PCクラスタの概要

同志社大学 知識工学科
知的システムデザイン研究室
廣安 知之

PCクラスタ誕生の背景

- **コモディティハードウエアの性能向上**

  様々な分野において，PCが普及している

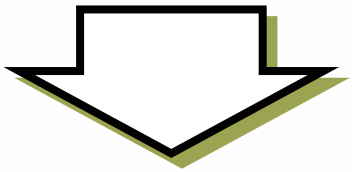
  ほとんどのPCはネットワークに接続されている

  - **飛躍的な性能向上**
  - **各種パーツの低価格化**
  - **高速ネットワークの開発と普及**

これらのPCをネットワークケーブルで結合して
使用すれば高性能な計算機になるのでは？

PCクラスタとは何か？

ぶどうなどの房、同種類のものの群れ



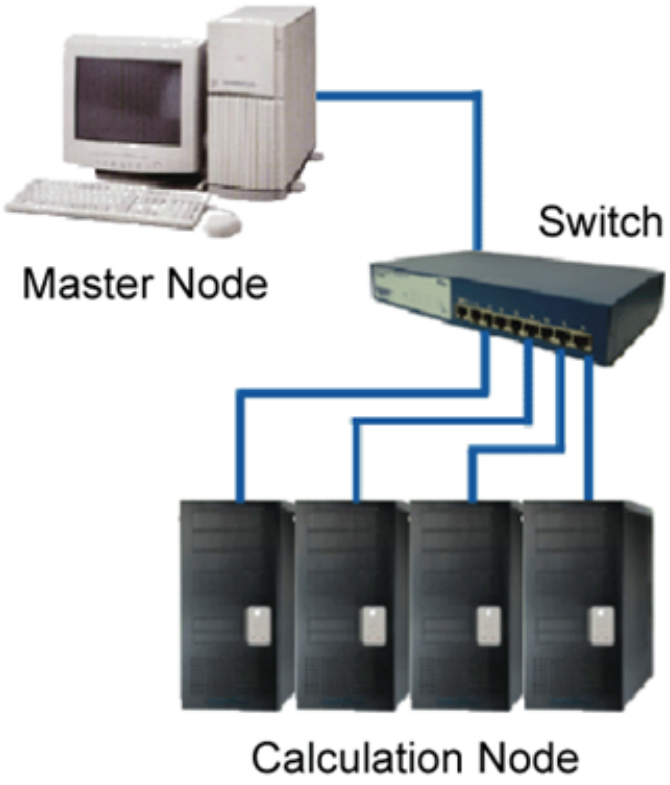
ネットワーク結合されたPC群

並行･並列･分散処理

PCクラスタ

- PCクラスタとは

  汎用のPCをネットワークで接続し，仮想的に1つの並列コンピュータとして利用できるシステム



- **最低限構築に必要な部品**
  - PC (**必要な台数**)
  - **スイッチングハブ**
  - **ネットワークケーブル**
  - OS (Linux**がよく用いられる**)
  - **ソフトウェア** (MPI, GNU**コンパイラ等**)

Beowulf クラスタ

- NASAのプロジェクト名
- 1998年末でほぼ終了
- 単一ホストからログインするクラスタ
- コモディティハードウエア
  - CPU
    - Intel
    - AMD
  - ネットワーク
    - Fast Ethernet
    - Myrinet
    - スイッチングハブ
- オープンソースソフトウエア
  - Linux
  - MPI
- コストの削減が可能（？）

