

並列コンピュータ内相互作用結合網 IP 化 による実行効率最適化方式の開発

東京大学
平木敬



のはずなのですが

のはずなのですが

**平木さんがなんだか別の用事でこれないそ
うなので**

のはずなのですが

**平木さんがなんだか別の用事でこれないそ
うなので**

代理で牧野がなんか話をします

のはずなのですが

**平木さんがなんだか別の用事でこれないそ
うなので**

代理で牧野がなんか話をします

**要素技術関係の話はよく知らないので
GRAPE-DR の話です**

GRAPE-DRと 超ハイエンドコンピューティング

国立天文台
理論研究部/天文シミュレーションプロジェクト (CfCA)
牧野淳一郎



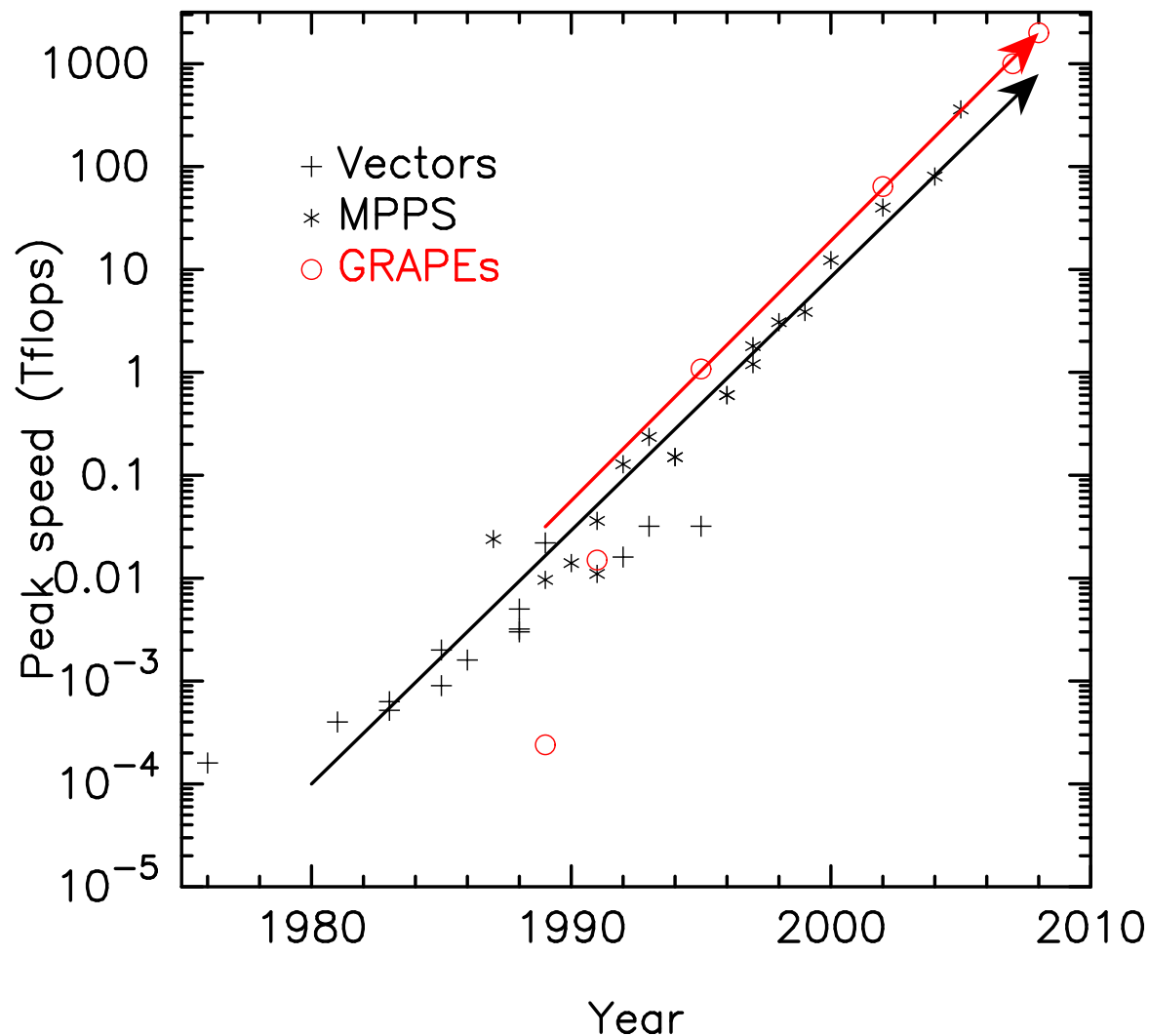
今日の話の構成

- GRAPE と GRAPE-DR
 - これまでの GRAPE
 - GRAPE-DR の考え方
 - 開発状況
 - ソフトウェアの実際
- これからの HPC

