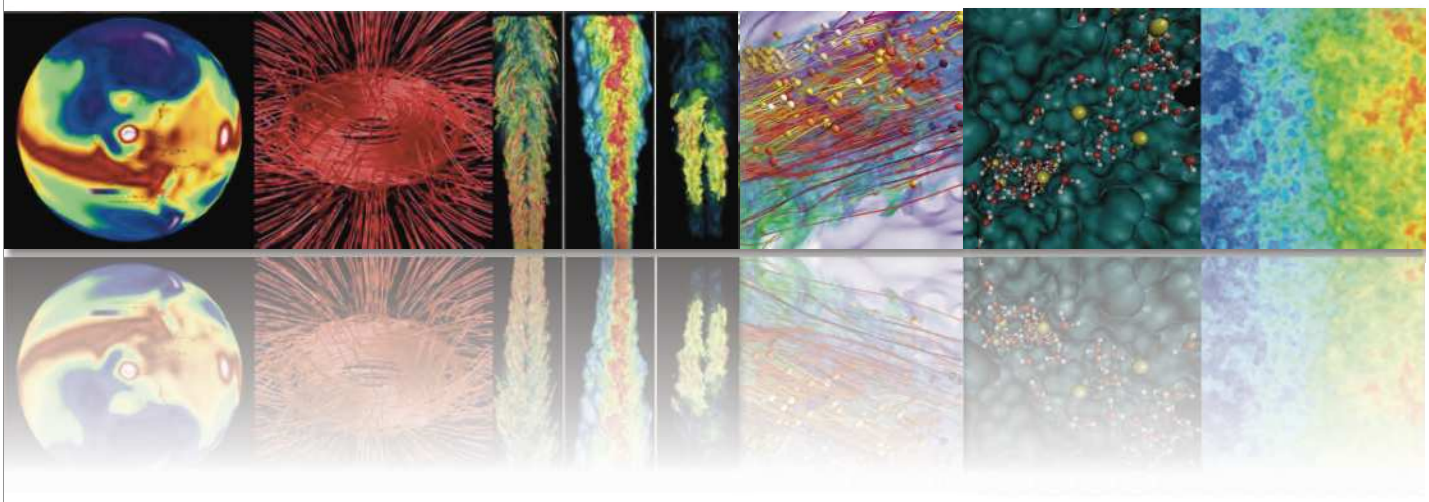


Prediction: All Top 10 systems in 2015 will be GPU/Accelerator based

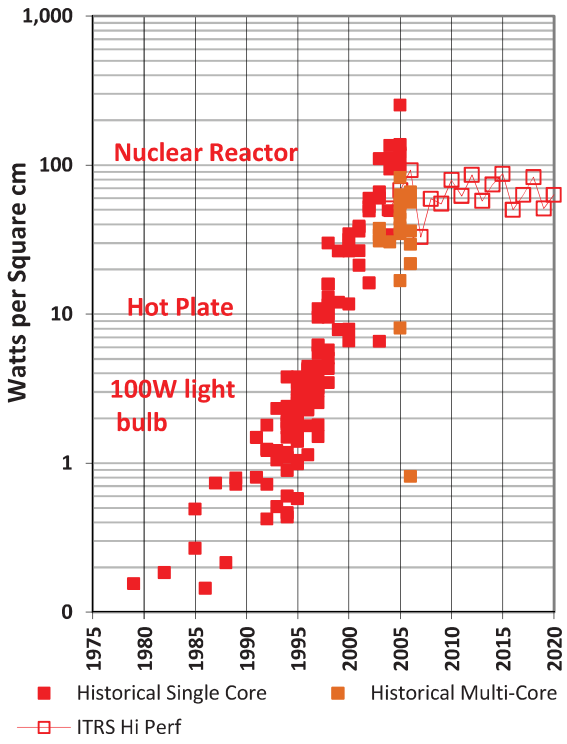
State of Supercomputing in 2013

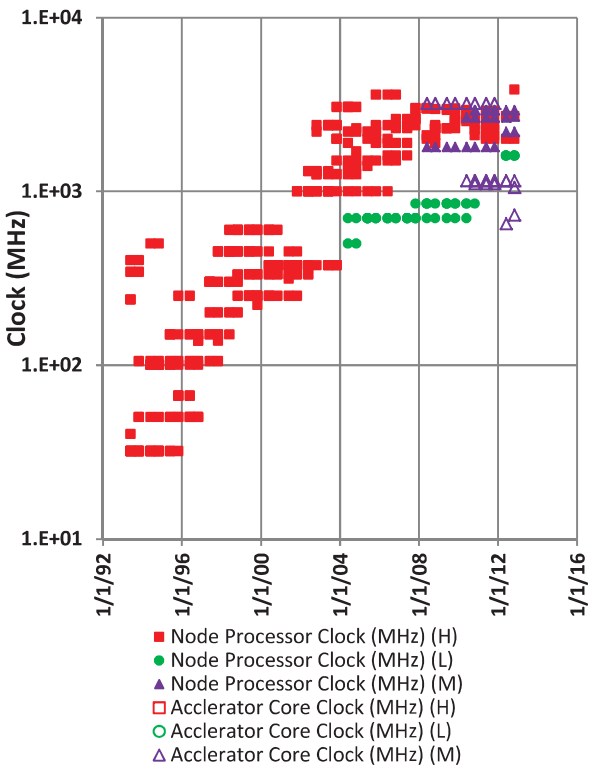
- Pflops computing fully established with more than 20 machines
- Three technology "swim lanes" are thriving
- Interest in supercomputing is now worldwide, and growing in many new markets
- Exascale projects in many countries and regions
- Rapid growth predicted by IDC for the next three years

Why we won't get to Exaflops by 2020



The Power and Clock Inflection Point in 2004

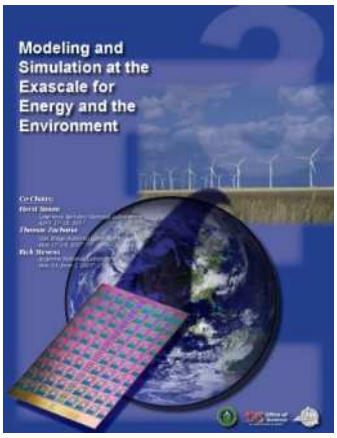


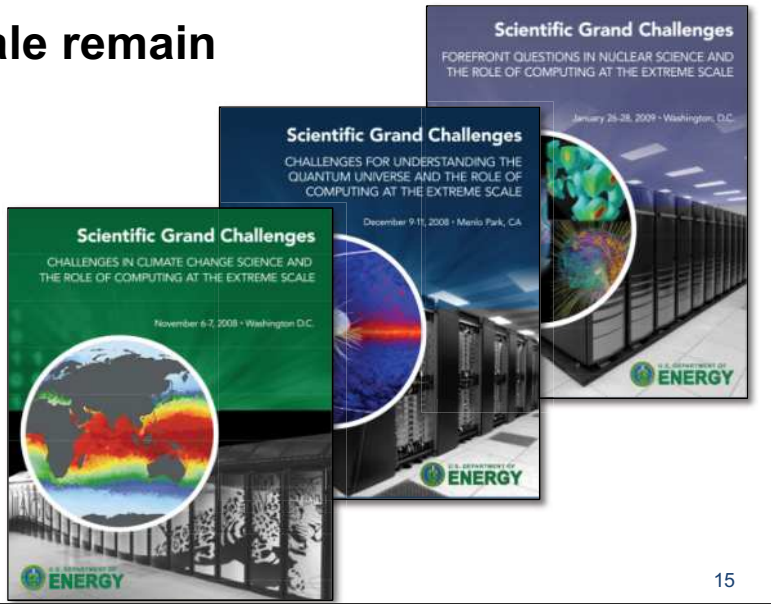


Source: Kogge and Shalf, IEEE CISE

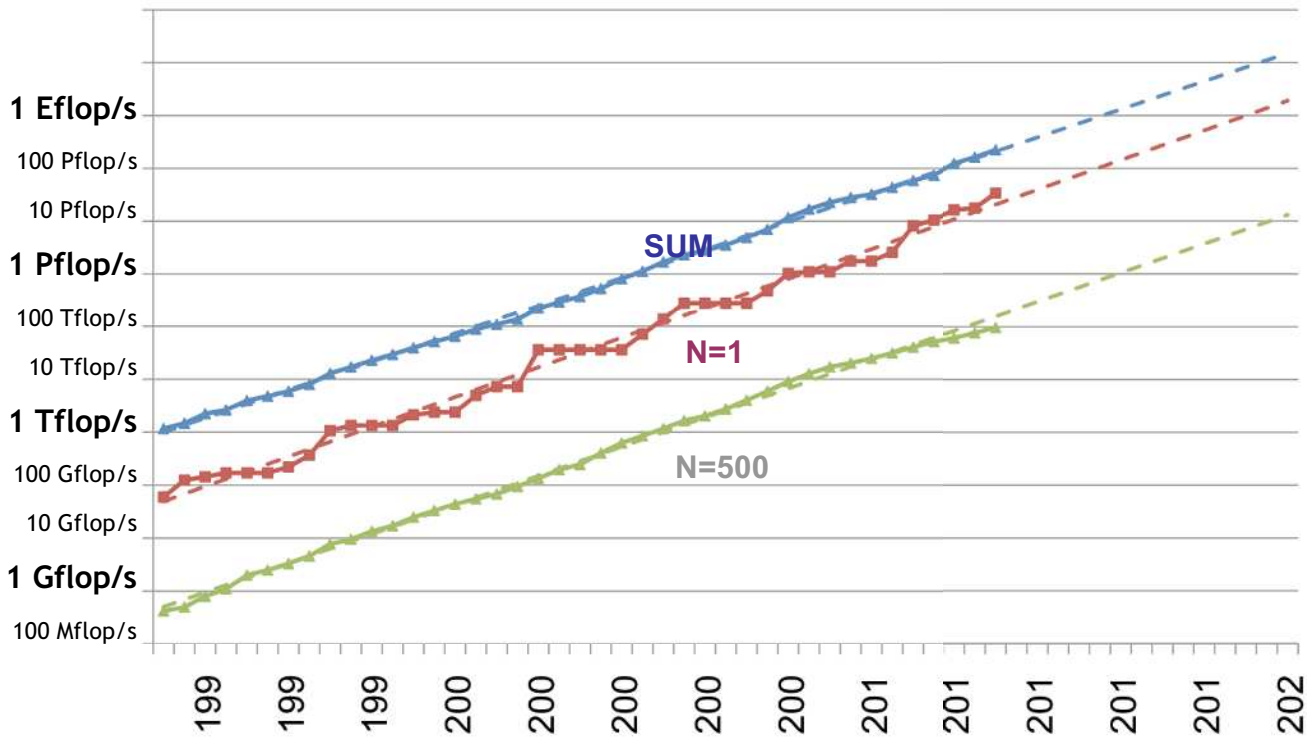
Towards Exascale

- **Exascale has been discussed in numerous workshops, conferences, planning meetings for about six years**
- **Exascale projects have been started in the U.S. and many other countries and regions**
- **Key challenges to exascale remain**





Projected Performance Development



Data from: TOP500 June 2013

The Exascale Challenge

- There is a lot of vagueness in the exascale discussion and what it means to reach exascale. Let me propose a concrete and measurable goal.
- **Build a system before 2020 that will be #1 on the TOP500 list with an Rmax > 1 exaflops**
- Personal bet (with Thomas Lippert, Jülich, Germany) that we will not reach this goal by November 2019 (for $2,000 or €2000)
- Unfortunately, I think that I will win the bet. But I would rather see an exaflops before the end of the decade and loose my bet.

