大規模PCクラスタ構築

新情報処理開発機構
石川 裕

新情報処理開発機構とは？

- 平成４年７月設立
- 通商産業省主導のもとに１０カ年計画で「リアルワールドコンピューティング（RWC）プロジェクト」を推進
- 体制
  - つくば研究センタ
  - 国内外の組合員２０社（機関）に存在する分散研究室

内容

- SCoreの簡単な紹介
- 1,024台プロセッサPCクラスタ SCore III
- PCクラスタ構築時の考慮点

SCore 型クラスタ

- 新情報処理開発機構で研究開発されたクラスタコンピューティングのためのシステムソフトウエアであるSCoreを利用したクラスタ


Aug. 1995


Feb. 1996


Oct. 1996


1997 - 1998


Oct. 1999


Oct. 2000


April. 2001

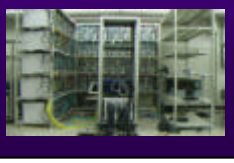
SCore

- University of Bonn, Germany
- University of Heidelberg, Germany
- University of Tuebingen, Germany
- Oxford University, England
- Warwick University, England









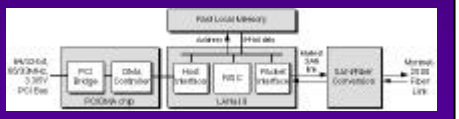
大規模PCクラスタ RWC SCore III

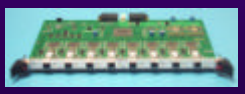
- Host
  - NEC Express Servers
    - Dual Pentium III 933 MHz
    - 512 Mbytes of Main Memory
- # of Hosts
  - 512 Hosts (1,024 Processors)
- Networks
  - Myrinet-2000
  - 2 Ethernet Links
- Linpack Result
  - 618.3 Gflops

This is the world fastest PC cluster at August of 2001

Myrinet-2000とは？

- 2 Gbps full duplex
- センダ・ルーティング
  - パケットの先頭にルーティング情報が格納されている
- NICの特徴
  - Lanaiプロセッサ
  - DMA Engines
    - HOST ←→NIC
    - Outgoing / Incoming Message
- スイッチの特徴
  - 16 port switch
  - warm hall routing





SCore Version 4 System Software

High Performance Communication Libs
- PMv2
  - 15.0 usec Round Trip time
  - 233 MB/s Bandwidth
- MPICH-SCore MPI Library
  - 24.4 usec Round Trip time
  - 228 MB/s Bandwidth
- PM/Ethernet Network Trunking
  - Utilizing more than one NIC

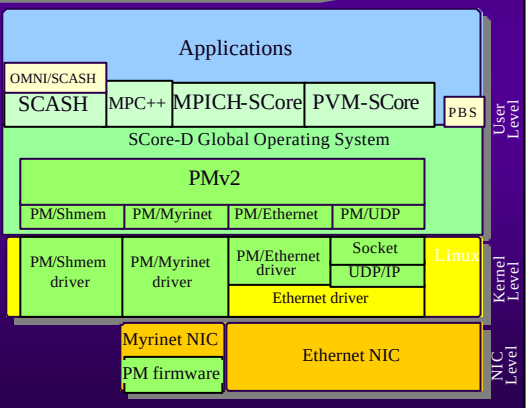
Global Operating System
- SCore-D
  - Single/Multi User Environment
  - Gang scheduling
  - Checkpoint and restart

