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

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

設定のオプション化

- Scorehosts.db
  - 主にホストグループの記述（オプション）
  - 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（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 β5 の新機能

SCore 7 β3 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 権限不要なのでビルドするだけで利用可能

Catwalk の使い方 (1)

% cat > *DIR*/foo.dat

% scout -g *HOSTGROUP*

scout は *HOSTGROUP* 上で動く並列シェル

% scout cat *DIR*/foo.dat

cat: *DIR*/foo.dat: No such file or directory

% cd *DIR*

% scout -catwalk *DIR* cat foo.dat

HOSTGROUP 上の各ホストで，サーバの *DIR*/foo.dat を読むことができる

Catwalk の使い方 (2)

% catwalk -nh *N* mpirun catwalk a.out

サーバ　　ノード数　　クライアント

% catwalk –path A:B:C –nh N ...

サーバのディレクトリ A,B,C を順に探す

% catwalk –if ib0 ...

Infiniband を使う

事前にコピーする必要がない

Catwalk-ROMIO の使い方

% catwalk –romio mpirun catwalk a.out

a.out で MPI_File_open("catwalk:/xxx/yyy") とするとサーバの /xxx/yyy がアクセスされる

server% catwalk –mpi –SSH comp32

Catwalk 環境と SSH port forwarding を設定し

comp32 に ssh でログイン

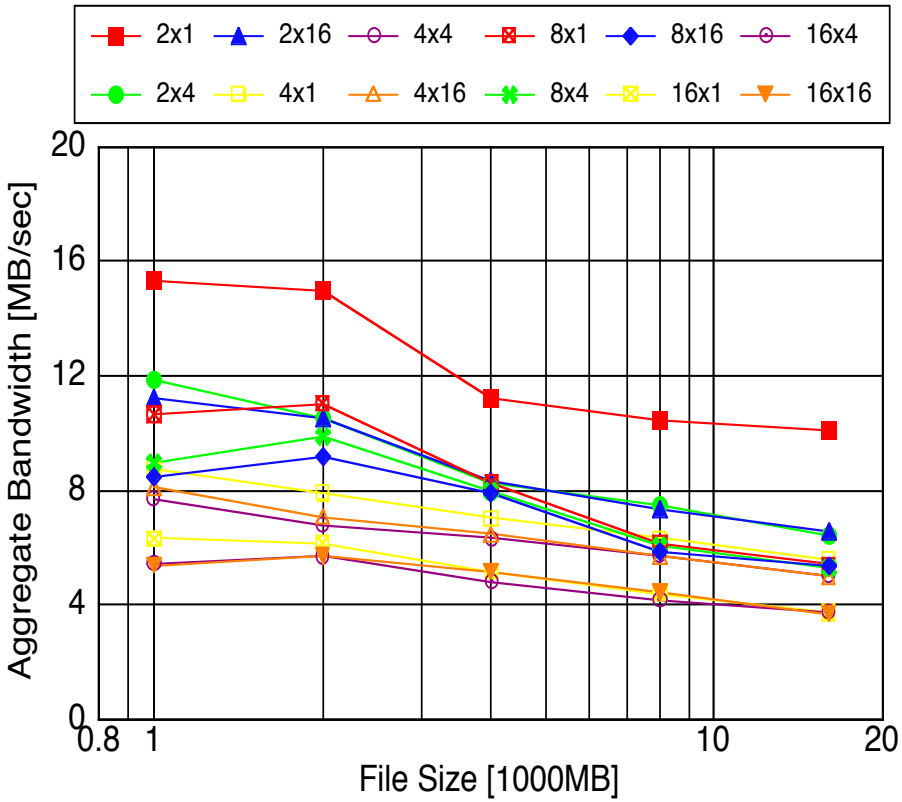
comp32% mpirun catwalk a.out

server 上のファイルを MPI-IO でアクセス.