At Exascale, even LINPACK is a Challenge

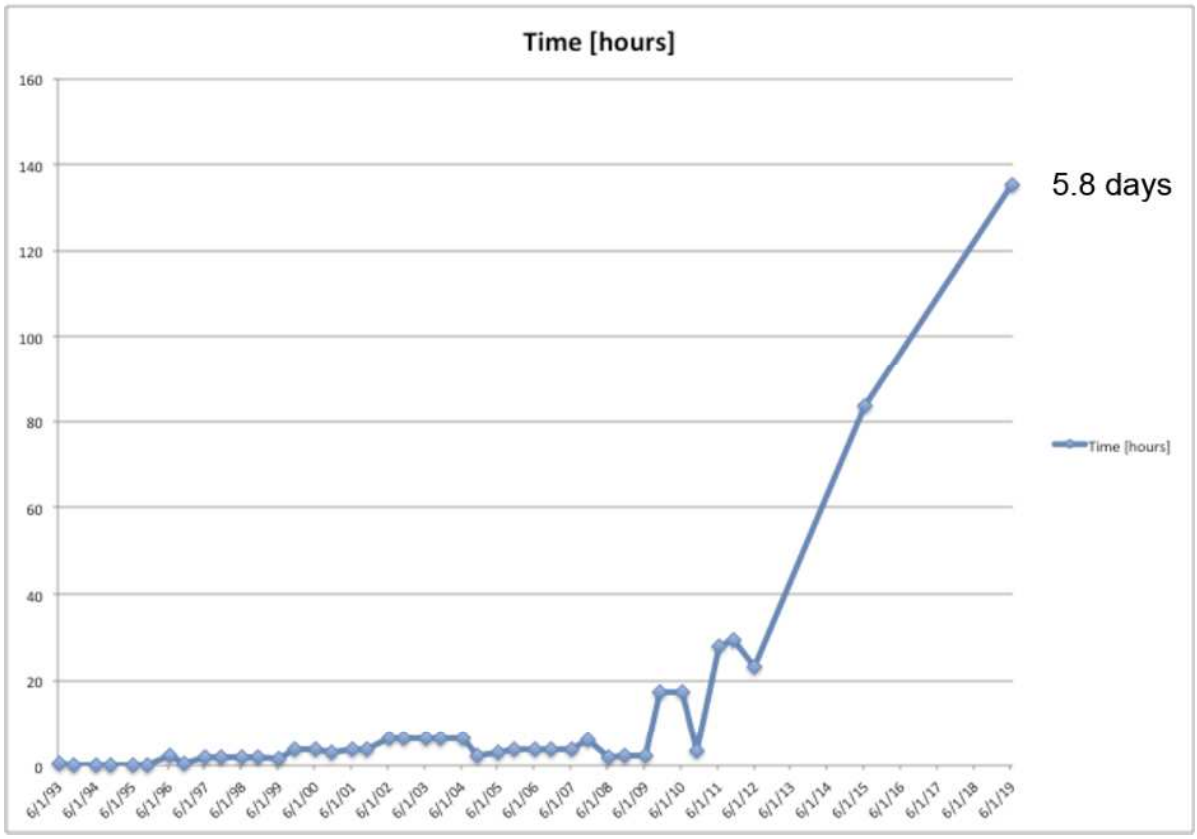
#1 System on the Top500 Over the Past 20 Years
(16 machines in that club)

9  6  2 

Top500 List	Computer	r_max (Tflop/s)	n_max	Hours	MW
6/93 (1)	TMC CM-5/1024	.060	52224	0.4	
11/93 (1)	Fujitsu Numerical Wind Tunnel	.124	31920	0.1	1.
6/94 (1)	Intel XP/S140	.143	55700	0.2	
11/94 - 11/95 (3)	Fujitsu Numerical Wind Tunnel	.170	42000	0.1	1.
6/96 (1)	Hitachi SR2201/1024	.220	138,240	2.2	
11/96 (1)	Hitachi CP-PACS/2048	.368	103,680	0.6	
6/97 - 6/00 (7)	Intel ASCI Red	2.38	362,880	3.7	.85
11/00 - 11/01 (3)	IBM ASCI White, SP Power3 375 MHz	7.23	518,096	3.6	
6/02 - 6/04 (5)	NEC Earth-Simulator	35.9	1,000,000	5.2	6.4
11/04 - 11/07 (7)	IBM BlueGene/L	478.	1,000,000	0.4	1.4
6/08 - 6/09 (3)	IBM Roadrunner -PowerXCell 8i 3.2 Ghz	1,105.	2,329,599	2.1	2.3
11/09 - 6/10 (2)	Cray Jaguar - XT5-HE 2.6 GHz	1,759.	5,474,272	17.3	6.9
11/10 (1)	NUDT Tianhe-1A, X5670 2.93Ghz NVIDIA	2,566.	3,600,000	3.4	4.0
6/11 - 11/11 (2)	Fujitsu K computer, SPARC64 VIIIfx	10,510.	11,870,208	29.5	9.9
6/12 (1)	IBM Sequoia BlueGene/Q	16,324.	12,681,215	23.1	7.9
11/12 (1)	Cray XK7 Titan AMD + NVIDIA Kepler	17,590.	4,423,680	0.9	8.2
6/13 (?)	NUDT Tianhe-2 Intel IvyBridge & Xeon Phi	33,862.	9,960,000	5.4	17.8

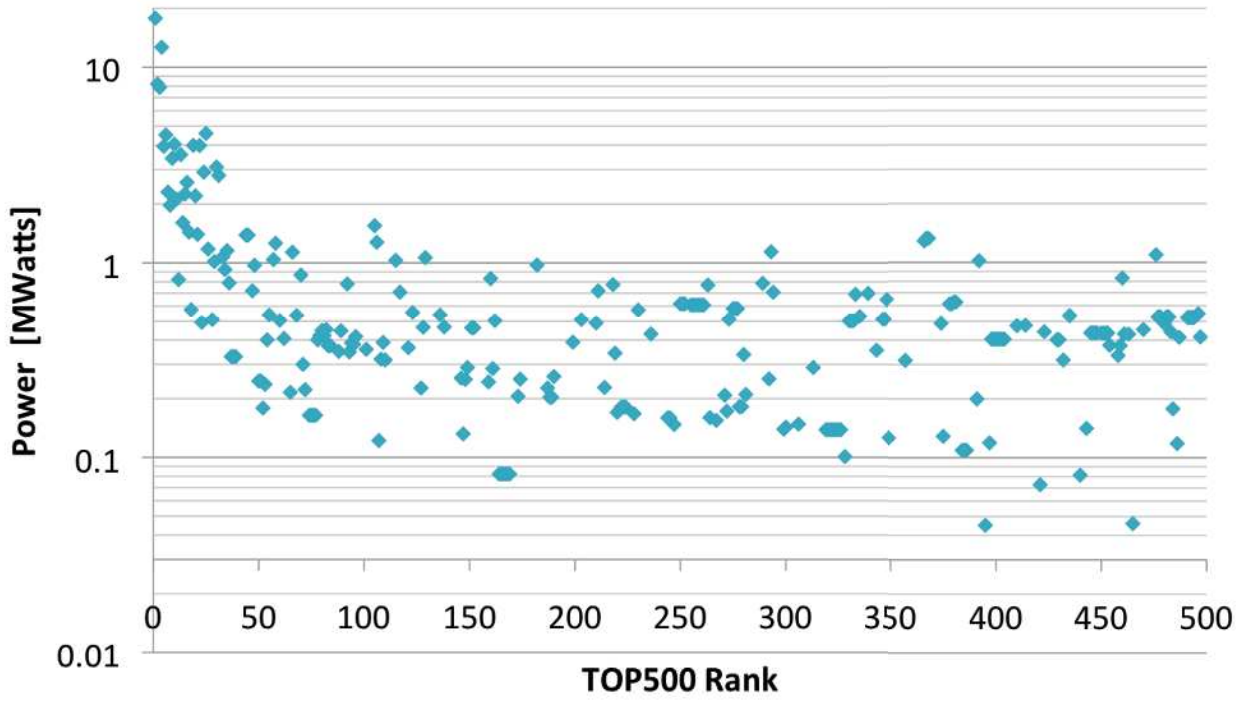
Source: Jack Dongarra http://bit.ly/hpcg-benchmark

