シームレス高生産・高性能プログラミング環境:「高効率・高可搬性ライブラリの開発」

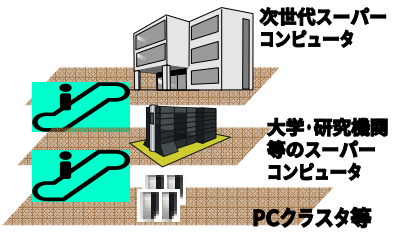
東京大学　　石川裕

T2K Open Supercomputer Alliance

全体背景

- 科学技術・学術研究における萌芽的研究は、研究室レベルで進められているが、利用できるコンピュータの能力程度の問題規模を解くに留まることが多い

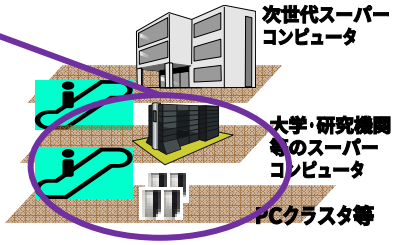
  理由:
  - 研究室レベルで使用されているコンピュータ上で開発されたアプリケーションを計算センターに設置されているスパコン上に持っていっても動かないことがある
  - また動いたとしても大規模問題を解くことができないことが多い
- 大規模問題を解くプログラムは、並列処理に関する深い知識に加えて、スパコンの性能を引き出す豊富なノウハウを持つ、限られたプログラマにしか作れない



並列アプリケーション生産性拡大のための道具
- 並列言語処理系
- 並列スクリプト言語
- ライブラリ＆チューニングツール

並列アプリケーション生産性拡大のための道具(1/2)

PCクラスタから大学情報基盤センター等に設置されているスパコンまで、ユーザに対するシームレスなプログラミング環境を提供

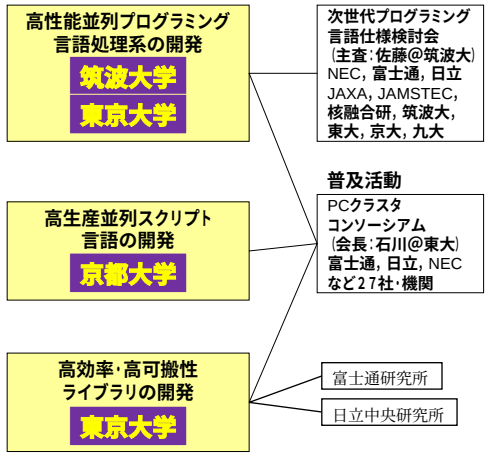


並列アプリケーション生産性拡大のための道具
- **並列言語処理系**
- **並列スクリプト言語**
- **ライブラリ＆チューニングツール**

並列アプリケーション生産性拡大のための道具(2/2)

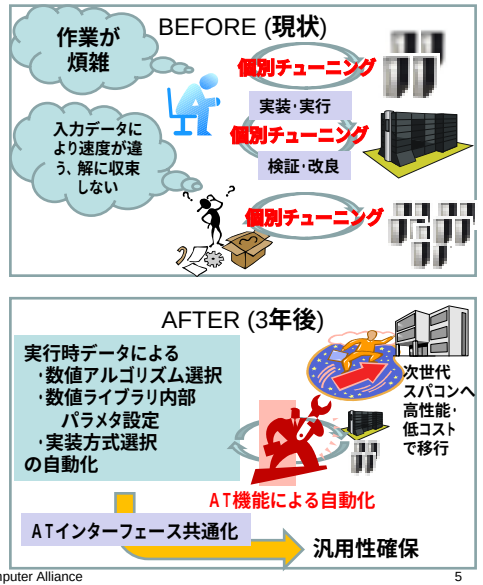
PCクラスタから大学情報基盤センター等に設置されているスパコンまで、ユーザに対するシームレスなプログラミング環境を提供

