

PCクラスタシステムへの 富士通の取り組み

富士通株式会社
株式会社富士通研究所
久門 耕一

PCクラスタワークショップin京都(2010年2月19日)

All Rights Reserved, Copyright © FUJITSU LIMITED 2009

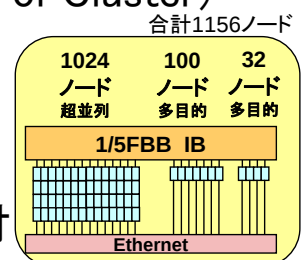
2009年度に富士通が提供する(した)大規模クラスタ **FUJITSU**

- 今年度はCPUとしてメモリバンド幅がNehalem, QDR-IB(片方向4GB/s)などPCクラスタにとって期待できる多くのコモディティコンポーネントが出現
- これら魅力ある素材を使ったシステムとして、2つのシステムをご紹介します

■ 理化学研究所様RICC(Riken Integrated Cluster of Cluster)

特徴:

- Nehalem1000+100(w/GPU)ノードクラスタ
- 上位スイッチのバンド幅を1/5に絞り、コスト圧縮スケジューラによりバンド幅有効利用のジョブ割付



■ 日本原子力研究開発機構 様 新スパコンシステム

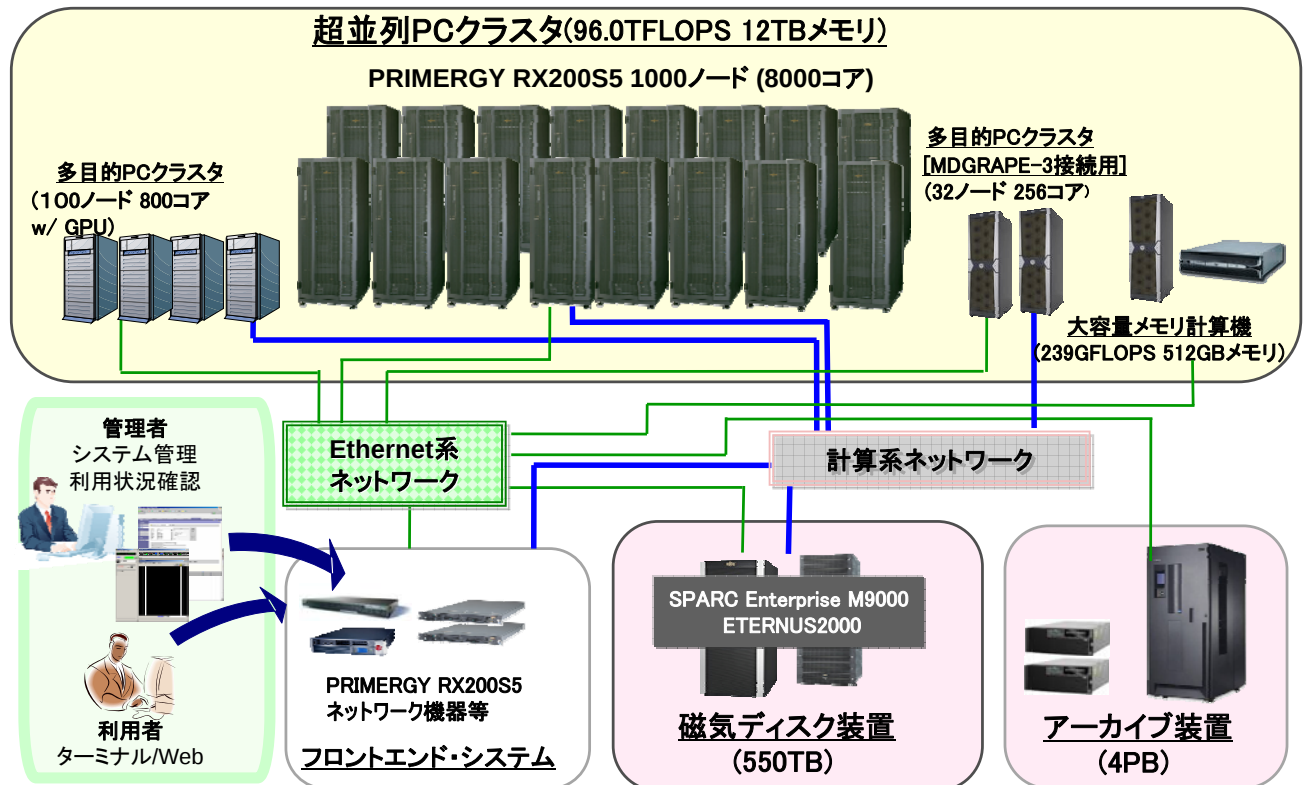
特徴:

