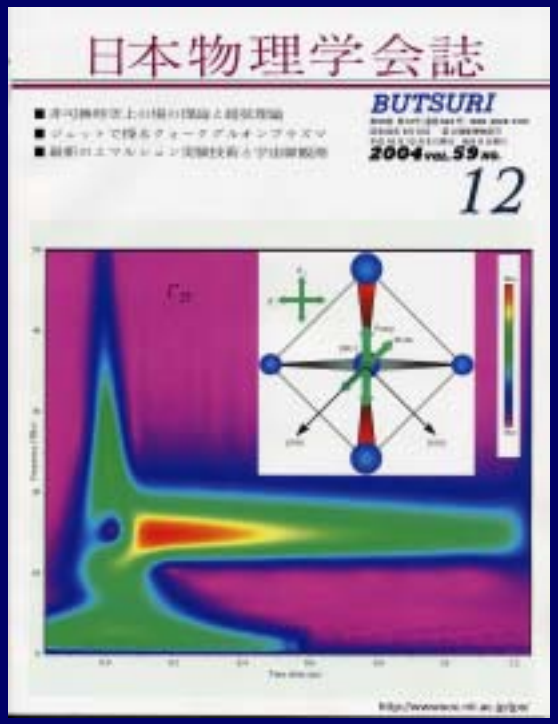
Beowulf (PC) Cluster Machine Equipped with Low-Latency Communication Software

M.Tanaka（NIFS, Japan）
http://dphysique.nifs.ac.jp/

Los Alamos Arxiv - physics 0407152 (2004)

手軽にできる研究室専用スーパーコンピュータ:
高速通信ソフトウェアを用いたPCクラスター計算機



2004年12月号
「話題」pp.898-902

参考文献

[1] G.Chiola and G.Ciaccio, http://www.disi.unige.it/
[2] M.Tanaka, Los Alamos Arxiv, physics /0407152 (2004)
[3] 田中基彦、物理学会誌、話題 (2004年12月)

高速計算に必要な3要素

● 単一プロセッサの高速化

1 高速な プロセッサ

2 プロセッサへの、高速なデータ供給

限界まで進歩

● 複数プロセッサの利用 による高速化

3 小さい応答待ち時間、大容量の プロセッサ間通信
高速なデータ供給

(4) 高速なデータの書出し、読み込み (Disk I/O)

