Parallel Computing

Katsuhiro Yamazaki

March 29th, 2007

1. What is parallel computing
2. Research projects
3. Research fields
4. Typical parallel machines
5. Earth simulator (with Prof. Shigeru Oyanagi)
6. Parallel programming
7. Parallel algorithms
8. Application fields
9. Grid computing (with Prof. Shigeru Oyanagi)
10. Summary

並列コンピューティング
山崎　勝弘
2007年3月29日

1. 並列コンピューティングとは
2. 研究プロジェクト
3. 研究分野
4. 代表的な並列マシン
5. 地球シミュレータ (小柳　滋教授と共同)
6. 並列プログラミング
7. 並列アルゴリズム
8. 応用分野
9. グリッドコンピューティング (小柳　滋教授と共同)
10. まとめ

1. Parallel Computing

- Solves large scale problems using parallel machines & supercomputers within a reasonable time
- Weather prediction, environment, aircrafts, biomedicine, human DNA, protein structure, education
- Same as high performance computing
- Massively parallel & highly parallel machines
- Execution time= computation time + communication time
- Load balancing & reduction of communication overheads

1. 並列コンピューティング

- 大規模な問題を並列マシンやスーパーコンピュータで高速処理して問題解決を図ること。
- 気象、環境、航空、医療、化学、遺伝子、蛋白質の構造、教育など
- 高性能コンピューティングと同じ
- 超並列･高並列マシン
- 処理時間= 演算時間 + 通信時間
- 負荷均衡、通信オーバーヘッドの軽減

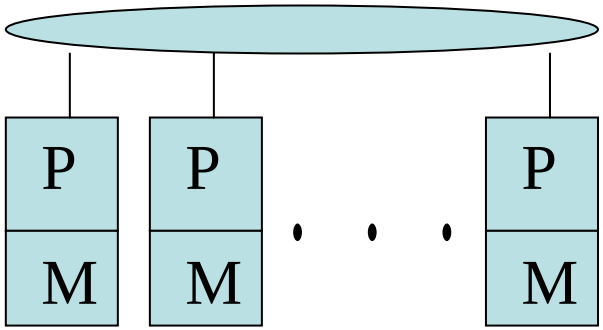
Outline of Parallel Computing

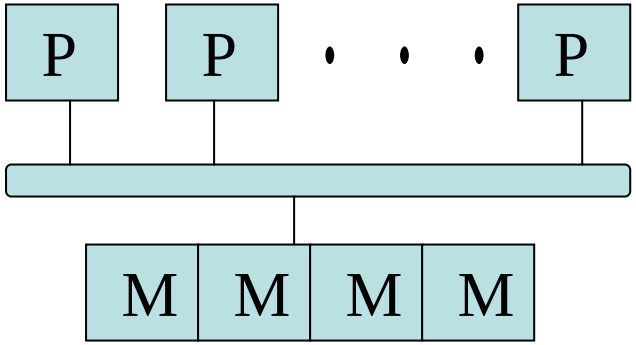
Applications with mass data & mass calculations

Parallel programming using MPI or OpenMP

Algorithms: How to solve problems

Parallel software: compiler, debugger, evaluator, OS





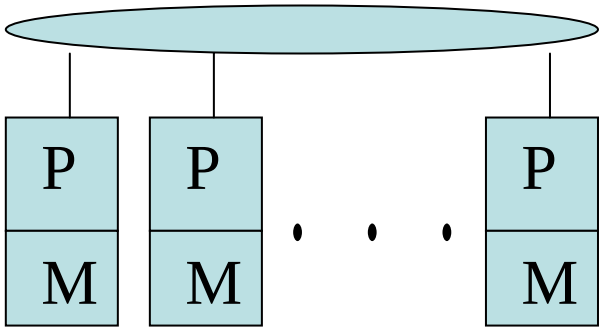
並列コンピューティングの概要

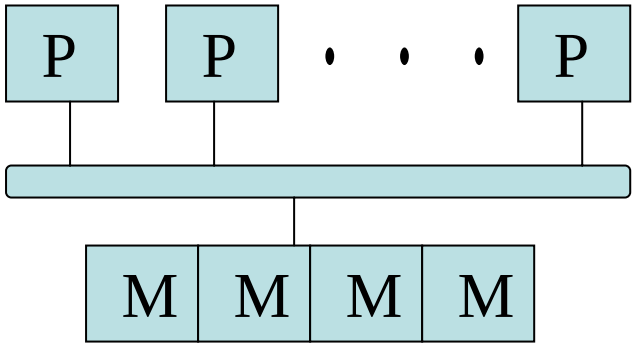
大量のデータと大量の計算を持つ応用

MPIまたはOpenMPを使った並列プログラミング

アルゴリズム: どのように問題を解決するか

並列処理ソフトウェア: コンパイラ、デバッガ、エバリュエータ、OS





Monumental Works

- ASCI Red (Intel) achieved 1Teraflops in 1996
  - 9216 200M PenPros, 580GB memory, 800MB/sec, 1.8Teraflops(peak)
- DeepBlue won Kasparov, the chess champion, in 1997
  - Started 1985, 50000 moves/sec
  - 200000000 moves/sec in 1997
  - IBM RS/6000 SP2
  - 32 Power2 super chips, 512 coprocessors
- Earth Simulator achieved 35.86Teraflops in 2002
  - World's fastest at the top500 supercomputers in June 2002.

画期的な業績

- ASCI Red (インテル) 1Teraflopsを達成、1996年
  - 200M PenPro 9216個, 580GBメモリ, 800MB/sec, 1.8Teraflops(ピーク)
- DeepBlueが チェスチャンピオンKasparovに勝利、1997年
  - 1985年開始, 50000手/sec
  - 200000000手/sec,1997年
  - IBM RS/6000 SP2
  - Power2 super chip 32個, 512個のコプロセッサ
- 地球シミュレータ 35.86Teraflopsを達成、2002年
  - TOP500スーパーコンピュータで世界一、2002年6月

Recent Trends

- Distributed shared memory
- Multi-threading: thread-level parallelism
- FPGA(Field Programmable Gate Array)
- PC cluster, SMP cluster
- Hybrid parallel programming
- Grid computing
- Multi-core processor

最近の動向

- 分散共有メモリ
- マルチスレッド: スレッドレベル並列処理
- FPGA(Field Programmable Gate Array)
- PCクラスタ、SMPクラスタ
- ハイブリッド並列プログラミング
- グリッドコンピューティング
- マルチコアプロセッサ

2. HPCC Project

- High performance computing act 1991
- Grand Challenges: High Performance Computing & Communications(1992,1993)
- Toward a National Information Infrastructure(1994)
- Technology for the National Information Infrastructure(1995)
- Foundation for America's Information Future(1996)
- Advancing the Frontiers of Information Technology(1997)
- Technologies for the 21st Century(1998)

2. HPCCプロジェクト

- HPC法 1991
- グランドチャレンジ：高性能コンピューティングと通信(1992,1993)
- 国家情報基盤に向けて(1994)
- 国家情報基盤のための技術(1995)
- アメリカの情報未来の基盤(1996)
- 情報技術の最先端の前進(1997)
- 21世紀の技術(1998)

- Networked Computing for the 21st Century(1999)
- Information Technology for the 21st Century: A Bold Investment in America's Future(2000)
- Information Technology: The 21st Century Revolution(2001)
- Networking and Information Technology Research and Development(2002)
- Strengthening National, Homeland, and Economic Security(2003)
- Advanced Foundations for American Innovation(2004)
- Networking and Information Technology Research and Development(2005-2007)

- 21世紀のネットワークコンピューティング(1999)
- 21世紀の情報技術：アメリカの未来に思い切った投資(2000)
- 情報技術：21世紀の革命(2001)
- ネットワークと情報技術の研究と開発(2002)
- 国家の強化と経済の安定(2003)
- アメリカの革新のための先進基礎 (2004)
- ネットワーキングとIT　研究開発 (2005-2007)

Real World Computing Project

- 1992~2002, MITI, 700 million dollars
- Flexible information processing
- Development of PC clusters
  - 1st: 1996 Pen 160M 32PCs
  - 2nd: 1998 PenPro 200M 128PCs, Myrinet
  - 3rd: 1999 Pen**III** 16PCs + Alpha21264 16PCs
  - 4th: Dual Pen**III** 933M 512PCs, Myrinet, Gigabit Ethernet
- PAPIA(Parallel Protein Information Analysis) system

リアルワールドコンピューティング（RWC）プロジェクト

- 1992年~2002年、通産省、700億円
- 柔らかな情報処理
- PCクラスタの構築
  - 1号機: 1996年 Pen 160M 32台
  - 2号機: 1998年 PenPro 200M 128台, Myrinet
  - 3号機: 1999年 PenIII 16台 + Alpha21264 16台
  - 4号機: 2001年 Dual PenIII 933M 512台, Myrinet, Gigabit Ethernet
- 並列タンパク質情報解析(PAPIA) システム

3. Research Fields of Parallel Computing

- Applications
  - Weather prediction, fluid dynamics, digital image processing, data mining, n-body problems
- Algorithms
  - Divide & conquer, processor farms, process networks, iterative transformation
- Software
  - Parallel programming languages & environment, parallelizing compiler, parallel OS
- Hardware
  - Processor, memory, I/O, interconnection networks

3. 並列コンピューティングの研究分野

- 応用
  - 気象予測、流体計算、デジタル映像処理、データマイニング、多体システム
- アルゴリズム
  - 分割統治法、プロセッサファーム、プロセスネットワーク、繰り返し変換
- ソフトウェア
  - 並列プログラミング言語・環境、並列化コンパイラ、並列OS
- ハードウェア
  - プロセッサ、メモリ、I/O、相互結合網

4. Typical Parallel Machines

- ILLIAC **IV**, Univ. of Illinois
  - Pioneer of parallel machines, 1972
  - 64(PE+PEM), mesh, SIMD
  - Semiconductor memory, ECL, multilayer printed circuit board
  - Matrix calculations, differential equations
- Connection machine, Thinking machines corp
  - 1985, CM-1, 65536 1bit PEs
  - Binary 12 dimension hypercube, 16PEs×4096
  - 1987, CM-2, FPU
  - 1992, CM-5, 1024 SPARC chips

4. 代表的な並列マシン

- ILLIAC **IV**, イリノイ大
  - 並列マシンのパイオニア, 1972年
  - 64(PE+PEM),メッシュ, SIMD
  - 半導体メモリ, ECL, 多層プリント回路基板
  - 行列演算, 微分方程式
- コネクション・マシン, 米シンキングマシンズ社
  - 1985年, CM-1, 65536個の1ビットPE
  - 2進12 次元ハイパーキューブ, 16PE×4096
  - 1987年, CM-2, FPU追加
  - 1992年, CM-5, 1024個のSPARC

- Transputer, Inmos corp.
  - 1979, Inmos corp. was built by British government
  - Processor, memory, link on a chip
  - Building blocks of parallel machines
  - 1985, T414, integer
  - 1988, T800, floating
  - 1994, T9000, virtual channels, superscalar
- AP-1000, Fujitsu
  - 1991, 16～1024 cells, SPARC
  - Torus+broadcast+synch networks
  - AP-1000+, AP-3000(1995)

- トランスピュータ,英Inmos社
  - 1979年, Inmos 英政府により設立
  - プロセッサ・メモリ・リンク／チップ
  - 並列マシンのビルディングブロック
  - 1985年, T414, integer
  - 1988年, T800, floating
  - 1994年, T9000,仮想チャンネル,スーパスケーラ
- AP-1000,富士通
  - 1991年, 16〜1024セル, SPARC
  - トーラス+ブロードキャスト+同期
  - AP-1000+, AP-3000(1995年)

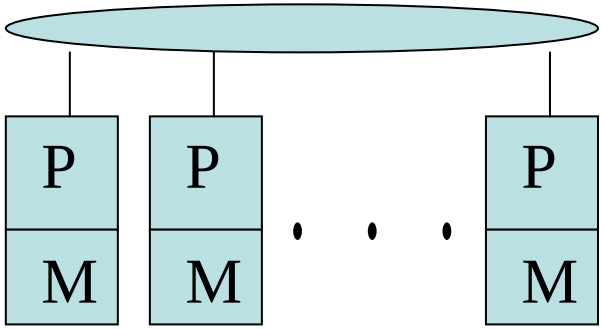
- KSR-1, Kendall Square Research
  - 1991, virtual shared memory parallel machine
  - 1 terabytes virtual address space
  - ALLCACHE memory, on demand data transfer
  - COMA: Cache Only Memory Architecture
- Cenju-3, NEC
  - 1993, max 256 PEs, VR4400
  - Multistage network(baseline network)
  - 1997, Cenju-4, VR10000
- Exemplar, HP
  - 1996, PA8000, 16CPU/node
  - Distributed shared memory parallel machine
  - Cc-NUMA(cache coherent Non-Uniform Memory Access)

- KSR-1, **米ケンダールスクエアリサーチズ社**
  - 1991**年**, **仮想共有メモリ並列マシン**
  - 1 **テラバイト仮想アドレス空間**
  - ALLCACHE**メモリ**, **要求によるデータの移動**
  - COMA: Cache Only Memory Architecture
- Cenju-3, NEC
  - 1993**年**, **最大**256 PEs, VR4400
  - **多段接続網**(**ベースライン網**)
  - 1997**年**, Cenju-4, VR10000
- Exemplar, HP
  - 1996**年**, PA8000, 16CPU/**ノード**
  - **分散共有メモリ並列マシン**
  - Cc-NUMA(cache coherent Non-Uniform Memory Access)

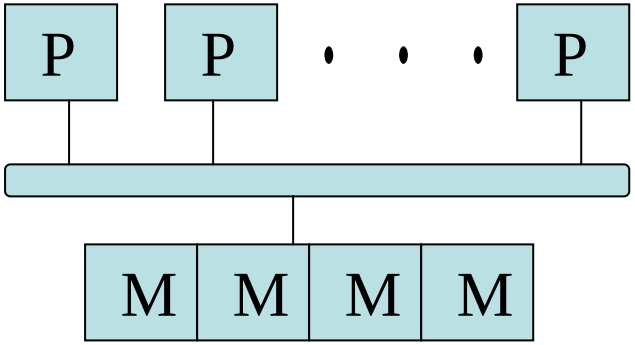
- Enterprise10000, Sun
  - UltraSPARC, 4～64 processors
  - 4 processor SMP per node
  - 16×16 crossbar switch for data & address bus
  - Cc-NUMA
- Blue Gene, IBM
  - 1999-
  - Dual PowerPC440/node, 1024 nodes/rack
  - 512MB SDRAM/node, 3D torus network
  - 2004.11 Blue Gene/L 70.72 Teraflops, world fastest
  - 2005.10 65536 processors, max 360 Teraflops, actual 280 Teraflops
  - 2010 Blue Gene/P 1 Petaflops, 2012 10 Petaflops

- Enterprise10000, Sun
  - UltraSPARC, 4〜64プロセッサ
  - 4プロセッサのSMP /ノード
  - データは16×16 クロスバスイッチ、アドレスはバスを使用
  - Cc-NUMA
- Blue Gene, IBM
  - 1999-
  - Dual PowerPC440/ノード, 1024 ノード/ラック
  - 512MB SDRAM/ノード, 3D トーラスネットワーク
  - 2004.11 Blue Gene/L 70.72 Teraflops, 世界最速
  - 2005.10 65536 プロセッサ, 最大 360 Teraflops, 実働 280 Teraflops
  - 2010 Blue Gene/P 1 Petaflops, 2012 10 Petaflops