Avalonクラスタ

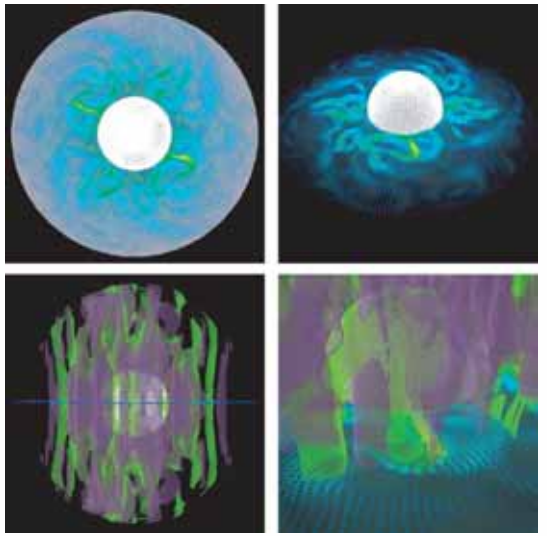
- Los Alamos National Laboratory
- **アルファ**(140)＋Myrinet
- **最初の**Top 500 **ランキング（**2000**年　３６４位)**
- Beowulf

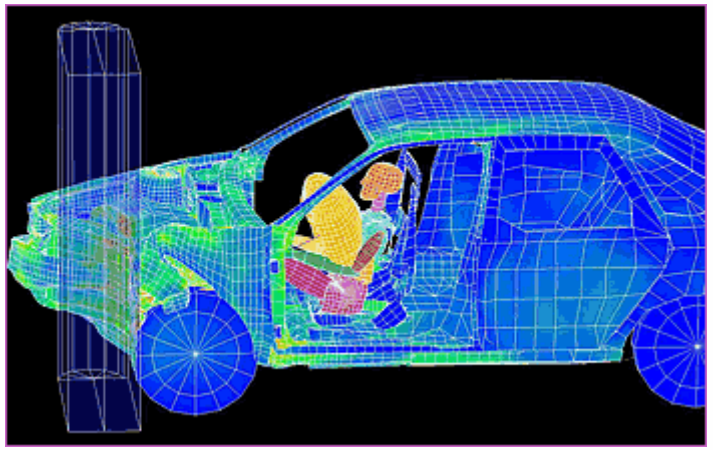
RWCPクラスタ

- 経済産業省リアルワールドコンピューティングプロジェクトを推進した技術研究組合 新情報処理開発機構
- 日本のクラスタのさきがけ
- Score, Open MP
- Myrinet
- PCクラスタコンソーシアム Cluster Consortium)

計算機の利用方法　1

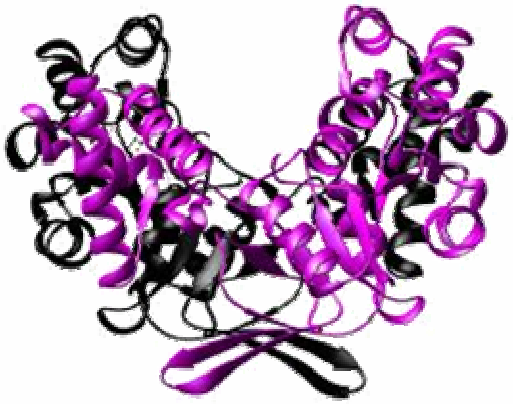
- 気象予測、地球環境予測、地質探査
  - 大気モデルと経験パラメータによる現象予測、地質調査による石油埋蔵予測
- 天体力学、素粒子、原子核物理
  - 銀河形成シミュレーション、素粒子・原子核物理論のシミュレーション
- 物質シミュレーション
  - 原子100個レベルの第一原理シミュレーションにより、物質の構造や特性を解析
- 構造解析、流体解析
  - 自動車衝突の仮想実験。自動車、航空機の空気抵抗の解析など産業応用が盛ん。

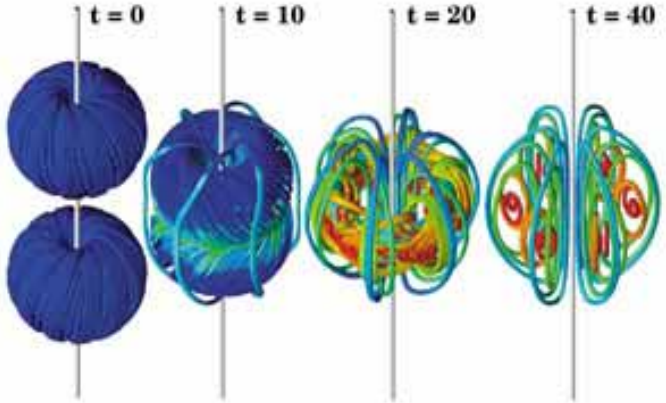




計算機の利用方法　2

- バイオインフォマティックス
  - 遺伝子解析、たんぱく質構造解析
- データマイニング
  - Web検索、顧客情報・売れ筋商品分析
- 経済予測・金融工学
  - マクロ経済予測、株価予測
- 核融合シミュレーション
  - 実験による知見と理論モデルに基づき核融合プラズマをシミュレーション





