スパコンの開発動向と ベンチマークの光と影

佐藤　三久

理化学研究所　計算科学研究機構

もくじ

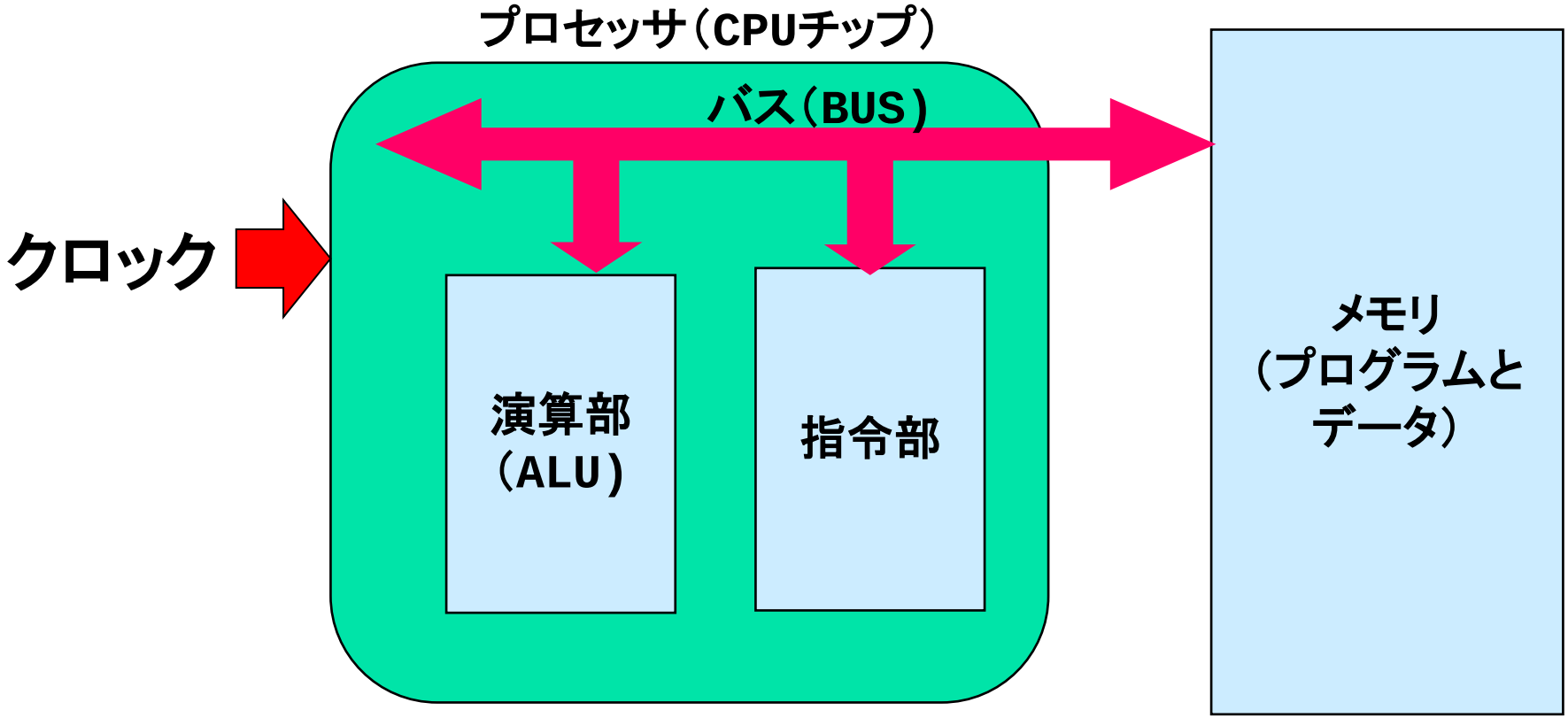
- コンピュータを早くするためには、…
- ベンチマーク・性能評価の目的
- 並列ベンチマークの現状と問題点
- スパコンの開発動向

コンピュータを速くするには、…

- ①　動作を速くする。
  - クロックを速くする（PCのプロセッサは2～3GHzの周波数）
  - 速いトランジスタ（回路）をつかう

コンピュータの基本的な構成

- コンピュータのもっとも基本的な要素は、メモリとプロセッサ（CPU)
  - メモリはプログラムやデータを格納する場所
  - クロックといわれる信号を基準にして、プロセッサはそのメモリからプログラムやデータを読み出して、プログラムを実行しています。
    - 指令する部分：プログラムを解釈（？）して指令する
    - 演算する部分：足し算や掛け算をする部分



半導体プロセスとクロック速度

- 半導体プロセスの向上
  - ムーアの法則、半導体の集積度は18ヶ月で2倍になる
  - 多くのトランジスタをつかうことができる
    ⇒いろいろな機能を盛り込む
  - Pentium4は、0.13μmプロセス ⇒ 90nm、今は20 nm
- クロック速度の向上
  - 明らか。(Pentium4は3GHz！)
  - 電圧を下げる (5Vから3V, さらに1V以下に)
  - 線幅が小さくなると回路を駆動する電流は小さくてすむが、電子の移動速度が遅くなる。

いろいろなマイクロプロセッサ

- マイコン（４ビットマイコン）
  - ４００４(世界初、1971年、750KHz)
- ８ビットマイコン
  - ８００８(1972年、500KHz、インテル)
  - ８０８０(1974年、2MHz、インテル)
  - ｚ８０(1976年、10?MHz、ザイログ)
  - MC6800（1974年、1MHz、モトローラ）
  - MC6809
- 16ビットマイコン
  - ８０８６（1978年、インテル）
    - IBM PC/MS-DOS
  - ８０２８６(1982年、インテル）
  - MC68000 （1979年、モトローラ）
    - UNIX
- ８ビット、１６ビットとは、バスの幅、メモリ空間のビット幅のこと。

- 32ビットプロセッサ
  - 80386 (1985年）、80486 (1989年、40MHz～)
  - MC68020(1984年）、MC68030 (1987年)
    - 仮想記憶
  - Pentium (1995年,100MHz～200MHz)
  - Pentium II (1998年,300MHz～)
    - SSE/MMX
  - Pentium III (1997年、900MHz～)
    - 1GHzを超える
  - Pentium 4 (2000年、～3.2GHz)
  - AMD K9, AMD Athlon
- 64ビットプロセッサ
  - Itanium(2000), Itanium II (2001)
  - AMD Opteron (2003)

