PCクラスターの光と陰

理化学研究所：姫野龍太郎、重谷隆之、黒川原佳
東大IML：小野謙二

Outline

- 光
  - PC性能
  - PCクラスタの性能
  - グラフィックスへの応用
- 陰
  - PC性能
  - 製品寿命
  - 保守・管理

1. PC性能

- PC単体性能
- 何で測るかによって結果は異なる
- 同じ問題でも問題の大きさで性能が異なる
- PC性能を他のWSやスパコンと比べたい
- 複数台での並列計算でも測定したい

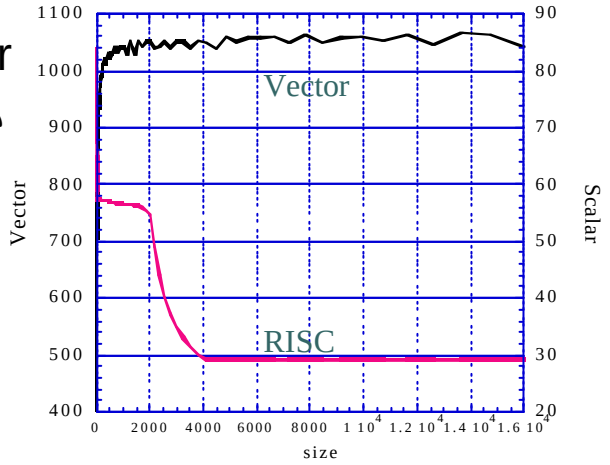
himenoBMT

PC用CPUの特性

RISCとVectorの特性の違い

cache_testBMT
http:w3cic.riken.go.jp

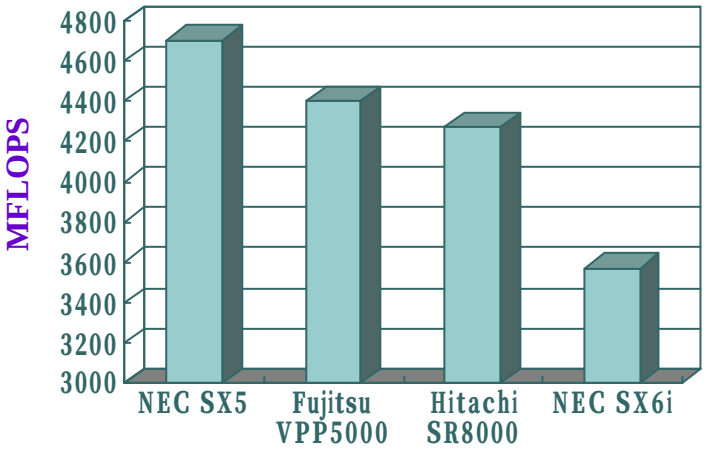
小さな計算は高速
大きな計算は遅い



himenoBMT

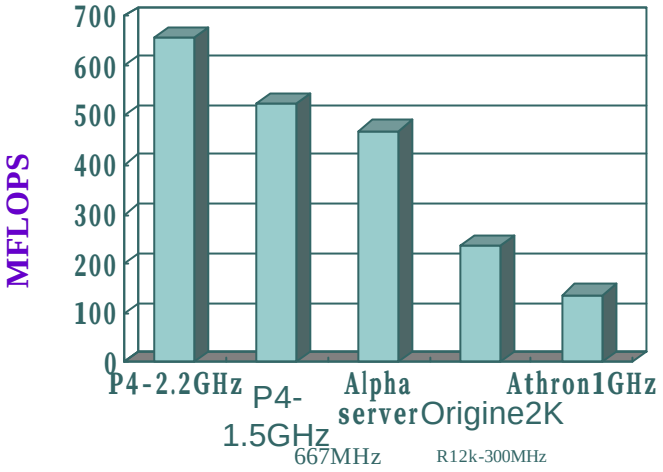
- 流体の数値解析コードのカーネル部分
- 全体の計算時間の５０%
- FORTRAN/C
- 実際の演算速度をMFLOPSで測定
- 領域分割による並列化をサポート
  - ３次元分割もサポート
  - 必要メモリも分割によって小さくなる
- http://w3cic.riken.go.jp

1CPUでのhimenoBMT結果(1)



