

並列 Linux-Cluster

PCクラスタの現状と今後

2001年12月 14日

日本電気(株)

HPC エンジニアリング センター

中西 浩一

目次

1. PCクラスタとは
2. PCクラスタ市場動向
3. NECのPCクラスタ事業への取り組み
4. PCクラスタの今後

PCクラスタとは

目的によって2種類のクラスタ

- High Availability Clusters for Mission critical application
強靱なシステム構築（商用SMPクラスタにより実現）

- High Performance Cluster for Computational science
並列処理による科学技術計算（「PCクラスタ」により実現）

従来のHPCシステムと何が違うのか

→ 以下の機能が比較的安価に得られること

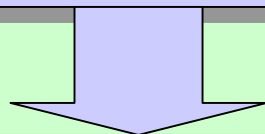
- 高性能
- 高スケーラビリティ
- 高スループット
- 高可用性

米国のHPC市場の背景

かつてMPP (Massively Parallel Processors) が流行

実用HPCシステムとして課題多かった

高価、複雑なアーキテクチャ(OS,HW)、
世代間の互換性問題(システムソフト、AP開発環境)



94年ごろからNASA周辺の研究コミュニティ中心に
コモディティ(HW/SW/NW)製品によるクラスタ開発PJ

Beowulf プロジェクト→Beowulf 型HPCシステム

Beowulf: 中世叙事詩の戦士 (研究者をMPPの悪夢から解放する??戦士)

クラスタリングソフト開発経緯

1990年代前半

コモディティのコンポーネント (CPU, Disk, Network) 性能が
HPCのボトルネックになる。

問題はOS、開発環境、
クラスタソフトが
高価かつ非標準

1994年ごろから標準フリーソフトが充実

OS → Linux が本格化

開発環境 → GNU, MPICH, OpenMP

1990年前半日米ほぼ同時にPCクラスタプロジェクトが発足

RWCP(経産省) → SCore(エスコア)

NASAグループ → Beowulf(ベオウルフ)

科学技術計算

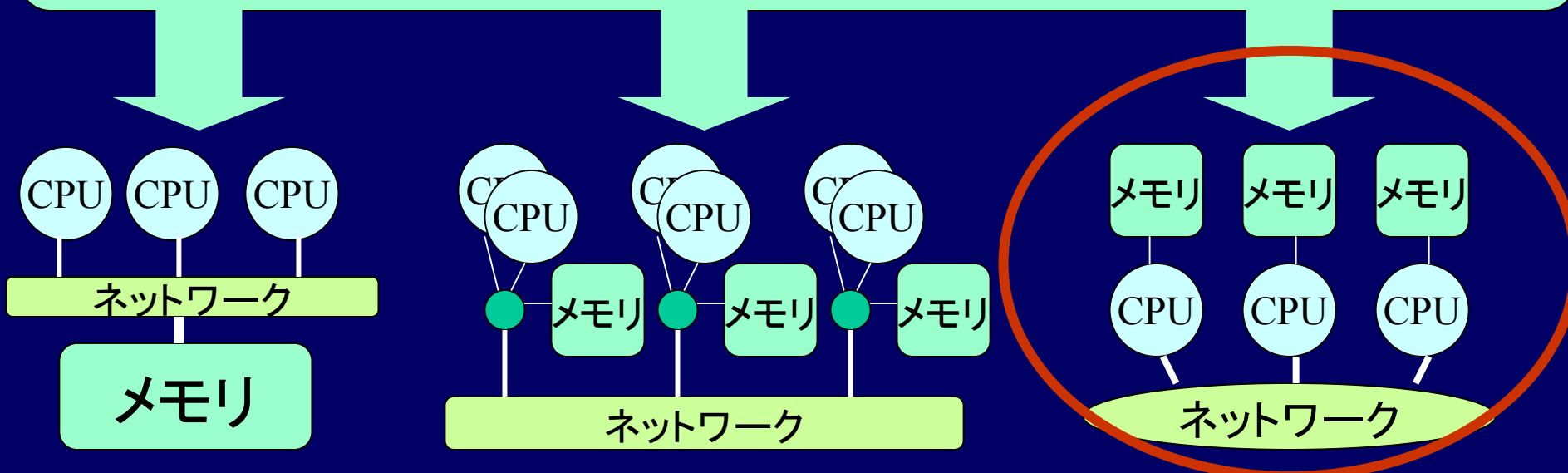
限りのない高速化要求

大規模シミュレーション

- ・地球環境/災害
- ・バイオテクノロジー
- ・衝突解析、構造解析、流体解析

プロセッサ単独の計算速度の限界

(プロセッサ速度に十分な)メモリアクセス速度が困難



コモディティPCクラスタの利点

SMP,分散メモリ,MPPに比べ
格段に優れた価格性能

PCクラスタ:\$1/Mflops、
ベクトル:\$10-100/Mflops

システム要素(HW/SW/NW)は
既存コンポーネントを中心に構成

→利用者にとって選択肢が広い
→ベンダーが示唆する制限から
自由

優れた
価格性能

構築・更新
の容易さ

独自強化
が可能

安定した
技術

運用者、ユーザに
よるシステム構築
(本体、周辺)、運用操作の
カスタマイズが容易

◎リサーチコミュニティにとって
大きな魅力(導入の強い動機)

安定した並列システム
アーキテクチャを提供

長期に渡るプラットフォームを提供し、
持続的な投資を保証。
(MPPの反省)