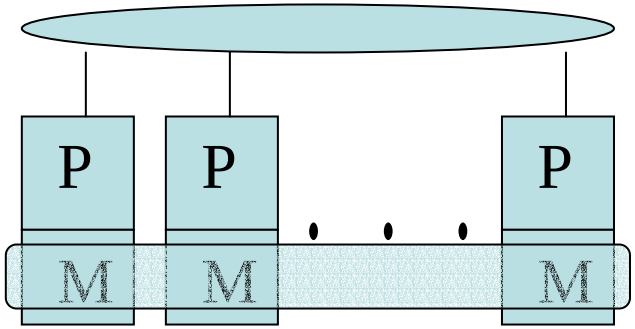
Memory Model



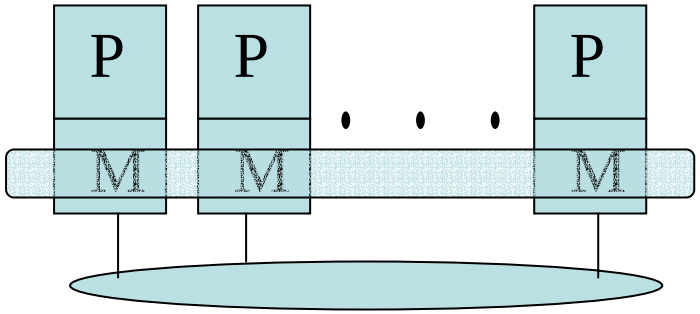
(a) Distributed memory



(b) Shared memory



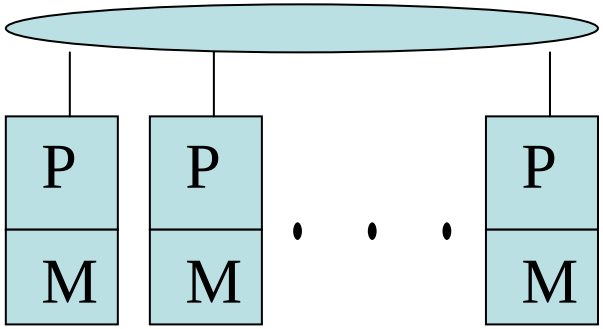
(c) Distributed shared memory



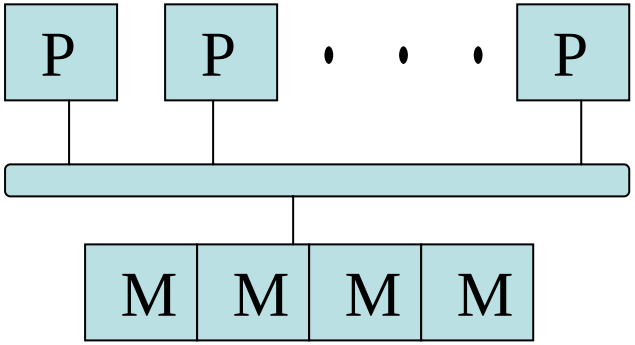
Shared & enlarged

(d) Virtual shared memory

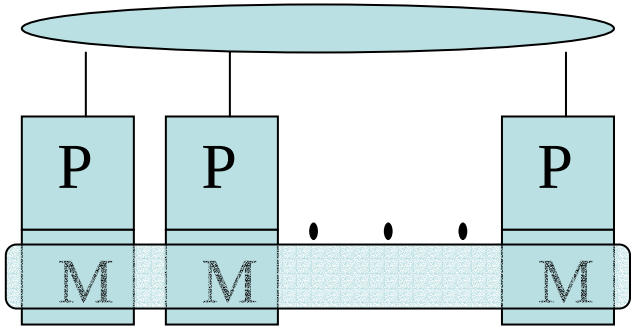
メモリモデル



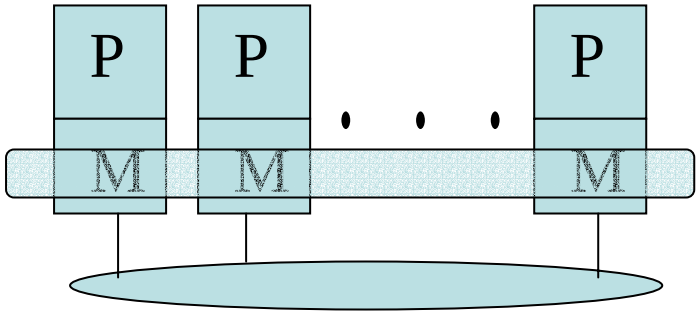
(a) 分散メモリ



(b) 共有メモリ



(c) 分散共有メモリ



共有と仮想化

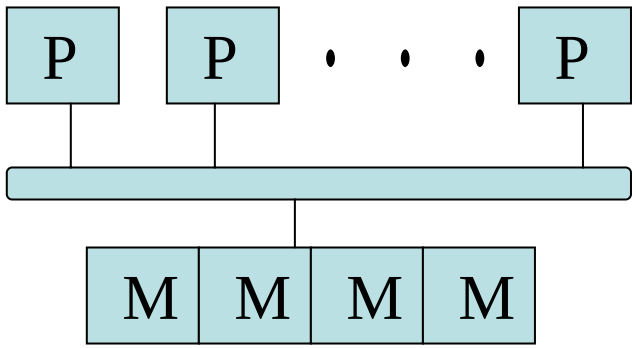
(d) 仮想共有メモリ

UMA and NUMA

Uniform Memory Access

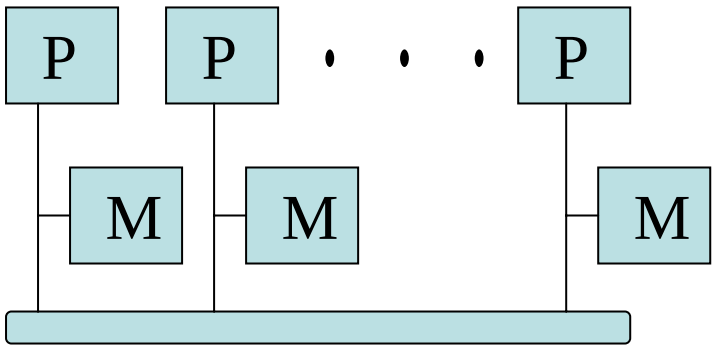
-memory access time is constant

-SMP: Symmetric Multiprocessor



Non Uniform Memory Access

- memory access time is variable

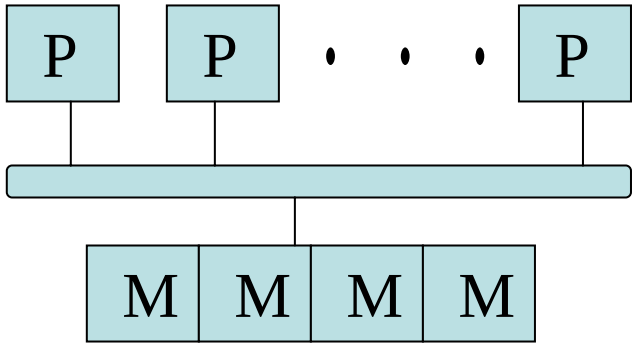


UMA and NUMA

均質メモリアクセス

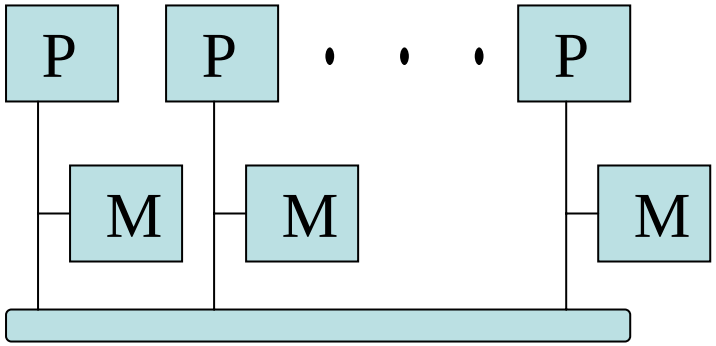
-メモリアクセス時間が一定

-SMP(Symmetric Multiprocessor)
対称型マルチプロセッサ

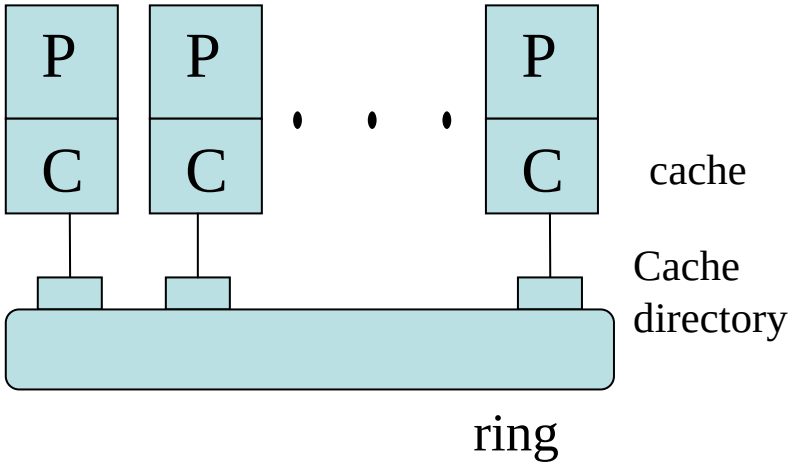


非均質メモリアクセス

- メモリアクセス時間が可変

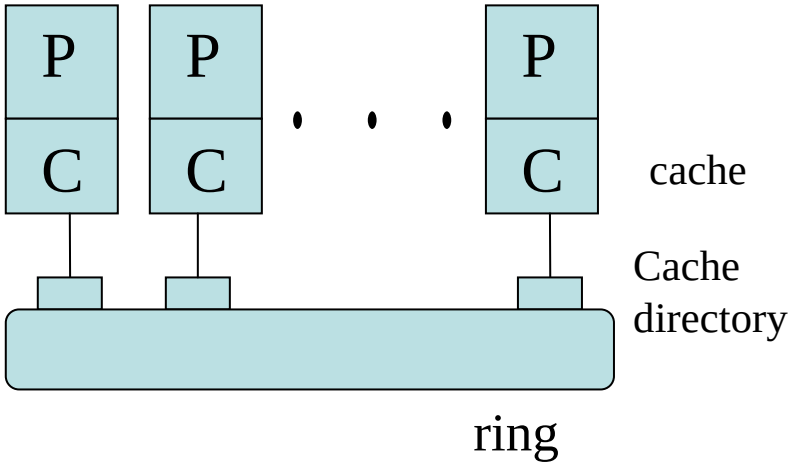


KSR-1: a virtual shared memory machine



- ALLCACHE memory
- Physical address space
  - 32MB✕32= 1GB
- Virtual address space
  - 1 Terabytes
- 1000 times bigger address space

KSR-1: 仮想共有メモリマシン



- ALLCACHE メモリ
- **物理的なアドレス空間**
  - 32MB×32= 1GB
- **仮想アドレス空間**
  - 1テラバイト
- 1000 倍のアドレス空間

Interconnection Networks

- Static (direct) networks
  - Directly connected via links
  - Ring, mesh, torus, tree, hypercube, pyramid, fully connected, …
- Dynamic (indirect) networks
  - Connected via switching elements
  - Crossbar, shuffle exchange network, omega network, …
- Communication performance
  - Bandwidth
  - Latency
  - Total number of links

相互結合網

- 静的 (直接) ネットワーク
  - リンクを通して直接つながっている
  - リング, メッシュ, トーラス, トリー, ハイパーキューブ, ピラミッド, 完全結合網, …
- 動的 (間接) ネットワーク
  - スイッチング素子を通してつながっている
  - クロスバー, シャッフル交換網, オメガネットワーク, …
- 通信パフォーマンス
  - バンド幅： 最大データ転送速度
  - レイテンシ： 遅延時間
  - リンク総数： ハードウェア量

Details of Interconnection Networks

- Refer to CSE431 Computer Architecture
- Lecture 27 Network Connected Multi's
- Prof. Mary Jane Irwin, PenState Univ.
- Adapted from Computer Organization and Design, Patterson & Hennessy, 2005

相互結合網の詳細

- CSE431 Computer Architecture 参照
- 講義27 Network Connected Multi’s
- Mary Jane Irwin教授, PenState Univ.
- パターソン＆ヘネシー、コンピュータの構成と設計、2005から改作

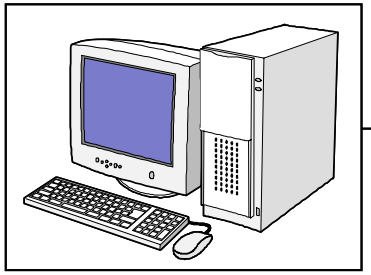
PC Cluster

- Beowulf
  - The Beowulf project 1993～
  - 16PCs(80486), 1GFLOPS, 1994
  - Built from commodity hardware & software
  - Ethernet, TCP/IP, MPI
- SCore
  - Global operating system
  - High communication functions, single system image, high availability, seamless environment/parallel programming
  - Myrinet, PM, MPITCH-SCore
  - Shared memory parallel programming support: OpenMp

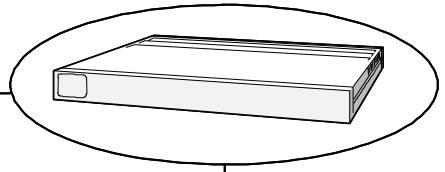
PC クラスタ

- ベオウルフ
  - ベオウルフプロジェクト 1993年～
  - 16PCs(80486), 1GFLOPS, 1994年
  - 日常品のハードウェアとソフトウェアから構築
  - Ethernet, TCP/IP, MPI
- SCore
  - RWC　1995年～
  - グローバルオペレーティングシステム
  - 高度な通信機能, 単一のシステムイメージ, 高い可用性, シームレス環境/並列プログラミング
  - Myrinet, PM, MPITCH-SCore
  - 共有メモリ並列プログラミングのサポート: OpenMp

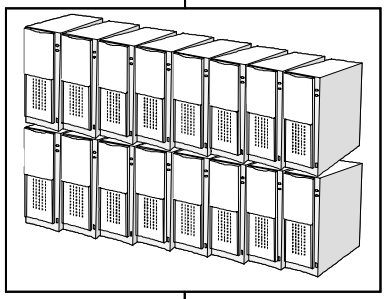
PC Cluster in HPC Lab 2000 Autumn~



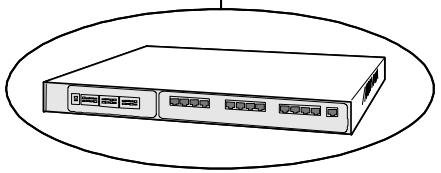
Server Host
-Pentium3 450MHz
-256MB SDRAM
-13.0GB HDD



Ethernet Switch
-maximum bandwidth: 100Mbps
-maximum latency: 80 μs

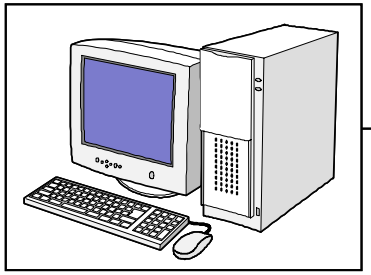


Compute Host (16PCs)
-Pentium3 500MHz
-512MB SDRAM
-6.4GB HDD

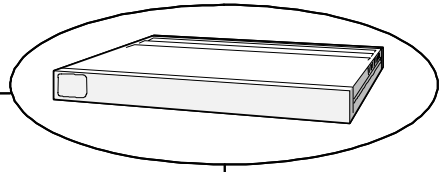


Myrinet-2000 Switch
-maximum bandwidth: 2Gbps
-maximum latency: 9 μs

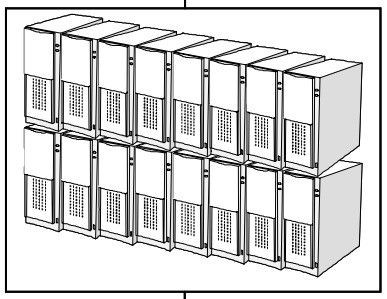
山崎研のPCクラスタ



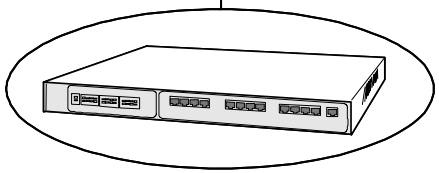
Server Host
-Pentium3 450MHz
-256MB SDRAM
-13.0GB HDD