- **高性能並列プログラミング言語処理系**
  - 逐次プログラムからシームレスに並列化および高性能化を支援する並列実行モデルの確立とそれに基づく並列言語コンパイラの開発
- **高生産並列スクリプト言語**
  - 最適パラメータ探索など粗粒度の大規模な階層的並列処理を、簡便かつ柔軟に記述可能で処理効率に優れたスクリプト言語とその処理系の開発
- **高効率・高可搬性ライブラリの開発**
  - 自動チューニング（AT）機構を含む数値計算ライブラリの開発
  - PCクラスタでも基盤センタースパコン(1万規模CPU)でも単一実行時環境を提供するSingle Runtime Environment Image環境の提供



高性能高可搬性ライブラリ[自動チューニング付き数値計算ライブラリ] (1/3)

- 現状と課題
  - 数値計算プログラムの性能チューニングコストの増大
    - PCクラスタから計算センターに設置されているスパコンなど、様々なアーキテクチャ毎にチューニングしなければならない
  - 実行時に与えられるデータセットの内容により、最適な数値計算アルゴリズムや実装方式が異なる
- 目標
  - 自動チューニング（AT）機構を含む数値計算ライブラリを開発し、数値計算コード部分のチューニングに関わる開発時間をなくす



高性能高可搬性ライブラリ[自動チューニング付き数値計算ライブラリ] (2/3)

- 研究項目
  - 実施責任者の片桐の既存研究成果である自動チューニング(AT)機能付き数値計算ライブラリABCLibの方式を大規模計算シミュレーションプログラムに適用・評価
  - AT機能のAPI化、軽量化、および、マルチコア向けAT方式
    - 日立中央研究所と共同研究（予定）
  - 対象プログラムは、固有値解析・連立1次方程式の反復解法の数値計算ライブラリ
- 期待される成果
  - AT機能をAPI化することにより、他の数値解法ライブラリにAT機能を追加しやすくなる
  - 開発する数値計算ライブラリを用いることで、PCクラスタからセンターマシンに至る広範な計算機環境においても、コード修正なしで高い性能を達成
  - 実行時に与えられるデータの変化に追従する自動チューニング機能により、従来の数値計算ライブラリでは達成できなかった高速化が実現

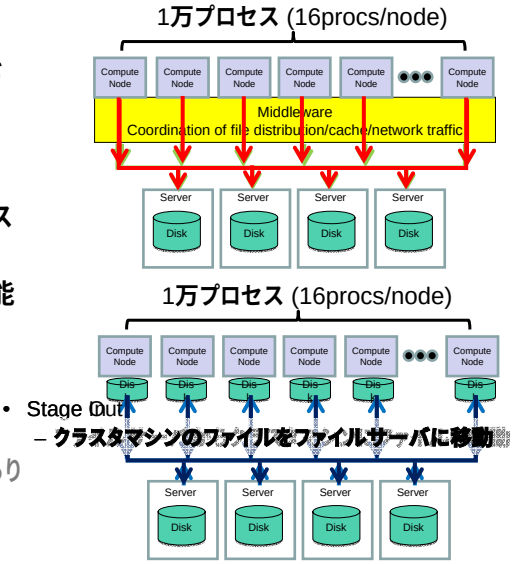
高性能高可搬性ライブラリ[自動チューニング付き数値計算ライブラリ] (3/3)

(e): e-Scienceプロジェクトで独自に研究開発するＡＴ機能
(注)：ソースコードは全て新規開発

- 平成20年度開発ライブラリ
  - ABCLib_LANCOZ
    - 標準固有値問題の求解のための、リスタート付きランチョス法ライブラリ
  - ABCLib_GMRES
    - 連立一次方程式の求解のための、GMRES(m)法ライブラリ
  - (e) OpenATlib
    - 共通自動チューニングインターフェースライブラリ
- 実用問題行列を集めた＜フロリダ行列＞から各ライブラリ6種 (計12種) を選別し、実行時ＡＴで基準値以上の速度向上達成を保証
- 以下のＡＴ機能を平成２０年度プロトタイピングで実装
  - リスタート周期を入力行列の特性を考慮し自動調整
  - 疎行列-ベクトル積の実装方式を入力行列の特性を考慮し自動選択
  - (e)起動時スレッド数ごとの最適実装方式選択 (マルチコア向けＡＴ機能)
  - (e)実行時に定まるメモリ残量を考慮した最適実装方式選択
  - (e)ＡＴ機能の実装を容易にする＜共通自動チューニングインターフェース＞