並列処理技術

アプリケーション

Problems, Algorithm

言語・コンパイラ・ツール

VM, MPI, C, C++, HPF

TotalView, Vampir

システムソフトウェア

Job and Resource management, Parallel job scheduling

Global File System and Parallel File System

Software Raid, Fault Tolerance manager

高性能、高効率、安全な 並列処理システム

コミュニケーション

TCP/IP, Sockets, RPC,

Active messages

Lightweight and

Multi path communication

Load balancing over network

Heterogeneity in networks

オペレーティングシステム

UNIX/Linux

++Device and network drivers

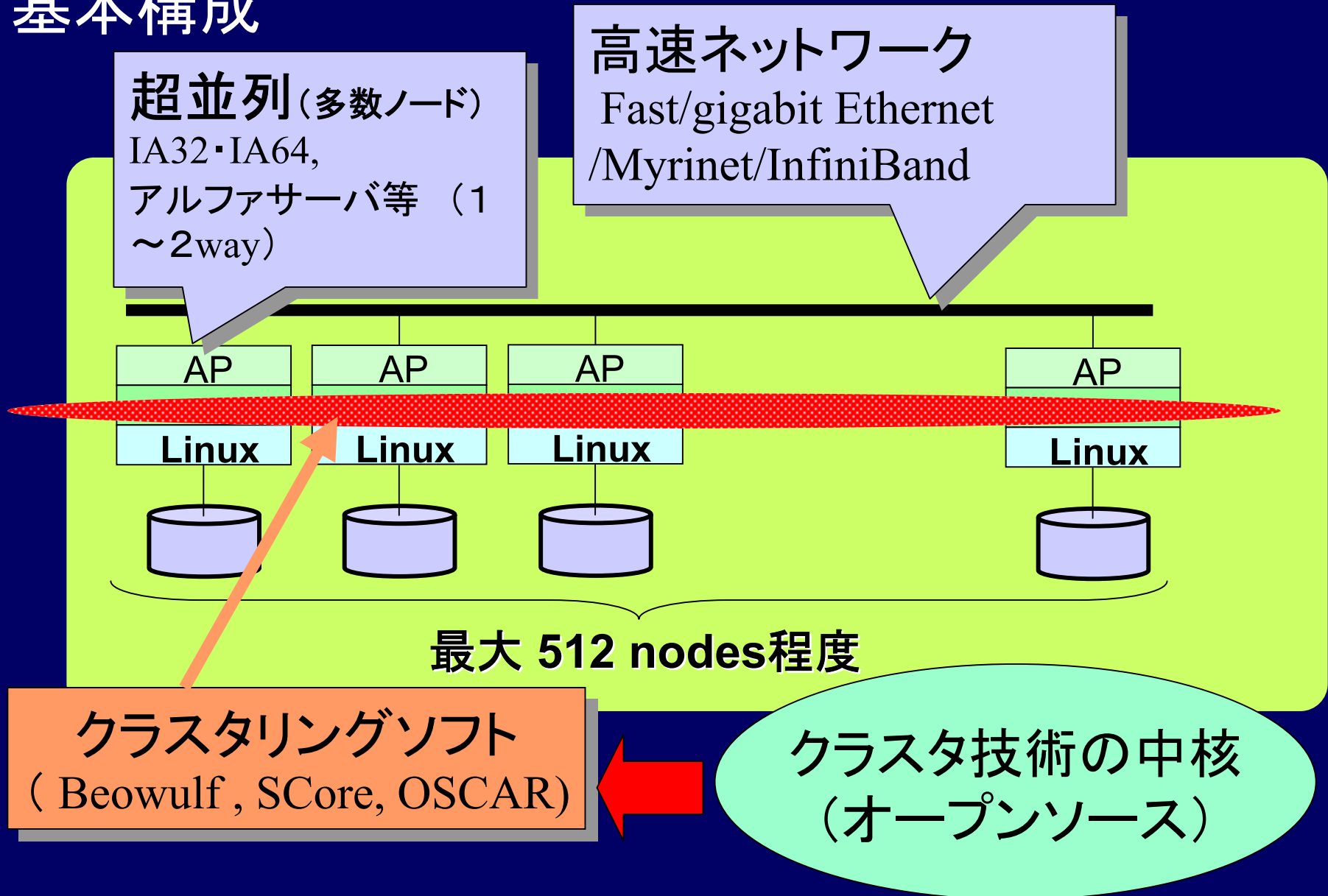
ネットワークハードウェア

HIPPI, Gigabit Ethernet, SCI, Myrinet

InfiniBand, ATM, Broadband network

PCクラスタの基本技術

基本構成



Beowulf型クラスタとは

- 1994年 NASAの2人の研究者 (Thomas Sterling, Donald Becker) によって開発された並列処理システム
- 特定システムをあらわすよりも既成の安価な機材、オープンソフトウェアを用いた並列処理アーキテクチャ
市販コンピュータ+ネットワーク+TCP/IP上の通信ソフトウェア
+フリーソフトウェア
OSはLinux
並列処理プログラミング環境としてPVM, MPI
- 開発・強化・サポート
Scyld Computing Corporationによる次世代Beowulfの開発
最初のコマーシャルリリース (The Scyld Beowulf distribution)
必須のスーパーコンピューティング機能の強化
(Gang, Checkpoint, InfiniBand)

SCore

- COTS ベースの PC クラスタシステムソフトウェア
- 経産省 Real World Computing Partnership
によって開発
- ハードウェア性能（特にノード間ネットワーク）を
引き出し、ユーザフレンドリな
並列プログラミング環境を提供
- Freely distributed as an Open Source
- URL <http://www.rwcp.or.jp/lab/pdslab>

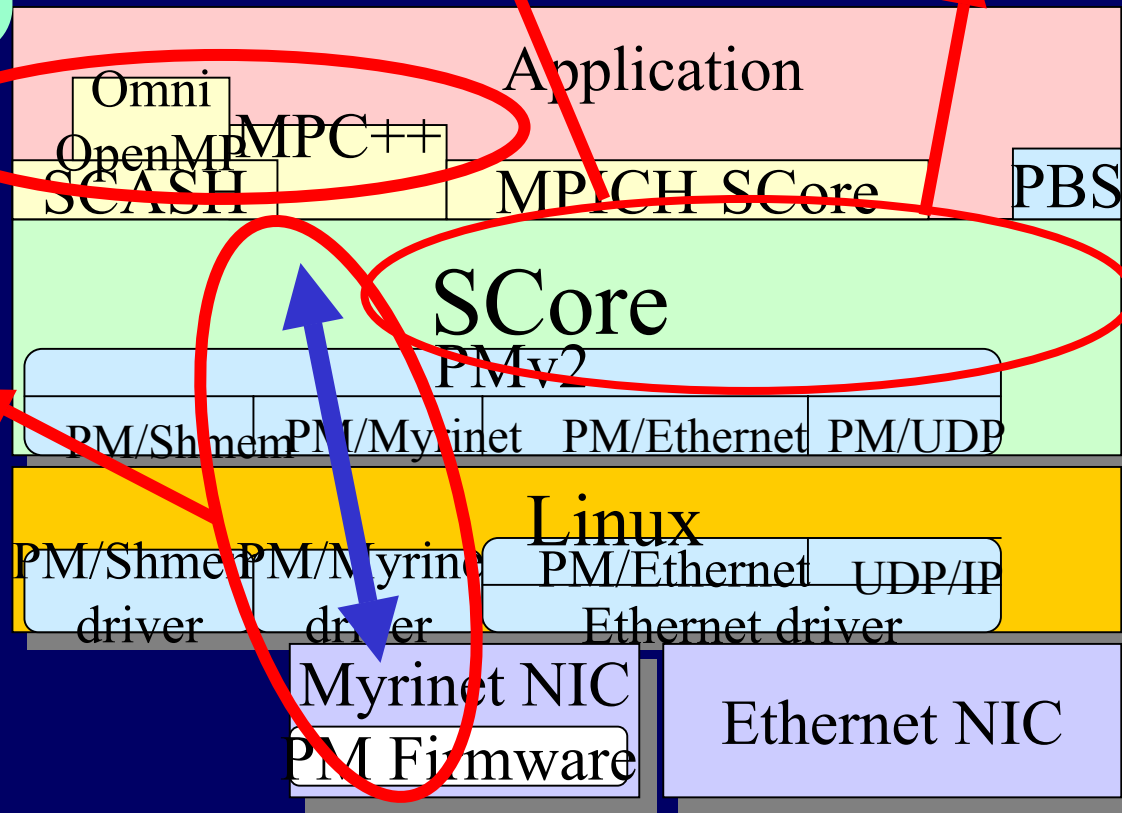
SCore (経産省、RWCPが開発) の先端機能

豊富な並列
プログラミング
環境

シングルシステム
イメージの実現

並列ジョブの効率的
スケジューリング
HA機能

OS介さない
ゼロコピー
高速通信



ユーザ
レベル

カーネル
レベル

NIC
レベル

SCore の特徴 (1/4)