高速演算、データ処理の試み

単一プロセッサでの高速計算 -- 高クロック、多重パイプライン化、には限界
複数プロセッサの並列利用による高速化は、自然な発展

- ベクトル処理 並列型

  Cray, Fujitsu, NEC 「スパコン」
  共有メモリによる密結合

  依存関係が少なく高いベクトル化が可能な、流体型の計算で超高速
  十分な予算のもと、ミッション目的の利用 （気象予報－時間が勝負）

- スカラー処理 並列型

  Thinking Machine, Beowulf machine 「クラスター計算機」
  分散メモリによる疎結合（＋SMP）、多重並列を指向

  予算に応じた規模が選べる、高いコストパフォーマンス
  分子動力学、データ処理 向き（スカラー演算が多い）

PCクラスター計算機 ー 研究室専用のスーパーコンピュータ

スカラー処理プロセッサの多数並列

○ システムを占有できる: キューでの待ち時間が無い（小さい）

○ 価格が安く 高性能 ¥15万/ PC

市販の汎用PCを使う --- Pentium 4, Opteron

NIC（ネットワークインターフェースカード）

[性能] / [価格] の比が、きわめて大きい

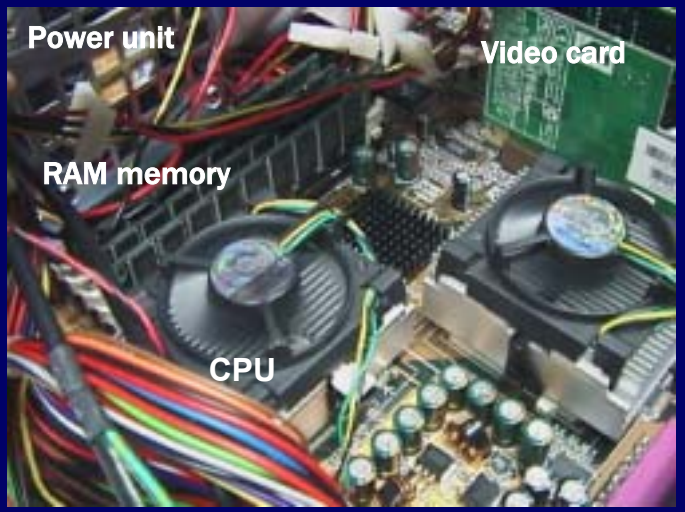
× 通信の遅さ

TCP/IP プロトコル -- 待ち時間（Latency）が非常に大きかった

Active Message /GAMMA（無料ソフトウェア）で解決できた

Beowulf PC Cluster: How to make and use it

PC units mounted in a 19" rack



Pentium III and M/B (chipset w/ bios)

1. Collect commercial PC parts ¥150-200k/unit
   - Mandatory: CPU, memory, M/B, LAN card, power unit
   - Optional: HDD, CD drive, Video card
   - Need 1 set: monitor, K/B, mouse, TCP/IP hub
2. Assemble screw drivers, glove
   - Fix M/B
   - Fix parts CPU → memory → cards →HDD
   - Connect cables signal, power, LAN, monitor
3. Install O/S and basic software
   - boot and install Linux + MPI (free)
   - Fortran f90、HPF (expensive !)
4. Set system environments
   - TCP/IP network
   - NFS mount: export the master's /home
   - NIS: enable the master node functionality
   - MPI, PBS
5. Programming
   - F90 + MPI for distributed memory
   - or HPF

プロセッサ間通信の高速化の方法

「PCクラスターの価格 に見合う範囲で」

1. ハードウェアで:

Grape ー 超高速だが中心力の計算に限られる
高価 ＞¥150-200万

Myrinet ーPC自身よりも高価、性能は良い
高価 ¥150万 + ¥20万x 台数

2. ソフトウェアで:

手持ちの Ethernet /NIC を活用する
投資費用がわずか！
増設 NIC（¥1-2万/ PC） + スイッチングハブ（¥2万）

通信の待ち時間（Latency）が小さい: GAMMA ＞ SCore
高い転送能力: SCore ＞= GAMMA

Active Message Communications Interface (UC Berkeley)

Direct application-network interface interactions, which bypass the operating system.

TCP/IP Layer	TCP/IP protocol	Active Message
Application	Telnet NFS	MPI / PVM
Transport	TCP UDP	
Network	IP	RPC / Sockets
Data Link	Ethernet (IEEE 802.3)	
Physical Layer	100/1000 Base-T, Myrinet, ...	
Latency	80μs	~10μs

