Post-Exaの
HPCシステムアーキテクチャ

牧野淳一郎
理研 AICS/東工大 ELSI

第十五回 PC クラスタシンポジウム パネルディスカッション 2015/12/18

Post-Exaの
HPCシステムアーキテクチャ

牧野淳一郎
神戸大学惑星学専攻 (3/1から)/理研 AICS/東工大 ELSI

第十五回 PC クラスタシンポジウム パネルディスカッション 2015/12/18

お題

中島さんからのメイル (12/12)

上記の件、「将来の」だと少し漠然なので、「Post-Exa のＨＰＣシステムアーキテクチャ」にしましょう。要するに FS2020 とか各国の Exa-challenge の「後」の展望（2025 ぐらい）を皆さんに語っていただく、ということで。

皆様わかってると思いますが、、、

- 問題は「**Post-Exa** がどうなるか」ではない

皆様わかってると思いますが、、、

- 問題は「**Post-Exa** がどうなるか」ではない
- 単に、
  - **Xeon** 以外に生き残るものはあるか？
  - **Xeon** 自体は破綻しないか？
  - **Xeon** が破綻したら世界は崩壊するのか？

32プロセッサの法則

リッチコアな物理共有メモリは **16-32** コアで破綻する
(それを超えると極端に **B/F** が下がる)

- マルチボード: **32** プロセッサ: **Cray T-90, NEC SX-7**
- マルチソケット: **Cray CS6400 (Sun Starfire)**, 富士通 **HPC2500**
- ソケット内: **Xeon Phi, ...**
- ソケット内超並列: **PEZY-SC?** ちょっと別物
  - リッチではない
  - **B/F** 小さい

32プロセッサの法則

リッチコアな物理共有メモリは **16-32** コアで破綻する
(それを超えると極端に **B/F** が下がる)

- マルチボード: **32** プロセッサ: **Cray T-90, NEC SX-7**
- マルチソケット: **Cray CS6400 (Sun Starfire)**, 富士通 **HPC2500**
- ソケット内: **Xeon Phi, ...**
- ソケット内超並列: **PEZY-SC?** ちょっと別物
  - リッチではない
  - **B/F** 小さい

これも皆様わかっていると思うけど

Xeon Phi は **Xeon** が破綻した後の姿

つまり

- **Xeon** は破綻する
- 我々は破綻後に備えないといけない
- 「備える」というのは 「**Xeon** の代わり」を用意することではない (それはあらかじめ破綻している)

ということで。

Xeon なきあとの世界

- 実際問題として Xeon そのものがなくなるか？
  - なくなる: サーバ系は ARM とかに移行
  - なくならない: 性能向上は今以上にスローダウン
    - AMD APU 的あれ: Intel には本気の製品作れない気が
- いずれにしても、「物理共有外付けメモリ」あたりの処理能力はサチる
- HPC 向けには「Xeon でないなにか」がはいりえるニッチは存在
- 結局アクセラレータ的なものを「誰かが作ることはできる」

2つの可能性

いずれにしても、汎用プロセッサが面倒くさい色々＋ネットワークを提供する構成でないと開発は不可能

- 別チップ: **IBM+NVIDIA** 的あれ
  - これにかぎっていうと多分上手くいかない
  - **Intel** はどうしたいか不明
  - まあ **PCIe Gen4** とかはある
- 同一チップ: **NVIDIA** が断念したあれ
  - **IP** 使うにしても高すぎる、、、
  - 開発リスクも巨大

逆にいうと

- 他に手を出す人があんまりいない
- 実はそれほど難しいことはない
- でもこういうのには研究費は、、、

どうなるか？ではなくて、
何をするべきか。

- これまでの計算機の進歩と現状の困難 (今日は省略)
- 解決の方向
- 可能なアーキテクチャ
- まとめ

これまでの計算機の進歩と現状の困難：まとめると

- 半導体技術は 2000 年頃に CMOS スケーリングが終焉、さらに微細化自体のスローダウンが進行中であり、もはや微細化が技術進歩を牽引していない。
- 昔の Cray-1 みたいな「スパコン」の進歩は、プロセッサコア 32 個くらいの並列度で限界、破綻した
- 現在のマイクロプロセッサは、それくらいのコア数に達してきており、破綻が近い
- この 2 重の困難をどう解決するか、という展望が必要。