- ブレードサーバBX900と内蔵QDR-InfiniBandによるPCクラスタ

理化学研究所様 新スパコンシステム

FUJITSU

■ 2009年8月から稼働開始 Riken Integrated Cluster of Cluster (RICC)



日本原子力研究開発機構 様 新スパコンシステム概要

FUJITSU

■ 2010年3月稼働予定

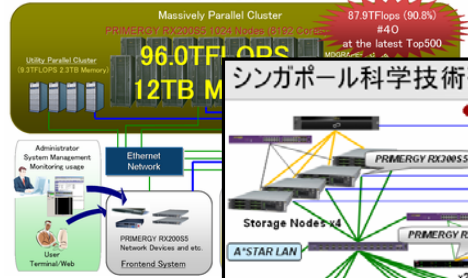


2009年度大規模スパコン商談

FUJITSU

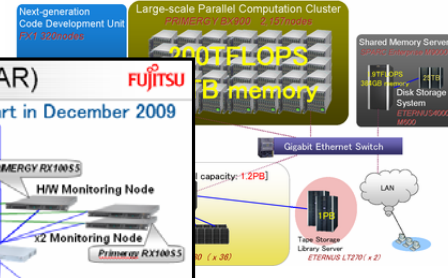
理化学研究所(RIKEN)

● 2009年8月1日運用開始



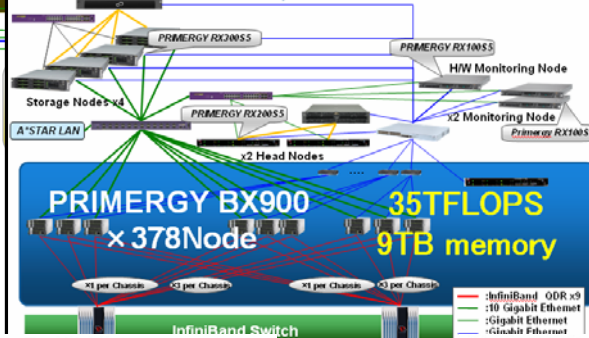
日本原子力研究開発機構(JAEA)

● 2010年3月運用開始予定



シンガポール科学技術研究庁(A*STAR)

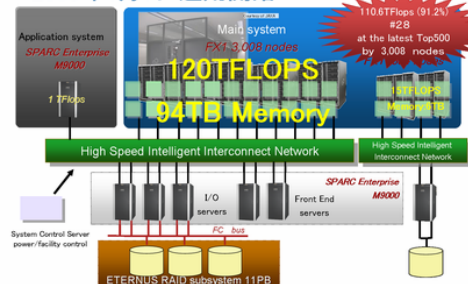
● Operation will start in December 2009



合計7500ノード
530TFlopw

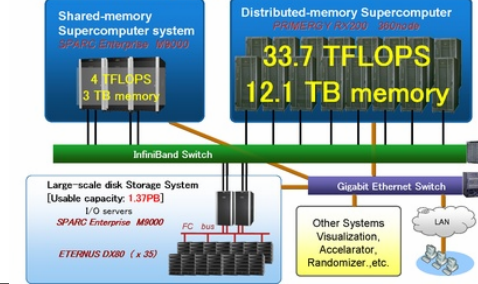
宇宙航空研究開発機構(JAXA)

● 2009年4月1日運用開始



統計数理研究所(ISM)