計算機の能力評価方法

- ベンチマークプログラム

  計算機の処理速度を計測するソフトウェア
  客観的な数値としてハードウェア性能の指標になる

- コンピュータの演算性能

  FLOPS (フロップス) という単位で表される
  一秒間に何回浮動小数点演算が行えるかの値

  世界で一番高速な計算機は
  約70TFlops (テラフロップス)
  Tは1兆という単位なので,
  一秒間に70兆回計算できる

Top500 SuperComputer Site

- Top500**とは**

世界のコンピュータのランキング http://www.top500.org/

順位	コンピュータ名	CPU数	性能
1	Blue Gene/L	32768	70.72 TFlops
2	Columbia	10160	51.87 TFlops
3	**地球シミュレータ**	5120	35.86 TFlops
4	MareNostrum	3564	20.53 TFlops
5	Thunder	4096	19.94 TFlops
6	ASCI Q	8192	13.88 TFlops
7	System X	2200	12.25 TFlops
8	BlueGene/L	8192	11.68 TFlops
9	eServer pSeries 655	2944	10.31 TFlops
10	Tungsten	2500	9.819 TFlops

Top500 SuperComputer Site

- Top500**とは**

世界のコンピュータのランキング http://www.top500.org/

順位	コンピュータ名	CPU数	性能
1	Blue Gene/L	32768	70.72 TFlops
2	Columbia	10160	51.87 TFlops
3	地球シミュレータ	5120	35.86 TFlops
4	MareNostrum	3564	20.53 TFlops
5	Thunder	4096	19.94 TFlops
6	ASCI Q	8192	13.88 TFlops
7	System X	2200	12.25 TFlops
8	BlueGene/L	8192	11.68 TFlops
9	eServer pSeries 655	2944	10.31 TFlops
10	Tungsten	2500	9.819 TFlops

世界のPCクラスタ

- **米バージニア州立工科大学** - System X -

System	Apple XServe G5 2.3 GHz
Processors	2,200
Network	InfiniBand + Gigabitether
Performance	12,25TFlops (7**位**)

Apple Computer**の**1U**サーバ**Xserve G5 **をクラスタ化したもの**

導入コストは地球シミュレータが
2**億**5**千万ドルに対し，**
初期System X**はわずか**520**万ドル**

日本のPCクラスタ

- RIKEN Super Combined Cluster (**理化学研究所**)

System	Intel Xeon 3.06GHz
Processors	2,048
Network	InfiniBand + Myrinet
Performance	8,029TFlops (14**位**)

- AIST Super Cluster P-32 (**産業技術総合研究所**)

System	AMD Opteron 2.0 GHz
Processors	2,200
Network	Myrinet
Performance	6,115TFlops (28**位**)

同志社大学のPCクラスタ（1/2）

- Supernova Cluster

System	AMD Opteron 1.8 GHz
Processors	512
Memory	PC2700 Registered ECC 2GB
OS	Turbolinux 8 for AMD64
Network	Gigabit Ethernet