- PM II : 高性能・低レベル通信ライブラリ

PM II API は、多くの種類のネットワークや共有メモリに同一の方法でアクセス可能。 Myrinet、Ethernet、Shmem

-TCP/IP protocol handling ではなく Zero-copy message transfer

-Ethernet ネットワークトランッキング 複数NICで通信バンド幅向上

-シームレスコンピューティング

100 Ethernet を使った小規模クラスタからGigabit Ethernet 又は Myrinet を使った大規模クラスタまで(APをリコンパイルすることなく)連続的に利用可能

MPICH-SCore の性能

Network (*)	Bandwidth (Mbytes/sec)	RTT Latency (microsecond)
Myrinet (Lanai9 in 33MHz PCI64)	226.9	24.2
Gigabit Ethernet	73.4	61
Beowulf/Gigabit Ethernet	46.7	89.6

(*)Measured using NEC Parallel PC-cluster:64 NEC Express Server
(Dual Pentium-IIIS,1GHz and 1GByte Main Memory) with SCore 4.2

SCore の特徴 (2/4)

SCore-D

ユーザレベルのグローバル オペレーティング システム
(各計算ノードでデーモンプロセスとして走行)

(1) Single System Image の実現

- Single Point of Entry

利用者は単一ノードとして接続

PM ネットワーク デバイスは SCore-D によって管理

SMP クラスタや異種混在クラスタも、単一種クラスタと同様にサポート

- Single Job Management System

透過的なジョブ投入システムにより外部からジョブ投入

並列ジョブ スケジューリング機構を提供

ユーザの並列ジョブは同時にかつ効率的に時間と空間が共有

先制的で一貫したチェックポイント

- Single Point of Management and Control

単一ツール、単一ウィンドウから全体の管理・監視制御

SCore の特徴 (3/4)

(2) ジョブスケジューリング

・ジョブ属性 (scrunch)

ホスト数、CPUタイプ・スピード、ネットワーク、優先度 (0,1,2)、最大CPU時間
メモリ/ディスク制限

・待ち行列属性

優先度

スケジューリングモード

マルチユーザモード: 複数のジョブを時分割方式で実行(Gang Schedule)

シングルユーザモード: ユーザがクラスタを排他的に利用可能

最大 CPU 時間、待ち行列残留時間、メモリ制限、ディスク制限

・Score-D によるスケジューリング

(3) チェックポイントリスタート

- ユーザレベルでの実装

数値計算に代表される大規模並列処理において実用的

プロセスメモリイメージを計算ノードのローカルディスクへ退避

ネットワーク状態もファイルに退避し復帰

→ チェックポイント時間はメモリ使用量とファイルシステムの
バンド幅に依存

SCore の特徴 (4/4)

(4) 並列プログラミング環境

- MPICH-SCore

MPICHバージョン1.2 をベースにつくられたMPIライブラリ
(MPI-1準拠)。

SCore-Dのもとで動作→SCore-Dによる、ジョブスケジューリング
PBSによるバッチ処理も可能

- PVM-SCore

- Omni OpenMP

ソフトウェアDSM(SCASH)を使って OpenMP プログラムが
クラスタ上で走行可能

SCASH: 分散メモリシステム上でソフトウェアにより仮想的な
共有メモリを提供。(Myrinet/PMにより低レイテンシ、広通信帯域)

- MPC++

C++の拡張 並列プログラミング環境サポート

PCクラスタコンソーシアム

URL <http://www.pccluster.org>

コンソーシアム設立目的

技術研究組合新情報開発機構 (RWCP) が開発した SCore クラスタシステムソフトウェアを中心にした クラスタシステムソフトウェアの開発の継続、維持、普及を通して、PC クラスタシステム市場育成に貢献する。

- **2001年10月5日設立**
- **非営利の任意団体**

経産省 技術研究組合新情報開発機構 (RWCP) を母体

- **国内外のユーザ、ベンダーにより組織**
- **NECはコンソーシアムのメンバーの一員として**
積極的に参加し、SCoreの強化をはじめ優れたPCクラスタ並列処理環境の開発、提供により市場発展に貢献

活動内容

- ・SCore ソフトウェアの開発、維持
- ・SCore ソフトウェアの標準配布
 - Linuxスタンダードカーネルに対するSCoreソフトのためのパッチ及び
 - SCoreソフト一式
- ・新しいハードウェアへの対応

- ・PCクラスタに関する情報収集 & 公
 - 年3-4回の会合
- ・PCクラスタに関するベンチマーク & 結果の公開
 - ・他研究団体との交流

開発部会 調査・
評価部会
AP技術 普及部会
支援部会

- ・アプリケーション移植支援
- ・SCoreに関する技術支援
- ・アプリケーション拡大に必要な活動
 - 年数回の会合によるknow-how蓄積

- ・WWW運営
 - PCクラスタのポータルサイト
- ・PCクラスタ普及推進のための講習会、技術交流会、展示会の実施

開発および普及が大きな柱

全ての情報は、まとめて、WEB、シンポジウム、交流会等で発信

法人：正会員17団体、準会員4団体 個人：正会員6名、準会員1名 計28団体・名

PCクラスタ市場動向

何ができるか、どう構築するかを模索

米国:DOE ASCI クラスタシステム(COMPAQ Alpha 1,2000基, ピーク30 Tflops)から
Beowulf 型のPCクラスタまで広く一般的に採用
研究開発コミュニティに急速に拡大(米国で>数千サイト)