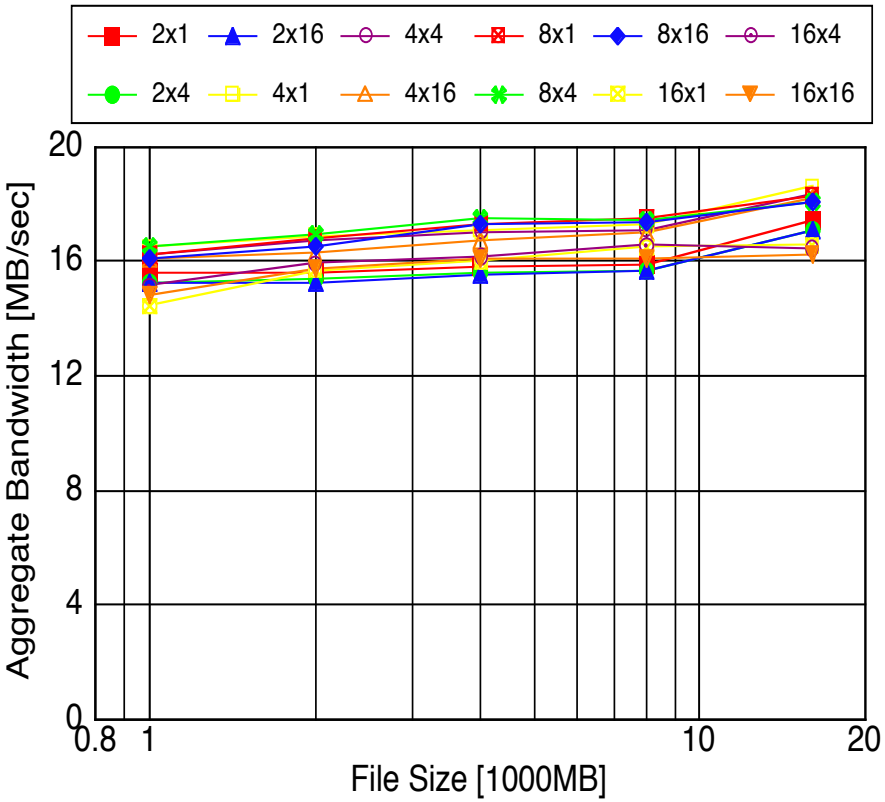
Catwalk vs. NFS (1)