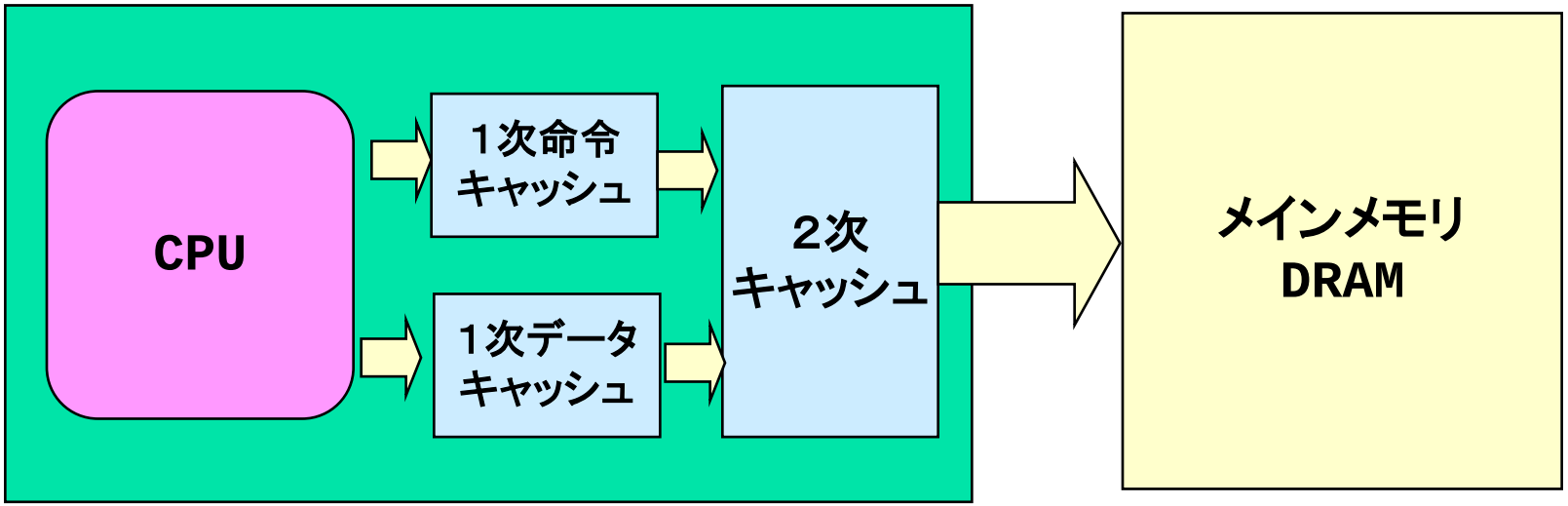
- 30年間で、1MHzから1GHz、1000倍の進歩

コンピュータを速くするには、…

- ②　コンピュータの中を工夫する
  - データのアクセスが速くなるように工夫する
  - 一度に、多くの命令を実行できるようにする
  - よく使われる演算を高速化する命令・仕組みを入れる

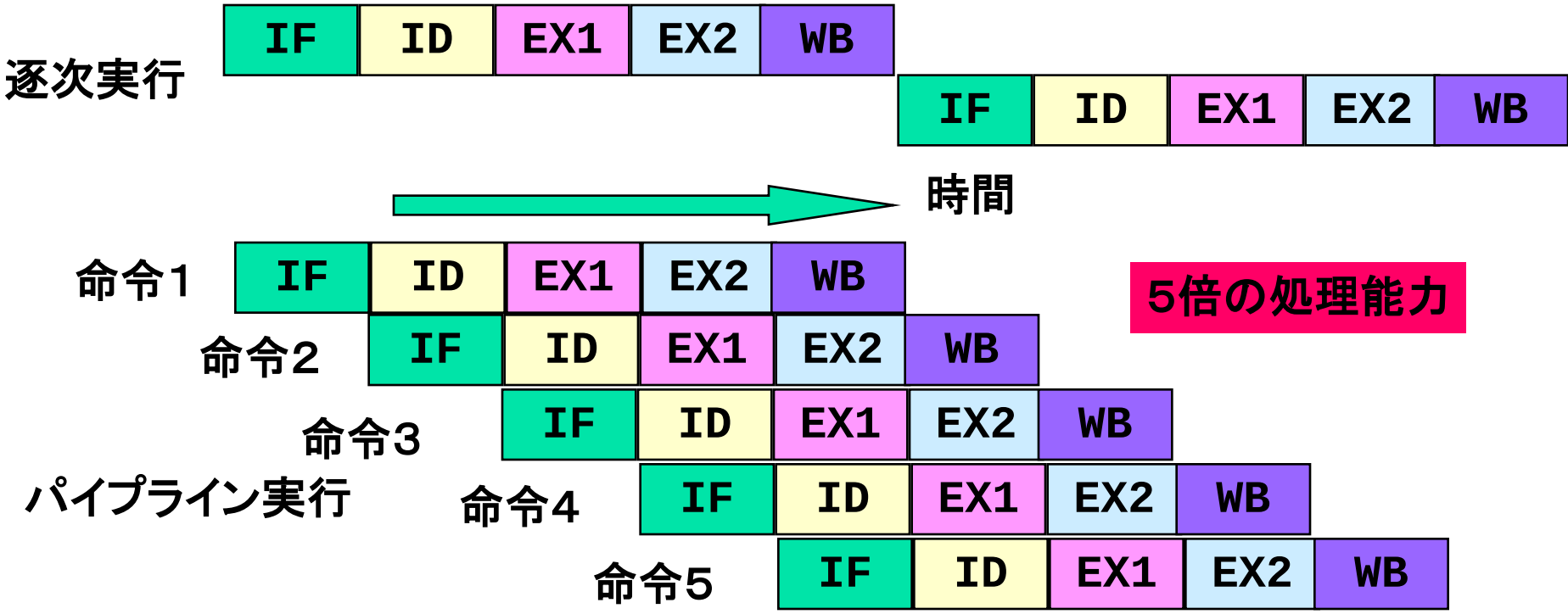
キャッシュメモリ

- 早い少量のメモリをCPUに近いところ（チップの中）におく。
  - キャッシュメモリ：SRAM(static RAM)、アクセス速度が速いが、少容量
  - メインメモリ：DRAM(dynamic RAM)、アクセスが遅いが大容量、安価
- 命令をおくための命令キャッシュとデータをおくためのデータキャッシュに分かれている
- １次キャッシュと2次キャッシュ、３次キャッシュも
- CPUのクロック速度が速くなっている現在、必須の技術



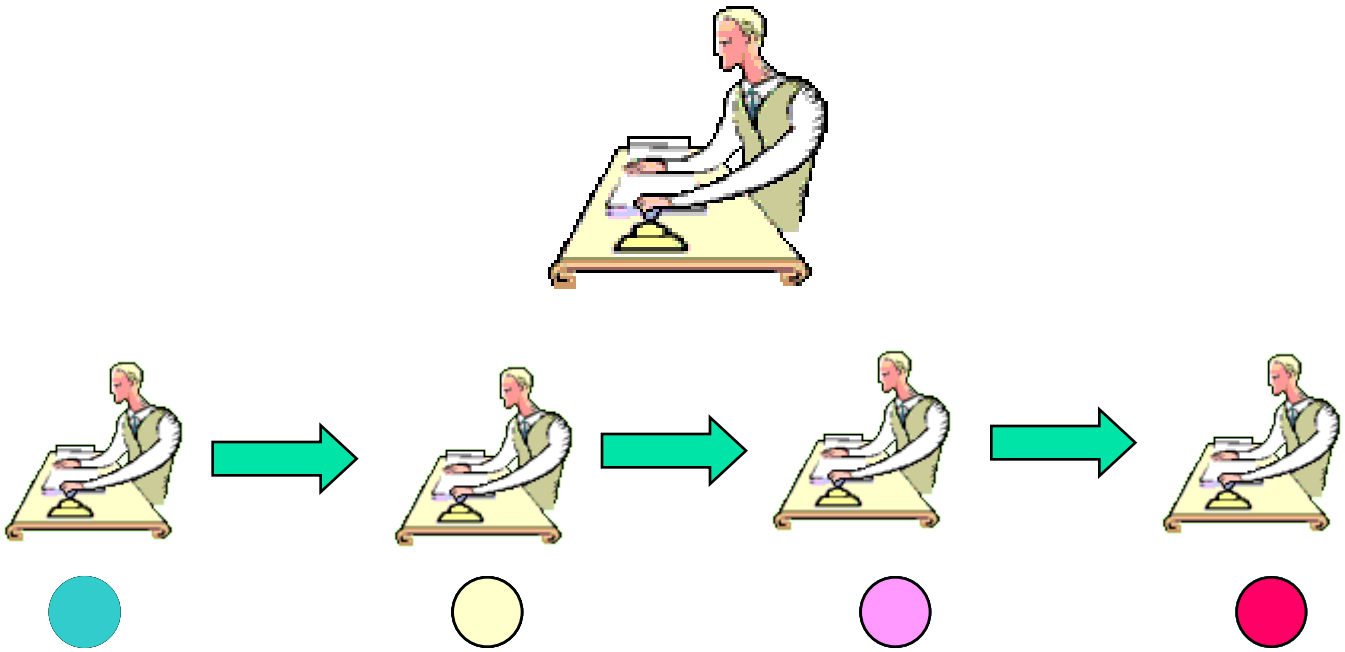
パイプラインアーキテクチャ

- 処理のフェーズをオーバーラップして、同時に複数の命令を時分割して実行する方式
- 同じ時刻ではそれぞれの命令は異なるフェーズを実行している。
- 細かくすればするほど、速度は向上する。
- 早いクロックのプロセッサでは各フェーズは細かくなる。



