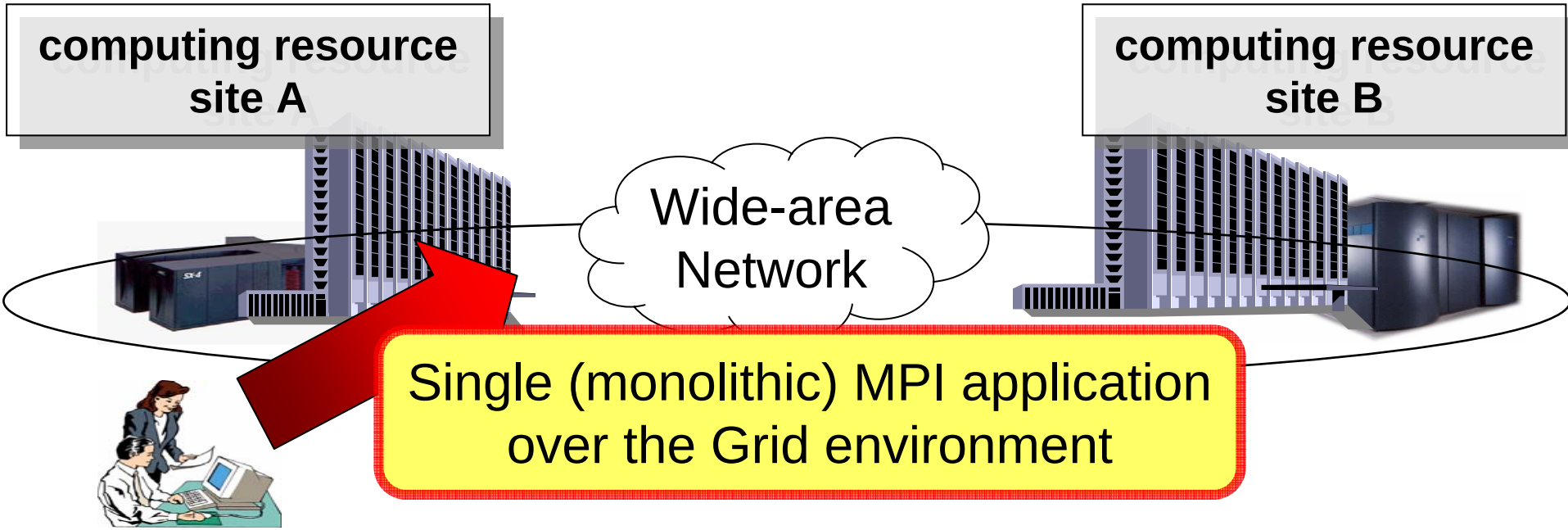
高性能計算のためのグリッド環境実現に向けてのフィジビリティスタディ

石川裕、松葉浩也

東京大学

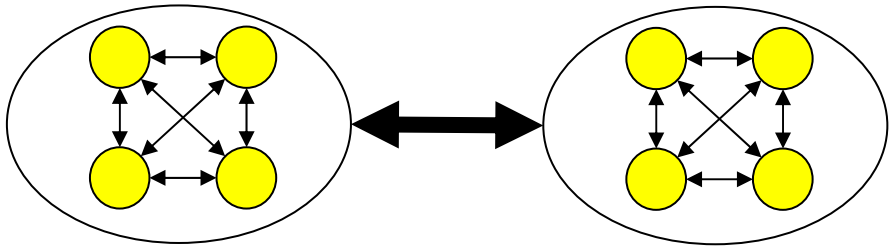
背景＆動機