高性能高可搬性ライブラリ[Single Runtime Environment]　概略

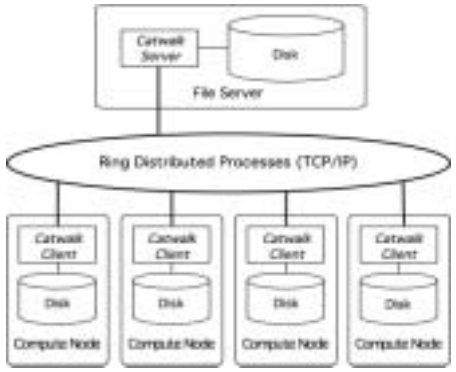
- ファイルシステム
  - 1万プロセスからのファイルI/Oを効率よく実現する
  - ファイルI/O API
  - ベンチマークプログラム集
- バッチジョブシステムインターフェイス
  - ステージングの導入
  - バッチジョブコマンド変換機能
- MPI通信ライブラリ
  - ABI (Application Binary Interface)が定義されていない
    例：OpenMPIにおける MPI_Comm型はアドレス型であり、他の実装は32bit integer



Catwalk: An Overview[Hori09]

- Transparent File Staging
  - The users do not take care of the file staging commands, but the Catwalk midleware takes care of it
    - At a file open, the Catwalk copies the file from the file server to the local disk if the file does not exist in the local disk
    - At a file close, the Catwalk copies the file from the local disk to the file server
- Assuming Environment
  - TCP/IP connection between the file server and the cluster
    - Requires some coordination of network traffic
  - No requirement of highly network bandwidth
  - No requirement of the administrator mode to install Catwalk

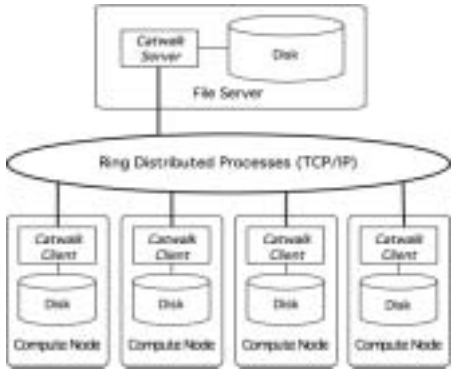
- New Oxford America Dictionary
  A ***narrow walkway*** or platform extending into an auditorium, esp. in an industrial installation, along which models walk to display clothes in fashion shows.
- **http://en.wikipedia.org/wiki/Catwalk**
  Typically, catwalks are located in positions ***hidden from audience*** view or directly above an audience, and are considered "behind-the-scenes".



Catwalk: An Overview[Hori09]

- Transparent File Staging
  - The users do not take care of the file staging commands, but the Catwalk midleware takes care of it
    - At a file open, the Catwalk copies the file from the file server to the local disk if the file does not exist in the local disk
    - At a file close, the Catwalk copies the file from the local disk to the file server
- Assuming Environment
  - TCP/IP connection between the file server and the cluster
    - Requires some coordination of network traffic
  - No requirement of highly network bandwidth
  - No requirement of the administrator mode to install Catwalk

- Catwalk consists of
  - user library
  - Client process
  - Server process

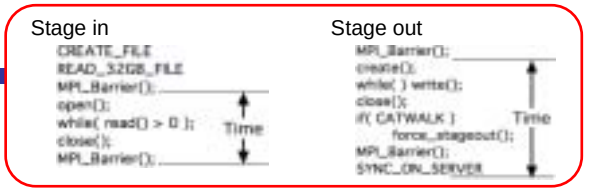


CatWalk: Evaluation

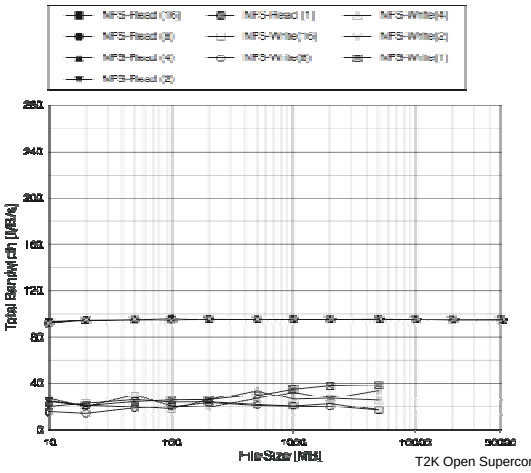
- T2K Open Supercomputer
- 17 nodes
  - One for file server and 16 for compute nodes
- Network
  - 1Gbps Ethernet