Parallel Programming Language
- MPC++ Multi-Thread Template Library

Shared Memory Programming Support
- Omni OpenMP on SCASH

10 times faster than Fast Ethernet + TCP/IP

Three times as fast as Gigabit Ethernet + TCP/IP

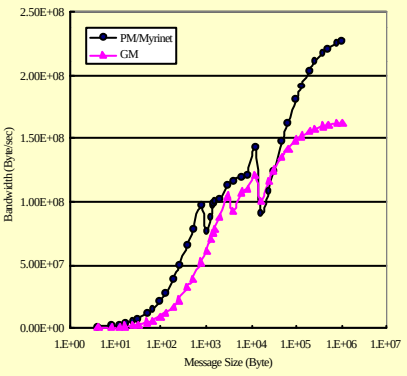


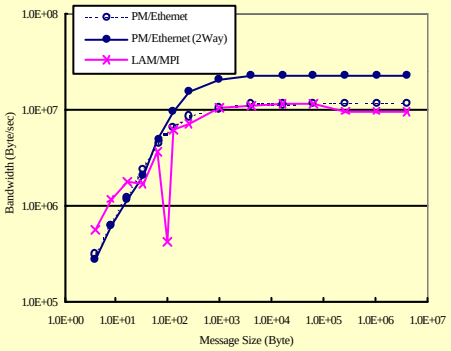
PM vs. GM

- PM
  - ４　DMAエンジンを同時利用
  - スタティックルーティング
- GM
  - 状態遷移に基づき、同時に複数のDMAエンジンに対して命令を発行していない
  - 自動ルーティング

MPI基本性能

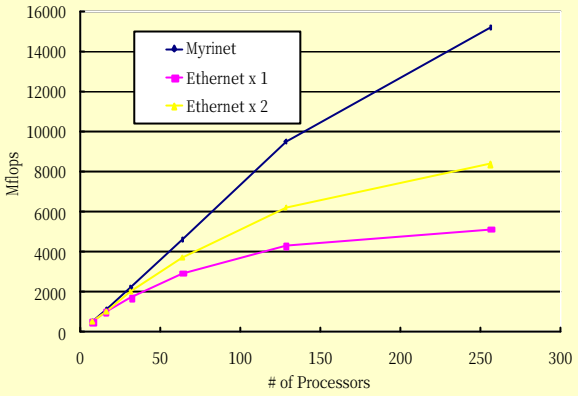
Point to Point MPI Communication Bandwidth





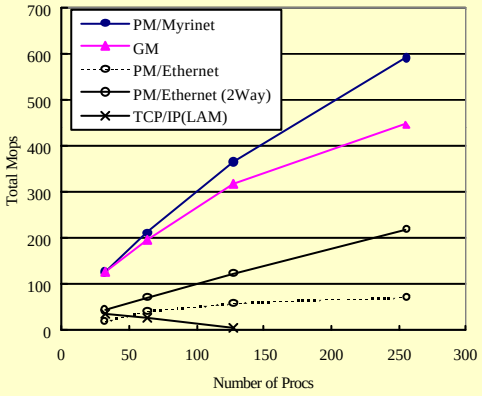
Application Benchmark

姫野ベンチマーク（512x256x256）



Application Benchmark

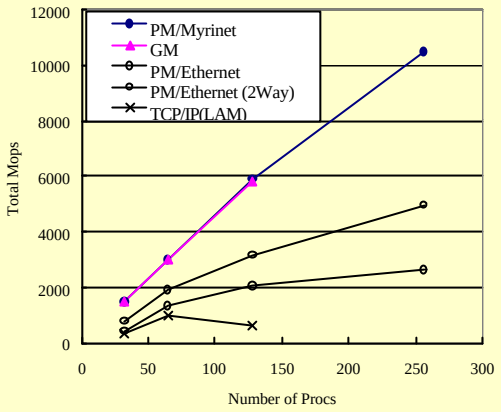
IS (Class C)　整数並べ替え



Application Benchmark

FT (Class C)

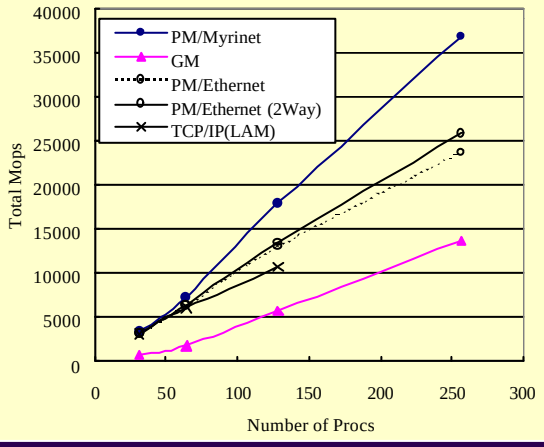
FFTを用いた３次元偏微分方程式の解法



Application Benchmark

LU (Class C)