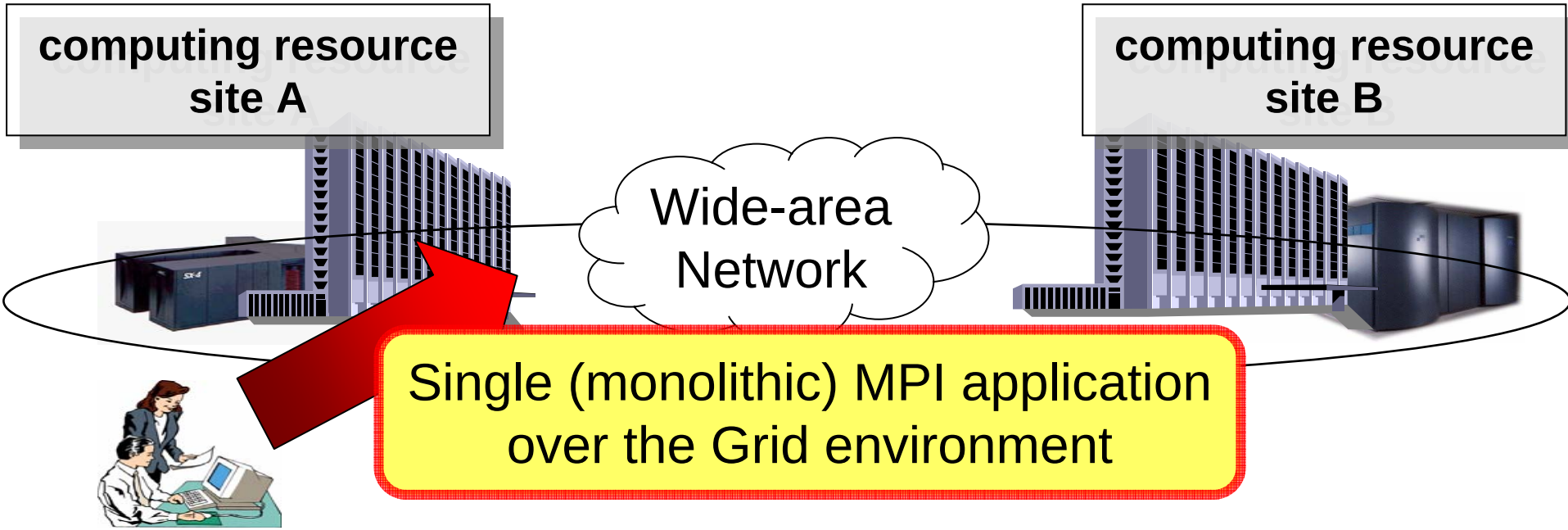
- 並列アプリケーションをグリッド環境で実行させたい
- 並列アプリケーションはMPI通信ライブラリを使ってプログラムされている
- 現状のMPI通信ライブラリを使ってもインターネット上で高性能通信を達成できない