CPU	AMD Barcelona, 2.3GHz, 4x4 cores
Memory	32 GB
Local Disk	SATA
Network	Intel E1000, Myrinet 10G
OS	RHEL5 5.1
File System	EXT3, NFS Ver.3

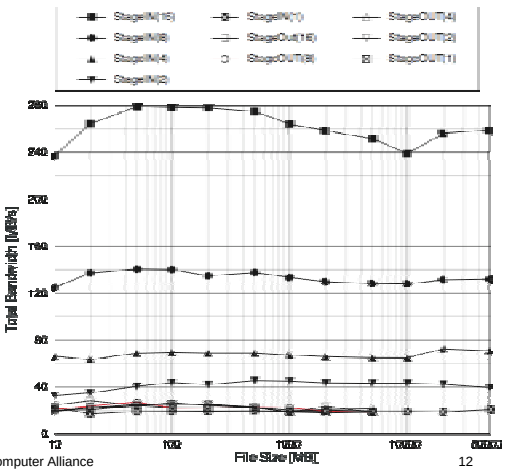
CatWalk: Evaluation



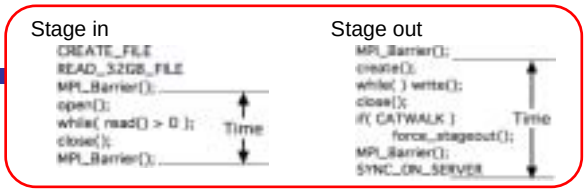
NFS



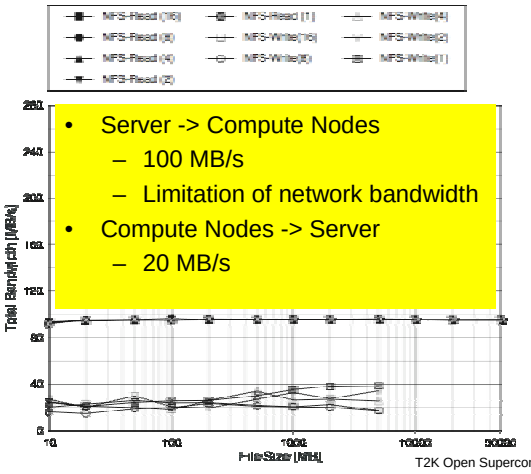
CatWalk



CatWalk: Evaluation

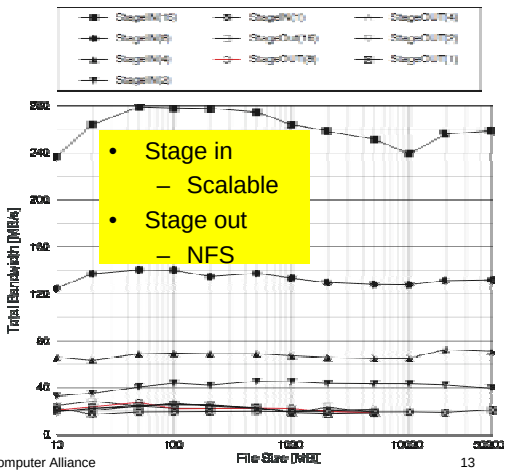


NFS



- Server -> Compute Nodes
  - 100 MB/s
  - Limitation of network bandwidth
- Compute Nodes -> Server
  - 20 MB/s