ネットワークを　データ用と　管理用に分離する

(PCでなく) NICに、ipアドレスとホスト名が割り当てられる

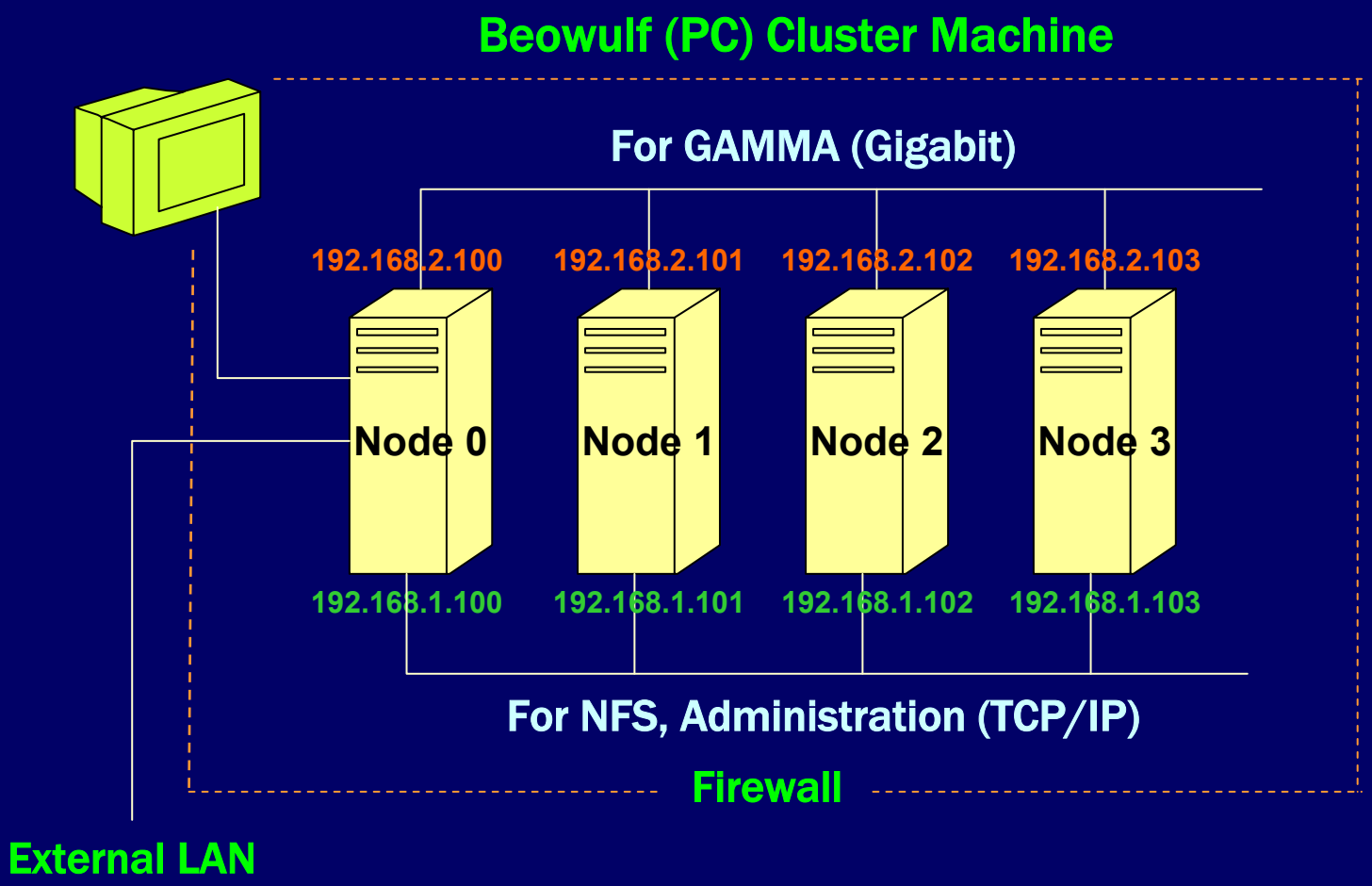


Fig.1 M.Tanaka

GAMMA (Genoa Active Message Machine)

Active Message Communications Interface (UC Berkeley):

OSを経由せず、物理層ーアプリケーション層間で直接に高速通信

制約

- OSはLinux（種類の指定はなし）、カーネルは 2.4.21　または 2.6.x
  gcc/ assembler のため

- ○ Gigabit NIC: bcm5700 ドライバーを利用したもの
  3Com996、Netgear GA620/621　＋現行機種の追加
  Fast Ethernet (100Mbps): 3Com905, Intel Pro e100

手順

0. ローカルネットワーク の2重化
1. Linux カーネル のアップグレード
2. GAMMA のインストールと設定
3. MPI/GAMMA のインストール
4. 高速C/Fortran の利用準備

