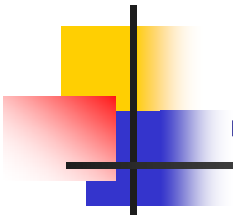
平成15年度

# 並列分散システム論
## Part III

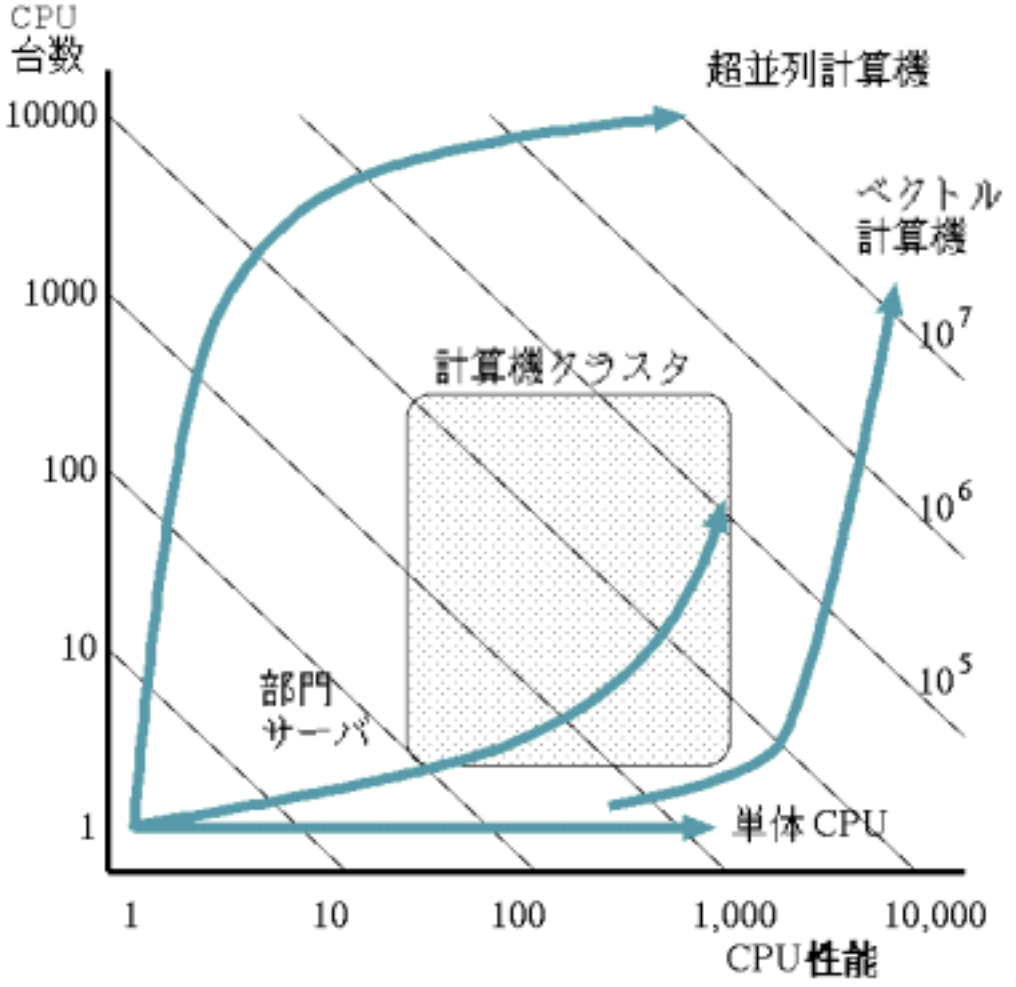
計算機システム講座
計算機アーキテクチャ分野
森眞一郎



# 並列･分散システムの実現

- **計算機クラスタ**
- **分散共有メモリシステム**
- Grid (??)

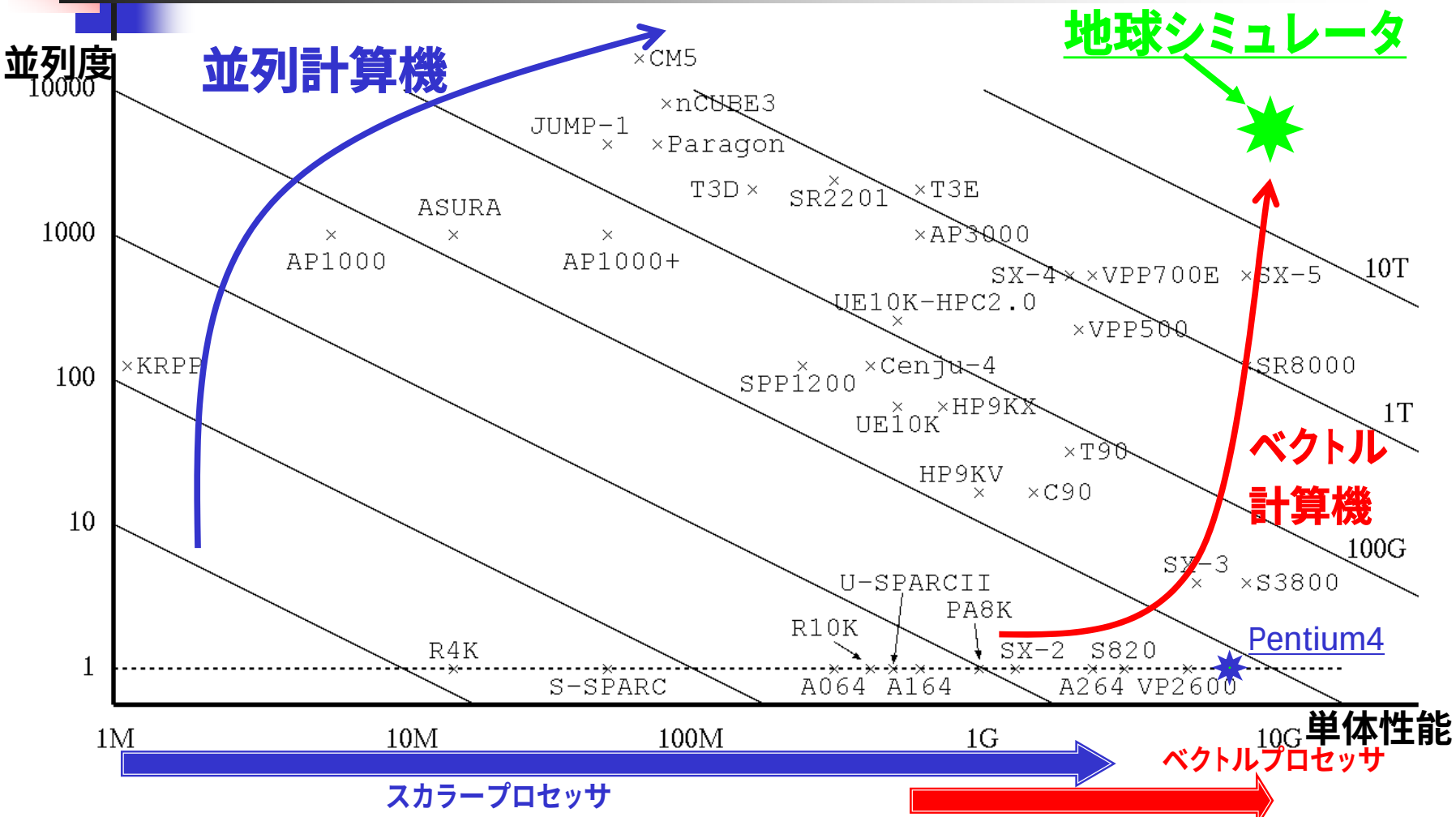
# 計算機クラスタの位置付け



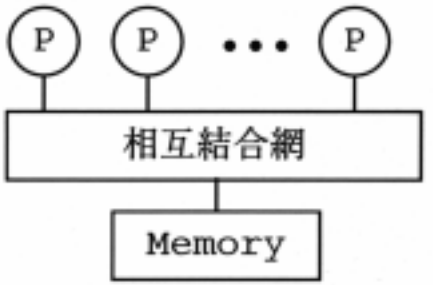
# 「計算機クラスタ」の追い風は？

- CPUの高速化/低価格化…… Alpha/Pentium (PCの追い上げ)
  - 8GFlopsの壁???……Alpha(612MHz)の6倍，UltraSPARC(336MHz)の12倍
    Pentium4(3.06GHz)の1.5倍
    9.6GF??
- 通信媒体の高速化
  - 2Gpbsの壁??? ….多重化 (波長,位相,振幅) ， 並列化，光のWave Pipeline
- 標準通信ライブラリの普及
  - Posix Thread, OpenMP, HPF, MPI, AM, FM, PM, …..
- 米国 (だけ??) での並列計算機メーカの不振……(最近は??….CMPで挽回？)
  - 地球シミュレータ…… 5120プロセッサ，40TF (2002.03@海洋科学技術センタ 横浜)
  - Cray Inc. …… 単体性能12.8GFlopsのベクトルプロセッサを発表(2002.11)
    - 3.2GFlops のスカラプロセッサ4台＋ベクトル処理向けに強化した同期機構
      ⇒仮想ベクトル処理
    - 4096プロセッサで52.4TFlops (未稼働)、 2010年までに 1 PetaFlops を目指す
  - Blue Planet …. 2005年を目処にIBM Power5(10GF) x 16,384 =150TFlops