背景＆動機

- インターネット環境
  - インターネットバンド幅　≥ クラスタにおけるネットワークバンド幅
    - 10 Gbps vs. 1 Gbps

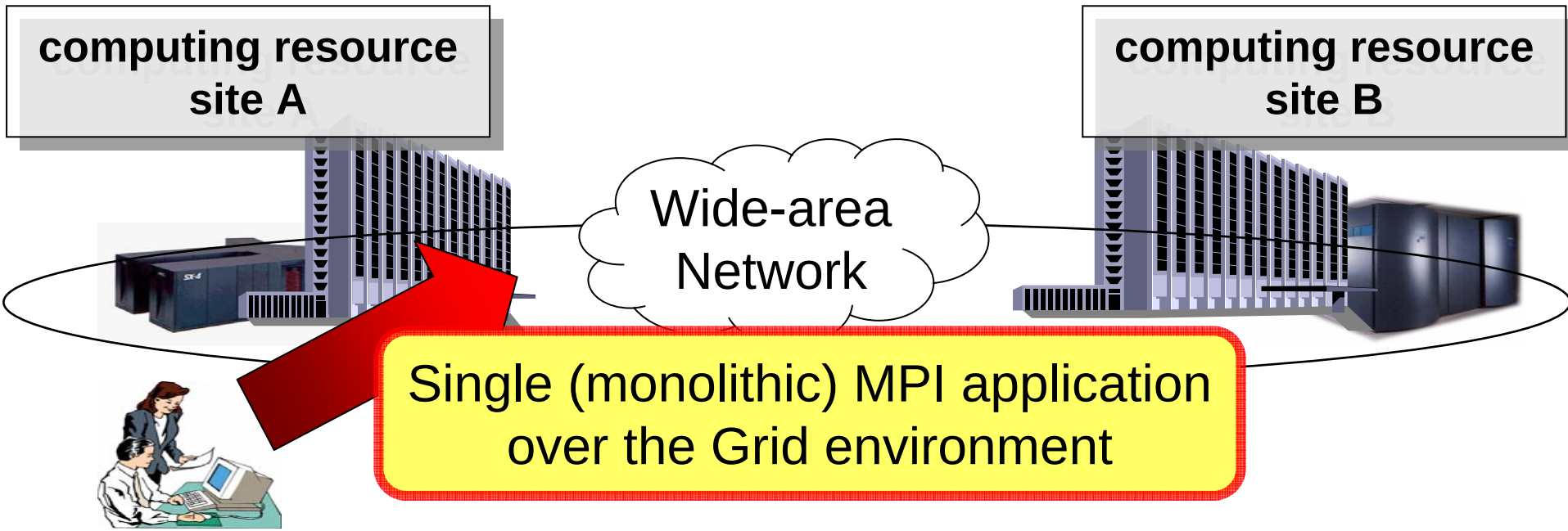




背景＆動機

- インターネット環境
  - インターネットバンド幅　≥ クラスタにおけるネットワークバンド幅
    - 10 Gbps vs. 1 Gbps

• Motohiko Matsuda, Yutaka Ishikawa, and Tomohiro Kudoh, ``Evaluation of MPI Implementations on Grid-connected Clusters using an Emulated WAN Environment," CCGRID2003, 2003

