アックスの
計算/分散クラスタ
たけおかラボLinux

2012/MAR/09

たけおか
(株)アックス/たけおかラボ(株)

アックス入ってる

■(独)産業技術総合研究所
知能システム研究部門
ヒューマノイド研究グループとの
実時間Linux共同開発

■ オリンパス デジカメ

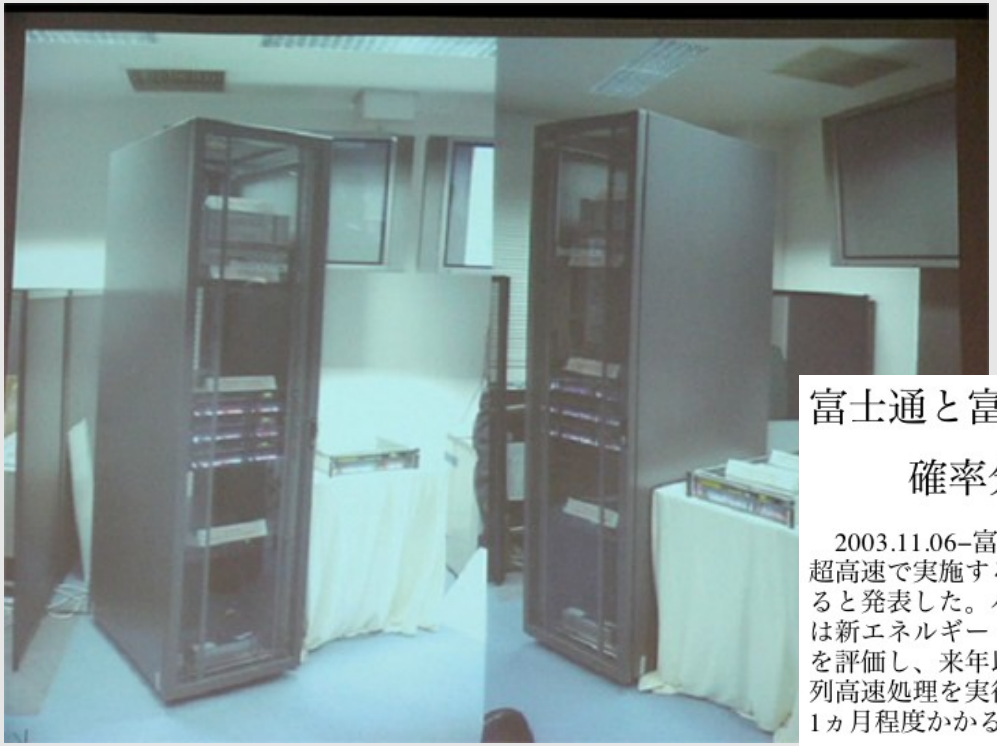
■ シャープ ザウルス

■実時間Linux
航空自衛隊で計測に使用

■ パナソニック
プロジェクタ

アックス入ってる

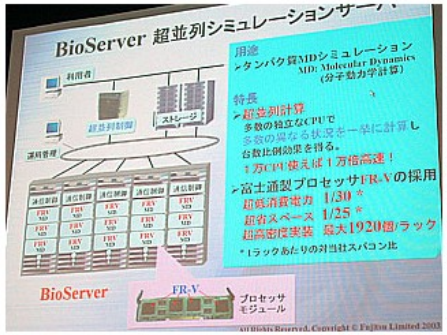
- バイオサーバ



富士通と富士通研究所がたん白質解析の専用サーバーを開発

確率分割法で高並列処理を実現、実証実験を開始

2003.11.06–富士通は5日、富士通研究所と共同でたん白質の立体構造シミュレーションを超高速で実施する専用サーバー「バイオサーバー」（開発コード名）を開発、実証実験に入ると発表した。バイオインフォマティクス分野での共同研究相手であるゾイジーン、さらには新エネルギー・産業技術総合開発機構（NEDO）プロジェクトを通してシステムの実用性を評価し、来年以降に製品化の検討に入っていく。プロセッサー（CPU）の数に比例した並列高速処理を実行できるのが特徴で、計算で60年以上かかっていた処理、あるいは実験で1ヵ月程度かかる解析を12日間で行うことができるという。