● 2010年1月運用開始予定



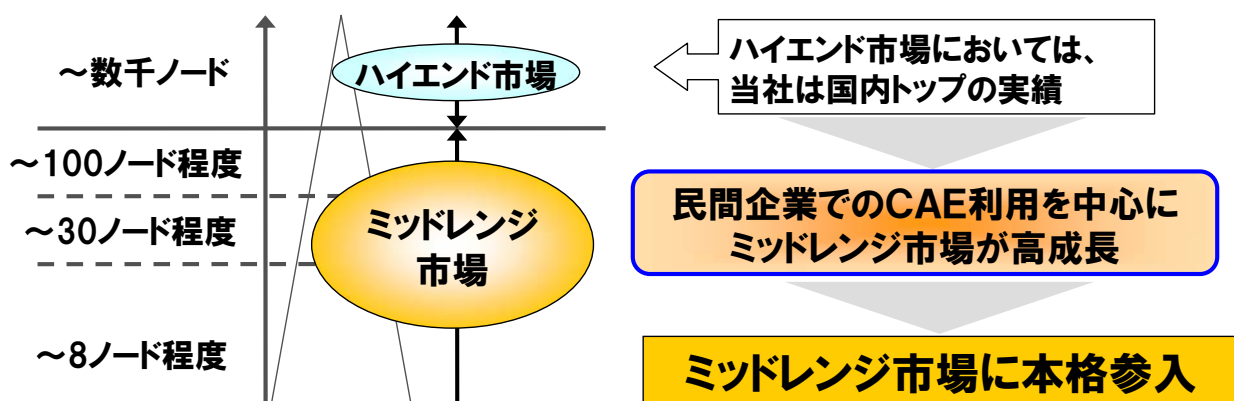
PCクラスタワークショップin京都(2010年2月19日)

4

All Rights Reserved, Copyright © FUJITSU LIMITED 2009

PCクラスタビジネス推進室を設置 (2009.12.21)

FUJITSU



- ・ PCクラスタのエキスパートによる専任組織化
- ・ ハイエンド市場で培った高度なノウハウを裾野へ拡大
- ・ おすすめモデルによる簡単導入
- ・ OSSなどを活用した新たなサービス提供
- ・ FTS (富士通テクノロジーソリューション) との連携によるグローバル展開

PCクラスタワークショップin京都(2010年2月19日)

5

All Rights Reserved, Copyright © FUJITSU LIMITED 2009

10Uに18ノード搭載可能 大規模クラスターや、部門間サーバ集約などに最適なブレードサーバ

■ 業界最高クラスの実装効率と設置性

- 18ブレード(2CPU/ブレード) /10Uシャーシ
- 高効率部品の採用による低電力設計
- 新静音ファンの開発による高い静粛性を実現

■ 最先端のHPC向け仕様

- Intel Xeon5500シリーズ(Nehalem)を2ソケット搭載
- InfiniBand™ QDR(4GB/s)をメザニンカードで搭載
- 10Gbシリアル通信可能な高品質バックプレーンで接続
- 内蔵36ポートSWにより外部18ポート出力。数シャーンまでの小中規模クラスターを外部SWレスで構築可能

■ 高信頼設計

- ECCメモリ (SDDC対応)、RAIDディスク、冗長電源、リモート管理機能等

SDDC (Single Device Data Correction)



PRIMERGY BX900



BX920 S1
サーバブレード



SX940 S1
ストレージブレード

理研システム新システムの性能(LINPACK)

- 2009年6月 40位 1032ノード(8256コア)
87.890TF(90.83%)

国内PCクラスター位(発表時)

- 富士通研開発のDGEMMを用い、事業部がHPLをチューニング

- 2009年11月 47位 1131ノード (9048コア)
97.940TFlops(92.36%)