「解決」の方向

- 90 年代にベクトル並列機が向かった方向を追いかける。
  - VPP500 を 1 チップ化。チップ内でも物理共有メモリを断念して、分散メモリにする
  - GPU は (外部に共有メモリはあるが) これに近い。
  - もっと極端なアーキテクチャが望ましい (GRAPE-DR、ポスト「京」向け「加速部」)
  - これはまあいいんだけど私あきたので今日はあんまりこの話はしない方向で、、、
- 専用回路にいく
  - 半導体が進歩しなくなるので専用回路が陳腐化しない。
  - 同じテクノロジーで、アプリケーションによるが GPU の 5-50 倍程度は電力性能あげられる。
  - 汎用プロセッサに比べると開発が非常に容易 (「加速部」に比べても楽)。

もうちょっと他の応用は？
人工知能とかGRAPEではできなかろ？

- それはそうである。
- では人工知能向けの専用回路って？
- そもそも最近の人工知能研究ってどんな計算しているの？

キー概念：深層学習

深層学習における計算

- 深層学習は多段ニューラルネット。
- ニューラルネット: 入力に係数掛けて足したものが次段のニューロンの入力。係数は同じ段でニューロン毎に違う = 行列ベクトル積。入力が複数あれば行列積。

  なので:

  - 計算の $\mathbf{99.9\%}$ くらいが行列乗算 (BP 学習でも)
  - GPU が使われてるのは結局単精度行列乗算専用マシンとして
  - NVIDIA が倍精度なしの Maxwell も高く売るのを始めたくらいには需要がある。
- 学習は多段ニューラルネットに対して BP が最適かどうかは (私には) 良くわからない。まあでも多分行列乗算ができればいいような方法が今後も使われる。

つまり: 深層学習専用回路=行列乗算専用回路

行列乗算専用回路

- 意外に重要なアプリケーションが色々ある
  - 量子化学計算一般
  - 深層学習
  - **FEM (DDM** とか使うと結局、、、)
  - **stiff** な系が多数ある話 (東大の心臓シミュレーションとか)
- **28nm** で、倍精度乗算器+加算器だけなら **0.02** 平方ミリくらい。**300** 平方ミリで **1** 万ペア。**500MHz** で回して **10TF**、**100-150GF/W** くらいまでいけるはず。単精度なら **40TF**、**300-500GF/W**。
- **28nm** の **GPU** の **10** 倍くらいの電力性能。
- 多分 **28nm** で **2025** 年くらいまで使える。
- **10 nm** にすれば **2035** 年くらいまでいけるのかも。

粒子と行列乗算はいいとして、あとは？

- 「あと」の典型: 規則格子・不規則格子。それらの上での陰的解法。
- 以下個人的意見
  - 陰解法には将来がない。そもそも大規模並列化困難。基礎方程式を書換えて陽解法にしたほうが幸せになれる。
  - 粒子法専用ハードウェアは不規則格子も扱えるように作ることはできる。
  - 規則格子は **AMR** 化が必須になっていくのでそれなら不規則格子で空間差分定式化するのと同じでは？というところも。
  - まあ遅くて高い計算機が使いたい人は使えばいいのでは？それで国プロが方向を誤ったところで私はもう知らない。

まとめ

- 半導体の進歩がスローダウン、終焉を迎えつつあり、また汎用マイクロプロセッサのアーキテクチャの進歩も、90年代に共有メモリベクトル並列プロセッサが辿った道と同じところで限界にきている。
- 今後の計算機アーキテクチャはとにかく電力性能が第一。GPU/加速部的アプローチか専用回路アプローチ。
- 例えば深層学習ならおそらく圧倒的に専用回路 (行列乗算専用回路) が有利。消費電力 1/10 以下、1 ケースの学習を現状の100倍高速化することも可能。GPU では将来になってもあんまり速くなりそうにない。
- チップ開発の費用は安くはないが、先端プロセスではないので比較すれば安価。

まとめ(続き)

- さらに、深層学習でありがちな問題サイズで **1PF** くらいのものなら基板１枚(日立サイズを想定、、、)でできるかも。

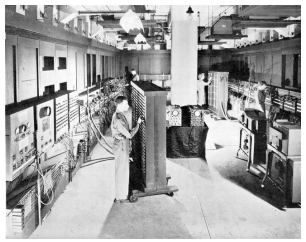
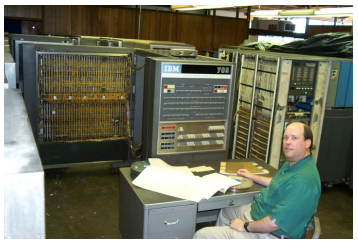
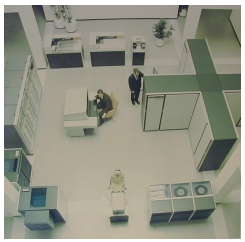
予備スライド