特に米国市場で成功、高い評価

米国政府プロジェクト(大規模Beowulf型クラスタ)の構築
バイオ中心に適用領域が急速に拡大
→ バイオ/化学/物質工学 エンジニアリングシミュレーション

日本国内でもバイオ中心に立ち上がる

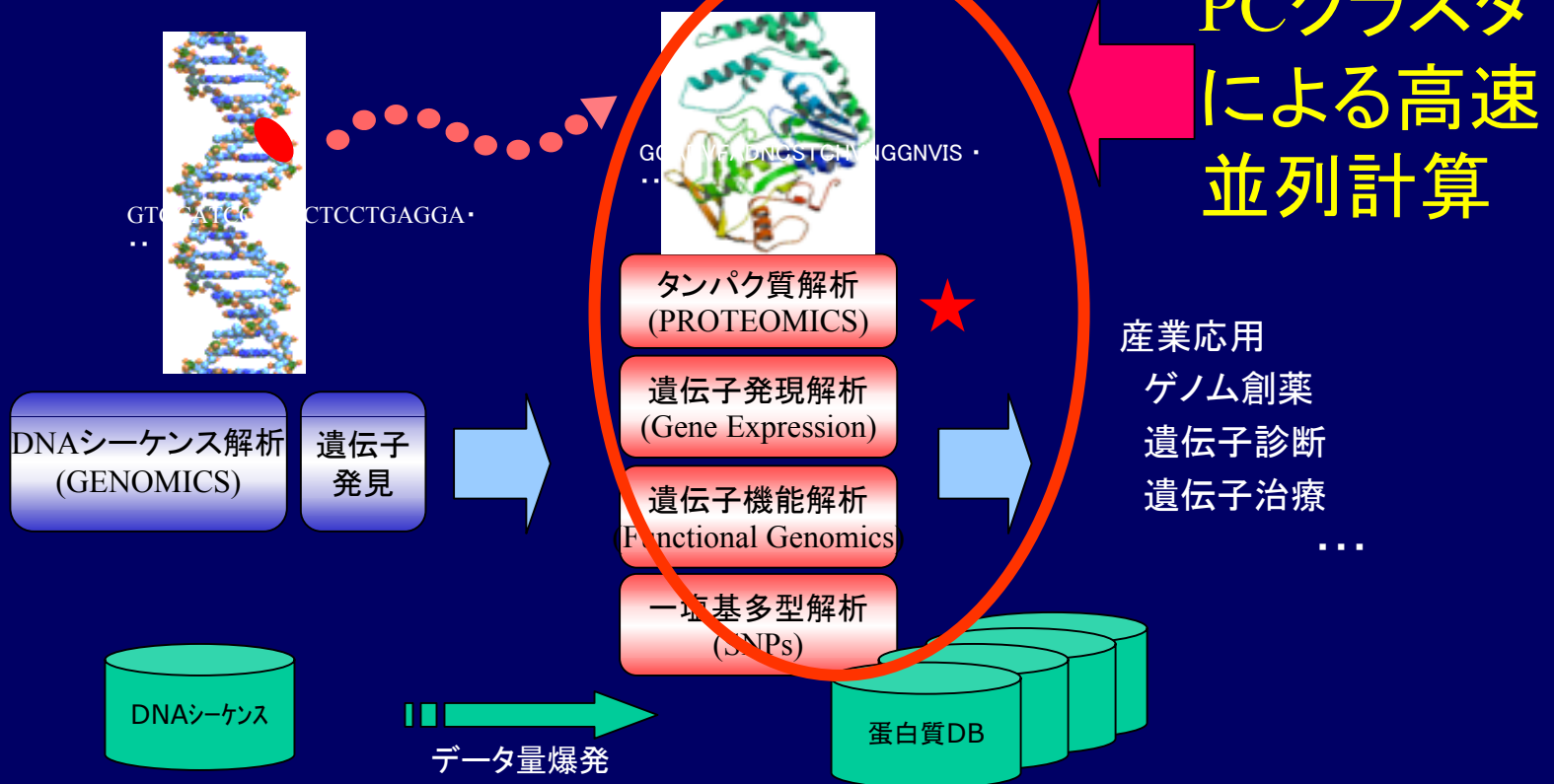
バイオインフォマクス市場を中心にPCクラスタ市場が
急速に立ち上がる

実用APの充実へ

アプリは物理、化学、生命情報工学等での大規模データ検索、構造解読ソフト等
研究分野から多様なインダストリ分野(自動車、環境)へ拡大(さらにビジネス分野へ)
(欧州では衝突解析での利用が一部で開始へ)

不足するCPUパワー (ポストゲノム)

- ・バイオ関連データが爆発的に増大。
- ・蛋白質の立体構造から、蛋白質-蛋白質(蛋白質-化合物)の化学反応を高精度で予測。生体内反応現象をコンピュータ上で再現
- ・4-5年スケールの研究で10万個CPU(Pentium4換算)



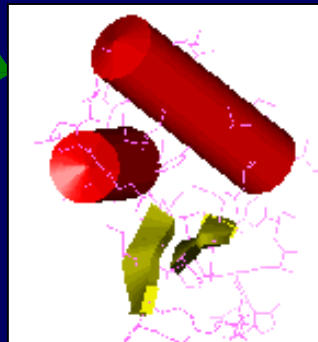
PCクラスタの活用事例；バイオ

産業技術総合研究所 生命情報科学研究センター(CBRC)様

薄型Expressサーバ×520台(1040CPU)、SCoreクラスタシステム、
Myrinet2000ネットワーク



背面ケーブリング



タンパク質立体図イメージ
(PCクラスタで計算)

バイオ研究領域でPCクラスタを利用。

<主な研究項目>

- ・ 遺伝子発見(ゲノム解析)
- ・ タンパク質単粒子画像解析
- ・ 分子動力学計算
- ・ タンパク質構造予測

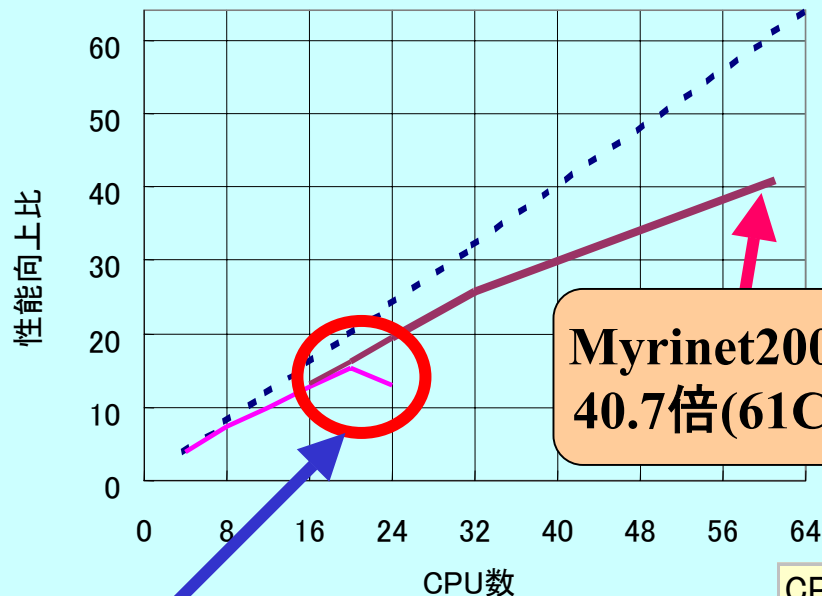
スーパーコンピューターTop500
において商用PCサーバのクラスタ
として世界最高性能を記録。

適用分野の拡大

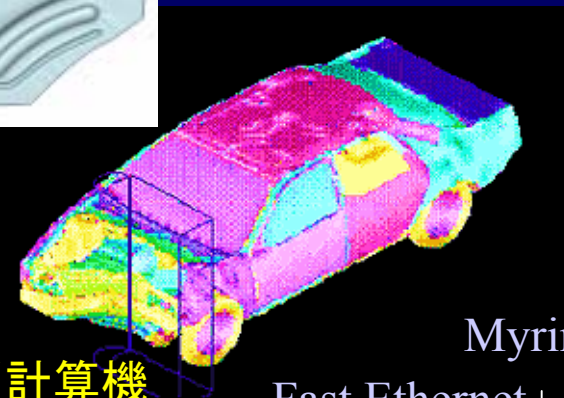
構造解析、衝突解析、計算流体力学(CFD)等

LS-DYNA

(解析モデル:NUMI99)



オフセット衝突解析



Myrinet

UNIX 並列計算機
との比較(単位:時間)

Fast Ethernet

100Ethernetでは16~24cpuで
通信のボトルネックが顕著

測定結果A~Dは '00 LS-DYNAユーザー会
論文集より (MPPDYNA Ver.940.2a 使用)
ケース1, 2 (MPPDYNA Ver.960 使用)

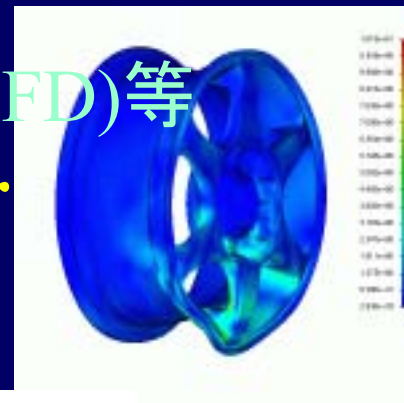
CPU数	A	B	C	D	ケース1	ケース2
4	19.31		16.78	17.5	22.22	
8	10.34		9.31		12.25	
12			6.57	7.22		
16	5.57	6.6			6.97	6.7
20					5.78	
24					6.85	
32	3.02	3.99		3.05		3.45
61						2.18
64		2.27				