1.169 TFlops 達成
世界：93位，国内：6位
PCクラスタ：1位（2003/11 時点）

同志社大学のPCクラスタ（2/2）

- Core Cluster

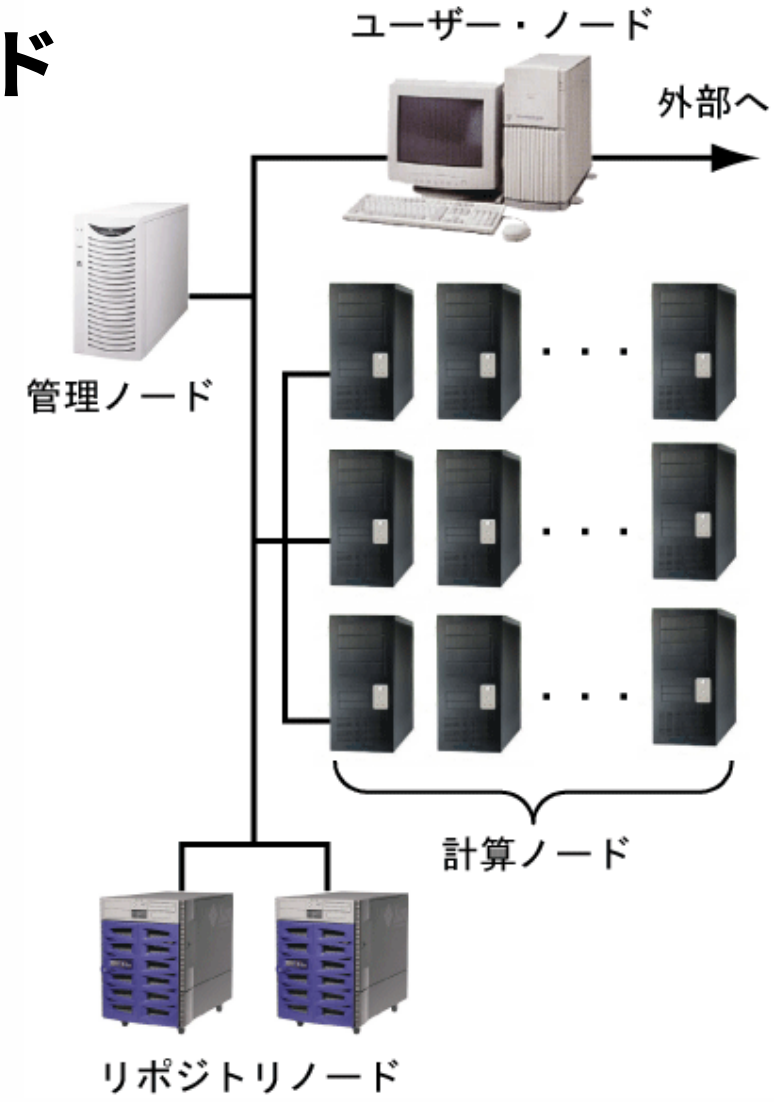
System	IBM PowerPC 970 1.6 GHz
Processors	252
Memory	ECC DDR SDRAM　2.5 GB
OS	SuSE Linux Enterprise Server
Network	Myrinet

