第５章　マルチプロッサ

５．１　概論

５.１.１分類

「超」の規模：１TFLOPSを実現できる

プロセッサ台数

地球シミュレータ：４０TFLOPS

ASCIプロジェクト

（１）プロセッサアーキテクチャ

SPP（Scalar Parallel Processor）

1TFLOPSの実現：１千台程度

マイクロプロセッサ：１GFLOPS

①スーパスカラ方式

②VLIW方式

③スーパパイプライン方式

④マルチスレッド方式

VPP（Vector Parallel Processor）

メモリ系単純化

ベクトルプロセッサの性能：１０GFLOPS程度

「超」並列：１００台程度

台数の少ない構成方式：実行性能優

制御単純、理解しやすい、ユーザからの連続性

VPPが優位

SPP：最新鋭のマイクロプロセッサを安価に利用

通信やメモリのレイテンシ（遅延）に耐え得

るアーキテクチャ

（２）メモリアーキテクチャ

①共有か非共有か

・共有メモリ方式

・非共有メモリ方式

②集中か分散か

集中共有、分散共有、

分散非共有、集中非共有の４つの場合

（３）ネットワークアーキテクチャ

トポロジ

交換方式（ストアド・フォワード、ワームホール）

並列計算機：特定の並列アーキテクチャと

特定応用との結び付き

数種の並列アーキテクチャに収斂する必要

５．１．２基本方式

（１）粒度

①プログラム分割

②プロセッサ割付け

負荷の分散と粒度の関係

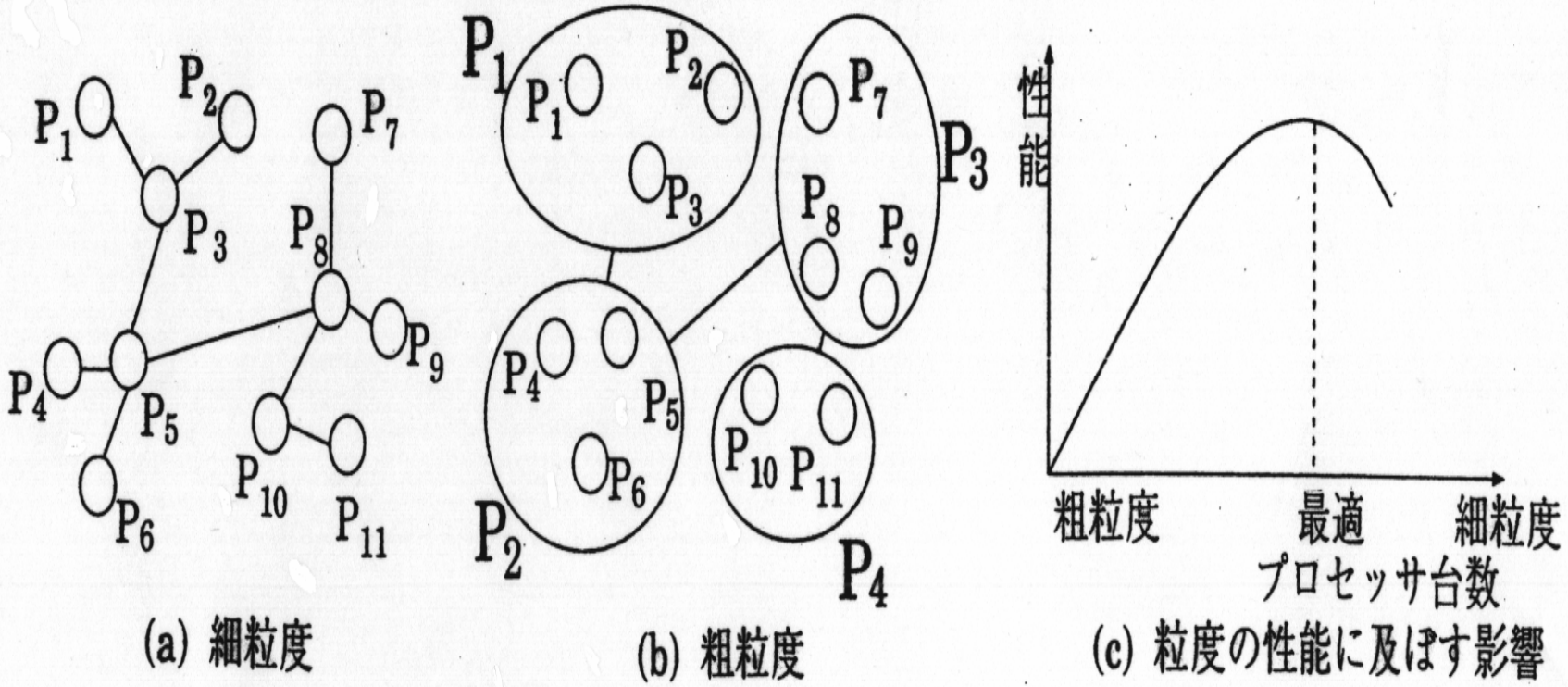
（２）プログラム分割

①メモリ共有方式

分割：命令系列のみ

データ：共有

処理内容・データ構造：



(a) 細粒度　(b) 粗粒度　(c) 粒度の性能に及ぼす影響

非定型、コンパイル時にはデータ分割不可

のような非定型応用

共有メモリ：

物理的に集中／分散共有方式で実現

細粒度並列処理向き

小規模のマルチプロセッサ：

集中共有メモリ方式

大規模なシステム：

分散共有メモリ方式

メモリ共有を分散非共有方式で実現

ワークステーションを用いたネットワークコンピューティング

仮想（論理）アドレス空間

ページフォールト

②メッセージ交換方式

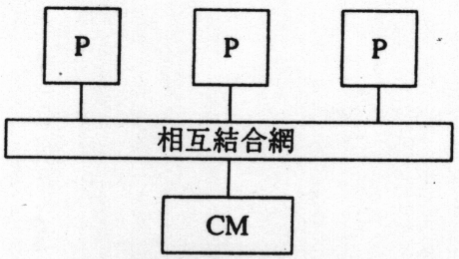
分割：命令系列とデータ

処理内容：定型的応用

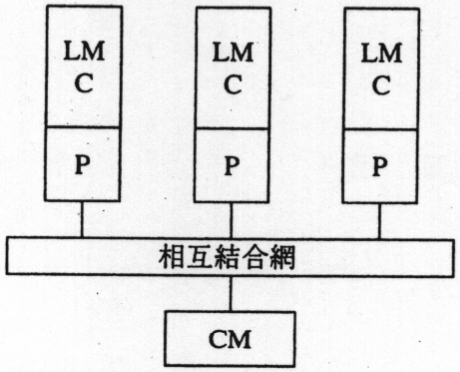
オブジェクト指向モデル

離散系シミュレーション

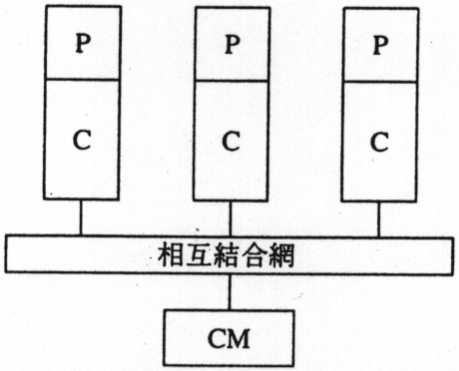
粗粒度並列処理



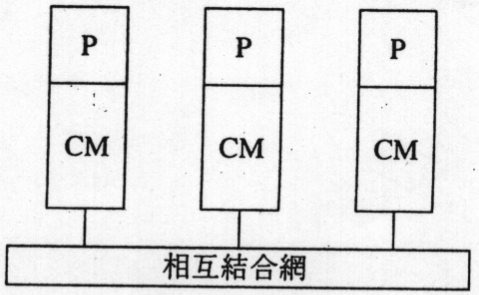
(a) 集中共有メモリ



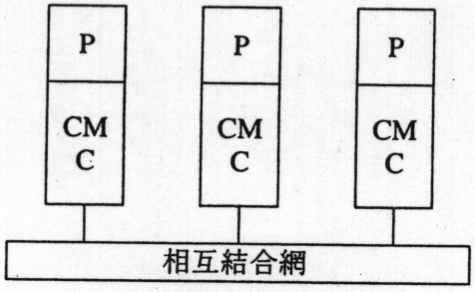
(b) 集中共有メモリ (ローカルメモリ付)



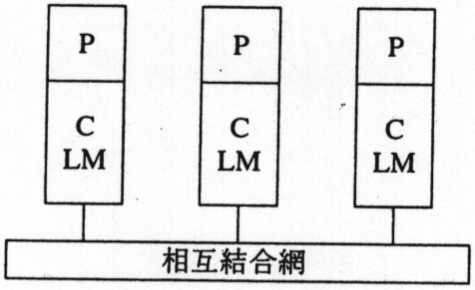
(c) 集中共有メモリ(キャッシュ付)



(d) 分散共有メモリ



(e) 分散共有メモリ (キャッシュ付)



(f) メッセージ交換 (分散非共有メモリ)

P : プロセッサ
CM : 共有メモリ
LM : ローカルメモリ (非共有メモリ)
C : キャッシュメモリ

表 5.1　マルチプロセッサ例

機種	結合網	台数(プロセッサ)	性能	メモリ
CM-5	トリー	16000ノード(1ノード: 1SPARC+ 4ベクトル装置)	2T	非共有
GC	3次元トーラス	1024クラスタ(1クラスタ: 16台のT-9000)	400G	非共有
Paragon	2次元メッシュ	4096 (iPSC860)	300G	非共有
nCUBE-3	ハイパキューブ	65000 (独自プロセッサ)	6.5T	非共有
TERA-3D	3次元トーラス	2048(DEC α)	300G	共有
SP2	多段網	512ノード(1ノード: 16台のRS/6000)	136G	非共有
KSR	階層リング	1088 (独自プロセッサ)	43.5G	共有キャッシュ(キャッシュオンリ)
Exemplar	リング	16クラスタ(1クラスタ: 8台のPA, クロスバ)	198M	共有キャッシュ(SCI プロトコル)
AP-1000	2次元トーラス	1024 (SPARC)	8.53G	非共有
VPP-500	クロスバ	222 (1.6Gベクトル,300Mスカラ (VLIW))	355G	共有
SX-4	クロスバ(ノード内) 光接続(ノード間)	16ノード(1ノード: 32台のPE, 1PE: 2G ベクトル)	1T	共有(ノード内) 非共有(ノード間)
Cenju-3	多段網	256 (V_R4400)	12.8G	非共有
ADENART	ハイパクロス	256 (独自)	2.56G	非共有
Cyberflow	2次元トーラス	64 (データフロー)	640M	非共有
CP-PACS	ハイパクロスバ	1024 (PA+機能追加)	300G	非共有
SR2201	ハイパクロスバ	1024 (PA+機能追加)	300G	非共有
JUMP-1	RDT	512 (SuperSPARC+)	30G	共有キャッシュ

T: TFLOPS, G: GFLOPS, M: MFLOPS, 共有: 共有メモリ方式,
共有キャッシュ: キャッシュコヒーレンス制御付き共有メモリ方式,
非共有: メッセージパッシング方式, PA: HP社製マイクロプロセッサ,
SPARC: サンマイクロシステムズ社製マイクロプロセッサ,
V_R : MIPS社製マイクロプロセッサ

bus to an Omega network.

Institution	Name	Number of nodes	Basic topology	Data bits/link	Network clock rate (MHz)	Peak BW/link (MB/sec)	Bisection (MB/sec)	Year
Thinking Machines	CM-2	1024 to 4096	12-cube	1	7	1	1,024	1987
Intel	Delta	540	2D grid	16	40	40	640	1991
Thinking Machines	CM-5	32 to 2048	multistage fat tree	4	40	20	10,240	1991
Intel	Paragon	4 to 2048	2D grid	16	100	175	6,400	1992
IBM	SP-2	2 to 512	multistage fat tree	8	40	40	20,480	1993
Cray Research	T3E	16 to 2048	3D torus	16	300?	600	122,000	1997
Intel	ASCI Red	4536 (× 2 CPUs)	2D grid			800	51,600	1996
IBM	ASCI Blue Pacific	1336 (× 4 CPUs)				150		
SGI	ASCI Blue Mountain	1464 (× 2 CPUs)	fat hypercube			800	200 × nodes	1998
IBM	ASCI Blue Horizon	144 (× 8 CPUs)	multistage Omega			115		1999
IBM	SP	1 to 512 (× 2 to 16 CPUs)	multistage Omega			500		2000
IBM	ASCI White	484 (× 16 CPUs)	multistage Omega			500		2001

Figure 8.18 Characteristics of interconnections of some commercial supercomputers. The bisection bandwidth is for the largest machine. The 2D grid of the Intel Delta is 16 rows by 35 columns and the ASCI Red is 38 rows by 32 columns. The fat-tree topology of the CM-5 is restricted in the lower two levels, hence the lower bandwidth in the bisection. Note that the Cray T3E has two processors per node, and the Intel Paragon has from two to four processors per node.

５．２メモリ共有型

キャッシュの装備を前提

①データ参照の局所性を生かすこと

②無駄な通信を極力削減すること

５.２.１構成方式

①集中共有メモリ

②集中共有メモリ（ローカルメモリ付き）

③集中共有メモリ（キャッシュ付き）

④分散共有メモリ

⑤分散共有メモリ（キャッシュ付き）

共有メモリ

①UMA（Uniform Memory Access）モデル

②NUMA（Non-Uniform Memory Access）モデル

4．5 メモリコンシステンシモデル

書込みの順序づけ (ordering)

プロセッサA,Bが書込み

プロセッサC：Aの後Bが到着

プロセッサD：Bの後Aが到着

これでよいのか？

4.5.1プロセッサコンシステンシモデル

プロセッサA	プロセッサB
①data＝new;	③while (flag！＝set) {}
②flag＝set;	④datacopy＝data;