At Exascale HPL will take 5.8 days



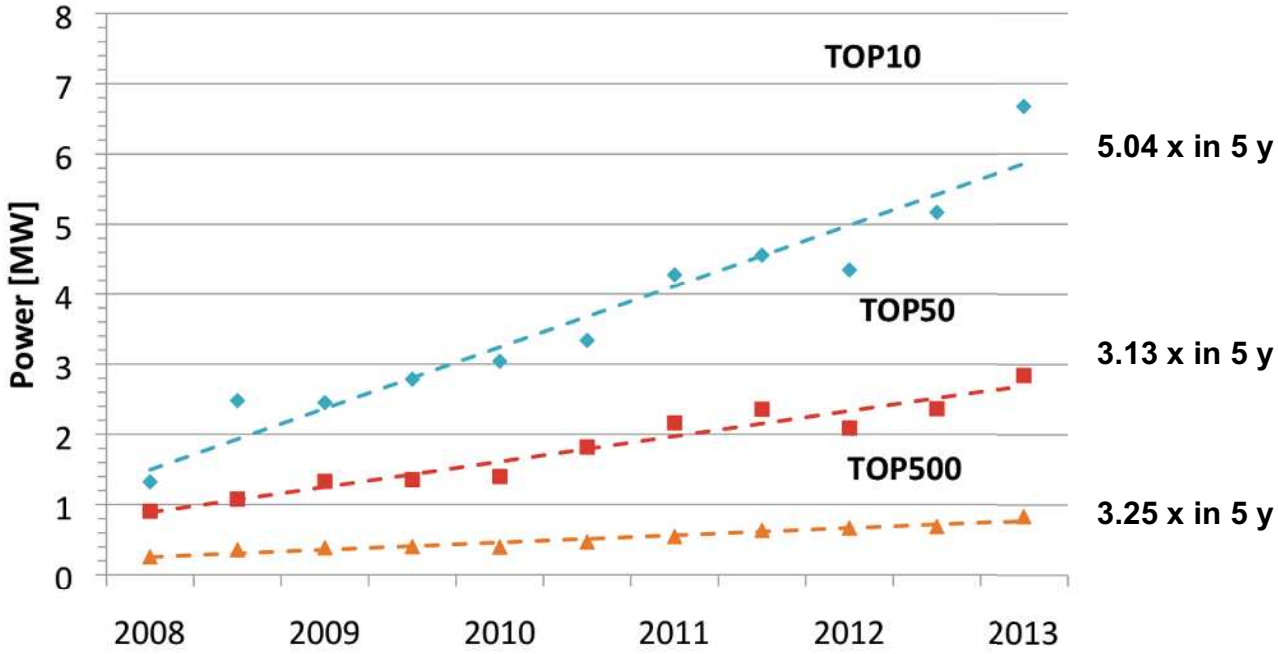
Source: Jack Dongarra, ISC'12

Total Power Levels (MW) for TOP500 Systems



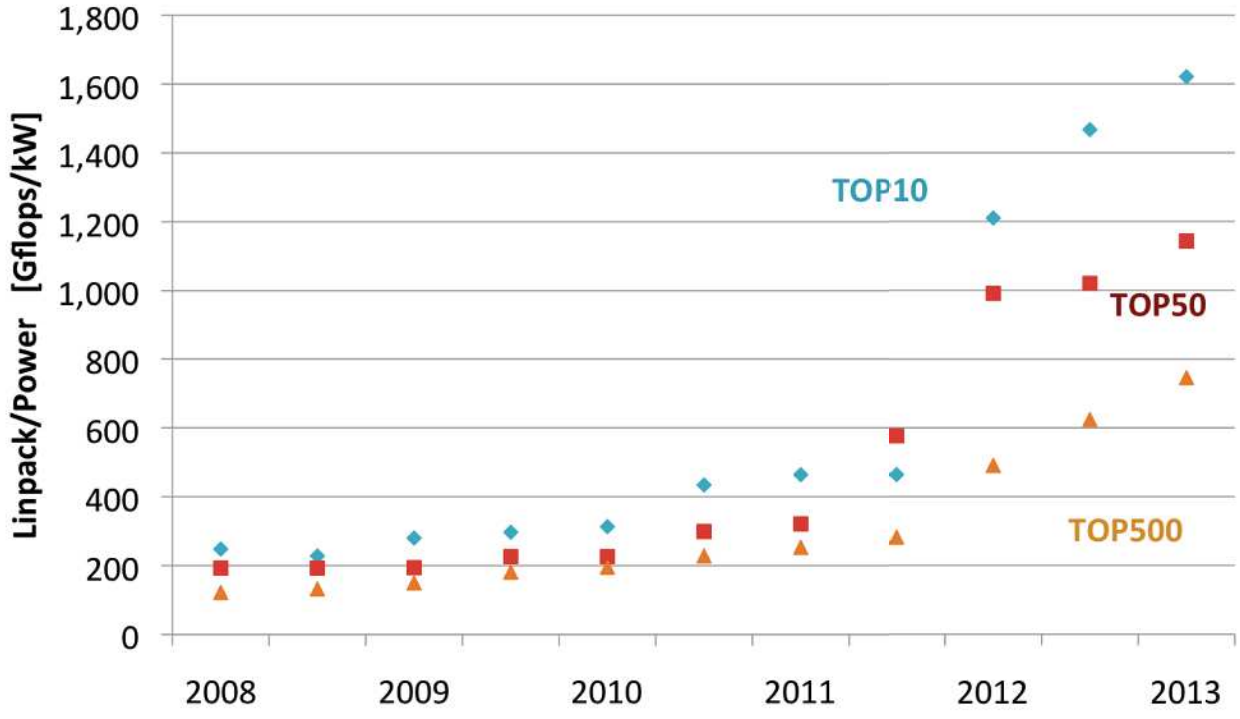
Source: TOP500 June 2013

Power Consumption



Source: TOP500 June 2013

Power Efficiency went up significantly in 2012



Data from: TOP500 June 2013

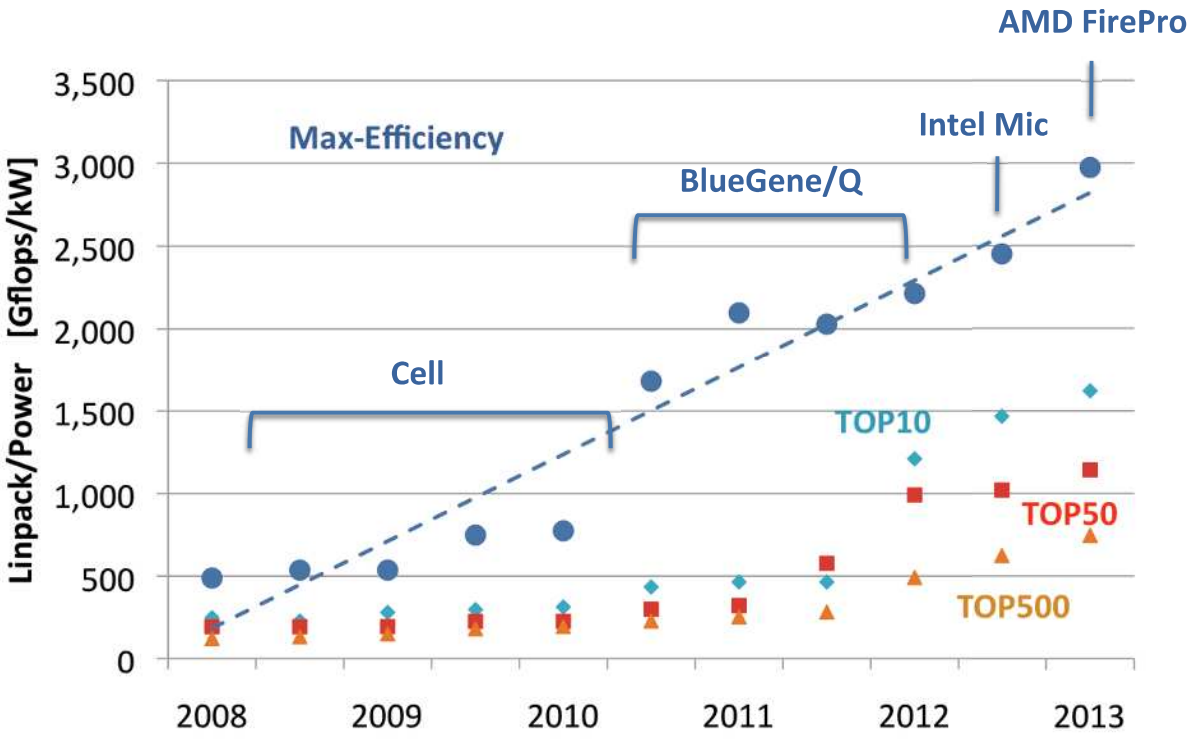
Most Power Efficient Architectures

Computer	Rmax/ Power
Adtech, ASUS, Xeon 8C 2.0GHz, Infiniband FDR, **AMD FirePro**	**2,973**
Appro GreenBlade, Xeon 8C 2.6GHz, Infiniband FDR, **Intel Xeon Phi**	**2,450**
BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	**2,300**
Cray XK7, Opteron 16C 2.1GHz, Gemini, **NVIDIA Kepler**	**2,243**
Eurotech Aurora HPC, Xeon 8C 3.1GHz, Infiniband QDR, **NVIDIA K20**	**2,193**
iDataPlex DX360M4, Xeon 8C 2.6GHz, Infiniband QDR, **Intel Xeon Phi**	**1,935**
Tianhe-2, NUDT, Intel Xeon 6C 2.2GHz, TH Express-2, **Intel Xeon Phi**	**1,902**
RSC Tornado, Xeon 8C 2.9GHz, Infiniband FDR, **Intel Xeon Phi**	**1,687**
SGI Rackable, Xeon 8C 2.6GHz, Infiniband FDR, **Intel Xeon Phi**	**1,613**
Chundoong Cluster, Xeon 8C 2GHz, Infiniband QDR, **AMD Radeon HD**	**1,467**

[Mflops/Watt]