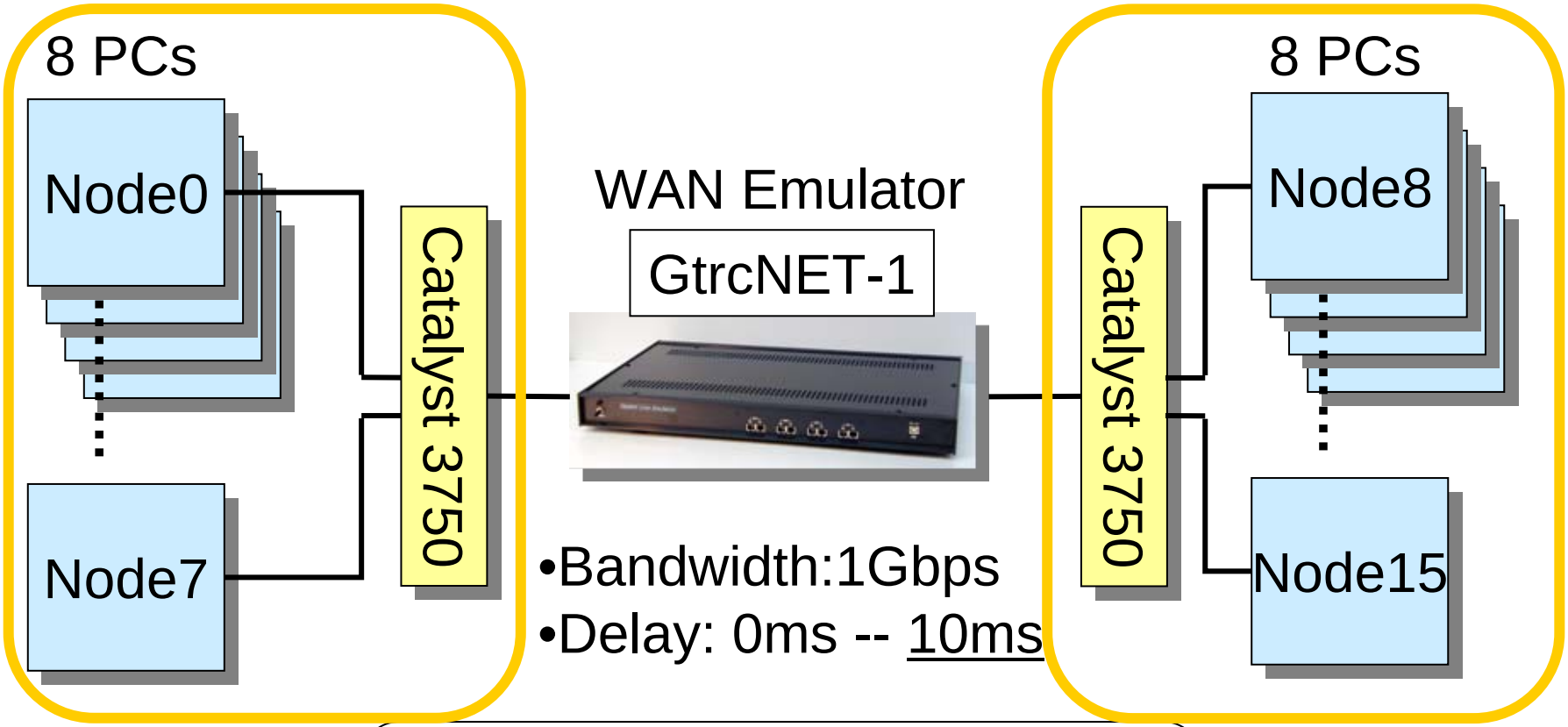


GridMPI

- グリッド上での通信バンド幅および遅延を考慮した高性能通信MPI
  - 文部科学省リーディングプロジェクト「超高速コンピュータ網形成プロジェクト（National Research Grid Initiative：通称NAREGI）」における産総研再委託において実施
- 課題
  - TCPプロトコル処理層
    - M Matsuda, T. Kudoh, Y. Kodama, R. Takano, and Y. Ishikawa, “TCP Adaptation for MPI on Long-and-Fat Networks,” IEEE Cluster 2005, 2005.
  - ソフトウェアによるペーシング
    - R. Takano, T. Kudoh, Y. Kodama, M. Matsuda, H. Tezuka, Y. Ishikawa, “Design and Evaluation of Precise Software Pacing Mechanisms for Fast Long-Distance Networks”, PDLFnet2005, 2005.
    - R. Takano, Y. Kodama, T. Kudoh, M. Matsuda, F. Okazaki, Y. Ishikawa, “Realtime Burstiness Measurement,” PDLFnet2006, 2006.
- ネットワークトポロジ＋通信バンド幅および遅延を考慮した通信アルゴリズム
  - 現在取り組み中

