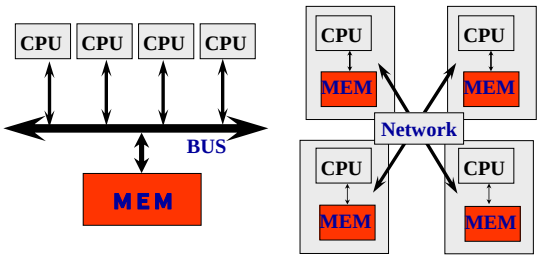
並列分散システム特論

計算機クラスタコンピューティングの技術動向

佐藤 三久
*1

並列コンピュータ（1）

- 共有メモリ型
- 分散メモリ型



並列コンピュータ（2）

- 共有メモリ型並列コンピュータ
  - プロセッサが同じメモリにアクセスできる。
  - プロセッサ数が少ない（１０から１００まで）
  - プログラミングが容易
- 分散メモリ型並列コンピュータ
  - プロセッサが別々にメモリを持っている
  - プロセッサ数が多くできる（１０００以上も可能）
  - メッセージ通信によるプログラム
  - プログラミングが難しい

高性能並列コンピュータの動向

- ベクトル並列コンピュータ
  - ベクトル処理ができるスーパーコンピュータを並列に結合したシステム　（数台から数十台）
  - VPP500,VPP700,C90,T90,SX-4,...
- 超並列コンピュータ
  - 多数のマイクロプロセッサを結合した分散メモリ型の並列コンピュータ　（数千台も可能）
  - SR2201、Cray T3E、Intel Paragon、...

高性能マイクロプロセッサの登場

並列処理の利点

- 計算能力が増える。
  - １つのＣＰＵよりも多数のＣＰＵ。
- メモリの読み出し能力（バンド幅）が増える。
  - それぞれのＣＰＵがここのメモリを読み出すことができる。
- ディスク等、入出力のバンド幅が増える。
  - それぞれのＣＰＵがディスクを読み出すことができる。
- キャッシュメモリが効果的に利用できる。
  - 単一のプロセッサではキャッシュに載らないデータでも、処理単位が小さくなることによって、キャッシュを効果的に使うことができる。
- 低コスト
  - マイクロプロセッサをつかえば。

ワークステーションクラスタ

- 通常のワークステーションを並列コンピュータとして使うシステム
- 高性能マイクロプロセッサを使ったワークステションが使える（数台から数十台）
- コストパフォーマンスがよい
  - 超並列コンピュータの down sizing!
- ネットワーク
  - イーサネット、ATM、FDDI
  - 高速なネットワークを使ったシステム　IBM SP-2
- 通信の少ないアプリケーションが適している（？）

クラスタコンピューティング

- クラスタシステム：既存のワークステーションやＰＣを（既存の）ネットワークで結合して，並列計算を行うシステム
- 第１世代：既存のＬＡＮで並列計算するシステム
  - Poorman's supercomputer
  - 遊休のワークステーションを利用
  - お茶大 Sun IPC cluster （計算センタのワークステーションを利用）
  - イーサネット
- 第２世代：クラスタ専用の計算機システム
  - etlwiz Alpha cluster, 東大富士川研クラスタ
  - 100 BASE-TX SWITCH，ATM
  - beowulf class クラスタとも呼ばれる．
- 第３世代：高速のネットワークによるクラスタ
  - 高並列計算機（ＭＰＰ）なみの性能
  - RWCP PC cluster
  - Myrinet，Gigabit Ethernet, Fiber Channel, DEC Memory Channel, IBM SP2 network
- その他
  - ＳＭＰクラスタ （UCB CLUMPS, RWC COMPaS)

Etlwiz: DEC alpha Cluster

SNOW: Wiz (1996 Sep. -) Scalable, Special Purpose, and Single user Network of Workstations at ETL