適用分野の拡大

構造解析、衝突解析、計算流体力学(CFD)等

商用・大規模並列CAEシステムADVentureCluster

分散メモリ型並列コンピュータ上で稼動



(例)モデル: エンジン・シリンダヘッド

—4面体1次、270,934要素、67,580 節点、約20万自由度

■ 解法: Lanczos法 ■ 求める固有値: 最低次より10個

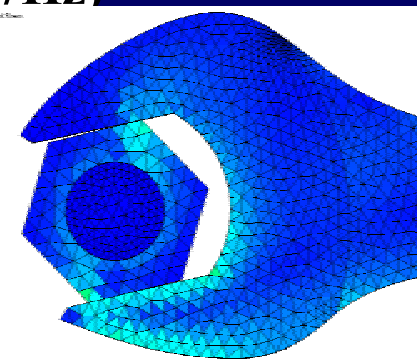
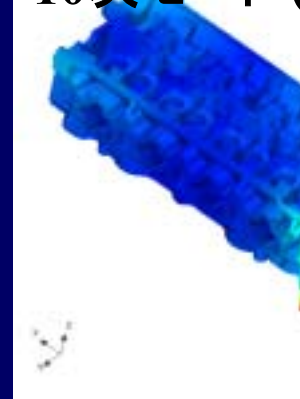
■ 質量マトリックス: 集中質量マトリックス

● NEC HPCエンジニアリングセンタ 64ノード

クラスタマシン (Pentium III 1GHz 1GB)

の16ノード(分散型)を使用

10次モード (4257Hz)



固有値解析のパフォーマンス

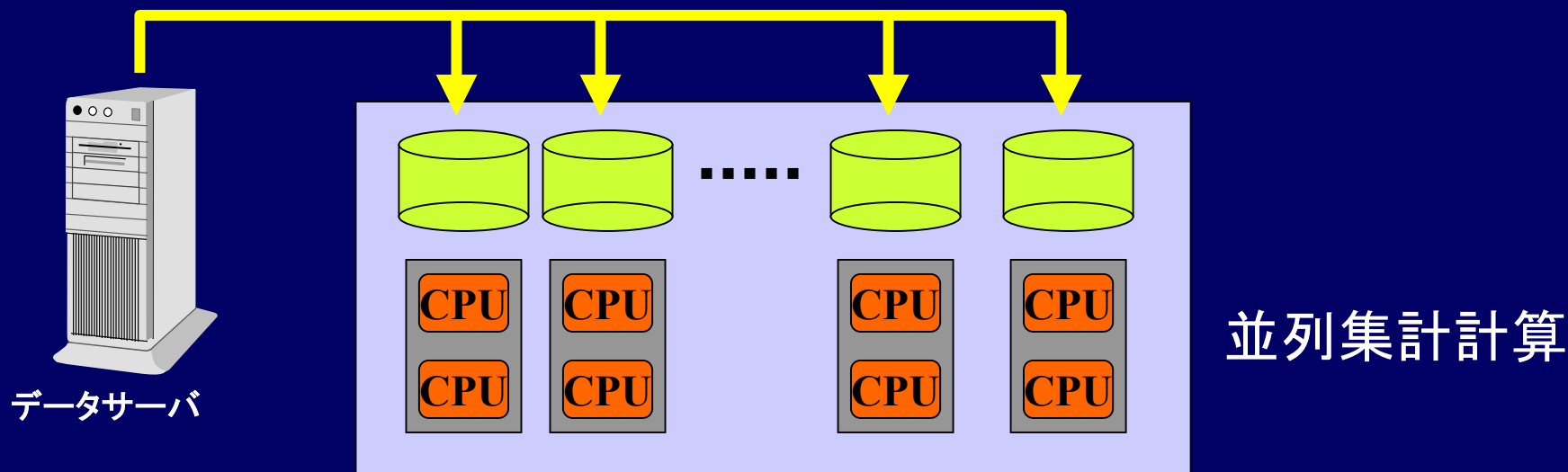
モデル	要素	自由度	求めた固有値の数	CPU数	メモリ	計算時間
シリンダヘッド	4面体1次	202,740	10	16	256MB	27分
エンジン部分(2)	4面体2次	4,273,260	5	32	10GB	9時間20分
エンジン部分(3)	4面体2次	1,316,685	10	32	2.4GB	7時間

適用分野の拡大

ビジネス系APへの展開

PCクラスタによる超高速並列集計システム

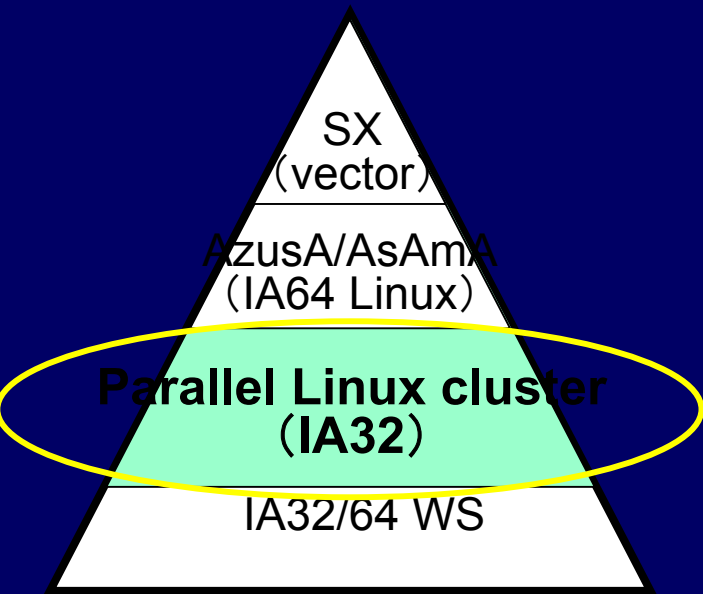
データベース構造を持たない膨大なフラットデータ
の高速並列処理技術



高速LAN接続された多数PCによる
並列集計処理, 並列計算処理(MPIでデータ分散)