今までの取り組み

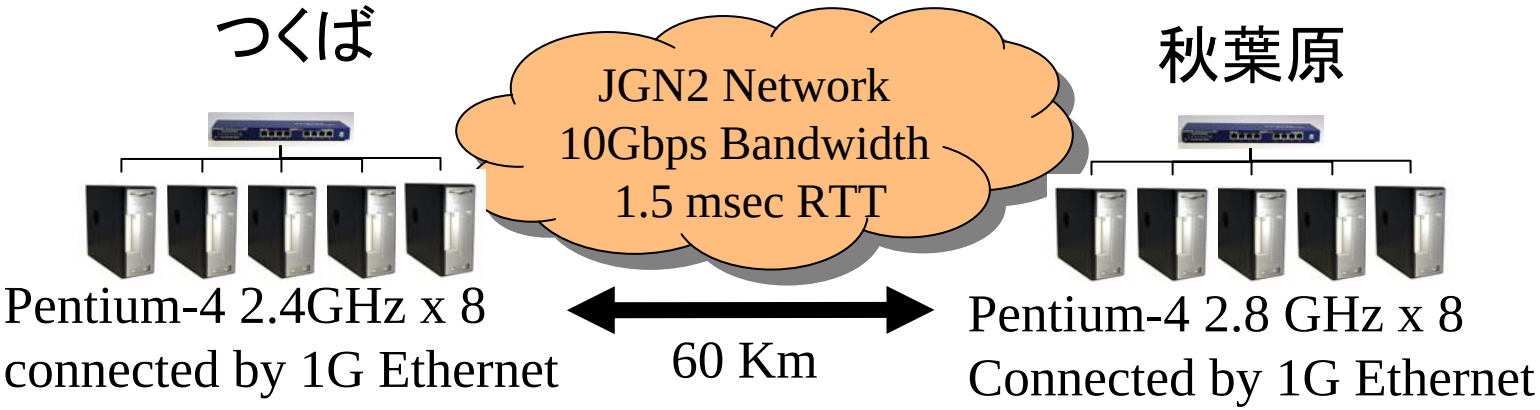
- WAN Emulatorによる性能解析



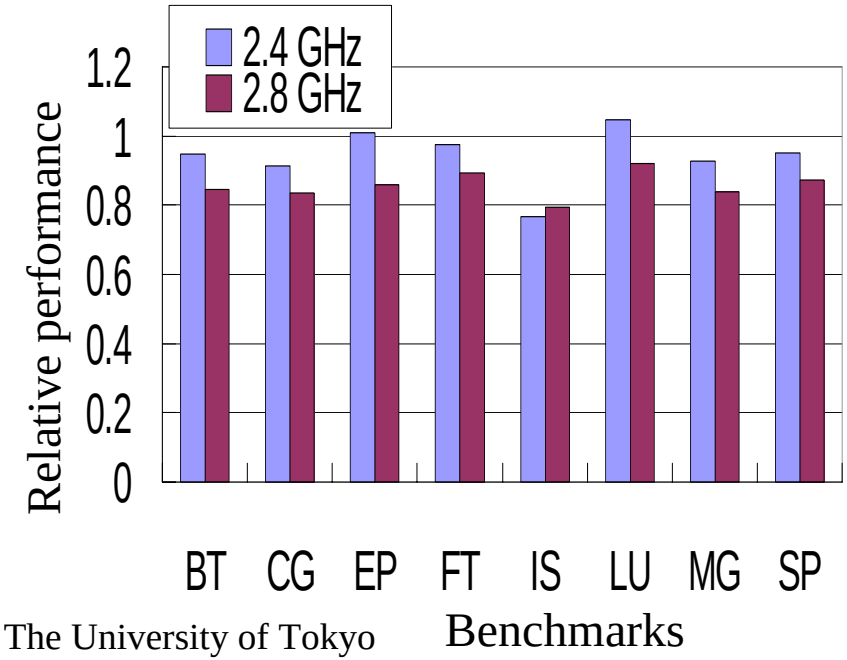
- CPU: Pentium4/2.4GHz, Memory: DDR400 512MB
- NIC: Intel PRO/1000 (82547EI)
- OS: Linux-2.6.9-1.6 (Fedora Core 2)
- Socket Buffer Size: 20MB