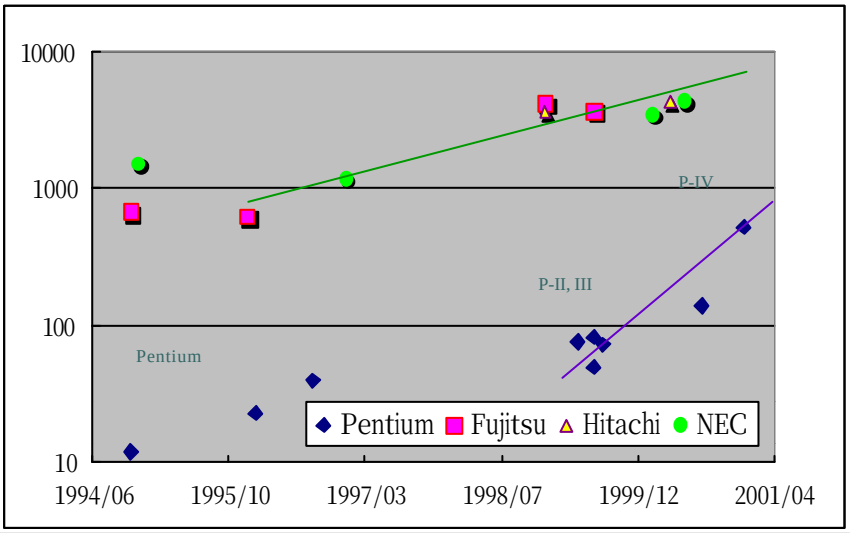
From http://w3cic.riken.go.jp/

1CPUでのhimenoBMT結果(2)



From http://w3cic.riken.go.jp/

性能向上のトレンド



比較すると

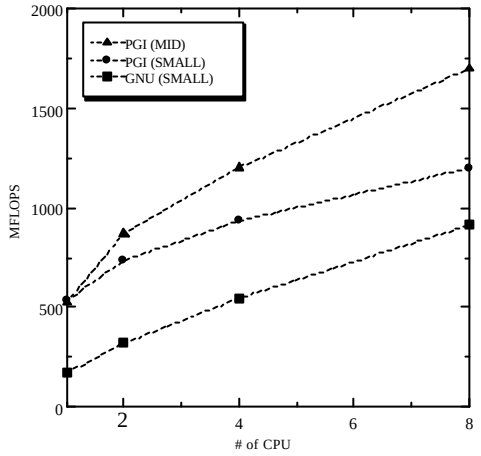
- PCの性能はWSの性能を越えた
- ベクトル計算機の性能は最新のP4の6-7倍しかない
- PCをクラスター化すると簡単にベクトルの１CPU並になりそう
- 価格性能比は１０倍以上

実際にやってみると・・・

- ２００1年　4月：Pentium 4 1.5GHz × ８台

- Hardware
  - CPU: Pentium 4 1.5GHz
  - Memory: 512MB
  - # of hosts : 8+1
  - Network : Fast Ethernet x 1 (100Mbps)
- Software
  - OS : Redhat 7.1 (Linux 2.4.0)
  - Compiler : GNU, PGI, Fujitsu

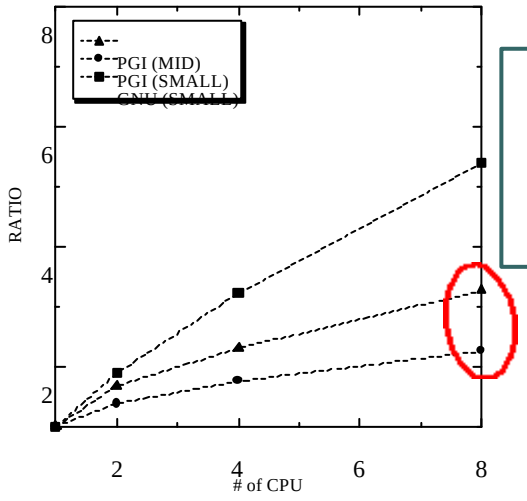
Pentium4/8の性能 (1)



- 8台で 1.7GFLOPS (PGI, SSE)

2-3倍くらいしか、出ない

Pentium4/8の性能 (2)