Nehalemのターボモードを使い単体性能向上

■ Nehalemのターボモードとは、

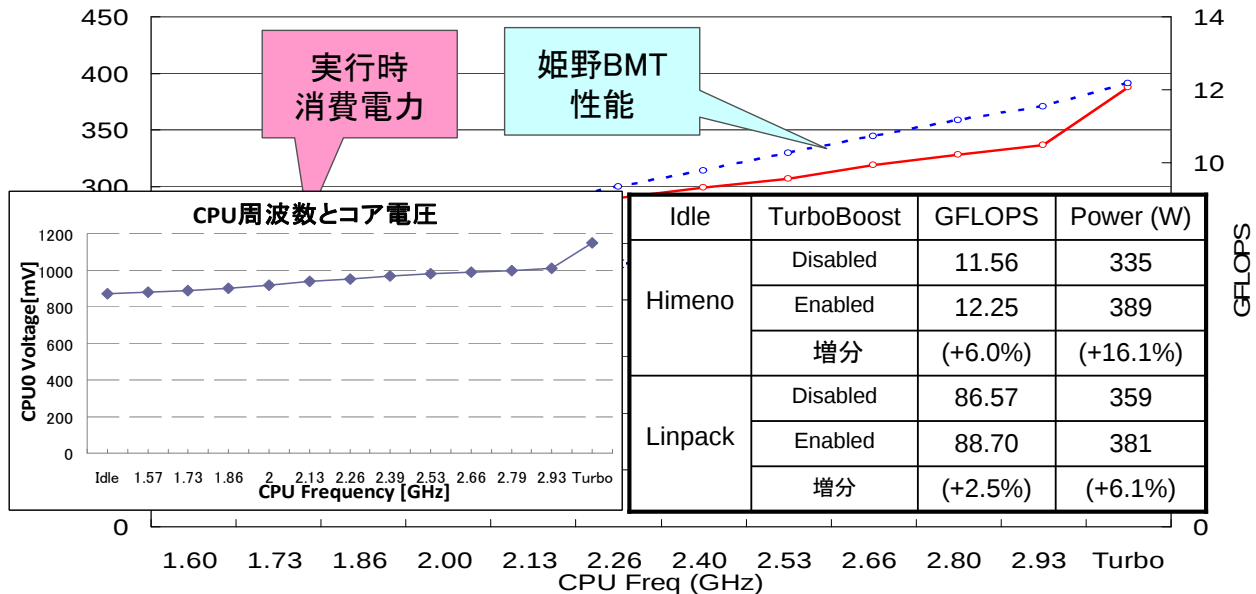
- CPUに熱的余裕があるときに、コアクロック周波数を自動で上げ性能を向上させるNehalem CPUの機能

- どのくらいよくなる？

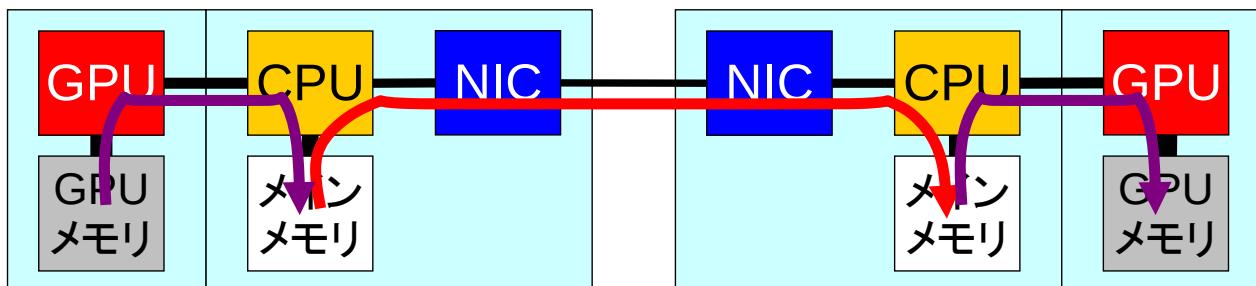
■ 性能は増加するが、電力も増加

- どのくらい増える？

- ターボモード時の実効周波数は不明
- 性能から見て、1ステップ分(3.06GHz相当)向上
- 電力は性能向上以上に増加、その理由は？
- Nehalemの低消費電力性は、アイドル時のクロック停止とパワーゲーティング(C6ステート)による(逆に言うとC6に入らないと電力はあまり減らない)