GRAPE の考え方

- **重力多体問題: 粒子間相互作用の計算が計算量のほとんど全部**
- **効率のよい計算法 (Barnes-Hut tree, FMM, Particle-Mesh Ewald(PPPM) ...): 粒子間相互作用の計算を速くするだけでかなり加速できる**
- **そこだけ速くする電子回路を作る (「計算機」というようなものではない)**

GRAPE のピーク性能の進歩



GRAPE-4 以降、完成した時点で世界最高速を実現

まあ、結構上手くいったのでは (自画自讃)

GRAPE-DR の基本的な考え

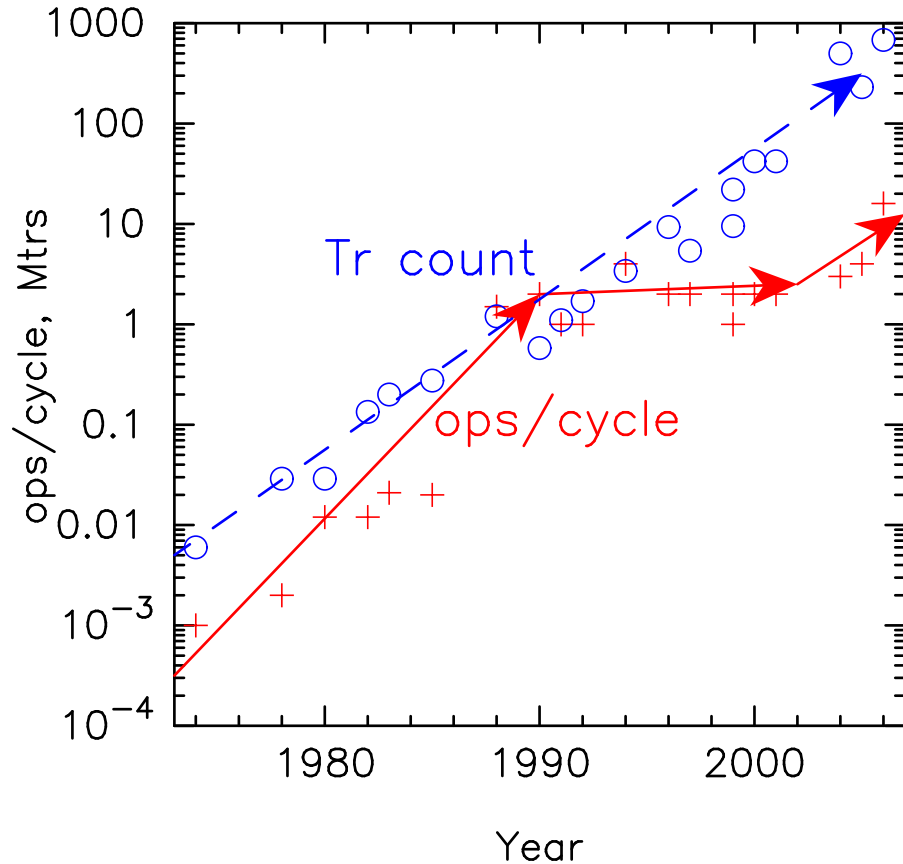
- 応用に特化し、多数の演算器を1チップに集積、並列動作させて高い性能を得た専用計算機の特徴を生かす
- しかし広い応用範囲を実現する
- 2004年度から5年計画
- 目標ピーク性能: 2 Petaflops
- チップ数 4096
- 単体チップ性能 0.5Tflops