同一のプロセッサからのライト：

その順で反映

2つ以上のプロセッサで発せられた書込み：

順序については何も制約なし

プロセッサ A

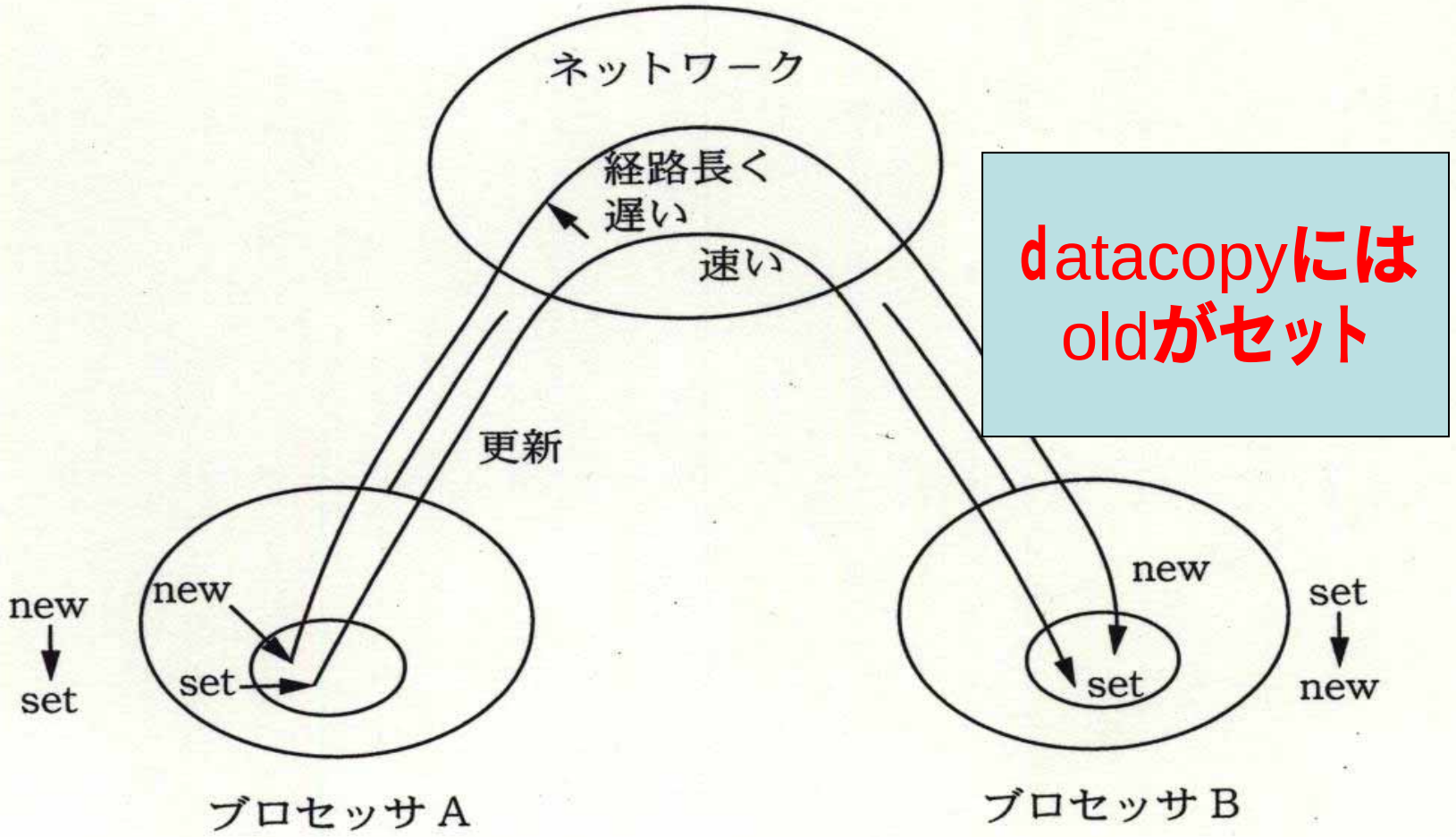
① data＝new；

② flag＝set；

プロセッサ B

③ while （flag！＝set）{}

④ datacopy＝data；



datacopyには
oldがセット

(a)　プロセッサコンシステンシモデル違反

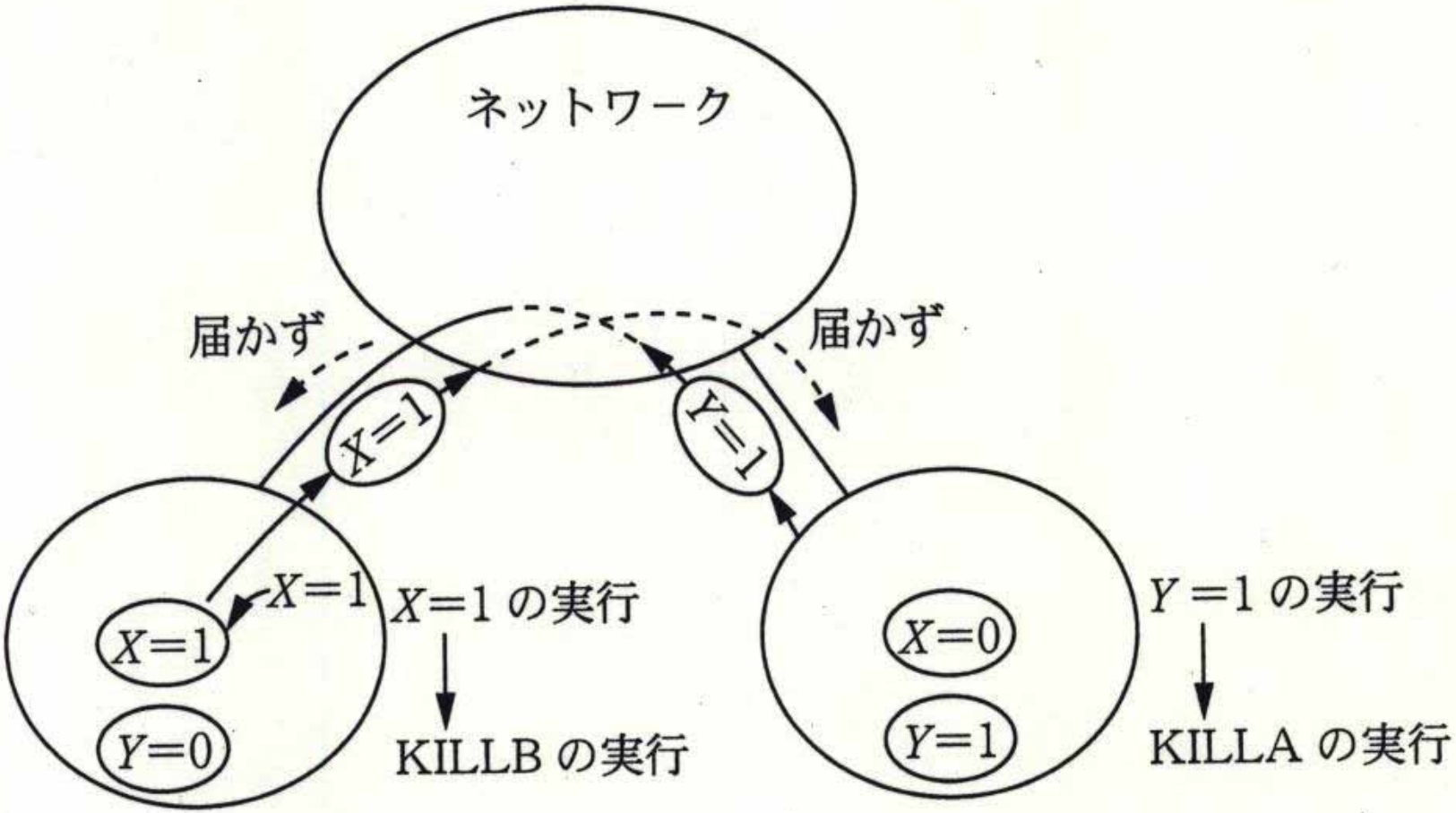
4.5.2逐次 (sequential) コンシステンシ

```
          X＝Y＝0
プロセッサA          プロセッサB
X＝1                 Y＝1
IF Y＝0 KILLB    IF X＝0 KILLA
```

