スーパーコンピューティングの将来

牧野淳一郎
東京大学理学系研究科
天文学専攻
2006/6/1から: 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プロジェクト長

詳しくは

http://jun.artcompsci.org/articles/future_sc/face.html

どういう話をするか

ベクトルを買ってはいけない。

- **何故か？**
- **ではどうすればいいか？**

大体の話の構成

- スパコンの歴史
- PC クラスタ以外のものが割高になった理由
- PCクラスタの問題点
- それ以外の可能性
- まとめ

スパコンとは何か

歴史

- 1960 年代 単に、 高くて速い計算機
  - 代表例: CDC 6(7)600 (Cray 設計)
- 1970年代 ベクトル プロセッサ
  - 代表例: Cray-1, CDC-Star
- 1980年代 いろいろあった
  - 日本のベクトル
  - クレイの並列
  - いろんな並列計算機

歴史のつづき

- 1990年代
  - ベクトル 衰退
  - 並列計算機 会社倒産
  - PC クラスタ 生き残る

何故そうなったか？

何故そうなったか？

- ベクトルは高くなった
- 並列計算機 やっぱり高くなった

何に比べて？

- PC クラスタ

高い計算機と安い計算機を比べると

1975:

Cray-1	100Mflops	10M$	Cray-1 50倍お得
PDP-11/70	10kflops?	50K$?	

1985:

Cray XMP	1Gflops	10M$	XMP 20倍お得
PC-AT	30kflops?	5K$	

1995:

VPP-500	100Gflops	30M$?	VPP 3倍損
Dec Alpha	300Mflops	30K$	

2005:

SX-8	10TF	50M$?	SX-8 60 倍損
Intel PD	12 Gflops	1K$	

30年間で 3000倍変わった

スカラ並列はどうか？

- 日立 SR-11000 (IBM製)
  SX-8 と変わらない
  東大、北大
  (IBM から p5-570 を買うとだいぶ安いかも)
- 富士通 HPC-2500
  ノーコメント

x86 サーバはどうか？

- 1 ノード 100 万円
- SX-8 より 10 倍まし
- 普通の PC より５倍損

ベクトルは何故高くなったか

- メモリシステムが高い
  - 超多バンク数(PC の1000 倍)
  - ランダムアクセスできる
  - バンド幅も大きい:PC の 10 倍
  - 共有メモリにしてるのでもっと高い
- 売れないから高い

というのも

スカラ並列は何故高いか

- チップ当りでベクトルより遅い
- 共有メモリなので高い
- 売れないから高い

x86 サーバは何故高いか

- 高くても買う人が少しはいるから儲けは大きい
- 保守とかでお金がかかる、ともいう
- 信頼性が高いとかいうのは嘘
- 計算機は高いほど良く壊れる
  - 作る台数が少ないと問題点をつぶせない
  - 日本のロケットが良く失敗するのと同じ

PC クラスタを買えばいいか？

そうでもない

- **ネットワークどうするか**
- **プログラムどうやって書くか**
  - MPI?
  - HPF?
  - **ソフトウェア分散共有メモリ？**

現実には MPI で書かないと性能でない

MPI でプログラム書く？

MPI なんかで書きたくないのが人情

- 別にサボリたいわけじゃない
- MPI でプログラムするのは生産性低い
  - ちゃんとそういう研究結果もある
  - 同じことをするのに余計に時間がかかる
  - サイエンスする時間が無くなる

何故 MPI で書かないといけないか

- そもそもそれ以外コンパイラがない
- コンパイラがあっても性能でない
  - 性能でないから誰も使わない
  - 誰も使わないから性能上がらない
  - 悪循環

昔は良かった？

1990年代
並列言語結構使えていた

- CM-5
- Cray T3D

PC クラスタでは初めから MPI とか

- 商用コンパイラが高かったせいも
- 性能でないのが大きい
- 性能でないのは通信レイテンシのせい

Cray T3D Shmem　　3μs
その頃の PC　　500 μs くらい
PC は 100 倍遅い

並列プログラムのスケーラビリティ

通信レイテンシが効く

台数増えると

- 計算量はもちろん減る
- 通信量は普通減る
- 通信回数は同じか増える

レイテンシが最後は効く

- 100倍遅いと並列化は「できない」

PC クラスタでの並列プログラム

- MPI で頑張って通信回数減らすと
  ある程度は性能がでる
- コンパイラが通信入れるのでは
  なかなか性能でない

どうすればいいか？
良くわからない

本来、計算機屋さんが考えるべきこと

- 速くて安いネットワークを作る
- まともなコンパイラを作る
- でも、あんまり役に立つ話はない

ネットワークが遅いのはどうして？

普通の PC クラスタ: ギガビットイーサネット

実はソフトウェアが遅いだけ

OS 層を全部飛ばして
ネットワークカードを直接アクセス

レイテンシ 5μs くらいにはできる

でも、スイッチがはいると 5μs 増える。

スイッチ重ねるとやっぱり遅くなる

専用ネットワークカード

高いけど速い

例: Quadrics QsNet-II

- 多段スイッチ通して 2μs
  GbE の 10 倍速い
- 値段も 10 倍
  PC 1台の 2倍以上する
  ネットワークつけると 3 倍損

何故高いか？

- 台数が少ないから。
- ハードウェアは GbE よりずっと単純。
- でも高い
- それでもベクトルより10倍まし
- スカラ並列よりも10倍まし
- MPI でもちょっとは楽