・ひとつのファイルを各プロセスが 1/N 読む

NFS



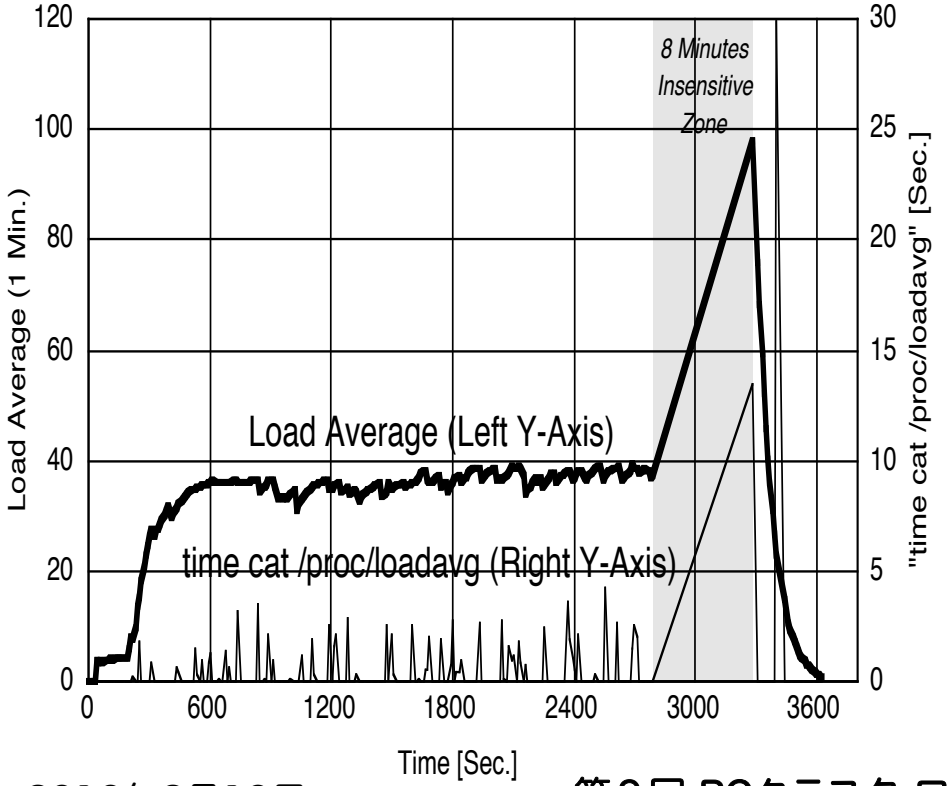
Catwalk



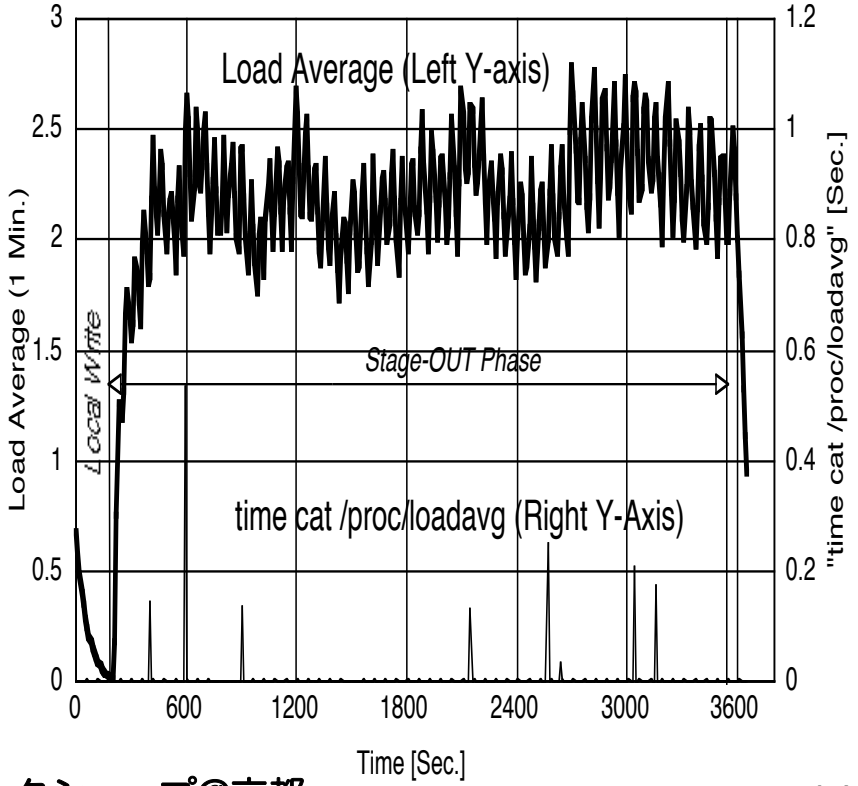
Catwalk vs. NFS (2)