計算性能 ≫ ネットワーク

高速なネットワーク (Myrinet) が必要

P4クラスタでは

- 演算速度が速いので、小さな構成でも通信がネックに
- Myrinetなどの高速インターフェースが必要
- 発熱が大きく、コンパクトな筐体がない

大規模PCクラスタになったら・・・

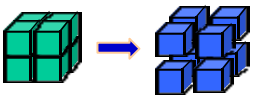
- 計算機センターでの運用は可能か？
  ⇒多数の利用者・ジョブ
- 本当に性能は出るのか？
  ⇒ネットワーク（Myrinet）、SCore

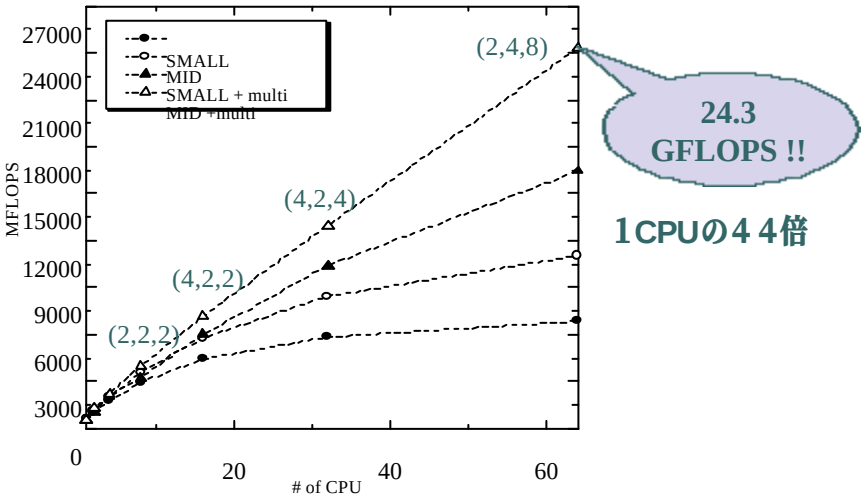
Pentium 4/64 クラスタ

RICE (RIken pc Cluster Enviroment)

- Hardware(ラックマウント 4U) **Fujitsu PRIMERGY CL460J**
  - ラック１０台（8台/1ラック）
  - CPU: Pentium 4 1.7GHz
  - Memory: 256MB, HDD: 40GBx65 + 70GBx1
  - # of hosts : 64+1
  - Network : Myrinet 2000, Fast Ethernet x 1 (100Mbps)
- Software
  - OS : Redhat 7.1 (Linux 2.4.3)
  - System Software : SCore 4.2
  - Compiler : GNU, PGI, Fujitsu
- Peak Performance：217.6 GFLOPS

３次元分割の結果





24.3 GFLOPS !!

1CPUの４４倍

実際には

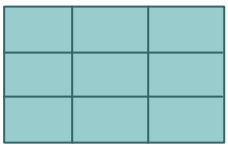
- 単純にMyrinetにしただけではだめ
- 性能を出すにはソフトを書き直す
  - １次元分割から３次元分割
  - 更に通信と演算を同時実行
  - 最終的に３３GFLOPS（60倍）
- でかい筐体（スペース効率の悪さ）
- 発熱がでかい（冷却用クーラー）

P3vsP4

- P3:２CPU/１U
- P4:１CPU/4U
- P4はスペース効率で言うと８倍悪い
- P3の８CPUなら１００Baseでok
- P4だと８CPUでもMyrinetなどの高速インターフェースが必要

Graphic PC Cluster

- PCのグラフィック性能はGWS並がそれ以上
- クラスター化するともっと高速な描画可能となる可能性
- 特に時間のかかるボリュームレンダリングや大スクリーンの分割表示、複数スクリーンで効果大（のはず）