これまでの計算機の進歩

1940 年代から現在までの 70 年間、ほぼ 10 年で 100 倍。何故そのような指数関数的進歩を長期に続けたのか？(今後はどうか？)

基本的な理由:

- 使うスイッチ素子が高速になった
- 使うスイッチ素子が小型、低消費電力になって、沢山使えるようになった
- 使うスイッチ素子が安くなって、沢山使えるようになった (スパコンの物理的大きさは 70 年代が最小。そこまで段々小さくなって、そこからまた大きくなった)


素子の高速化

- といっても、真空管でもそれなりに速かった。
- スイッチング速度が重要でないわけではないが、配線を信号が伝搬する速度のほうが昔から重要。
- 昔は信号はほぼ光の速さでつたわった。
- 最近の LSI 上の配線は非常に細く (抵抗が大きく)、キャパシタンスを充電しないといけないことによる RC 遅延のため、信号が伝わる速度は光速度よりはるかに低い。
- 太い配線に大電流を流せば速いが、大量の電力消費になる

素子の小型化

- 真空管 $\rightarrow$ トランジスタ $\rightarrow$IC という進化は 1970 年代までは重要
- サイズだけでなく、消費電力が下がることが重要
- 80 年代から重要になったのは CMOS LSI の微細化。10 年でサイズが 1/10 になる
- CMOS 素子では (2000 年くらいまでは) 微細化すると電圧を下げることができ、消費電力が下がり、速度は向上した。いわゆる CMOS スケーリング。
- 2000 年頃からは電圧が下がらないので、電力はちょっと下がるが速度は上がらなくなった。いわゆる CMOS スケーリングの終焉。
- そろそろ微細化も困難になってきた。また、トランジスタの構造・製造工程が複雑になり、微細化するとかえって価格上昇するようになった。いわゆるムーアの法則の終焉。

素子の性能向上、サイズ低下と スパコンの性能向上

「スパコンの進歩」は 4 つの時期に分けられる

I スパコンが完全パイプライン化した乗算器をもたない時期 **(1969** 年まで**)**

II スパコンは 1 つ以上の演算器をもつが、1 チップにはまだ演算器 1 つが入らない時代 **(CMOS** では **1989** 年まで**)**

III 1 チップに複数の演算器がはいるが、微細化すると動作クロックが上がり、電力が下がった時代 **(2001** 年頃まで**)**

IV チップに多数の演算器がはいるが、微細化してもクロックが上がらず、電力がへらない時代 **(2001-)**

このそれぞれで、素子の性能向上がどう使われたかは違う。

I期: 1演算器未満の時代 ~ 1969

- この時期には、メモリアクセスのほうが演算より速い
- 演算器の性能向上が最重要課題
- CDC 6600 くらいまで。
- 計算機の構成方法もまだ手探り。
- CDC 7600 で1演算器はいるようになる

II期: 1演算器以上の時代 ~ 1990

- 1つ以上の演算器を有効に使うことが課題
- パイプライン化 (ベクトル命令)、複数の演算ユニットの SIMD and/or MIMD 並列実行が必要になる
- 演算器の数が増えると、メモリとどうつなぐかが課題になる。
- 演算器 16-32 個で破綻する (1992 年頃): メモリと演算器の間のデータ移動回路が大規模・複雑になり過ぎるため
- 末期の計算機: Cray T-90, 日立 S-3800
- 富士通 VPP500 では、分散メモリにすることで当座しのごうとした:ある程度の成功
- 他の方向: 1チップマイクロプロセッサでの分散メモリ。プロセッサ間の通信は細い線でいいことにする。これが 90 年代以降の主流になった