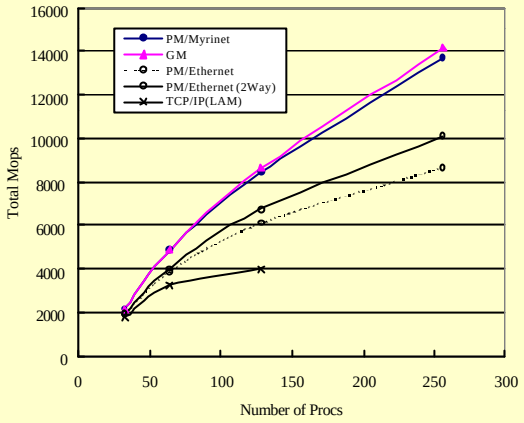
上下三角行列システムをSSOR(Symmetric Successive Over-Relaxation)法で解くCFD



Application Benchmark

MG (Class C)

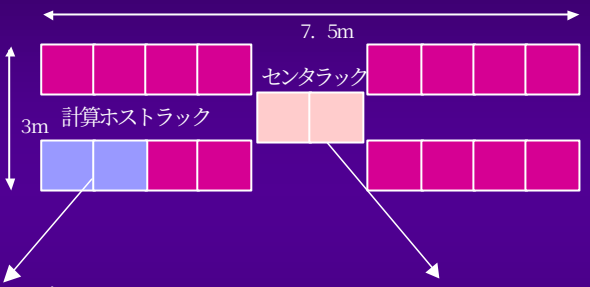
3次元ポアソン方程式を簡略化したマルチグリッド法のカーネル



大規模PCクラスタ

- ネットワーク構成
- ハードウエア設置
- ソフトウエアインストール
- ハードウエアテスト

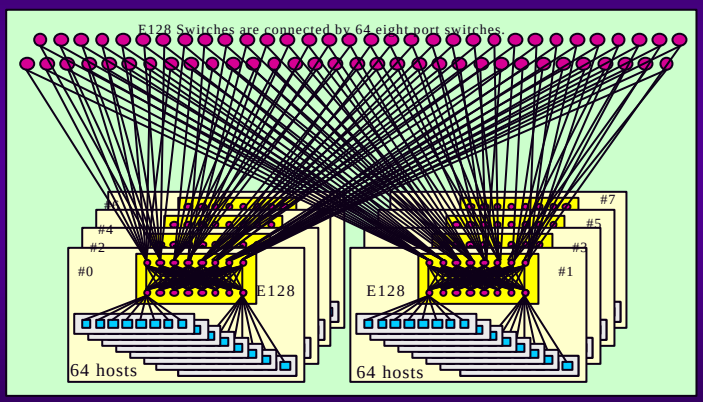
SCore III のレイアウト



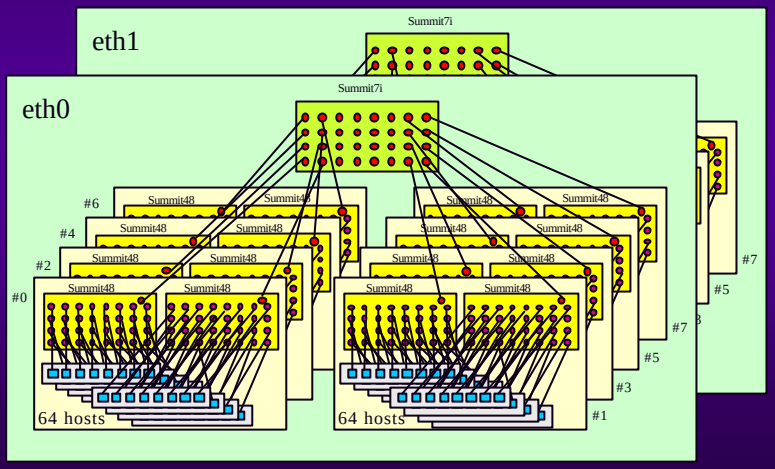
- 計算ホストモジュール
  - 2ラック
    - 64台PCサーバ
    - 2 Ethernetスイッチ
    - Myrinet Clos128