PC クラスタより安いのはないか？

- FPGA
- ゲーム機
- IBM BG/L
- Cray の超並列
- GPGPU
- GRAPE

FPGA

Field Programmable Gate Array

論理回路を書き換えられる LSI

- **プログラム:ハードウェア設計しないといけない**
- **数値計算向けのソフトウェアない**

ここ 15 年くらい「未来の計算機」

FPGA のプログラム環境

PGR システム
濱田・中里 (理研・戎崎計算宇宙物理研究室)

- パイプラインの演算器を書く
- 後はソフトウェアが自動でなんかする
- GRAPE 程度なら 2-30 行で書ける
- 割合使えるかも？
- 但し、計算精度低いものでないと厳しい

ゲーム機

- 発表当時、PS2 は使えそうだった
- 実は使えなかった

理由:

- 5 年間新製品がでない
  スパコン並
  5 年たつと 10 倍遅くなる
- 単精度しかない
- 丸めが変
- その他色々

使えなかった

PS3, Xbox も倍精度遅い
多分使えない

BM BG/L

オリジナル:
コロンビア大の QCD 専用計算機 QCDOC

ちょっといじったのがBG/L

QCD 屋さんはとても偉い

BG/L の特徴

- 高いネットワーク+PC 並の性能
- 電気は割合喰わない
- スペースも喰わない
- IBM が保守にくる

計算センター向け

但し:

- プログラム書くのはとても大変
- 「性能でた」というのが論文になる

Cray の超並列

XT3 / XD1

- XT3 はT3D の子孫
- Opteron + 速いネットワーク
- 結構よさげ
- でも高い。高いx86サーバの倍くらい
- ベクトルに比べると10倍いい

GPGPU

グラフィックカードを計算に使う

- 単精度
- 値段高い
- 計算機科学者の論文ネタにはいいかも
- 使って嬉しいかどうかは結構疑問

GRAPE は？

GRAPE-6 (2002年完成、予算 5 億): 64 Tflops
— 当時の PC (ネットワーク抜き) の 10倍いい

今作ってるのは実は「汎用」
理由: LSI 設計・試作の費用が高くなった

- GRAPE-4 の時は 2500万
- GRAPE-6の時は 1億ちょっと
- 今なんか作ると 4 億くらい (TSMC)
  - IBM だと10 億
  - 国内某社だと 20億
  - 特別推進でもそんなの無理

GRAPE-DR の考え方

天文専用GRAPE で予算とるのは無理

ではどうするか？

- プログラム可能に!
  - でも FPGA では駄目
- SIMD のお化け
- 1 チップに512 プロセッサ
  - マイクロプロセッサの200倍

512プロセッサも入る理由

- 昔のプロセッサはそんなもの
- $2\mu\mathrm{m}$ でプロセッサ作れた
- トランジスタ数今の $1/500$

沢山入れないのはいれても性能でないから

- メモリバンドが足りない

GRAPE-DR ではメモリバンド幅は?

今までのGRAPE:
メモリバンド幅あまりいらなかった。
GRAPE に似たことをやるなら同じでいい。

- GDR ではメモリバンド幅は頑張らない
- あまりいらない計算がメイン
  - GRAPE をエミュレーション
  - 行列計算もできる
  - 粒子法なら流体計算 だってできる
- メモリアクセス当りの計算量多いなら性能でる
- FFT は駄目

GRAPE-DR チップ

- チップ単体性能512Gflops
- 倍精度だと半分 — でも 256 Gflops
- 先週サンプルチップが届いた
- システム完成は 2009/3
  (振興調整費が5年だから)

GRAPE-DR システム

- 完成した時の速度 (予定) 2 Pflops
  地球シミュレータの 50 倍
  (あまり本当とは思えない)
- チップ 4000 個、512 ノード PC クラスタがホスト
- ネットワークは未定
- 消費電力 300KW
- 電気代払えないかも — 出来てから考える

京速計算機

- 今年度予算35億
- 何に使うのか私は知らない
- 文科省計画は総額1100 億
- 2011/3 に10Pflops の「特定処理部」先に完成
- 全体完成は2012/3
  - ベクタ 0.5P?
  - スカラー 2P?
  - 特定 10P?
- 実施機関:理研 (理研の責任者:魔球の姫野さん)

「特定」って何？

- 公式には決まってない
- 非公式にも多分決まってない
- まだ何ができるかわからない
- 夏くらいには見えてくるかも
- GRAPE-DRの技術をもとにとかいう資料も
- そもそも 3種混合を本当にするかどうかもまだ不透明

これからの計算センターは？

- ベクトル使いやすいけど遅い
- PC クラスタピークは高いけど性能でるとは限らない
- GRAPE-DR ピークはもっと高い。「汎用」とはいいがたい。

まとめ

- これからどうなるかはよくわからない

まとめ

- これからどうなるかはよくわからない
- よく考えましょう

おしまい