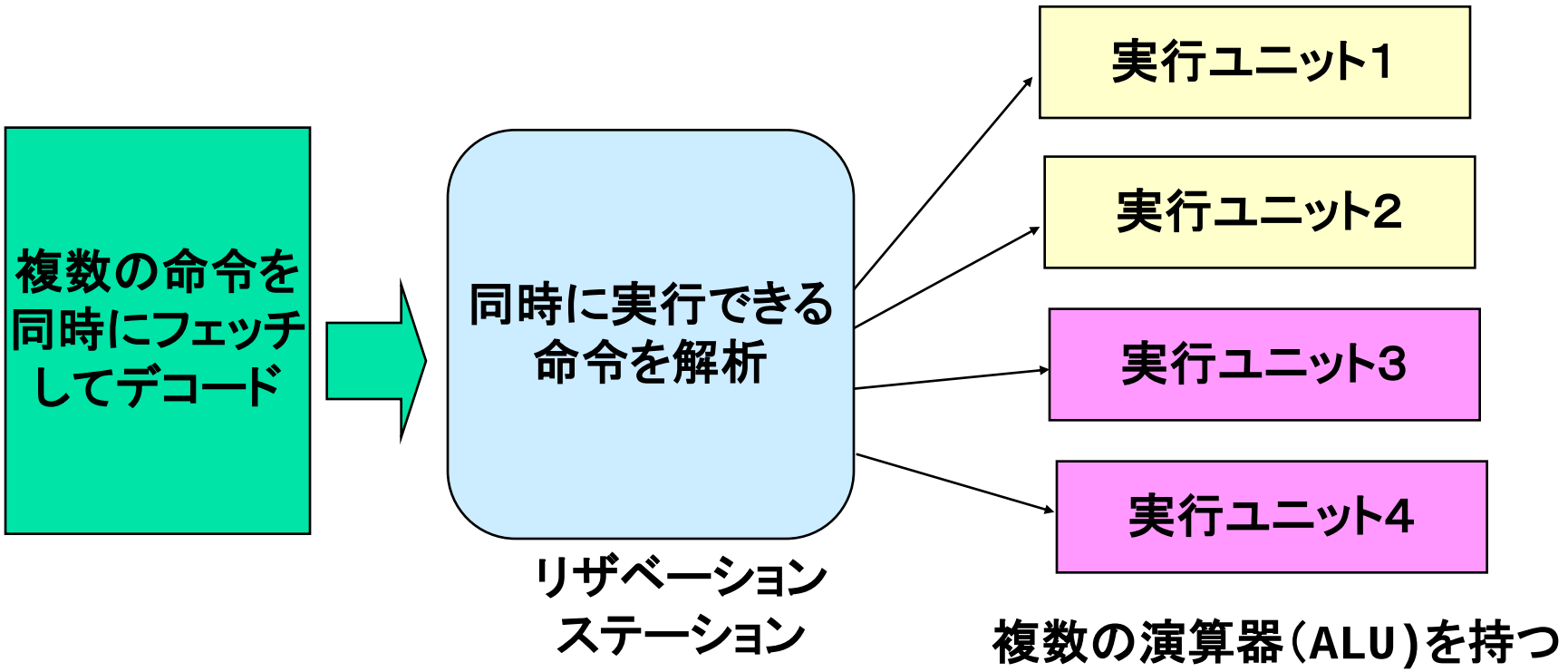
パイプライン

- いわば、流れ作業
- 一人の人がやるよりも、機能分担して、流れ作業をすればいいということ。



スーパースカラ

- 複数の命令をフェッチして、並列に実行できる命令を見つけて、複数の実行ユニットを使って、複数の命令を同時に実行する。
  - Out of Order 実行



特別な用途の命令

- 数値計算に多く使われる行列計算を効率的に処理する仕組みを入れる
  - ベクトルコンピュータ

- グラフィック処理など、特別なアプリケーションに有効な命令をつくる
  - たとえば、3x3の行列演算、座標変換などに用いる
  - MMX (Multi-Media eXtension) ⇒ SSE (Streaming SIMD Extensions)
  - 3DNow! (AMD)
  - AltiVec (IBM, PowerPC)

コンピュータを速くするには、…

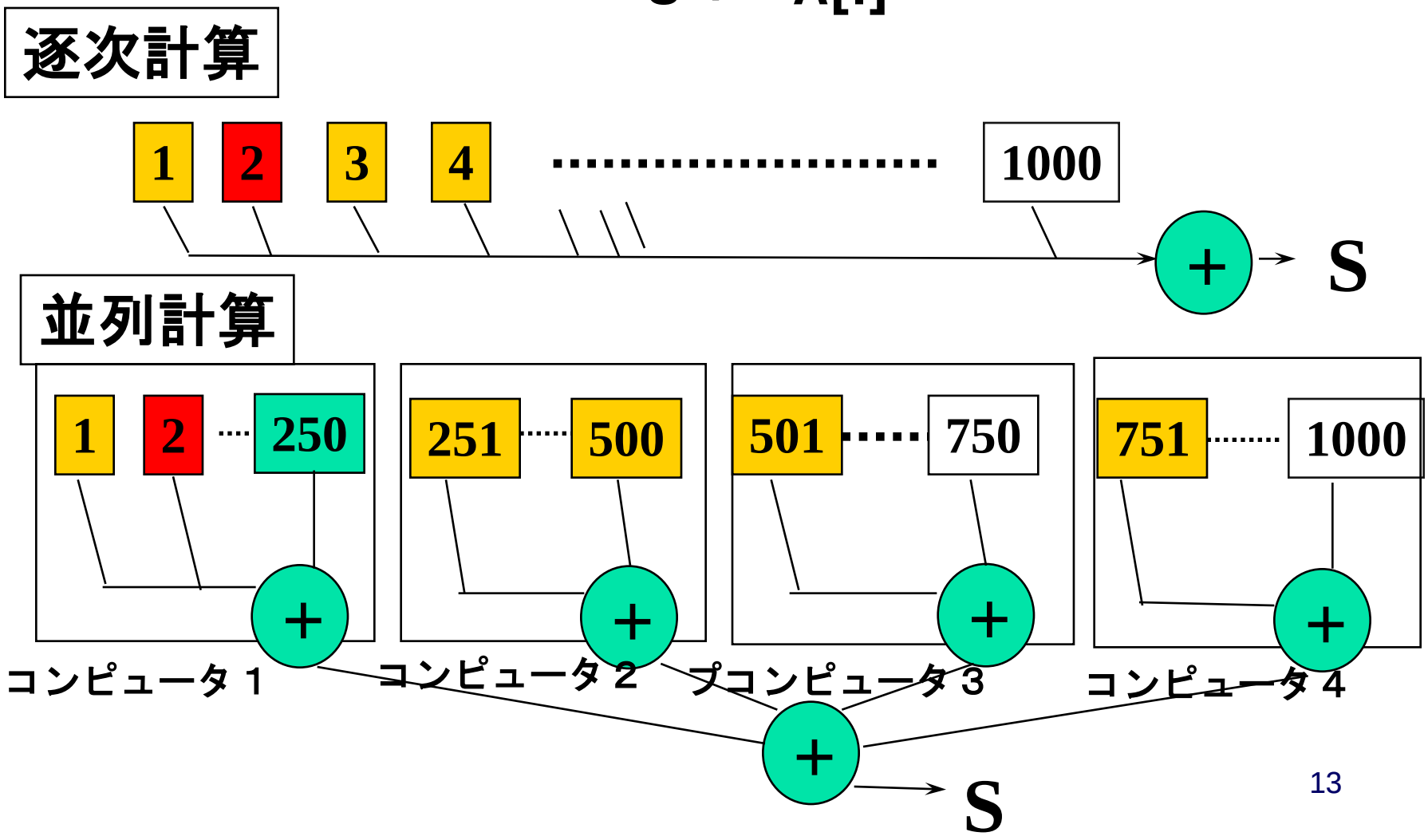
- ③　たくさんのコンピュータを同時に使う。
  - 並列コンピュータ
  - 今のスパコンの主流はこれ！