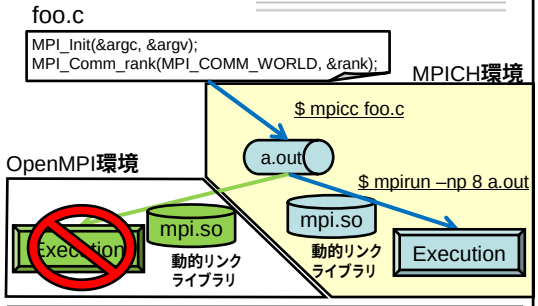
CatWalk



- Stage in
  - Scalable
- Stage out
  - NFS

高性能高可搬性ライブラリ[Single Runtime Environment] 概略

- ファイルシステム
  - 1万プロセスからのファイルI/Oを効率よく実現する
  - ファイルI/O API
  - ベンチマークプログラム集
- バッチジョブシステムインターフェイス
  - ステージングの導入
  - バッチジョブコマンド変換機能
- MPI通信ライブラリ
  - ABI (Application Binary Interface)が定義されていない
    例：OpenMPIにおける MPI_Comm型はアドレス型であり、他の実装は32bit integer

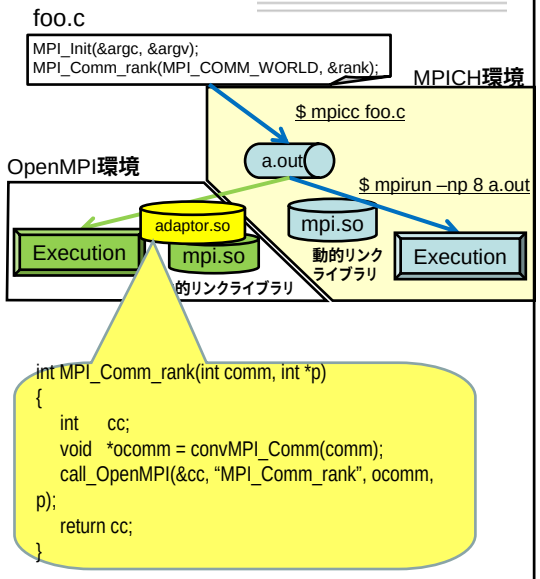


Constant	MPICH2	OpenMPI
MPI_COMM_WORLD	0x44000000	&ompi_mpi_comm_world
MPI_INT MPI_INTEGER	0x4c000405 0x4c00041b	&ompi_mpi_int &ompi_mpi_integer
MPI_SUCCESS MPI_ERR_TRUNCATE	0 14	0 15
MPI_COMM_WORLD	0x44000000	0
MPI_INTEGER	0x4c00041b	&ompi_mpi_integer
MPI_SUCCESS MPI_ERR_TRUNCATE	0 14	0 15

MPI-Adaptor [Sumimoto09]

- ファイルシステム
  - 1万プロセスからのファイルI/Oを効率よく実現する
  - ファイルI/O API
  - ベンチマークプログラム集
- バッチジョブシステムインターフェイス
  - ステージングの導入
  - バッチジョブコマンド変換機能
- MPI通信ライブラリ
  - ABI (Application Binary Interface)が定義されていない