- 最新の技術情報は、随時 http://dphysique.nifs.ac.jp/ に掲載

高速C/Fortranコンパイラの利用

- GAMMA通信のライブラリと応用プログラムの間で、整合性をとる
- 数値計算ライブラリを再コンパイル

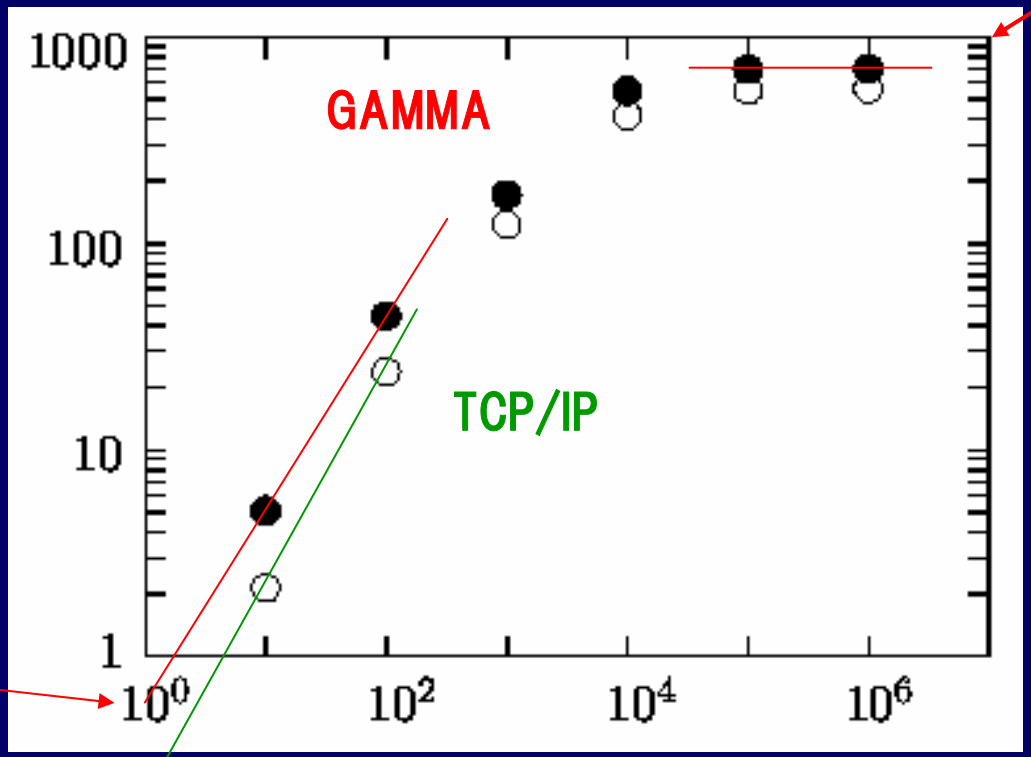
(i)　GNU C/Fortranにあわせ、一般のコンパイラでも生成されるオブジェクトに
　2個のアンダースコア(__)を添付、
(ii)　一般のコンパイラとGAMMAの2系統のインクルードファイルをこの順序で引用、
(iii)　線形計算ライブラリBLAS、LAPACK、その並列計算への拡張版BLACS、
　SCALAPACK を、条件(i)でコンパイル、
(iv)　引数参照を仲介するfarg.fをリンク、
(v)　論理定数の定義の整合性を確認する、である。

とくに、pgcc /pgf90コンパイラを利用する場合、標準的なコンパイル・スクリプトは、

```
pgf90  –o  ax1.out  –Msecond_underscore  –Mvect（プログラムの名前） ¥
-I/usr/local/pgi/linux86/…/include  ¥
-I/usr/local/mpich/build/LINUX/gamma/include  ¥
farg.o  -L/usr/lib/libgamma.a  ¥
-L/usr/ local/mpich/build/LINUX/gamma/lib  –lfmpich  –lmpich  ¥
/usr/local/BLAS/libblas.a  /usr/local/LAPACK/liblapack.a
```