Ethernet Switch
-maximum bandwidth: 100Mbps
-maximum latency: 80 μs



Compute Host (16PCs)
-Pentium3 500MHz
-512MB SDRAM
-6.4GB HDD



Myrinet-2000 Switch
-maximum bandwidth: 2Gbps
-maximum latency: 9 μs

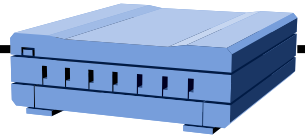
Raptor：PC Cluster　2003~

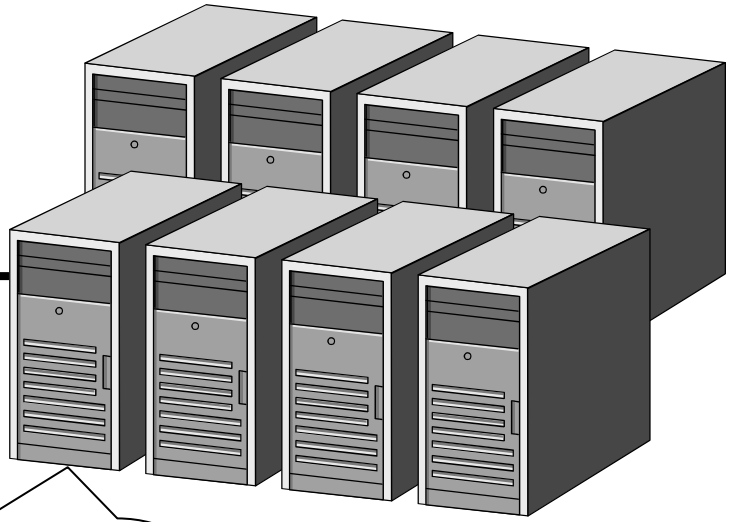
SCore Cluster System Software 5.8.3

Raptor00〜Raptor07

Spino

Gigabit Ethernet Switch





CPU: Xeon 2.8GHz

Memory: 2GB

HDD: 500GB

Pentium4 3.2GHz

2GB RAM

50GB HDD

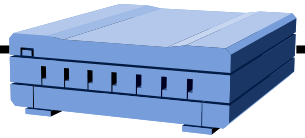
MPI: MPICH-SCore

Diplo: SMP Cluster 2006~

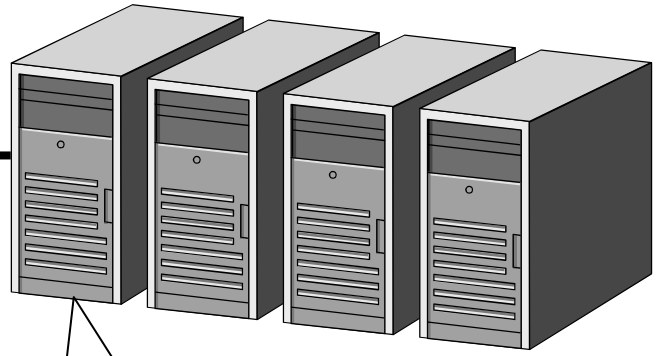
Rocks Cluster System 4.2

Tarbo

Gigabit Ethernet Switch



Diplo00〜Diplo03



CPU: Dual Xeon 3GHz

Memory: 4GB

HDD: 500GB

MPI: MPICH 1.2.7
OpenMP: Intel C/C++ Compiler 9

Dual Xeon 3GHz

Dual Xeon 3GHz

4GB RAM

320GB HDD

5. Earth Simulator

- Global warming, El Nino
- Temperature increases 2 degrees & the surface of the sea rises 50 cm in 2100
- High precision weather prediction: 1～10 km per mesh
- 1000 times bigger memory capacity & computing power are required
- Actual performance 5 Teraflops & peak performance 40 Teraflops
- Started 1997, and completed in Feb. 2002

5. 地球シミュレータ

- 地球温暖化, エルニーニョ
- 2100年　気温2度上昇、海面50cm上昇
- 気象予測の高精度化: 1〜10 km /メッシュ
- メモリ量、計算能力とも現行の1000倍以上必要
- 実行性能 5 Teraflops 、 ピーク性能 40 Teraflops
- 1997年から開始, 2002年2月完成

- 8 processors per node & 640 nodes in all
- 8 Gigaflops per processor & 64 Gigaflops per node
- 10 Terabytes memory in all & 16 Gigabytes shared memory in each node
- 35.86 Teraflops for Linpack, the best in the world
- ASCI White system 12.3 Teraflops in 2000
- Each node consists of 8 arithmetic processors, 16GB shared memory, a remote control unit and an I/O processor
- Arithmetic processor consists of vector processors, a scalar processor and a memory access unit

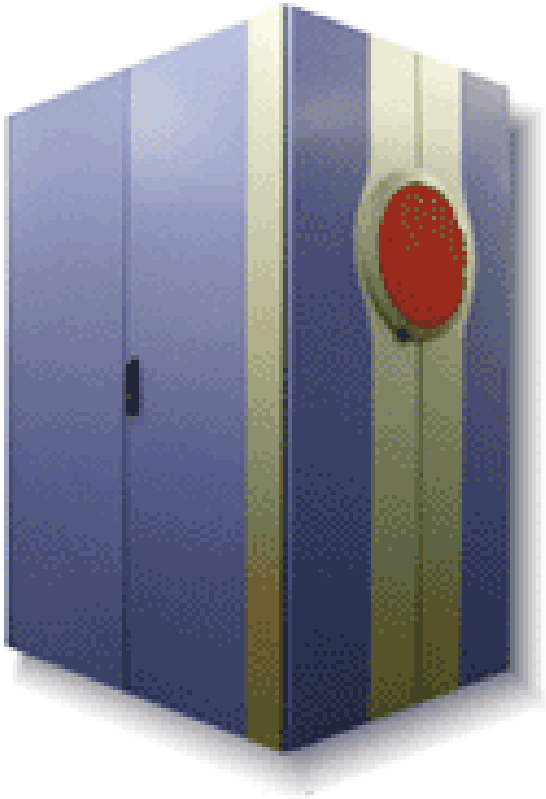
- 8 プロセッサ/ノード、全 640 ノード
- 8 GFlops /プロセッサ、 64 GFlops /ノード
- 全メモリ 10TB、ノード内共有メモリ 16GB
- 35.86 Teraflops (Linkpackベンチマーク) を達成、世界一 (Top500)
- 2000年、ASCI White system 12.3 Teraflops
- 計算ノード=ベクトルプロセッサ8台+共有メモリ16GB+リモート制御装置+入出力プロセッサ
- ベクトルプロセッサ(AP)=ベクトル処理部+スカラ処理部+メモリアクセス処理部

- Crossbar switch: max 12.3GB/sec
- Simulator building: 50m×65m×17m
- 83200 cables & total length is 2800 km
- Parallelism = vector processing in AP & shared memory parallel processing in a node & distributed memory parallel processing among nodes
- MPI, HPF, OpenMP, Hybrid

- クロスバスイッチ: 最大12.3GB/sec
- シミュレータ棟: 50m×65m×17m
- 83200 本のケーブル　全長 2800 km
- 並列性 = AP内ベクトル処理、ノード内並列処理 (共有メモリ) 、ノード間並列処理 (分散メモリ)
- MPI, HPF, OpenMP, ハイブリッド

http://ascii24.com/news/i/hard/article/2002/06/14/636530-000.html

http://ascii24.com/news/i/hard/article/2002/06/14/636530-000.html



計算ノード筐体

Processor Node Cabinet

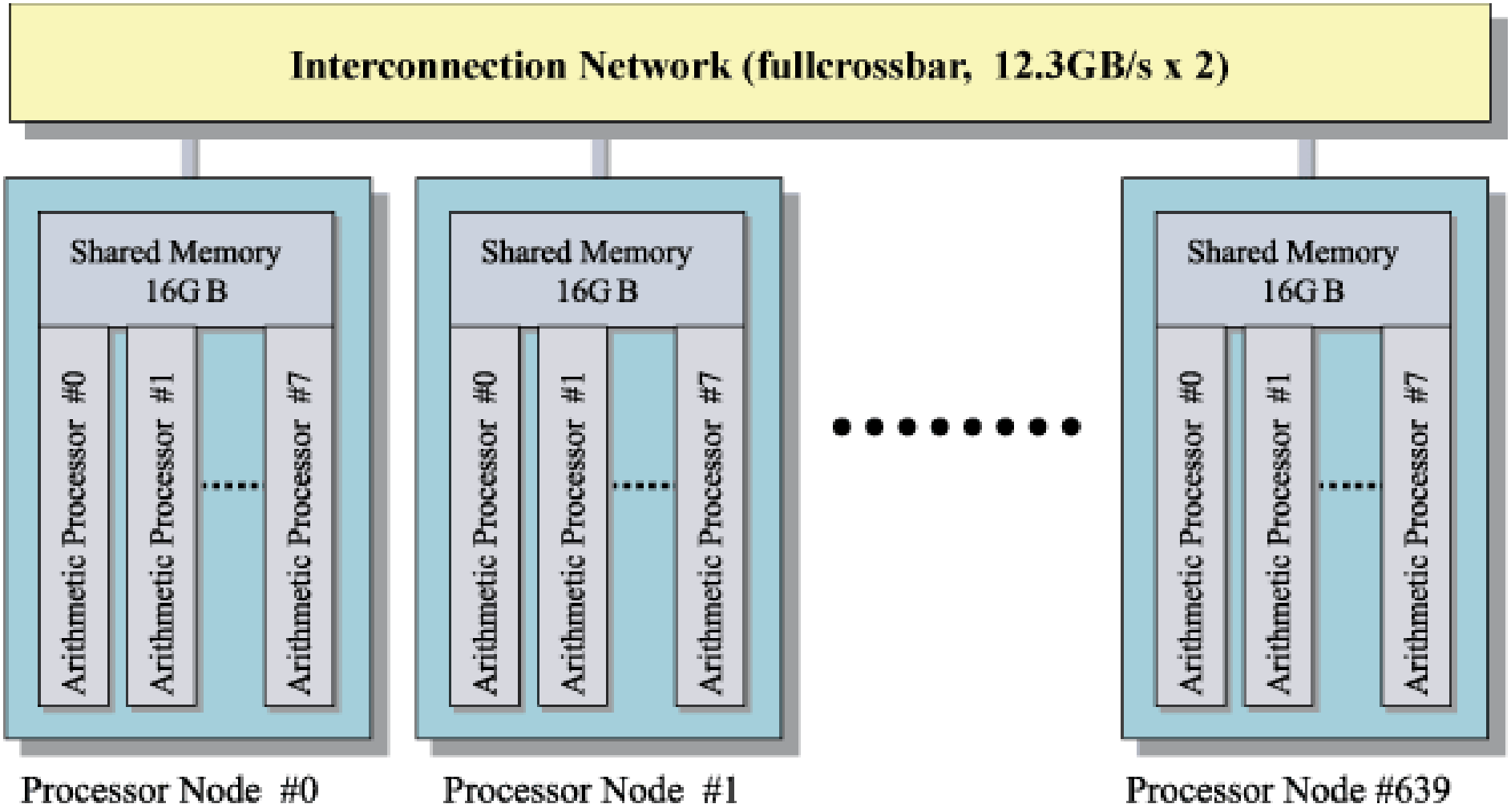
結合ネットワーク筐体

Interconnection Network Cabinet

http://www.es.jamstec.go.jp/esc/jp/index.html

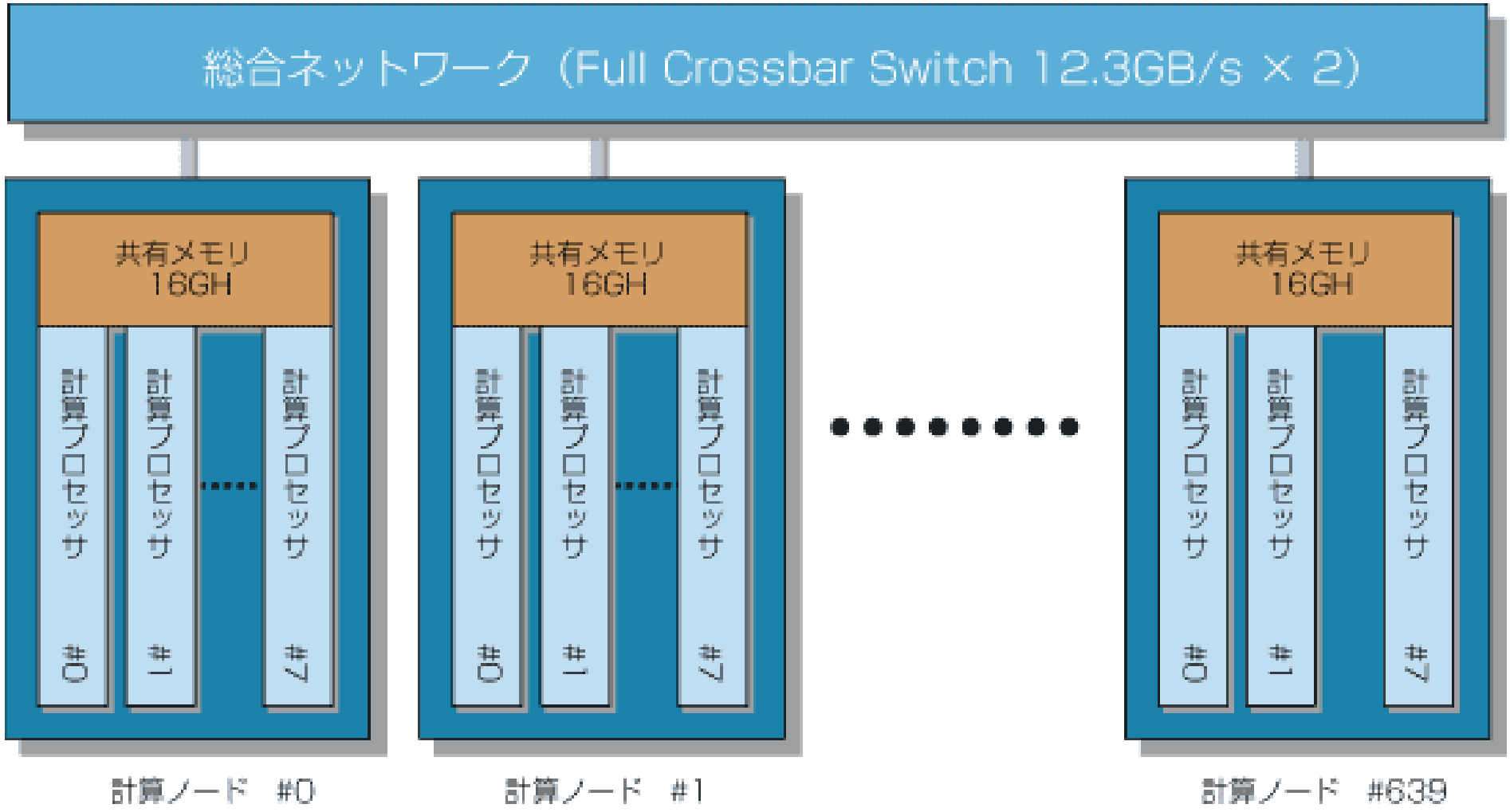
http://ascii24.com/news/i/hard/article/2002/06/14/636530-000.html

System Configuration



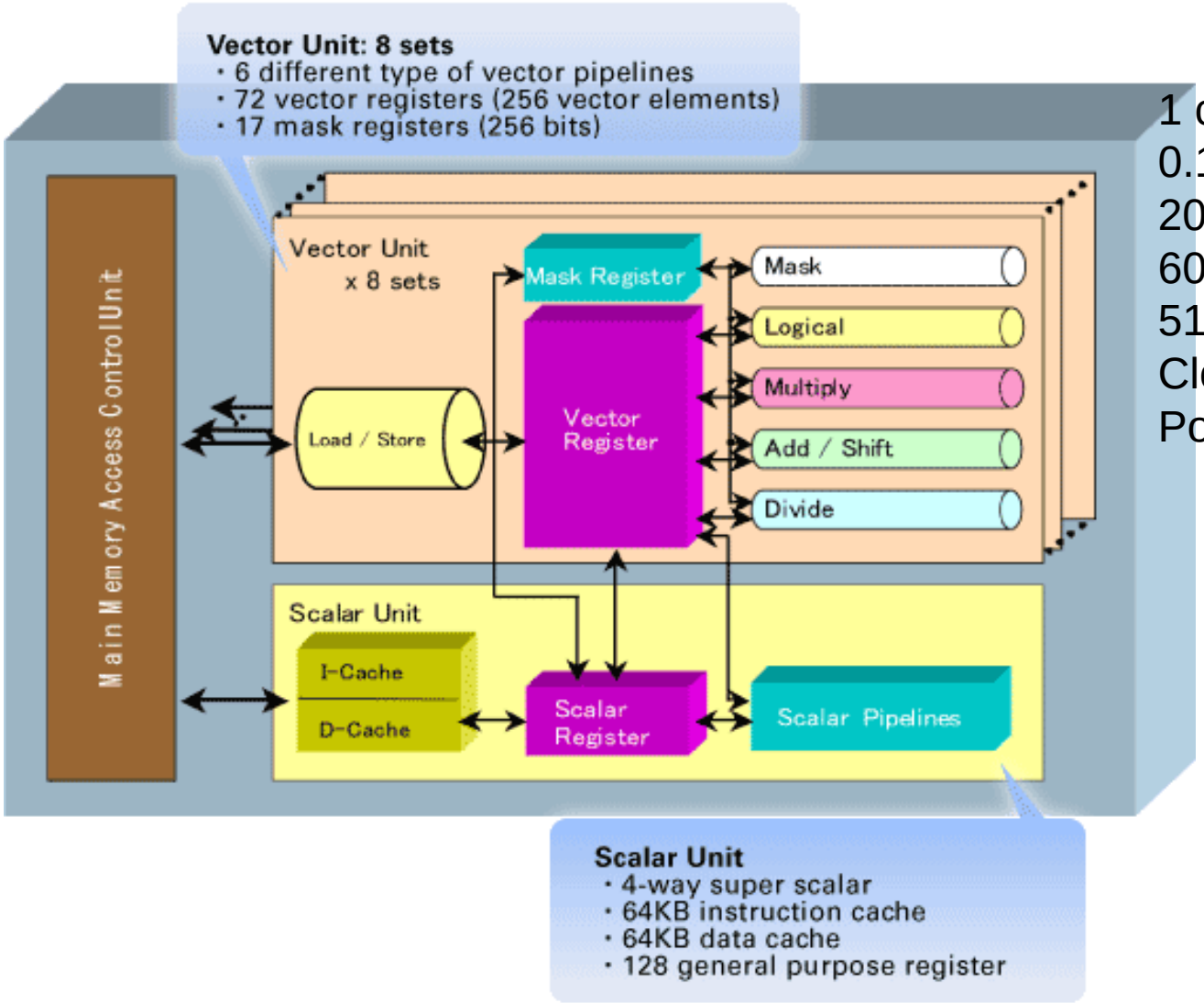
http://www.es.jamstec.go.jp/esc/eng/index.html

