

SCore 7 最新状況

PC Cluster Consortium 開発部会
堀 敦史

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

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

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 (要カーネル)
 - 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

コマンドの組み合わせ可能

```
% 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 !!



SCore 7 4の新機能

- SCore 7 3に加わった機能

メニイコアへの対応

- ・ 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
```

- ・ MPI と OpenMP のハイブリッド

```
% scrun -openmp=4 ./a.out
```

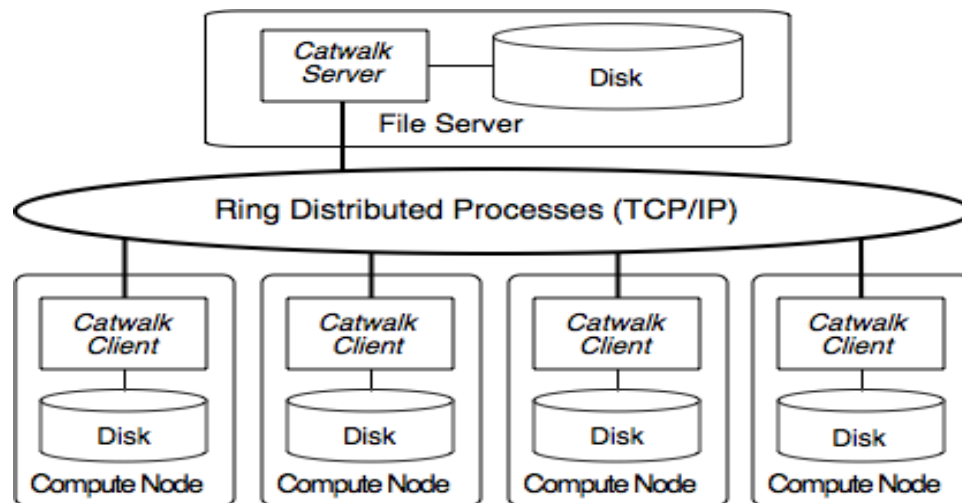
- ・ NUMA に関しては現在研究中

Catwalk ファイルシステム

- ・ eScience プロジェクト
- ・ Catwalk
 - On Demand File Staging
 - ・ ステージングの記述が不要なので、記述を間違えない
 - Catwalk-ROMIO: MPI-IO インターフェイス
- ・ 既存の分散 / 並列ファイルシステムに独立なので、共存可能
- ・ root 権限不要なのでインストールするだけで利用可能

Catwalk のプロセス構造

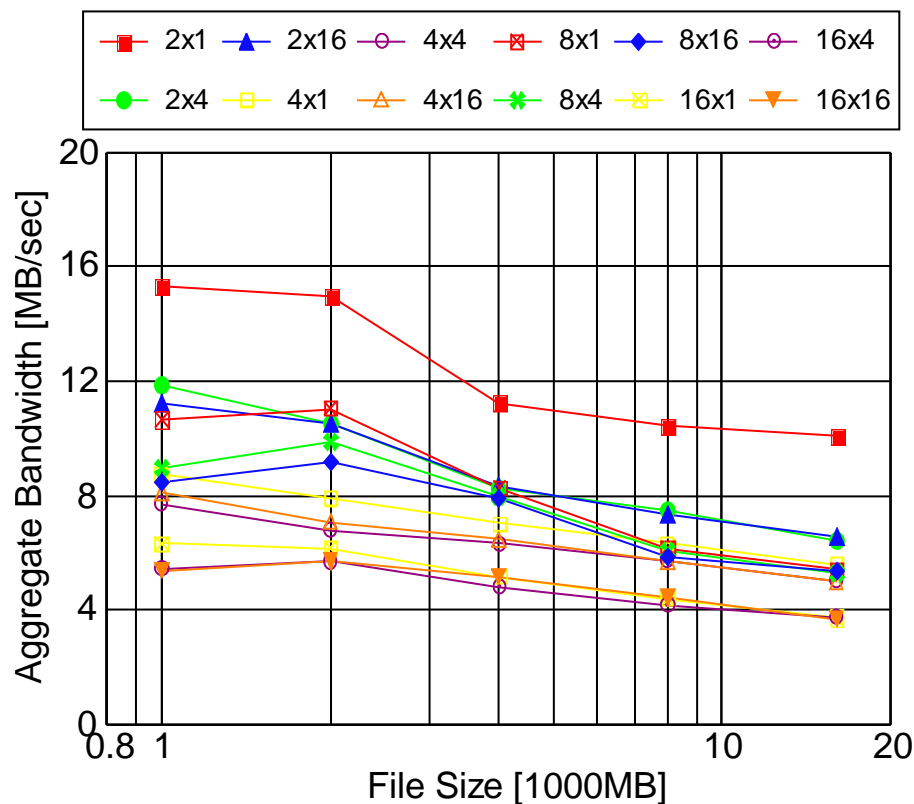
- ・ Catwalk サーバ
 - 最終的なファイルの在処
- ・ Catwalk クライアント
 - 一時的なファイルの置き場 (キャッシュ)