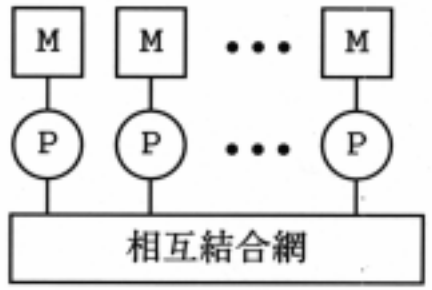
# いつまで続くCPUの高速化



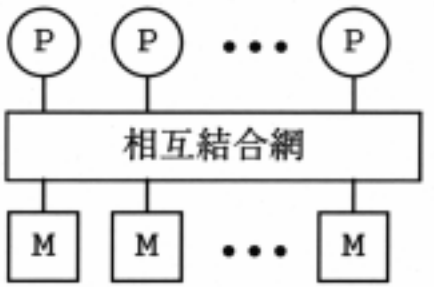
# 並列計算機の分類



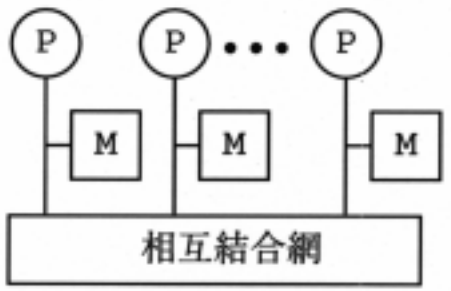

(a)集中メモリ型
(UMA or SMP)




(c)分散メモリ型
(NORA)



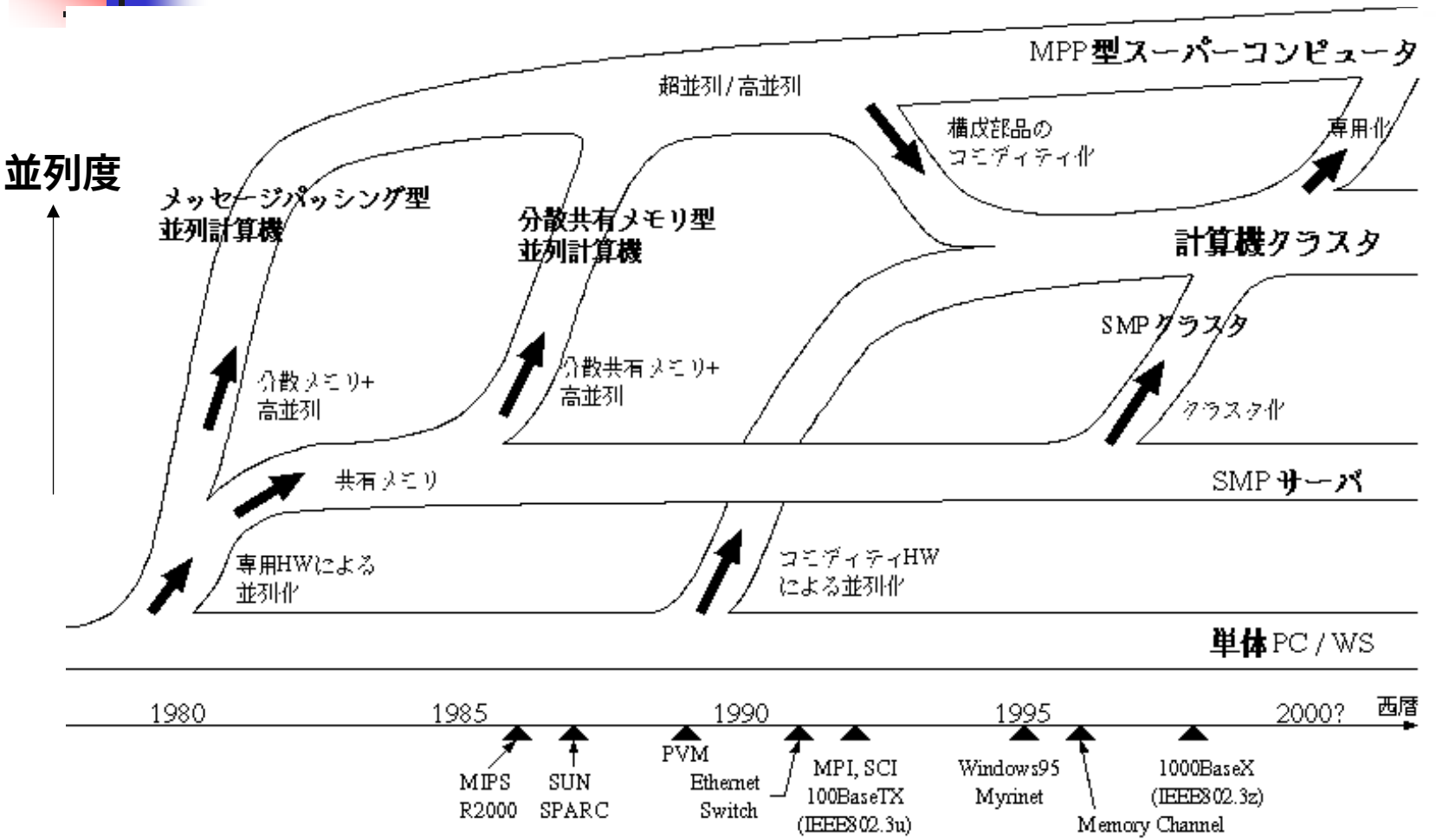

(b)分散メモリ型
(UMA/NUMA)



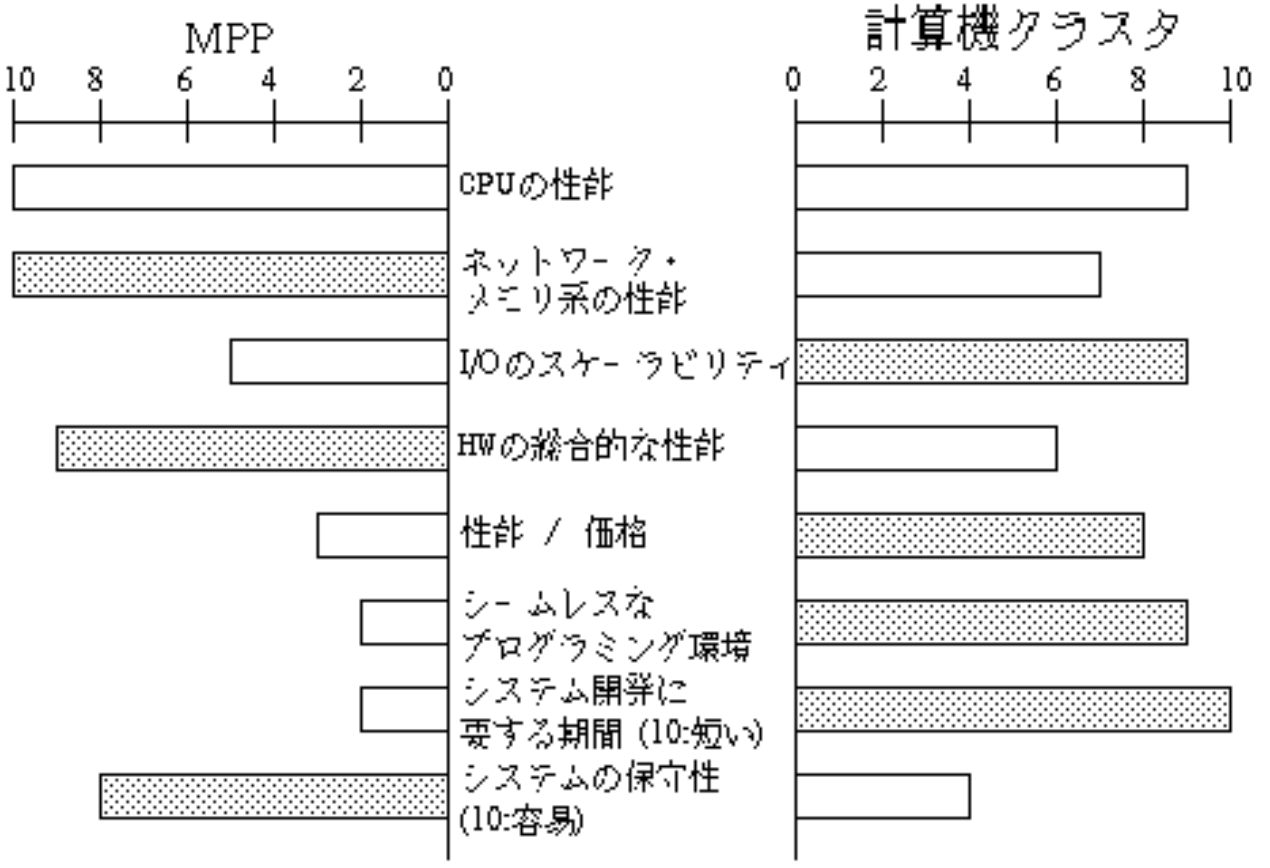

(d) 分散メモリ型
(NUMA)