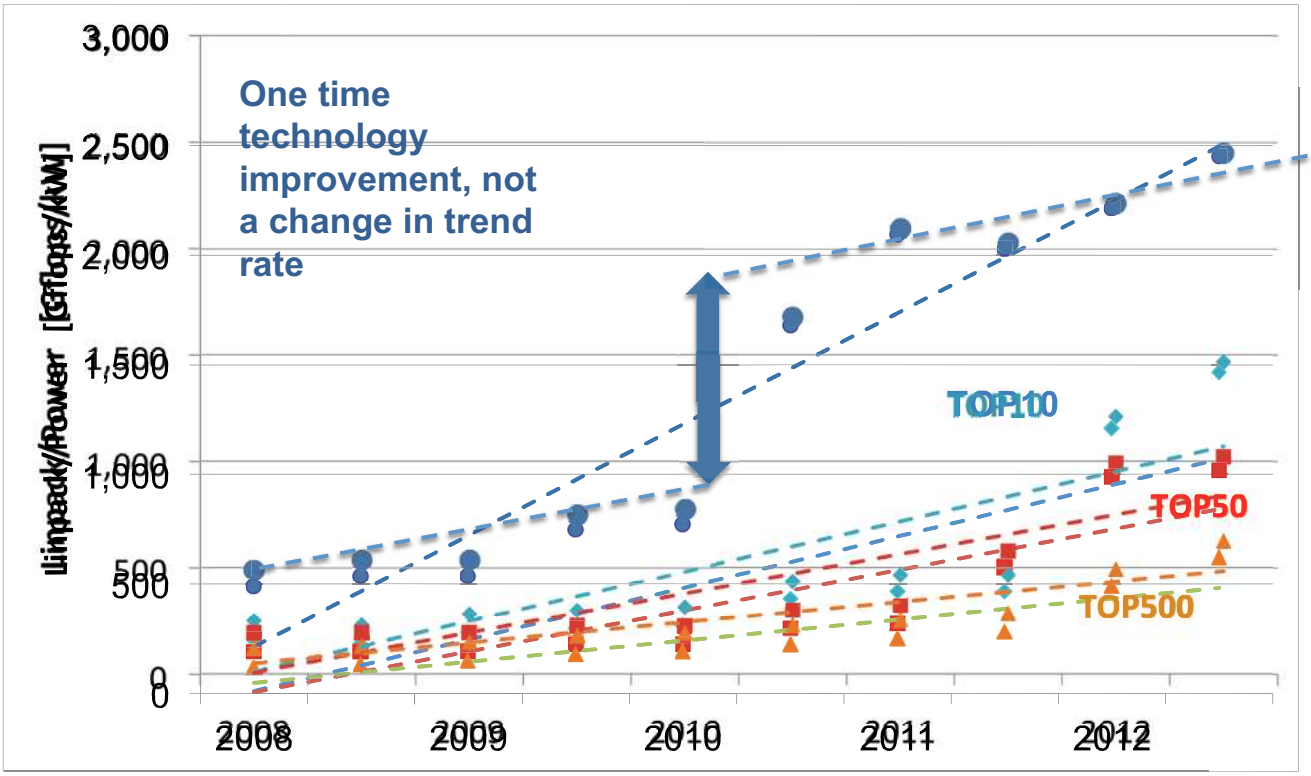
Data from: TOP500 June 2013

Power Efficiency over Time



Data from: TOP500 June 2013

Power Efficiency over Time



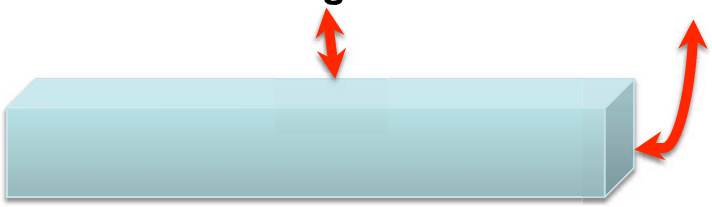
BERKELEY LAB

Data from: TOP500 November 2012

The Problem with Wires:

Energy to move data proportional to distance

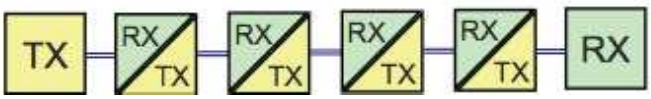
- **Cost to move a bit on copper wire:**
  - **power= bitrate * Length / cross-section area**



- **Wire data capacity constant as feature size shrinks**
- ***Cost to move bit proportional to distance***
- ***~1TByte/sec max feasible off-chip BW (10GHz/pin)***
- ***Photonics reduces distance-dependence of bandwidth***

Photonics requires no redrive and passive switch little power

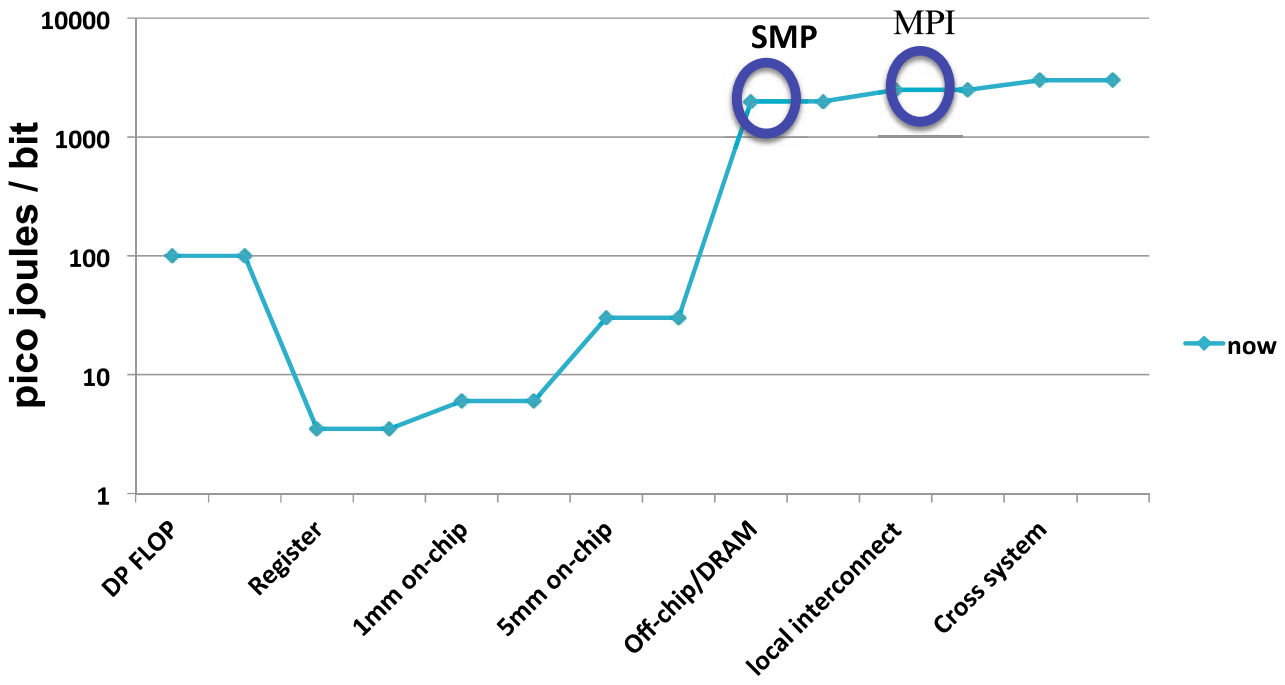
Copper requires to signal amplification even for on-chip connections