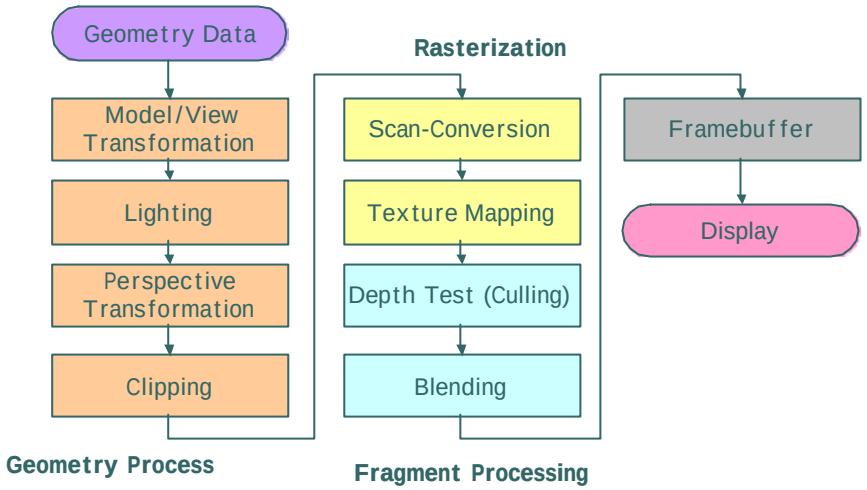


１台で表示するより９台で
表示する方が速い!!

World wide でのPC clusterの現状

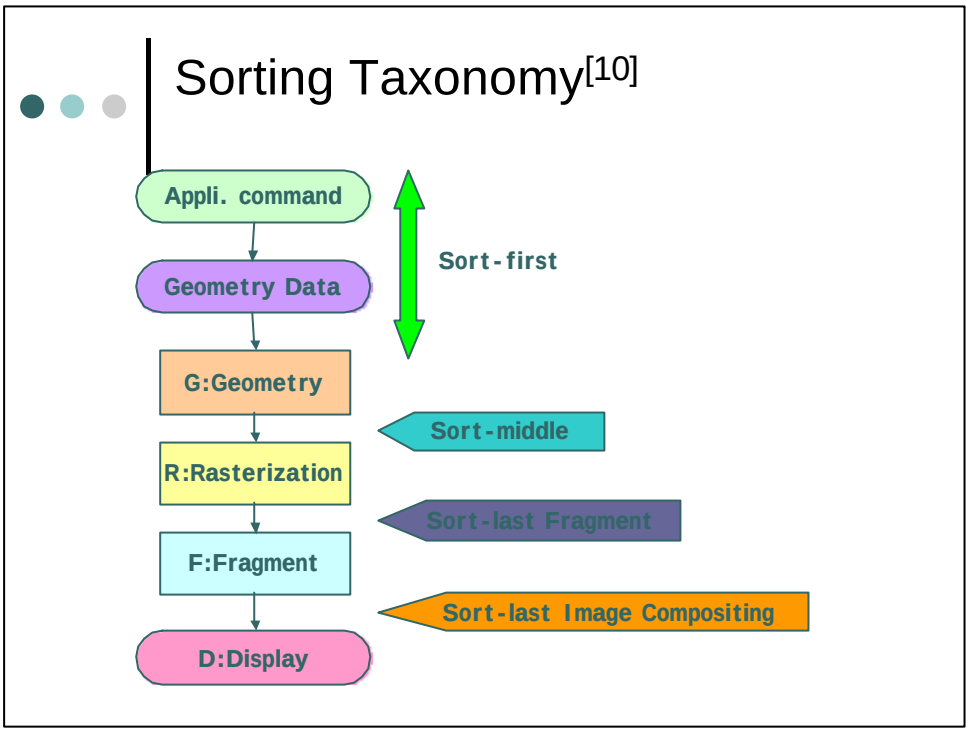
Site	Target	Cluster Features	node	Performance	Remarks
Univ. of Stuttgart[1]	Vol. Rendering	Dual P3@650MHz, 1GB, Myrinet, GeForce2GTS-64MB	96	1024^2pix., 256^3 vox., 4fps	Software blending
Stanford[2]	Tiled Display	Dual P3@800MHz, 256MB, Quadro2Pro, Myrinet, Lightning-2			WireGL, Chiromium
Caltech & Compaq[3]	Tiling, Compositing	Intel or alpha CPU, Volume Pro 500, Sepia-2	8	1024^2 pix., 512^3 vox., 24-28 fps	
Princeton[4]	Display wall				
Sandia NL [5]	Tiling	Dual P3@800MHz, GeForce2-32MB DDR, 512MB	64		
AIST[6]	Vol. Rendering	Dual P3@1GHz, 512MB, Vpro500, GeForce3, Myrinet, MergerH/W	9	256^2pix.,500^3 vox., 19fps, 128^3vox., 88fps	blending, renderingはH/W
LANL[7]	Vol. Rendering	Dual P3@800MHz, 1GB RDRAM, G-Ether, Wildcat4210 128MB	36		Software blending
Univ. of Utah[8]	Vol. Rendering Immersive Interactivity Unsteady data	Trex: Origin 2000, 16-pipe, IR-2, Direct I/O	128 SGI	5fps(10fps stereo), 1024^3 vox.	Software blending
		Trex2: Windows, wildcat4210	32 win	1.9fps	