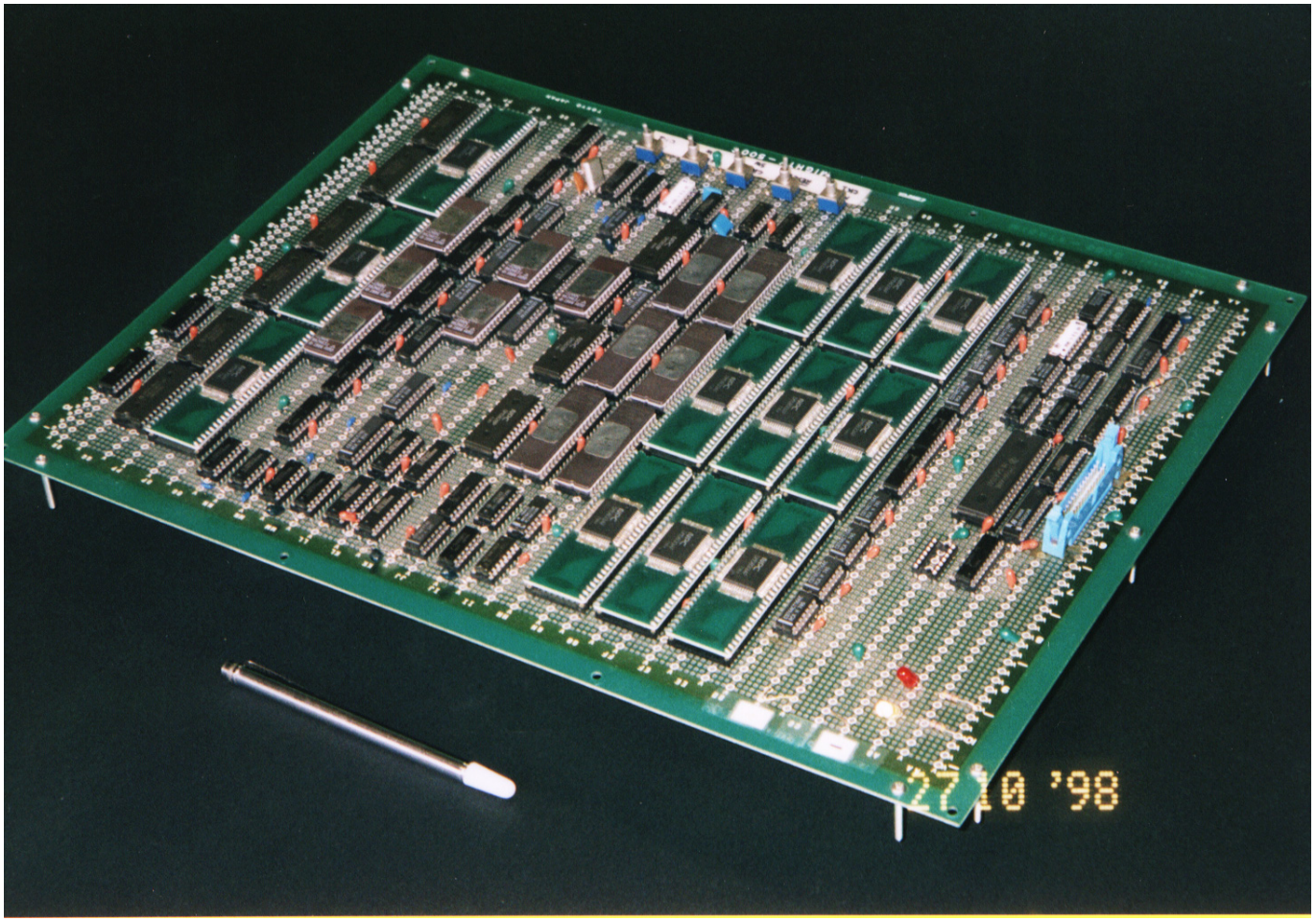
III期: 1チップ高速化の時代 ~ 2005

- 1チップマイクロプロセッサを分散メモリで多数つなぐのは II期末期から
- 1チップに1演算器以上がはいるが、ありあまるトランジスタを演算器を増やすのにはつかわないで動作クロックの向上、キャッシュメモリの大型化に使った時代
- 1990年代。 牧野の意見としては「失われた10年」
- 計算機全体としては II期に起こった問題がチップレベルでも起こるのを先送りにした
- メモリとの接続は速度が不足になった(memory wall): キャッシュでごまかす方針
- 迷走したプロセッサ開発も多い:各社の複数チップ共有メモリプロセッサ。(失敗プロジェクト: ASCI Red 以外の ASCI マシン)
- 末期の計算機: Intel Pentium 4, DEC Alpha 21264

IV 期: 1 チップ多演算器化の時代 ～ 2020?