今までの取り組み

- 実環境での評価

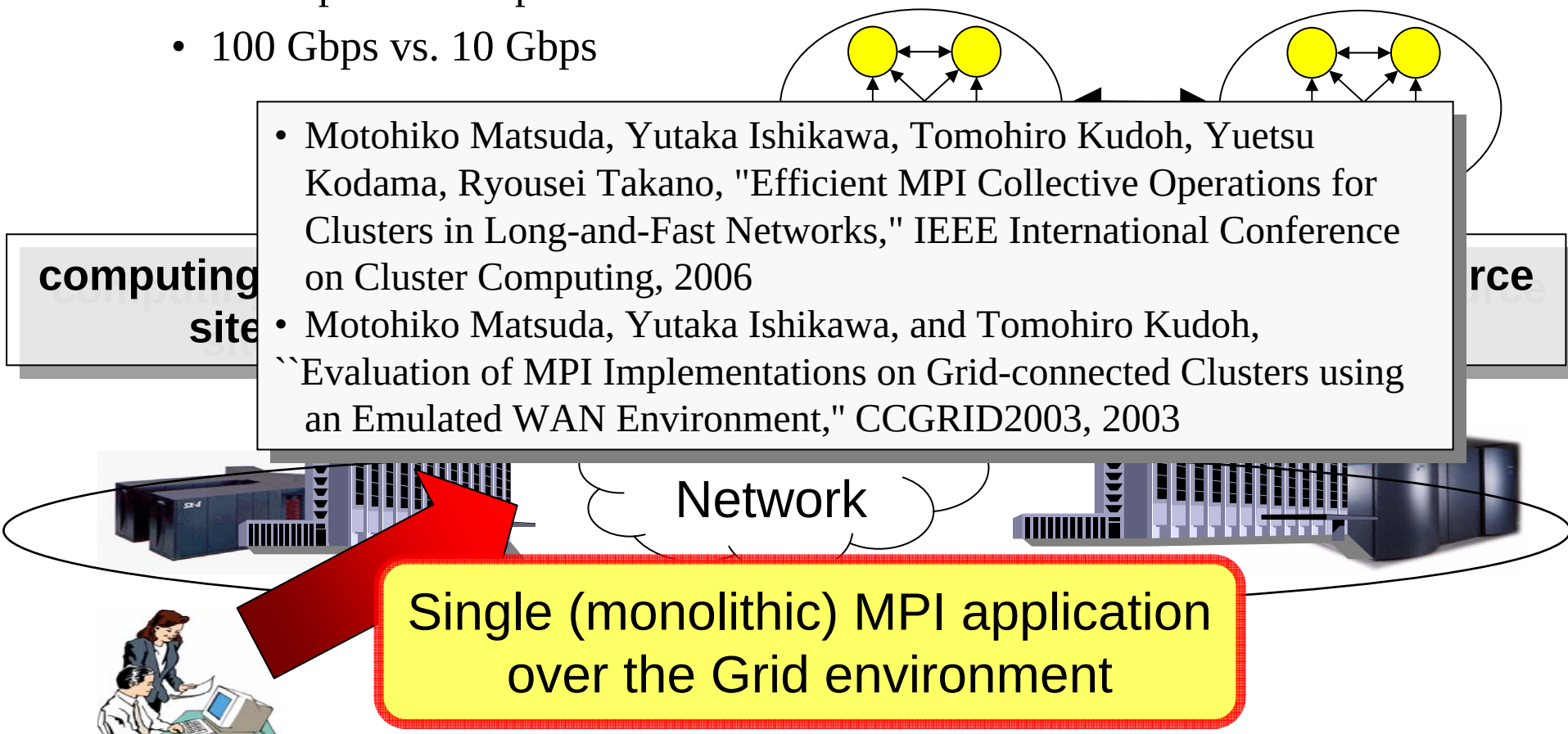


- NAS Parallel Benchmarks run using 8 node (2.4GHz) cluster at Tsukuba and 8 node (2.8GHz) cluster at Akihabara
  - 16 nodes
- Comparing the performance with
  - result using 16 node (2.4 GHz)
  - result using 16 node (2.8 GHz)



背景＆動機

- 将来のインターネット環境を想定したい
  - インターネットバンド幅　≥ クラスタにおけるネットワークバンド幅
    - 10 Gbps vs. 1 Gbps
    - 100 Gbps vs. 10 Gbps