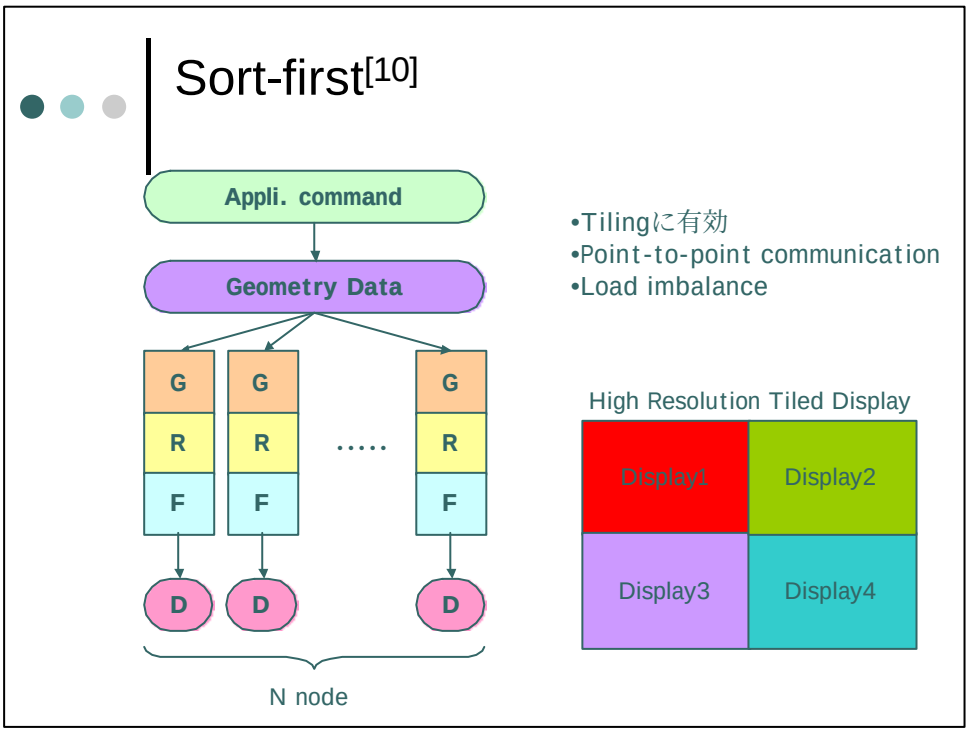
Rendering Process[12]



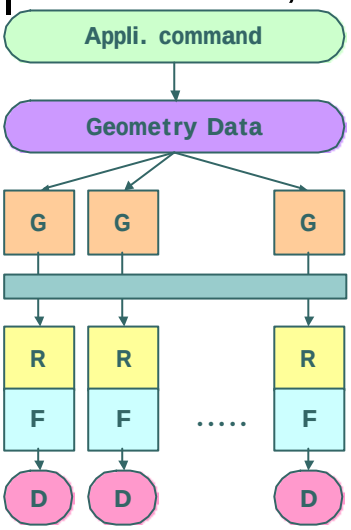
Source Parallelism[9]

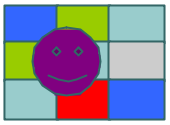
- Task
  - Functional parallelism (pipelining)
- Data
  - Primitive parallelism
    - Object space parallelism
    - Pixel space parallelism
- Sorting Taxonomy
  - Sort-first
  - Sort-middle (Interleaved, Tiled)
  - Sort-last (Fragment, Image compositing)
- Communication
  - Load Balance


Sorting Taxonomy[10]
Appli. command
Geometry Data
G:Geometry
R:Rasterization
F:Fragment
D:Display
Sort-first
Sort-middle
Sort-last Fragment
Sort-last Image Compositing

