PC クラスタWG報告

姫野龍太郎　理化学研究所

Outline

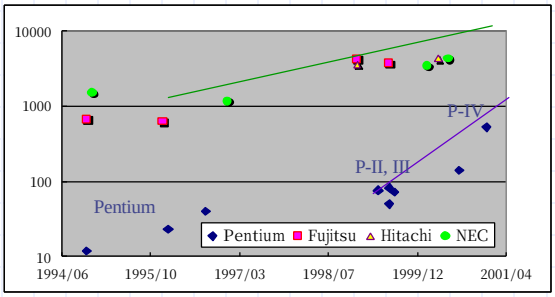
- PCクラスターとは
- 背景
- 目的
- メンバー
- WGの検討内容
- 活動
- 今後の予定

PCクラスターとは

- 複数台のPC＋NICネットワーク＝分散メモリー型並列計算機
- 市販PCをそのまま集めて構成
- ラックマウント型PCで構成

背景

- 安い・速い
- 科学技術計算分科会:PCクラスタは Poorman's Supercomputeer
- 合同分科会でのLinux討論会



問題点

- 一から十まで、自己責任
  - 性能が出ない組み合わせ
  - 新しいのは避ける
- 自分の作業量を含めたトータルコストは大
- 大規模な構成では小規模では気づかない問題も

2001年11月1日 SS研合同分科会 5

WG設置の目的

- PCクラスタの設計や保守に関する要望をまとめ、実現を図る
- 中小システムは効果が明らかなので、普及を図る
  - PR活動
  - システム設計や運用の手引書作成
  - アプリケーションの移植の要望や支援

2001年11月1日 SS研合同分科会 6

WGメンバー

- [会員]○姫野 (理研)、真鍋(KEK)、庄司(原研) 、石井 広島大学)
- [富士通]○ 石田、大空、久門、佐藤、市川

WGで話題として取り上げるもの

(1)大規模なシステムにおける問題点の抽出
(2)今後のクラスタ用ハード/ソフトの動向調査
(3)規模による運用特性の違い
(4) PCクラスターのリスク

(1)大規模なシステムにおける問題点

- ジョブのモニタリング、ロードレベライザーなどの管理ソフトが十分でない
- シリアルコンソールかモニター切り替え機か、
- リセットの手間
- 資産管理上の問題
- ベンダーのサポートが受けられない
- データ管理(NFSは数百台規模では動かなくなるなど)

(2)クラスタ用ハード/ソフトの動向調査

- サーバー用CPUの動向
- Myrinet、GbitEthernetその他、高速ネットワークの動向
- Linux/SCORE/MPI-CH(RWCPの今後なども含む)
- コンパイラー性能
- メモリーやチップセットの動向

(3)規模による運用特性の違い

- 基本的にマルチ・ユーザーには向かないのか
- ユーザーに占有時間を与える運用方法の検討
- 規模による適正化は可能か
- 小規模のシステム導入のポイントの整理
- 中規模のシステム導入の注意点

(4)PCクラスターのリスク

- ハードウエアのトラブル
- ソフトウエアのトラブル
- トラブルにあわないための仕様書の作り方、ベンチマークの仕方

活動

(1)調査 ・ディスカッション
(2)講演会の企画
(3)WG会合

(1)調査 ・ディスカッション

- 外部(新情報開発機構などの機関や、先端ユーザー、LINUX協会、インテルなど)に協力を求め、レクチャーやディスカッションを行ってゆく。
- この予定は事前に公開し、興味を持つ会員であればWGに限らず、参加できるものとする。

(2)講演会の企画

- 広くPCクラスターを普及し、さらに関係者への影響力を増すために、SS研会員に限らない、広い層を対象としたPCクラスタに関する講演会を企画・運営する。
  - 富士通本社
  - 200人規模の講演会
  - 2002年5月9日(SS研総会の前日)

2001年11月1日 SS研合同分科会 15

(3)WG会合

- これまでに2回開催
- #1は自己事例紹介と活動内容の検討
- #2から実質的な活動
  - 富士通研究所・久門氏“PCクラスタ～その性能と今後～”
  - RWCP・住吉氏“RWCPでのPCクラスタへの取り組みと構築上のポイントについて”

2001年11月1日 SS研合同分科会 16

久門氏の講演要約

- 性能の向上が著しいが、スペック通りの性能を出すには、慎重な選択が必要
- 知識が性能を決める
- 差は増大傾向
- CPUはP4が当面最適
- インターコネクトの比較
  - 専用：Myrinrt, cLAN, Infiniband
  - コモディティ：性能が確認されたNICを選択する事が必須

住吉氏の講演要約(1)

- PCクラスタを確実に動かすためには、ハードウエアの適切な選択が大事
- 空調、電力、ケーブリングなど細かく気を使う必要。これがあとの稼働に大きく響く。
- 自分でどこまでやるのか/やれるのかをよく考えて、仕様書を書こう
- 導入当初の試験が初期不良出しや、所定の性能を出すために重要

住吉氏の講演QA#1

Qケーブル配線が多数あるが、ラックの場所を移動させる場合はどうするのか?