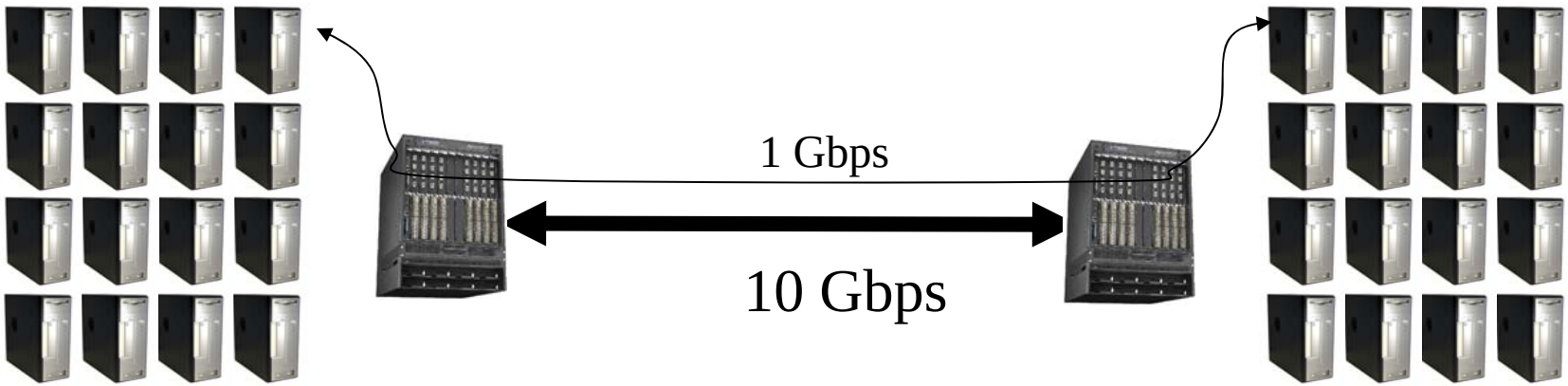


- Motohiko Matsuda, Yutaka Ishikawa, Tomohiro Kudoh, Yuetsu Kodama, Ryousei Takano, "Efficient MPI Collective Operations for Clusters in Long-and-Fast Networks," IEEE International Conference on Cluster Computing, 2006
- Motohiko Matsuda, Yutaka Ishikawa, and Tomohiro Kudoh, ``Evaluation of MPI Implementations on Grid-connected Clusters using an Emulated WAN Environment," CCGRID2003, 2003

Single (monolithic) MPI application over the Grid environment

従来の考え方

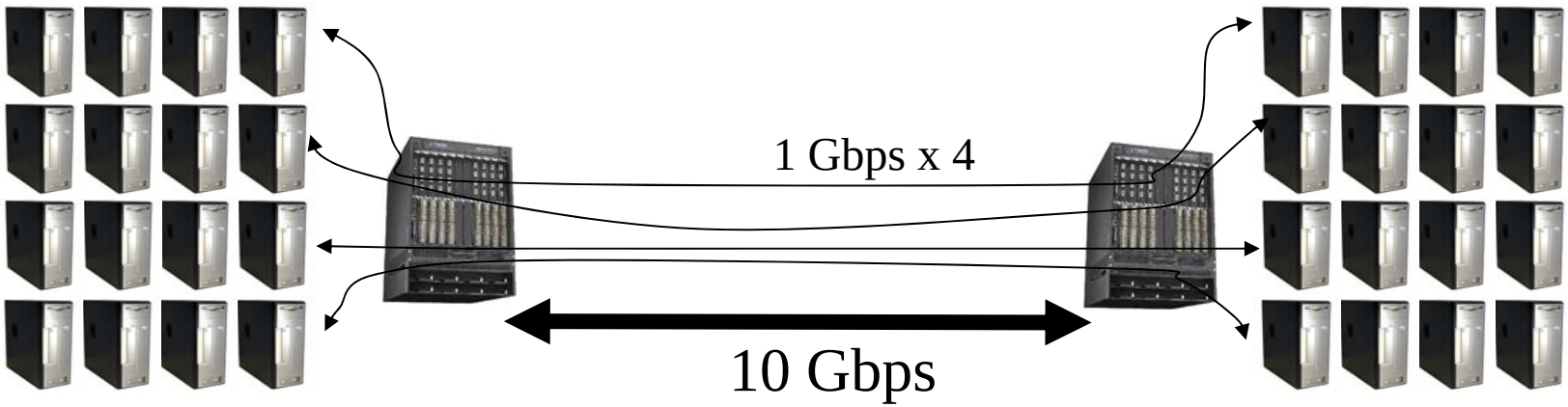
- インターネットは低帯域かつ通信遅延が大きいと仮定
  - MPI通信ライブラリが提供する集団通信（Barrier, Bcast, AlltoAll, Allreduce, …)の実装の仮定
  - 代表ノードによる通信
    - MPICH-G2、MagPIe、PACX-MPI



