SCore7 の紹介

PC Cluster Consortium 開発部会
堀　敦史

SCore 7 に向けて

反省と現状認識

SCore 6 の反省

- やり過ぎだった？？？
  - 速さを追求し過ぎた
    - カーネルパッチ (SCore5)、ドライバパッチ (SCore6)
    - 脆弱なギャングスケジューリング
  - 排他的だった
    - Myrinet: SCore と GM/MX は共存できなかった
  - 抱え込み過ぎ
    - Myrinet firmware など
  - インストールが面倒

やり過ぎの原因

- 性能至上主義
  - HPC クラスタなんから速くなければならない
    - 独自プロトコル (Myrinet, Ethernet)
  - ギャングスケジューリングも高速でなければ！
    - 研究目的
    - 問題発生時の影響が大きい
- 他のシステムとの共存はあまり考慮しなかった
  - 自前のバッチスケジューラ

SCore の良いところ

- マルチネットワークのサポート
  - 同じバイナリで様々なネットワークで動く
- ギャングスケジューリング
  - SCore のユニークな機能のひとつ
- All-in-one パッケージ
  - 複数パッケージはバージョン管理が大変
- ネットワークファイルシステムに依存しない
  - メニイコアで 100 プロセスは当たり前
  - NFS がボトルネックになるのは自明

SCore の実際の使われ方

- SCore は one of them
  - MPI_SELECTOR
    いくつかの MPI の実装からユーザが選ぶ
  - 排他的な実装 (SCore6以前) は嫌われる
- バッチスケジューラの下で動く
  - SGE、PBS、NQS、…
  - バッチスケジューラとタイトに連携できていない

SCore 7

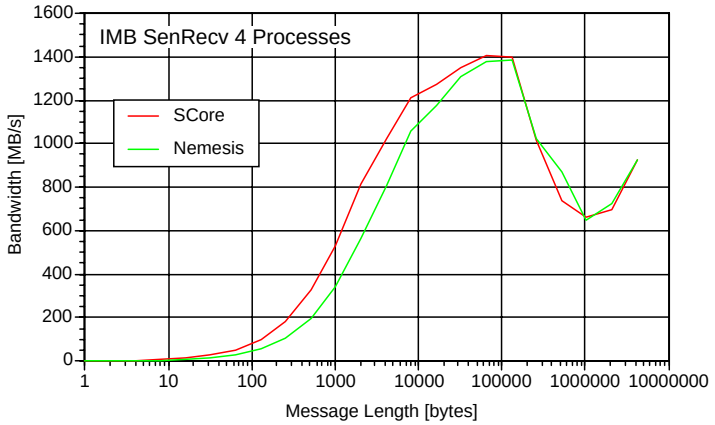
SCore 7 の新機能

SCore 7 開発キーワード

- ほどほどに
  - 性能至上主義からの脱却
  - 3rd-party ライブラリの利用
  - 多少遅くなっても安定性、簡便さを優先
- もっと簡単に，便利に，簡潔に
  - 設定ファイル不要，インストールに root 不要
  - 一体となったパッケージからツールの集合へ

とはいっても速度は大事

- MPICH2: Shmem 性能 Nemesis vs PMX



メニイコアへの対応

- Shmem の改良によるコア間通信の高速化
- CPUソケットの指定

  % scrun -nodes=8x2x4 ./a.out

  % scrun -hosts=64/2/4 ./a.out
- プロセスとコアのバインディング

  % scrun -corebind=0x1:0x2:0x4:0x8 ./a.out

  改良の余地あり

PMX ネットワーク

- 新たなサポート
  - Infiniband　　OFED
  - Myrinet/Myri10G　MX
- Ethernet - 3つの PMX デバイス
  - PMX/SCTP　SCTP (Linux 標準) プロトコル
  - PMX/Ethernet　　ドライバ･パッチ不要
  - PMX/EtherHXB　　ドライバ･パッチ必要 (より高性能)

  上記3つは全て同時使用可能

設定のオプション化

- Scorehosts.db
  - 主にホストグループの記述
  - 変更に際し，デーモンの再起動は不要
- PMネットワーク設定ファイル
  - 従来の pm-ethernet.conf 等は全く不要
- SCOUT
  - Ssh 対応，並列化 => scoutd 不要