  - 一つのチップにたくさんのコンピュータを入れる
    - マルチコア・メニーコア
    - PCでも、スマホでも2，3個のコンピュータがはいっている

  - たくさんのコンピュータをネットワークつなぐ
    - PCクラスタ＝PCに使われるプロセッサをつなぐ
    - 並列スパコン＝専用のコンピュータ（ノード）をつなぐ

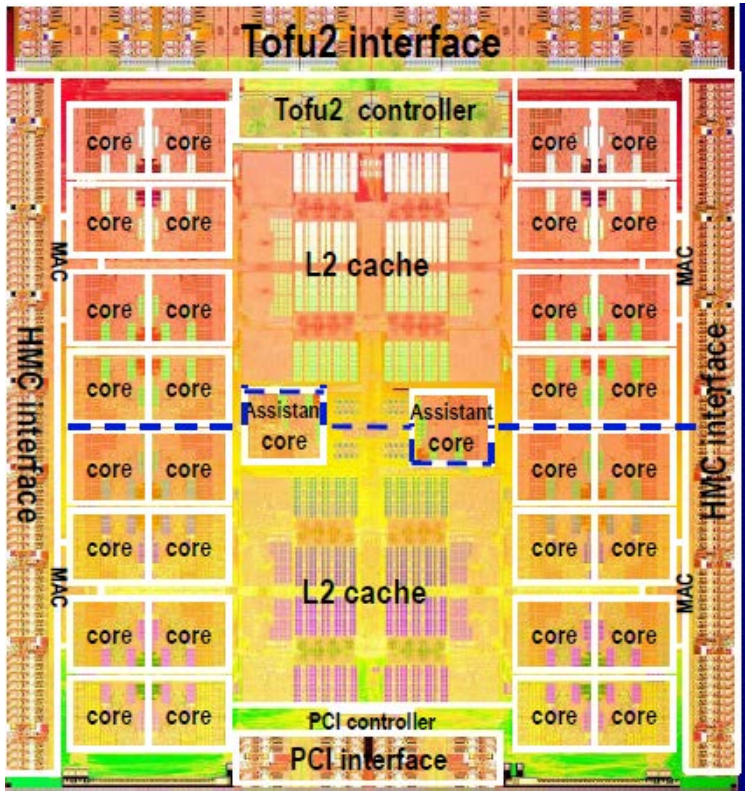
並列処理の簡単な例

```
for(i=0;i<1000; i++)
    S += A[i]
```



マルチコア

- マルチコアプロセッサ
  - 1つのチップに多くコア（CPU)を搭載したプロセッサ
  - 現在のプロセッサでは、2～4のコアがあり、主流となっている。
  - 現在は、16コア以下
- メニコア・プロセッサ
  - コアの多いプロセッサ
  - 50コア以上のことが多い



富士通のスパコンFX100の
SPARC64 Xifxのチップ写真
（32+2コア）

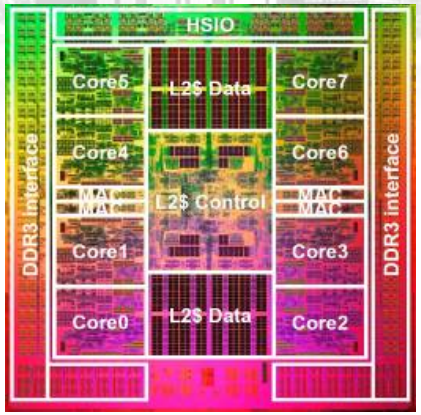
PCクラスタ

- PCクラスタ＝普通のディスクトップPC・サーバーPCをネットワークでつないだシステム。

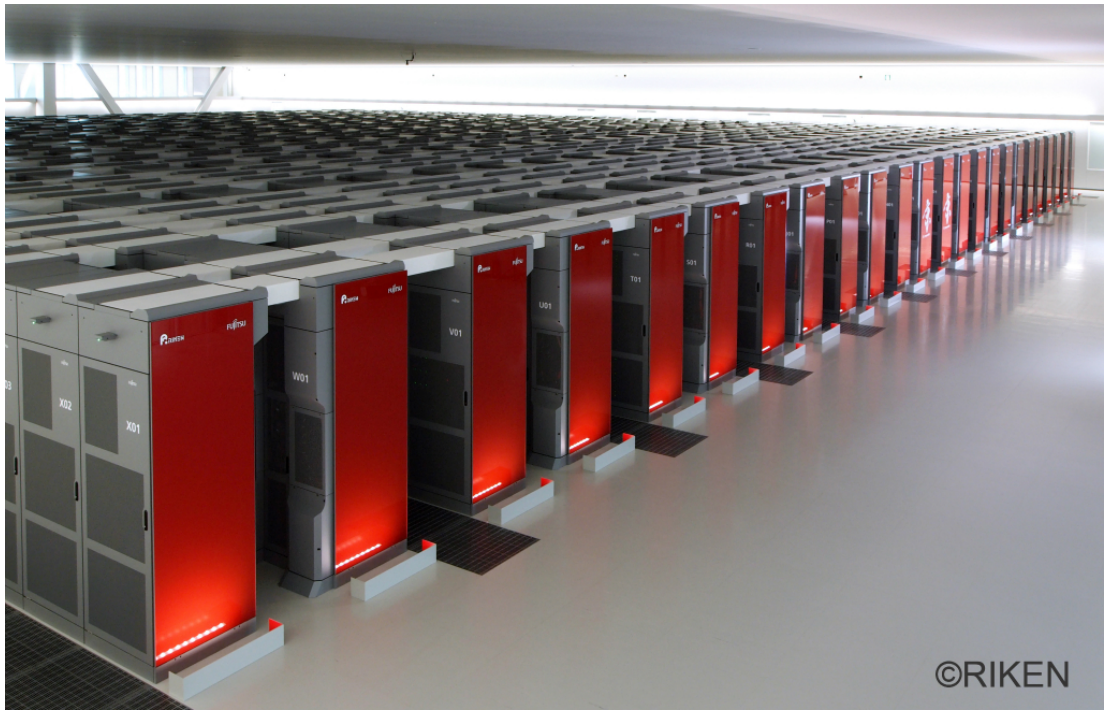
京コンピュータ

- 1つのチップに8個のコンピュータ(コア)
- 1つのコンピュータの性能は、16GFLOPS (2GHz), チップあたり、128GFLOPS
  - PCとかわらない？

元画像提供:富士通



- 筐体数　864
- チップ数:　82,944
- コンピュータ数:　663,552
- 性能　Linpack
  10.51PF
  (電力12.66MW)
  2011/11月



演算能力はどうやって「表す」のか

- 1秒（単位時間）あたりの演算可能回数
  - 数値は「浮動小数点」という形式で格納されている。
  - MFLOPS: Millions of FLoating Point OPerationS.（1秒間に10^6回の浮動小数点処理）
  - GFLOPS：10^9回，TFLOPS：10^{12}回，PFLOPS：10^{15}回
  - 理論ピーク性能＝常に演算器が使えた場合の性能
    - 演算器の数　x　クロックあたりの演算回数
  - 並列システムでは、台数xプロセッサあたりの理論ピーク性能

- では、演算能力はどうやって「測る」のか

ベンチマーク

ベンチマーク・性能評価の目的

- ユーザ： （自分の）アプリケーションがどのくらいの性能で実行できるかの目安を得る。
- システム調達・運営： 調達するシステムの所望の性能を指定する手段として用いる。
- システム設計： システム設計する場合の代表的なワークロードとして用いる
- （他のシステムとの比較）
  - 多くのシステムでのデータがあったほうが望ましい。