- 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

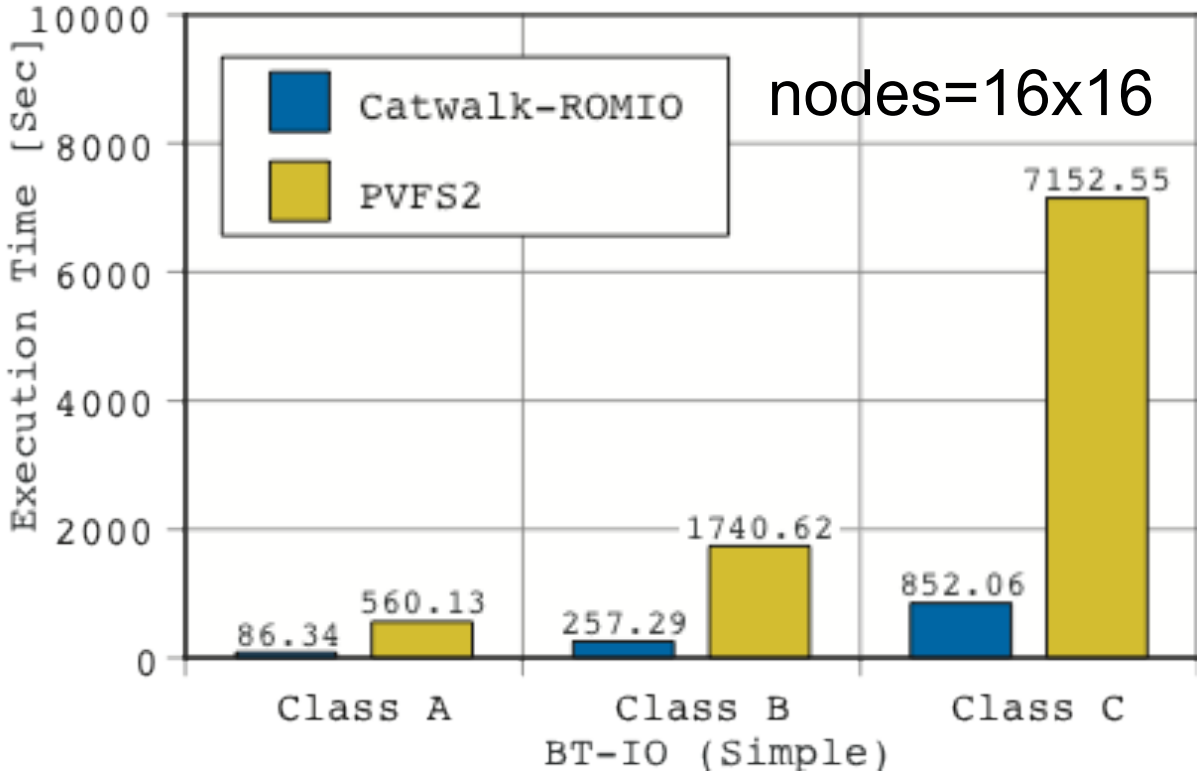


Catwalk



Catwalk-ROMIO vs. PVFS2

- Catwalk-ROMIO
  サーバ1台
  Ethernet (1 Gb/s)
- PVFS2
  サーバ4台
  Myri10G (10 Gb/s)



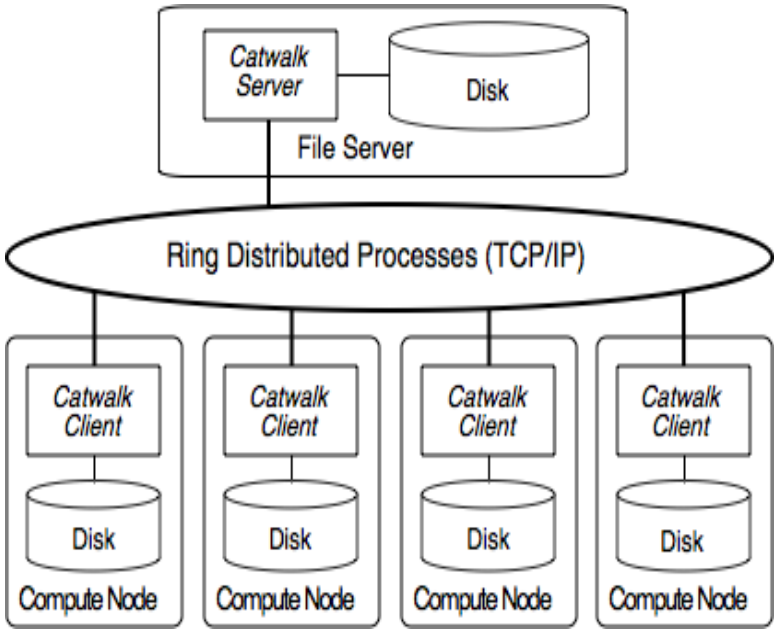
Catwalk の内部

リング分散プロセス構造

並列アクセスを出来るだけ逐次アクセスに

→ シークを減らして高速化

- Stage-IN
  - 対象ファイルを全てローカルディスクにコピー
- Stage-OUT
  - 逐次全体コピー
- MPI-IO Read
  - Stage-IN と同じ
- MPI-IO Write
  - 並列書込要求を逐次書込に変換



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 とは？

メモリ不足の時にカーネルが発動する

「ロシアンルーレット」

・ OOM_KILLER の対応

– 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β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

eScience プロジェクト

- 今回リリースに含まれるもの
  - STG, Catwalk　　ファイルステージング
  - libioc　　ファイル I/O のトレース
  - MPI-Adapter　　異 MPI 環境での互換性を保つ
  - Xabclib　　自動チューニング数値 Lib.
  - Xcrypt　　ジョブ投入スクリプト言語
- 今後のリリースに含まれる予定のもの
  - XcalableMP　　新並列プログラミング言語