=> ソフトをインストールするだけでSCoreの実行が可能，ビルド，インストールに root 権限も不要

ツールセット

- SCoreの内部機能を別途ツールとして独立

Scorehosts	ホストグループの指定
Papion	PAPI による計測
Scan	デバッガのアタッチ
Scratch	行単位でヘッダを付加
Catwalk	On Demand File Staging
Windup	リモートプロセス起動（ssh/rsh）

ツールセットの応用例

```
% scrun  scratch  ptrace  ./a.out
% scrun  scratch  valgrind  ./a.out
% scrun  scratch  papion -f  ./a.out
```

これらは SCore 6 以前ではできなかった !!

ファイルステージング(1)

- eScience プロジェクト
- On Demand File Staging: CATWALK
  - ステージングの記述が不要
  - 記述を間違えない
- 実装

  Exec() や open() 等に hook
- 分散／並列ファイルシステムと共存し，互いに補完するもの

ファイルステージング(2)

- SCore6
  - 実行ファイルのみ自動でステージング
  - 入出力データファイルは陽に記述

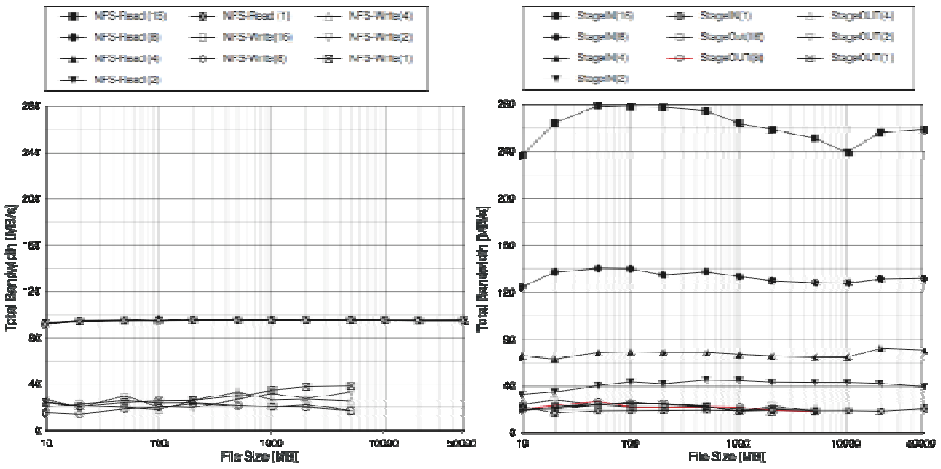
  % scrun stgin infile :: a.out :: stgout outfile
- SCore7

  % scrun -catwalk=DIR0:DIR1:DIR2 a.out

  Infile を DIR0:DIR1:DIR2 から探してコピー

  終了後に outfile をコピー

Catwalk vs. NFS



MPI

- MPICH2 ベースの2つの実装
  - Libmtmi/PMX
    - YAMPIから派生した独自チャネルデバイス
  - NEMESIS/PMX
    - NEMESISベースの実装

今後の予定

ファイナルリリース

- SCore フル機能を実装
  - バッチスケジューラとのタイトな連携
  - One-sided 通信
  - チェックポイント・リスタート
    - デバッガとの連携も視野
  - ギャングスケジューリング
- リリース：2009年秋を予定

eScience 成果の取込み

東大, 京大, 筑波大 - 文科省プロジェクト

- Catwalk - ファイルステージング (SCore7 β2に組込)
- MPI-Adaptor - バイナリコンパチビリティ
- ファイル I/O プロファイラ
- 並列ファイルシステム用キャッシュ
- xCalableMP - 新並列言語 (OpenMP+HPF+α - β)
- パラメータサーチ問題用スクリプト言語

  などなど

リリー:ススケジュール

2008/11	SCore 7 β0	SC@Austin
2008/12	SCore 7 β1	at 東京
2009/3	SCore 7 β2	at 大阪 (3月末)
2009/5	SCore 7 β3	at 広島-メディア配布予定
2009/11	SCore 7	SC@Portland