A,Bともに相打ちは起こるか？

(1)Lamportの逐次コンシステンシ定義

マルチプロセッサ上での
　並列プログラムの実行結果＝
並列プログラムを単一のプロセッサで
　時分割で実行したときと同一



(b)　逐次コンシステンシモデル違反

図 4.29　メモリコンシステンシ問題

(2)十分条件

並べられた命令順に実行

メモリ操作の大域的完了後，後続メモリ操作発行

大域的書込み完了：

書込み操作がすべてのプロセッサに反映

大域的読込み完了：

リードデータが大域的に書込み完了

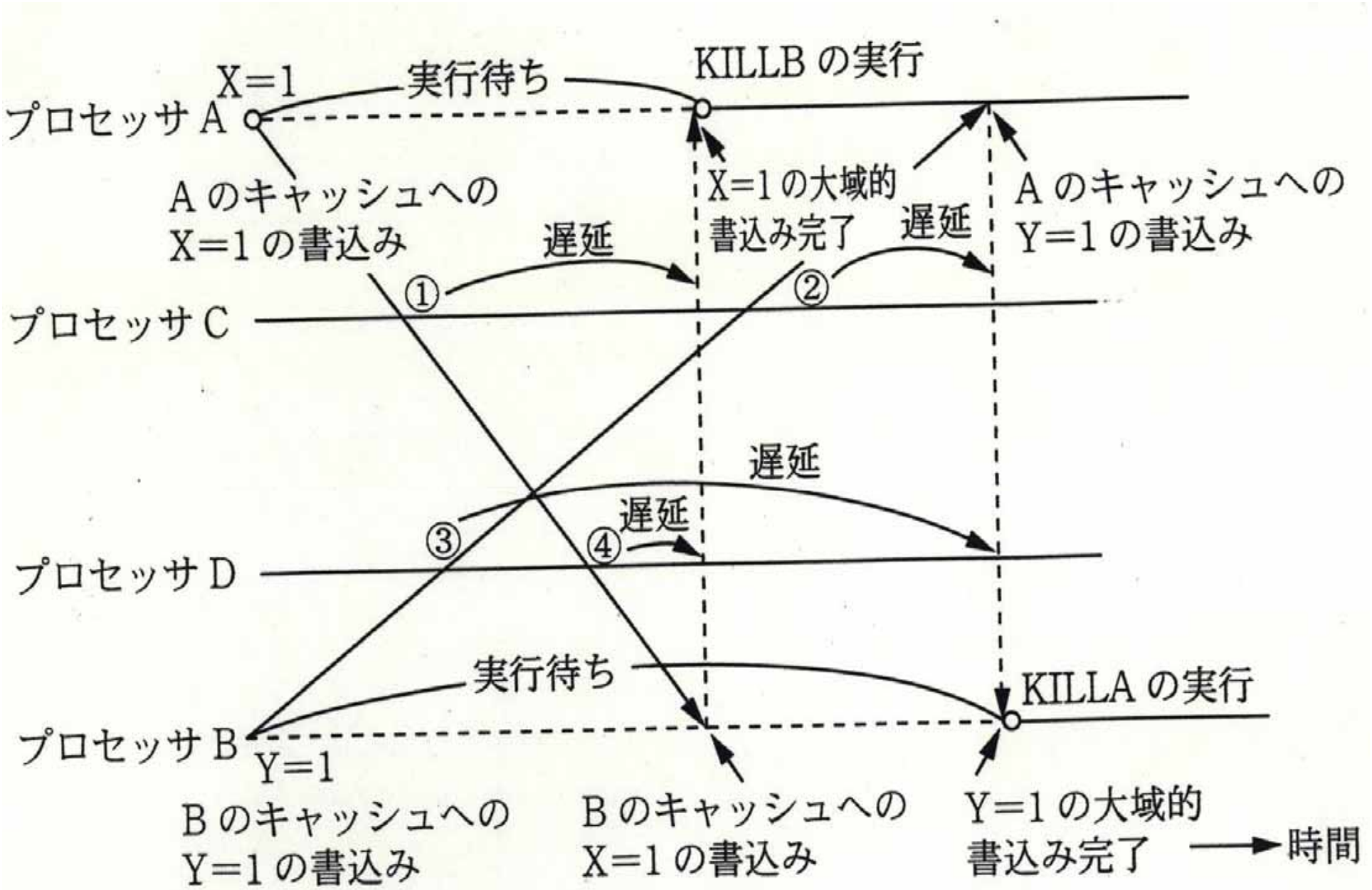


図 4.30 逐次コンシステンシの実現

(3)逐次コンシステンシの緩和

臨界領域へのアクセス

異なる領域へのアクセス

Weakコンシステンシ

同期命令

先行命令の大域的完了後発行

後続命令は同期命令の大域的完了後発行

Releaseコンシステンシ

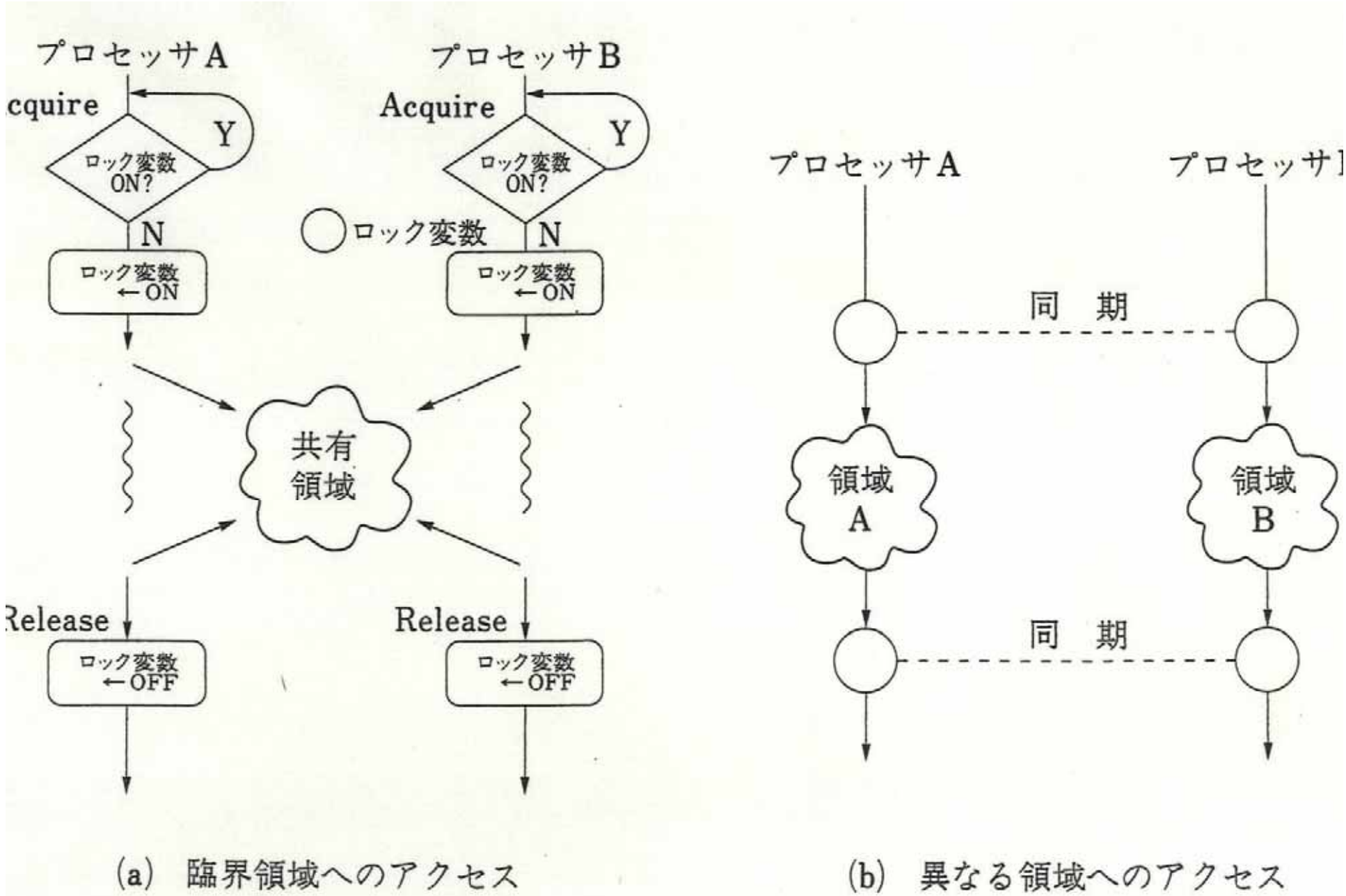
Acquire同期命令

後続命令はAcquireの大域的完了後発行

Release同期命令

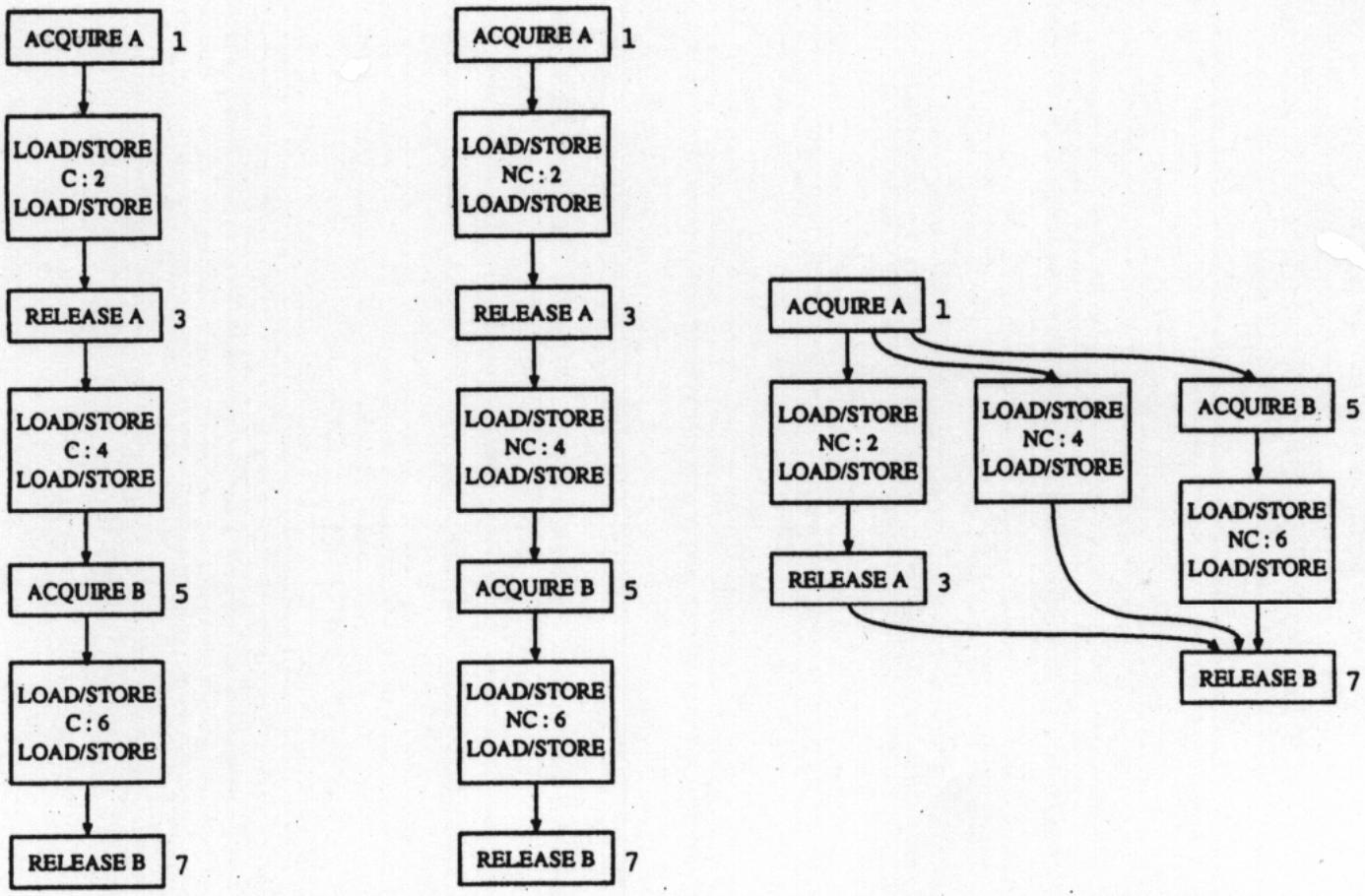
先行命令の大域的完了後発行

後続命令は発行可



(a)　臨界領域へのアクセス　　(b)　異なる領域へのアクセス

図 4.31　共有データへのアクセス



(a)逐次コンシステンシ　(b)ウイークコンシステンシ　(c)リリースコンシステンシ

ノンコンシステントなメモリ操作　逐次コンシステントなメモリ操作

4.4 キャッシュコヒーレンス

(1)ハードウェアによる方式

スヌープキャッシュ方式

ディレクトリ方式

(2)ソフトウェアによる方式

4.4.1スヌープキャッシュ法

(1)単純な方式の場合

スヌープコントローラ

(2)各ブロックにタグを持たす方式

ライト時に無効化または更新

キャシュ間転送時の主記憶に書き戻し

(3)バスの特徴

放送能力：分散制御

排他制御：逐次コンシステンシ

①スヌープ方式

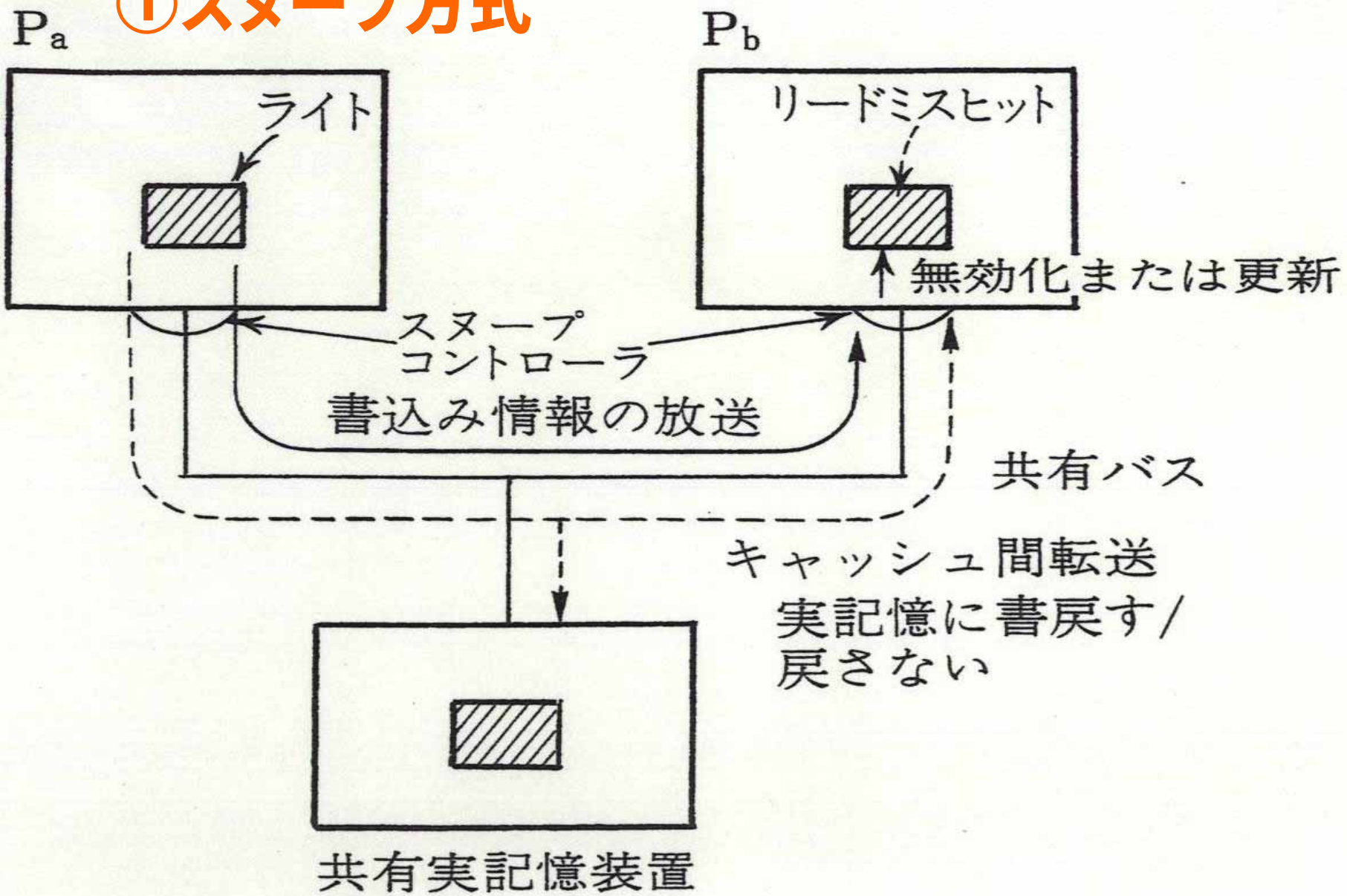
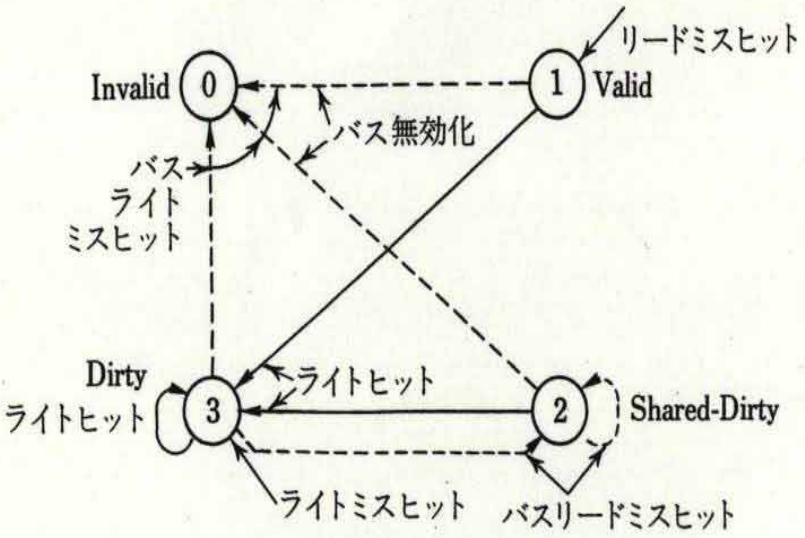


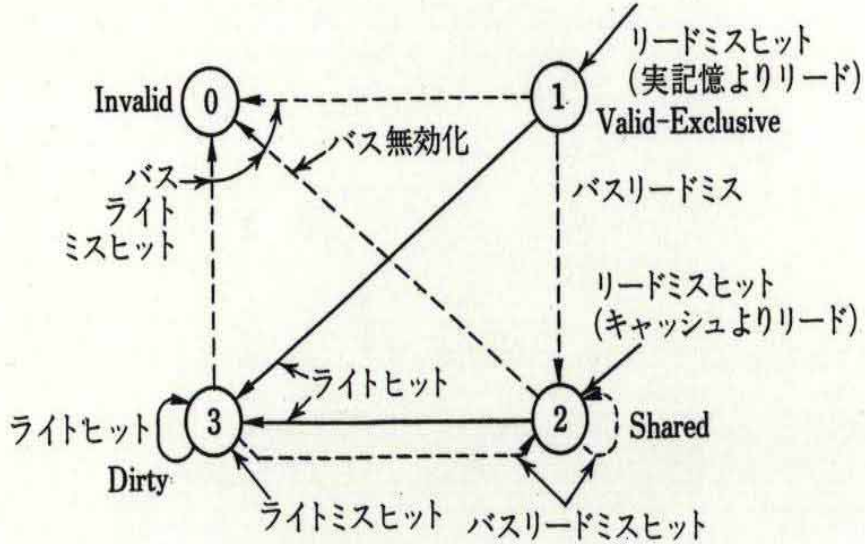
図 4.24　スヌープキャッシュの基本方式

表 4.3 スヌープキャッシュ方式

キャッシュ間転送時の主記憶更新 ＼ 共有ブロックへの書込み時処理	ブロードキャスト無効化	ブロードキャスト更新
変更のあるブロックのキャッシュ間転送時に主記憶へは書き戻さない	Berkeley State 0 : Invalid State 1 : Valid (clean, potentially shared, unowned) State 2 : Shared-Dirty (modified, potentially shared, owned) State 3 : Dirty (modified, only copy, owned)	Dragon State 0 : Valid-Exclusive (clean, only copy) State 1 : Shared-Clean (clean, one or more copy) State 2 : Shared-Dirty (modified, one or more copy) State 3 : Dirty (modified, only copy)
変更のあるブロックのキャッシュ間転送時に主記憶へも書き戻す	Illinois State 0 : Invalid State 1 : Valid-Exclusive (clean, only copy) State 2 : Shared (clean, possibly other copies) State 3 : Dirty (modified, only copy)	Firefly State 0 : Valid-Exclusive (clean, only copy) State 1 : Shared (clean) State 2 : Dirty (dirty, only copy)

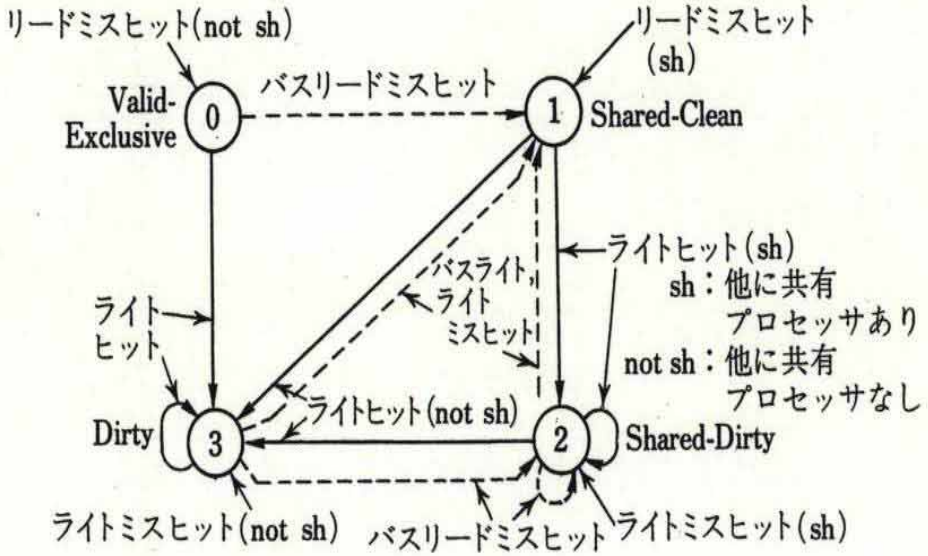


(a) バークレイプロトコル

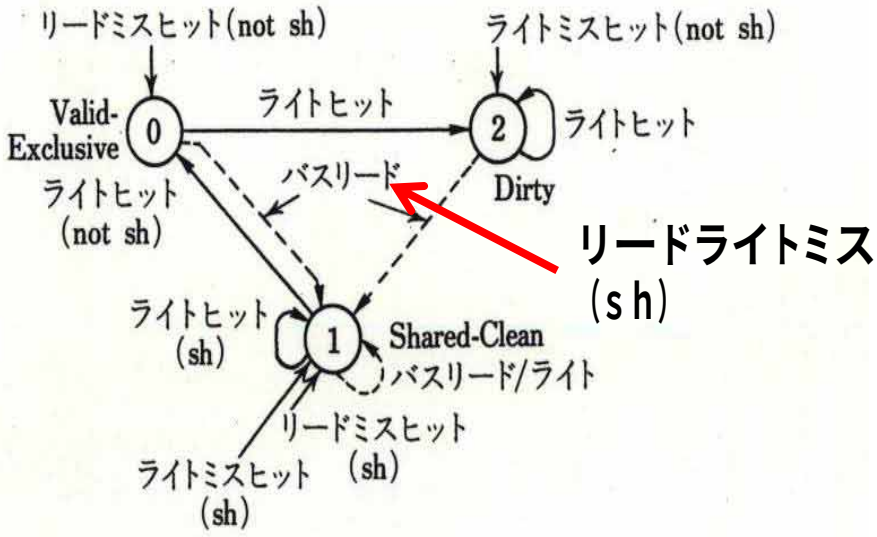


(b) イリノイプロトコル

MESI Pentiumで採用



(c) ドラゴンプロトコル

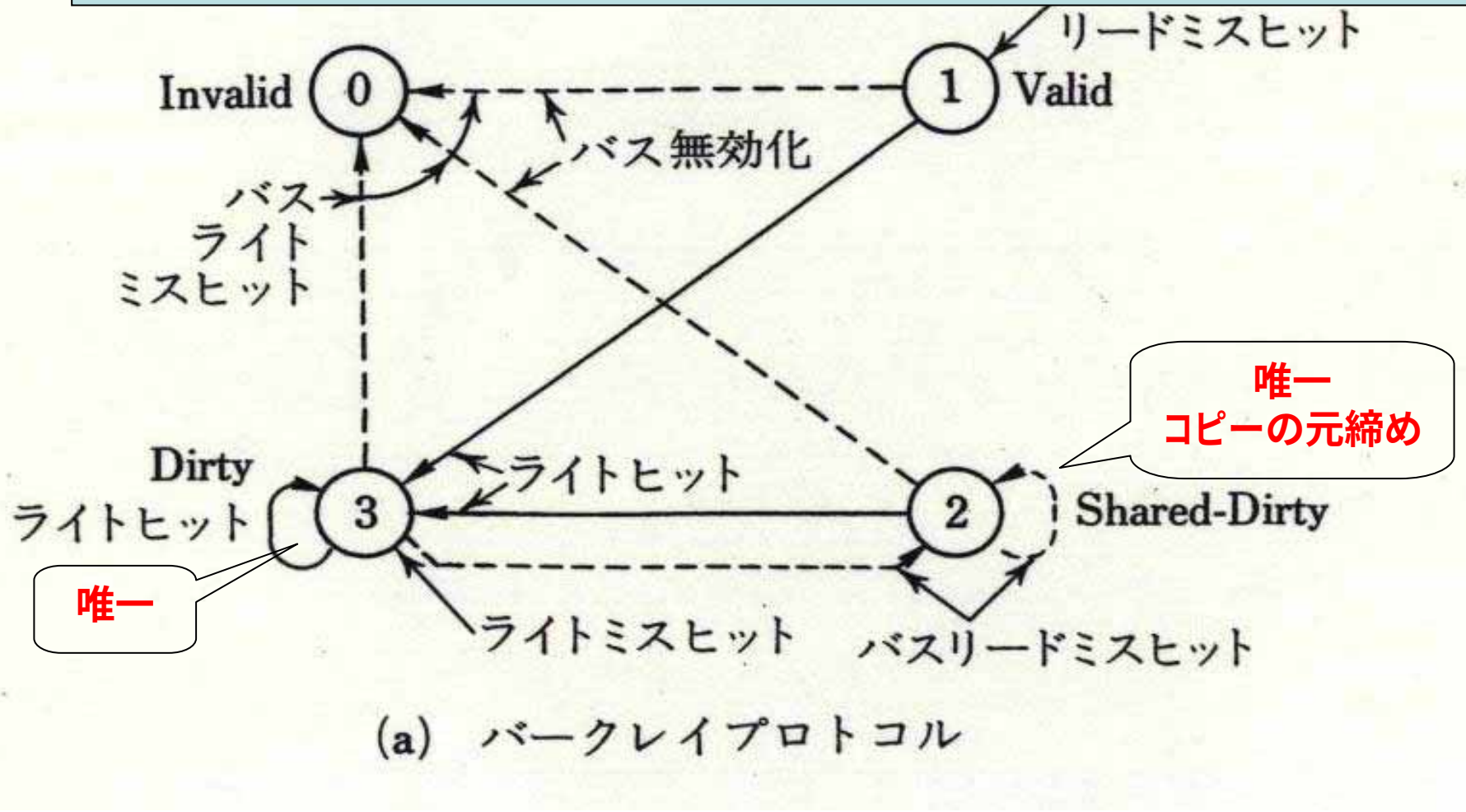


(d) ファイアフライプロトコル

DEC Firefly
Multiprocessor

図 4.25 種々のプロトコル[13)]

プロセッサA	状態		プロセッサB	状態	プロセッサC	状態
LOAD	Valid					
STORE	Dirty					
	Shared	Dirty	LOAD	Valid		
					LOAD	Valid
	InValid		STORE	Dirty		InValid



(a) バークレイプロトコル

MOESI　ADM６４で採用

- Modified　Dirty　唯一、メモリ内容と不一致
- Owned　唯一、メモリ内容と不一致、コピーを持つのはＳｈａred
- Exclusive、Clean、唯一
- Shared　他にOwnedがあればメモリ内容と不一致、なければメモリ内容と一致
- Invalid　無効

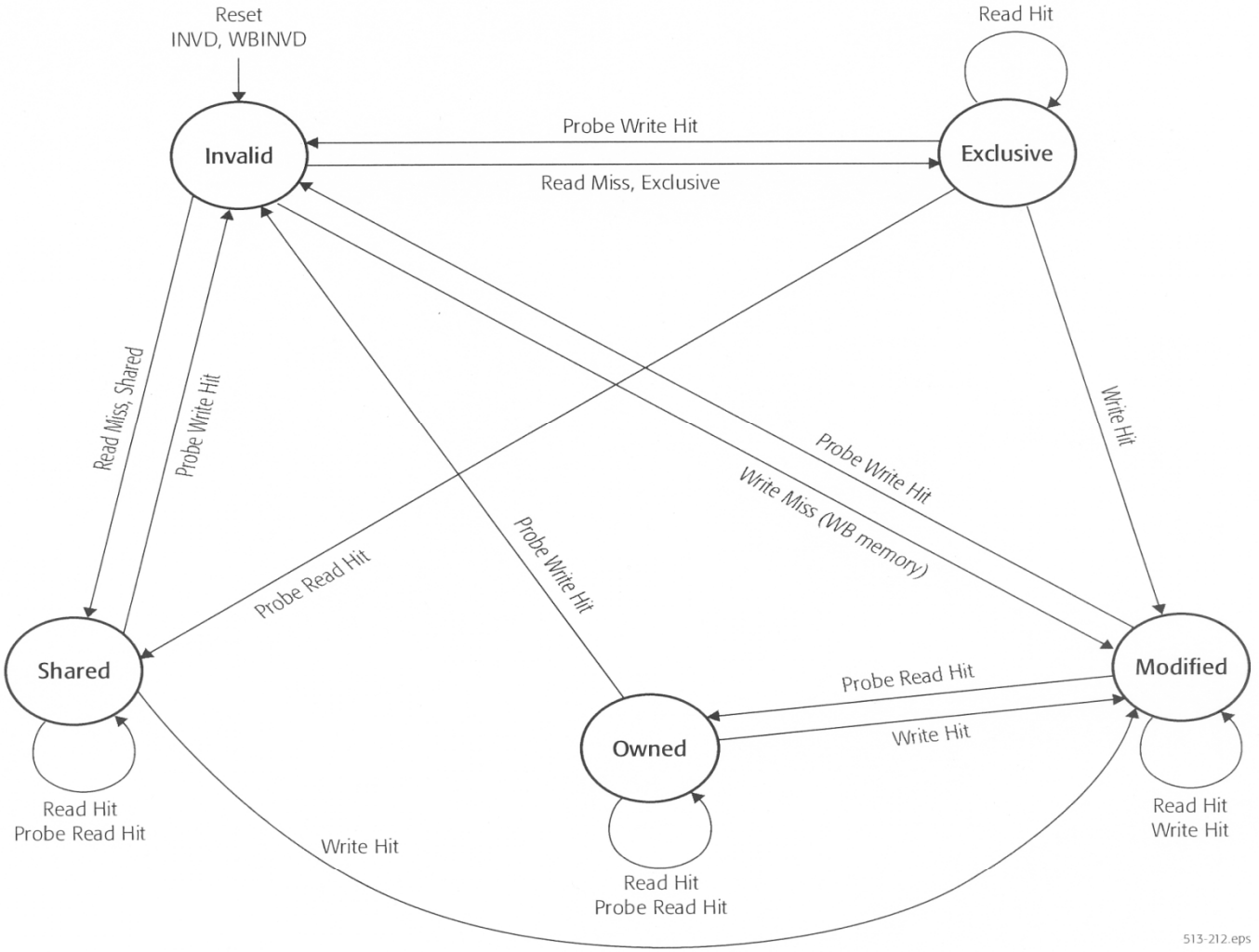


Figure 7-2. MOESI State Transitions

To maintain memory coherency, external bus masters (typically other processors with their own internal caches) need to acquire the most recent copy of data before caching it internally. That copy can be in main memory or in the internal caches of other bus-mastering devices. When an external master has a cache read-miss or write-miss, it *probes* the other mastering devices to determine whether the most recent copy of data is held in any of their caches. If one of the other mastering devices holds the most recent copy, it provides it to the requesting device. Otherwise, the most recent copy is provided by main memory.