発想を変える理由

- GRAPE あきた
- カスタムチップにお金かかりすぎるのでやってられない
- プログラムできるのはやはり嬉しい

トランジスタは沢山ある

マイクロプロセッサの「進歩」



- トランジスタ: 15 年で 1000 倍
- 演算器の数: 同じ期間に 8 倍
- 100 倍分がどこかに消えた
- 最も良かった時でもチップ上の演算器の割合は 5% くらい

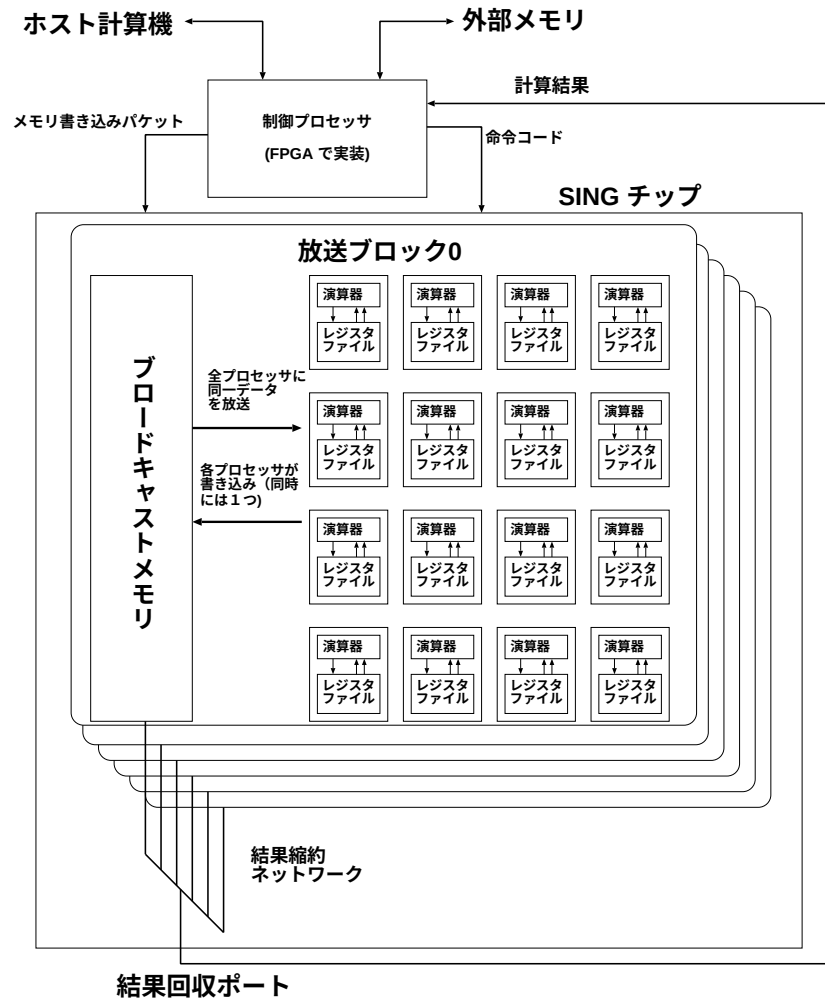
多数の演算器を詰め込む方法

境界条件: メモリバンド幅は増やしたくない (システムコストはほぼメモリバンド幅で決まる)