1.009 TFlops 達成
（2004年11月）
世界392位，日本22位
PCクラスタ7位

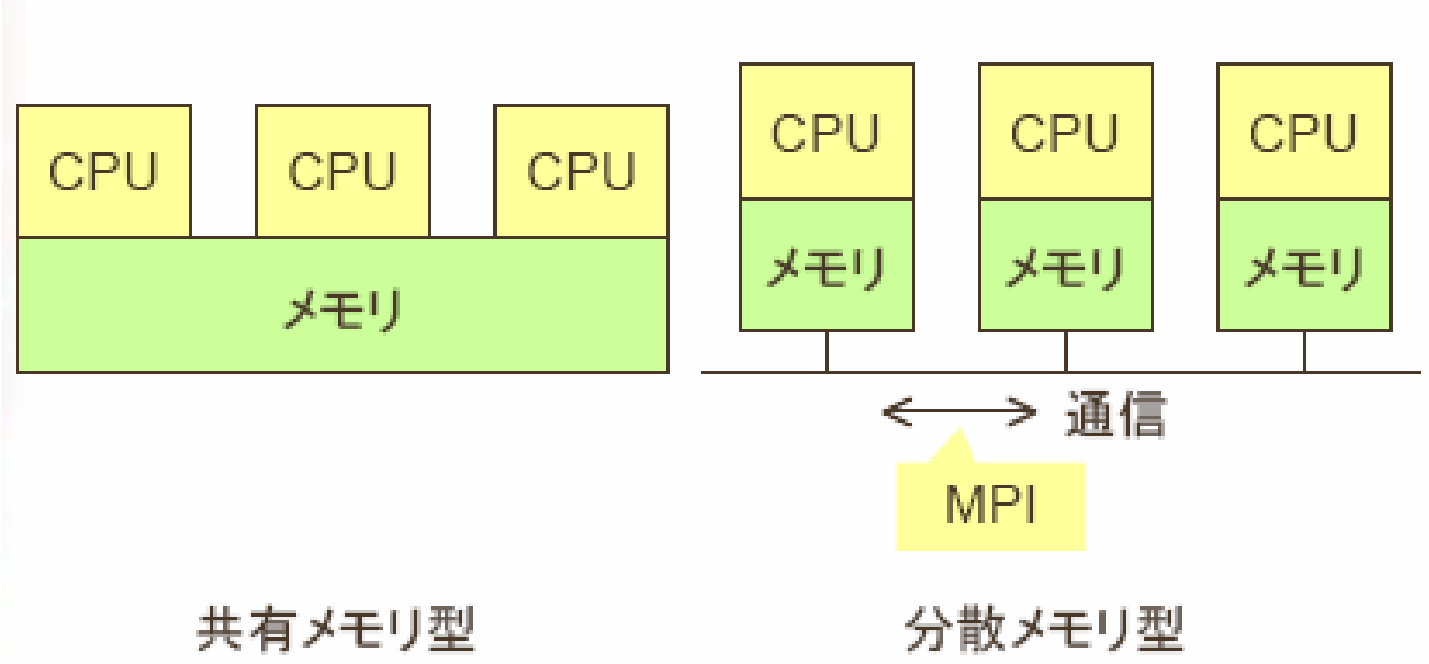
大規模PCクラスタの構成図

- PCクラスタを構成するノード
  - ユーザー・ノード
  - 管理ノード
  - 計算ノード (最も数が多い)
  - リポジトリノード

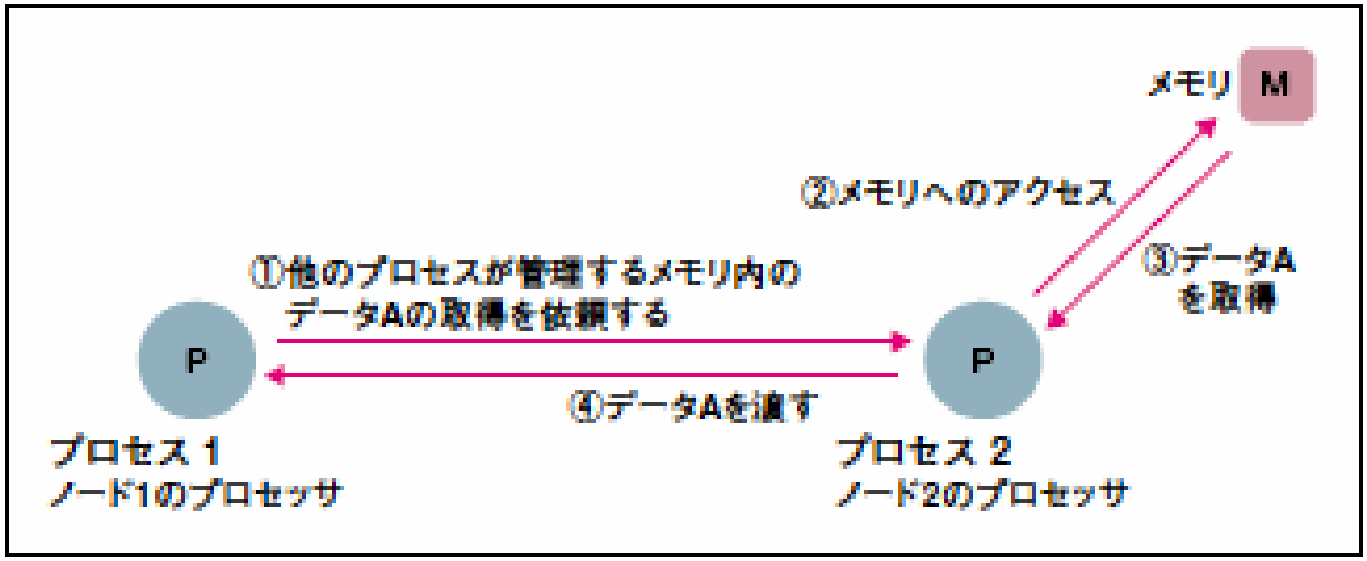
小規模な場合は1台のノードが,
複数のサービスを提供し,
構成を簡略化する場合がある