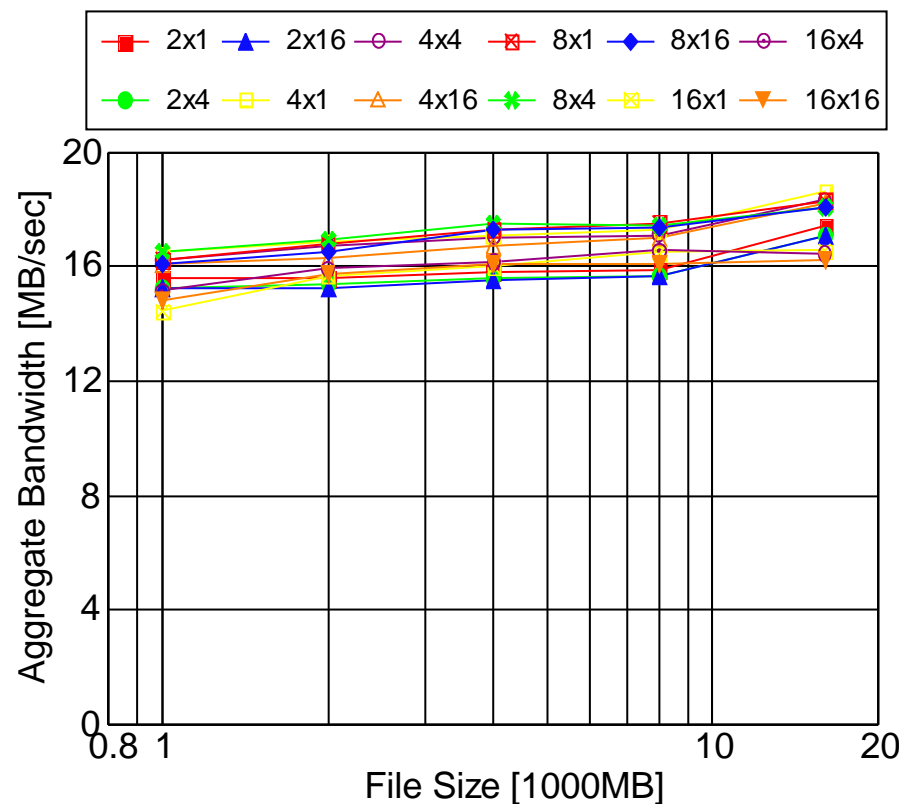
Catwalk vs. NFS (1)