- Dedicatedなクラスタ環境
  - アプリケーション:virtual micro scope
- 高性能マイクロプロセッサ（綿第時）のワークステーションの利用
  - コンパクトなラックマウント
  - Digital Unix
- 100Base/TX 全結合可能なスイッチモジュールの採用
- メモリを無駄にしない構成
  - 8,16,32台のいづれにおいても 4GB使用可能
- 保守性の向上
  - スイッチひとつによる立ち上げと停止
  - 障害時，点検時に有効，UPSと連携

DEC Alpha Station 333MHz x 33

Cisco Catalyst 5000
Fast Ethernet switch
12x3 + 2 port
1.6Gbps Back Plane

RWC PC Cluster II

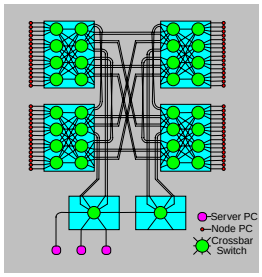
- 高性能なマルチユーザ並列プログラミング環境
  - 並列ソフトウェアのテストベッド
- 市販部品 + フリー OS による PC クラスタ
  - 産業用 PC（PICMG）
  - ギガビットネットワーク（Myrinet）
  - NetBSD・Linux
- 筐体のみ独自仕様
  - 実装密度
  - 保守性
- ＭＰＰに匹敵する性能を達成

Pentium Pro 200MHz x 64

モニタ PC ✕ 2
サーバ PC✕ 3

RWC PC Cluster: Myrinet ネットワーク

- ハードウエア
  - リンク： 160MB/s ✕2
  - スイッチ間： 4 リンク
  - ラック間： 16 リンク
  - Bisection バンド幅 5120MB/s
- PM通信ライブラリ
  - :メッセージ
    - 50〜119 MB/s (8KB)
    - 7.5 μs (8B)
  - リモートメモリライト
    - 79〜109 MB/s (1MB)
- MPICH-PM
  - 76.7〜104 MB/s (1MB)
  - 11 μs (8B)

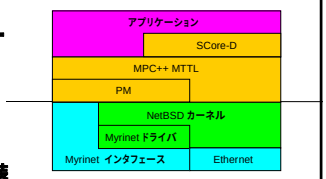


RWC PC Cluster II: PM 通信ライブラリ

- ユーザレベル･Myrinet ドライバ
- 非同期メッセージパッシング
- リモートメモリライトによるゼロコピー通信
  - ピンダウン･キャッシュ
- 複数の通信チャネル
- ネットワークコンテキストスイッチ
  - ギャングスケジューリングのサポート
- 通信負荷の分散

RWC PC Cluster: システムソフトウェア

- ローカルオペレーティングシステム
  - NetBSD/linux
- PM 通信ライブラリ
- 言語: MPC++(Level 0)
- マルチユーザ並列オペレーティングシステムSCore-D
  - ギャングスケジューリング
  - アイドル検出
  - Myrinet を介した I/O
  - ロードモニタ
  - デーモンプロセスとして実装
- MPICH-PM



並列分散システム特論

COMPaS: Cluster Of Multi Processor System

- 背景
  - PC、マイクロプロセッサの高性能化と普及、低価格化
  - 計算機クラスタによる高性能計算の可能性
- SMPクラスタ
  - SMPがクラスタのノードとして使われつつある
    - SMPの低価格化
    - コンパクト、管理が容易
    - ネットワークが少なくてすむ
  - SMPクラスタ利用技術の研究
    - プログラミング
    - 性能モデル

Node (Toshiba GS700)
Eight quad-processor
Pentium Pro 200MHz

Networks
Myricom Myrinet
100Base-T Ethernet

並列分散システム特論

クラスタコンピューティングを支える技術

- ハード，ソフトのコモディティ化，高性能化，標準化
- ハードウエア
  - プロセッサテクノロジロードマップの恩恵
    - 急激な高性能化，価格性能比の向上
  - ネットワークの高性能化
    - ethernet : 10Mbps から 100Mbps そしてGigabit either
    - MyrinetなどのSAN Network
  - 高性能 I/Oインタフェースの標準化
    - ＰＣＩなど
- ソフトウエア
  - 並列通信ライブラリの発展・標準化
    - PVM, P4, TCGMSG，MPI，MPI2
  - 標準ライブラリ上に並列ソフトウエアが開発できる．
  - フリーなオペレーティングシステムの普及