ベンチマークの歴史と種類

- （昔の）単一プロセッサのベンチマーク
  - Whetstone (1972): 浮動小数点演算のベンチマーク。多項式演算や行列演算、算術関数などを組み合わせた、人工的なベンチマーク。
  - Dhrystone (1984): 一般的な処理性能を測るベンチマーク。関数呼び出しや文字列の比較、分岐などのコードを組み合わせた人工的なベンチマーク。
  - Livermore カーネルベンチマーク (1986): 科学技術計算のコードの中から典型的なループを取り出して作ったベンチマーク。ベクトル計算機のベンチマークとして使われた。
  - Linpack ベンチマーク （1979）: Linpackという数値計算ライブラリの一部として、設定された密行列1次連立方程式を解くプログラム。はじめは、固定サイズ（はじめは100, のちに1000)の行列を解くプログラムをベンチマークとして用いた。

ベンチマークの歴史と種類

- （最近の）単一プロセッサ（マルチプロセッサも含む）のベンチマーク
  - SPEC:　非営利団体 Standard Performance Evaluation Corporation（SPEC、標準性能評価法人）が、設定したベンチマークで、1988年に設立された。「現実の」状況をテストすることを目指しており、その結果はwebサイトに公開される。ベンチマークテストとして、メモリやプロセッサの性能を測定するSPEC CPU, webサーバー用のSPEC Webを始め、スパコン用のSPEC HPCもある。有料。
  - EEMBC: 非営利団体　Embedded Microprocessor Benchmark Consortiumが決めた組み込み用のベンチマークセット。
  - Stream ベンチマーク:　メモリ性能を計測するプログラム
  - PC用には、グラフィックを測る 3DMarkや、PCMark、演算速度を測るスーパπが有名。
  - TPC:　データベースのトランザクション性能を計測するプログラムで、いくつかの種類がある。

スパコンのハードウェアの歴史

- 1983年：1 GFLOPS，1996年：1 TFLOPS…
- 1990年以前は、特別なスパコン（ベクトル型）が主流
- 1990年代以降は、多数のコンピュータを結合した並列計算機が主流に。
- 2000年以降は、PCに使われてマイクロプロセッサを使ったが並列計算機（PCクラスタ）が主流に。
  - PCに使われているマイクロプロセッサ（1つのチップでできたコンピュータ）の急激な進歩
  - 1.5年に2倍の割合でトランジスタの集積度が増加（ムーアーの法則）
  - 30年間で、1MHzから1GHz、1000倍の進歩
- 2008年にはIBM RoadRunner, 1Peta Flops (@Linpack) を達成
- そして、2011年「京」が世界1(@Linpack) に！

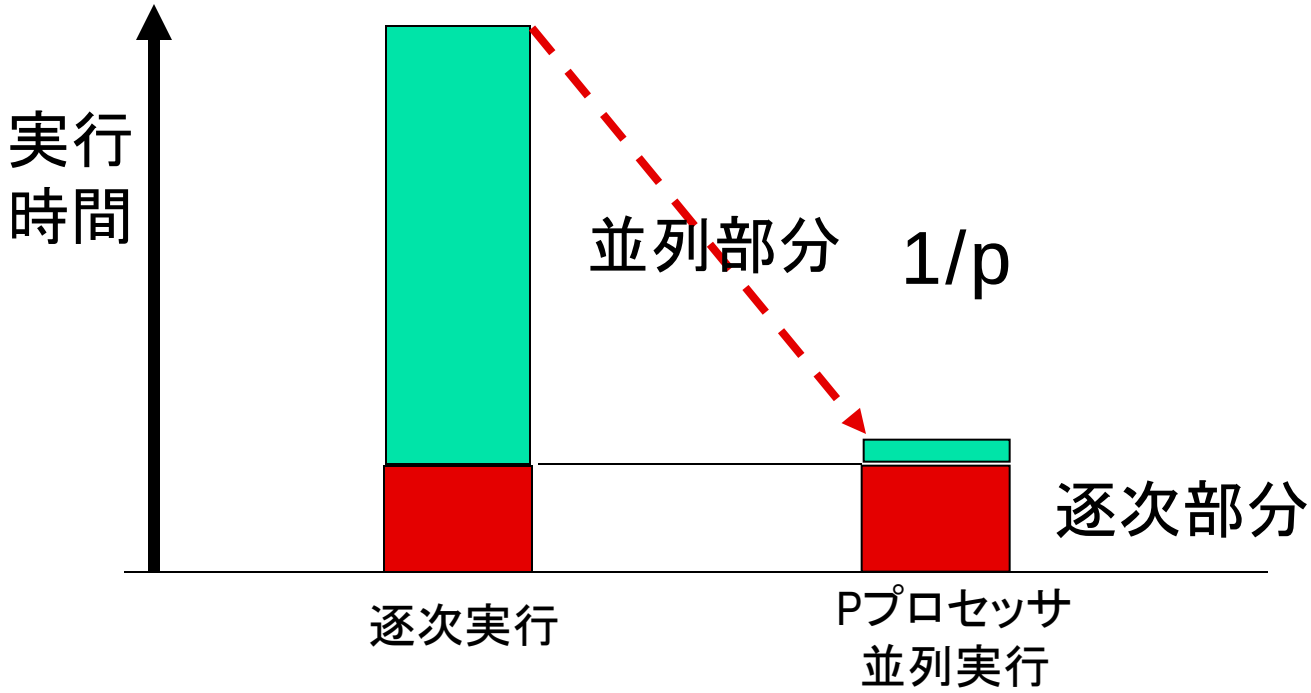
スパコンのベンチマーク

- Linpack (HPL): 並列システム向けの連立1次方程式を直接法で解くベンチマーク。サイズは自由。1秒当たりの演算数で結果を表示する。
- HPCチャレンジ(HPCC)ベンチマーク: HPL, RandomAccess, FFT, Stream の4つのベンチマークを1つのプログラムとして実行し、夫々の値から演算性能や通信性能、メモリバンド幅の総合性能を計測するプログラム
- HPCCG: 並列システムにおいて、疎行列の連立1次方程式をCG法による反復解法により解くプログラム。メモリバンド幅が性能に大きく影響を与える。
- Graph 500 ベンチマーク: グラフ探索問題を解くベンチマークプログラム。
- NAS並列ベンチマーク: 流体解析計算で用いられるいくつかの典型的な計算を抜き出したベンチマークセット。