可能な方策

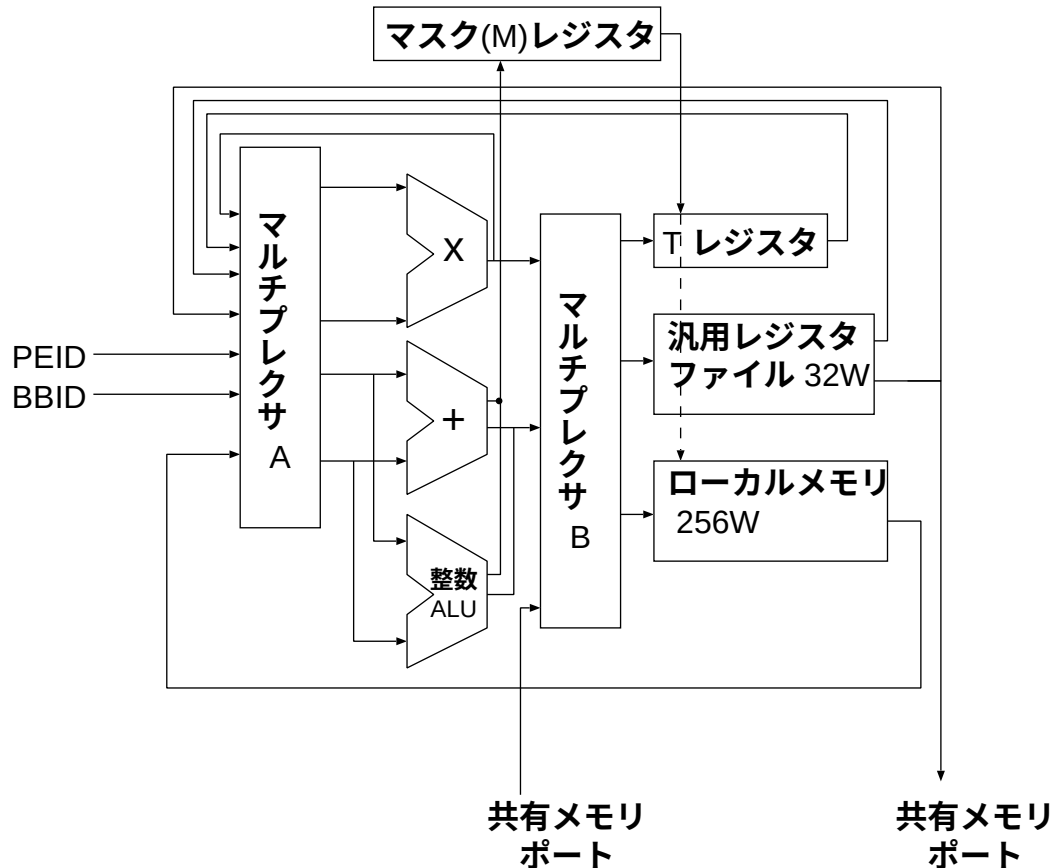
1. GRAPE 的専用パイプラインプロセッサ
2. 再構成可能プロセッサ (FPGA とか)
3. SIMD 並列プロセッサ

GRAPE-DR における SIMD



- 非常に多数のプロセッサエレメント (PE) を 1 チップに集積
- PE = 演算器 + レジスタファイル (メモリをもたない)