並列分散システム特論

クラスタとMPP

- もはや違いはないのだけれど，．．．
- MPPの利点：スペースコスト，実装密度，いろいろなサポート
- MPP専用のハードによる最適化
  - ＤＳＭ，再帰度通信，…..
- クラスタは開発期間が短い
  - 高性能な部品がすぐに使える
- コスト(price-performance)
  - MPPなどのハイエンドコンピュータは，500〜1000万円/GFLOPS
  - Pentium II ８台クラスタ 200万/GFLOPS (PC30万円*8＋etherswitch 150万円＋その他 50万＝440万円で300*8=2.4GFLOPS)
  - 自分の作業コストは高い ？
- クラスタの構築の問題点
  - 組み合わせたときの動作確認が必要（電気的特性の相性，デバイスドライバのサポート）
  - 組立作業が大変
  - 製品サイクルが短い（保守、拡張時に同じ部品が入手できない）
  - システム管理が面倒

並列分散システム特論

クラスタテクノロジの課題

- 並列ソフトウエアは，ＭＰＰとそれほど違いはない．ソフトは依然として問題．
- スケーラビリティは重要か？
- 通信インタフェースの高性能化
  - 既存のプロトコルでは通信性能を使いこなせない
    - Gigabit Ether 100MB/s以上だが，TCP/IP/Gigabitでは30MB/s
  - リモートメモリ通信
    - たとえば，VIA
  - 標準化，コモディティ化は依然として重要
- 並列プログラムのジョブ管理
  - マルチユーザ並列プログラム環境
    - たとえば，Score-D/gang scheduling(RWCP), implicit co-scheduler (UCB)
- 再び，LAN環境へ
  - 光インタコネクトを使った分散並列環境
  - ＳＭＰクラスタ，ヘテロ環境，windows-NT，software DSM

並列分散システム特論

動向

- 大規模システム
  - BigMac (Mac G5)を１０００ノード（Top500 で3位), InfiniBand
  - AIST スーパークラスタ Opteron 1000ノード, Myrinet
  - RIKENクラスタ P4? 1000ノード、InfiniBand
- プロセッサ
  - P4、Xeon 3.xx GHz
  - Itanium 2
  - Opteron , EM64T (Xeon)
  - Multi-Core Petinum-D …
- メモリ
  - FSBの高速化 (400MHz -> 800MHz, 1066Mhz@P4 XE 3.64GHz)
  - DDR2（Double Data Rate Synchronous Dynamic RAM）
    - DDR４００Mhz（３．２GB/ｓ） ⇒ ６６７Mhz（４．２GB/ｓ@５３３Hz)
    - 低電圧、低消費電力
  - Rambus (600MHz-700MHz, 600MB/s-700MB/s)

並列分散システム特論

動向

- 内部IOバス
  - PCI (33MHz, 32bit, 133MB/s ⇒ 66MHz, 64bit, 533MB/s)
  - PCI−X（１３３MHｚ、６４bit、１.06GB/ｓ、PCIの倍）
  - PCI-Express (シリアルリンク、2.5Gbps/レーン, 250MB/s x 2 両方向)
    - 16レーンで、4GB/s x 2 (双方向)
  - HyperTransport （AMD他)
    - 6.4GB/s
    - メモリ接続バスにも使える
- ネットワーク
  - Mryinet XP (2+2Gbps, 250+250MB/s)
  - InfiniBand （2.5Gbps/channel, 16channelで60Gbps=8GB/s)
  - Quadrics QsNet II (920MB/s)
  - トランキング・マルチリンク技術
  - Gbit Ethernet
  - 10Gbit Ethernet