スパコン＝並列システムの
ベンチマークは何が難しいか？

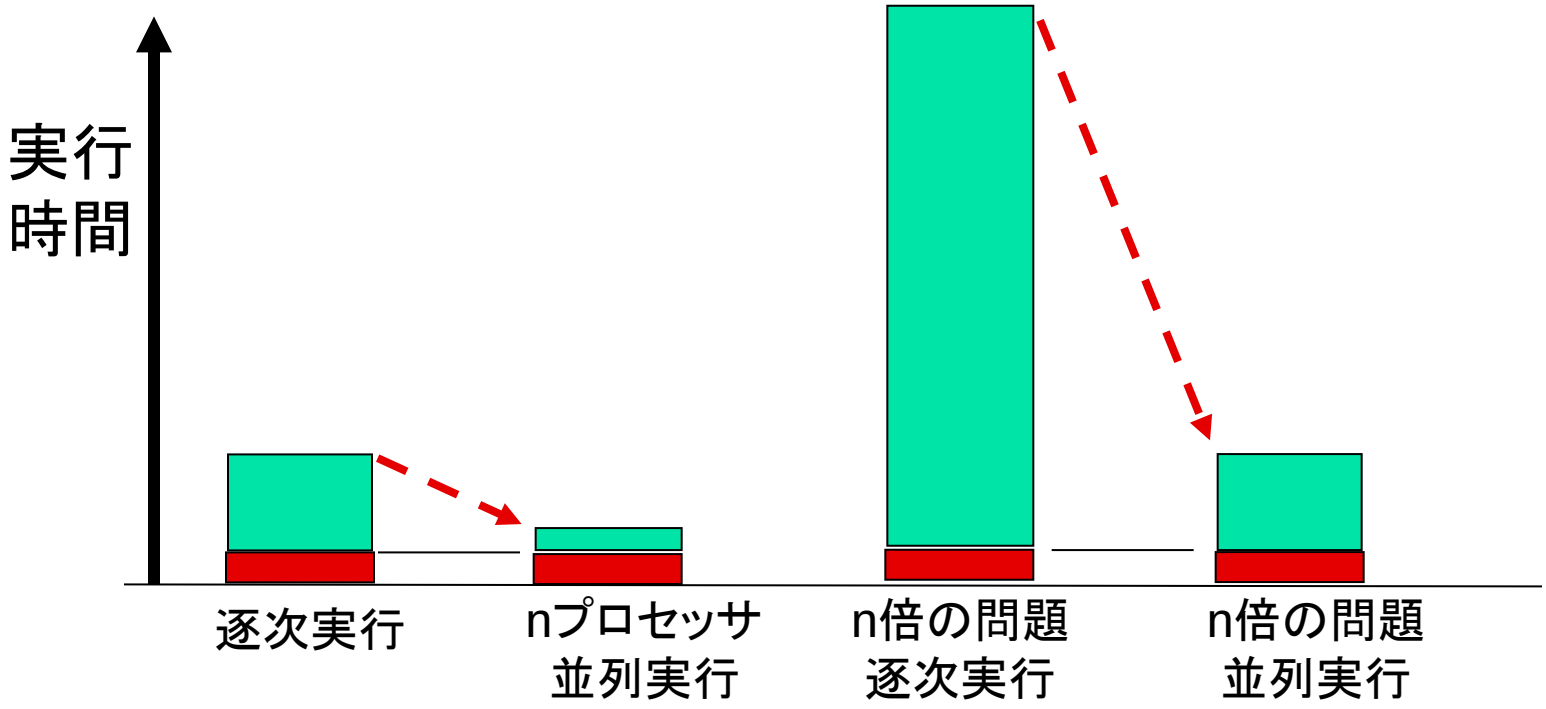
並列処理の問題点:「アムダールの法則」の呪縛

- アムダールの法則
  - 逐次処理での実行時間をT_1, 逐次で実行しなくてはならない部分の比率がαである場合、pプロセッサを用いて実行した時の実行時間(の下限)T_pは、$T_p = \alpha * T_1 + (1-\alpha) * T_1/p$
  - つまり、逐次で実行しなくてはならない部分が10%でもあると、何万プロセッサを使っても、高々10倍にしかならない。



並列処理の問題点:「アムダールの法則」の呪縛

- 「Gustafsonの法則」:では実際のアプリではどうか？
  - 並列部分は問題規模によることが多い
  - 例えば、ノード数nの場合、n倍の大きい問題を解けばよい。n倍の問題は、計算量がnになると、並列処理部分は一定
  - Weak scaling – プロセッサあたりの問題を固定←大規模化は可能
  - Strong scaling －問題サイズを固定　←こちらはプロセッサが早くなくてはならない。



ベンチマークの分類

- Real benchmark vs. Synthetic benchmark
  - Synthetic（人工的、合成）とは、よく現れるパターンの組み合わせなど、計算の結果には意味がない。
- Application benchmark vs. kernel benchmark
  - アプリベンチマークは1つのアプリ全体だが、kernelは一部分の計算だけ。
- Single value results vs. vector (multiple) results
  - 結果が1つの値になるかどうか。
  - 複数のベンチマークの結果を集計するベンチマークではその意味が問題になる。（たとえば、SPEC CPUのSpecINT)
- Fixed size or variable size
  - システムの規模にしたがって問題のサイズを変えることができるのか

Linpack (top500)についてのコメント

- 良いところ
  - 大規模連立一次方程式を解くという、一応意味のある計算を行っており、結果の検証ができる。(real benchmark)
  - 並列システムのプログラムであり、任意のサイズで超大規模なシステムまで同じベンチマークで計測できる。(variable size)
  - 結果がsingle valueであり、結果に意味があり、わかりやすい。
  - これまでの結果が蓄積されており、また、多くのシステムの結果が集計されており、システムの比較評価が可能。

- 悪いところ
  - アプリケーションベンチマークではなく、アプリの一部でしかない。
  - 大規模システムで解く問題はもはや現実的ではないサイズで、結果にはもはや意味がない。
  - 演算の最適化手法がすでに成熟しており、特定の演算パターン(行列積)の性能にバイアスしている。結果として、ベンチマークの結果が反映できるアプリの範囲が限定されている。

最先端のスパコンを作る時の問題は、…

- いまのスパコンの性能は、並列処理から
  - つまり、性能　＝　プロセッサの性能　x　台数
  - Top500の性能向上は、Mooreの法則による伸びを越えている。
- すなわち、コンピュータ数
- ということは、性能は結合するコンピュータの数を増やせばいい
- が、電力が限界
- 2008年から、電力消費量を表示するようになった
  - これからのスパコンは電力が大切

Top500の性能の伸び

June/2011

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	548352	8162.00	8773.63	9898.56
2	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C / 2010 NUDT	186368	2566.00	4701.00	4040.00
3	DOE/SC/Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
4	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU / 2010 Dawning	120640	1271.00	2984.30	2580.00
5	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU Linux/Windows / 20 NEC/HP				
6	DOE/NNSA/LANL/SNL United States	Cielo - Cray XE6 8- Cray Inc.				
7	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix Xeon HT QC 3.0/Xe Infiniband / 2011 SGI				
8	DOE/SC/LBNL/NERSC United States	Hopper - Cray XE6 Cray Inc.				
9	Commissariat a l'Energie Atomique (CEA) France	Tera-100 - Bull bull S6010/S6030 / 201 Bull SA				
10	DOE/NNSA/LANL United States	Roadrunner - Blade Cluster, PowerXCel 1.8 GHz, Voltaire In IBM				

