



eScience プロジェクトと SCore 7

PC Cluster Consortium 開発部会
堀 敦史

2009年5月27日 ワークショップin広島 1



eScience と SCore 7

eScience プロジェクト

2009年5月27日 ワークショップin広島 2



T2K から eScience へ

- T2K プロジェクト
センターマシンの調達コストを下げるために、3大学共通仕様とした。しかし、
- ユーザにとって同じ使い勝手ではない

	バッチスケジューラ	ファイルシステム	MPI
東大	Torque	HSFS	MPICH-MX
筑波大	SGE	Lustre	MVAPICH
京大	NQS	SRFS	富士通 MPI

2009年5月27日 ワークショップin広島 3



eScience の目的

- クラスタの生産性をもっと上げよう！
 - できるだけ同じクラスタ環境であって欲しい
 - 面倒なチューニングは止めたい
 - 同じようなジョブの投入を自動化したい
 - より簡単に、高性能な並列プログラムを書きたい

2009年5月27日 ワークショップin広島 4



同じクラスタ環境の提供

- ファイルシステム
 - 並列／分散ファイルシステムの問題点
 - ファイルステージングを見直そう
 - Catwalk (後述)
 - STG モジュール化された汎用ステージング機構
 - ファイルアクセスのトレース
- MPI の違い
 - MPI-Adaptor が違いを吸収
 - 再コンパイル, 再リンク不要

2009年5月27日 ワークショップin広島 5



手動チューニングは大変

- コンパイラのパラメータ, CPU の性質, ネットワークの特性等々...
 - クラスタ毎にチューニングするのは大変
- 自動チューニング機構付き並列数値ライブラリ
 - ABCLib_LANCOZ
 - 標準固有値問題: リスタート付きランチョス法ライブラリ
 - ABCLib_GMRES
 - 連立一次方程式: GMRES(m)法ライブラリ

2009年5月27日 ワークショップin広島 6

ジョブスクリプト言語

- Xcrypt(エクスクライプト)
 - Perl ベース
- 目的
 - 同じようなジョブの大量投入を自動化したい
 - バッチスケジューラの違いを意識したくない

2009年5月27日

ワークショップin広島

7

新並列プログラミング言語

- xCalableMP(エクスケラブルMP)
 - これまでの並列言語の「いい所取り」
 - Fortran HPF, CAF, ...
 - C UPC, ...
 - MPI や OpenMP とも共存可能
 - 新規性よりも使えるものを
 - 今年度中に最初の動く版が完成

2009年5月27日

ワークショップin広島

8

eScience と SCore

eScience の成果は
フリーソフトウェアとして
SCore と共にリリースされる

2009年5月27日

ワークショップin広島

9

SCore 7

SCore 7 の新機能

2009年5月27日

ワークショップin広島

10

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

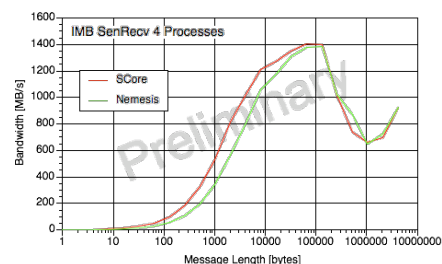
2009年5月27日

ワークショップin広島

11

とはいっても速度は大事

- MPICH2: Shmem 性能 Nemesis vs PMX



2009年5月27日

ワークショップin広島

12

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

2009年5月27日

ワークショップin広島

13

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-高性能
- 上記3つは全て同時使用可能

2009年5月27日

ワークショップin広島

14

設定のオプション化

- ・ Scorehosts.db
 - 主にホストグループの記述
 - 変更の際し、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

2009年5月27日

ワークショップin広島

15

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	デバッグのアタッチ
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)

2009年5月27日

ワークショップin広島

16

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 !!

2009年5月27日

ワークショップin広島

17

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

2009年5月27日

ワークショップin広島

18

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

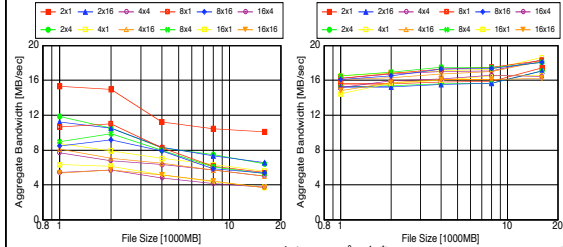
2009年5月27日

ワークショップin広島

19

Catwalk vs. NFS (1)

- ひとつのファイルを各プロセスが 1/N 読む



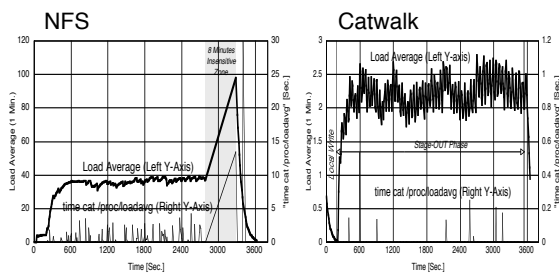
2009年5月27日

ワークショップin広島

20

Catwalk vs. NFS (2)

- 4x16 の各プロセスが 1GB のファイルを書込む



2009年5月27日

ワークショップin広島

21

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

2009年5月27日

ワークショップin広島

22

SCore 7

今後の予定

2009年5月27日

ワークショップin広島

23

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

2009年5月27日

ワークショップin広島

24



リリーススケジュール

2008/11	SCore 7 β 0	SC@Austin
2008/12	SCore 7 β 1	at 東京
2009/5	SCore 7 β 2	at 広島
2009/11	SCore 7	SC@Portland