- ・大量データの超高速集計ソフト
- ・大量データの超高速検索エンジン

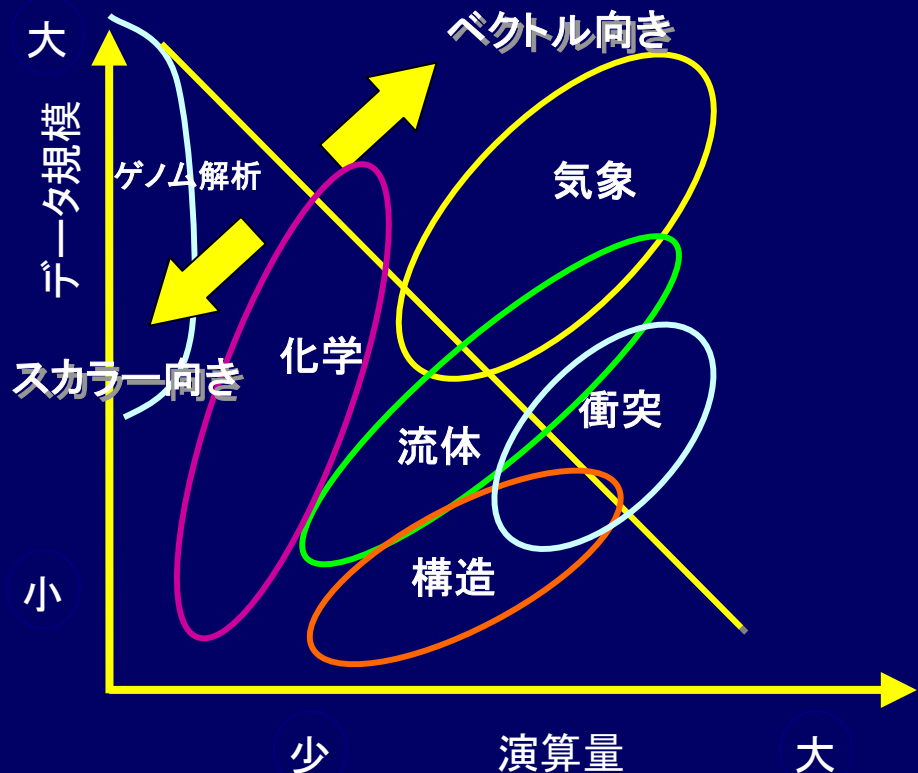
NEC 並列 Linux クラスタ事業



- シームレスな HPC 製品の提供
- 柔軟な HPC 環境
- HPC 市場の拡大

- Cluster Platforms

- IA32 Server (1 – 2 way)
- IA32 Server (4 – 8/16 way)
- IA64 Server (2 – 4 way)
- IA64 Server (> 4 way)



Ex. Rack mount

- 1Uラックサーバ「Express5800/120Ra-1」
2Uラックサーバ「Express5800/120Rc-2」

- オンボード100Mbpsイーサネット
Myrinet2000インタフェース
Gigabit Ethernet

- RedHat Linux
SuSE
Turbo Linux

- SC Dual Pentium III**
Beo Max1.4GHZ

Max Mem:4GB

Max Disk: 72.6GB

PCクラスタ構築に必要な **Max Disk: 72.6GB** を
検証し、短期間で確実にシステム構築

[illegible]

小、中規模クラス

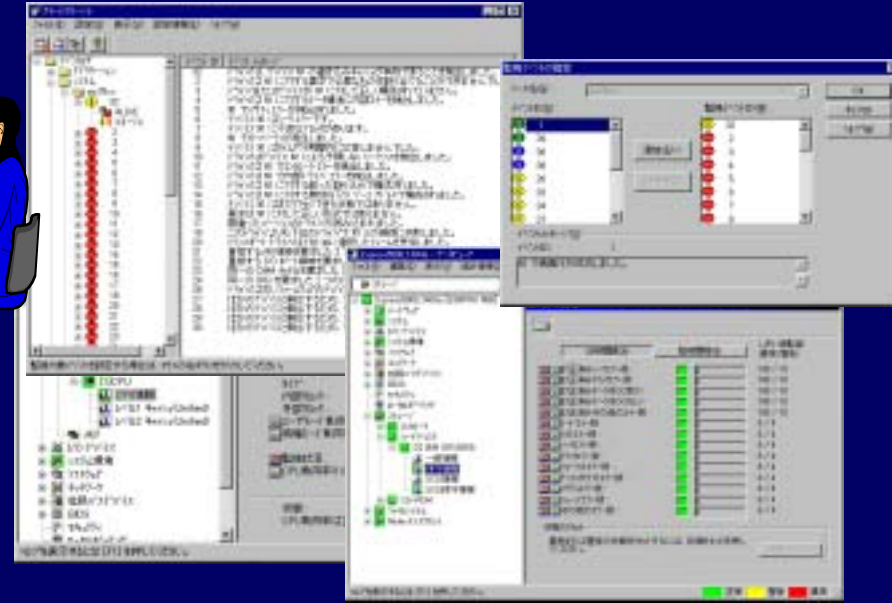
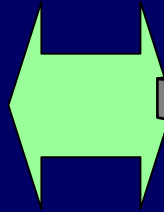
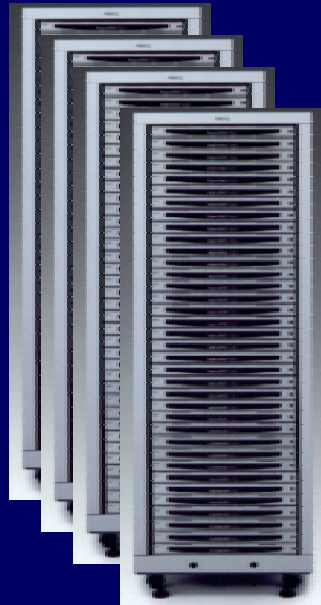
容易なスタートアップパッケージと
迅速な技術サポート(並列APチューニング含む)

大規模クラスタ

(トップ性能 and/or センター運用重)

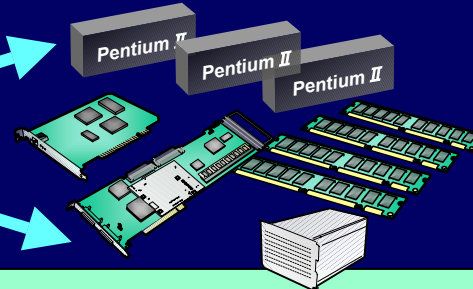
ユーザと一体になった高性能システム
(大規模)クラスタの構築・運用

システム監視



マザーボードにオンボード 運用管理チップ(BMC)を搭載

オンボード
運用管理チップ

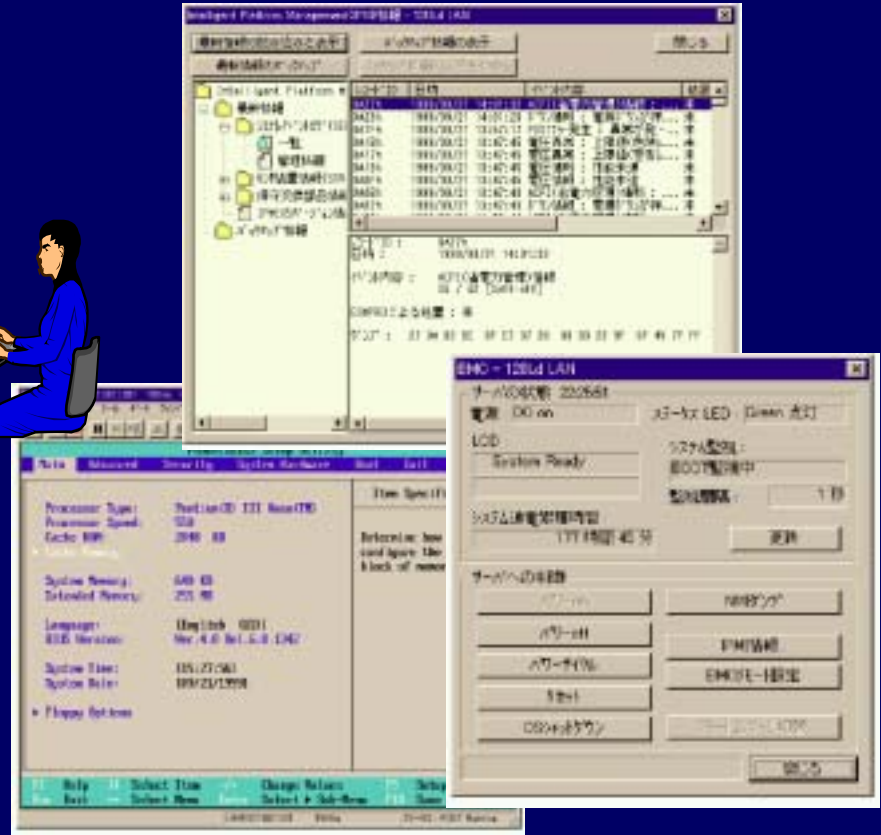
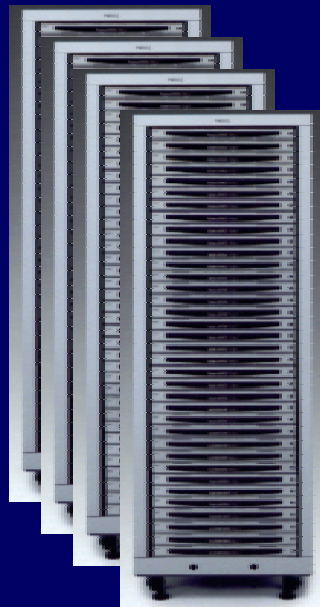