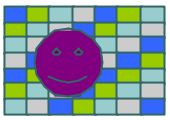

Sort-first[10]
Appli. command
Geometry Data
G
R
F
D
N node
•Tilingに有効
•Point-to-point communication
•Load imbalance
High Resolution Tiled Display
Display1
Display2
Display3
Display4

Sort-middle Interleaved, Tiled[10]



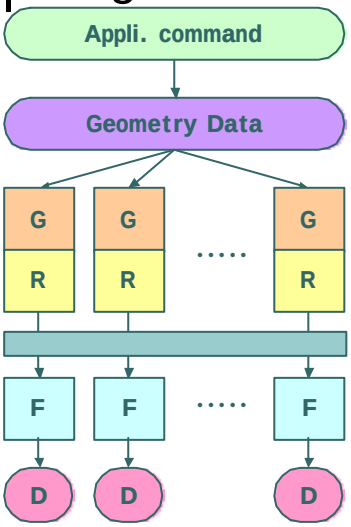


- Tiled
  - Load imbalance (RF)
  - Primitive overlap
  - Point-to-point comm.



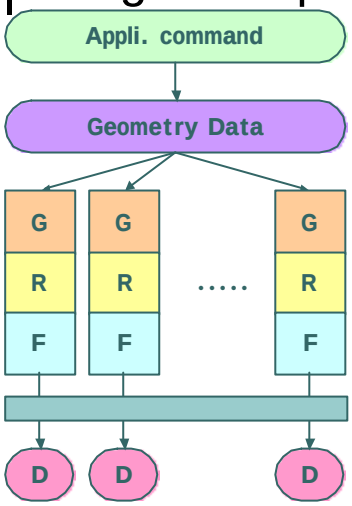
- Interleaved
  - Load balance
  - Broadcast comm.

Sort-last Fragment[10]



- Load imbalance
- Point-to-point comm.
- Fine grain tiling insures good load balance

Sort-last Image Compositing[10]



- Load imbalance may occur
- Requires more bandwidth
  XGA-RGBA, 30fps > 90MB/s
- H/W or S/W implementation

Graphic PC Clusterの現状

- まだ研究開発段階
- 本命は？
- 現状では性能を上げるためには何らかのハードウェア（同期、映像信号の合成など）が必要
- アプリケーションソフトは未対応
- AVSが近々発表予定のようす

PCクラスターの問題点

- ハードウェアの組み合わせで性能が出る場合と出ない場合がある
- 新しいハードを選ぶのはリスキー
- 性能の陳腐化が激しい
  - ３年の使用が限度？
  - ２年で補修部品が手に入らなくなる
- 運用の手間
  - 障害の切り分けが困難
  - 分かる人間が少ない
  - ジョブの運用管理や制限などの機能の機能不足（PCCCが改良してくれる？）

References

[1] M. Magallón, Parallel Volume Rendering Using PC Graphics Hardware, Pacific Graphics 2001.
[2] G. Humphreys, Chromium: An Open-source Cluster Rendering System, Course Note 37, SIGGRAPH 2001.
[3] S. Lombeyda, Scalable Interactive Volume Rendering Using Off-the-Shelf Components, ?
[4] A. Finkelstein, Tiled Display/ Early Experiences with an Inexpensive Scalable Display Wall System, Course Note 37, SIGGRAPH 2001.
[5] B. Wylie, DOE/ASCI-Lab Research Efforts, Course Note 37, SIGGRAPH 2001.
[6] S. Muraki, Next-generation Visual Supercomputing using PC Clusters with Volume Graphics hardware Devices, SC2001, Nov. 2001.
[7] A. McPherson, Los Alamos Cluster Visualization, Course Note 37, SIGGRAPH 2001.
[8] J. Kniss, et al., Interactive Texture-Based Volume Rendering for Large Data Sets, IEEE Computer Graphics and Applications, 21(4):52-61, 2001.
[9] S. Molnar, et al., A Sorting Classification of Parallel Rendering, IEEE Computer Graphics and Applications, 14(4):59-68, 1994.
[10] M. Eldridge, Parallel Graphics: Scalability and Communication, Course Note 37, SIGGRAPH 2001.
[11] R. Frank, Commodity-based Scalable Visualization: Graphics Cluster Components, Course Note 37, SIGGRAPH 2001.
[12] Wolfgang Heidrich, course note, http://www.cs.ubc.ca/~heidrich/