Aそのラックに接続されているケーブルを全部はずして移動するしかない。

Q配線をすっきりということだが、一本ずつケーブル長が違うのか?

A1本ずつ違う。ただし、PCを搭載する構成が決まれば何mが何本という形で業者に作ってもらうことができる。

Q電源の線は各PCから全部だしているのか、UPSはどうしているのか?

A各PCから出している。UPSは数が多いため設置していない。

2001年11月1日 SS研合同分科会 19

住吉氏の講演QA#2

Q.セキュリテイ対策はどうしているのか?

A.PCクラスタをプライベートネットワークに接続し、対外接続はファイアウォール経由とする、各ノードへのtelnetを禁止する、といった対策が考えられる。

Q.大きな計算をする場合各ノードに結果ファイルが分散することになるが、それらの管理はどうすればよいのか?

A.計算後にバラバラのデータを1ヶ所へ集めて処理する方法、クラスタファイルやNFSのようなファイルを利用する方法、計算ノードで最終処理(可視化など)までする方法、などが考えられる。（WGで今後継続して、調査する）

2001年11月1日 SS研合同分科会 20

住吉氏の講演QA#3

Q.並列コンパイラはMPIを不要にするものではないのか?
A.理想はそうだが、現実はMPIで開発されている方が多い。
C.富士通はVPPのコンパイラでそれを実現しているのだから、そういう技術をPCクラスタでも実現してほしい。
Q.1000台規模のクラスタの利点と欠点は?
A1.利点はトラブル切り分けの時、半分のシステムに分割して同一プログラムを流すことにより問題個所を切り分けられる。
A2.欠点は数が多いのでハードトラブルが色々出る
C.トラブルが出たときベンダーの対応力がわかる。

2001年11月1日 SS研合同分科会 21

住吉氏の講演QA#4

Q.トラブル発生率の高い順にいくつかの問題を教えてほしい
A.ケーブルの接続不良が多い、その他は熱問題によるCPUのトラブル、メモリエラー、DISK障害、ネットワークスイッチ障害、という感じ。
Q.某ベンダーのPCはファンがよく壊れたがそういうことはないのか?
A.RWCPではファンが壊れたということはない。
Q.途中でノードを追加したとき、SCoreはそこだけ最新バージョンにできるのか?
A.全ノードでバージョンを合わせる必要がある。

2001年11月1日 SS研合同分科会 22

Know-Howまとめ#1

- 構築上のポイント
  - アプリケーションの性質の把握　：実際に測定
  - ネットワークの選択　:実際に測定
  - ラックレイアウト：移動を考慮したレイアウト
  - 試験　:一晩くらいの全ノード通信試験が必須、NASベンチマークで値の検証する

2001年11月1日　SS研合同分科会　23

Know-Howまとめ#2

- 購入時のポイント
  - 購入業者にどこまで頼むかを明確に、技術力があり実績を持つ業者選定が重要。
- 運用面
  - 1000台規模の故障率はベンダーのエラー率程度
  - トラブル切りわけは半分ずつ切りわけて確認

2001年11月1日　SS研合同分科会　24

WGでの議論#1

- ファイル　:PCクラスタを構成するCPU廻りは安価にできても、ファイル廻りにお金がかかる。
- プリ/ポスト処理との連携　:PCクラスタで数値シミュレーションさせた時のプリ/ポスト処理システムとの連携で使いやすい接続、大量のシミュレーション結果データの取り扱い
- 計算センターとしての大型PCクラスタの運用　:運用管理機能(特に課金)、エンドユーザのデータの扱い

2001年11月1日　SS研合同分科会　25

WGでの議論#2

- ハードスペック向上速度にどう対応するか?　:価格低下、性能向上が激しく、仕様書を作って調達をかけても導入時にはものがない可能性がある。また、増設時には同じものが手に入らない。テスト環境が必要。
- 計算だけでなく、計算結果出力の画像処理などの処理が必要。これらを効率的におこなうにはどうしたらよいか。
- 同様に ファイルI/Oは全体のスループットとして重要であるが、これを、並列化された計算性能に見合うだけ実現しなくてはならない。どうしたらよいか。
- その他・・・・・

2001年11月1日　SS研合同分科会　26

まとめ

- 現在、熱い議論の最中
- 今後、クラスタによるグラフィックや ファイルについて調査
- さらにガイドラインなどを整備し、公開予定
- 来年5月に講演会を開催

2001年11月1日 SS研合同分科会 27