表1:キャッシュラインの状態遷移の例

	事象	P0	P1	P2
0	初期状態	Invalid	Invalid	Invalid
1	P0 Read	Exclusive	Invalid	Invalid
2	P1 Read;P0応答データ供給	Shared	Shared	Invalid
3	P2 Read;P0応答データ供給	Shared	Shared	Shared
4	P0 Write;Invalidate応答を待ってWrite	Modified	Invalid	Invalid
5	P2 Read;P0応答データ供給	Owned	Invalid	Shared
6	P1 Write;Invalidate、P0(Owner)応答データ供給、P2 Invalidate応答	Invalid	Modified	Invalid
7	P2 Read;P1応答データ供給	Invalid	Owned	Shared
8	P1 Writeback	Invalid	---	Shared

4.4.2**具体例**

4.4.3**ディレクトリ方式**

一般のネットワーク利用

メモリ側で集中管理

フルマップ

リミテッド

チェイン

一般のネットワーク

Snoop**方式はX**

放送機能が弱い

排他制御が困難

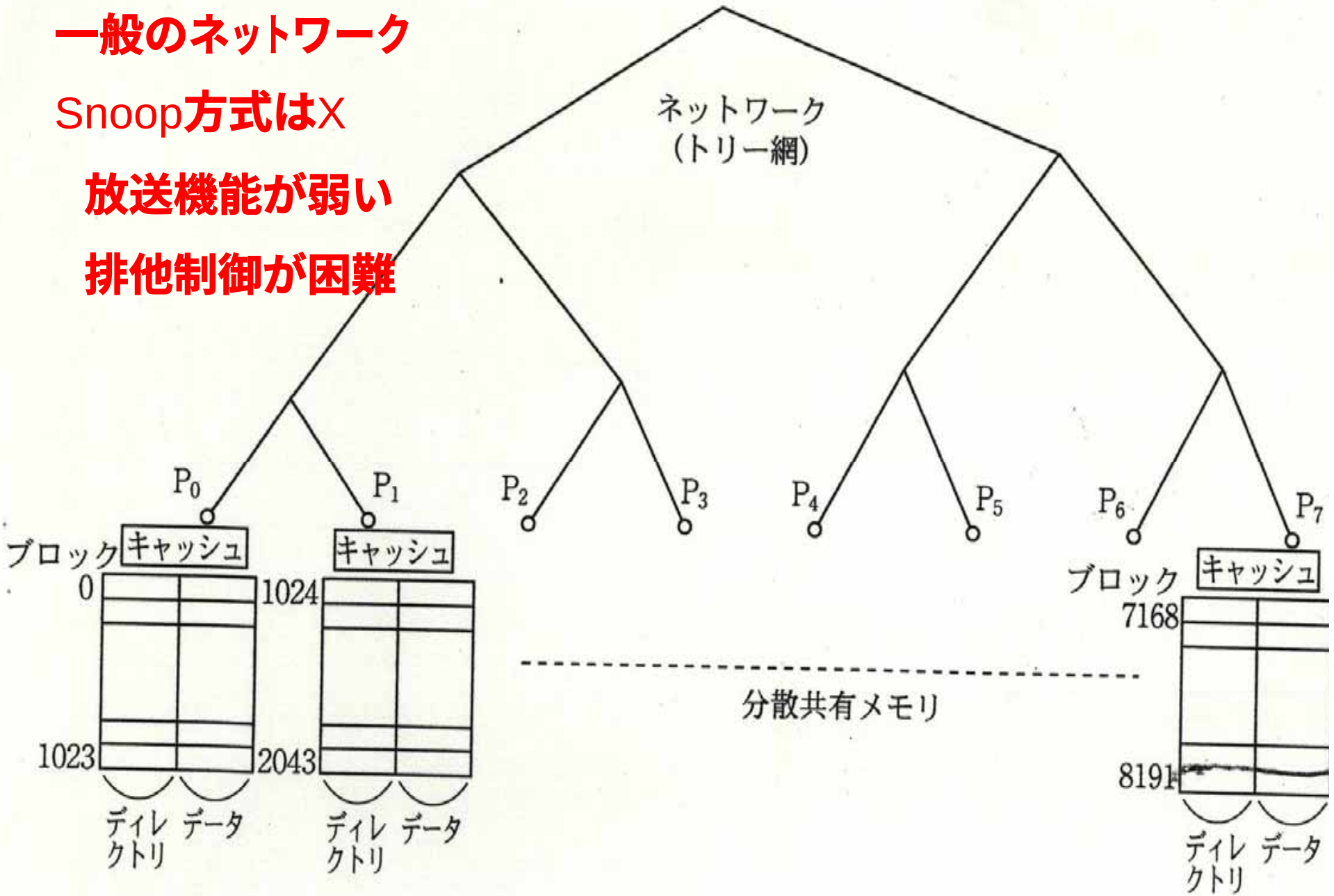


図 4.26 ディレクトリ方式

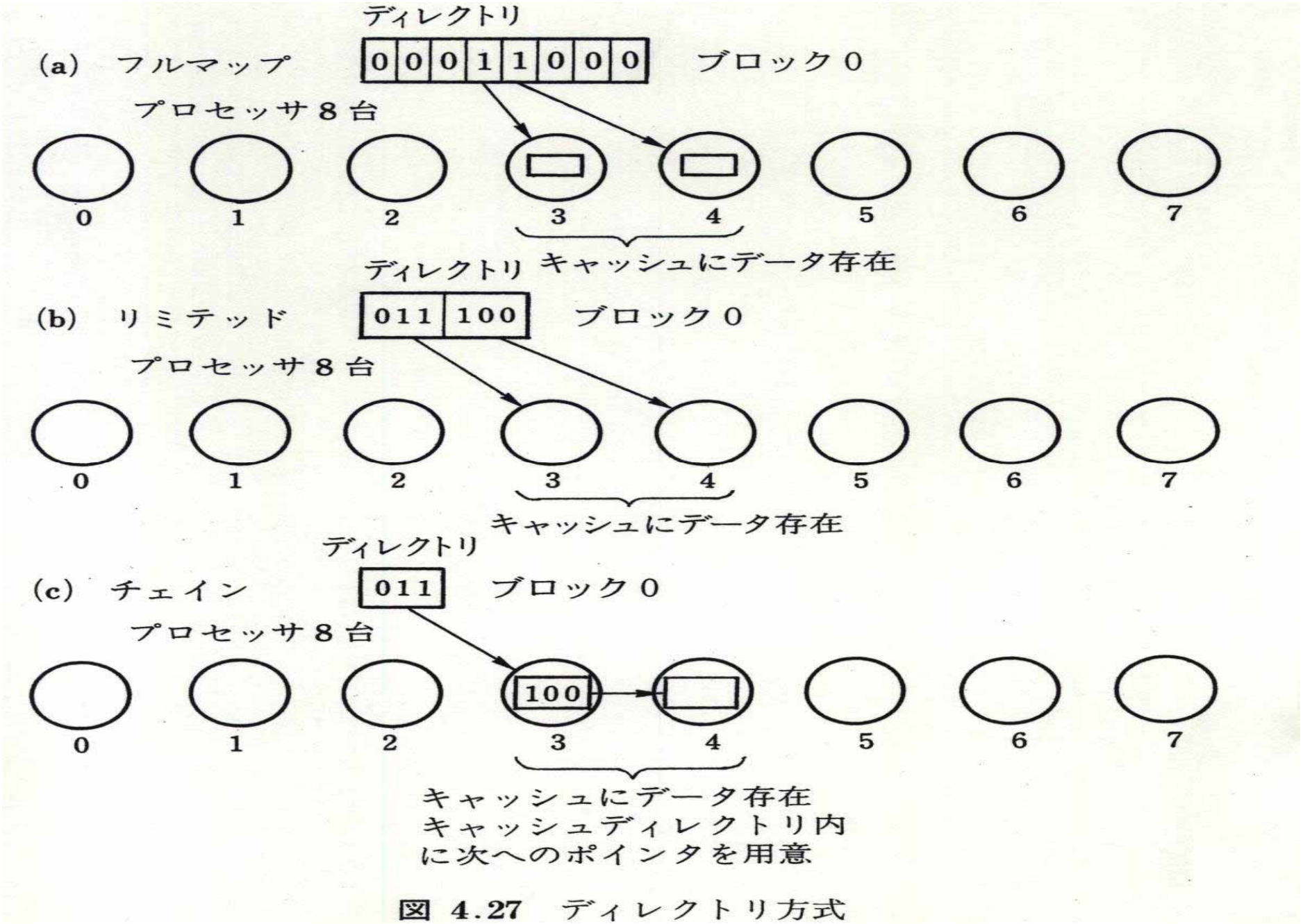
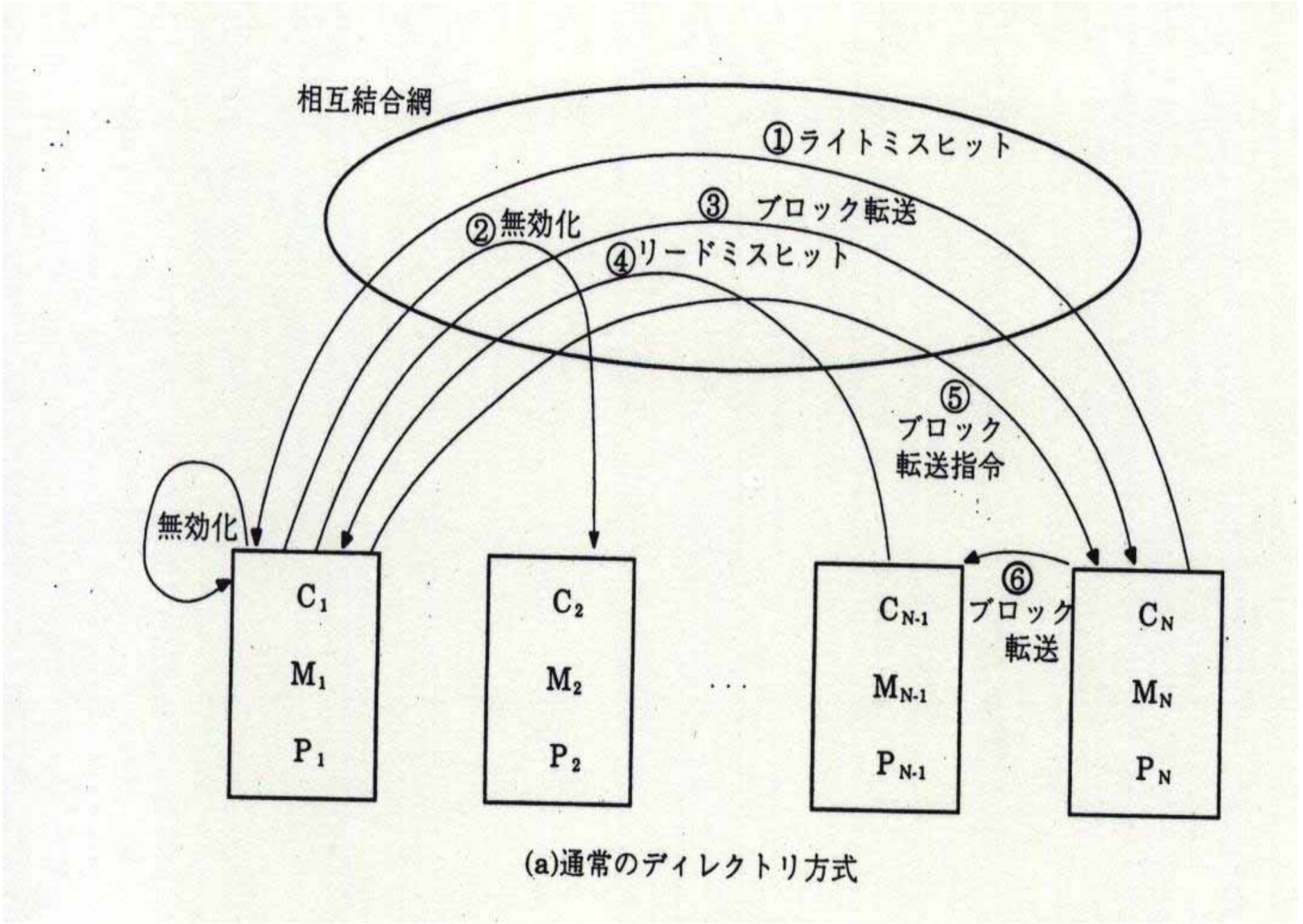
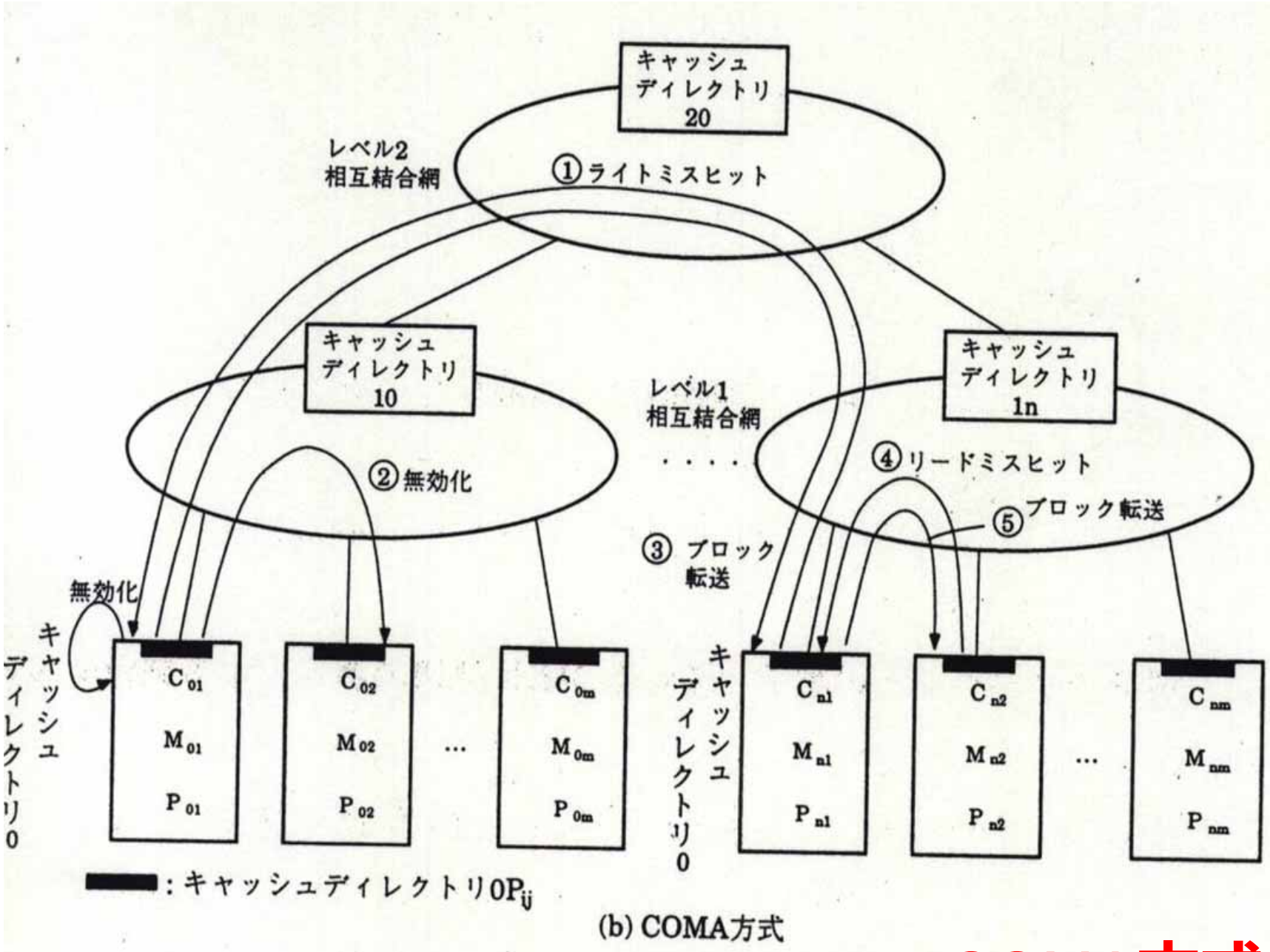


図 4.27 ディレクトリ方式



(a)通常のディレクトリ方式



(b) COMA方式

図5.11　キャッシュオンリメモリ方式

COMA方式

コヒーレンスの動作を応用例で見ると

線形1次方程式の反復解法

$AX=a \rightarrow X=b+BX$

$X=D^{-1}((D-A)X+a)$：ヤコビ法

直接法：ガウス消去法、LU分解法

スヌープキャッシュの2つのプロセッサで並列実行

同期命令が必要：先走り禁止

新しい値／古い価を使ってしまう

無効化／更新どちらがよいか

この場合は更新

コヒーレンス制御が必要のないものがある

```
      DO 30 I=1, M
      DO 20 J=1, 4
      XTEMP(J)=b(J)
      DO 20 K=1, 4
      XTEMP(J)=XTEMP(J)+B(J, K)*X(K)
20    CONTINUE
      ----------------------------------
      DO 10 J=1, 4
      X(J)=XTEMP(J)
10    CONTINUE
30    CONTINUE
```