BERKELEY LAB

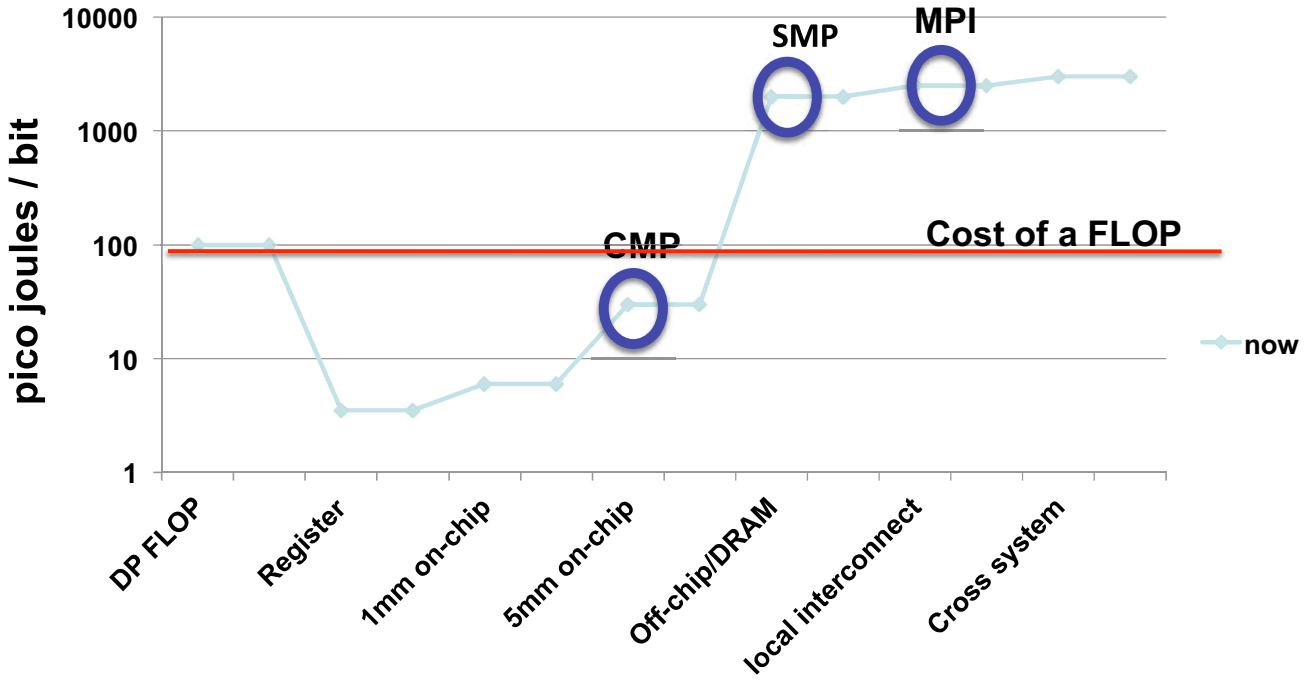
Adapted from John Shalf

The Cost of Data Movement



Adapted from John Shalf

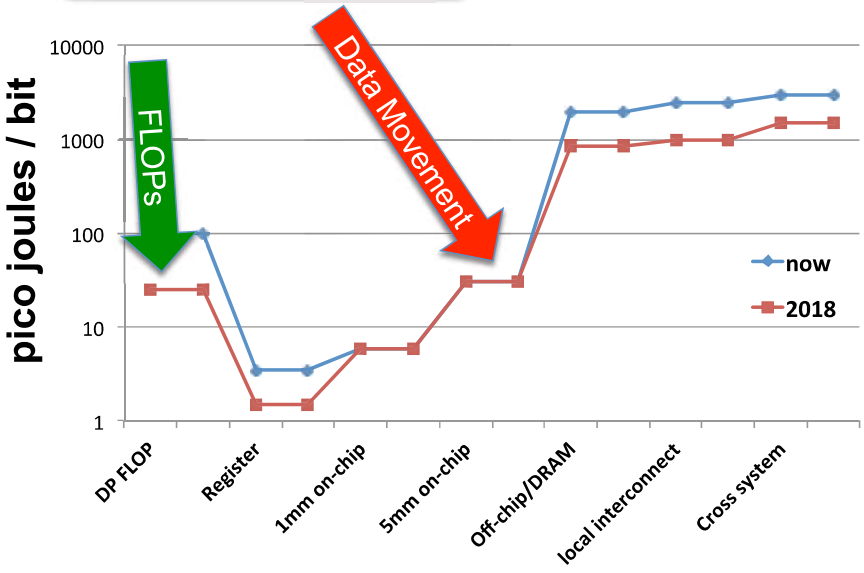
The Cost of Data Movement



Adapted from John Shalf

The Cost of Data Movement in 2018

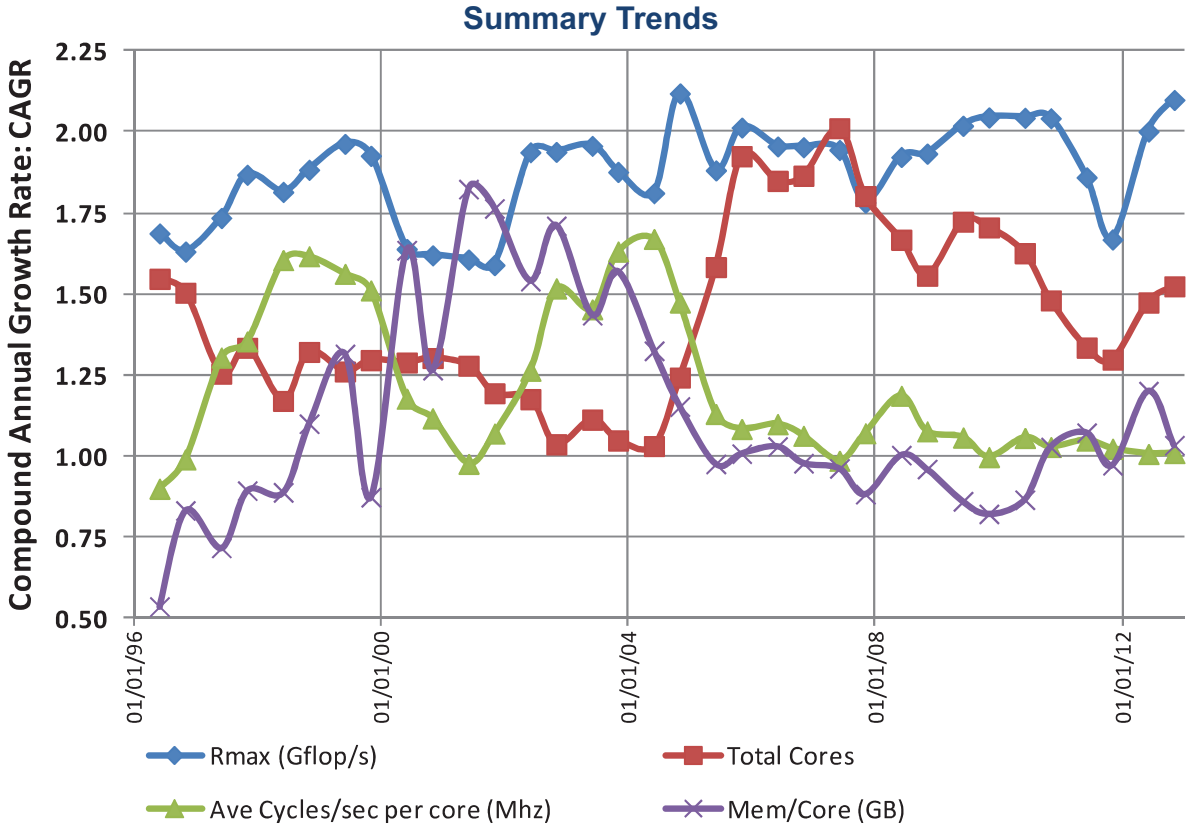
FLOPs will cost less than on-chip data movement! (NUMA)



None of the fast forward projects addresses this challenge

Adapted from John Shalf

It's the End of the World as We Know It!



Source: Kogge and Shalf, IEEE CISE

Why I believe that I will win my bet

1. HPL at Exaflop/s level will take about a week to run challenge for center directors and new systems
2. We realized a one time gain in power efficiency by switching to accelerator/manycore. This is not a sustainable trend in the absence of other new technology.
3. Data movement will cost more than flops (even on chip)
4. Limited amount of memory, low memory/flop ratios (processing is free)

The Logical Conclusion

If FLOPS are free, then why do we need an "exaflops" initiative?

"Exa"-anything has become a bad brand

- Associated with buying big machines for the labs
- Associated with "old" HPC
- Sets up the community for "failure" if "goal" can't be met

It is not just "exaflops" – we are changing the whole computational model

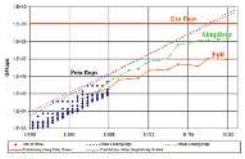
Current programming systems have WRONG optimization targets

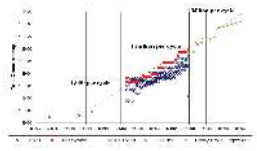
Old Constraints

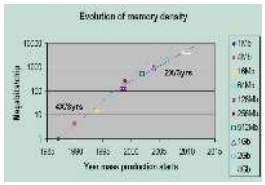
- **Peak clock frequency** *as primary limiter for performance improvement*
- **Cost:** *FLOPs* are biggest cost for system: *optimize for compute*
- **Concurrency**: *Modest growth of parallelism by adding nodes*
- **Memory scaling***: maintain byte per flop capacity and bandwidth*
- **Locality***: MPI+X model (uniform costs within node & between nodes)*
- **Uniformity**: *Assume uniform system performance*
- **Reliability***: It's the hardware's problem*

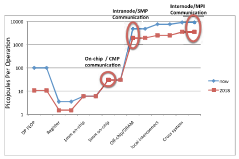
New Constraints

- **Power** *is primary design constraint for future HPC system design*
- **Cost:** *Data movement dominates: optimize to minimize data movement*
- **Concurrency:** *Exponential growth of parallelism within chips*
- **Memory Scaling:** *Compute growing 2x faster than capacity or bandwidth*
- **Locality***: must reason about data locality and possibly topology*
- **Heterogeneity**: *Architectural and performance non-uniformity increase*
- **Reliability***: Cannot count on hardware protection alone*





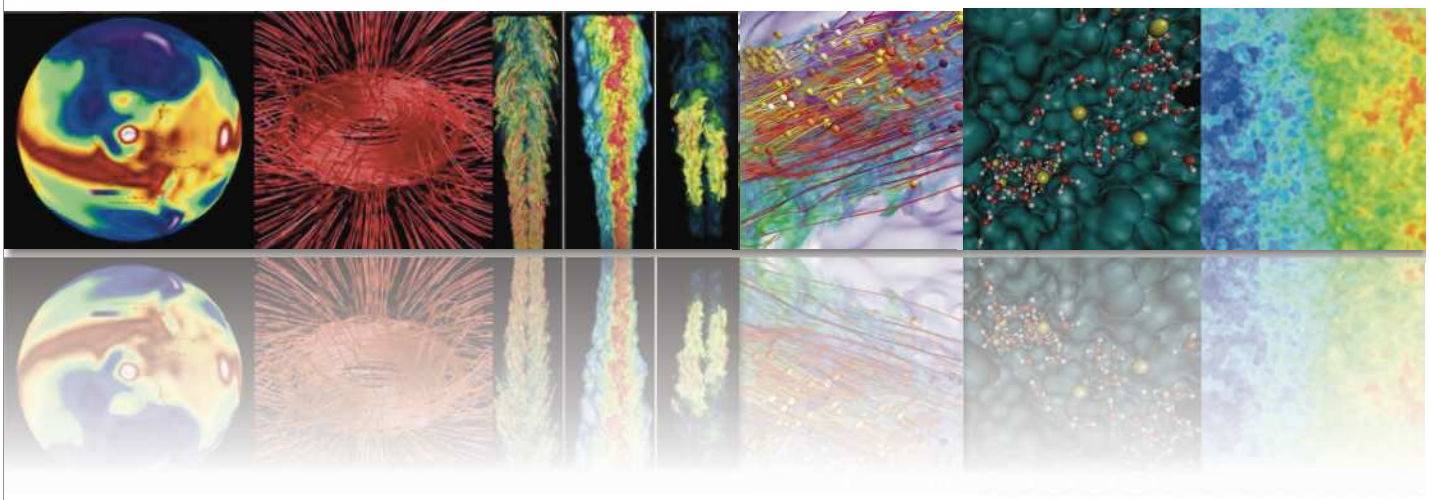




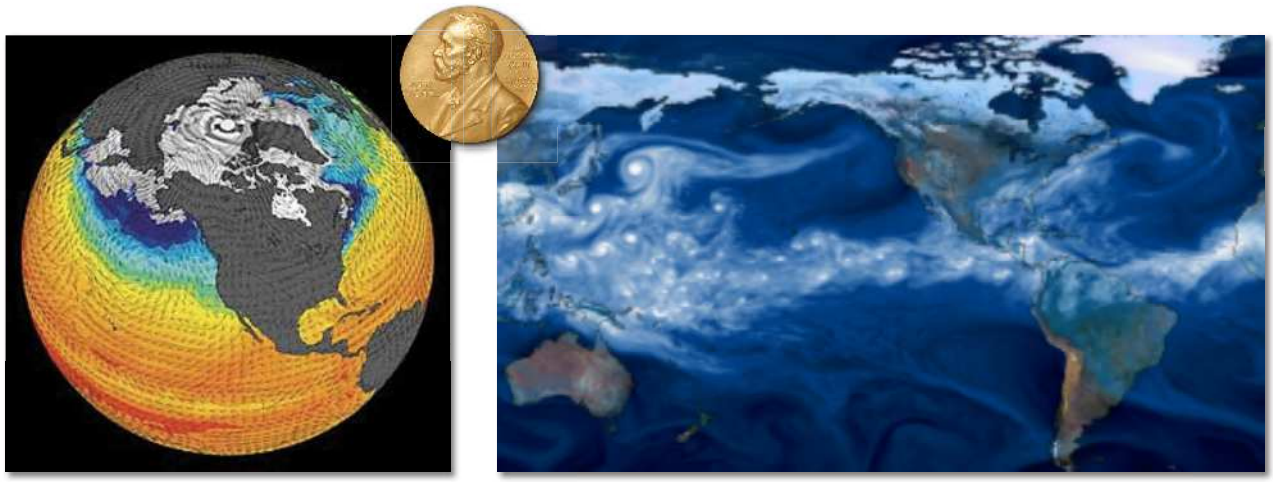
Fundamentally breaks our current programming paradigm and computing ecosystem

Adapted from John Shalf

The Science Case for Exascale (old)