今回開発した「バイオサーバー」は、CPUに富士通の組み込みプロセッサーである「FR-V」を採用しており、1ラックに最大1,920個搭載することが可能。これは、最大8命令を同時に実行できるVLIW（ベリーロングインストラクションワード）型プロセッサーで、浮動小数点演算でも4命令の同時実行が可能であり、1ワットという低消費電力で1.33ギガFLOPSのピーク性能を発揮する。CPU当たり256メガバイトのメモリーを積んでおり、OS（基本ソフト）としてはアックス（本社・京都市、竹岡尚三社長）が製品化した組み込み系Linuxである「axLinux」を採用している。

三菱化学の100％子会社であるゾイジーンとの共同研究で使用する1号機は1,920個のFR-V

http://homepage2.nifty.com/ccsnews2/2003/4q/2003_4Qfujitsubioserverptype.htm

アックスの「たけおかラボ」

- アックスの子会社(連結対象)として「たけおかラボ(株)」創立
  - スパコン/並列計算技術/サーバ技術の提供を担当
  - 2012/FEB/14創立
    - 現在、ひとりぼっち
  - GUIは、Web技術
    - CGIは、Lispで記述
  - 大規模データ分析
    - Big data
    - 自然言語、記号処理
    - Hadoop+Lisp

たけおかの超並列計算機実績

超並列計算機「SM-1」の開発

- http://www.yuasa.kuis.kyoto-u.ac.jp/study/sm1/sm-1.html
- 京都大学 湯浅研(当時:豊橋技術科学大学)＋住友金属工業
- 1024PEのSIMDマシン開発
- 演算LSI開発、マイクロコード開発、開発環境の開発
- 並列計算Fortran「BeeFortran」の開発
- 1990～1993
- マイクロコード開発環境は

Lispで記述

たけおかラボの並列計算実績

ATOM 16CPU機の試作

- 東京エレクトロンデバイス社と
- Intel ATOM 16CPUでクラスタ計算
- Linuxをディスクレスでクラスタリング管理
- 簡単なノード管理
  - 数億CPUまでスケール
- Hadoop
- MPI