システム構成



http://www.es.jamstec.go.jp/esc/jp/index.html

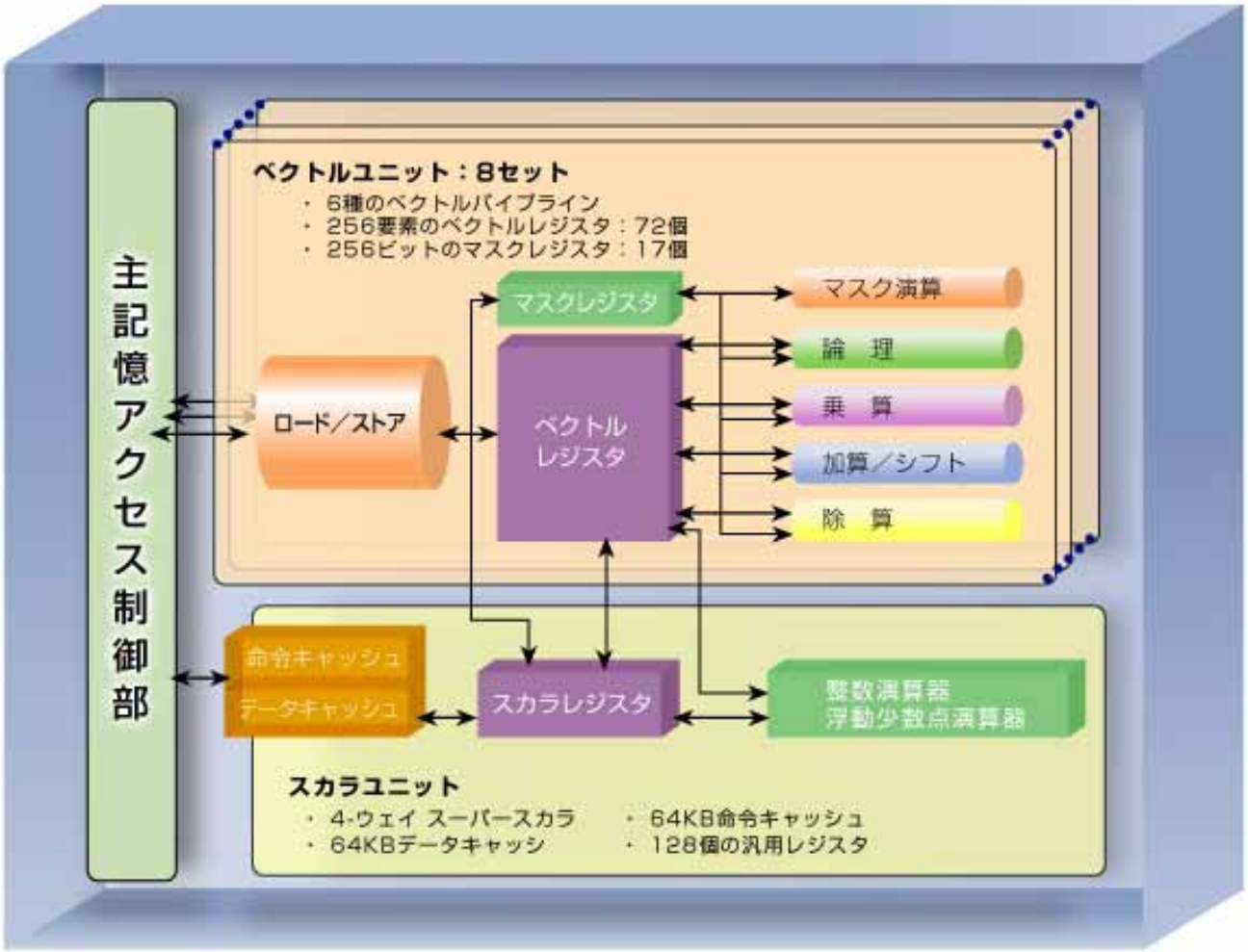
Arithmetic Processor (AP)



1 chip LSI: 8Gflops
0.15µm CMOS technology
20.79mm x 20.79mm
60 Million trangister
5185 pins
Clock frequency: 500MHz
Power dissipation 140W (Typ.)

http://www.es.jamstec.go.jp/esc/eng/index.html

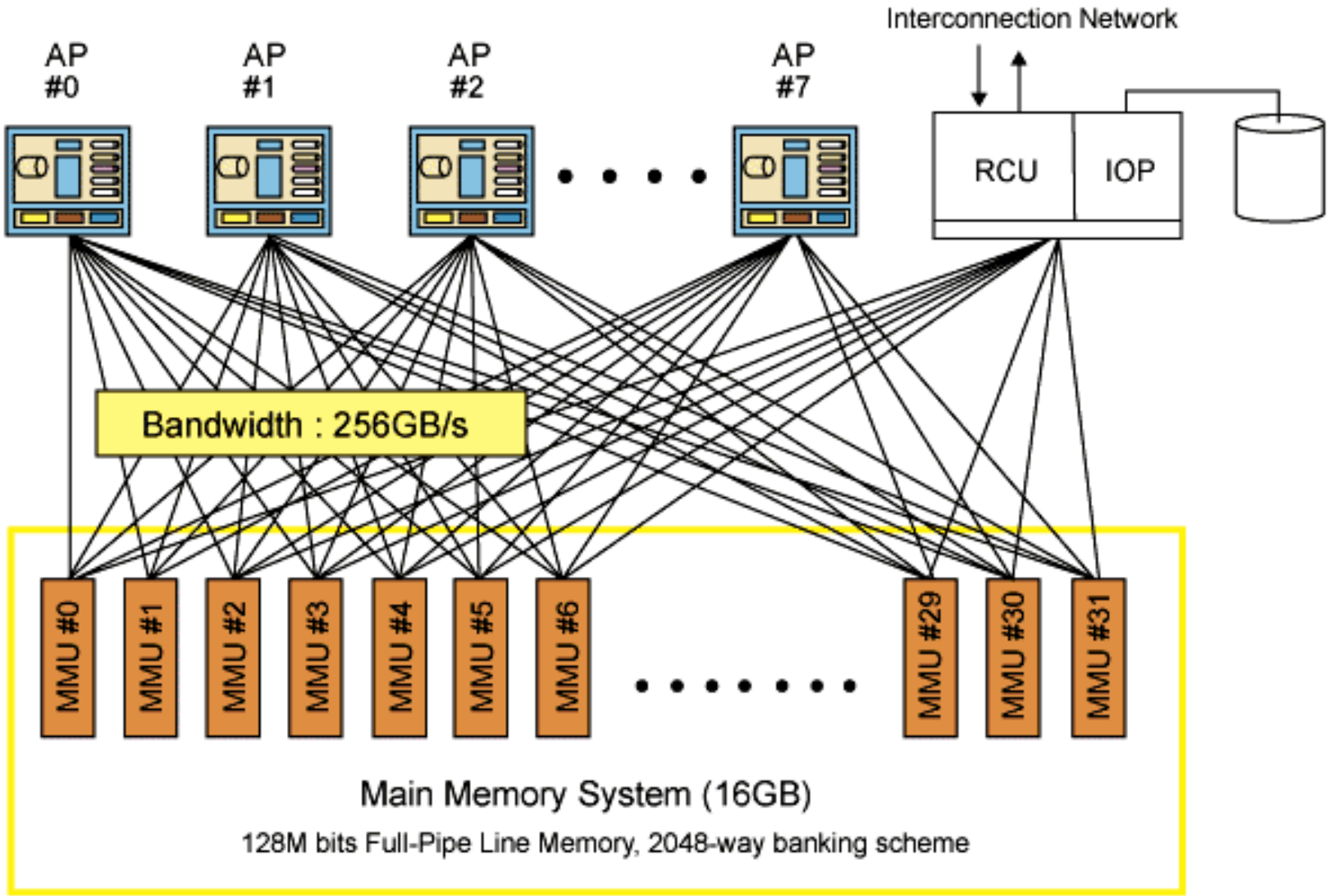
計算プロセッサ(AP)



1チップLSI: 8Gflops
0.15µm CMOS**テクノロジ**
20.79mm x 20.79mm
6000**万トランジスタ**
5185 **ピン**
クロック周波数 500MHz
消費電力 140W (Typ.)

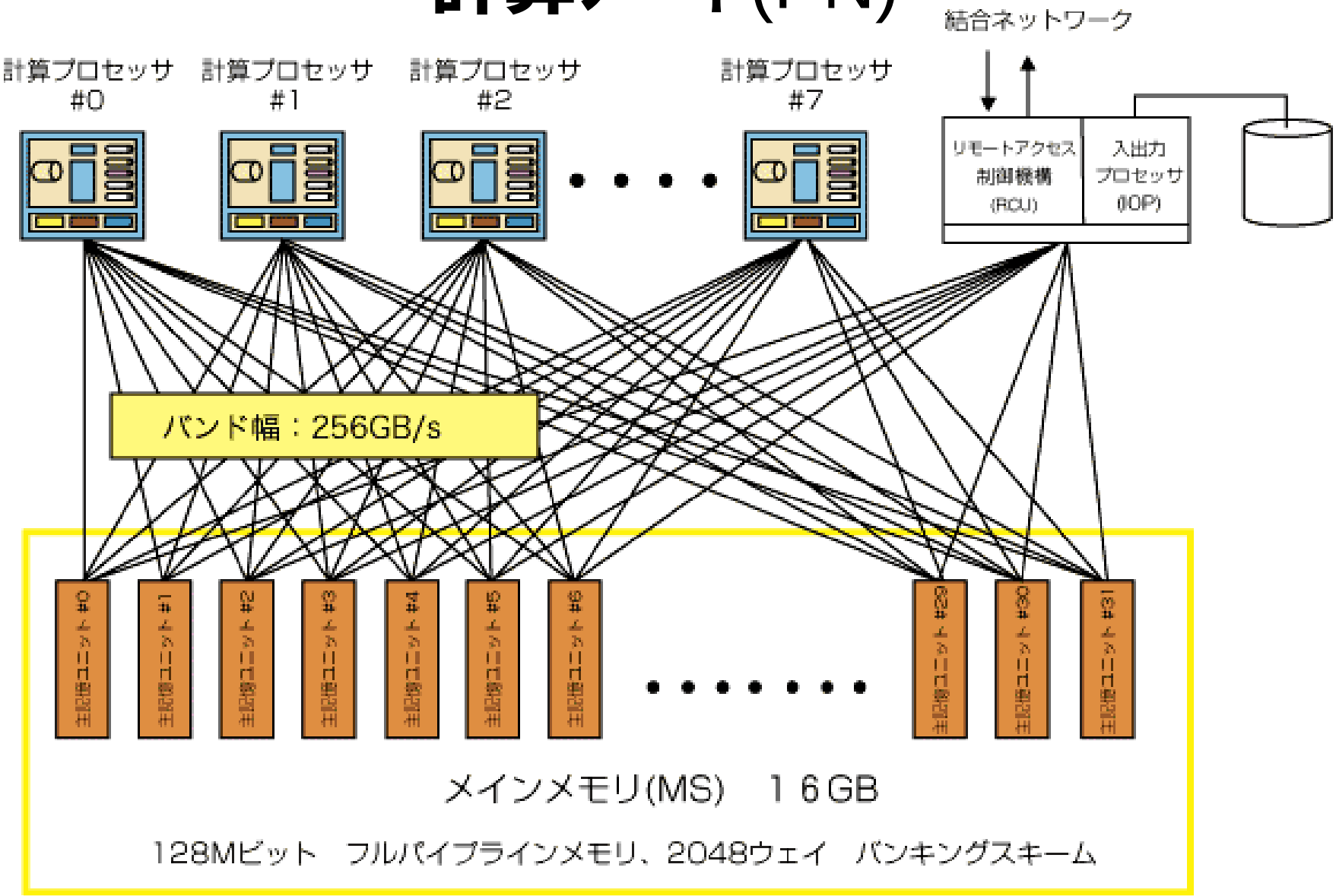
http://www.es.jamstec.go.jp/esc/jp/index.html

Processor Node (PN)



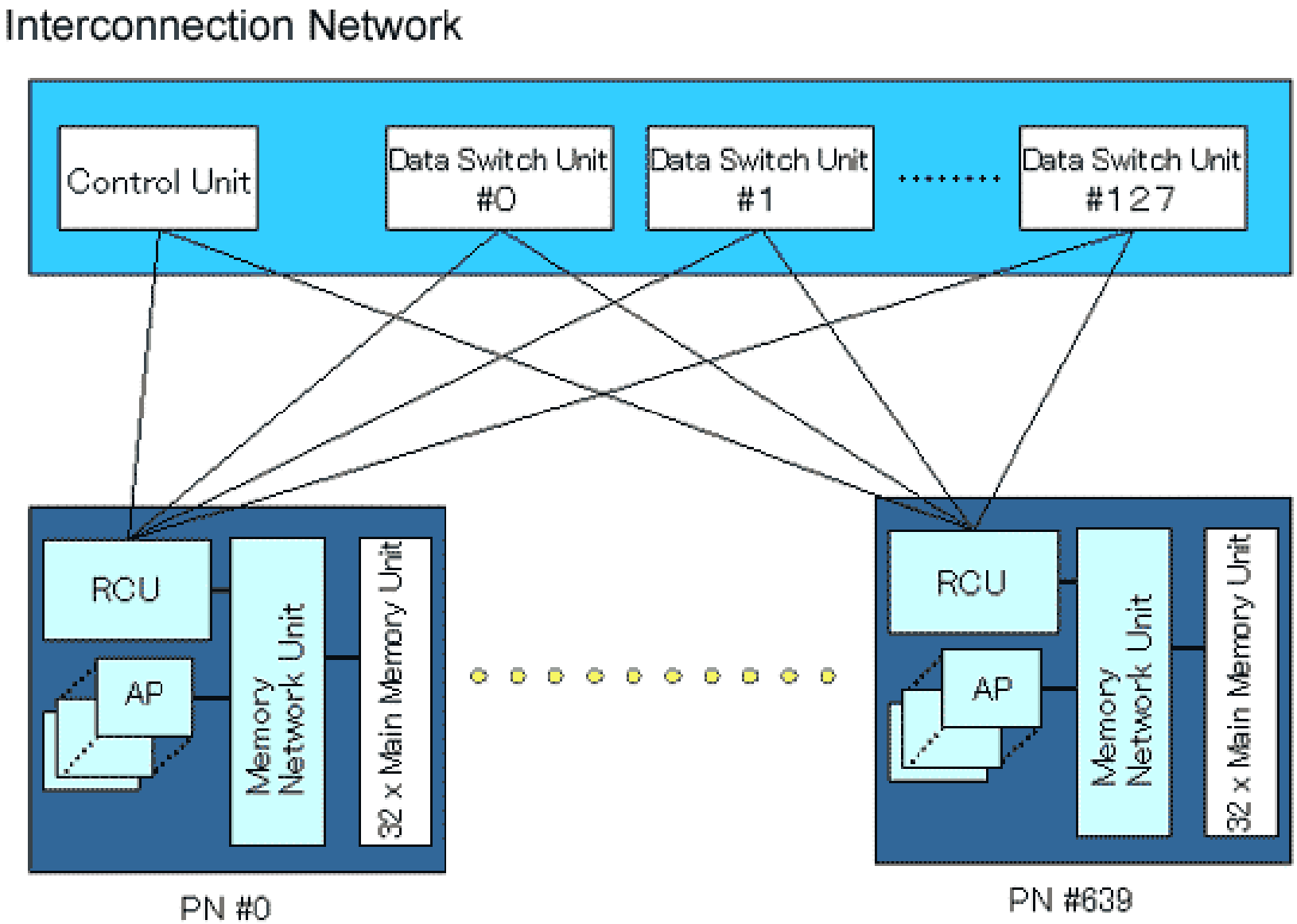
http://www.es.jamstec.go.jp/esc/eng/index.html