拡張RAS機能

- ・ H/Wログ自動採取
- ・ H/W構成管理/障害監視
- ・ 稼動状況監視
- ・ 温度/ファン監視
- ・ リモート保守機能

- ・ RAS管理 (CPU監視, メモリ監視, 温度監視, 筐体電圧監視, SMB監視・・・)
- ・ ストレージ管理 (ファイルシステム監視, ストレージデバイス監視)
- ・ ネットワーク管理 (LAN監視)
- ・ イベント管理 (スツール監視, システムエラー(パニック)監視, イベントログ監視)

リモートコンソール



OSを介在せずにマシンを制御

- リモートコンソール機能 (BIOS設定, Self Test画面表示)
- リモート電源制御 (Power On/Off, OSシャットダウン)
- リモート情報収集 (ハードウェアログ, 構成情報表示)

HPC エンジニアリングセンター

(HPEC (ハイペック))

・HPCクラスタコンピューティング技術の研究開発センタ (コンピュータソフトウェア事業本部内)

PCクラスタ技術の開発

(RWCPプロジェクトの成果である
システムソフトウェア(SCore)を継承し、発展させる)

HPC関連APのポーティング・チューニング

SI/SE 技術支援

HPCユーザとの交流、世界への情報発信

センター内設備 (64ノード,128CPU)を
外部利用者(研究機関,ベンダー,ISV)へ開放

- ・ベンチマーク
- ・システム検証
- ・APポーティング・チューニング

● 問い合わせ

・ URL : <http://www.nechpec.com/>

・ E-mail : hpec-all@hpec.nec.co.jp

・ TEL : 03-3798-9991

・ 住所 : 東京都港区芝浦4-16-23
アクアシティー芝浦ビル5F

HPEC 活動内容

1. クラスタ技術開発・バイオ等並列AP開発の交流の場

- ・ HPC 研究者・組織・ISVベンダーとの交流、情報交換

大学、国立研究機関（並列処理、バイオ、Grid 等）
研究会、学会への参加等

- ・ PC Cluster Consortium を通しHPC市場発展に貢献

4つの専門部会で積極的に活動

・開発部会 ・調査・評価部会 ・AP・技術支援部会 ・普及部会

- ・ ユーザ・プロスペクト（企業、大学、研究所）によるセンタ利用

2. システム支援

システム構築,技術支援

ユーザと一体になった高性能
システム化
(大規模) クラスタの構築
・ 運用支援

3. クラスタソフトの開発、 並列APの移植・性能チューニング

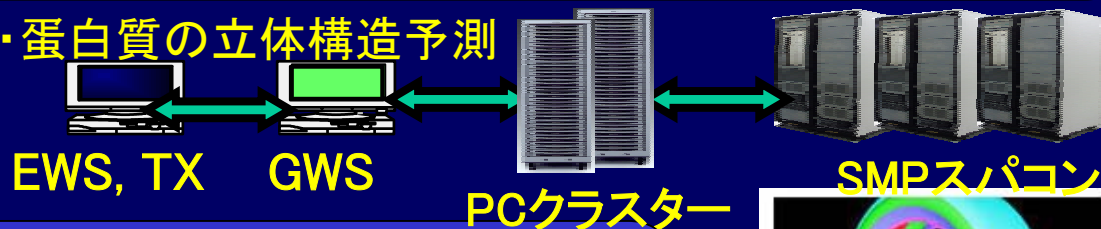
PCCCを通したSCore機能の開発

+付加価値機能の開発
AP移植/ ベンチマーク

PCクラスタ の今後

コストパフォーマンスの優れた
大規模計算・データサーチエンジン
(バイオ・ナショナルプロジェクト関連)

・蛋白質の立体構造予測



PCクラスタ
ビジネス拡大
・高性能/高効率
・適応領域拡大
・グローバル

民需市場への拡大

(衝突、構造、流体
シミュレーション等から
ビジネス系APへ)



Grid 関連インフラ(計算ノード、データベース)
(ナショナルプロジェクト、キャンパスネットワーク)

広域分散コンピューティング

地球規模での巨大データを扱う大規模計算

本格化するバイオビジネス

ヒトゲノム配列の解読(2001年)＝ヒトの設計図がデジタルデータ化
ターゲット探索 リガンド探索・最適化

遺伝子・蛋白質データ解読
個人差解読(SNPs)

ターゲット蛋白質と薬剤候補との
分子レベルでの結合シミュレーション

データ量・情報処理量の級数的増大

大規模計算の必要性

大規模DB, 並列サーバ, データマイニング

スパコンによる大規模シミュレーション

ゲノム解読によって蓄積される
膨大なデータからの知識発見

分子レベルのシミュレーション
による生体機能の予測

PCクラスタ



スパコン

「バイオ」はIT応用の巨大市場

Grid Computing と Cluster Computing

GRID computing とは (= Global computing)

electric power GRID(電力供給網)

- ・遠くの高性能計算機を手元にあるかのように使う
- ・地理的に離れた異なる種類の資源を組み合わせる

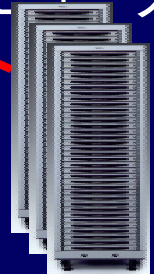


PC

記憶装置



スーパー
コンピュータ

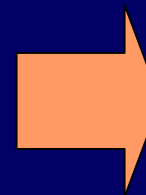


PCクラスタ



測定装置

- ・単一アプリの並列分散化による高速化
- ・逐次プログラムの分散バッチジョブ (parameter sweep 等)
- ・遊休計算資源の有効利用 (SETI 等)
- ・複合シミュレーションのカップリング
- ・遠隔地に配置した多数のセンサのリアルタイム処理
- ・巨大データを扱う大規模計算
- ・巨大データのデータベース処理 (data mining)
- ・遠隔協調作業



計算力供給網

GRIDコンピューティングの現状

- 現状認識

- 現在は、HPC(High Performance Computing)ドメインでの研究が主
- アプリケーションが未成熟(どのように、GRIDインフラを使いこなすのか。どんな新しい可能性を生むのか)
- インフラ技術としては、どんどん良いものが生まれている
- これまでは計算機科学者だけの研究対象だったが、応用の研究者がGRIDに挑戦する下地ができた : Globus