  例：OpenMPIにおける MPI_Comm型はアドレス型であり、他の実装は32bit integer



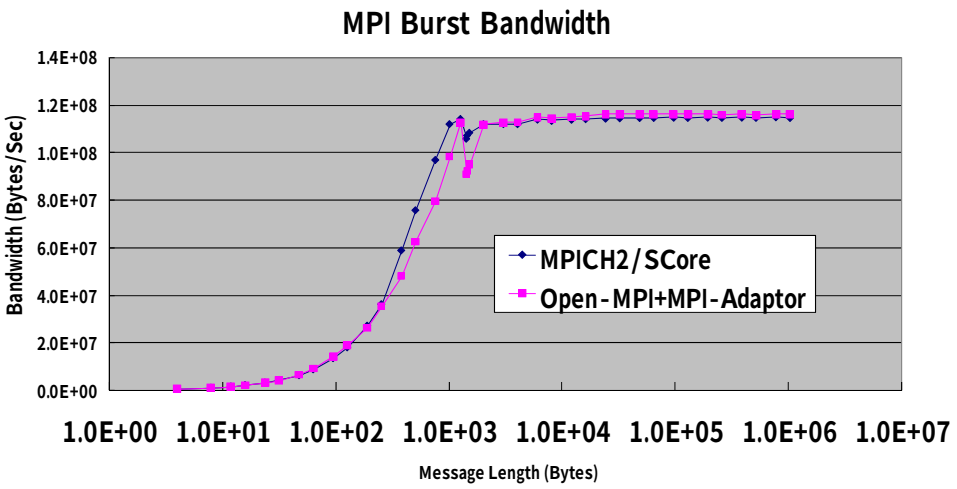
MPI-Adaptor: Evaluation

- MPI-Pingpong(mpi_rtt)
- MPICH2/SCore
  - MPICH2/SCore 環境でコンパイル
- OpenMPI+MPI-Adaptor
  - OpenMPI環境でコンパイル
  - MPI-Adaptor on MPICH2/SCore 環境で実行

RX200S2 Cluster (Xeon 3.8GHz, SCore7.0)
Network: Intel E1000 NIC,
Netgear 48Port Switch
MPI MPICH2/SCore w/ PMX/Etherhxb

	RTT(usec)	Ratio
MPICH2/SCore	43.328	100%
OpenMPI+ MPI-Adaptor	43.440	100.2%

MPI-Adaptor: Evaluation



MPI-Adaptor: Evaluation

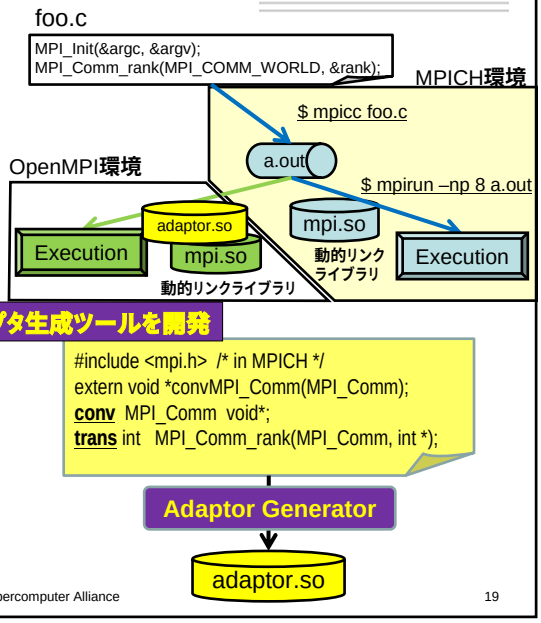
	NPB IS Class A
MPICH2/SCore	50.75 Mops
OpenMPI+MPI-Adapter	48.47 Mops

- 4.7% **の性能低下**
- **詳細な検討は今後**
- **他のベンチマークプログラムによる評価**
  - Fortran**向け**MPI Adaptor**の実装が必要**

MPI-Adaptor

- ファイルシステム
  - 1万プロセスからのファイルI/Oを効率よく実現する
  - ファイルI/O API
  - ベンチマークプログラム集
- バッチジョブシステムインターフェイス
  - ステージングの導入
  - バッチジョブコマンド変換機能
- MPI通信ライブラリ
  - ABI (Application Binary Interface)が定義されていない

  例：OpenMPIにおける MPI_Comm型はアドレス型であり、他の実装は32bit integer



おわりに

- 開発したシステムはSCore Version7に統合
  - CatWalk, MPI-Adaptorは5月下旬αリリース予定
- ファイルシステム
  - 今回紹介しなかったが、ファイルキャッシュシステム[Ohta09]、ファイルステージングシステムも開発中
- ファイルアクセストレーサ
  - 今回紹介しなかったが、アプリケーションファイルアクセスパターンをログするシステムも開発中

参考文献：

[Hori09] 堀、鴨志田、松葉、安井、住元、石川、「ファイルステージング再考：オンデマンド化と高速化に向けたプロトタイプ実装」、情報処理学会、HOKKE'09、2009.

[Sumimoto09] 住元、中島、成瀬、久門、安井、鴨志田、松葉、堀、石川、「並列プログラムの実行可搬性を実現するMPI通信ライブラリ」、情報処理学会、HOKKE'09、2009.

[Ohta09] 太田、石川、「ファイルサーバ独立な並列ファイルキャッシュ機構」」、情報処理学会、HOKKE'09、2009.