共有メモリ型と分散メモリ型



メッセージパッシング


メモリ M
②メモリへのアクセス
③データA
を取得
①他のプロセスが管理するメモリ内の
データAの取得を依頼する
P
P
④データAを渡す
プロセス 1
ノード1のプロセッサ
プロセス 2
ノード2のプロセッサ

メッセージ通信ライブラリ

- 並列アプリケーションで通信を行うソフトウェア
  - ネットワークに依存しない通信がしたい
    - いちいちソケットを使って書くのは大変
    - ネットワークプロトコルがTCP/IPとは限らない
- かつては各社が独自に提供
  - IBM: MPL
- 可搬性のあるプログラムを書きたい
- 標準規格が欲しい
  - MPI Forum: MPI

MPI

- Message Passing Interface
- 1992**年** MPI Forum
  - **並列計算機ベンダーと学術組織**
- **仕様のみ策定。実装は各ベンダーに任せる**
  - **ベンダーがそれぞれの計算機固有の方式を用いて実装**
- MPICH, LAM
- 1997**年** MPI-2
  動的プロセス生成、並列I/O**などの仕様を追加**

MPIの実装

- MPICH
  - アルゴンヌ国立研究所・ミシシッピ州立大
  - 現在の主流
- LAM
  - オハイオ　スーパーコンピューティングセンター
- CHIMP
  - エジンバラ並列計算センター

MPIプログラミング

- SPMD (Single Program Multiple Data)モデル
- **基本的には6つの関数で書ける**
  - MPI_Init()
  - MPI_Comm_size()
  - MPI_Comm_rank()
  - MPI_Send()
  - MPI_Recv()
  - MPI_Finalize()