エポック1

エポック2

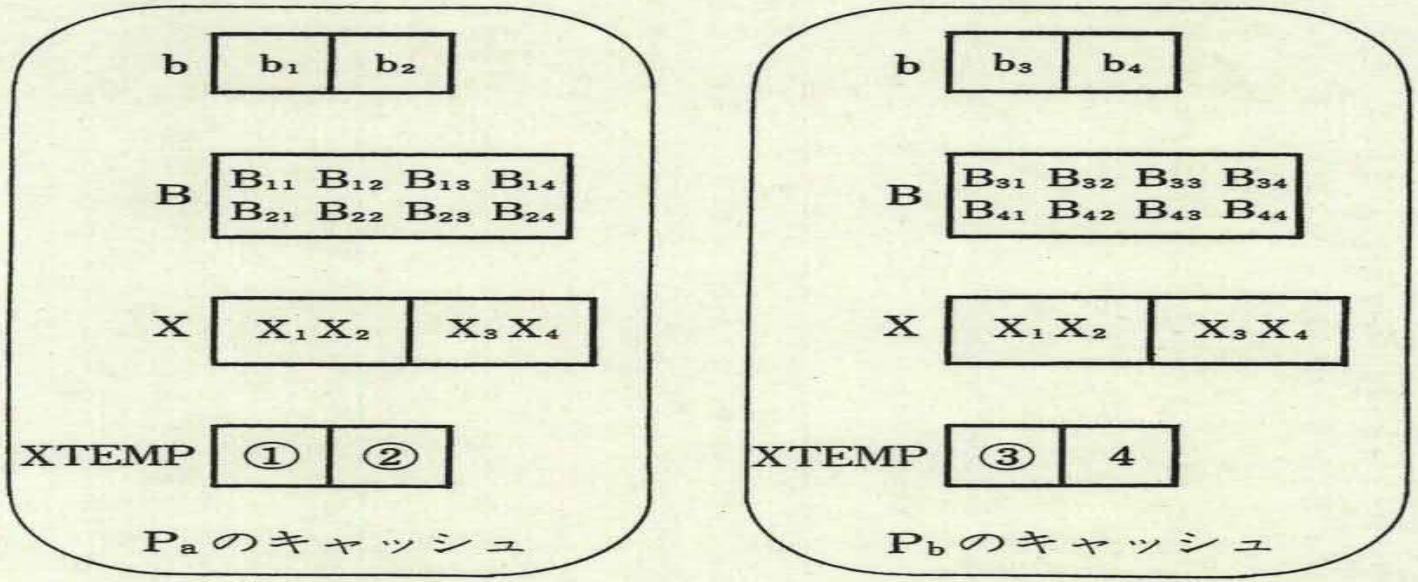
同期操作

同期操作

J=1,2：Paで分担
J＝3,4：Pbで分担

(a)　線形方程式の反復解法

$$\begin{pmatrix} X \\ \hline X_1 \\ X_2 \\ \hline X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} b \\ \hline b_1 \\ b_2 \\ \hline b_3 \\ b_4 \end{pmatrix} + \begin{pmatrix} B \\ \hline B_{11}B_{12}B_{13}B_{14} \\ B_{21}B_{22}B_{23}B_{24} \\ \hline B_{31}B_{32}B_{33}B_{34} \\ B_{41}B_{42}B_{43}B_{44} \end{pmatrix} \begin{pmatrix} X \\ \hline X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}$$



(b)　キャッシュの状態

図 4.28　ソウトウェアによるキャッシュコヒーレンス制御方式

4.4.5ソフトウェアによる方式

コンパイラで共有データを検出
　リードオンリ、排他的専有利用データ：対象外
ストアスルーを前提：最新データはメモリに存在、ハード簡単
ソフト的な選択的無効化
① マーキング法：共有データにマークを付け、キャッシュしない
②キャッシュ制御命令：共有データブロックをひとつ
　ひとつ命令で無効化
③ハードウェアによるキャッシュ全面的無効化
　同期命令でキャッシュ全面無効化
　リードオンリデータはCache　Read命令で復活、有効化
　共有データはMemory　Readでミスヒット、
　　メモリからキャッシュへ転送、その後ヒット
④ハードウェアによる選択的無効化機構の設置
　キャッシュブロックごとに無効化ビットを持たせ、一斉に選択的に無効化

５．３．５通信量の削減法

（１）データ属性に基づいたプロトコル切替え

①読出し専用（リードオンリ）データ

②専有データ

③無効化が有利なデータ

④更新が有利なデータ

⑤最終書込み時ブロードキャストが

有利なデータ

⑥バリア同期変数

⑦ロック変数

不可分操作

ＴＳ（Test and Set）命令

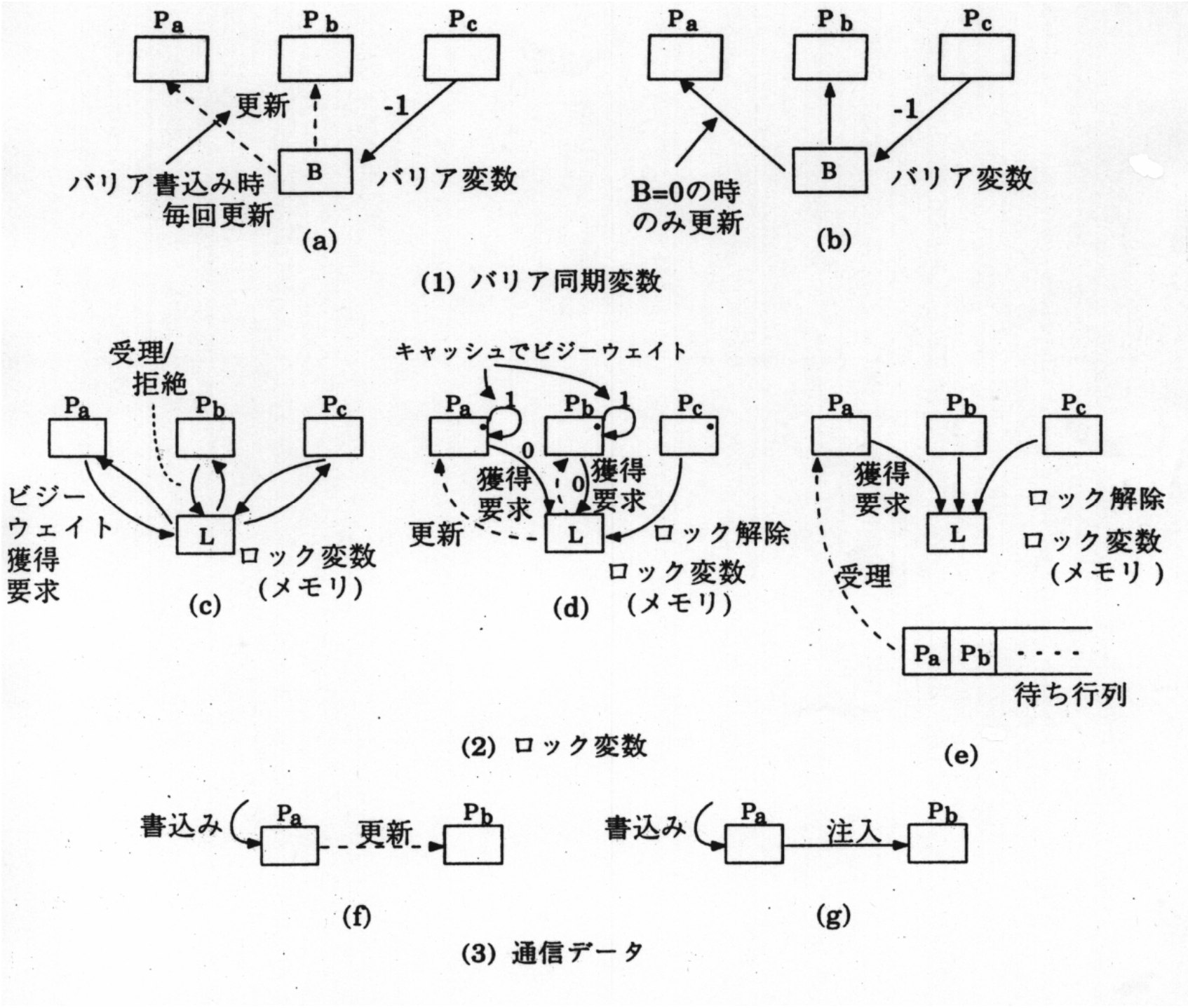
原始的な方式

ＴＳ命令で行う方式

ビジーウエイト

少し効率のよい方式

更新方式


Pa
Pb
Pc
更新
-1
バリア書込み時
毎回更新
B
バリア変数
(a)
B=0の時
のみ更新
(b)
(1) バリア同期変数
受理/
拒絶
ビジー
ウェイト
獲得
要求
L
ロック変数
(メモリ)
(c)
キャッシュでビジーウェイト
1
0
獲得
要求
更新
ロック解除
(d)
(e)
ロック解除
受理
待ち行列
(2) ロック変数
書込み
(f)
注入
(g)
(3) 通信データ