すなわち、いくら高帯域ネットワークがあっても
代表ノードから出て行くネットワークの性能しか
出せない。

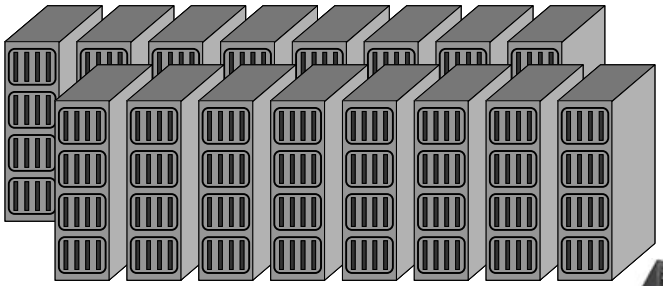
将来

- インターネットは高帯域、通信遅延に起因するホップ数は最小限と仮定
- 帯域を使うようなアルゴリズム開発
  - GridMPI



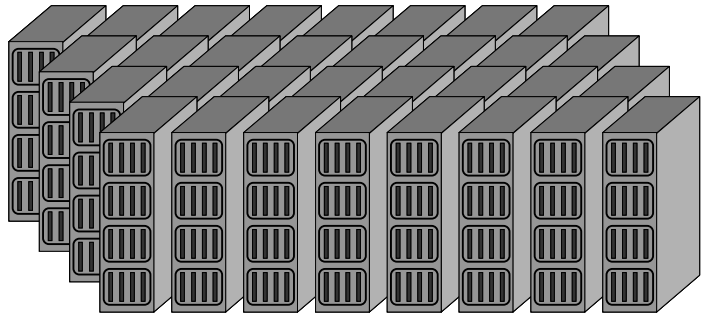
T2K Open Supercomputer

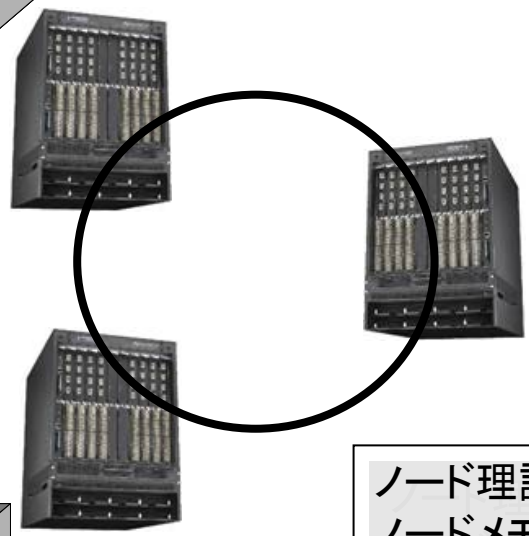
筑波大：80 TFlops～



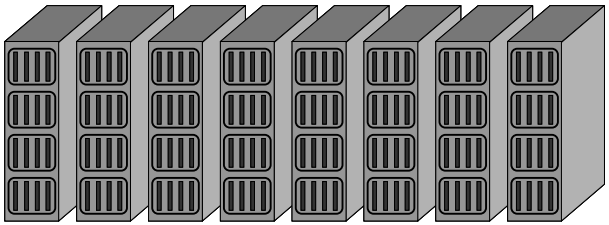
2008年3月下旬稼動開始

東大：150 TFlops～





京大：66 TFlops～



ノード理論演算性能: 160GFlops以上
ノードメモリ：32GB以上
ノードアーキテクチャ：64bit x86
ノード間ネットワーク：5GByte/sec以上
MPI通信性能：5GB/sec以上、6usec(往復遅延)

まとめ

- 10 Gbps & over 10 Gbpsを利用したグリッド環境テスト
- スパコン間をつなげたテスト

100 Gbpsの環境が欲しい