計算ノード(PN)



http://www.es.jamstec.go.jp/esc/jp/index.html

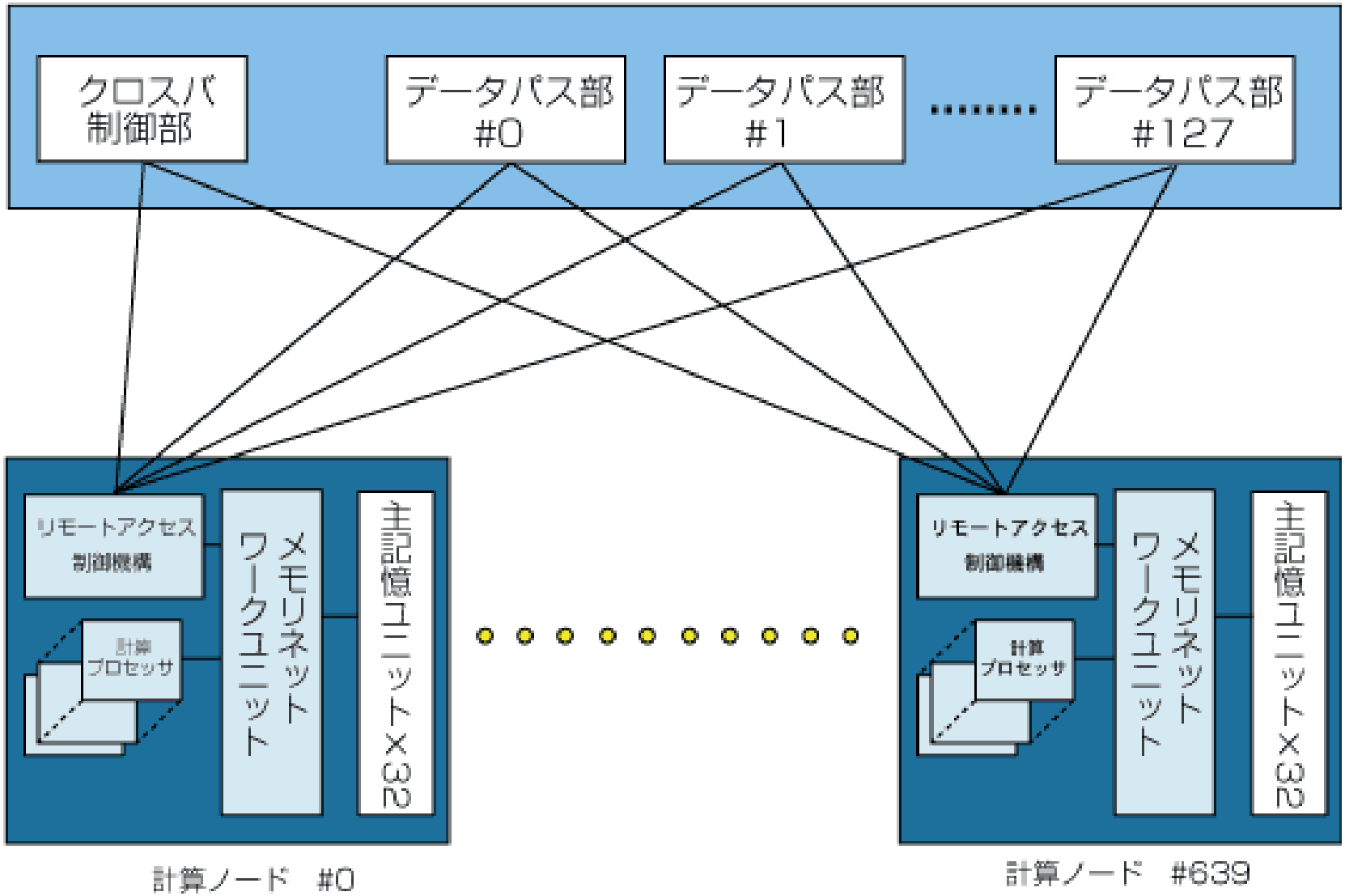
Construction of Interconnection Network



http://www.es.jamstec.go.jp/esc/eng/index.html

結合ネットワーク接続の構成

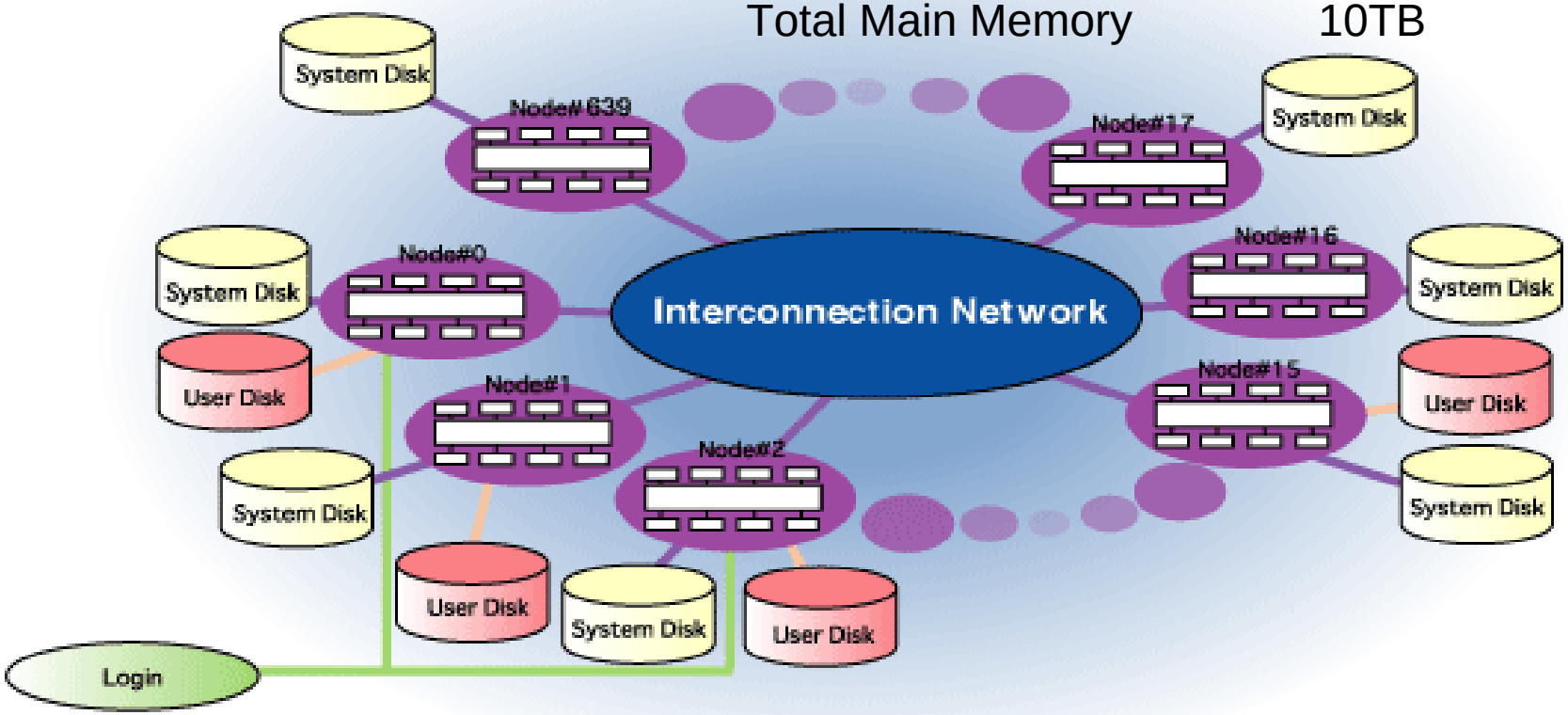
結合ネットワーク(IN)部



http://www.es.jamstec.go.jp/esc/jp/index.html

System Performance

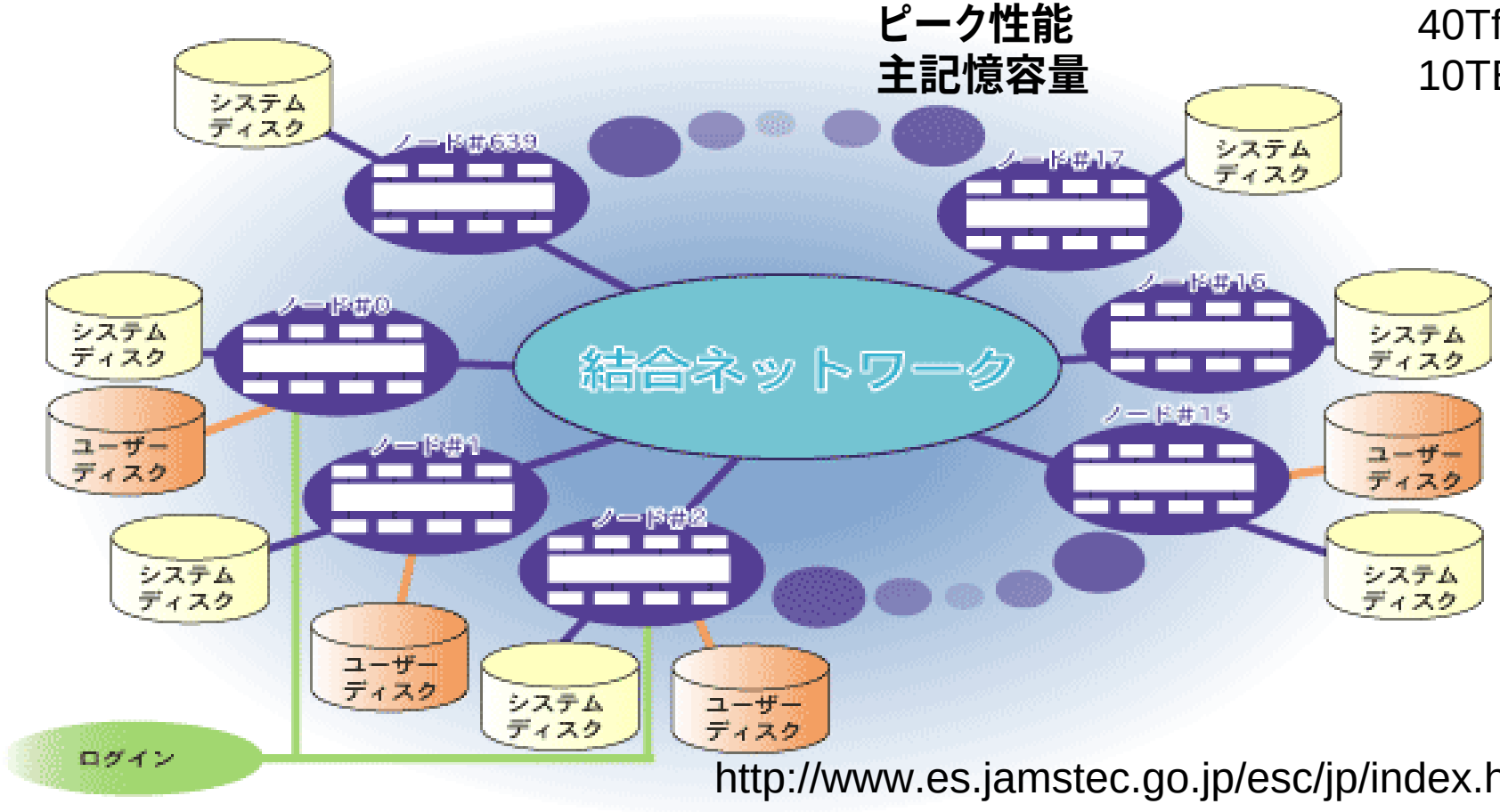
Peak Performance / AP	8Gflops
Total number of APs	5120
Peak Performance / PN	64Gflops
Total number of PNs	640
Shared Memory / PN	16GB
Total Peak Performance	40Tflops 64Gf×640
Total Main Memory	10TB



http://www.es.jamstec.go.jp/esc/eng/index.html

システム性能

計算プロセッサのピーク性能	8Gflops
総プロセッサ数	5120
計算ノードのピーク性能	64Gflops
総計算ノード数	640
計算ノードの主記憶容量	16GB
ピーク性能	40Tflops
主記憶容量	10TB



http://www.es.jamstec.go.jp/esc/jp/index.html

Milestones of Development

- In July 1996, the promotion of research & development for the Earth Simulator was reported to the Science Technology Agency, based on the report titled "For Realization of the Global Change Prediction" .
- In April 1997, the budget for the development of the Earth Simulator was authorized. The Earth Simulator Research and Development Center was established. The project started.
- The conceptual system design of the Earth Simulator proposed by NEC Corporation was selected by bidding.
- Manufacturing the Earth Simulator started in March 2000.
- At the end of February in 2002, all 640 processor nodes (PN's) started its operation for check up. The Earth Simulator Center (ESC) started the actual operation in March, 2002.

http://www.es.jamstec.go.jp/esc/jp/index.html

開発経緯

1996年7月、科学技術庁 航空･電子等技術審議会 地球科学技術部会の報告書
「地球変動予測の実現に向けて」により地球シミュレータの研究開発推進が提言された。

1997年度科学技術庁で地球シミュレータ研究開発予算が認められ、
地球シミュレータの開発が開始された。

1997年11月、複数の概念設計の提案から、日本電気 ㈱ の提案を採用した。

2000年3月より製作開始。

2002年2月末　640ノードの全計算ノードが稼働し、運用を開始した。

http://www.es.jamstec.go.jp/esc/jp/index.html

6. Parallel Programming

- C, Fortran+Directives
  - Optimizing compiler, additional information by a user
- HPF(High Performance Fortran)
  - HPF1.0(1993), HPF2.0(1997)
  - Data division by a user & others by a compiler
  - Research on optimizing compiler
- PVM(Parallel Virtual Machine)
  - 1989～Oak Ridge Lab, Ver.2(1991), Ver.3(1993)
  - Easy parallel programming on LAN
  - Dynamic process management, resource management

6. 並列プログラミング

- C, Fortran+**並列構文**
  - **最適化コンパイラ**, **ユーザが補助情報を与える**
- HPF(High Performance Fortran)
  - HPF1.0(1993**年**), HPF2.0(1997**年**)
  - **データ分割はユーザが、その他をコンパイラが行う**
  - **コンパイラ最適化の研究が進展中**
- PVM(Parallel Virtual Machine)
  - 1989**年～**Oak Ridge **国立研**, Ver.2(1991**年**), Ver.3(1993**年**)
  - LAN**環境で容易に並列プログラミング可能**
  - **動的なプロセス管理**, **資源管理**

- MPI(Message Passing Interface)
  - The MPI Standard(1994), MPI-2(1997)
  - MPMD: Multiple Program Multiple Data
  - Standard message passing library
- OpenMP
  - 1997～ shared memory parallel programming model
  - Parallel directives, runtime libraries, environment variables
  - Fork-join model, processor-farm algorithm
  - Parallelization is done by a user, but simple loops can be automatically parallelized
  - Portability

- MPI(Message Passing Interface)
  - The MPI Standard(1994年), MPI-2(1997年)
  - MPMD: Multiple Program Multiple Data
  - メッセージ通信ライブラリの事実上の標準
- OpenMP
  - 1997年～ 共有メモリ並列プログラミングモデル
  - 並列化指示文, 実行時ライブラリ, 環境変数
  - Fork-join モデル, プロセッサファーム型のアルゴリズム
  - 並列化はユーザの責任。ただし、データ並列が明らかなループは自動並列化
  - 移植性あり

- Posix thread
  - Thread based on Posix, portable operating system interface determined by IEEE
  - Multithreads: multiple threads run on one process
  - Address space(program code & data area) is shared among threads. Each thread has PC, a stack pointer and a stack.
  - Creation & deletion of threads, synchronization among threads, mutual exclusion
  - Each thread is executed by each processor
  - Programming is extremely difficult. For experts only.