通常のロード命令でキャッシュからロック変数読み出し

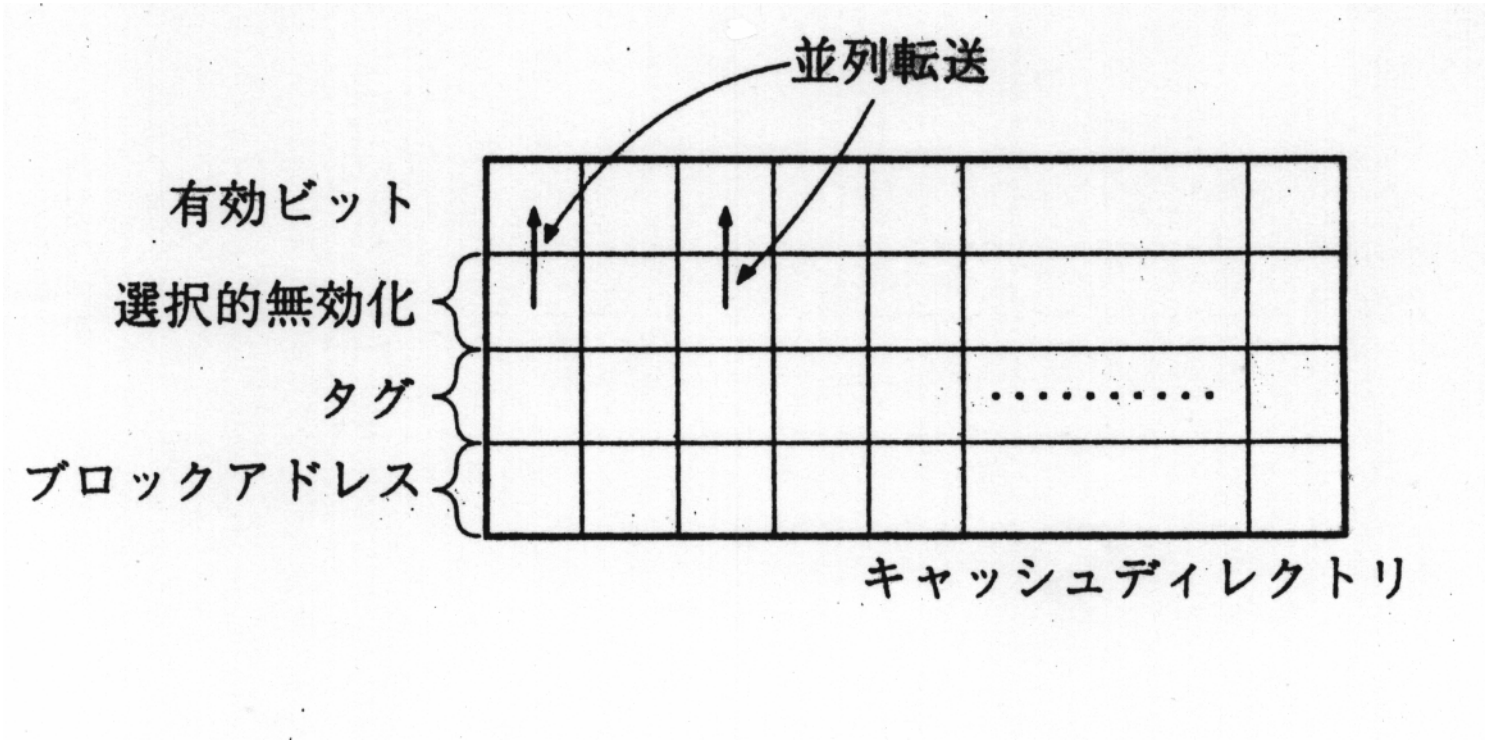
０であれば、ＴＳ命令によるロック獲得

要求の待ち行列

⑧通信データ

注入（injection）

（２）キャッシュ管理の積極的導入


並列転送
有効ビット
選択的無効化
タグ
ブロックアドレス
キャッシュディレクトリ

５.３.６キャッシュ

オンリメモリ

プログラムやデータ：

システム内のどこかのキャッシュに分散格納

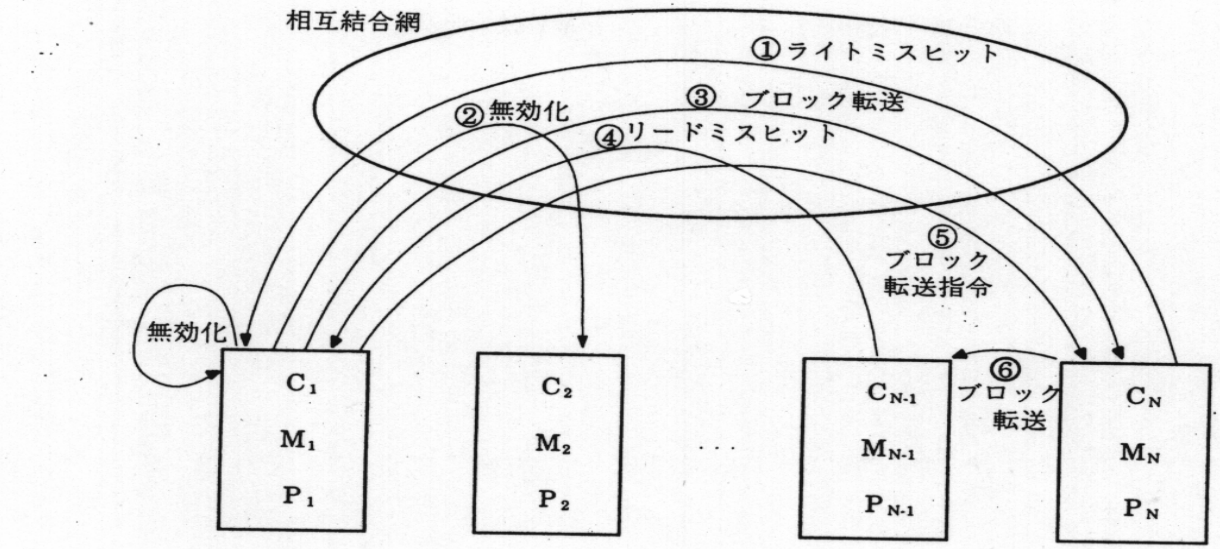
階層構造

キャッシュディレクトリのハードウェア構成：

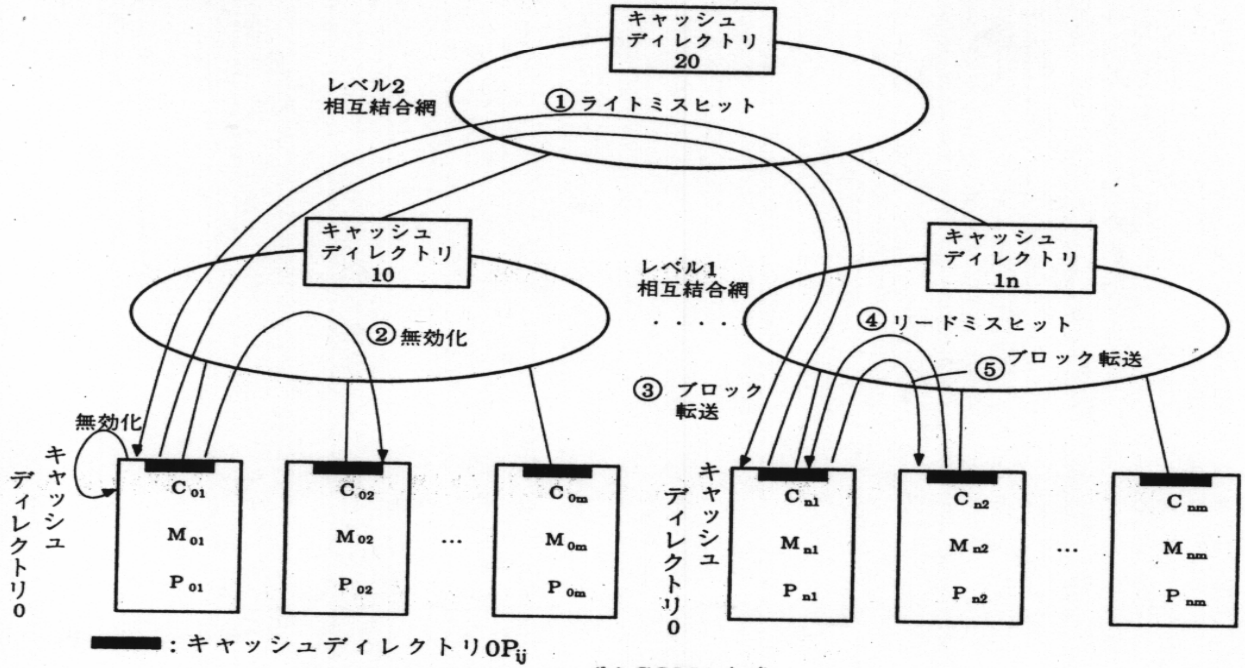
連想メモリで構成

上位包含性（multi-level inclusion**）**

データ参照の局所性



(a)通常のディレクトリ方式



(b) COMA方式

図 5. 11　キャッシュオンリメモリ方式

５．４ＪＵＭＰ－１

JUMP-1、スタンフォード大学のDASH、

MITのAlewifeなど

JUMP-1

クラスタ構成：１２８クラスタ

クラスタ：

４台のSuper SPARC Plus（動作周波数は６０MHz）

２台のメモリ（各６４MB）、

２台の知的メモリ機構（MBP）

数本の高速シリアルリンク（STAFF-Link）

プロセッサ：各１MBの２次キャッシュ付き

階層トーラス網（以後、RDT網とよぶ）

５．４．１特長

（１）キャッシュコヒーレンス制御機構を

内蔵したメモリ共有方式

（２）通信オーバヘッドの削減を図れる

MBP機構の導入による細粒度並列処理方式

①レイテンシ削減

②無駄な通信の削減

（３）スケーラブルで直径の小さい階層トーラス網

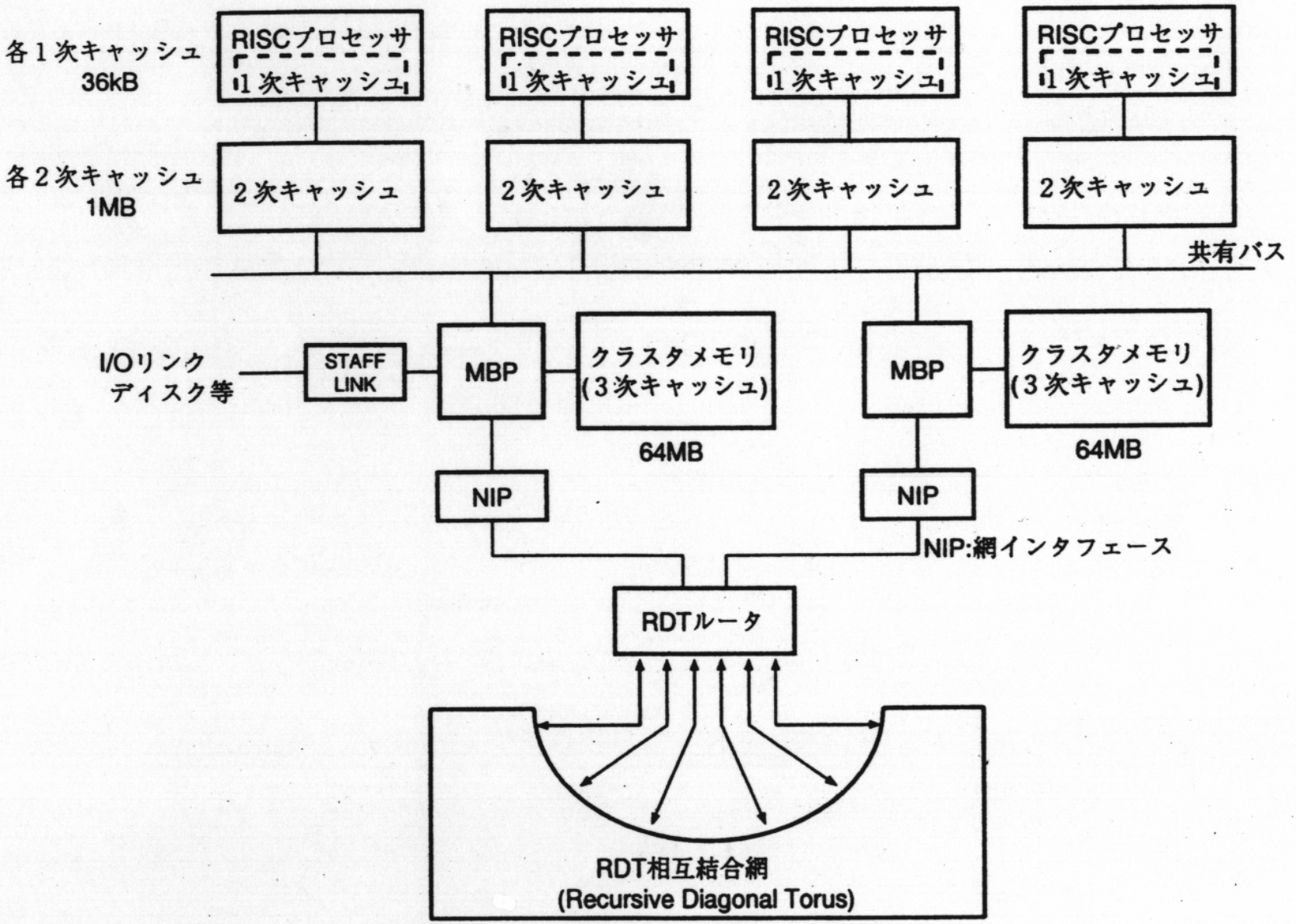


図 5. 12　JUMP-1 のクラスタ内構成

（4）高速な入出力機構

5.4.2メモリアーキテクチャ

（1）仮想共有メモリ

クラスタメモリ：

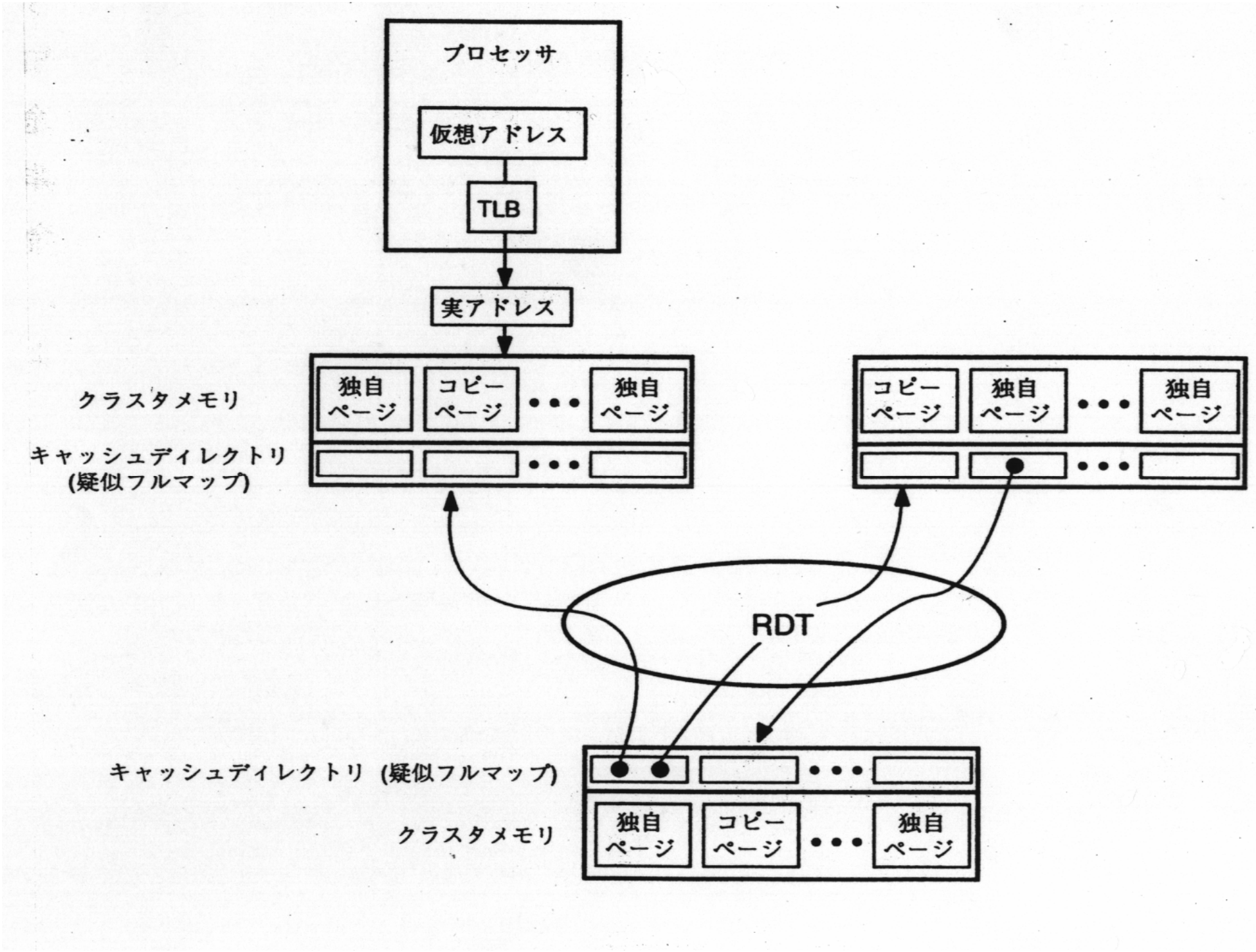
独自（original）空間とコピー（copy）空間

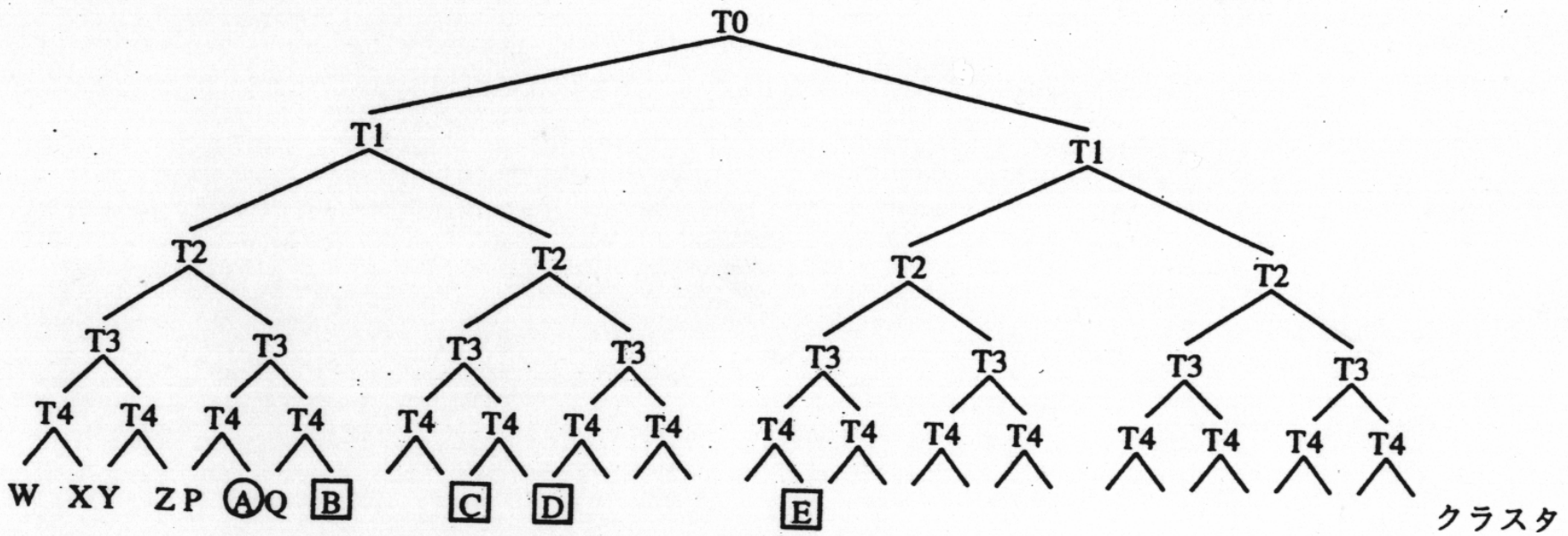
３次キャッシュ

疑似フルマップ方式

（2）クラスタ間キャッシュコヒーレンス

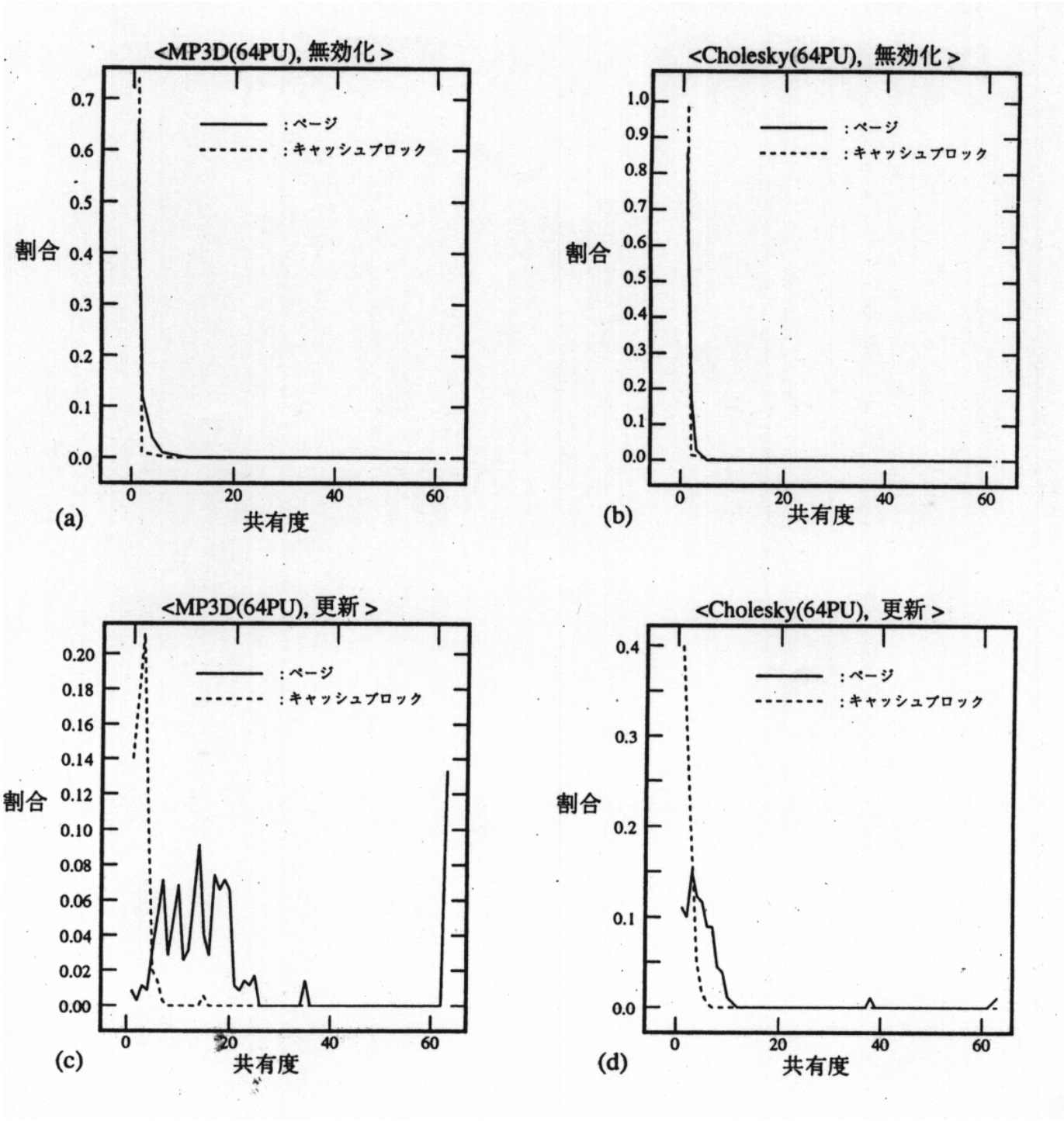
(a)疑似フルマップ方式

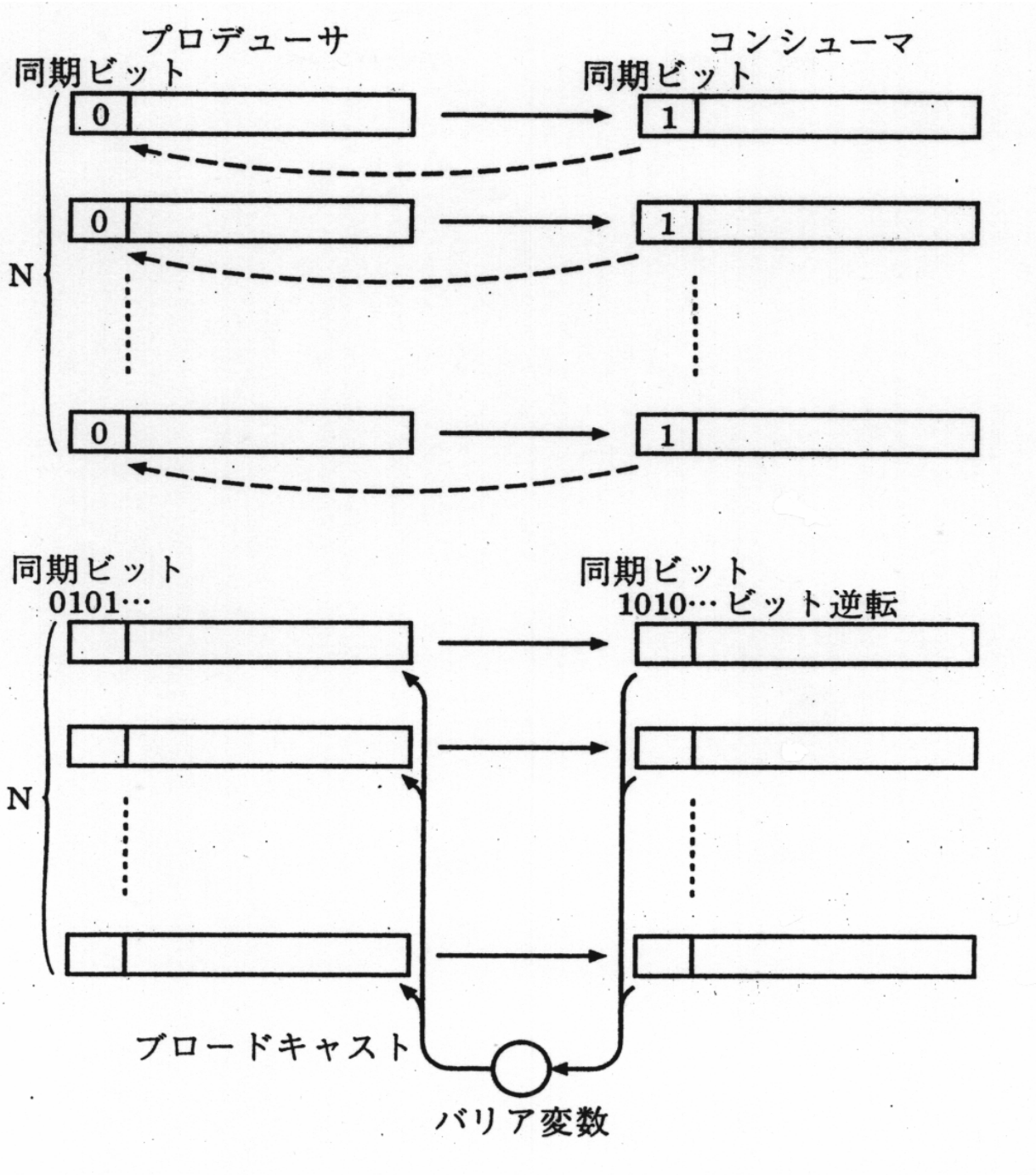

プロセッサ
仮想アドレス
TLB
実アドレス
クラスタメモリ
独自ページ
コピーページ
独自ページ
キャッシュディレクトリ
(疑似フルマップ)
コピーページ
独自ページ
独自ページ
RDT
キャッシュディレクトリ (疑似フルマップ)
クラスタメモリ
独自ページ
コピーページ
独自ページ



	T4	T3	T2	T1	T0	MD
A->B	0:0	0:1	0:0	0:0	0:0	0
A->B,C,D	0:0	0:1	0:0	0:1	0:0	0
A->B,E	0:0	0:1	0:0	0:0	0:1	0
A->ブロードキャスト	1:0	0:1	1:0	0:1	0:1	0
A->E	0:1	1:0	1:0	1:0	0:1	1

図 5. 14　疑似フルマップ [188]


<MP3D(64PU), 無効化 >
<Cholesky(64PU), 無効化 >
<MP3D(64PU), 更新 >
<Cholesky(64PU), 更新 >
:ページ
:キャッシュブロック
割合
共有度
(a)
(b)
(c)
(d)


プロデューサ
コンシューマ
同期ビット
同期ビット
0
1
0
1
N
0
1
同期ビット
0101…
同期ビット
1010… ビット逆転
N
ブロードキャスト
バリア変数

放送機構

結合操作（Combining）

(b)疑似フルマップ方式の評価

・共有ブロックの大きさ

・書込み時のポリシ

リミテッド、チェイン方式と比較

・ディレクトリ容量

・ネットワークを流れるトラフィック量

５．４．３高速同期・通信機構

（１）キャッシュ

タグ：INV(Invalid),EX-Dty(Exclusive Dirty),
LS-Cln(Locally Clean),
LS-Dty(Locally Dirty),
GS-Cln(Globally Clean)の５状態