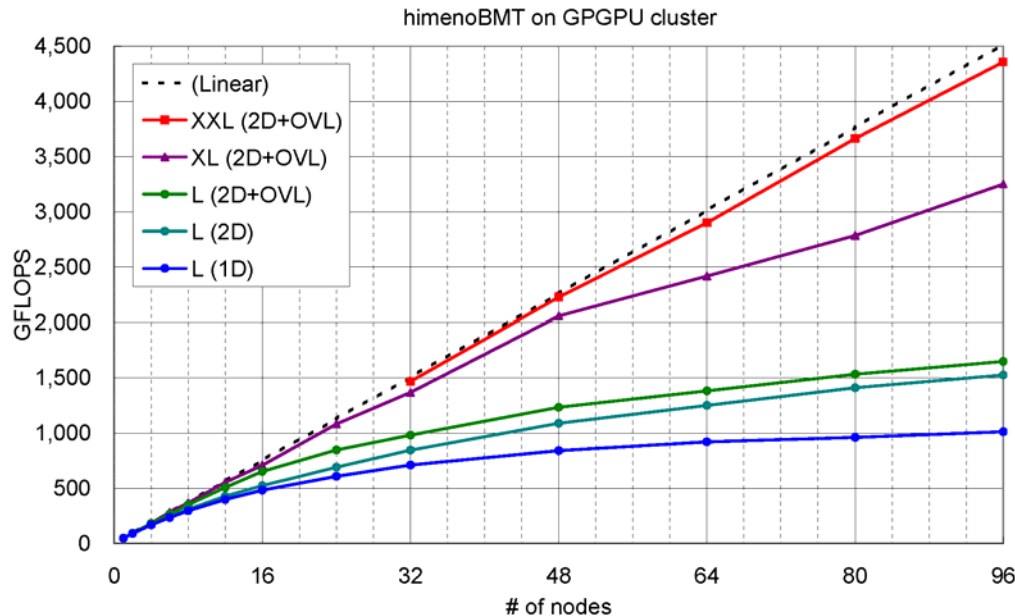


GPU搭載PCクラスターの課題

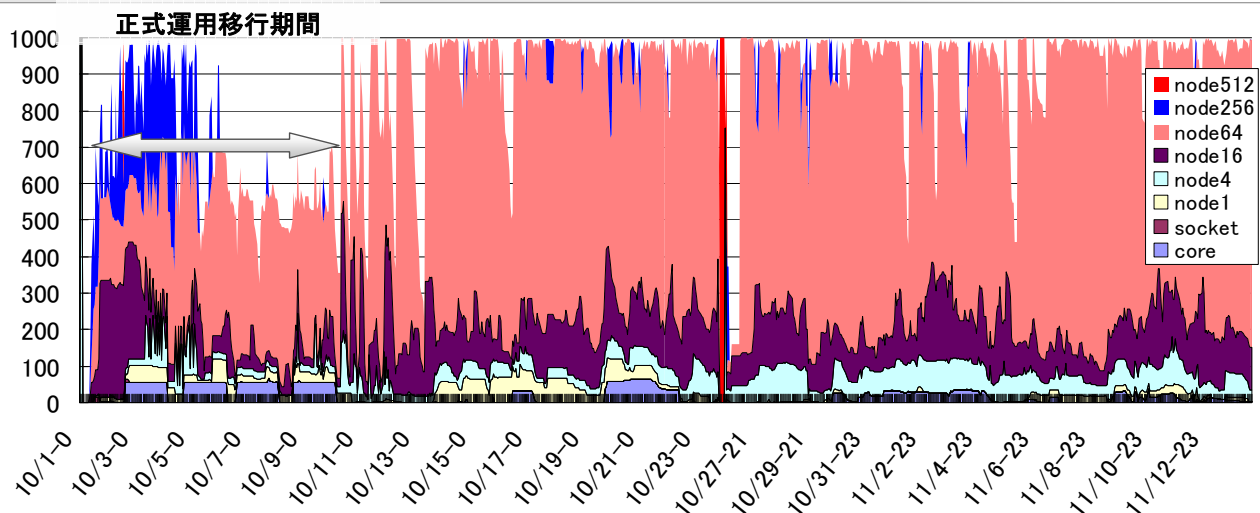


- CPUだけ利用するPCクラスターでのクラスター間データ転送
 - メインメモリ→(CPU)→NIC ⇒ NIC→(CPU)→メインメモリ
- GPUを使うPCクラスターでのクラスター間データ転送
 - GPUメモリ→(GPU)→(CPU)→メインメモリ→(CPU)→NIC ⇒ NIC→(CPU)→メインメモリ→(CPU)→(GPU)→GPUメモリ
- GPUクラスターでは、GPU演算性能は高い、が、レイテンシは長い、
 - ⇒ 計算時間 ≪ 通信時間
 - ⇒ Embarrassingly Parallelでない、と、並列効果が出にくい
- 通信と計算のオーバラップによる通信時間隠蔽も限界