- 米国、欧州では国家プロジェクトとして急速に立ち上がりつつある。

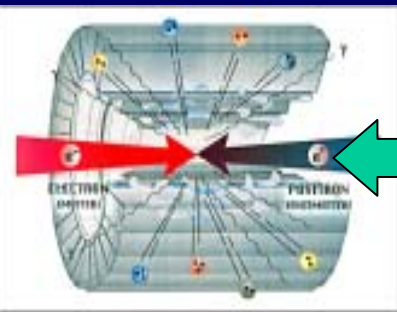
- 全世界で約30のGridシステム
- The TERAGRID (米国4主要研究センタをブロードバンドで接続)
\$53M, 11.6TFlops, 450Peta Byte, 40Gbps
- National Grid (英国)
- そして日本でも。。。 (産総研、高エネ研、東工大、東大、阪大、、、)

GRIDコンピューティングの応用

Grand Challenge 問題への適用

(Tera Flops, Peta Byte, Tera Bps)

- ・ バイオ(蛋白質構造決定、創薬)
- ・ 高エネルギー物理
- ・ 高精度気象予報、気候/災害シミュレーション
- ・ 天文学、地球物理学
- ・ ニューロサイエンス
- ・ 知的モバイル環境



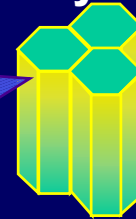
Data from
detector

Input:

1-100 GB/s

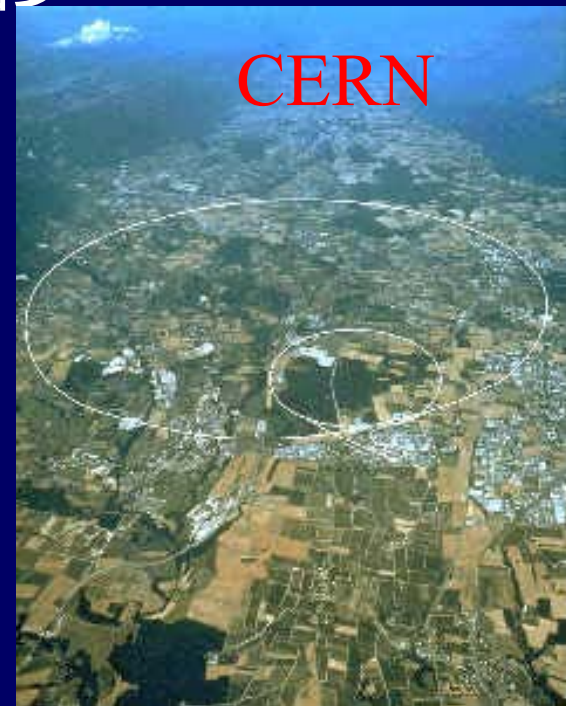
raw data

~ 1 Petabyte/year



summary data

1-200 TB/year



CERN

実用 Grid Computing への道

- ・実地的な問題

セキュリティ、バンド幅、 、 、

- ・New Technology への「実験装置」

Chipmaker から商用アプリケーション

今世紀IT革命のキーテクノロジー

- ・学術研究領域とビジネス両者の利益

世界のコンピュータメーカーの本格的参加

商用 Grid Computing ソリューション

既存の分散コンピューティングソフトウェア

(ジョブスケジューラ、システム管理、、、)

と Globus Toolkits との結合

今後の課題

(Cluster Computing の立場から)

大規模計算サーバ及び ストレージサーバとしての基盤

- 超高速ネットワーク
10/100GB/s, 1TB/s < LAN
OC12/48 ATM, Broadband network
- Peta-scale ストレージシステム
高速並列 I/O (リモートファイルシステムを
アクセスするための並列 I/O インターフェース)
並列ファイルシステム、ソフトウェア RAID

資源の共有

- API から connectivity までの
セキュリティメカニズム
TLS, Kerberos, , ->
多様なローカルセキュリティ方式の統合
- ディレクトリサービス
名前, 属性, availability 及び負荷による資源の
問い合わせ確保

動的システム構成

- Plug & play
- Heterogeneous nodes

システム/ ユーザ 管理

- アカウンティング, 資源の使用制限
- 動的資源監視及び管理

フォルトトレラント

- ヘテロ環境での Checkpointing
- ファイルオーバ

並列プログラミング環境

- MPI in heterogeneous
distributed computing

今後の課題 (Cluster Computing の立場から)

HPCユーザにとって最も重要なことは？ (IDC 2001)

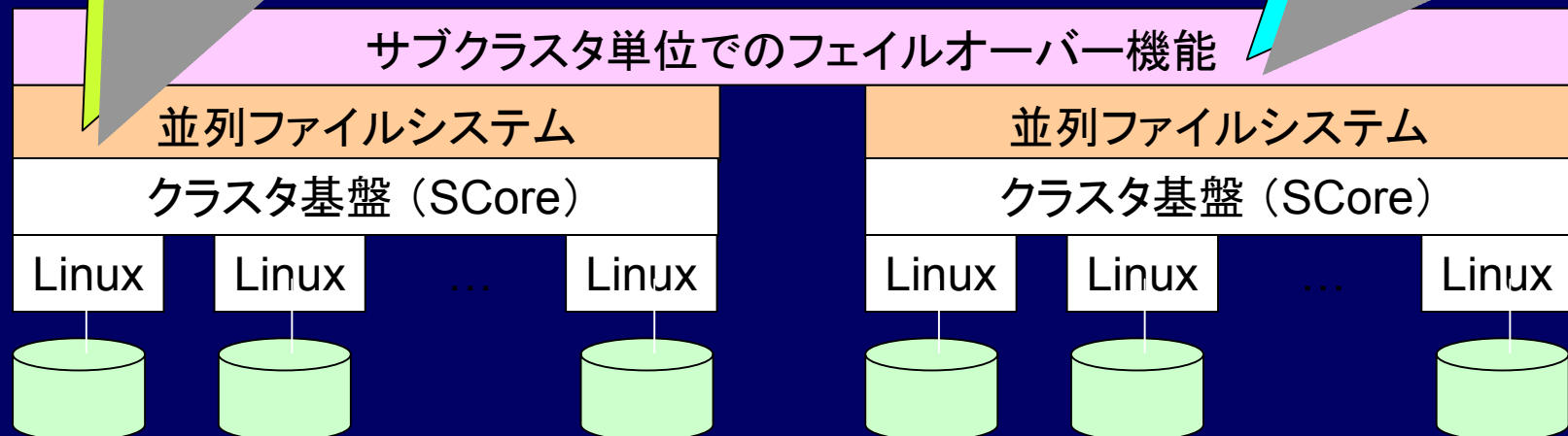
High Availability >> 共有メモリプログラミングモデル > SSI

並列ファイルシステム

- ・よりパワフルかつよりフリーな
並列ファイルシステムが必要
- ・並列ファイルを利用した
チェックポインティングの実装

サブクラスタ単位のフェイルオーバー

- ・サブクラスタ(計算タスクの実行単位)は
SCoreで構成
- ・サブクラスタ間で双方向待機型の
フェイルオーバー機能を実装



まとめ

- 並列PCクラスタは超並列アーキテクチャによる超高速処理性能を提供
- ソフトウェア(OS, システムソフト)はオープンソースがベース。利用環境の標準化、高性能化が進む
- 適用領域はポスト・ゲノム(タンパク質構造/機能などの解明)のみでなく、従来HPC領域の一部でも実用化が拡大そしてビジネス分野への展開も
- Grid Computing との融合
e-Science (これまでにない計算量、データ量を扱う)を支える
基盤技術へ発展

終了