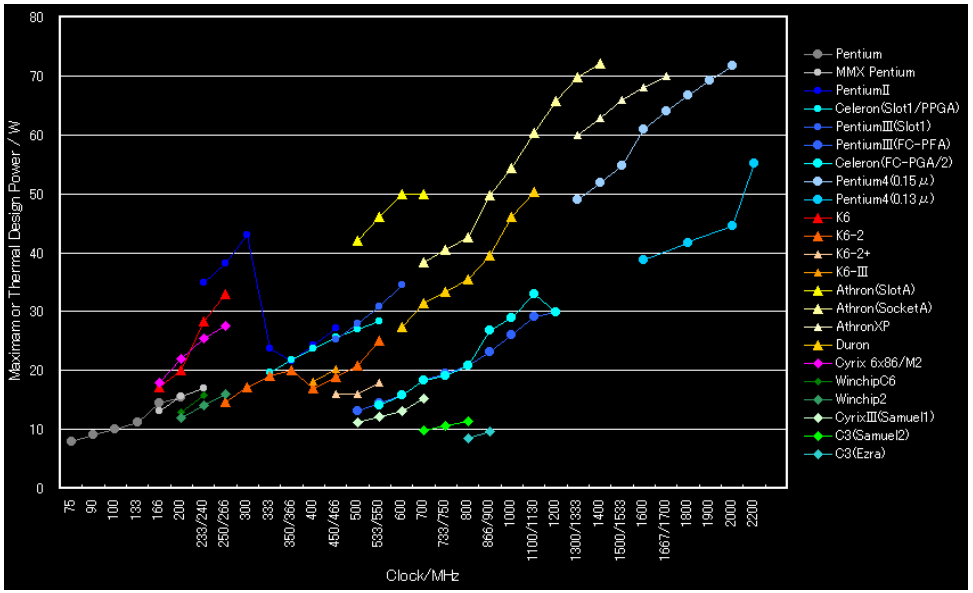
プロセッサの消費電力の遷移と省電力の重要性

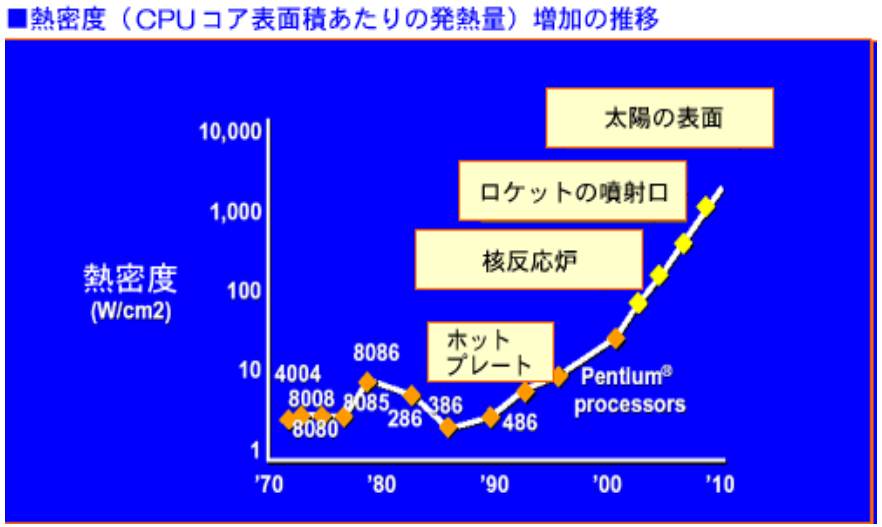
- 一般にCMOS回路の消費電力は以下の式で表現される。

$$P \propto CV^2F$$

P:消費電力　C:静電容量　V:電圧　F:周波数

- 近年の消費電力の増加は主に周波数の増加による
  - サーバーで100W以上、
- 集積度が上がると、発熱密度が上がり、冷却が難しくなる



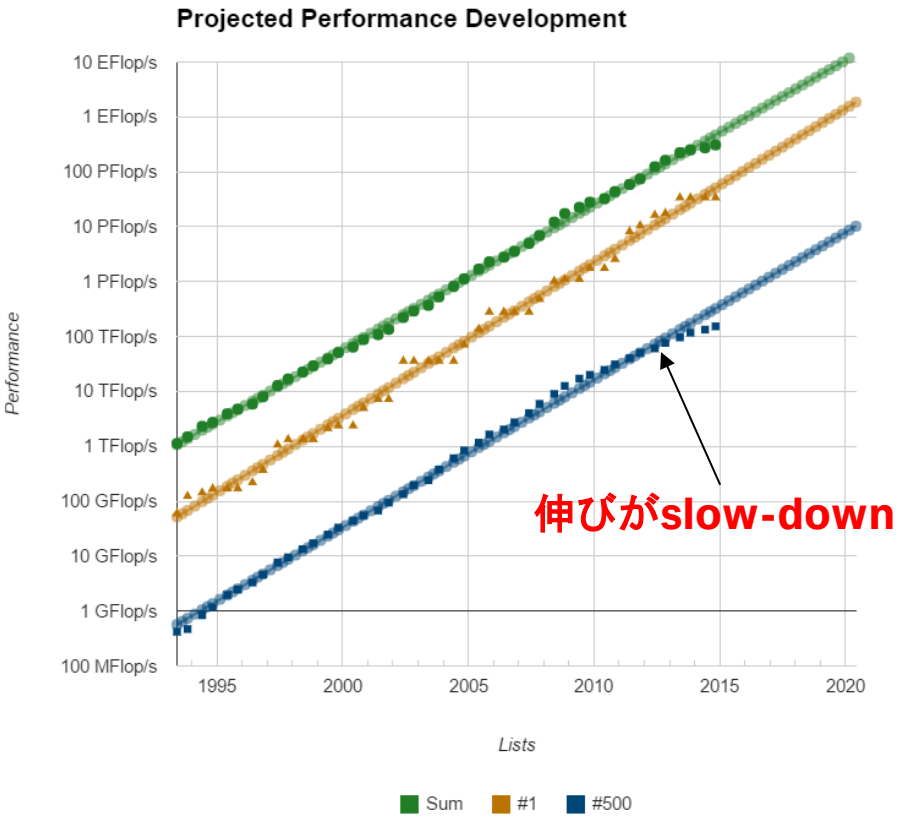


ISCCC 2001基調講演　より抜粋

Top500の動向 （1）

- Top500から、近年のスパコンの進歩の停滞が指摘されている。
  - 性能の伸びが、これまでの年率1.9%から1.2倍に。
  - 2014年11月(SC14)のリストでは、1位から9位までは変化がなかった。Top500に新規に入るシステムの数が激減（これまでの200～150システムから80程度に）
  - 2010年頃までは上位50～70システムの性能合計が全システムの性能合計の半分を占めるという状況であった。しかし、このところ上位10～30システムで半分を占めるという状況に。

Top500: 全世界のスパコンの性能をLinpackと呼ばれるプログラムでランキングしたもの。例年6月と11月に更新される。
http://www.top500.org



参考:
http://news.mynavi.jp/articles/2014/12/10/sc14_top500bof/
http://www.cnet.com/news/top500-supercomputer-race-hits-a-slow-patch/

top500の動向 （2）

- システムの性能の伸びに比べ、プロセッサの性能は伸びていない。
  - 性能は、プロセッサの個数の増加（大規模化）、アクセラレータ（メニ・コアを含む）による。
  - プロセッサ自体の性能は伸びが鈍っている。Intel メニーコア、NVIDA GPU等の最新プロセッサのデリバリの遅れ。ムーアーの法則のスローダウン
- 大規模化により、電力性能の重要性が顕著になっている
  - Top10システムでは2000GFlops/kW程度になっているのに対して、Top50システムでは1500GFlops/kW、Top500システム全体では1000GFlops/kWとTop10に比べて半分の効率でしかない。

- 米国、中国、日本で、2015年から2017年頃に、数10～数100PFlops級のスパコンの設置計画あり、現在のスローダウンは一時的なものであるという見方もある。
- 国別のシステム数では、米国がほぼ半分の46%を占め、中国が12%。日本、英国、フランス、ドイツの各国が5～6%。
- Top500は主にCPU性能のみで、ワークロードを反映していないという意見から、HPCGやgraph500での評価にも興味が集まっている。