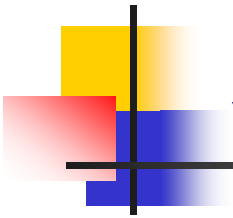
P: Processor, M: Memory

# 計算機クラスタへの道程



# MPPと計算機クラスタの比較





# ネットワークの急速な高速化

- 通信媒体自体の高速化
- 通信ソフトウェアの高速化

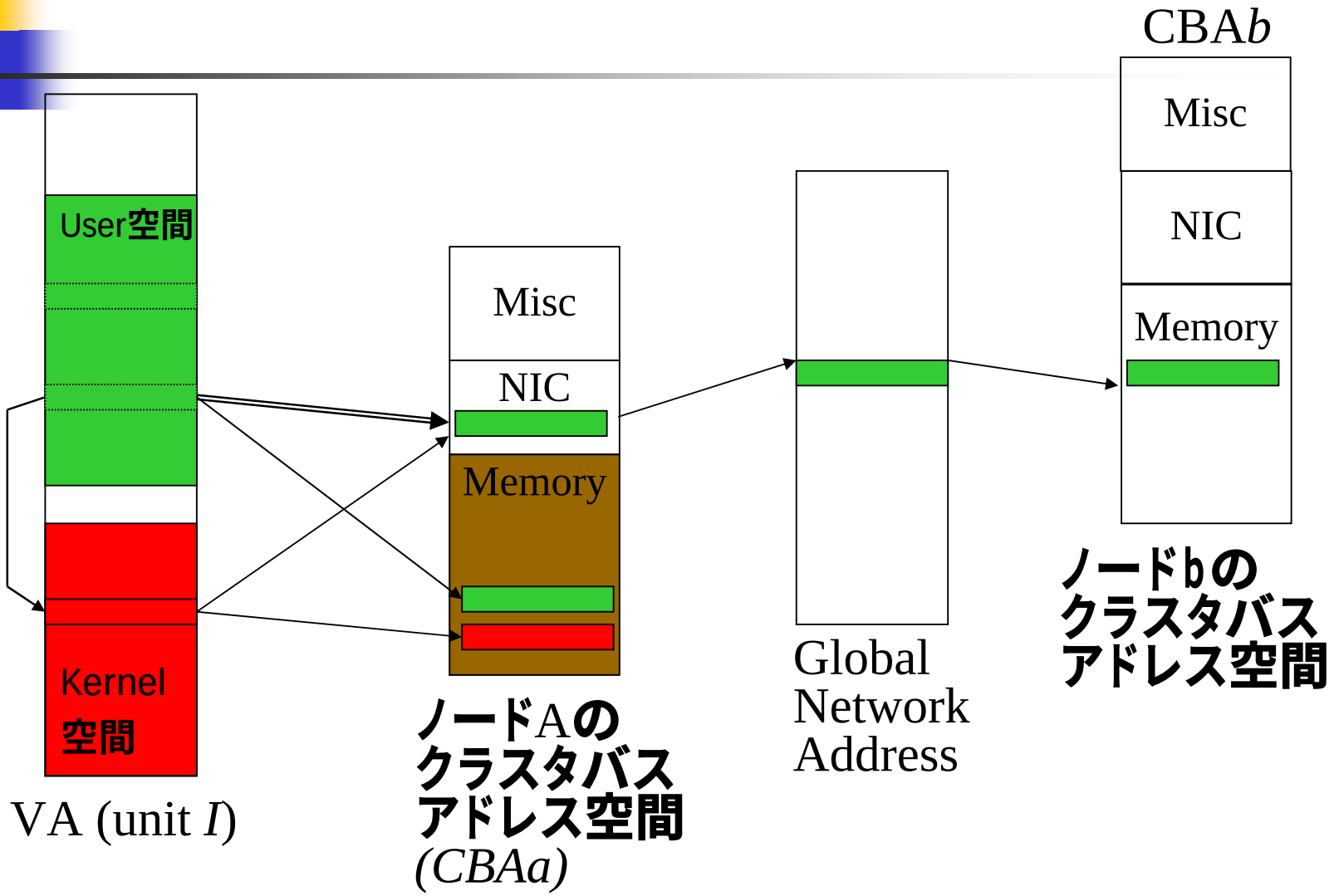
# 通信媒体自体の高速化

- **汎用の通信媒体として**
  - 10Base2/5 10BaseT 100BaseT Gbit Ether
    - **バス型結線 スター型結線 …… 相互干渉の軽減, 高周波設計の容易さ**
    - **電気 光 …… 物理的な距離に関する制約の緩和**
  - USB(Universal Serial Bus), IEEE1394**バス**
    - **家庭内ネットワークコンピューティングへの足掛かりになるか??**
- **専用の超高速通信媒体として**
  - Myrinet(Myricom**社**) **通信用コプロセッサを搭載**
  - Memory Channel(DEC) **仮想共有メモリ環境を提供する** PCI **バス用**
  - SCI (IEEE1596) CC-NUMA**環境**
  - Synfinity NUMA(Fujitsu) CC-NUMA**環境** + 1.6Gbps
  - QsNet (Quadrix) VA-to-VA Remote DMA, 340MB/s

# 通信ソフトウェアの高速化

- **0コピー通信**
- remote read/write (put/get)
- **軽量化プロトコル**

# Zero-copy通信


CBAb
Misc
NIC
Memory
User空間
Misc
NIC
Memory
Kernel
空間
ノードbの
クラスタバス
アドレス空間
Global
Network
Address
ノードAの
クラスタバス
アドレス空間
(CBAa)
VA (unit I)

# Myrinet

- USC/ISI+Caltec**の共同開発**(ATOMIC) **の** commercial version
  - **専用通信プロセッサ** LANai
  - **高速パケット交換スイッチ** (2Gbps + 2Gbps)x2@MyrinetXP
- Application Programming Interface **や** LANai chip **仕様 の公開**
  - **多くの** Hacker**達が競って最適化！！**
  - UC Berkeley(NOW), Illinois(Fast Messages), RWCP(PM)

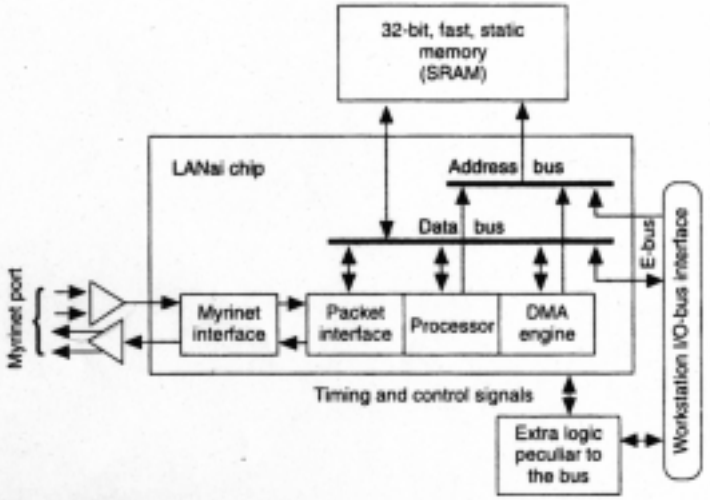



Fig6 通信ボードのブロック図 (文献 [1])

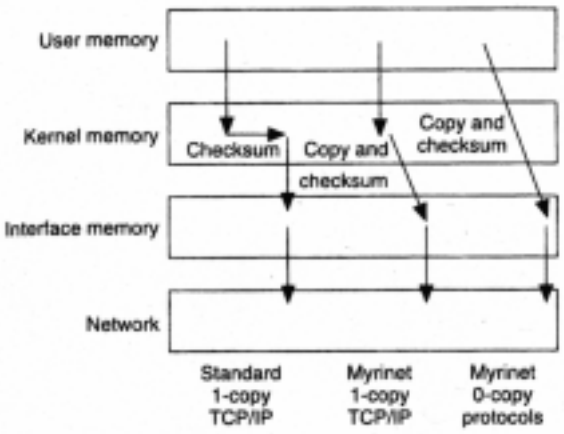
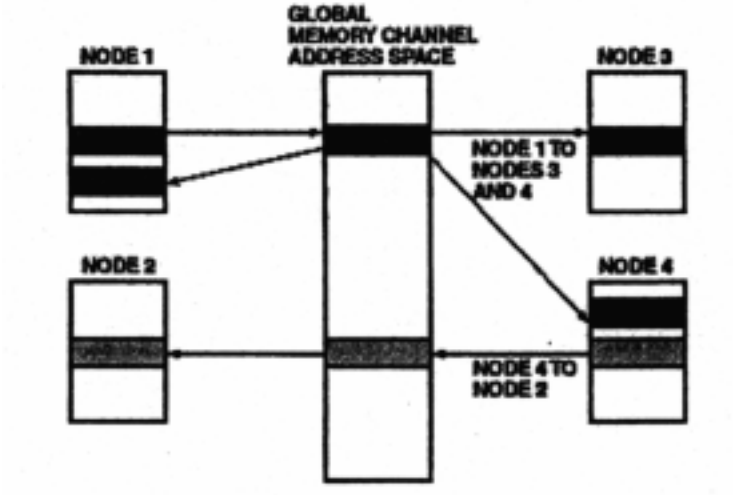



Fig7 Zero-copy プロトコル (文献 [1] より引用)

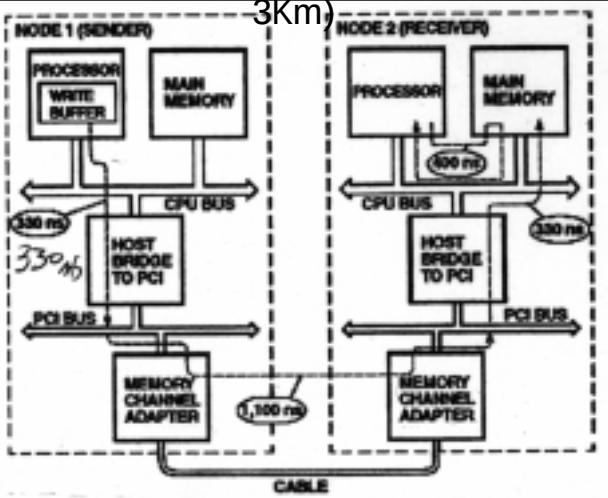
# Memory Channel

- **仮想共有メモリ環境を提供する** PCI **バス用通信ボード**
- **異なるノード上のプロセス間通信が**, load/store **で実現可能**

| **世代** | **転送速度**<br>**(ユーザプロセス間**sustained) | **遅延**<br>(store to load) | **ハブの構成**<br>(cabling) |
|---|---|---|---|
| **第一**(1996～ ) | 77MB/s (66MB/s) | 2.9μs | **バス構造(電線**4m) |
| **第二**(1997～ ) | 97MB/s (88MB/s) | 2.2μs | **スイッチ(電線**10m or **光** 3Km) |