- チップ内に小規模な共有メモリ (PE にデータをブロードキャスト)。これを共有する PE をブロードキャストブロック (BB) と呼ぶ。
- 制御プロセッサ、外部メモリへのインターフェースを持つ

PE の構造



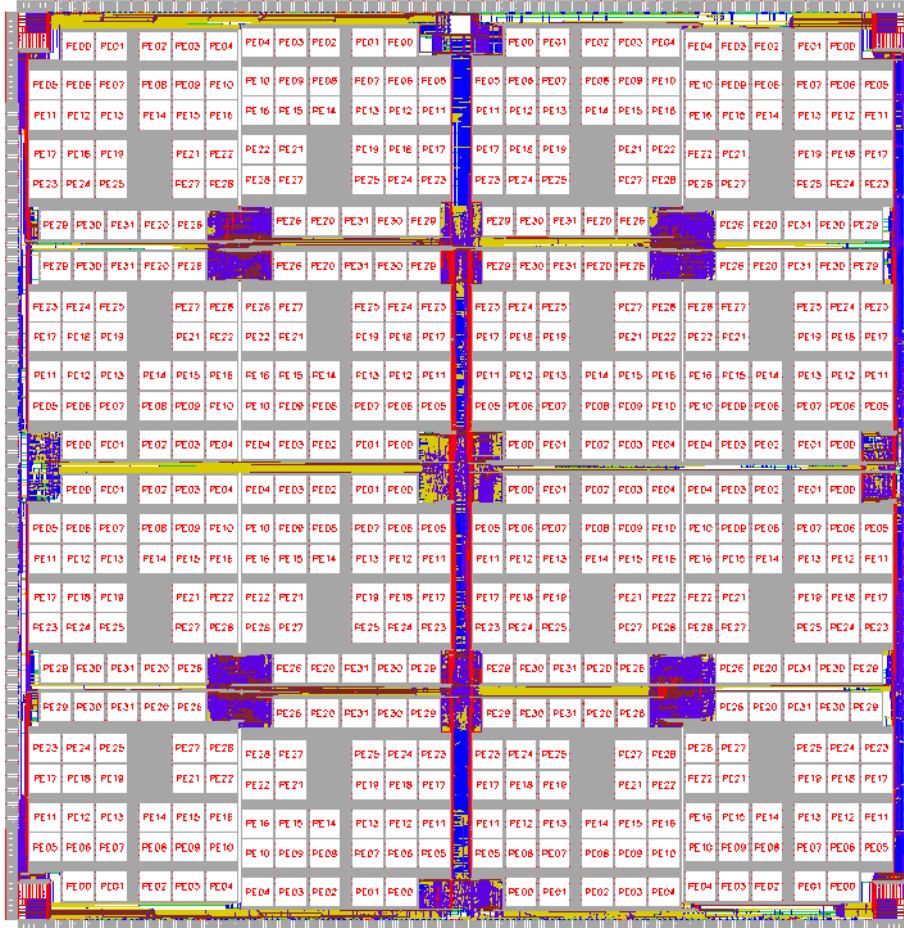
- 浮動小数点演算器
- 整数演算器
- レジスタ
- メモリ 256 語
(K とか M ではない。)