- Posix スレッド
  - IEEEが決めたポータブルOSインターフェースに基づくスレッド
  - マルチスレッドとは1つのプロセス中で複数のスレッドが動作すること
  - スレッド間でアドレス空間（プログラムコードとデータ領域）を共有。スレッド毎にPC、スタックポインタ、スタックを持つ
  - スレッドの生成・消滅、スレッド間の同期、相互排除
  - 1スレッドが1プロセッサで実行される
  - プログラミングは極めて困難。一般ユーザには不向き

History of Parallel Programming Languages

Conventional	Data-parallel	Message passing	Shared memory	
C, Fortran+ directives		PVM1.0		**1989**
		PVM2.0		**1991**
	HPF1.0	PVM3.0		**1993**
		MPI standard		**1994**
			Pthread	**1996**
	HPF2.0	MPI-2	OpenMP	**1997**
			C/C++ 1.0	**1998**
			Fortran 1.1	**1999**
			Fortran 2.0	**2000**
	HPF for the Earth Simulator		C/C++ 2.0	**2002**
		MPICH	V2.5 combined	**2005**

No portability

Standard

Portability

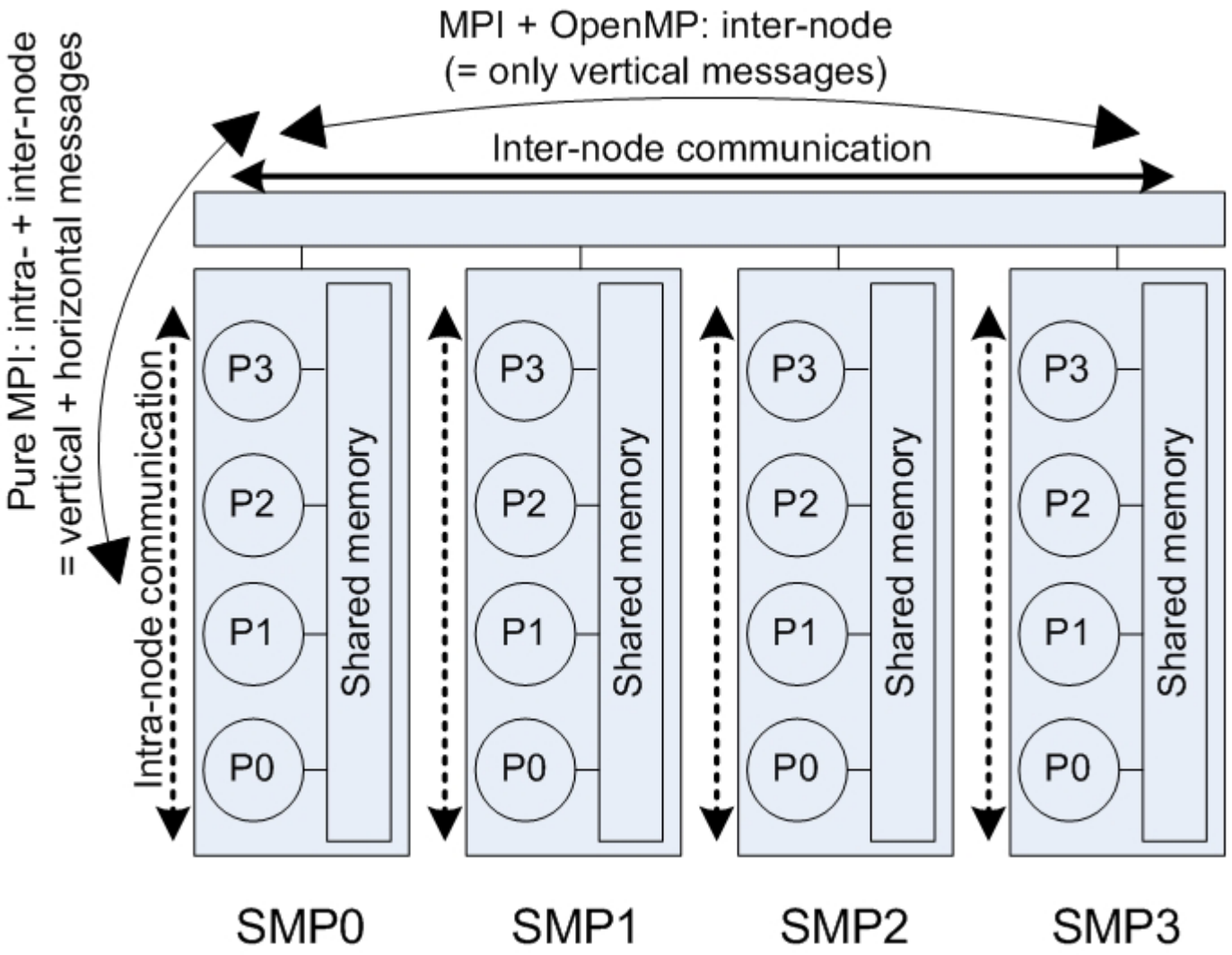
並列プログラミング言語の歴史

C, Fortran+
並列構文

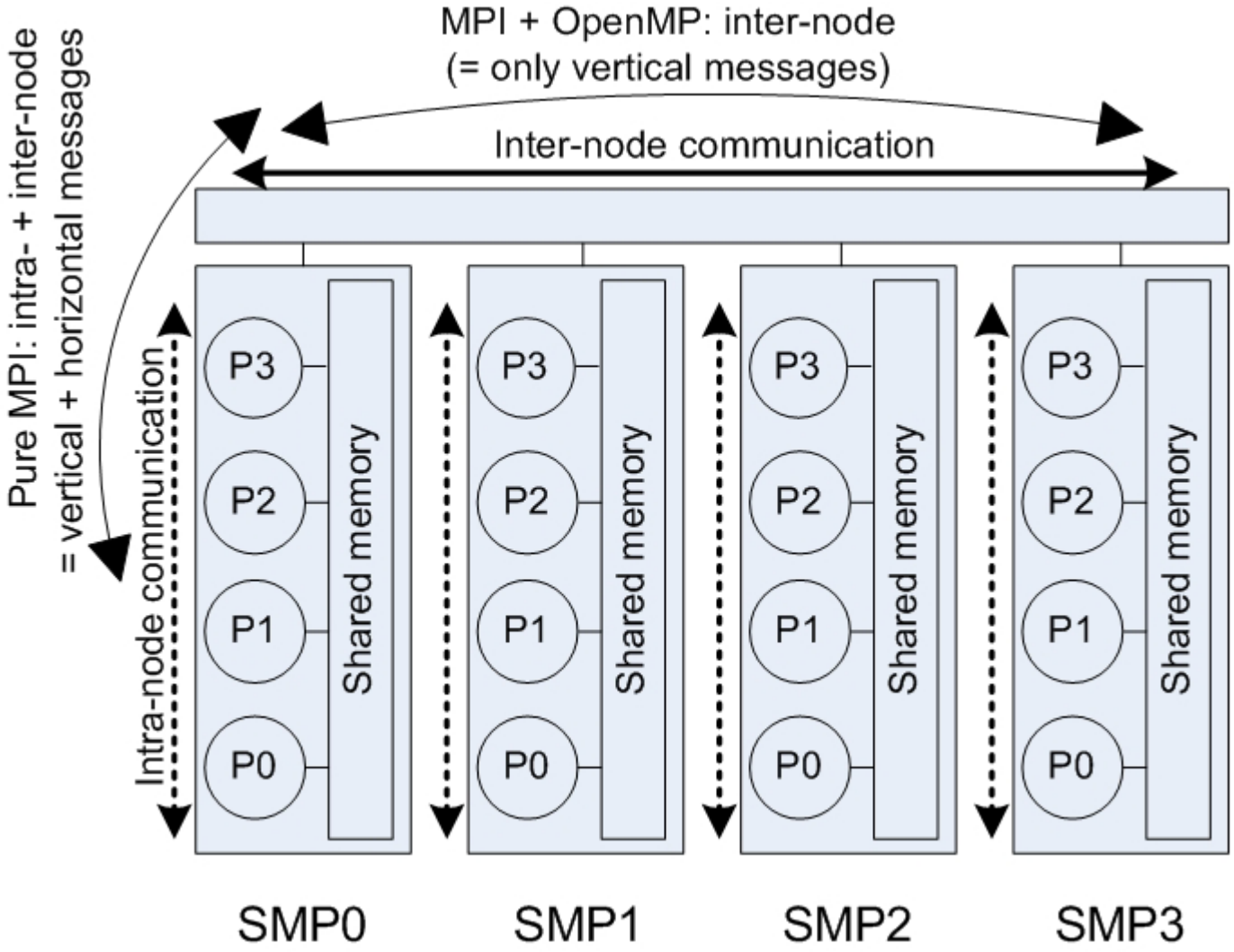
移植性なし

		PVM1.0		1989
		PVM2.0		1991
	HPF1.0	PVM3.0		1993
		MPI standard		1994
			Pthread	1996
	HPF2.0	MPI-2	OpenMP	1997
			C/C++ 1.0	1998
			Fortran 1.1	1999
			Fortran 2.0	2000
	HPF for the Earth Simulator		C/C++ 2.0	2002
		MPICH	V2.5 combined	2005
		標準	移植性あり	
従来	データ並列	メッセージ通信	共有メモリ	

Hybrid Parallel Programming



ハイブリッド並列プログラミング

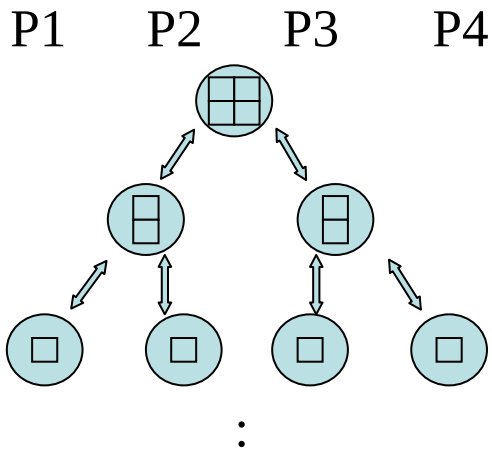


7. Parallel Algorithms

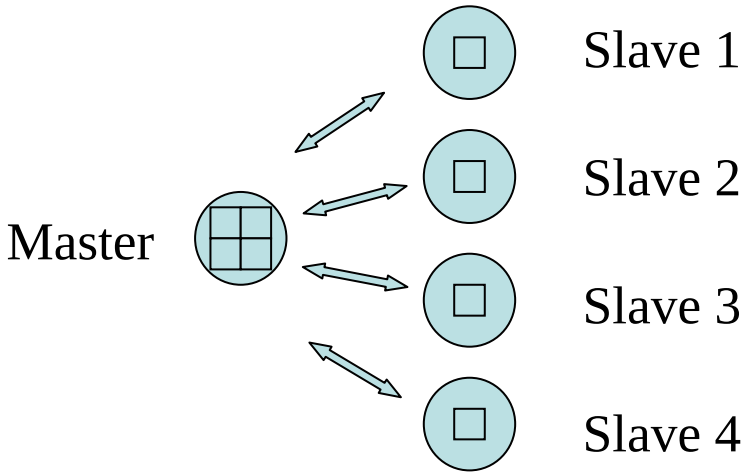
- Divide & conquer
  - Divides a problem into subordinate problems which are themselves recursively solved by dividing them further.
- Processor farms
  - Divides a problem into a number of independent computations and the results of these computations are combined.
- Process networks
  - Divides a computation into stages with the data flowing through the stages. Stages operate synchronously.
- Iterative transformation
  - Objects are transformed until the termination conditions are satisfied through several iteration steps.

7. 並列アルゴリズム

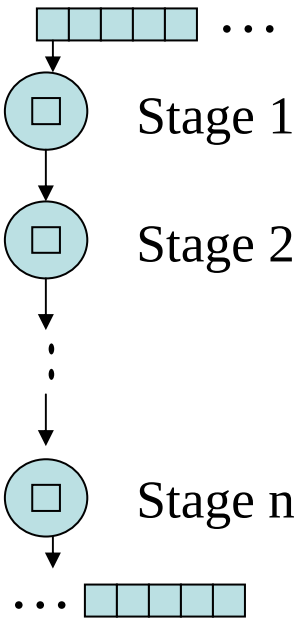
- 分割統治法
  - 問題は下位の部分問題に分割され、それら自身がさらに分割されて再帰的に解かれる
- プロセッサファーム
  - 問題は複数の独立な計算に分割され、それらの結果が結合される
- プロセスネットワーク
  - 計算を複数のステージに分け、データがステージを流れる。ステージは協調して動作する
- 繰り返し変換
  - 終了条件を満たすまで、各オブジェクトを繰り返し変換する