**クラスタワイドなアドレスマッピング**




Latency Contribution

# 汎用ネットワークの追い上げ

## 超並列計算機のネットワーク性能

| 製品名 | メーカ | 発表年 | Link速度 |
|---|---|---|---|
| CM-5 | T.M. | 1992年 | 1.06Gbps |
| Paragon | Intel | 1992年 | 1.6Gbps |
| T-3D | Cray | 1993年 | 1.2Gbps |
| T-3E | SGI | 1995年 | 4.8Gbps |
| SR2201 | 日立 | 1996年 | 2.4Gbps |
| VPP700E | 富士通 | 1997年 | 4.8Gbps |

| 製品名 | メーカ | 発表年 | Link速度 |
|---|---|---|---|
| SR8000 | 日立 | 1998年 | 1.0GB/s |
| SX-5 | NEC | 1998年 | 8(?)GB/s |
| Cenju-4 | NEC | 1998年 | 0.8GB/s |
| VPP800 | 富士通 | 1999年 | 1.6GB/s |
| AsamA | NEC | 2002年 | 12.8GB/s |
| 地球Sim | NEC | 2002年 | 12.3GB/s |

## ギガビット級ネットワーク@2003

| 名称 | 企画/企業 | Link速度 | 通信の信頼性@物理層 |
|---|---|---|---|
| Fibre Channel | ANSI X3 T11 | 1.06Gbps | Reliable |
| SCI | ANSI/IEEE 1596 | 1.6Gbps | Reliable |
| Memory Channel2 | HP (DEC) | 1.06Gbps | Reliable |
| Myrinet | Myricom | 2.56Gbps | Reliable |
| Gigabit Ethernet | IEEE 802.3z | 1.25Gbps | Unreliable |
| Synfinity | Fujitsu | 1.6Gbps | Reliable |

# 汎用ネットワークの追い上げ

## 超並列計算機のネットワーク性能

| 製品名 | メーカ | 発表年 | Link速度 |
|---|---|---|---|
| CM-5 | T.M. | 1992年 | 1.06Gbps |
| Paragon | Intel | 1992年 | 1.6Gbps |
| T-3D | Cray | 1993年 | 1.2Gbps |
| T-3E | SGI | 1995年 | 4.8Gbps |
| SR2201 | 日立 | 1996年 | 2.4Gbps |
| VPP700E | 富士通 | 1997年 | 4.8Gbps |

| 製品名 | メーカ | 発表年 | Link速度 |
|---|---|---|---|
| SR8000 | 日立 | 1998年 | 1.0GB/s |
| SX-5 | NEC | 1998年 | 8(?)GB/s |
| Cenju-4 | NEC | 1998年 | 0.8GB/s |
| VPP800 | 富士通 | 1999年 | 1.6GB/s |
| AsamA | NEC | 2002年 | 12.8GB/s |
| 地球Sim | NEC | 2002年 | 12.3GB/s |

## ギガビット級ネットワーク@２００４

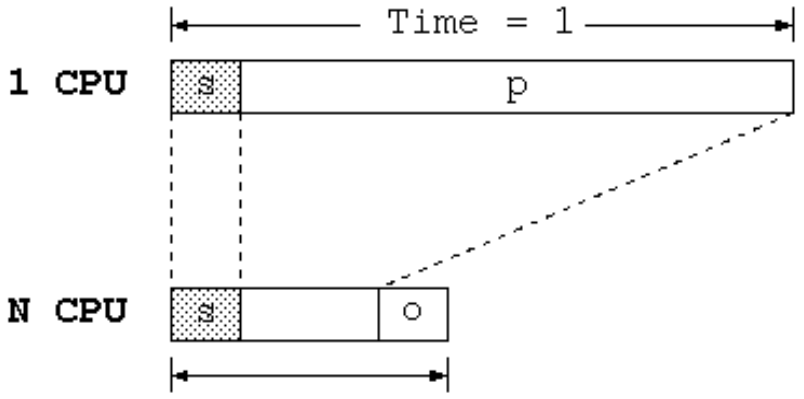
| 名称 | 企画/企業 | Link速度 | 通信の信頼性@物理層 |
|---|---|---|---|
| Fibre Channel | ANSI X3 T11 | 2.12Gbps | Reliable |
| SCI | ANSI/IEEE 1596 | 1.6Gbps | Reliable |
| Memory Channel2 | HP (DEC) | 1.06Gbps | Reliable |
| Myrinet | Myricom | 2.56Gbps | Reliable |
| 10Gigabit Ethernet | IEEE 802.3ae | 10Gbps | Unreliable |
| (参考)VisA PRO Link | 富田研 | 10Gbps | Reliable/Unreliable |

総務省の資料2003/05/07 によると数年後には40GbE/100GbEも登場！！

# プロセッサ間通信の遅延

| NIC | Latency | RATIO | I/F |
|---|---|---|---|
| DIMMnet-1 | 225ns(??) | 1 | PC133DIMM |
| ELAN | 2us | 10 | 66M64bPCI |
| Memory Channel | 2.2us | 10 | 33M32bPCI |
| SCI | 2.3us | 10 | 66M64bPCI |
| VIA on CLAN1000 | 3.5us | 16 | 33M64bPCI |
| PM on Myrinet | 7.5us | 33 | 33M32bPCI |
| GM on Myrinet2000 | 7.6us | 33 | 66M64bPCI |
| PM on GbE | 24.1us | 107 | 33M32bPCI |
| AsamA**カスタム**LSI | 300ns | 1 | ItaniumII bus |

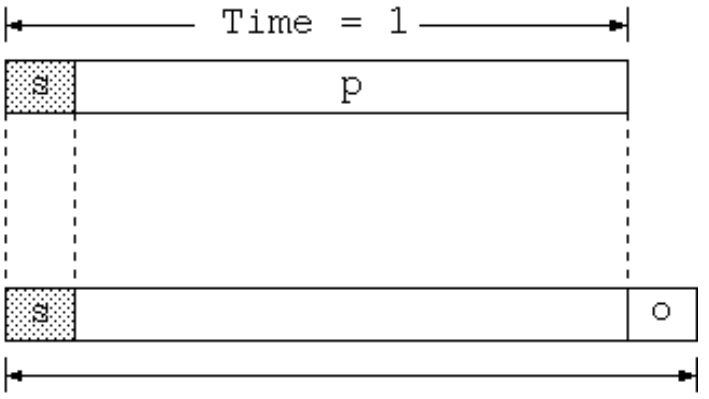
# スピードアップに関する2つの視点




Time = s + p/N + o

$$\text{Speeup} = \frac{1}{s + p/N + o}$$

(a) 問題サイズ固定



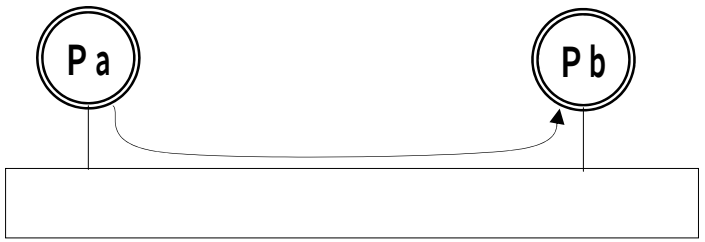

Time = s + p + o

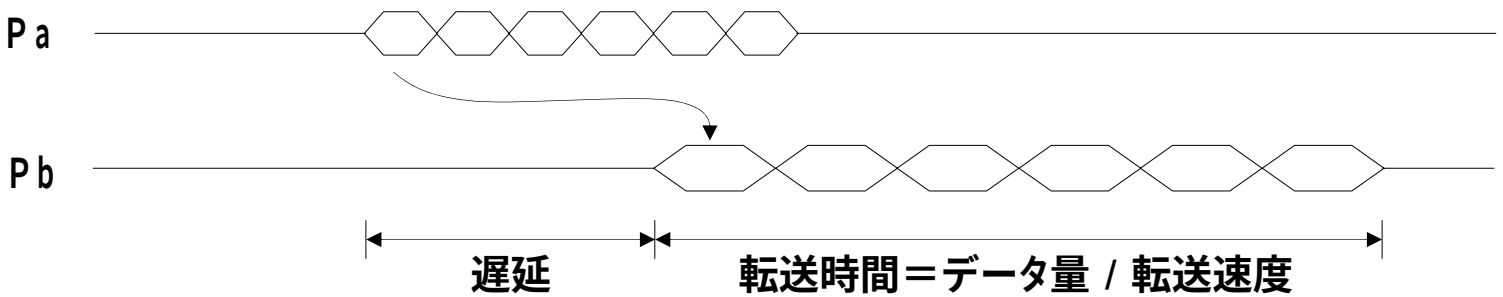
$$\text{Speeup} = \frac{1}{1 + o} \times N$$