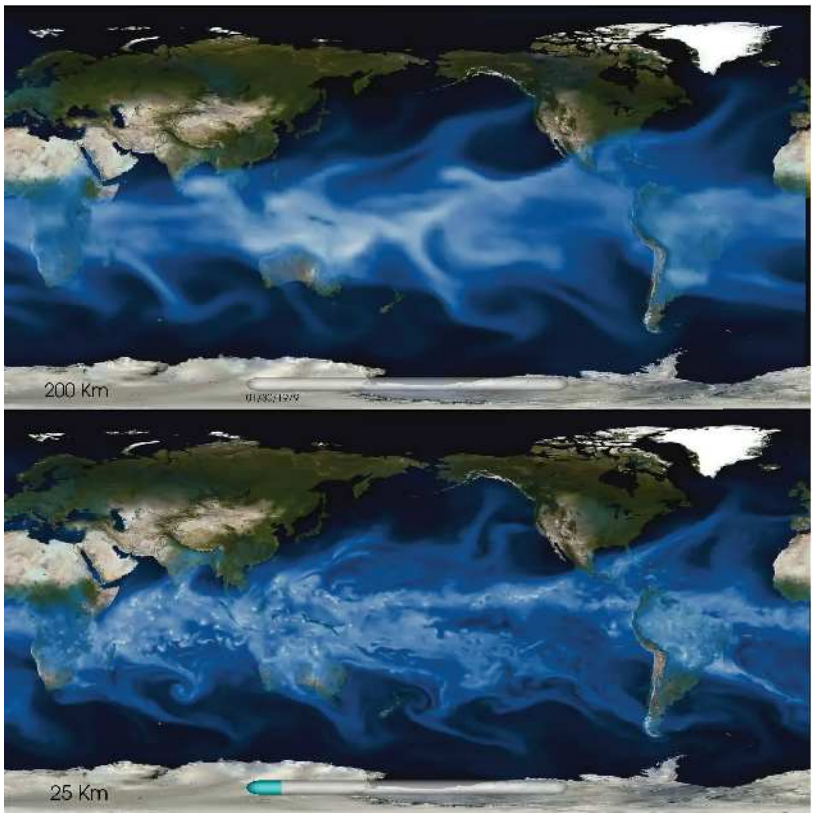
Climate change analysis



Simulations	Extreme data
• Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms • New math, models, and systems support will be needed	• "Reanalysis" projects need 100× more computing to analyze observations • Machine learning and other analytics are needed today for petabyte data sets • Combined simulation/observation will empower policy makers and scientists

BERKELEY LAB

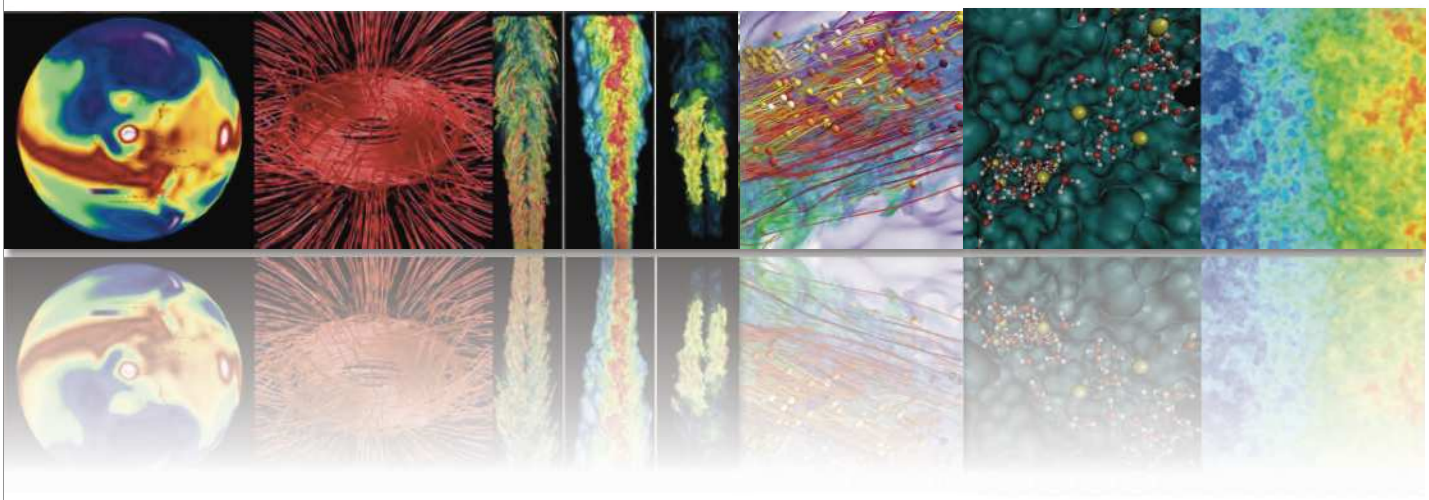
Qualitative Improvement of Simulation with Higher Resolution (2011)



BERKELEY LAB

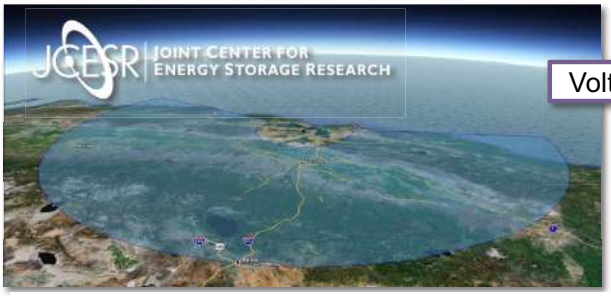
Michael Wehner, Prabhat, Chris Algieri, Fuyu Li, Bill Collins, Lawrence Berkeley National Laboratory; Kevin Reed, University of Michigan; Andrew Gettelman, Julio Bacmeister, Richard Neale, National Center for Atmospheric Research

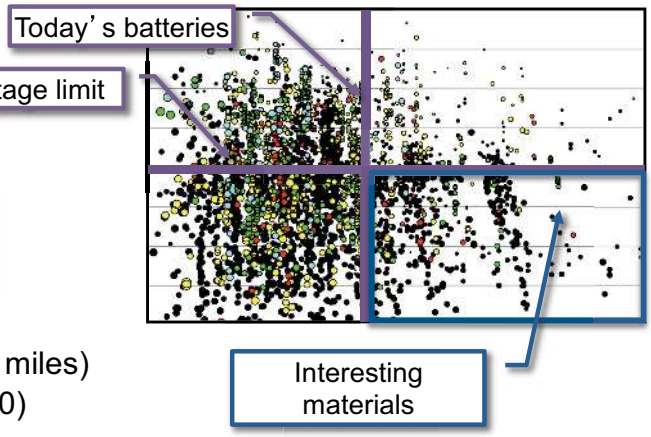
The Science Case for Exascale (new)



Materials Genome

Computing 1000× today	Data services for industry and science
• Key to DOE's Energy Storage Hub • Tens of thousands of simulations used to screen potential materials • Need more simulations and fidelity for new classes of materials, studies in extreme environments, etc.	• Results from tens of thousands of simulations web-searchable • Materials Project launched in October 2012, now has >3,000 registered users • Increase U.S. competitiveness; cut in half 18 year time from discovery to market

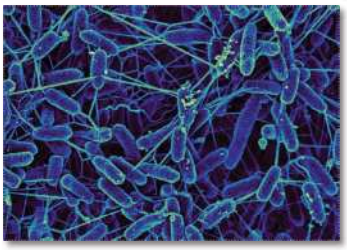




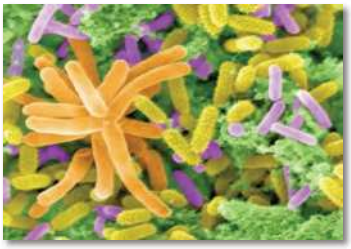
By 2018:
•Increase energy density (70 miles → 350 miles)
•Reduce battery cost per mile ($150 → $30)

DOE Systems Biology Knowledgebase: KBase

- **Integration and modeling for predictive biology**
- **Knowledgebase enabling predictive systems biology**
  - Powerful modeling framework
  - Community-driven, extensible and scalable open-source software and application system
  - Infrastructure for integration and reconciliation of algorithms and data sources
  - Framework for standardization, search, and association of data
  - Resource to enable experimental design and interpretation of results



Microbes



Communities

Plants

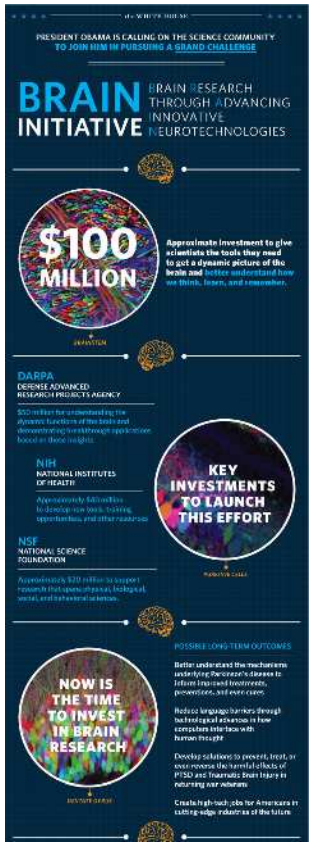
KBase
PREDICTIVE BIOLOGY

BERKELEY LAB

Obama Announces BRAIN Initiative – Proposes $100 Million in his FY2014 Budget

- Brain Research through Advancing Innovative Neurotechnologies
- Create real-time traffic maps to provide new insights into brain disorders
- "There is this enormous mystery waiting to be unlocked, and the BRAIN Initiative will change that by giving scientists the tools they need to get a dynamic picture of the brain in action and better understand how we think and how we learn and how we remember,"