- ひとつのファイルを各プロセスが $1/N$ 読む

NFS



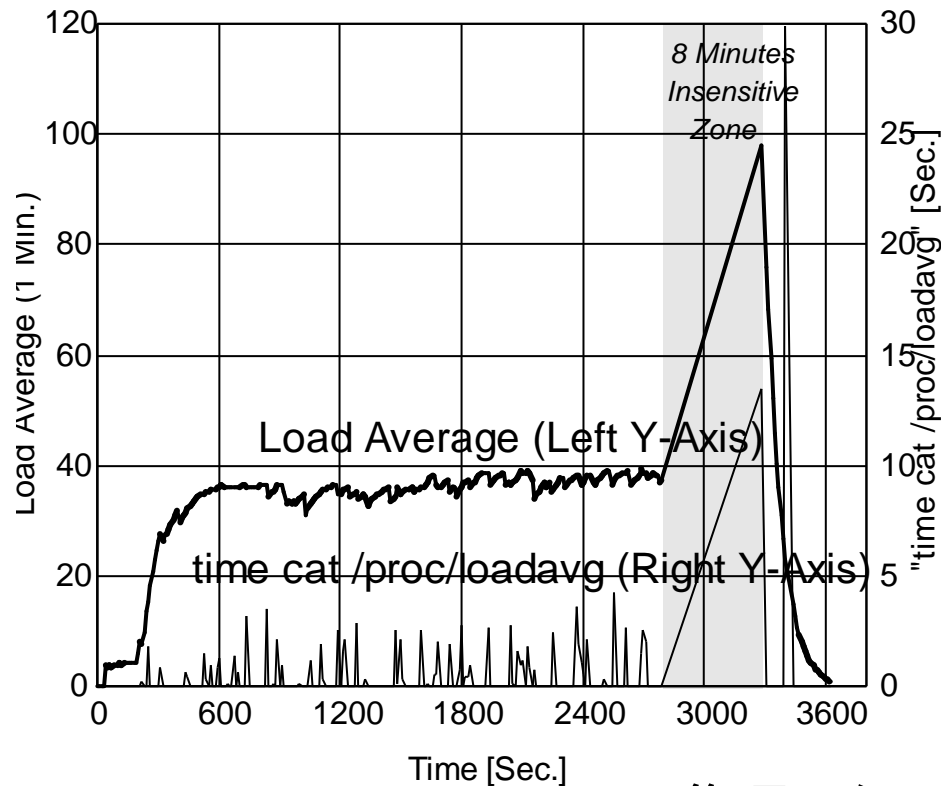
Catwalk



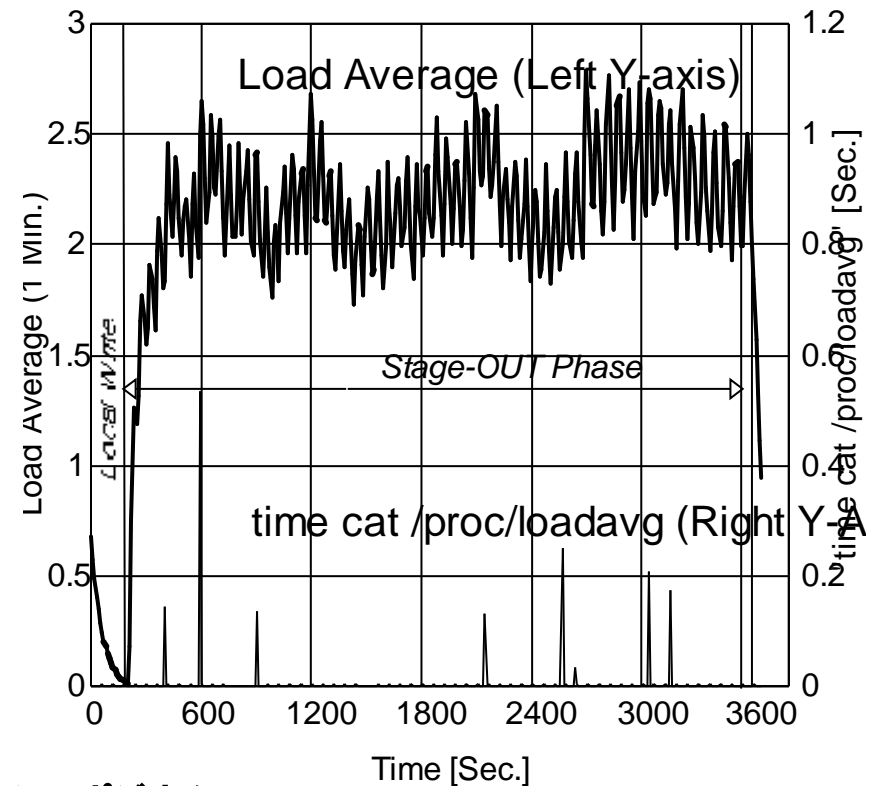
Catwalk vs. NFS (2)

- 4x16 の各プロセスが 1GB のファイルを書込む際のサーバの負荷

NFS



Catwalk



Catwalk の利用方法

- ・ その1

login% catwalk mpirun catwalk a.out

サー

バ

クライアント

- ・ その2 (ssh-agent に似た方式)

server% catwalk

サー

バ

環境変数が表示

login% export 環境変数

login% mpirun catwalk a.out

クライアント



事前にコピーする
必要がない



Catwalk の今後

- セキュリティへの配慮
 - Catwalk サーバにパスフレーズを設定
 - ssh 経由でクラスタにログインした場合に ssh のポートフォワーディングを使えるようにする

OOM_KILLER への対応

- ・ Linux でクラスタ運用する場合の大きな問題
 - OOM_KILLER
 - メモリが足りなくなった時にカーネルが発動する「ロシアンルーレット」
 - 今回のリリースで対応
 - ・ SCore の並列ジョブは、より OOM_KILLER の対象となるように自動的に設定
 - ・ これによりメモリが足りなくなってもノードが落ちる(使えなくなる)現象を回避



ファイナルリリースへ

- ・ 以下の機能を実装したら を外す予定
 - One-sided 通信
 - ギャングスケジューリング
- ・ 2010 年度中を予定

- 今回リリースに含まれるもの
 - ファイルステー징システム
 - STG, Catwalk
 - MPI-Adapter (本日 14:10～)
 - XAbclib 自動チューニング機能付き数値ライブラリ
- 今後のリリースに含まれる予定のもの
 - XcalableMP (本日 14:40～)
 - Xcrypt ジョブ投入スクリプト言語