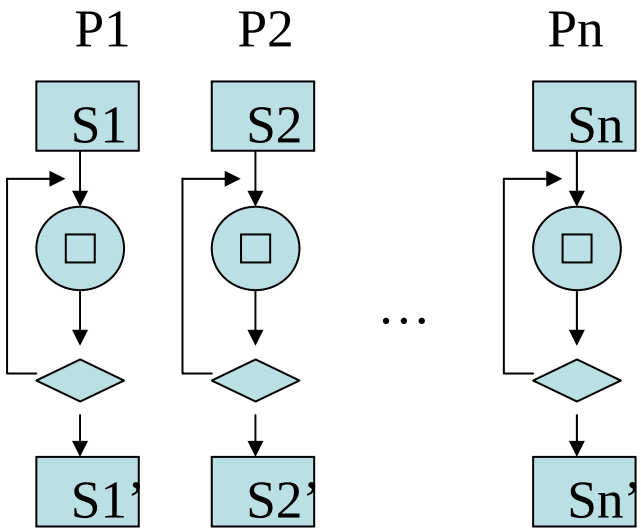
(a) Divide & conquer



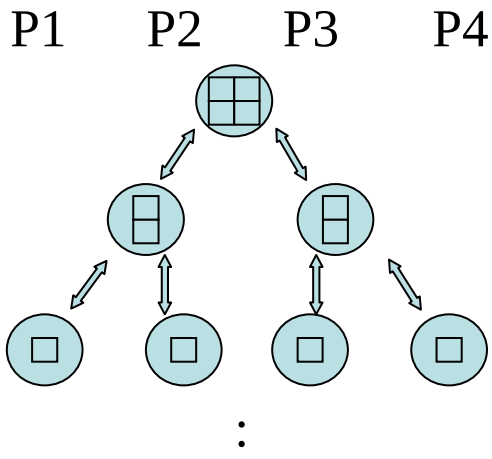
(b) Processor farms



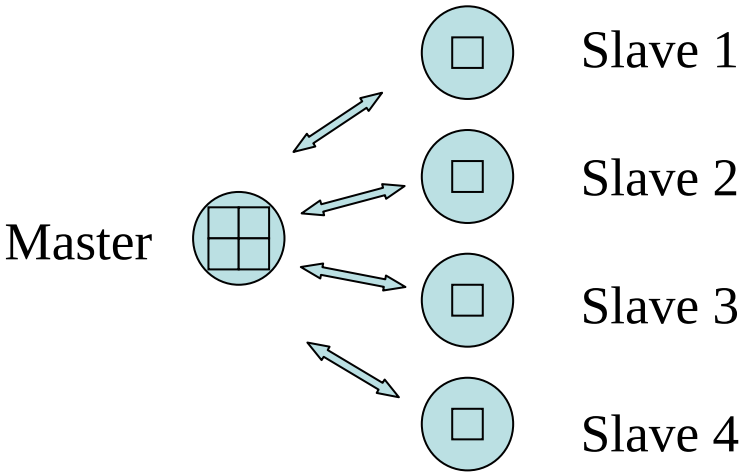
(c) Process networks



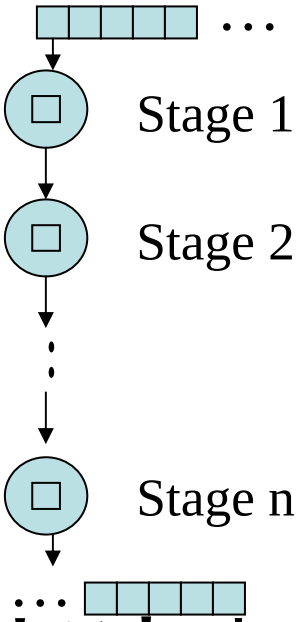
(d) Iterative transformation



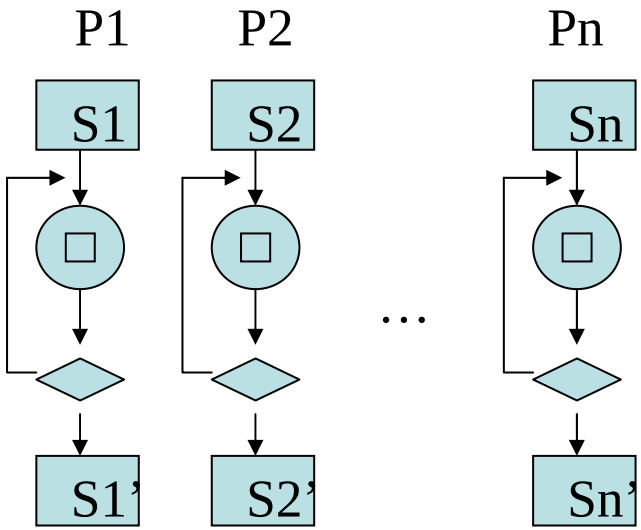
(a) **分割統治法**



(b) **プロセッサファーム**

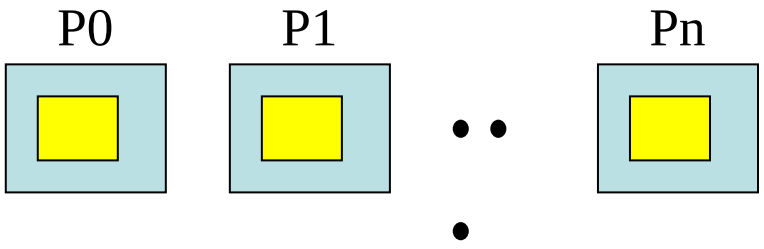


(c) **プロセスネットワーク**

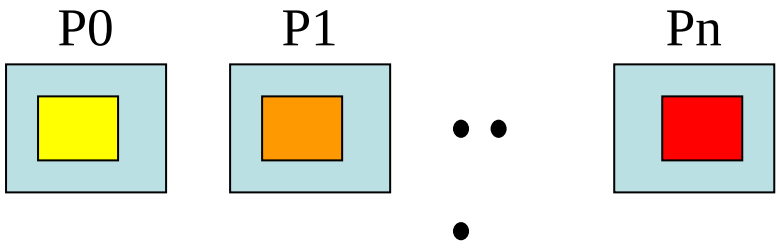


(d) **繰り返し変換**

SPMD vs. MPMD



Single Program Multiple Data

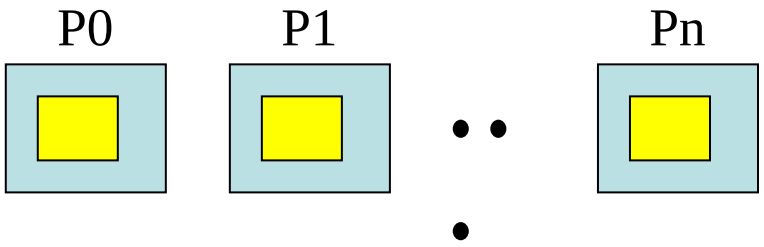


Multiple Program Multiple Data

- Single program runs on all PEs
- Instruction sequencing may be different

- Different programs run on different PEs
- Master & slave programs

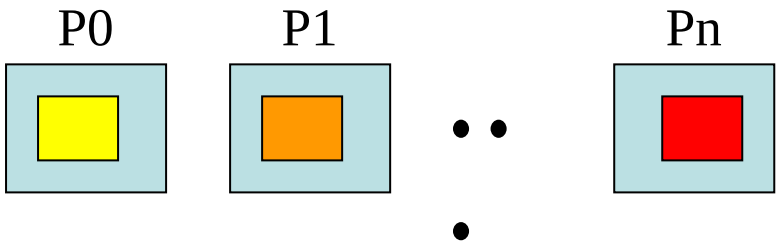
SPMD vs. MPMD



Single Program Multiple Data

単一プログラム複数データ

- **全PE上を1つのプログラムが走る**
- **命令の順序が異なる可能性もある**



Multiple Program Multiple Data

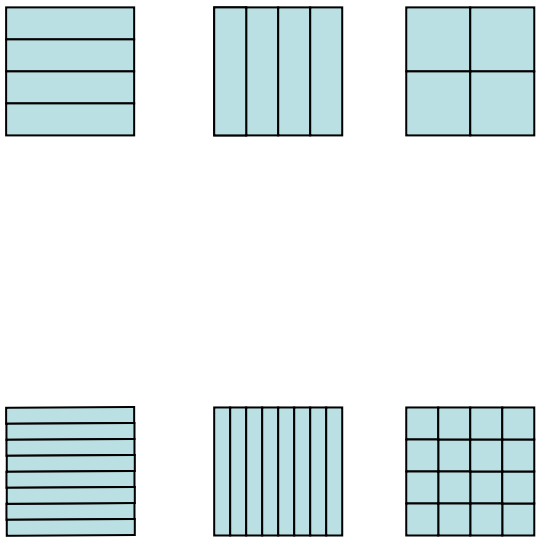
複数プログラム複数データ

- **各PE上を異なるプログラムが走る**
- **マスタ･スレーブプログラム**

Data Division

- Block division
  - Suitable for data parallelism
  - Divides the whole data into multiple equal sized blocks
  - number of blocks = number of PEs
- Cyclic division
  - Unequal loads on the data area
  - Divides the whole data into small blocks
  - Each small block is cyclically allocated to PEs

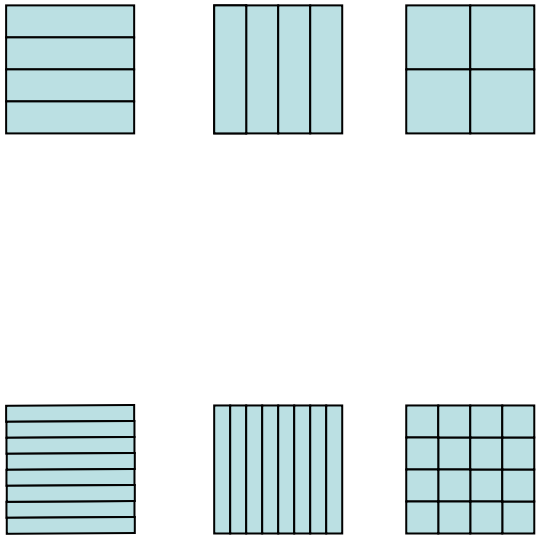
Four PEs



データ分割

- ブロック分割
  - データ並列性に適する
  - データ全体を複数の同じ大きさのブロックに分ける
  - ブロック数 = PE数
- サイクリック分割
  - 負荷が不均衡な場合
  - データ全体を小さなブロックに分ける
  - 各ブロックが各PEにサイクリックに割り当てられる

4台のPE



8. Application Fields

- Computational fluid dynamics
  - Atmospheric and oceanic modeling
  - Weather forecasting, pollution transport
  - Wind tunnel, airflow prediction
- Structural and filed analysis
  - Prediction of new semiconductor materials
  - Protein structure
- N-body problems
  - Evolution of galaxies comprising millions of stars
  - Astronomical simulation
  - Molecules, atoms and quarks

8. 応用分野

- 流体力学
  - 環境･海洋モデリング
  - 気象予測, 汚染の拡散
  - 風洞, 気流の予測
- 構造解析
  - 半導体材料の予測
  - たんぱく質の構造
- 多体システムモデリング
  - 無数の星からなる銀河系の進化
  - 天文シミュレーション
  - 分子、原子、クォーク

- High performance interfaces
  - High quality rendering
  - Image analysis, pattern recognition
- Very large databases
  - Data mining
  - Multi-media database
- Systems biology
  - Aims at system-level understanding of biological systems
  - Understanding of structure, dynamics, control methods and design ones of the system

- 高性能インターフェース
  - 高品質レンダリング
  - 画像解析、パタン認識
- 大規模データベース
  - データマイニング
  - マルチメディアデータベース
- システムバイオロジー
  - 生物をシステムとして理解することを目指す
  - システムの構成、ダイナミクス、制御方法、設計方法の理解

9. Grid Computing

- Distributed, high performance computing and data handling infrastructure
  - Geographically and organizationally dispersed, heterogeneous resources
  - Just as we plug into the electrical power network, we plug into the Internet/Intranet
- The layered Grid architecture
  - Applications
  - Programming tools/PSE:Problem Solving Environment
  - Common services
  - Fabric/infrastructure

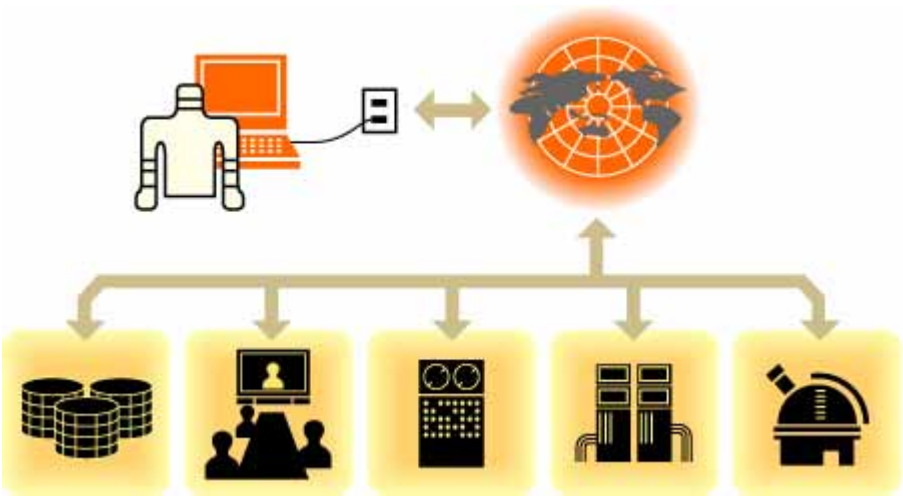
9. グリッドコンピューティング

- 分散･高性能コンピューティング／データ操作の基盤
  - 地理的･組織的に分散したヘテロなリソースを統合
  - 電力供給ネットとの類似
- グリッドの階層構造
  - 応用
  - プログラミングツール/PSE(Problem Solving Environment)：問題解決環境
  - 共通サービス
  - 構造/基盤

What is grid?

Grid is named for high-voltage power line network conducting electricity. (power grid)
The next-generation infrastructure to enjoy various information service through networks safely, stably and simply by only connecting with information outlets as we can get necessary electric power anytime and anywhere by plugging in.

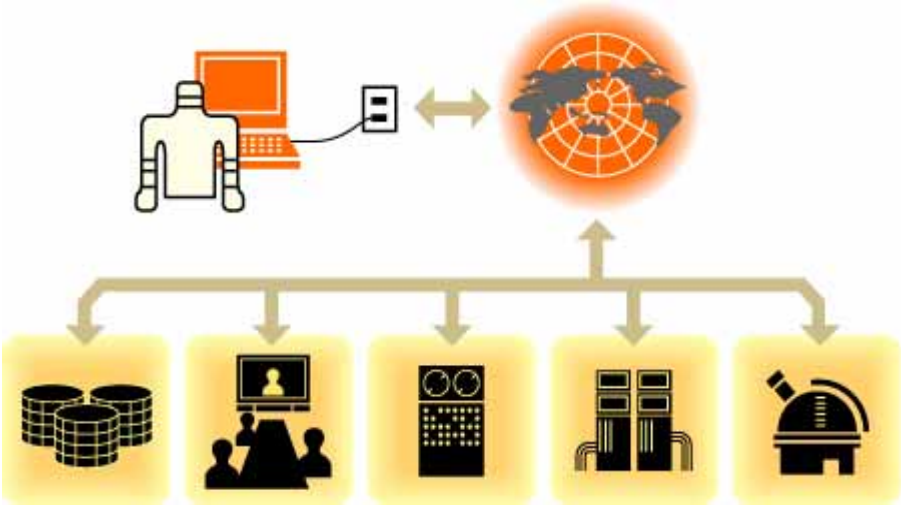
“ Safely “ care for the security
“ Stably “ possible to provide virtual resources when needed
“ Simply “ users can access information service without worrying about what happen over the network



Gridとは

グリッドとは、電気を伝える高圧送電線網（パワーグリッド）に由来しています。
コンセントに差し込めばいつでもどこでも必要なだけ電力が得られるように、情報コンセントに接続するだけで、ネットワークを通して、安全に・安定して・安易に様々な情報サービスを享受できるようにするための次世代インフラです。

「安全に」 セキュリティ面での配慮がある
「安定して」必要な時に必要なだけ仮想化された資源が提供可能
「安易に」ネットワークの向こうで起きていることをユーザは気にすることなく情報サービスへのアクセスが可能



http://www.jpgrid.org/about/index.html

Definition of grid

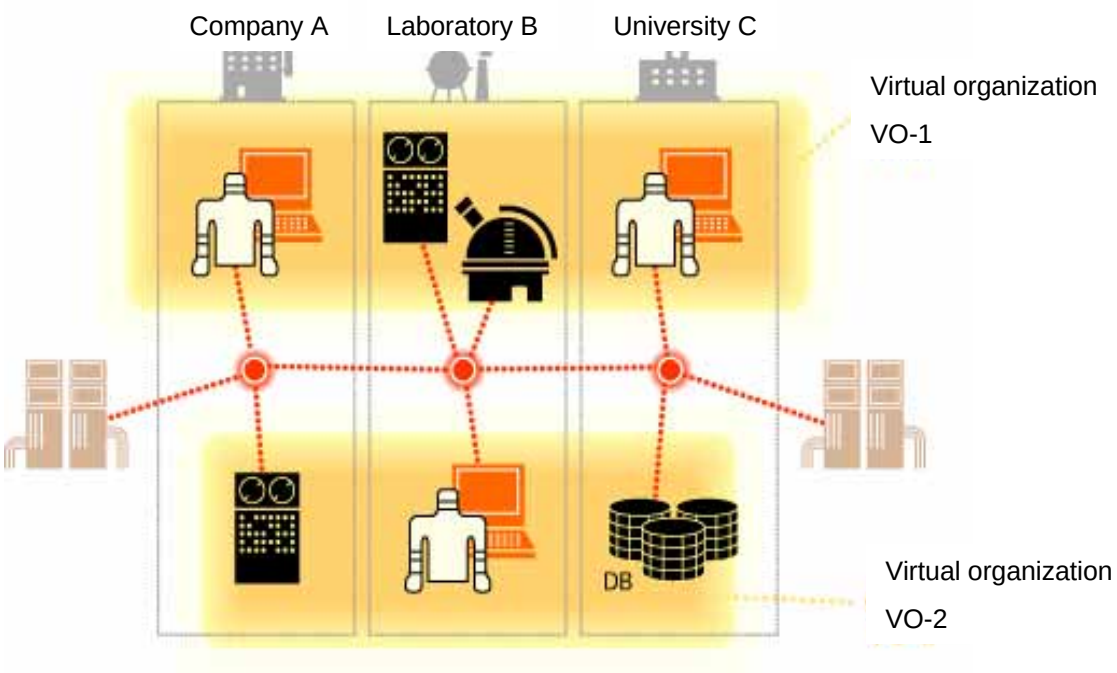
Grid is an infrastructure to build Virtual Computer and Virtual Organization dynamically if necessary by virtualizing and integrating resources such as calculations, data, experimental devices, sensors and human beings on wide area networks.