プロセッサチップとボード

チップ写真

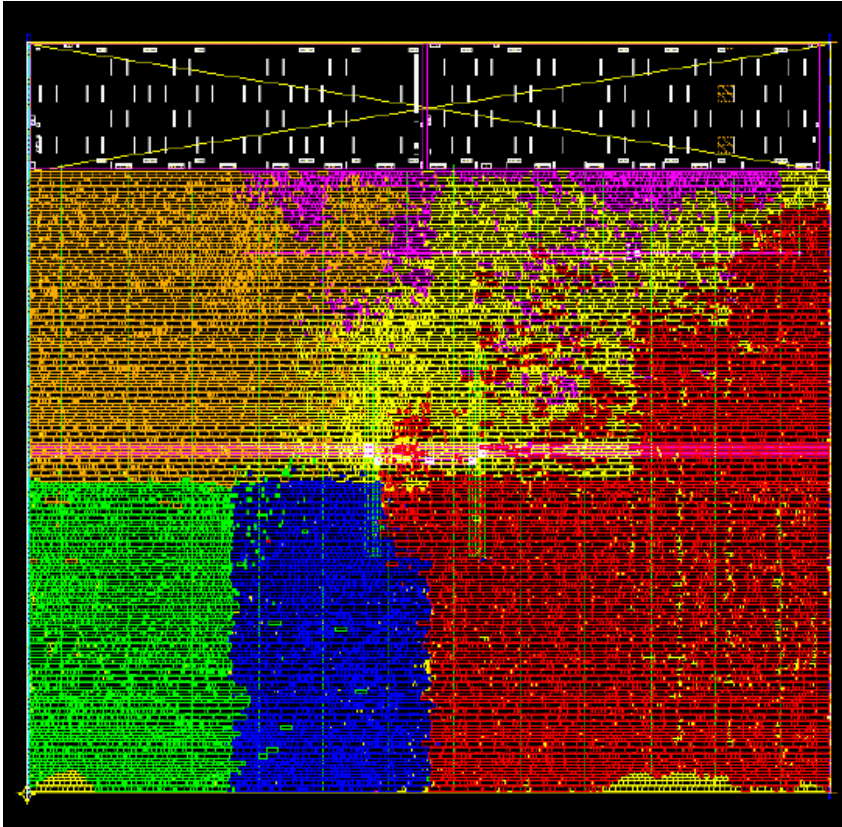


レイアウト



- 32PE(要素プロセッサ)のブロックが16個
- 空いているところは共有メモリ
- チップサイズは18mm 角

PE レイアウト



0.7mm by 0.7mm

Black: Local Memory

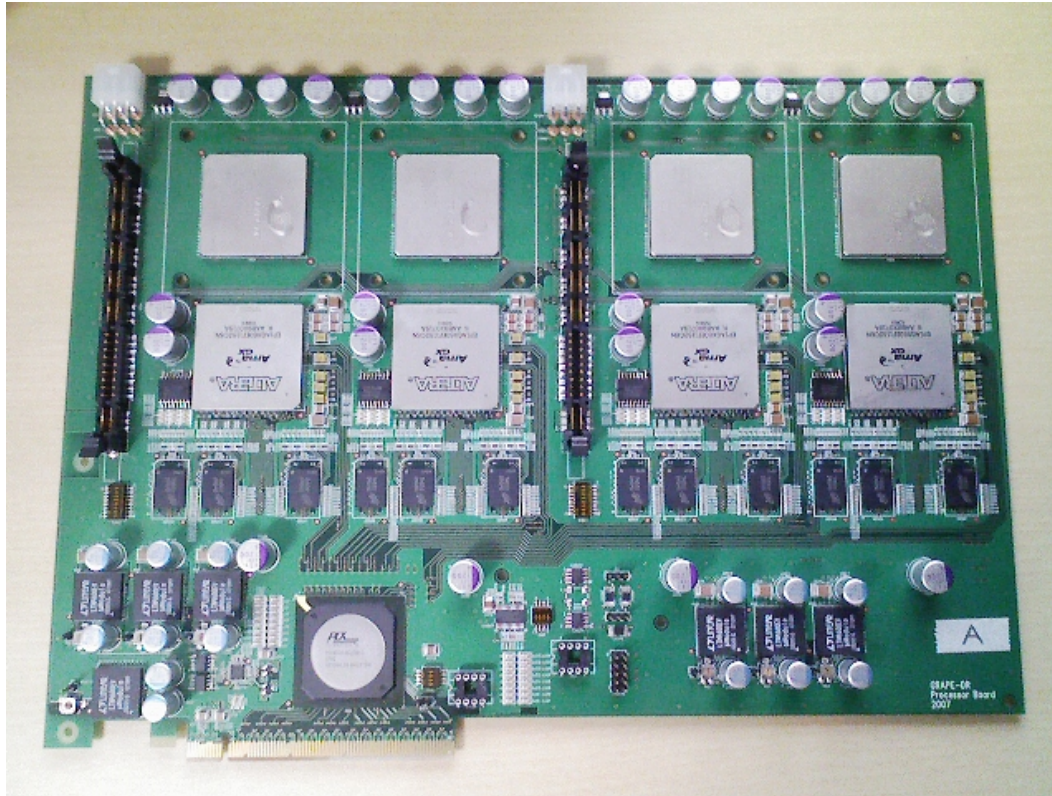
Red: Reg. File

Orange: FMUL

Green: FADD

Blue: IALU

GRAPE-DR プロセッサボード



- PCI-Express カード (16 レーン、通信速度 2GB/s)
- 4 GRAPE-DR チップ
- 理論ピークスピード 1TP(DP), 2TF(SP)

全体システム

- 目標:理論ピーク 1 Pflops = 4096 チップ
- ボード2枚のせた PC 512 台で構成する予定
- 製造は年次進行

今後のハイエンド計算機のために

- 3つの必要技術を、というのが座長からの司令
- 本当に必要なこと: ローエンドより C/P の良いハイエンド

1975:	Cray-1	100MF	10M\$	Cray	50 倍お得
	PDP-11/70	10kF?	50K\$?		
1985:	Cray XMP	1GF	10M\$	XMP	20 倍お得
	PC-AT	30kF?	5K\$		
1995:	VPP-500	100GF	30M\$?	VPP	3 倍損
	Dec Alpha	300MF	30K\$		
2005:	SX-8	10TF	50M\$?	SX-8	60 倍損
	Intel PD	12 GF	1K\$		