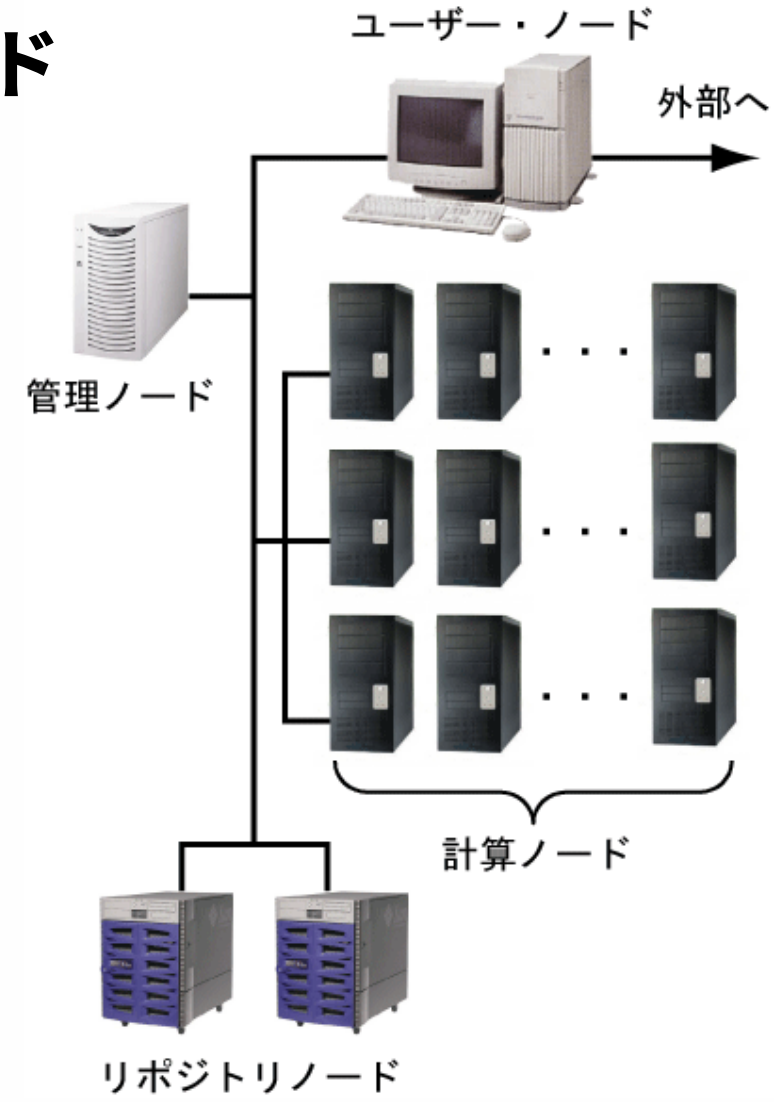
サンプルMPIプログラム

```
#include <mpi.h>

int main(int argc, char *argv[])
{
    int rank;
    MPI_Status status;
    char buf[256];
    char data[] = "Hello";

    MPI_Init(&argc, &argv);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);

    if (rank == 0) {
        MPI_Send(data, 5, MPI_CHAR, 1, 0, MPI_COMM_WORLD);
    } else if (rank == 1) {
        MPI_Recv(buf, 5, MPI_CHAR, 0, 0, MPI_COMM_WORLD, &status);
        printf("%s¥n", buf);
    }
    MPI_Finalize();
}
```

大規模PCクラスタの構成図

- PCクラスタを構成するノード
  - ユーザー・ノード
  - 管理ノード
  - 計算ノード (最も数が多い)
  - リポジトリノード

小規模な場合は1台のノードが,
複数のサービスを提供し,
構成を簡略化する場合がある



PCクラスタを設置するときの問題点

- **電源の確保**

  1ノード当たり500Wとすると1000ノードで5000A必要
  導入前に電源の工事が必要

- **空調の整備**

  適切に熱処理を行わないと，故障の原因になる

- **PCクラスタの維持費**

  電気代，故障ハードウェアの取替，メンテナンス費
  予算に組み込めない？

- **ノード数に比例した構築，管理コストが発生**

  すべてのマシンにOSとソフトウェアの設定を行う

無料でPCクラスタを作る

- 100ノードのPCクラスタを作る場合

  1ノード20万円とすると2000万円かかる・・

  無料で手に入らないか？

- 大学のPCを利用する

  昼は学生が使用しているから，夜のみ研究に利用

  広島大学が遊休PCを利用したキャンパスグリッドを構築 (2004年11月)

  - 医療画像の利用に向けた大量の計算処理
  - 地域の製造業に対するCAEの計算資源として提供

KNOPPIXクラスタ（1/2）

- KNOPPIX
  - CDのみでブート可能な Linuxディストリビューション
  - ドイツのKnopper氏によって開発
  - Debian GNU/Linuxがベース

  カスタマイズが容易であるため，様々な構成のKNOPPIXが存在する

  日本語版は産業技術総合研究所が開発

須崎 有康 (産業技術総合研究所 情報技術研究部門)

KNOPPIXクラスタ（2/2）

- KNOPPIXクラスタ

  KNOPPIXを用いてハードディスクにOSやソフトウェアをインストールすることなくPCクラスタが構築する

  【事例紹介】
  - 広島国泰寺高校科学部物理班
  - 柴田 良一（岐阜工業高等専門学校 建築学科）
  - 小西 史一（理化学研究所 ゲノム科学総合研究センター）

  PCクラスタをハードディスクにインストールして構築したいけど，マスタノードやソフトウェアのインストールが面倒

  - 中尾 昌広（同志社大学工学部知識工学科 修士2年）