(b) 問題サイズ N 倍

# 並列処理オーバヘッド

## (1) 通信オーバヘッド

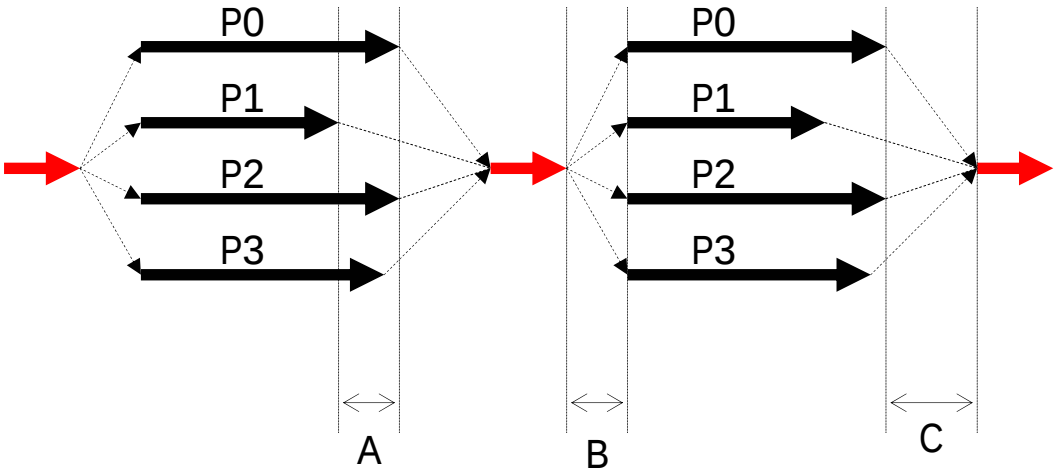



[技術的背景] 転送速度の向上は著しいが、遅延時間の短縮は極めて困難

# 並列処理オーバヘッド

## (2) 負荷のアンバランスと同期オーバヘッド




A: **負荷のアンバランスによるオーバヘッド**

B＋C: **同期オーバヘッドの例**

# 同期操作のオーバヘッド削減

- **スピンロック**
  - Test and Test&Set……**同期待ちトラフィックの軽減**

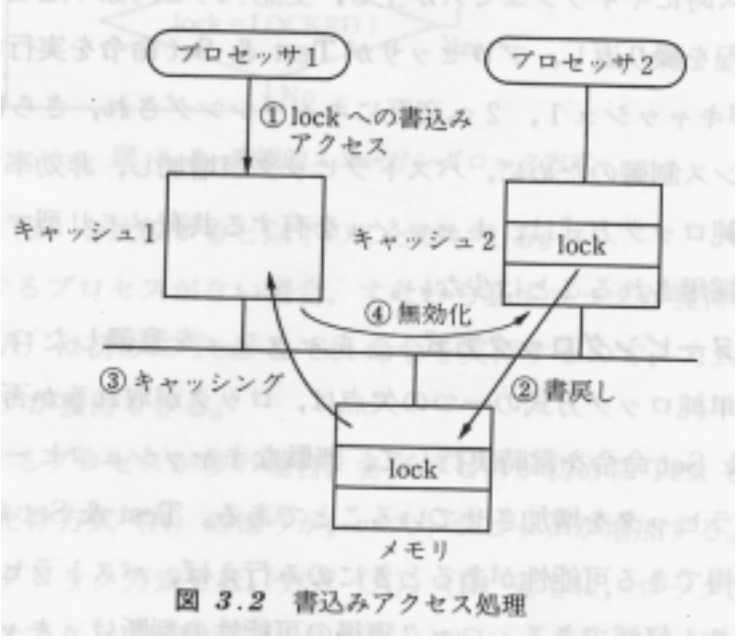


図 3.2 書込みアクセス処理

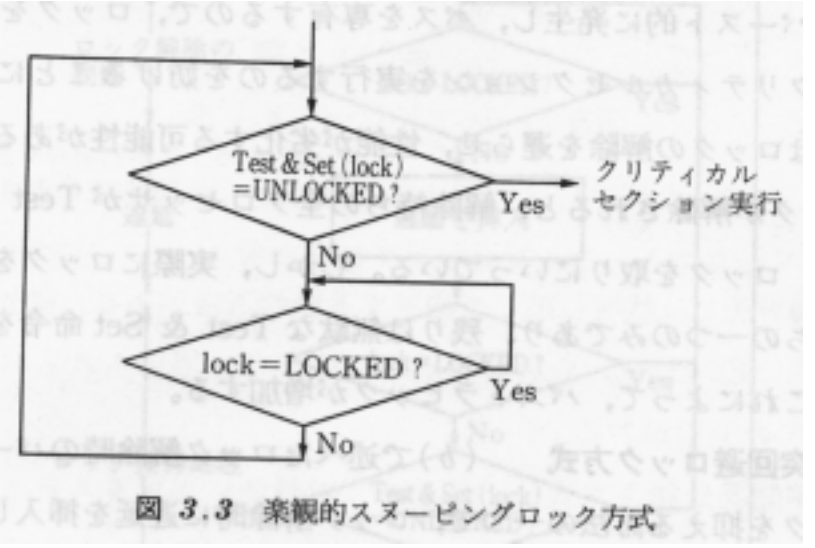


図 3.3 楽観的スヌーピングロック方式

# 集中型バリア同期機構の例

★カウンタベース

★再利用可能

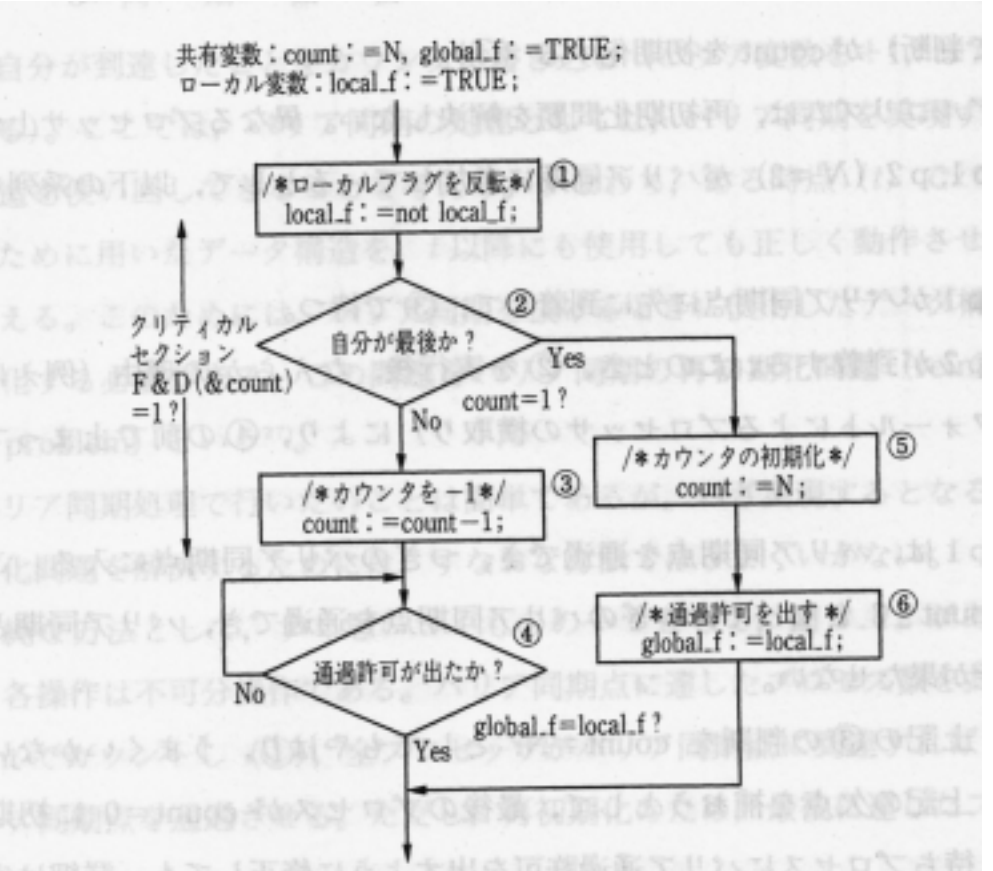



図 3.13 集中型バリア方式（反転方式）

# 分散バリア同期機構の例

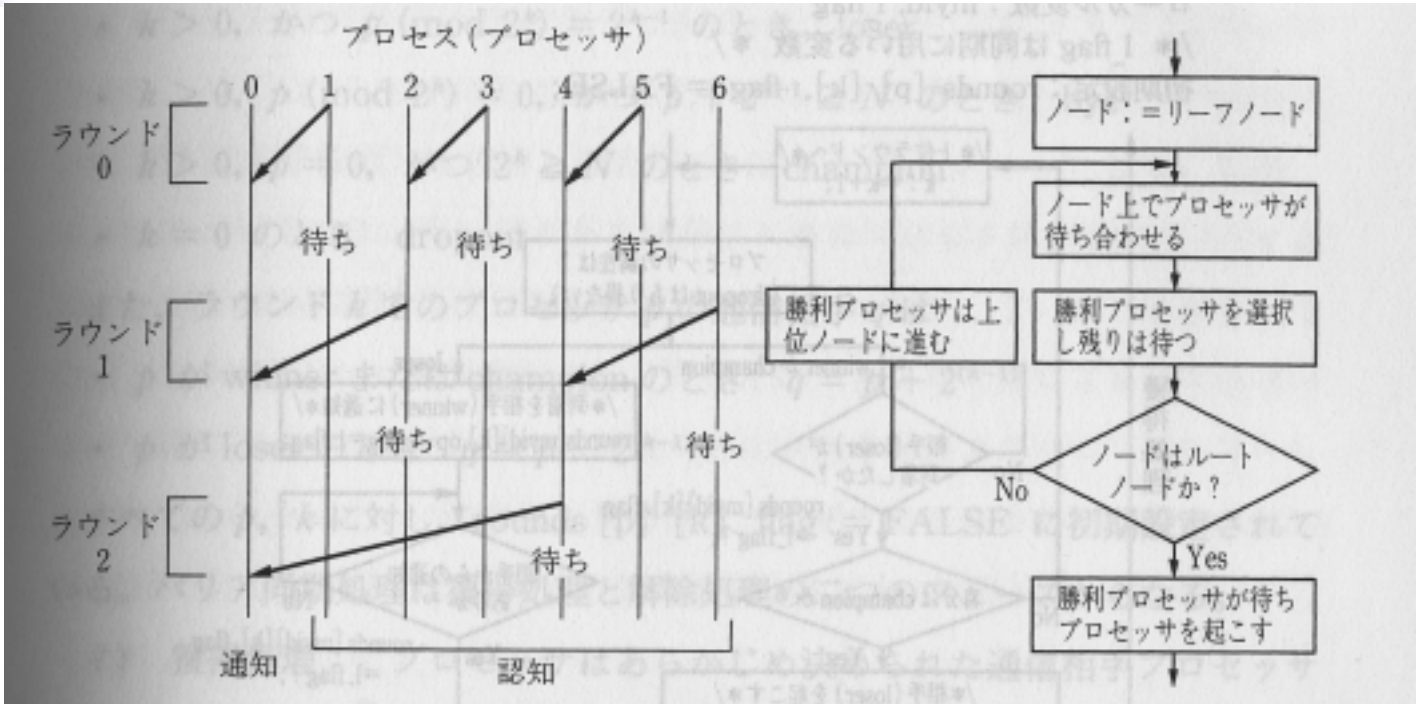



図 3.16 トーナメントバリア方式における通信パターン例 ($N=7$)