- センタラック:
  - 2ラック
    - サーバホスト x 4
    - Myrinetスイッチ、Gigabit Ethernet スイッチ

Myrinet-2000を使ったネットワークトポロジ



Ethernetを使ったネットワークトポロジ



計算ホストモジュール

モジュール内のスイッチ群

モジュール内の結線

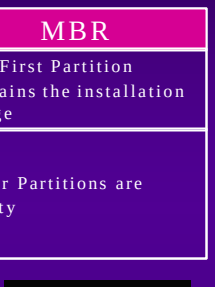
センタラック

ハードウエア設置

- 16ラックにEthernetケーブルを取り付ける作業
  - 4人の専門家で2日間
    - 全てのEthernetケーブルはカスタマイズ
    - ケーブルを切りコネクタ取り付け作業を現場で行う
- 4ラック毎にPCサーバ設置
  - 5人で４日間

ソフトウエアインストール(1/2)

- インストール用ディスクイメージの作成
  (800 Mbytes)
  - カーネル
  - RPMファイル
  - 改良版anaconda installation tool
  - Scoreインストール用スクリプト
  - ファイル生成に１０分から２５分かかる
- イメージファイルをディスクにコピーする
  - SCore IIを使用
    - 64 NEC Express Servers (2U type) with Myrinet
      - 5 SCSI disks
  - 32ホストにイメージをコピーするのに約2分
  - 128個のディスクにイメージをコピーするのに約6分
    - 1 minutes and half for one disk copy



ソフトウエアインストール(2/2)

- ディスクをクラスタに装着
- 電源投入
  - 1st stage
    - ディスクのパーティショニングおよびフォーマット
  - 2nd stage
    - ファイルシステム生成
    - IPアドレスの取得
    - Kickstartファイルを使ってanacondaの実行

テスト方法

- 配線テスト
  - Each rack, Each modules, and All racks
- Scoreが提供するrcstestによるall-to-all通信テスト
- キラーアプリケーションによるストレステスト
  - Stressing Myrinet network in terms of
    - network packets
    - memory and Lanai Processor of Myrinet NIC
  - Stressing processors and memory in hosts

Some initial hardware failures appear at the stress test !!

インストール時の問題

- Connection between Myrinet Card and PCI bus slot
  - Performance degradation
    - We have not found the reason
- Connection between Myrinet Line Card and back-plane
  - No communication
  - CRC errors
- Connection between Myrinet Card and Cable
  - No communication

PCクラスタ構築時の考慮点

- ハードウエア、ソフトウエアの知識のない人が実績のない新しいものを使うと後で痛い目に会うかもしれない
  - 沢山のハードウエアの組み合わせ
    - チップセット
    - Ethernetカード
    - Ethernetスイッチ
  - があるが、性能が出なかったり安定しない場合がある
    - Linuxドライバの問題
    - Ethernet NICの問題
    - Ethernet Switchの問題

PCクラスタ構築時の考慮点

- 8台や３２台の小規模でスケールアップしたからと言って、大規模クラスタを構築すると痛い目に会うかもしれない
  - ネットワーク性能の問題
  - アプリケーションの問題

PCクラスタ構築時の考慮点

- 空調、保守性を考えたデザイン
  - ケーブリング
    - ラベリング
    - 空気の流れを遮らない
  - ハードウエア
    - 容易にボード交換が出来る構造
    - リムーバブルディスク

PCクラスタ構築時の考慮点

- 現調の限界
  - NICの追加
  - メモリの追加
  - プロセッサの追加

初期不良の要因

大規模PCクラスタ設計時の考慮点

- 空調、保守性
- プロセッサvsメモリvsネットワーク＆トポロジ
  - 予算との相談
- 寿命
  - ２年レンタルあるいは３年レンタルで最後の一年はオーバラップできるのが理想では？
- 重量、電源

PCクラスタ市場の健全なる発展

- PCクラスタ構築に必要な技術的知識を持つ企業の育成
  - PCパーツ、ネットワーク機器
  - Linuxカーネル
  - SCore

PCクラスタコンソーシアム

www.pccluster.org

- メンバ
  - 19社, 2 研究機関, 7個人
- SCoreとOmni OpenMPのさらなる開発、サポート
- メンバ企業への技術移転
- PCクラスタに関する教育、市場育成

Real World Computing Partnership is over, but

SCore Development is continued