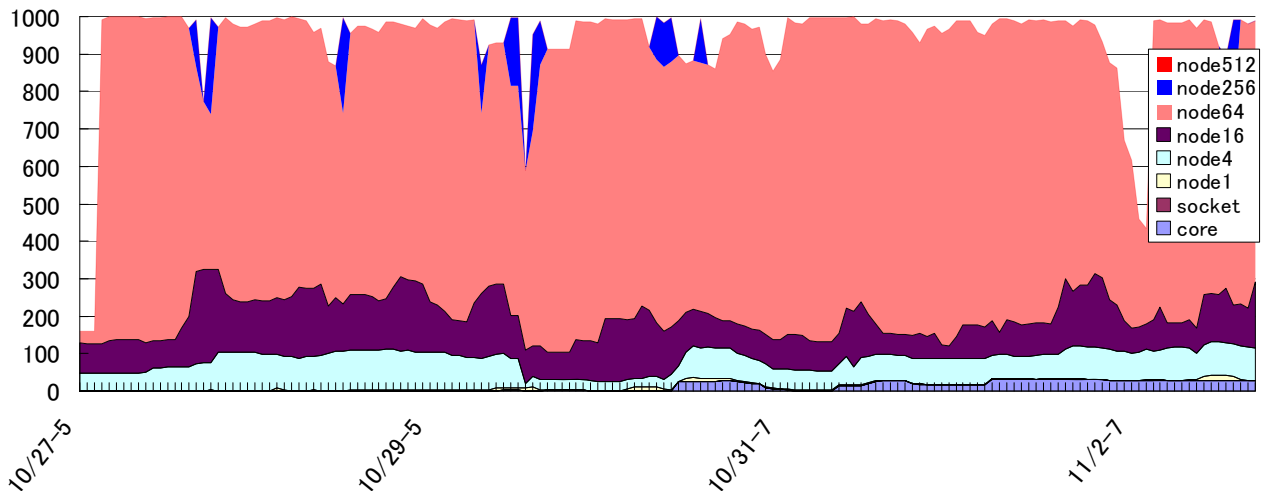
- himenoBMTのように、通信が多いプログラムでも、大きな問題(XXL)を解けば性能は出る



理研新システム(RICC)ジョブ実行率状況



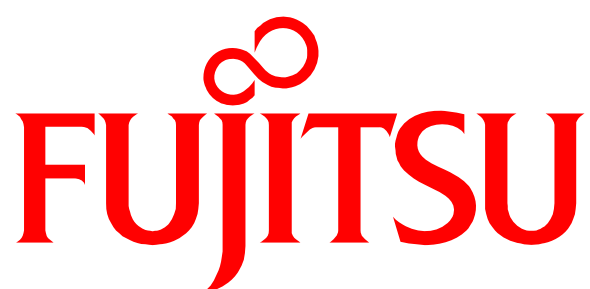
- 新システム(RICC)は2010/10/1に正式運用移行
- 全クラスタのノードを**メタジョブスケジューラ**で統一管理
 - ユーザ間フェアシェア、連成実行、順序実行、リソース指定(メモリ量、GPU,その他)
- 従来のメタジョブスケジューラを**マルチコア用に改良**
 - コア、ソケット、ノード単位と局所性を保ちながら、**空間+時間軸上で割付**、高効率に実行
 - 特定rankだけメモリ量を増大させる割り当ても可能
- 超大規模並列ジョブ(2000コア以上)実行前の稼働率低下防止が今後の課題
 - バックフィルできるジョブが十分には見つからない



- 新システム(RICC)は2010/10/1に正式運用移行
- 全クラスタのノードを**メタジョブスケジューラ**で統一管理
 - ユーザ間フェアシェア、連成実行、順序実行、リソース指定(メモリ量、GPU,その他)
- 従来のメタジョブスケジューラを**マルチコア用に改良**
 - コア、ソケット、ノード単位と局所性を保ちながら、**空間+時間軸上で割付**、高効率に実行
 - 特定rankだけメモリ量を増大させる割り当ても可能
- 超大規模並列ジョブ(2000コア以上)実行前の稼働率低下防止が今後の課題
 - バックフィルできるジョブが十分には見つからない

おわりに

- 今後、HPC計算は学術的計算、技術計算だけでなく、社会のあらゆる活動を支える基盤技術になると考えています
- 富士通は、BX900のようなコモディティであるシステム素材をベースに、コンパクトで高性能なHPCシステムの提供を行っていきます
- また、高性能演算ルーチンの開発、新たなGPUの利用技術の開発、構築運用管理の簡素化と高効率化を図る管理システムの提供を行ってゆきたいと思えます



THE POSSIBILITIES ARE INFINITE