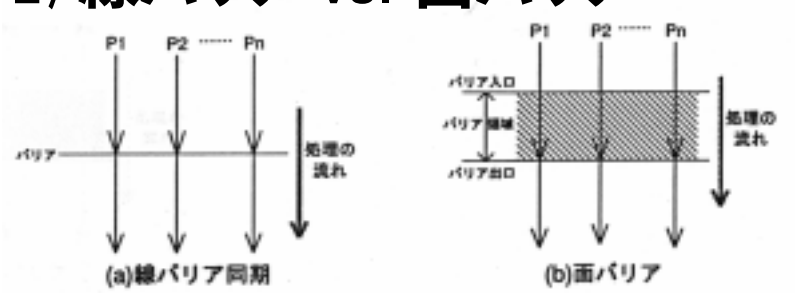

図 3.17 トーナメントバリア方式の概要

# バリア同期モデルの分類

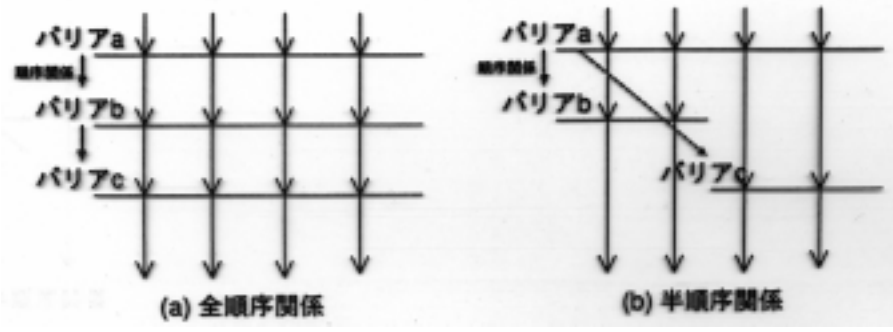
| 分類項目 | | | | バリア同期モデル |
|---|---|---|---|---|
| 線／面 | 強制参加／任意参加 | 全順序関係／半順序関係 | オーバラップ不可能／可能 | |
| 線 | 強制参加 | 全順序関係 | 不可能 | (普通の)バリア |
| | 任意参加 | 全順序関係 | | SBM |
| | | 半順序関係 | | HBM, LDBM, FDBM |
| 面 | 強制参加 | 全順序関係 | 不可能 | Fuzzyバリア |
| | | | 可能 | 重複可能バリア |
| | 任意参加 | 全順序関係 | 不可能 | 文献[104]で提案 |
| | | | 可能 | Elasticバリア |
| | | 半順序関係 | 不可能 | 文献[104]で提案 |
| | | | 可能 | 文献[104]で提案 |

# バリア同期モデルの分類

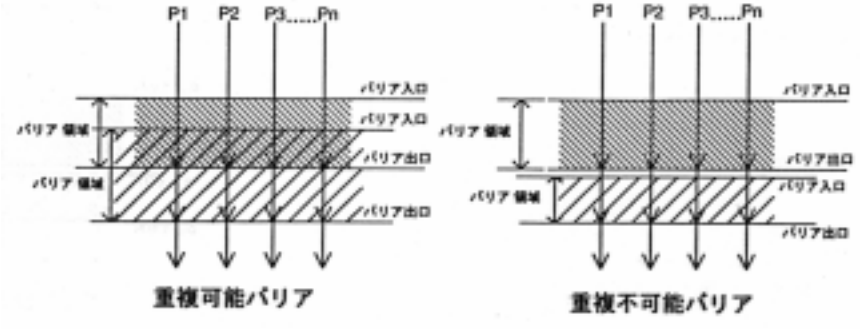
## 1) 線バリア vs. 面バリア




## 2) 全順序 vs. 半順序




## 3) 重複可能 vs. 重複不可能




## 4) 強制参加 vs. 任意参加

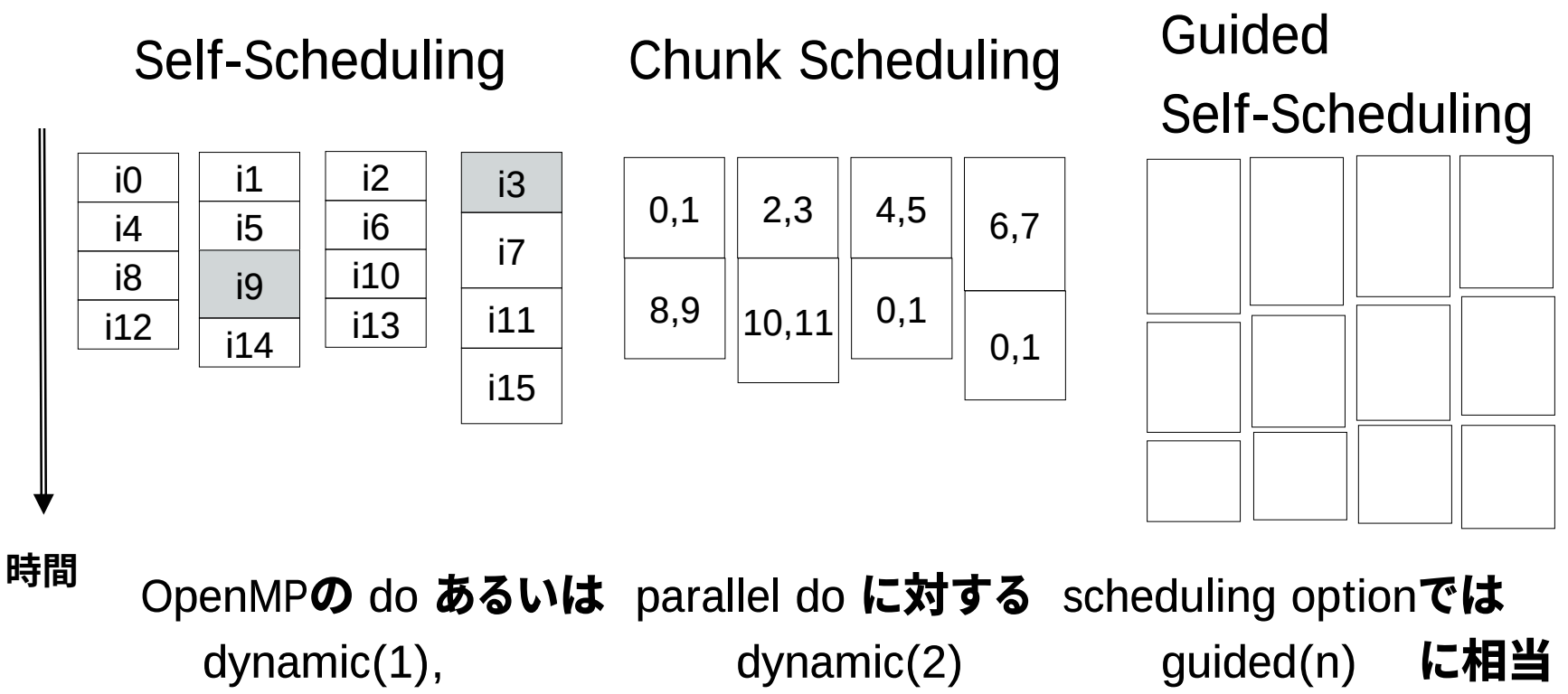
# 負荷の不均衡の解消

- **動的負荷分散方式**
  - Doall**型の例**
    - Self-Scheduling，Chunk Scheduling，Guided Self-Scheduling
  - **反復計算向けの負荷分散アルゴリズム**
    - **N回目の計算時の実行時情報からN+1回目の計算時の最適な処理の分割を予測する。**

      **⇒ 非均質な計算環境への適用可**(heteroTINPAR by **富田研**)
  - OS**レベルでの対応**
    - Cache Affinity Scheduling (**ラストプロセッサ方式、最小介入数方式，等**)
    - Memory Affinity Scheduling
      - 1)**ホーム**，2)**コピー**， 3)Minimum Load Cluster，4)Other
    - **ギャングスケジューリング**/Co-scheduling
  - **その他**……HTG**を利用した実行時粒度制御機構**