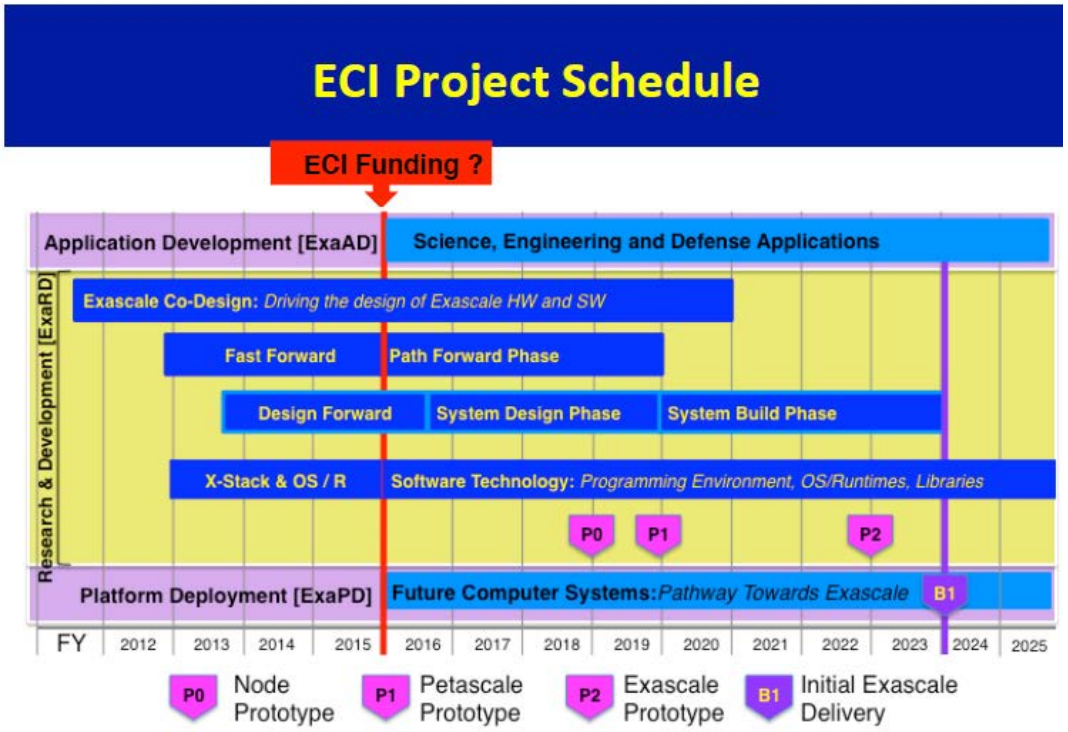
米国の動向　～エクサスケールに向けての計画～

- Exascale Deliveryは、2023年末（SC14での発表）
- 2011年より、3つのExascale Co-Design Center(LANL, ANL, SNL)に設置された。
- FastForward　Phase2(2014年7月～）:2020年～2023年の製品化時期を想定したノードアーキテクチャとメモリ技術への研究開発への企業へのファンデング、1億ドル規模。Phase1は、2012年から、２年。
- Design Forward: ネットワークを対象。2013年から。
- X-stack:プログラミングモデル

ランタイムなどのCSの研究開発。
2012年から。

- 2015: Preliminary Conceptual

Design for an Exascale Computing Initiative　(ECI)
が進んでいる模様



Bill Harrod、
http://wallaby.aics.riken.jp/isp2s2/program/

米国の動向　～システム設置の計画～

- Cori (LBNL, NERSC-8)　2016
  - Xeon Phi (KNL)ベースのシステム。Crayが受注。
  - ノード性能3～3.5TFとすると、全体性能は約30PF
- Trinity (NNSA) 2016
  - これもCayが受注。
- CORAL (Collaboration of Oak Ridge, Argonne and Livermore)
  - Summit (ORNL, LLNL) 2017～2018
    - IBMが受注。NVIDIAとMellanoxとともに、Power9とVolta (Maxwellの次の世代のGPU)によるノードをInfinibandネットワークで結合したシステム。Power9とGPUは、NVIDIAの専用リンクであるNV-Link（80～200GB/s）で結合。
    - ノードの性能は、40TFの計算性能で、ノード数は3400。
    - 40TFの性能のためには、ノードあたりのGPUの個数が、6～8個と予想される。
  - Auroa (ANL) 2018
    - Intelが受注。KNLの次の世代のメニーコアプロセッサKNH(Knights Hill) を、IntelのインターコネクトOmniPath2で結合したシステム。
    - システムのインテグレーションは、Cray

運用開始	2016年半ば
CPU Architecture	Knights Landing
Peak Performance	3TF/Node以上
Number of Node	9,300 Node以上
I/O Bandwidth	400 GB/sec以上
Storage	28 PB
Price	70 M$ (70億円)

Coriシステムの概要
Priceは全体予算。1ドル100円換算

	Aurora (ANL)	Summit (ORNL)
導入年	2018	2017-2018
システムピーク性能(PF)	180	150
電力(MW)	13	10
電力あたり演算性能（GF/W）	13.8	15.0

CORALシステムの概要

- **米国の購入コストおよびGF/Wの計算にはストレージシステムも含まれている。1ドル120円計算。**
- **電力あたり演算性能はGF/Wはピーク性能/ピーク電力で計算している。ピーク性能に対する消費電力は公表されていない。**

米国スパコンの情報は以下の情報から

- **http://www.datacenterknowledge.com/archives/2015/04/15/doe-taps-intel-cray-to-build-worlds-fastest-supercomputer/**
- **http://www.hpcwire.com/2015/04/09/argonnes-200-million-supercomputing-award/**
- **http://www.hpcwire.com/2015/02/04/obamas-2016-budget-request-holds-clues-exascale/**
- **https://asc.llnl.gov/CORAL/**

中国の動向

- 中国の今まで
  - 2002-2005: High Performance Computer and Core Software
  - 2006-2010: High Productivity Computer and Service Environment
  - 2010-2016: High Productivity Computer and Application Service Environment
  - Tianhe-2 and Sunway-NG (Shenwei processorを使う？)

2015年4月12日の記事

米商務省12日までに中国が保持する世界最速級のスーパーコンピューター2基が核爆発関連の研究に活用されていたことが判明したとして、米半導体大手の「インテル」と「エヌビディア」の2社に対しスパコン運営に携わる中国の4つの技術センターへのチップなどの輸出を禁止したことを明らかにした。

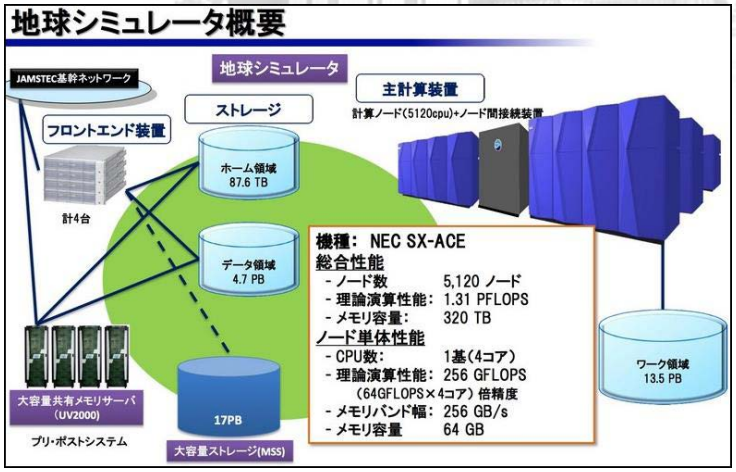
http://www.cnn.co.jp/business/35063042.html

- Tianhe-2: 現在、top500 1位のシステム
  - ノードは、Intel XeonにXeon Phi(KNC)を結合したもの。
  - インターコネクトは独自開発され、改良されている（MPI通信性能: 5GB/s ⇒12 GB/s、低遅延化および複数RDMA engine搭載、バリア同期高速化等）
  - 利用状況：N-body, CFD, Large-scale SNP(single nucleotide polymorphism), NEMO5などの利用例が紹介されている。クラウド利用も。

- 中国の今後
  - 国家三大主体計画（863計画、973計画、科学技術支援計画）の統合
    - 863計画：ハイテク産業技術の開発を目的とした応用技術研究開発プログラム。1986年3月に実施が決定されたことから863計画と呼ばれる
    - 973計画：将来の発展に役立つ基礎研究の強化を目的としている。1997年3月に実施が決定されたことから、973計画と呼ばれる。
  - 2015-2016 A transit period
  - Whether or not HPC will be a key is still open
    - ということで今後の計画は決まっていないもよう

国内・欧州の動向

- 3代目地球シミュレータ
  - https://www.jamstec.go.jp/es/jp/info/150601_es.html
  - 2015年6月から運用開始
  - NEC SX-ACE, 1.3PFLOPSとなり、メモリ容量は16倍の320TB
  - 東北大、阪大でもSX ACEが稼動



- Post T2K (東京大情報基盤センター・筑波大計算科学研究センター)
  - 両大学が、単一システムを柏キャンパスに共同設置
  - Intel Xeon Phi (KNL)ベースのシステムの導入を計画、20PF～30PF
  - プロセッサのデリバリが遅れ、導入は2016年にずれ込む見込み
- Tsubame3 (東工大学術国際情報センタ)
  - 資料招請「クラウド型ビッグデータグリーンスーパーコンピュータ」 2015年2月

- 欧州
  - ESSI2 : The European Exascale Software Initiative
    - MontBlanc – ARMベースの省電力スパコン
    - DEEP – メニーコア用のインタコネクトテクノロジの開発
  - システムについては、目立った動きはなし。

Thank you for your attention!!!