である。最後の2つのライブラリは、(i)の条件下でユーザー自身がコンパイルしたもの。

Point-to-Point Communication Speed of GAMMA Communications



1 Gbits/s

710 Mbits/s

Latency 15μs
3com996

Pentium 4 (3GHz): 3Com996 NIC /GAMMA

GAMMAのホームページ からの抜粋

- ハードウェア
  Myrinet（1.28Gbits/s）+ BIP通信
  待ち時間 4.3μs　　極限帯域幅 1005Mbits/s

- ソフトウェア
  GAMMA＋Netgear GA621 NIC
  待ち時間 8.5μs　　極限帯域幅 976Mbits/s

多元の巨大行列を並列計算で解く:
PC間通信が頻繁
→ 通信待ち時間の短さが高速演算の鍵

PCクラスターで効率的なプログラム

● 領域分割が容易

複数プロセッサの利用で、高速化がはかれる

● リスト演算（Table-Lookup）やスカラー演算が多い

（スパコンで）ベクトル演算が活用されない

○ Ab initio分子動力学

○ 巨大 線形方程式の解法

例）ポアソン方程式

第一原理（Ab initio）分子動力学法

- 量子的な多体系を扱うための理論手法

  Schroedinger方程式は、少数自由度系（N < 6）のみ解ける

- 電子の記述

  - 断熱近似： 電子と原子核の運動を分離する
  - 1電子近似： 複数の電子を1電子の集まり として記述

    相互作用ポテンシャルは？　　交換相関相互作用 V_{ex} ？

  - 電子は基底状態： 作用 S 最小化して、方程式を導く

（ある時刻の）与えられた原子核の座標群に対して、電子分布を
量子力学的に求める（SCF: Self-consistent field）

Schroedinger 方程式

波動関数 $\psi_n(x,t)$

$$H\psi(x,t) = i\hbar\,\partial\psi(x,t)/\partial t$$