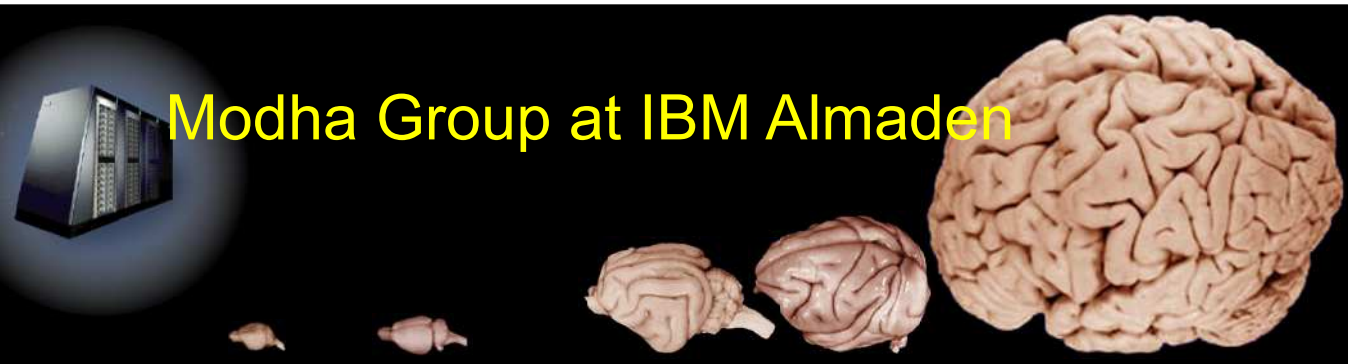
  Barack Obama



BERKELEY LAB

See Alivisatos et al., Neuron 74, June 2012, pg 970

Modha Group at IBM Almaden



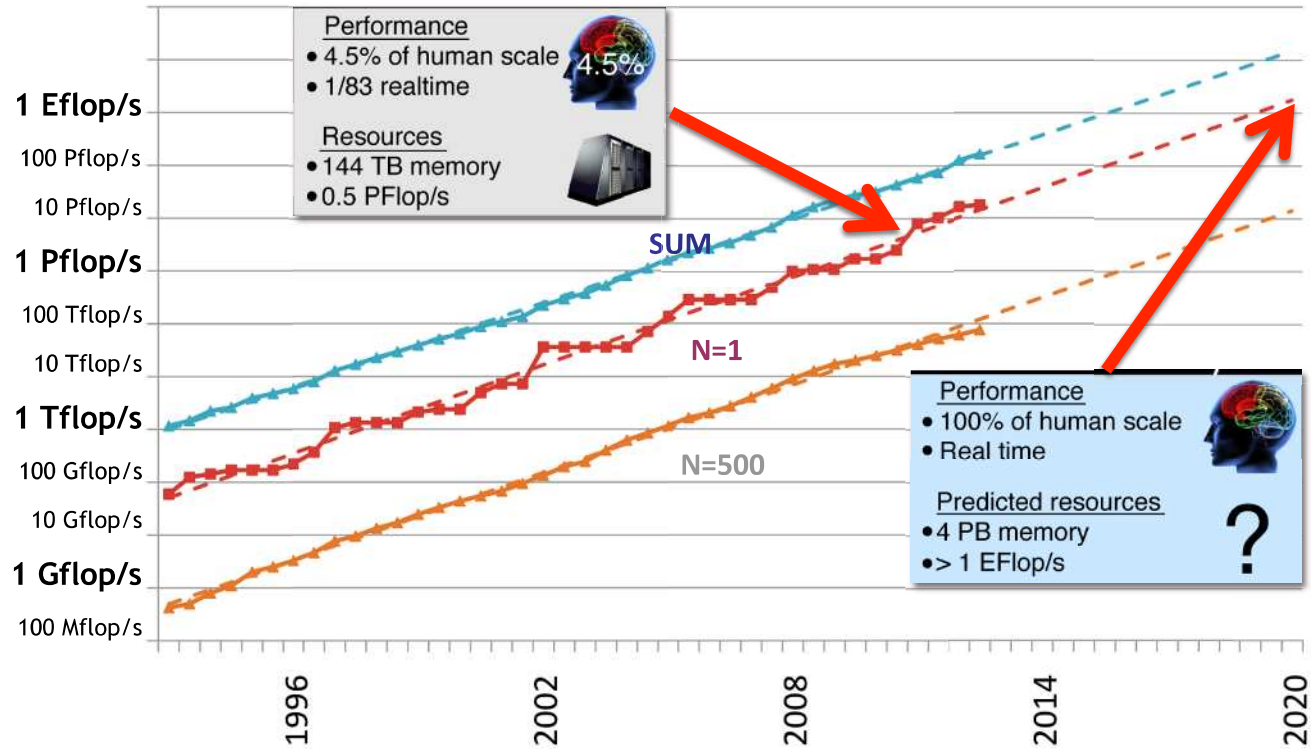
	Mouse	Rat	Cat	Monkey	Human
N:	16 x 10^6	56 x 10^6	763 x 10^6	2 x 10^9	22 x 10^9
S:	128 x 10^9	448 x 10^9	6.1 x 10^{12}	20 x 10^{12}	220 x 10^{12}

New results for SC12

Almaden	Watson	WatsonShaheen	LLNL Dawn	LLNL Sequoia
BG/L	BG/L	BG/P	BG/P	BG/Q
December, 2006	April, 2007	March, 2009	May, 2009	June, 2012

Latest simulations in 2012 achieve unprecedented scale of 65*10^9 neurons and 16*10^{12} synapses

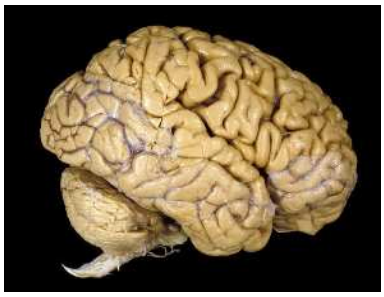
Towards Exascale



Data: TOP500 November 2012

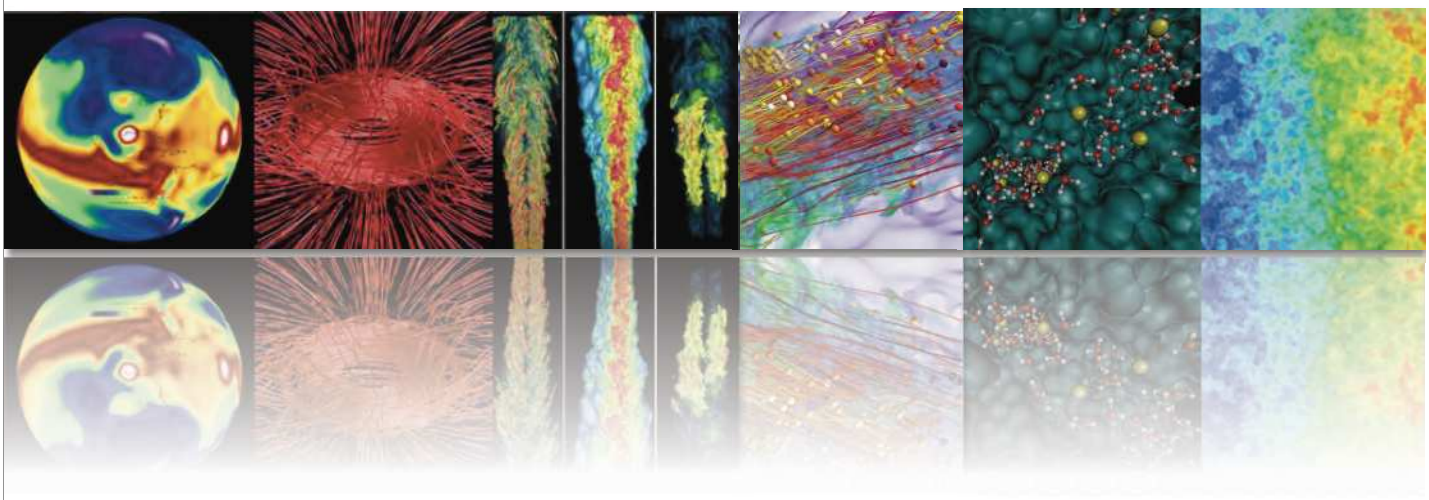
Towards Exascale – the Power Conundrum

- Straight forward extrapolation results in a real time human brain scale simulation at about 1 - 10 Exaflop/s with 4 PB of memory
- Current predictions envision Exascale computers in 2020 with a power consumption of at best 20 - 30 MW
- The human brain takes 20W
- Even under best assumptions in 2020 our brain will still be a million times more power efficient



BERKELEY LAB

Exascale is Critical to the Nation to Develop Future Technologies and Maintain Economic Competitiveness



BERKELEY LAB

U.S. competitive advantage demands exascale resources

- Digital design and prototyping at exascale enable rapid delivery of new products to market by minimizing the need for expensive, dangerous, and/or inaccessible testing
- Potential key differentiator for American competitiveness
- Strategic partnerships between DOE labs and industrial partners to develop and scale applications to exascale levels

BERKELEY LAB

Exascale computing is key for national security

"I think that what we've seen is that they may be somewhat further ahead in the development of that aircraft than our intelligence had earlier predicted."

—Defense Secretary Robert M. Gates,
(*The New York Times*, Jan. 9, 2011)

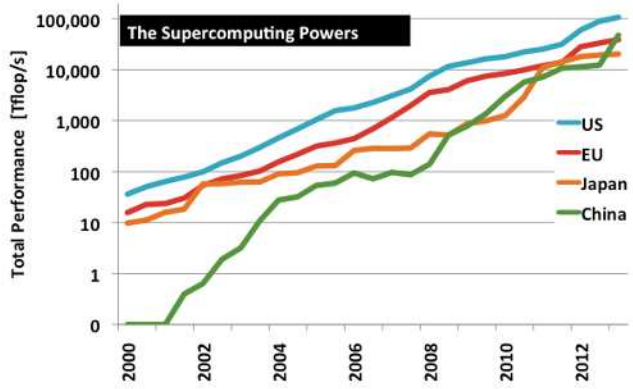
Potential adversaries are not unaware of this

BERKELEY LAB

Exascale technologies are the foundation for future leadership in computing

The U.S. is not the only country capable of achieving exascale computing. The country that is first to exascale will have significant competitive intellectual, technological, and economic advantages.

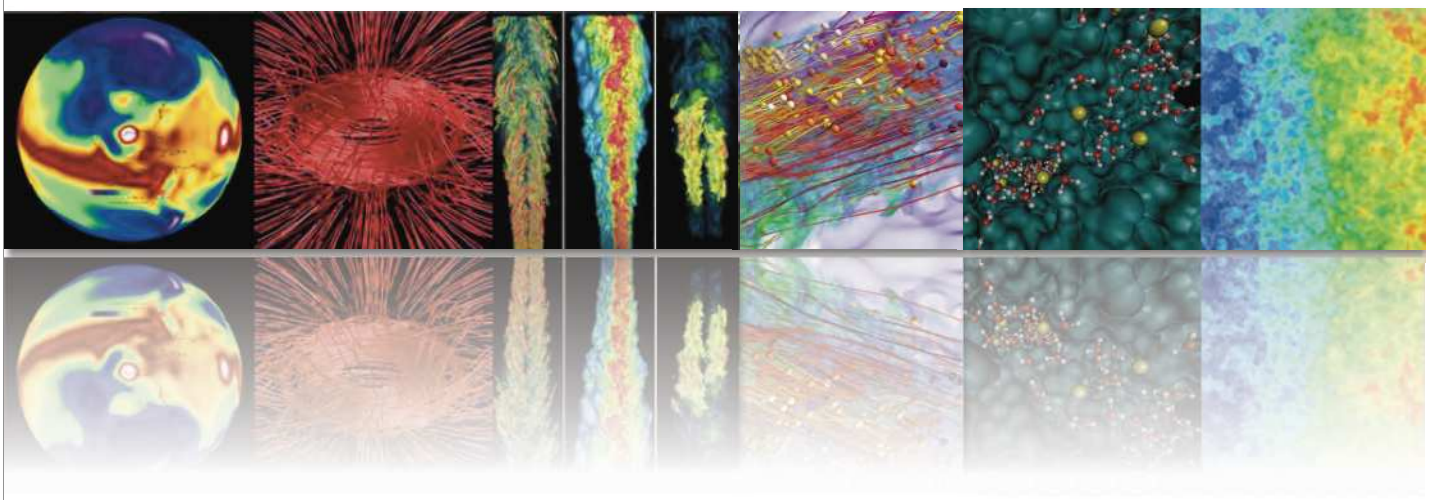
Achievement of the power efficiency and reliability goals needed for exascale will have enormous positive impacts on consumer electronics and business information technologies and facilities.