# 負荷の不均衡の解消

- (広義の)Self-scheduling方式




OpenMPの do あるいは dynamic(1),

parallel do に対する dynamic(2)

scheduling optionでは guided(n) に相当

# 非均質環境における動的負荷分散の例 (heteroTINPAR)

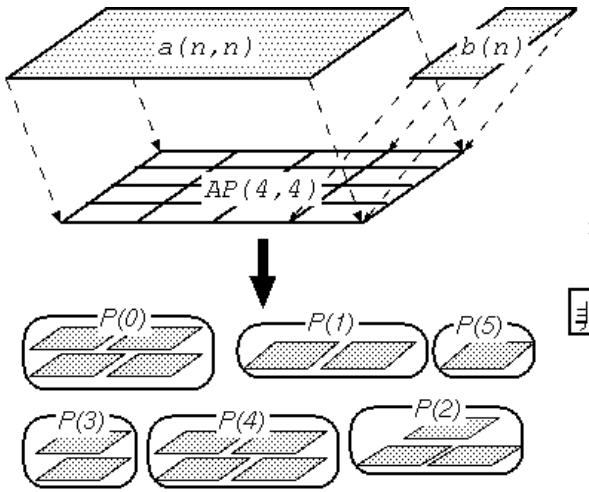
**While(収束条件成立) {**

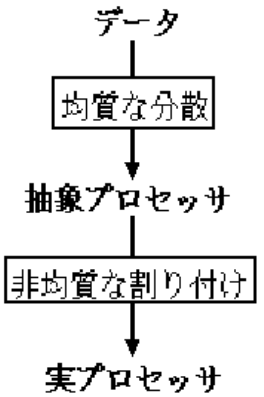
**for I = 0, n-1 do{**

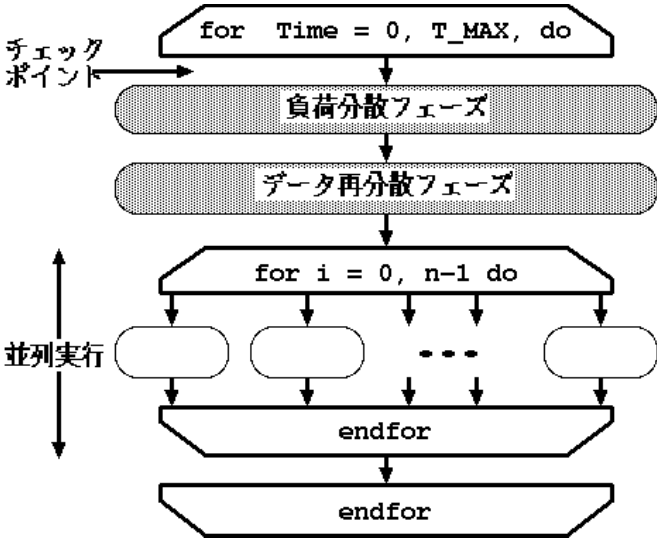
**……**

**}**

**}**

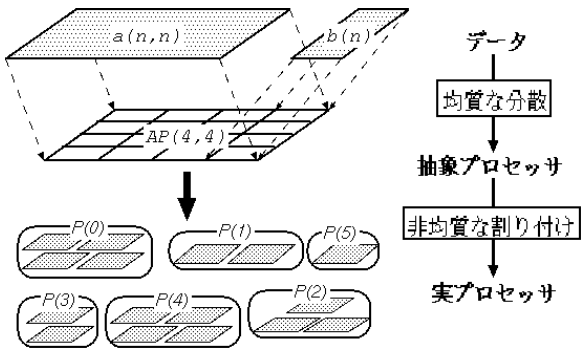


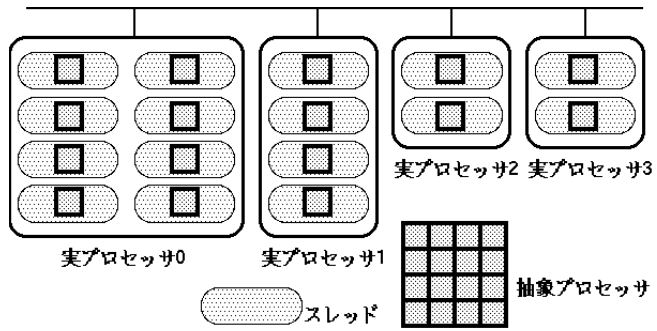



# 非均質環境における動的負荷分散の例 (heteroTINPAR)

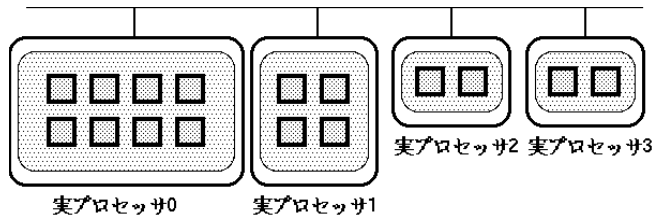
**実プロセッサにおける**
**複数抽象プロセッサの処理方式**




**マルチスレッド方式**




**シングルスレッド方式**



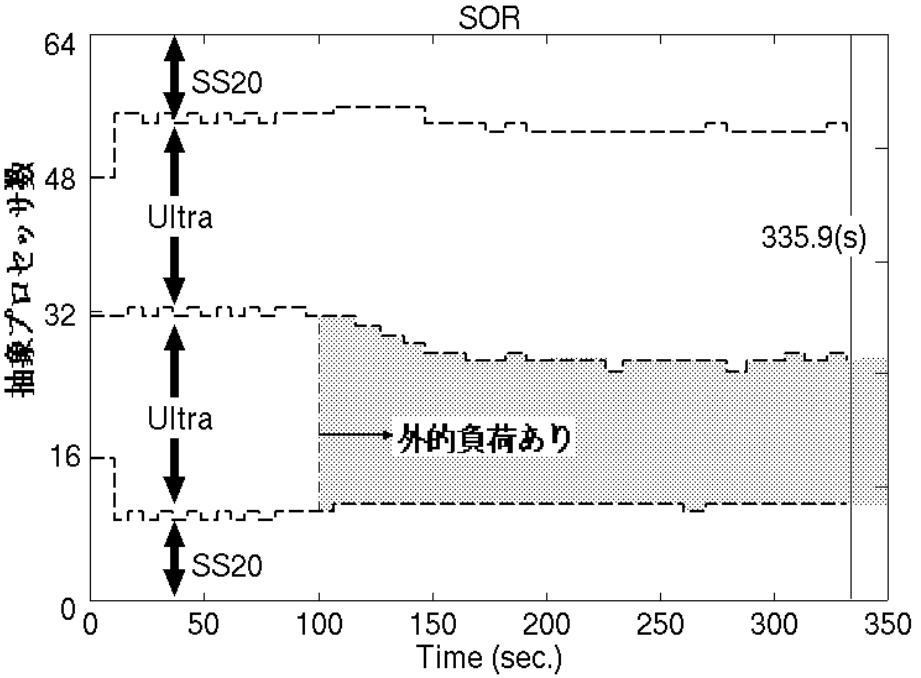
# 非均質環境における動的負荷分散の例 (heteroTINPAR)

**動的負荷分散による実行時間の変化**

**評価環境**： SS20x2台＋Ultra1 x 2台
**評価プログラム**：
SOR(1024x1024, 2000反復)
**抽象プロセッサ数**：64

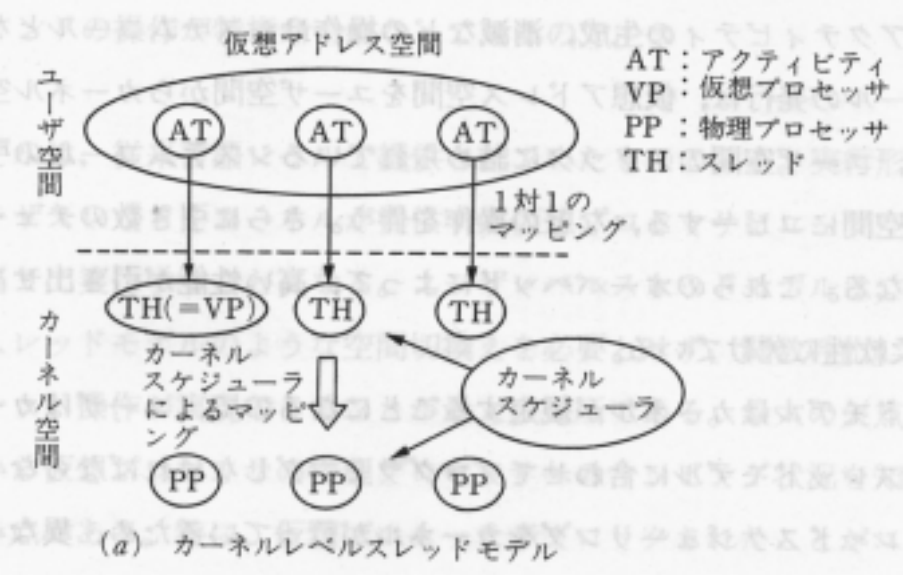
**実行開始100秒後に一方のUltra1に負荷をかける**

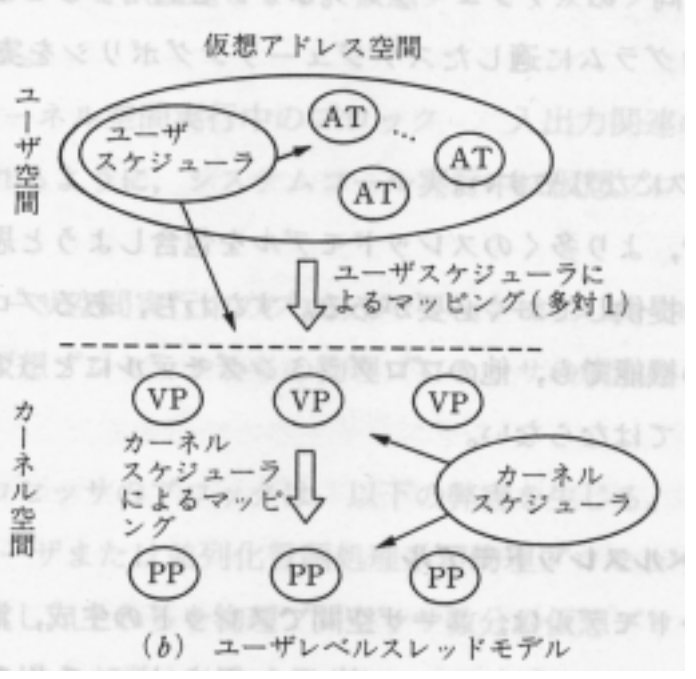
**負荷分散なしの実行時間**＝357.6(s)



# マルチスレッド環境

Kernel Level **スレッド** vs. User Level **スレッド**




[HW**実装(余談)**] Simultaneous Multi-threading(SMT) …Intel**用語では**HT

# SSIクラスタ

- Complete Transparency.... Shared Address Space
- Scalable Performance.....Small SSI Overhead
- Enhanced Availability..... Fault Tolerant