Ｉストラクチャのfull/emptyビット
（８バイトごと）

２次キャッシュの特長

・ページ属性として無効化と更新
データ属性に応じたプロトコル選択

・実アドレスの上位５ビットをコマンドとして

使用

①プリフェッチと注入

注入の例：FIFOキューの先頭が消費された場合

リードオール

②更新データのマージ

③早期共有解除

④自浄機構

（２）キャッシュコヒーレンスプロトコルの動的切り替え

（３）Ｉ－ストラクチャとFIFO機構

①１対１通信

Ｉ－ストラクチャ

異なるプロセッサ間（Ｎ個）で１対１通信

Ｎ個のコンシューマへのデータ到着

バリア同期機構で高速に検知

各プロデューサに放送

ビット逆転方式（最初のデータ転送では

０、次の転送では１、その次では０、

でデータの利用可能状態を示す）

②１対多通信

③多対１通信

FIFOキュー

④多対多通信

（４）バリア機構

５．４．３ネットワーク アーキテクチャ

ブロードキャスト

ランク０グループ

グループでランク１グループ

８進木構成

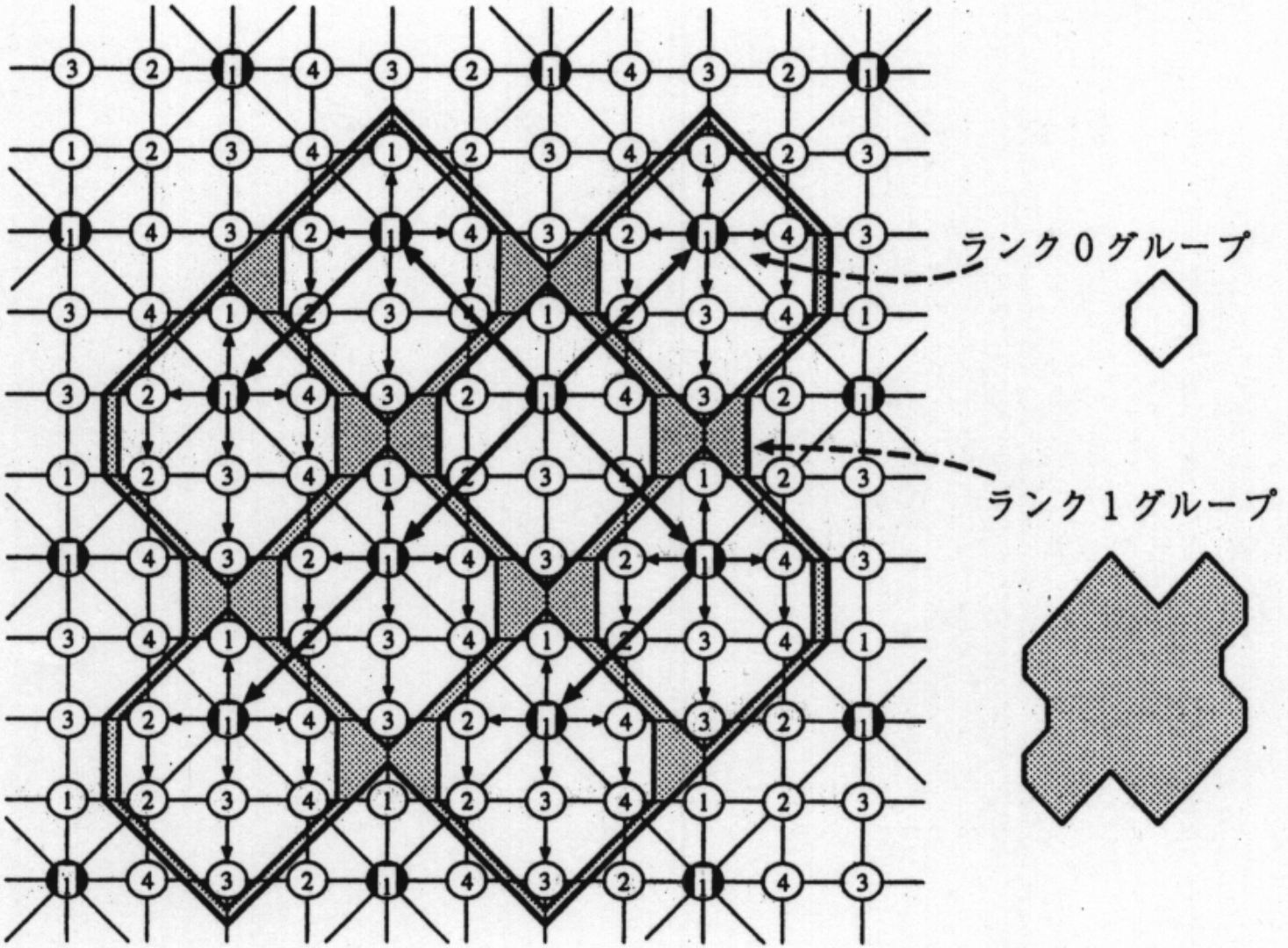
ブロードキャスト､バリア同期や疑似フルマップの制御

ルーティング：ベクトルルーティング

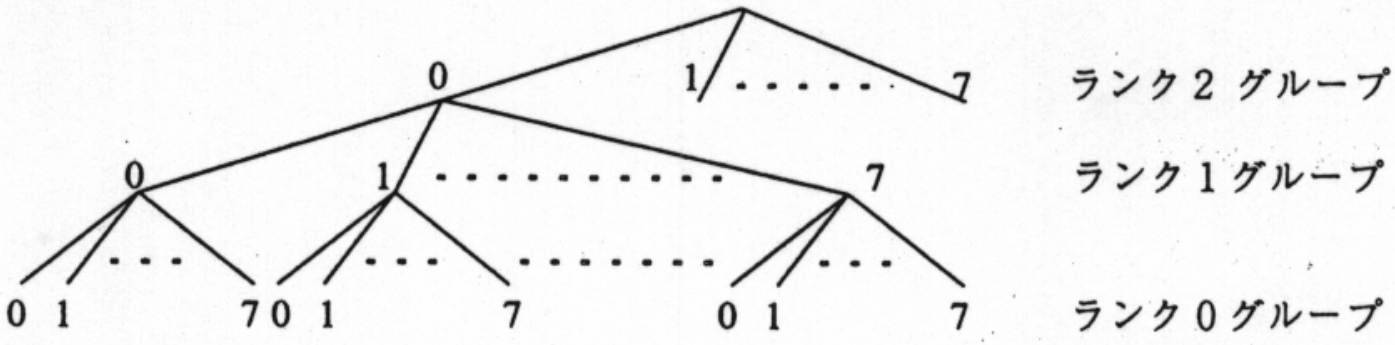
RDTの性能：

３次元トーラスとハイパキューブの中間的な特性

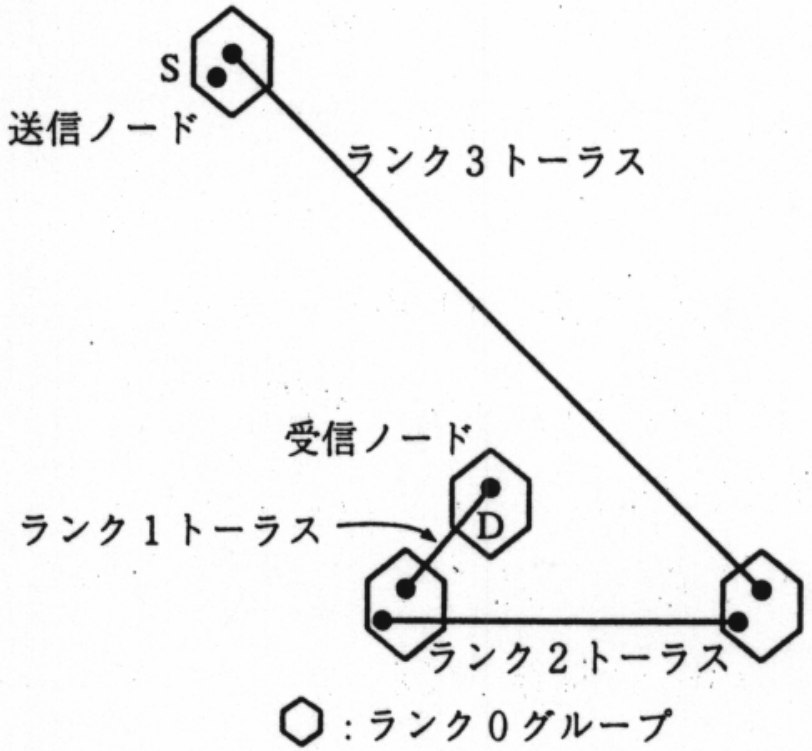
ハイパキューブ網やシャフル網の高速にシミュレート



(a) RDTの構造



(b) 8 進木構成



(c) ベクトルルーティング

５．５メッセージ交換型マルチプロセッサ

ＯＳ

１ ＯＳの必要性

① 複雑な機能の提供：命令セットの拡張

大容量記憶，ファイル操作，

ネットワーク，入出力

②リソースの共有利用：効率的な管理，

セキュリティ

多数のユーザプロセスの実行

多数のサーバプロセスの実行

２ ＯＳの機能

①記憶管理

　多重仮想記憶

②ファイル管理

　ディレクトリ管理，ファイル更新･編集

　効率のよい記憶方式

③入出力管理

　デバイスドライバ

　スケジューリング

①記憶管理

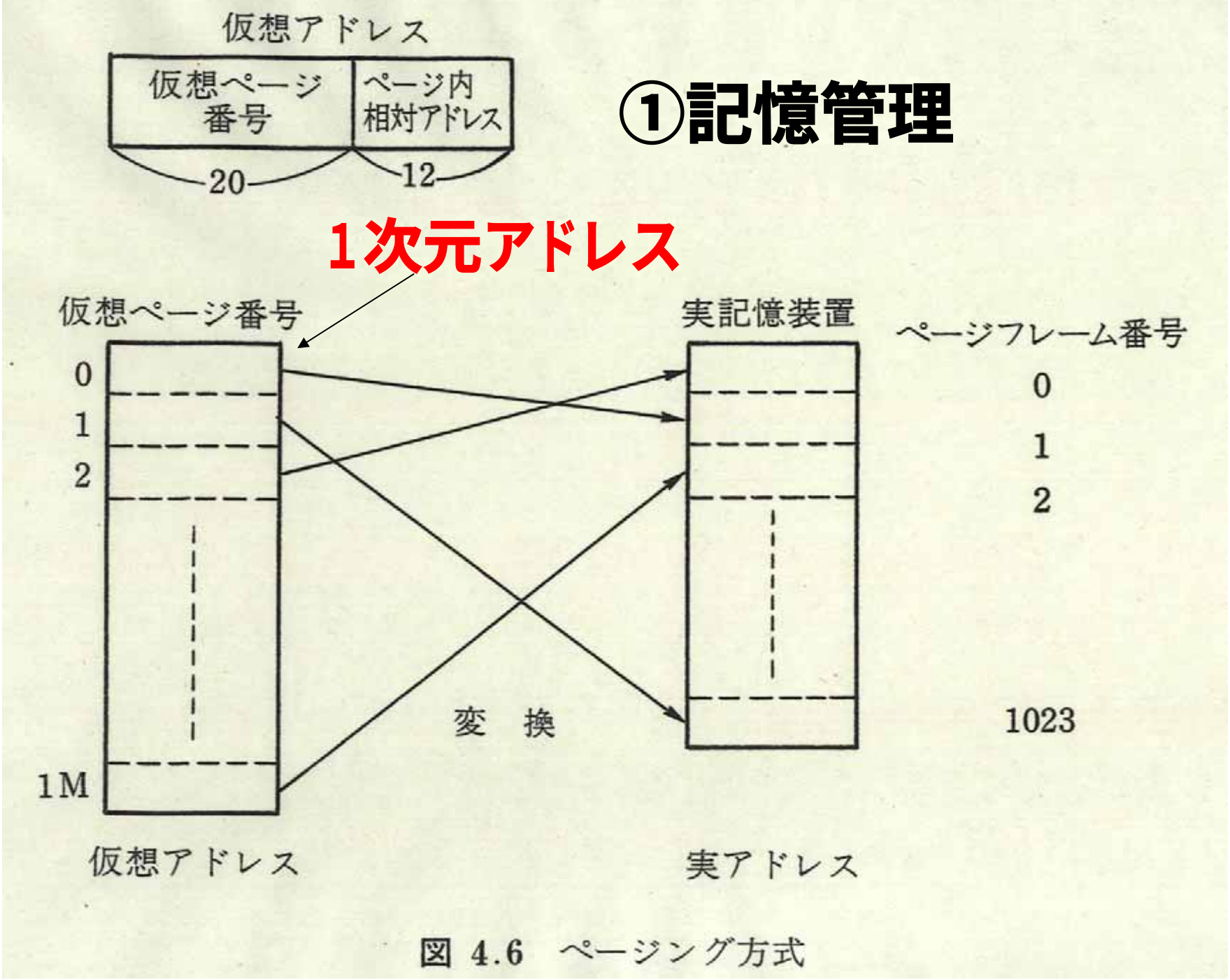
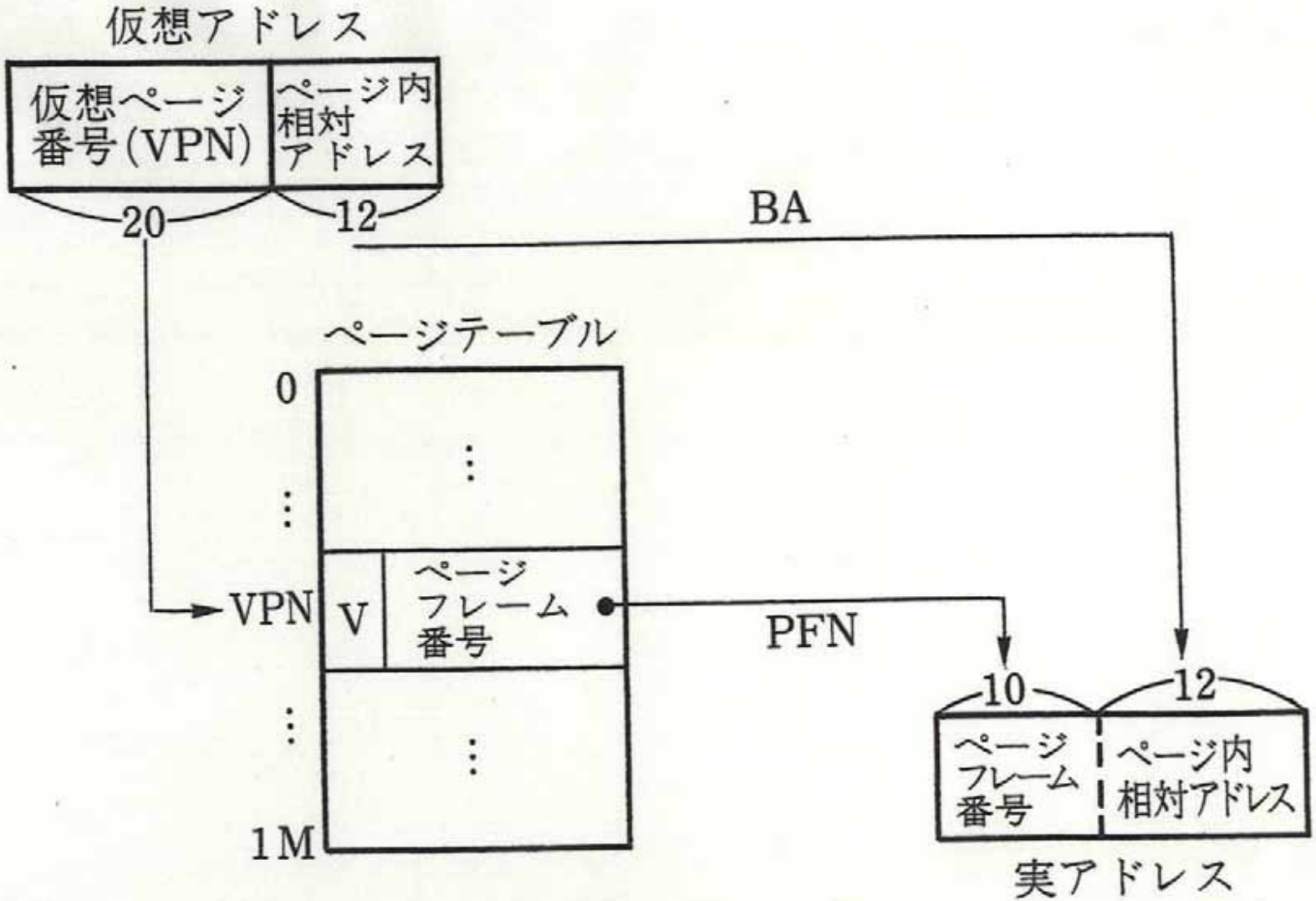