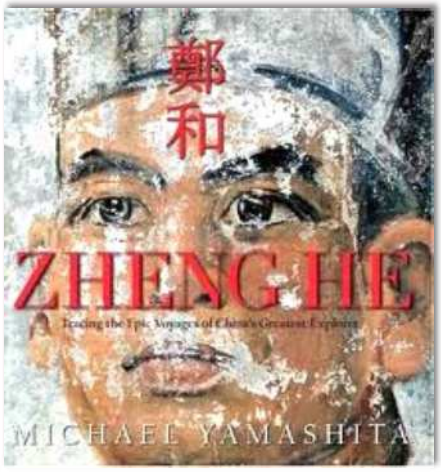
The gap in available supercomputing capacity between the United States and the rest of the world has narrowed, with China gaining the most ground.

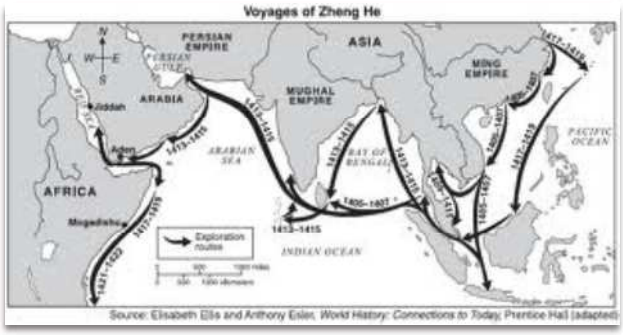
Data from TOP500.org; Figure from Science; Vol 335; 27 January 2012

Exascale is Discovery



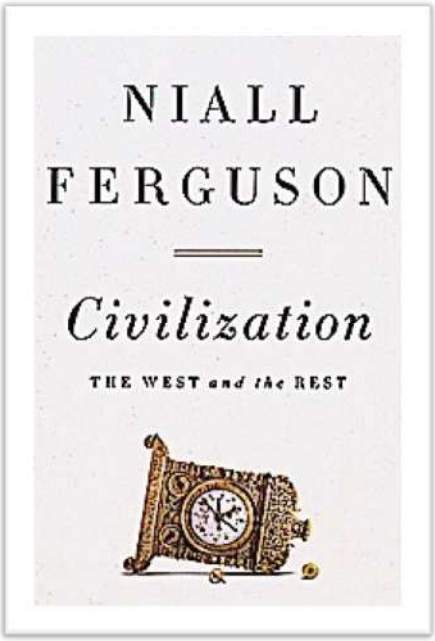
Admiral Zheng He's Voyages in 1416





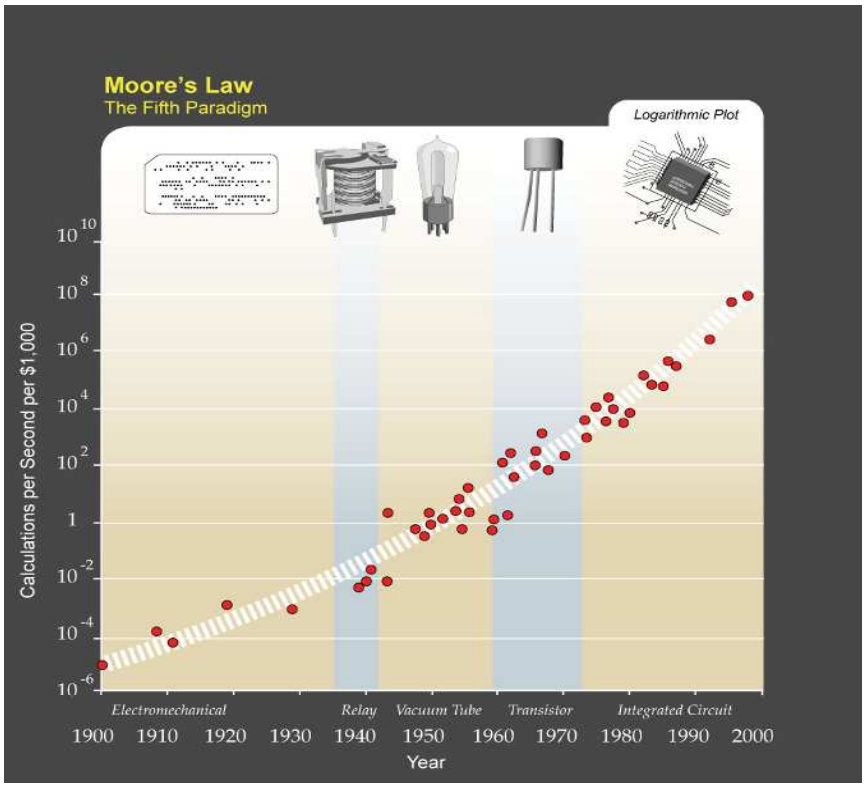
http://en.wikipedia.org/wiki/Zheng_He

The Rise of the West

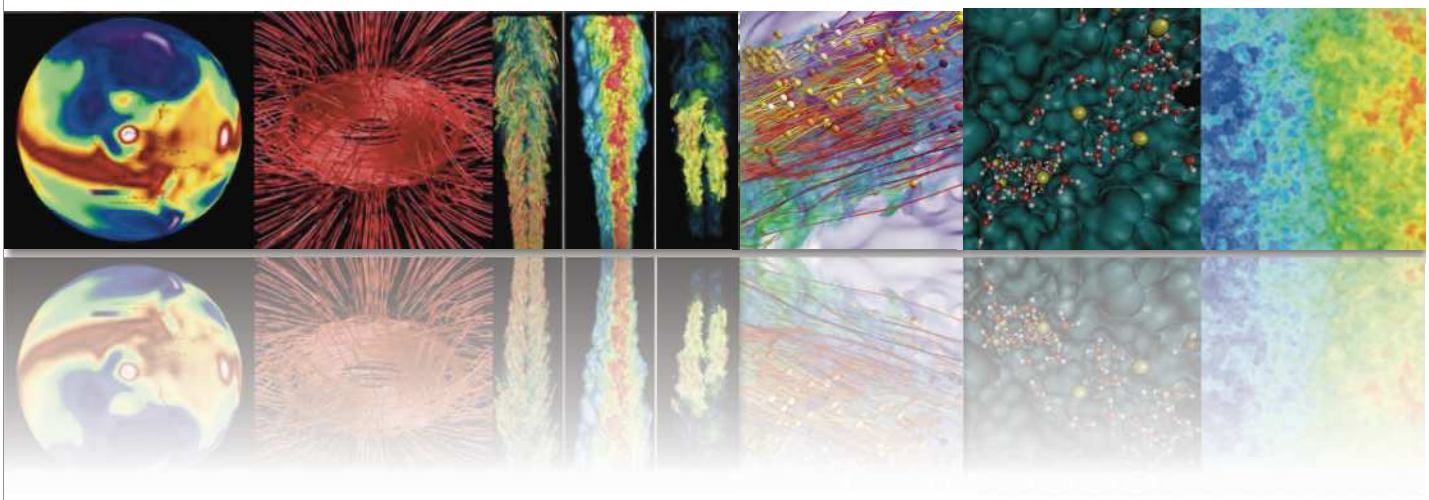


"Like the Apollo moon missions, Zheng He's voyages had been a formidable demonstration of wealth and technological sophistication. Landing … on the East African coast in 1416 was in many ways … comparable to landing an American astronaut on the moon in 1969. By abruptly canceling oceanic exploration, Yongle's successors ensured that the economic benefits remained negligible."

Exascale is the next step in a long voyage of discovery, we should not give up now and retreat



Discussion



Exascale rationale

- Exascale computing means performing computer simulations on a machine capable of 10 to the 18th power floating point operations per second – ~1,000× the current state-of-the-art.
- The U.S. is capable of reaching exascale simulation capability within ten years with cost-effective, efficient, computers. An accelerated program to achieve exascale is required to meet this goal, and doing so will enable the U.S. to lead in key areas of science and engineering.
- The U.S. is not the only country capable of achieving exascale computing. The country that is first to exascale will have significant competitive intellectual, technological, and economic advantages.
- Exascale systems and applications will improve predictive understanding in complex systems (e.g. climate, energy, materials, nuclear stockpile, emerging national security threats, biology, …)
- Tightly coupled exascale systems make possible fundamentally new approaches to uncertainty quantification and data analytics – uncertainty quantification is the essence of predictive simulation.
- American industry is increasingly dependent on HPC for design and testing in a race to get products to market. Exascale computing and partnerships with DOE laboratories could be THE key differentiator, a winning U.S. strategy for advanced manufacturing and job creation.
- The scale of computing developed for modeling and simulation will have a revolutionary impact on data analysis and data modeling. Much of the R&D required is common to big data and exascale computing.
- Exascale provides a ten year goal to drive modeling and simulation developments forward and to attract young scientific talent to DOE in a variety of scientific and technological fields of critical importance to the future of the U.S.
- Achievement of the power efficiency and reliability goals needed for exascale will have enormous positive impacts on consumer electronics and business information technologies and facilities.

Summary

- **There is progress in Exascale with many projects now focused and on their way, e.g. FastForward, Xstack, and Co-Design Centers in the U.S.**
- **HPC has moved to low power processing, and the processor growth curves in energy-efficiency could get us in the range of exascale feasibility**
- **Memory and data movement are still more open challenges**
- **Programming model needs to address heterogeneous, massive parallel environment, as well as data locality**
- **Exascale applications will be a challenge just because of their sheer size and the memory limitations**