→

Kohn-Sham 方程式

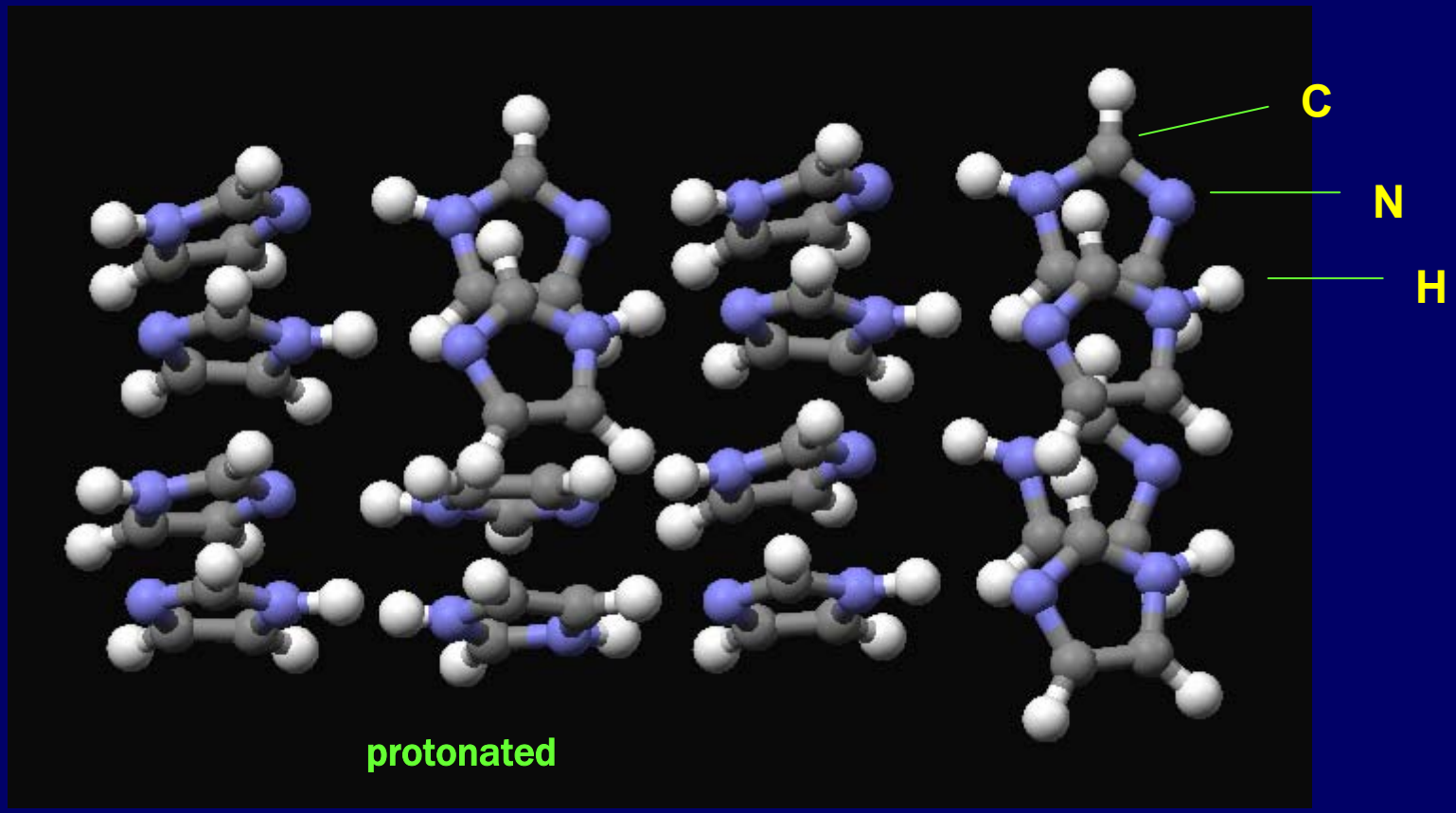
電子密度

$n_e(x,t) = |\psi(x,t)|^2$

Car- Parinello 法

- 原子核の運動は、古典的に解く

第一原理分子動力学： PCクラスター計算機が高い能力を発揮



Ion liquid imidazole which consists of organic molecules with high electrical conductivity

Fig.3 M.Tanaka

Computation Abilities of Various Methods

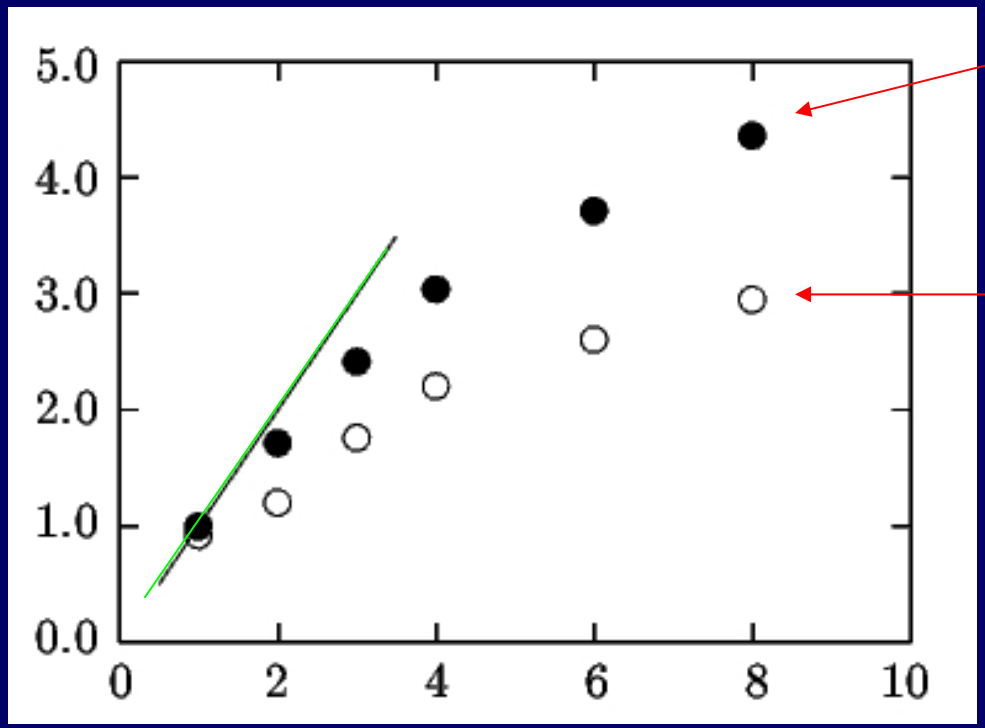
Ab initio Molecular Dynamics

	wallclock	cpu time	overhead time	ratio
TCP/ IP	93 sec	67 sec	26 sec	1.39
GAMMA	66 sec*1	66 sec	0.1 sec	1.00
RISC W/S*2	59 sec	59 sec	0.1 sec	1.00

Timing of first-principle molecular dynamics code Siesta, which makes many MPI communications for matrix diagonalization in every simulation time step. "Wallclock" is an elapsed time, and "overhead" is the difference of wallclock and cpu times. *1 is the cases with flow control. : [Pentium 4 (3GHz) + 3Com996] x 4 processors,*2 [IBM Power 4 (1.5GHz) + HP Switch] x 4 processors.