The checklist of grid

1： coordinate the distributed resources out of central control

2： use the open standard protocols and interfaces

3： provide the high quality service which could not get simply

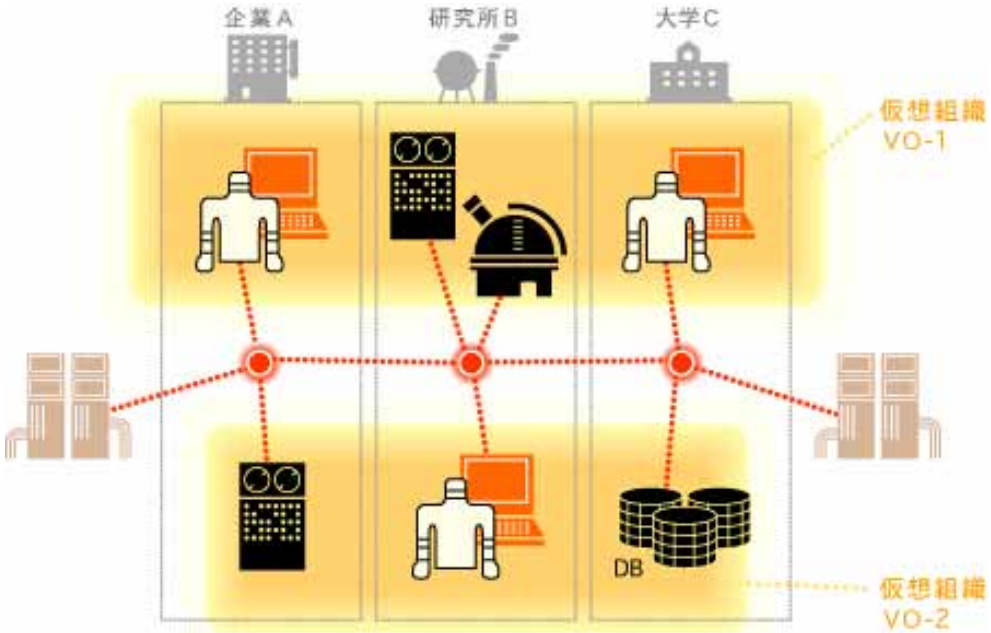


Gridの定義

グリッドは広域ネットワーク上の計算、データ、実験装置、センサー、人間などの資源を仮想化･統合し、必要に応じて仮想計算機 (Virtual Computer) や仮想組織 (Virtual Organization) を動的に形成するためのインフラです。

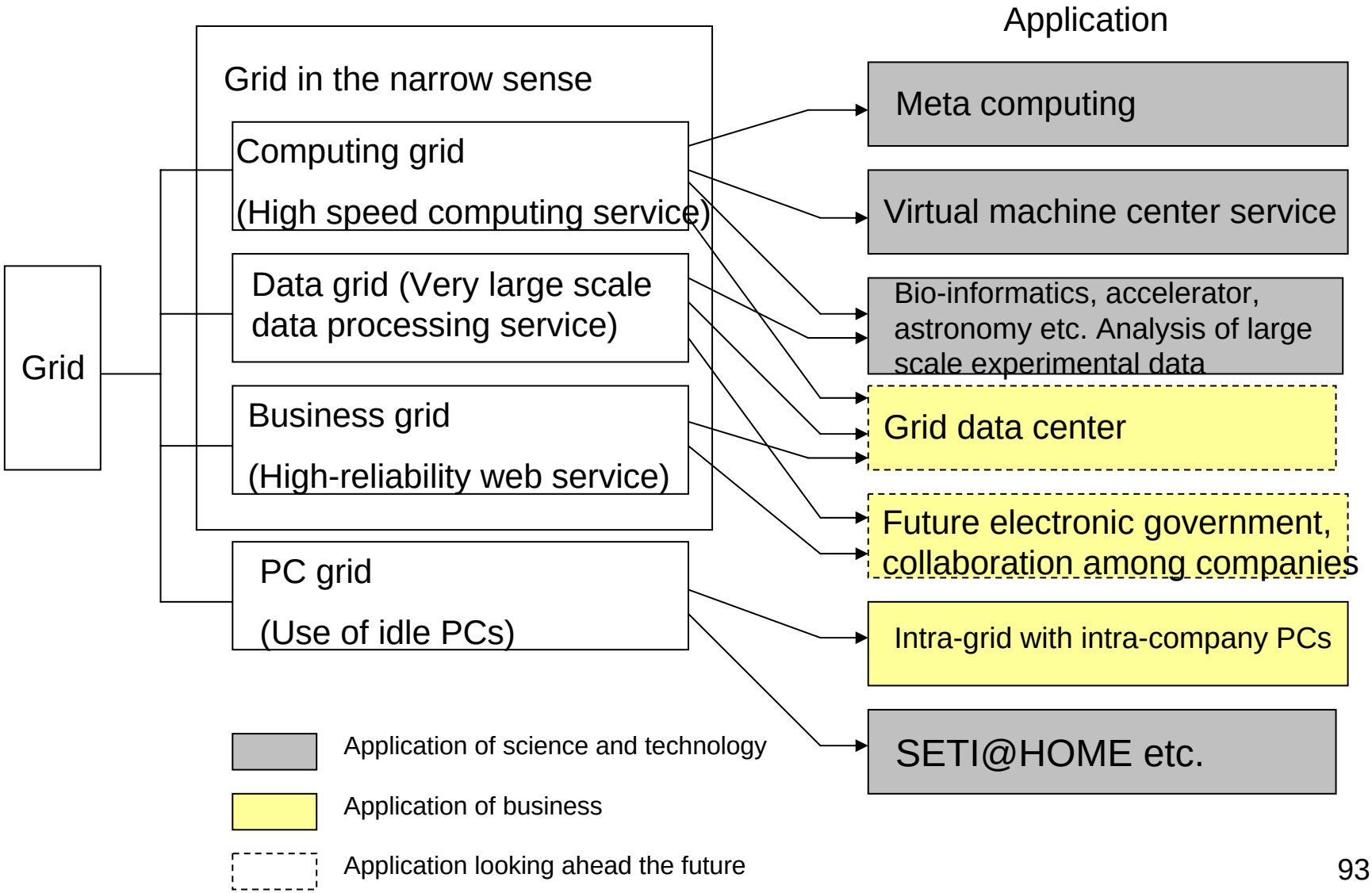
グリッドのチェックリスト

1： 集中管理されていない分散した資源のコーディネート
2： オープンスタンダードなプロトコルやインターフェースの利用
3： 単純には得られない質の高いサービスの提供

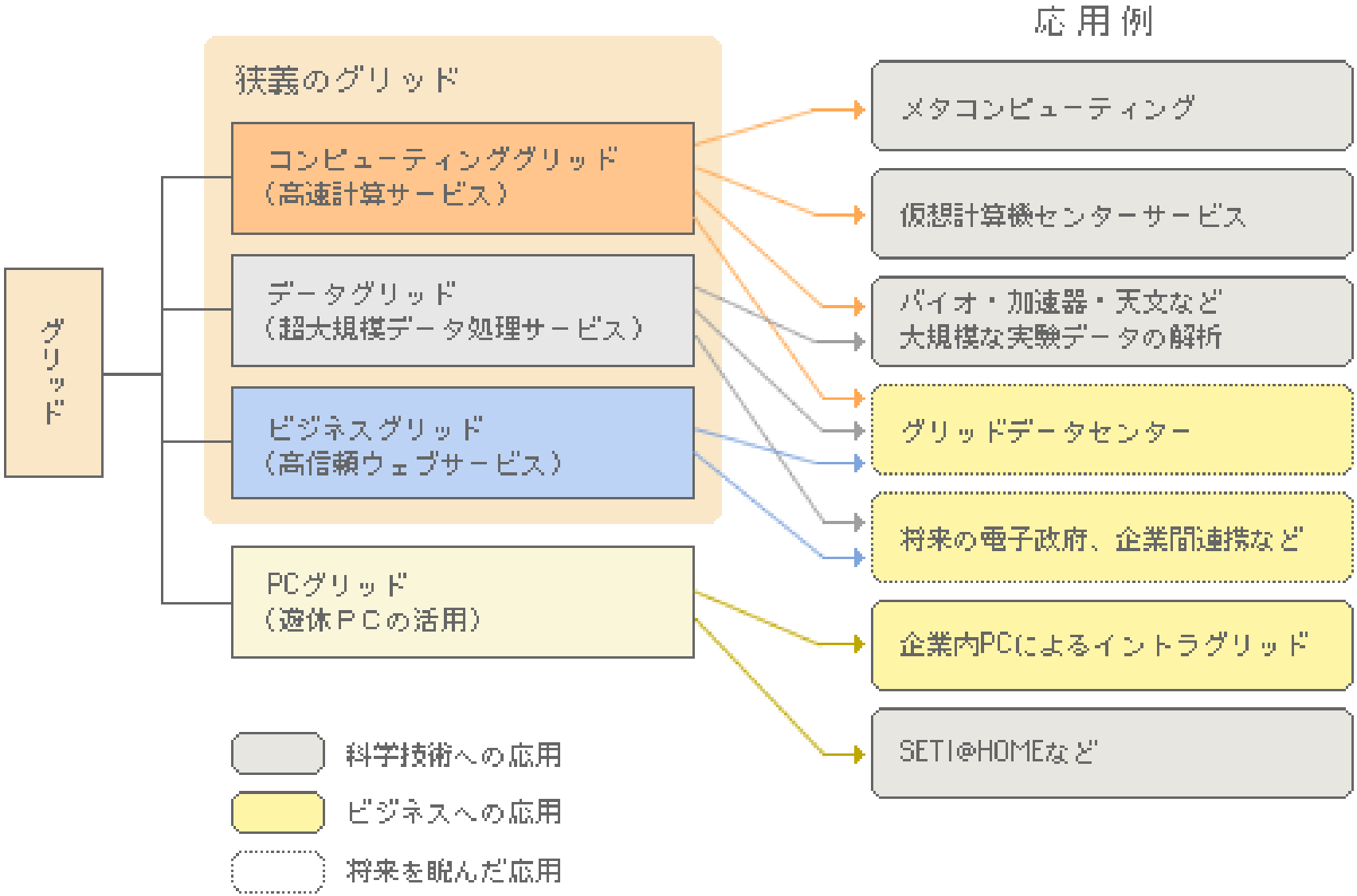


http://www.jpgrid.org/about/index.html

Classification of grid



Gridの分類



http://www.jpgrid.org/about/index.html

- Applications
  - Data-intensive, on-demand, distributed, collaborative & high-throughput computing
- Programming tools/PSE
  - Language, compiler, visualization, remote service/communication library, PSE
- Common services
  - Scheduler, security, remote data access, resource management, information, communication, fault tolerance
- Fabric/infrastructure

- 応用
  - データ集約、オンデマンド、分散、協調、高スループットコンピューティング
- プログラミングツール/PSE
  - 言語、コンパイラ、可視化、リモートサービス/通信ライブラリ、PSE
- 共通サービス
  - スケジューラ、セキュリティ、リモートデータアクセス、リソース管理、 情報、通信、フォールトトレランス
- 構造/基盤

- Development, preparation & standardization of middleware are the center of Grid activity
  - Globus Toolkit
- Problems of Grid
  - Network infrastructure, security, scheduling, load balancing
- Adaptability to applications
  - Multi-discipline, embarrassingly parallel
  - Extension of PC clusters & MPPs
- Examples

- ミドルウェアの開発･整備･標準化がグリッドの活動の中心
  - Globus Toolkit
- グリッドの課題
  - ネットワーク･インフラ、セキュリティ、スケジューリング、負荷均衡
- アプリケーションとの適合性
  - 学際的、驚異的な並列性
  - PCクラスタと超並列マシンの拡張
- 例

10. Summary

Parallel computing is to solve very large scale problems within a reasonable time using the fastest machines at that time. Research fields cover a wide range of issues such as parallel machines, parallel programming, parallel algorithms and applications.

Future issues

- Parallel programming software
- Parallel programming education
- Collaboration between researchers on parallel computing and the target application fields

10. まとめ

並列コンピューティングとは最大性能のマシンで大規模な問題を高速に（実用的な時間内で）解くこと。並列マシン、並列プログラミング、並列アルゴリズム、各種応用など、研究範囲は多岐にわたる。

今後の課題

- **並列プログラミング用ソフトウェア**
- **並列プログラミング教育**
- **並列コンピューティングと対象アプリケーションの研究者の協力**

References

(1)Introduction to Parallel Computing
http://www.llnl.gov/computing/tutorials/parallel_comp/#Designing

(2)ＰＣクラスタ超入門、超並列計算研究会、2000.
http://mikilab.doshisha.ac.jp/dia/smpp/cluster2000.

(3)B.Wilkinson and M.Allen: Parallel Programming, Prentice-Hall, 1999.
飯塚、緑川　訳：並列プログラミング入門、丸善、2000.

(4)B.Wilkinson and M.Allen: Parallel Programming, 2nd edition, Prentice-Hall, 2004.

(5)湯浅、安村、中田：はじめての並列プログラミング、bit別冊、共立出版、1998.

(6)HPCCプロジェクト、アメリカ、1992～　http://www.nitrd.gov/

(7)PAPIA: Parallel Protein Information Analysis system
http://mbs.cbrc.jp/papia/papiaJ.html.

(8)佐藤三久：Omni OpenMPコンパイラとCluster-enabled OpenMP、京大大型計算機センター第66回研究セミナー,2001.

(9)B.M.Barney: Tutorial: Introduction to OpenMP Programming, 2000.
http://www.llnl.gov/computing/tutorials/openMP/

文献

(1)Introduction to Parallel Computing
http://www.llnl.gov/computing/tutorials/parallel_comp/#Designing

(2)ＰＣクラスタ超入門、超並列計算研究会、2000.
http://mikilab.doshisha.ac.jp/dia/smpp/cluster2000.

(3)B.Wilkinson and M.Allen: Parallel Programming, Prentice-Hall, 1999.
飯塚、緑川　訳：並列プログラミング入門、丸善、2000.

(4)B.Wilkinson and M.Allen: Parallel Programming, 2nd edition, Prentice-Hall, 2004.

(5)湯浅、安村、中田：はじめての並列プログラミング、bit別冊、共立出版、1998.

(6)HPCCプロジェクト、アメリカ、1992～　http://www.nitrd.gov/

(7)PAPIA: Parallel Protein Information Analysis system
http://mbs.cbrc.jp/papia/papiaJ.html.

(8)佐藤三久：Omni OpenMPコンパイラとCluster-enabled OpenMP、京大大型計算機センター第66回研究セミナー,2001.

(9)B.M.Barney: Tutorial: Introduction to OpenMP Programming, 2000.
http://www.llnl.gov/computing/tutorials/openMP/

(10)石川　裕他：Linuxで並列処理をしよう、共立出版、2002.
(11)Ian Foster: Designing and Building Parallel Programs, Addison Wesley, 1995.
(12)D.E.Culler and J.P.Singh: Parallel Computer Architecture A Hardware/Software Approach, Morgan Kaufmann, 1999.
(13)Rajkumar Buyya: High Performance Cluster Computing, Vol.1 & 2, Prentice Hall PTR, 1999.
(14)J.L.Hennessy and D.A.Patterson: Computer Architecture A Quantitative Approach, third edition, Morgan Kaufmann, 2003.
(15) I. Foster and C. Kesselman, The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 1999.
(16) I. Foster and C. Kesselman, The Grid 2: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 2004.
(17) 特集　グリッドコンピューティング、情報処理、vol.44, no.6, 2003.
(18) D.A.Patterson and J.L.Hennessy: Computer Organization and Design -The hardware/software interface-, third edition, Morgan Kaufmann, 2005.

(10)石川　裕他：Linuxで並列処理をしよう、共立出版、2002.
(11)Ian Foster: Designing and Building Parallel Programs, Addison Wesley, 1995.
(12)D.E.Culler and J.P.Singh: Parallel Computer Architecture A Hardware/Software Approach, Morgan Kaufmann, 1999.
(13)Rajkumar Buyya: High Performance Cluster Computing, Vol.1 & 2, Prentice Hall PTR, 1999.
(14)J.L.Hennessy and D.A.Patterson: Computer Architecture A Quantitative Approach, third edition, Morgan Kaufmann, 2003.
(15) I. Foster and C. Kesselman, The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 1999.
(16) I. Foster and C. Kesselman, The Grid 2: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 2004.
(17) 特集　グリッドコンピューティング、情報処理、vol.44, no.6, 2003.
(18) パターソン、ヘネシー、成田　訳：コンピュータの構成と設計、ハードウェアとソフトウェアのインタフェース、第３版、日経BP社、2006.