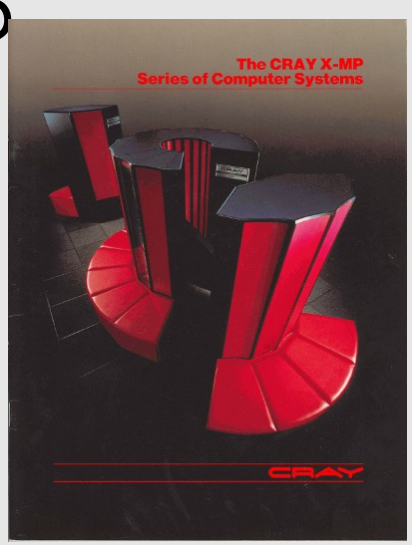
たけおかの一人プロジェクト
ベクトル計算機のオープンなドキュメント和訳

- CRAY X-MPなどのマニュアルがフリーに

http://www.bitsavers.org/pdf/cray/

HR-0032_CRAY_X-MP_Series_Model_22_24_Mainframe_Ref_Man_Jul84.pdf

- CRAY X-MPについて詳しく述べてある
  - 非常に勉強になる
  - ベクトル計算機の使用方法が分かる
  - ベクトル計算機の作り方もわかる



- 日本語翻訳　一人プロジェクト
  - CRAY X-MPについて

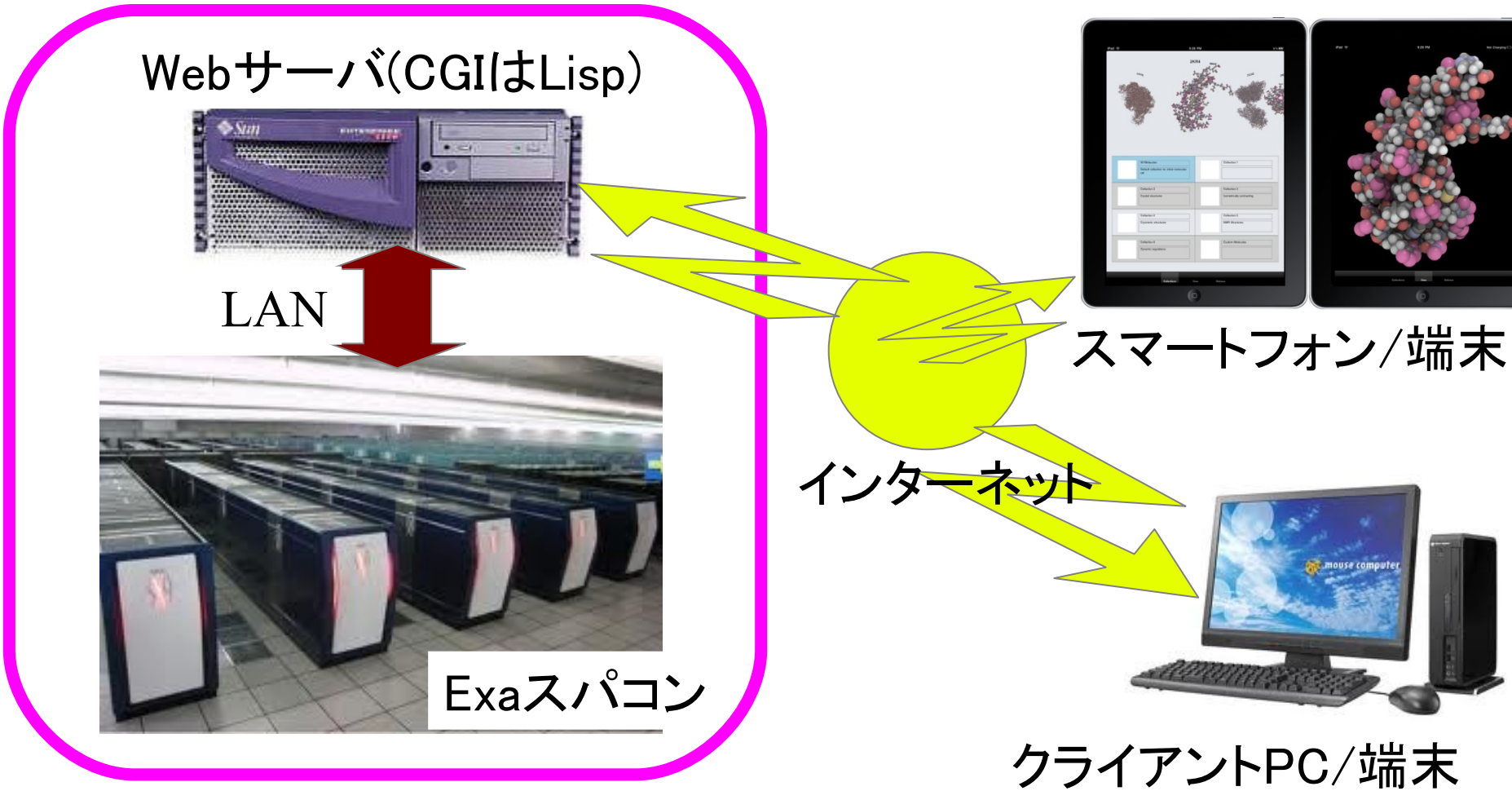
http://www.takeoka.org/~take/supercom/cray-xmp.html

たけおかラボの目指すユーザビリティ

Webベースの利用

Webサーバの後ろにスパコンがある

Webブラウザを使って気軽に利用



たけおかラボが目指すユーザ・エクスペリエンス

- **Webブラウザ、Excel,Scilabでスパコンを利用**
- **オフィス・スパコン**
  - **部門レベルで買えるスパコン**
  - **デスクサイド～1ラック**
  - **特殊な空調は不要**
  - **低消費電力PCクラスタ**

部門スパコン/部門クラスタ計算機

部内LAN

PC

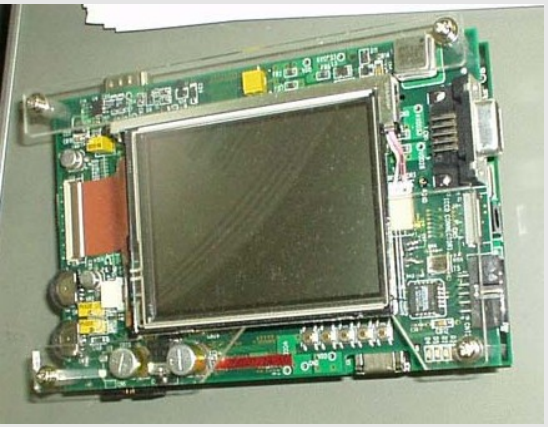
アックスのLinux技術

- 国産CPUへのLinuxポーティング実績 No.1
  - ルネサステクノロジ SH-Mobile, SH2
  - 富士通 FRV
  - 東芝 MeP
  - セイコーエプソン C33, C33ADV
  - IPFlex DAP-DNA2
  - シャープ, サンヨー,セイコーエプソン ARM core