30年間で 3000倍変わった

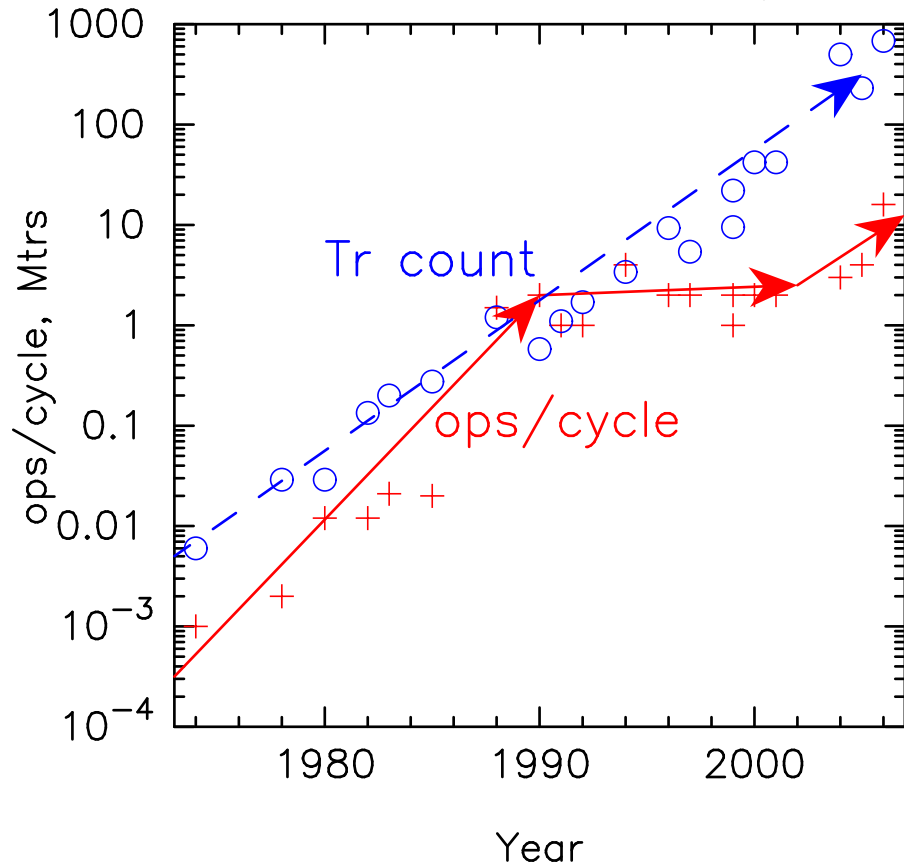
普通のスパコンは割高になった

一応3個あげてみた

- 多演算プロセッサチップ
- 多演算プロセッサ向けのアルゴリズム・アプリケーション開発
- 低レイテンシネットワーク

多演算プロセッサ

マイクロプロセッサの「進歩」



- もうちょっと演算器にトランジスタ使わないと性能向上は大変
- 電力が特に厳しい
- 5年後なら 5000 演算器くらいは入るはず。
- 64 とかは論外

電力あたり性能の例@2008

NEC SX-9	30KW/Tf
Cray XT4	5KW/Tf
GRAPE-6 (2002...)	1KW/Tf
GRAPE-DR	300W/Tf

- アーキテクチャで 100 倍は変わる
- トレンドとして、トランジスタ利用効率を落とさないことが重要

多演算プロセッサをどうやって使うか

- 特殊な技術を使わない限りメモリバンド幅は低い
- 特殊な技術を使っても、メモリバンド幅を増やすと電力が増える
- SX-9: 2.5B/F, Opteron 0.35B/F, GRAPE-DR 0.015B/F
- 一方、そこそこの量のメモリはオンチップで入る。5年後なら 256MB くらい
- 1980年代のスパコンの主記憶くらい

アプリケーション側の対応

- 演算のブロック化とか。アプリケーションによってはかなりできている。
 - 特に駄目なのは FFT がからむもの。
 - 陽解法流体とかはスキームによる

低レイテンシネットワーク

- 並列性能を決めるのはバンド幅ではなくてレイテンシ。
- バンド幅は沢山並べれば増える
- レイテンシは沢山並べると悪化

改善の方法

- ソフトウェアレイヤーをなるべく減らす
- スレッド数を減らす (チップ内 SIMD)
- 専用プロトコル？

まとめ

すみません、あんまりまとまってません

- 今後の発展は結局どれだけトランジスタを演算に使えるか
- 1チップに沢山演算器を入れる
- どうしてもチップ外の通信速度は落ちるので、アプリケーション、アルゴリズムからなんとかする
- B/F を落とさないように頑張る方向には将来はない