- 引き続き、1 チップマイクロプロセッサ、分散メモリ。
- デザインルール 130 nm あたりから、動作電圧低下、速度向上に限界
- マルチコア化、コア内 SIMD 化を同時に進めている
- 80 年代のベクトルスパコンの辿った道を追いかけている
- ということは、16-32 コアで破綻がくるはず
- 破綻がくることの予見: Intel Xeon Phi (60 コア)
- GPGPU はちょっと別。

GRAPE-1(1989)

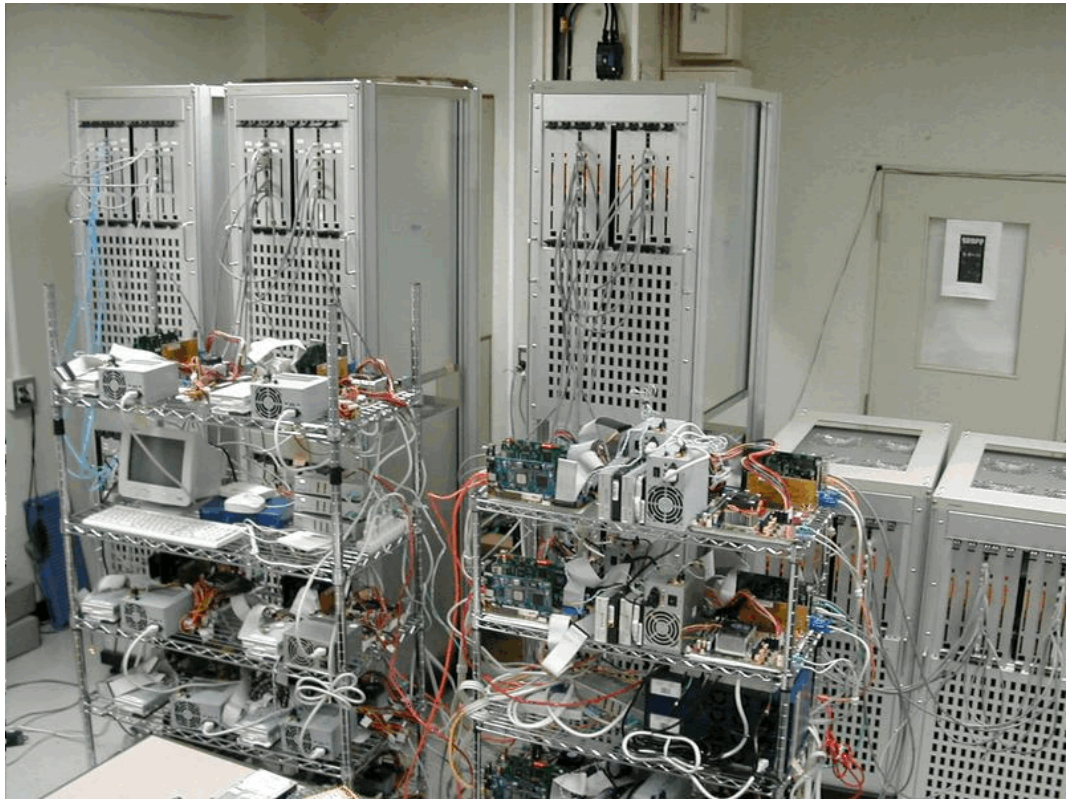


GRAPE-4(1995)

1Tflops を世界に
先駆けて実現

汎用で 1TF 超えた
ASCI Red の 2 年
前、 1/20 のコスト

GRAPE-6(2002)



2002 年の 64 Tflops システム

初代地球シミュレータの 1.5 倍の性能を 1/100 のコストで

深層学習専用にもうちょっと頑張ってみると、、、

数千 × 数千 の行列積を「可能な限り短い時間で」やりたい。

- 大量データに対する BP 学習はあんまり上手く並列化できてない。
- 1つのニューラルネットを多数の GPU でとかは全然無理
- チップ間を高速ネットワークでつなげばなんとかならないか？

例

- 1チップ 30TF (単精度)
- 行列サイズ 2048
- 36 チップを 6×6 の2次元グリッドに
- 行列乗算が 16 マイクロ秒で終わる。
- 1方向の転送速度が 16MB を 16 マイクロ秒で、 1TB/s。
  6 でわって 170GB/s
- すぐ隣のチップとならできるかも。5GHz x 300 本。
- GPU より 100 倍速い。

チップ間通信

- **170GB/s** はまあ今どきなら、、、 **25Gbps SERDES** なら **50** ペアくらい。隣のチップとなのでシングルエンドで **3GHz** くらいでつなぐのでも。
- データ圧縮は多分可能。音声や動画データでは時間相関があるので、予測演算が可能なはず。
- **DNN** の中間層でそんなことができるかどうかは自明ではないが、、、

コストの話

- **LSI** 開発の問題点は初期コストが高騰したこと。現在、**28nm** でマスク代だけで **1** 億はだいぶ超えるはず。
- 設計を渡すインターフェースを **RTL** として、高い **IP** あまり使わないとして **4-5** 億くらい。
- 製造コストはチップあたり **20-30** 万。行列乗算回路の他に色々いれるのでもうちょっと上がるかも。倍にはならない (ように設計する)。