  - M32RへのART Linux移植

国産CPUメーカとの協業

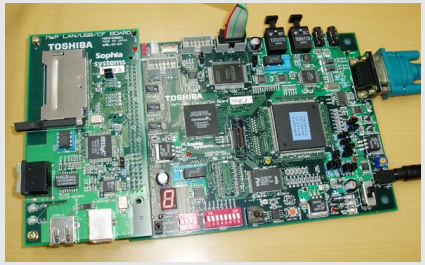
- 国産CPUへのLinuxポーティング実績 No.1



富士通 **FR/V**

ルネサス
SH-Mobile/SH-2A

NECエレ
V850



東芝 **MeP**

東芝CELL

セイコーエプソン **C33**

シャープ**LH795xx**

サンヨー**LC690132**

アックスのSCoreへの取り組み

- PCCCへコードを寄付
  - 2.6 kernel対応(PM通信含む)
  - Opteron, EM64対応

  の作業を行い、コードをPCCCへ寄付

- PCCCメンバー他社とともに、SCore強化のグループを作り、IPAオープンソース基盤整備事業から資金を得てScore強化開発

- アックスはPCクラスタ・コンソーシアム理事です

アックスの計算用クラスタLinux

- Linuxカーネルを変更
- スパコン向けスケジューリング
- 不公平スケジューリングを可能にした
- 指定した特定のプロセスがCPUを長期間得られる
- 計算を行うプロセスを圧倒的に有利にできる
  - キャッシュのヒットミス
  - ページ・フォールト発生
  - TLBミス

の軽減

※組込みLinuxで開発したQoSなどと同じ技術を
スパコンに適用

アックスのスパコン向けスケジューリング

- Linuxカーネルを変更
  - 特殊ファイル /proc/axesched/sched に
    pid チック数
    を書き込む。
  - 指定したプロセスのカンタムが、そのチック数となり、権利を得てから、その期間走行する。
  - その他は通常のLinuxのスケジューラのまま
    優先度の高いプロセスが走行可能になると、CPUを取られる

- 上の進化形として QoS がある
  - CPUの計算時間を、プロセスごとに割合(パーセンテージ)で指定
  - 家電で非常に効果あり
    一般にCPUは非力だが、マルチタスクで仕事を進める

アックスの QoSスケジューラ

- QoSスケジューリング設定
- /proc/axqos/sched デバイスに、
  プロセスID　CPU使用割合%
  の組を書き込む

```
# echo "100 30" > /proc/axqos/sched
# echo "123 10" > /proc/axqos/sched
```

- 存在しないプロセスIDが指定された場合は、何も設定されない。
- CPU使用割合が0または負の数の場合は、QoS設定が解除される。
- また、CPU使用割合が省略された場合も、QoS設定が解除される。
- CPU使用割合の合計が100以上の場合は、合計に対する設定値の割合で、スケジューリングされる。

アックスのQoS制御機能の実装

- 新機構はプロセスごとに、
  品質重み値(p)
  その集積した値(m)
  を管理する

- 全プロセスのカンタムが0になった時に、
  走行可能な全プロセスごとの 各 m に 各 p を加算
  m が100以上のプロセスの、カンタムを1にする
  そのプロセスの m から100を減ずる
  Linuxスケジューラへ制御を渡す

```
t->m += t->p;
if(t->m >= 100){ t->カンタム ++; t->m -=100; }
goto リスケジュール
```

たけおかラボLinuxの特徴

- ScientificLinux6, CentOS6.2 ベース
  - カーネルは axLinux
- Hadoopサポート
  - Big data時代のプラットフォーム
  - Java言語サポート
- GPGPU, IntelAVX 対応
  - x86 (80bit Floatマシン)での64bit計算ノウハウ
    - x86は80bit floatなので、精度の制御が難しい
  - ノード内では、GPGPU, AVXを使用
  - ノード内はOpenMP
- ノード間通信は、MPI, GASNet
  - XcableMP
- g95, gcc, Intel CC, Intel Fortran サポート

コンパイラ

- GNU コンパイラをサポート
  - バグ情報など
- Intel CC, FORTRAN(オプション)
- PGI Fortran (オプション)

- UPC (Unified Parallel C)

URL

- www.takelab.com
- www.axe-inc.co.jp
- www.axlinux.com
- www.sikigami.com
- www.takeoka.org/~take/