Computational Speed of PC Cluster against Allocated Number of Processors

First-principle molecular dynamics code Siesta: 181 atoms



Pentium 4 (3GHz) + 3Com996 NIC under TCP/IP and GAMMA

Comp. speed = $1/[\alpha + (1-\alpha)/N]$

α: non-parallel part

~ 0.10 for GAMMA

~ 0.23 for TCP/IP

研究のまとめ

- 並列化で、PCクラスター計算機を高速にできた

  OSを介さず高速通信を行う、ソフトウェアGAMMA の導入

  * 短い応答待ち時間 (Latency) を実現
  * 高いコストパフォーマンス　全コスト: ¥15万/cpu

  高速 C/Fortran コンパイラの併用が可能

  コンパイルで、GAMMA (gcc) と整合性をとる

- 並列プログラムでの効果

	性能比較
巨大線形方程式	$(2/3 - 1/2)\ \tau_{4*Pen4} = \tau_{4*RISC}$
第一原理分子動力学	$\tau_{4*Pen4} = \tau_{4*RISC}$

目的に応じて、最適な計算機を選択

- 緊急性のあるミッション目的　ー　気象予報、地震・津波解析
  流体型計算

  ベクトル並列型 計算機

- 計算負荷は高いが、非ベクトル演算が多い
  分子動力学、量子化学計算

  スカラー型 クラスター 計算機

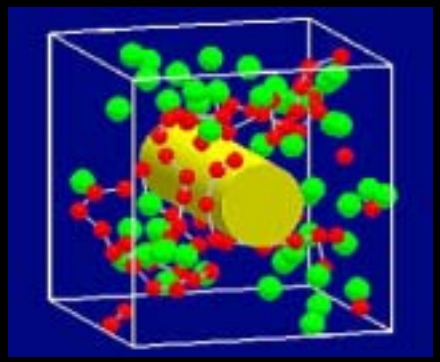
  資金に余裕？

  Yes → RISCマシン のクラスター

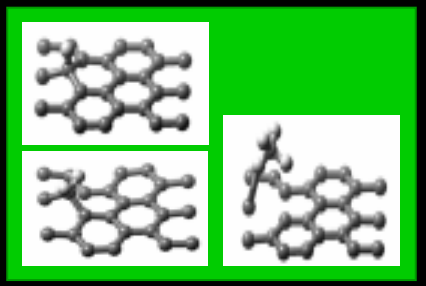
  No → PC のクラスター

Plasma and Ionic Condensed Matters by Molecular Dynamics Simulations

http://dphysique.nifs.ac.jp/



Charge inversion



Graphen destruction

Boewulf PC cluster

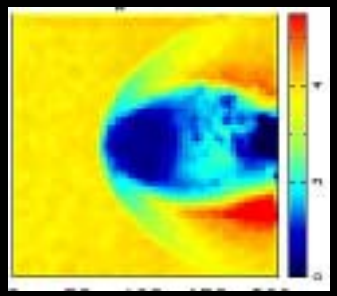
Publications
Cover Pictures

Ionic Soft Condensed Matters

First Principle (ab initio) Molecular Dynamics

Method and Tools of Molecular Dynamics

High Temperature Plasmas



Planetary shock

First proof of Collisionless Magnetic Reconnection
Development of Mesoscale Particle Code
Planetary Shocks

Scalapack on PGI & Red Hat Linux 7.3
Pentium 4 and its performance

1. Ionic soft condensed matters (Polymers, Charge inversion), 2. First principle molecular dynamics (Quantum mechanics), 3. High-temperature plasmas (Magnetic reconnection, Mesoscale particle code, Planetary shocks), 4. Method of molecular dynamics and Boewulf PC cluster, 5. Published papers and books (Cover pictures)
* Video movies of molecular dynamics simulations