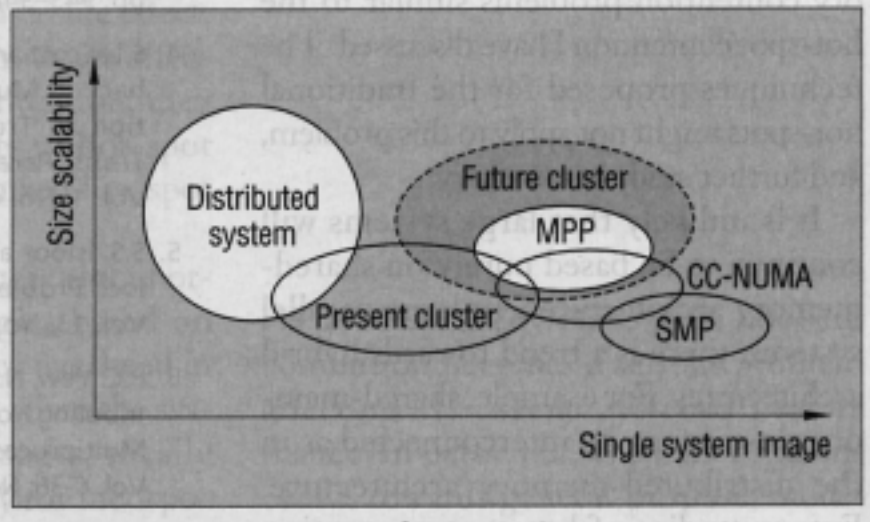


Figure 1. Design space of competing computer architectures.

Designing SSI Clusters with Hierarchical Checkpointing and Single I/O Space

Kai Hwang, Hai Jin, Edward Chow, Cho-Li Wang, and Zhiwei Xu

IEEE Concurrency, Vol.7, No.7, pp.60—69, January-March 1999

## マルチスレッド処理性能向上に注力

CPUの性能指標に1クロック当たりに実行可能な命令数（IPC：Instructions Per Clock cycle）がある。PentiumやAthlonといったx86互換CPUは，平均で1命令から3命令を同時実行する。しかし3命令以上にIPCを上げるのは難しい。命令同士には依存関係がある。例えばある演算結果を基に条件分岐する命令があると，実行した結果を破棄して分岐先の値で処理し直す必要がある。しかもメモリー・アクセス速度はCPUの動作周波数に比べて遅い。分岐予測に失敗した際に発生する遅延が性能向上の妨げとなっている。

そこで浮上してきたのが，複数スレッドを同時に実行するSMT（Simultaneous Multithreading）である（図）。命令の実行順を決める際に，二つのスレッドから実行可能な命令を抽出する。二つのスレッドは独立しており，片方のスレッドで分岐予測に失敗しても，もう一つのスレッドの実行を続けられる。実装例としては，米Intel社のHyper-Threadingがある。

ただSMTは物理的に一つのCPUである以上，演算器やキャッシュを共有しなくてはならない。複数のスレッドを同時実行するもう一つの手法がCMP（Single-Chip Multiprocessor）だ。米IBMの汎用機向けCPU「Power4」がすでに実装している。米Sun Microsystems社も2004年第1四半期に同社のCPU「UltraSPARC III」のコアを二つ集積した「UltraSPARC IV」を出荷する予定だ。

100mm²前後のダイサイズが求められるパソコン向けCPUでは，2005年の65nm世代以降でマルチコアに向かう見込みだ。とはいえマルチコア化による性能向上は，マルチスレッドで動作するアプリケーションでなければ生かせない。そこでマルチコアのメリットを高めるために，Intelはシングル・スレッドのアプリケーションを自動的にマルチスレッド化するコンパイラを開発中だという。

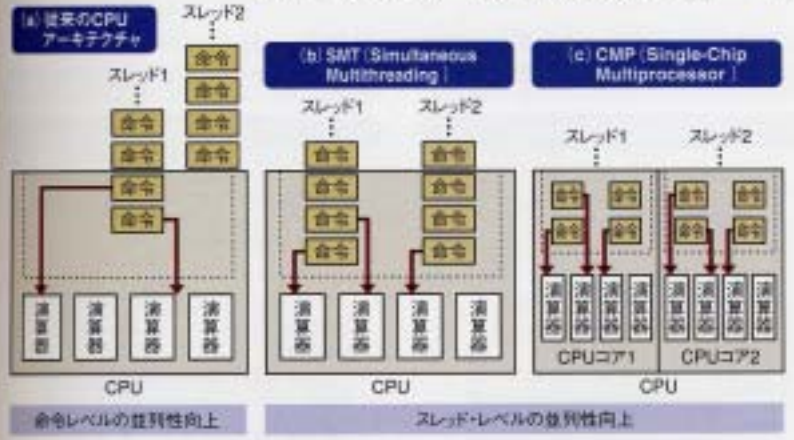
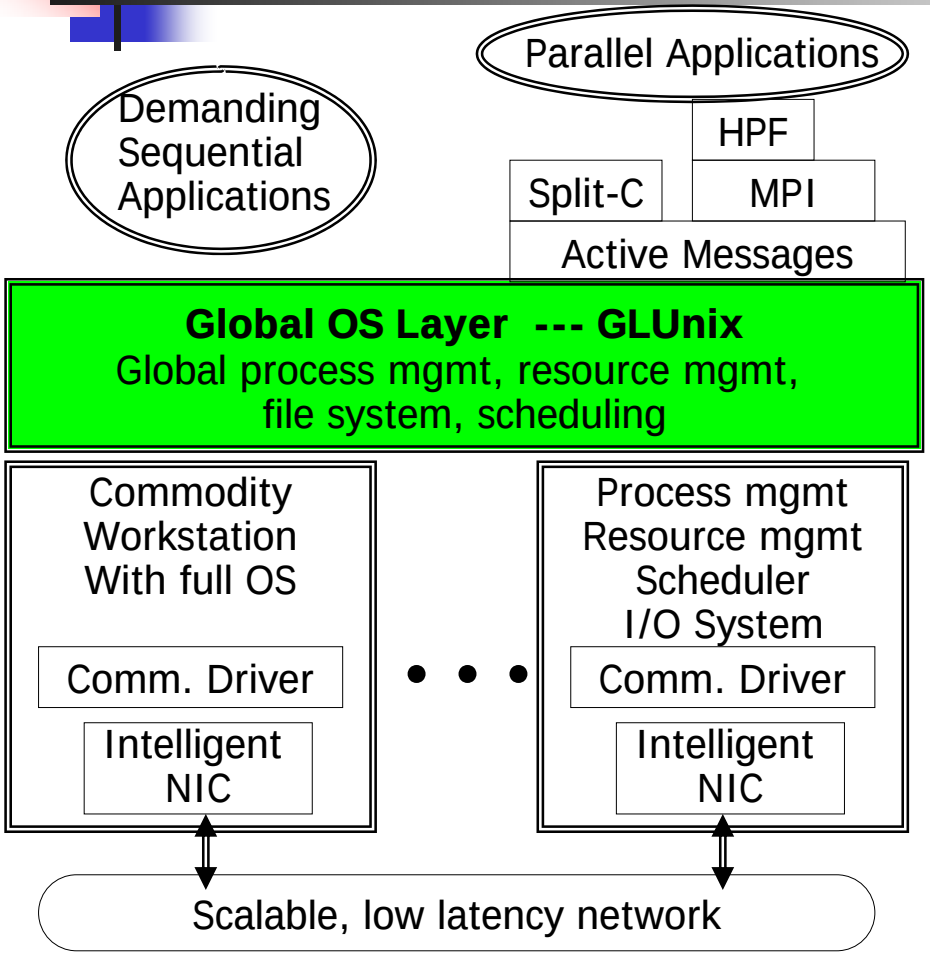



**図 並列処理性能を上げるアーキテクチャの種類**

従来は命令レベルの並列実行性能の向上に注力してきた。今後は微細化によって増えたトランジスタ数を生かし，米Intel社のHyper-Threadingに代表されるSMT（Simultaneous Multithreading）と，CPUダイに複数のCPUコアを作り込むCMP（Single-Chip Multiprocessor）が主流になる。

# UC Berkley “NOW”




GLUnix

UNIX のシステムコールを横取り

SSI環境を実現

xFS ….Serverless FS

Network RAID(Striping)

Cooperating Caching

1)Clientの主記憶をFile Cacheとして使用

2)MP用のCC機構を利用してFileを管理

# RAID:
# Redundant Array of Inexpensive Disks

| | RAID1 | RAID2 | RAID3 | RAID4 | RAID5 | RAID0 |
|---|---|---|---|---|---|---|
| ストライピングの有無 | 無 | 有 | 有 | 有 | 有 | 有 |
| ストライピングの単位 | --- | bit | bit | block | block | block |
| 故障検出単位 | Disk | 数bit | Disk | Disk | Disk | Disk |
| データ復元方法 | 2重化 | ハミング符号 | パリ | パリ | パリティ | --- |
| 冗長ディスク構成 | 固定 | 固定 | 固定 | 固定 | 分散 | --- |

(striping)

| DISK0 | DISK1 | DISK2 | DISK3 | DISK4 |
|---|---|---|---|---|
| P0(B0-B3) | B0 | B1 | B2 | B3 |
| B4 | P1(B4-B7) | B5 | B6 | B7 |
| B8 | B9 | P2(B8-B12) | B11 | B12 |

**RAID5の構成例**

# 近くのハードディスク vs 遠くのメモリ

- HDDの転送速度
  - SCSI系　～　320MB/s (ULTRA SCSI320)
  - Fibre Channel 1.06～4.24Gbps
  - IEEE1394系　～　400Mbps
- HDDのレイテンシ
  - 数ミリ秒のオーダ　(DISKの回転速度が支配)

- ネットワークの転送速度　……. 1～10Gbps
- ネットワークのレイテンシ……. 数十マイクロ秒