図 4.6　ページング方式



(a) ページテーブル

図 4.8 直接写像方式

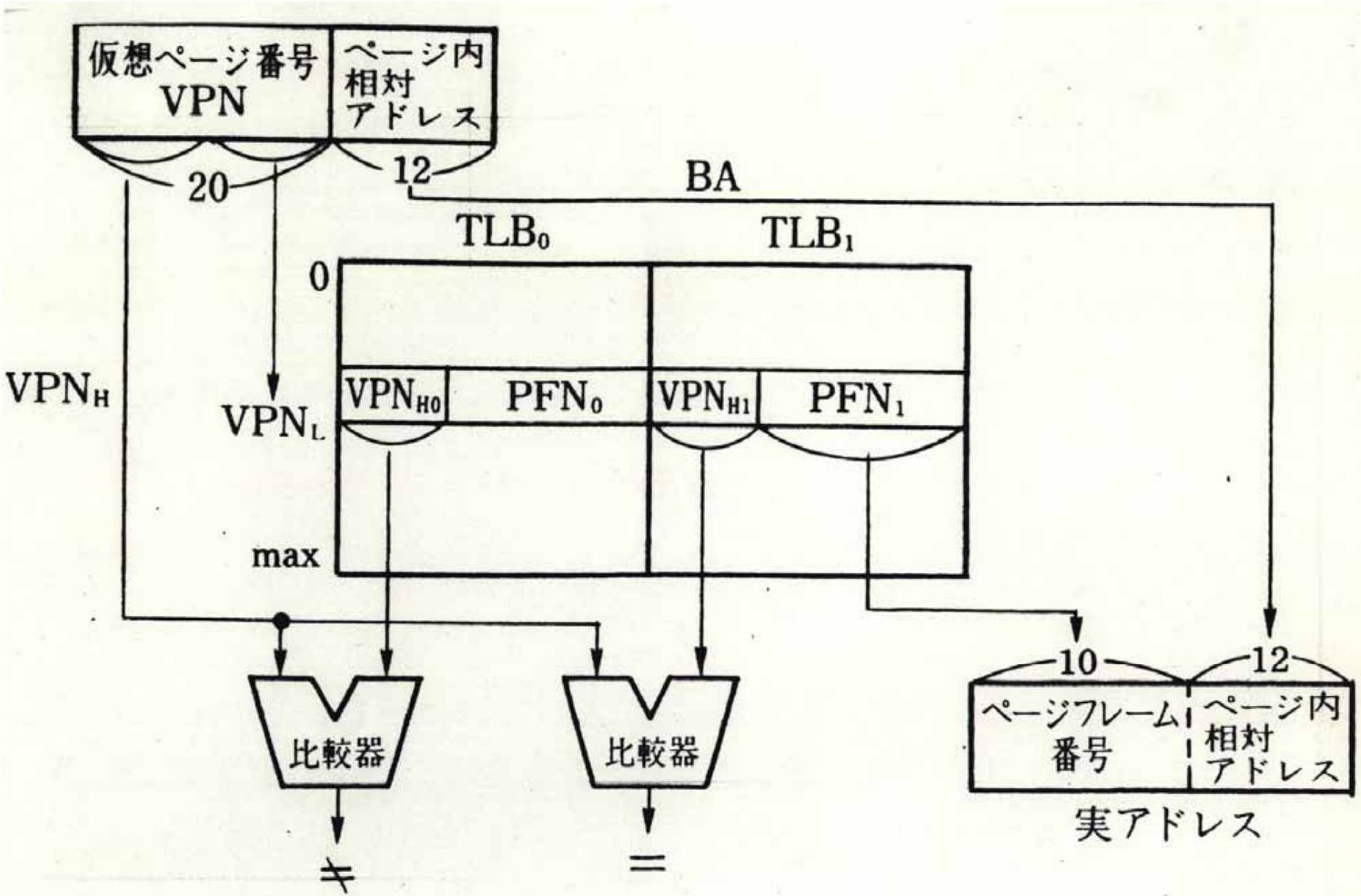

仮想ページ番号
VPN
ページ内
相対
アドレス
20
12
BA
TLB0
TLB1
0
VPNH
VPNL
VPNH0
PFN0
VPNH1
PFN1
max
比較器
比較器
キ
ニ
10
12
ページフレーム
番号
ページ内
相対
アドレス
実アドレス

図 4.11　TLB の構成

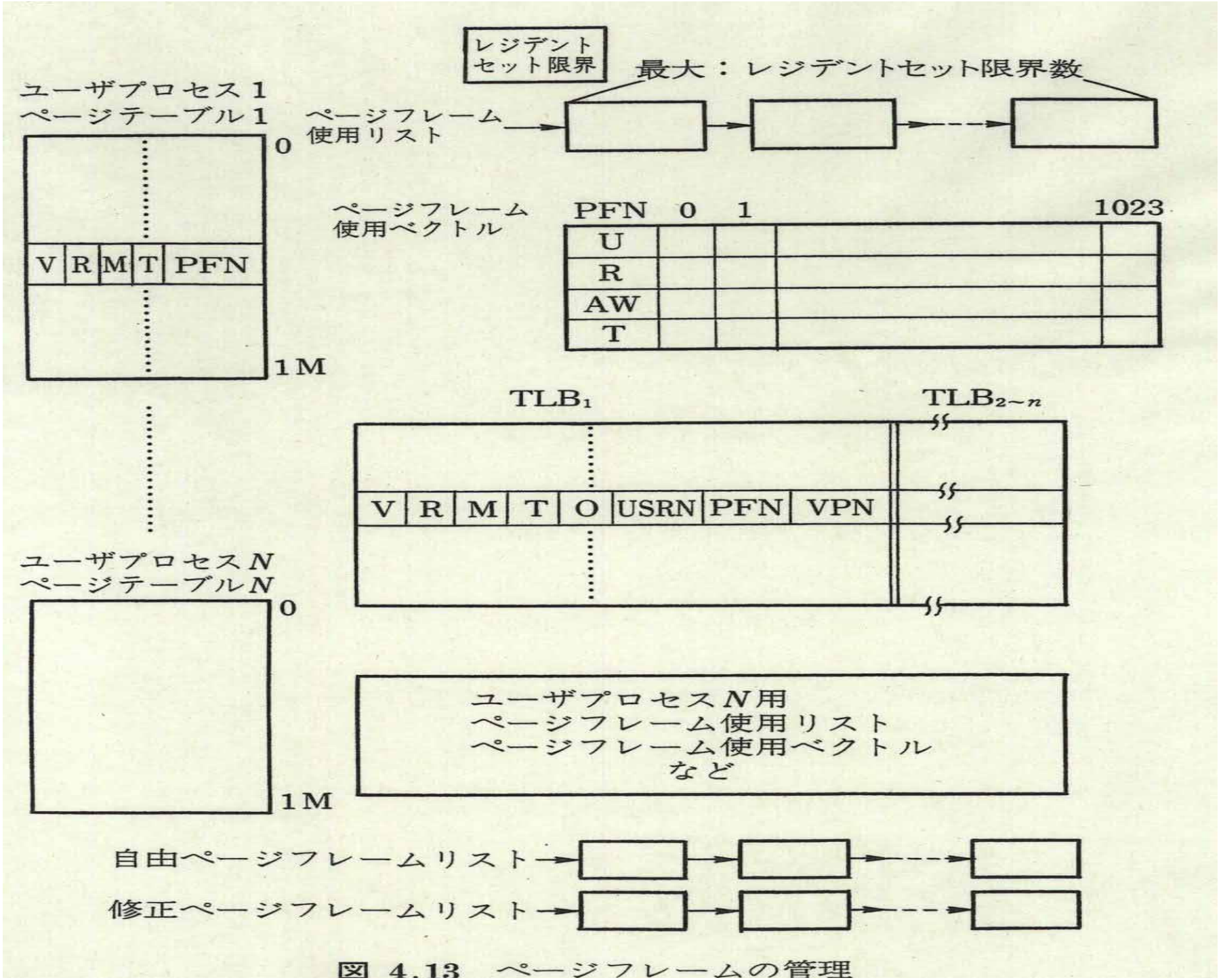


図 4.13 ページフレームの管理

②ファイル管理

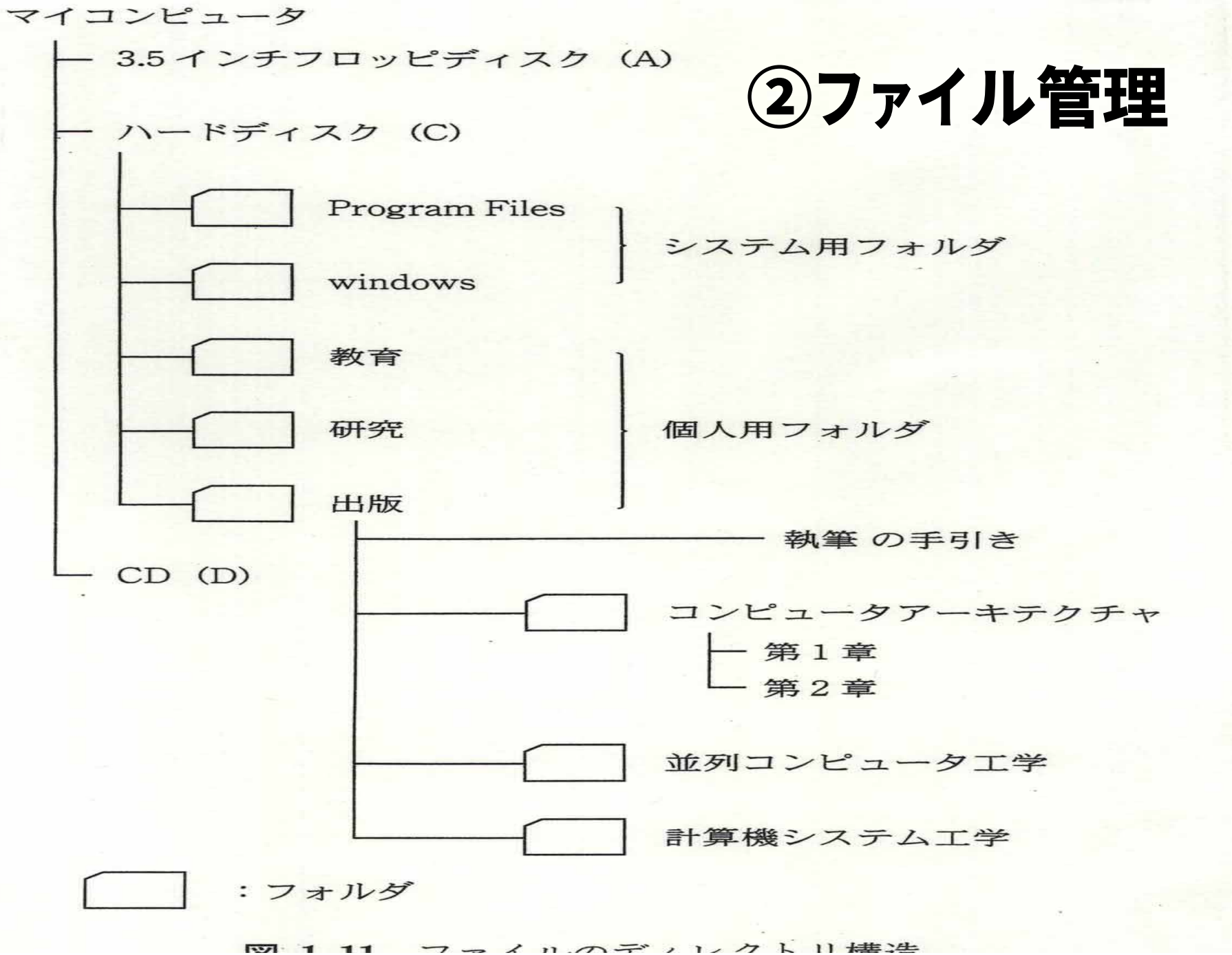


図 1.11 ファイルのディレクトリ構造

④通信管理

アプリケーション層

FTP, Telnet, WWW, SMTP, . .

TCP／UDP層

TCP：コネクション型

バーチャルチャネル

UDP：コネクションレス型

IP層

中継ノード間での通信制御

IPアドレス

インタフェース層

マルチキャスト機能

物理層

イーサネット, ATM

⑤プロセス管理
ＯＳの中のＯＳ

たくさんのプロセスの擬似的な並行実行

プロセス：ＯＳにより管理実行されているプログラム

並列プロセスの例

長時間かかるＣプログラムを実行させながら

ＷＯＲＤで文書作成し，プリントアウト

プリンタが動作し始め

ＥＸＣＥＬで表計算処理を開始していると

時計の表示

電子メールの受信通知の表示

プロセスの状態

実行中：プロセスがＯＳからプロセッサを
　　割り当てらてれ実行中の状態

レディ：プロセスは実行可能状態であるが，
　　ＯＳのプロセッサ割当てがなされて
　　いない状態

待ち：プロセスに必要なデータが揃って
　　いないので，待たねばならない状態

プロセススイッチ

契機：割込み

外部割込み：入出力，タイマ

マシンチェック

内部割込み：演算例外，

命令例外，ページ

フォルト，トレース

スーパバイザコール

多重プログラミング，

ＴＳＳ（Time Sharing System）

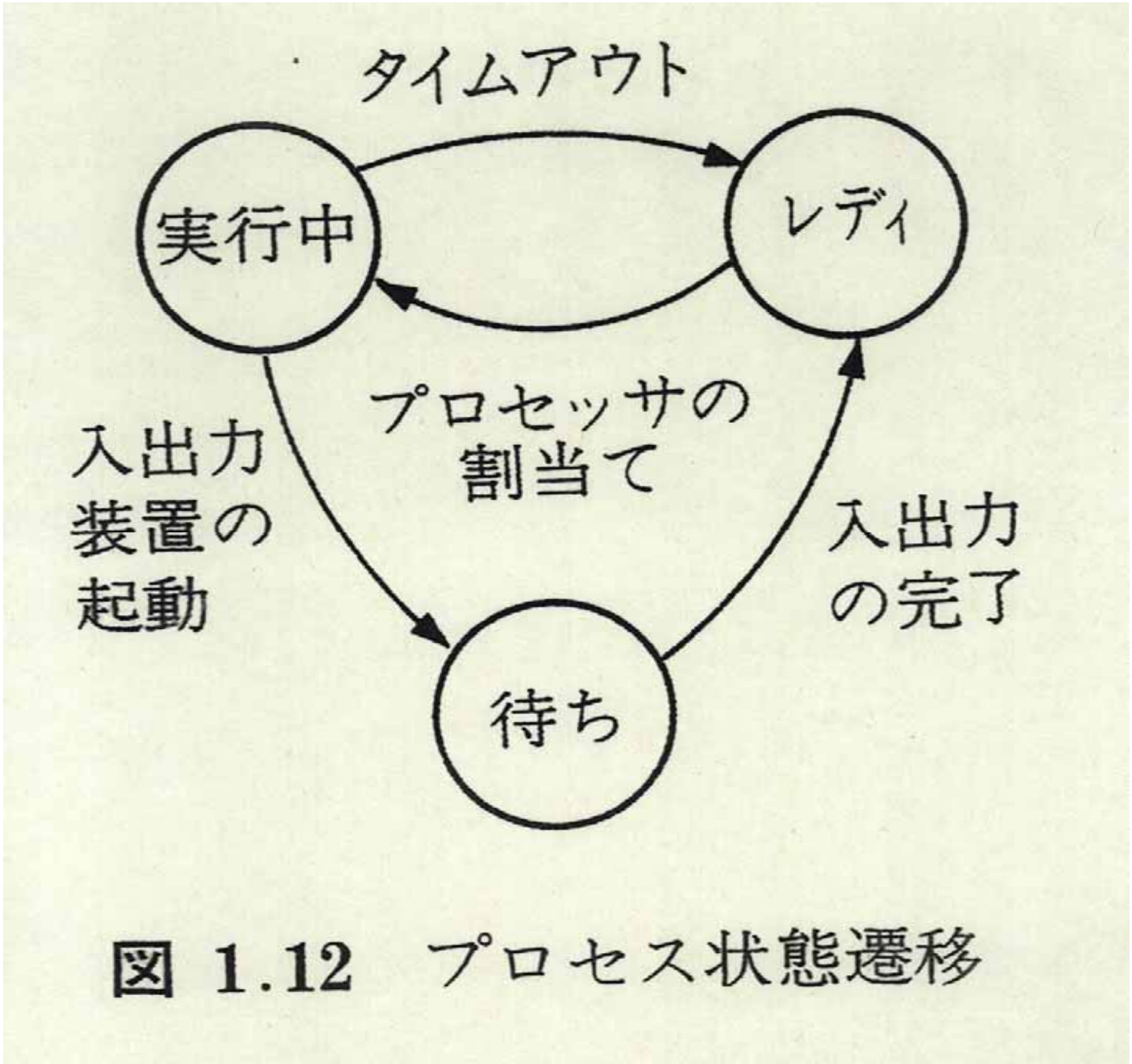
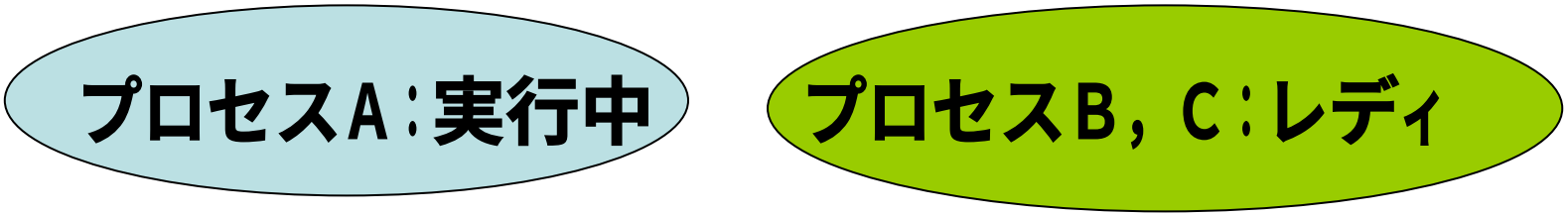
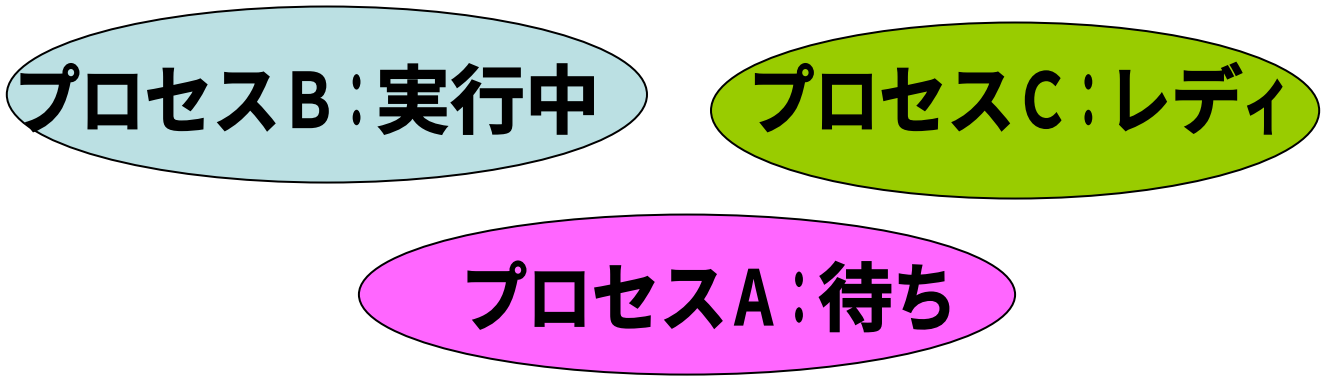


図 1.12　プロセス状態遷移

①プロセスA，B，C：レディ
　ＯＳ：プロセスAを選択し，実行させる



②プロセスA：ディスクアクセス
　Ｉ／Ｏ命令実行でＯＳに割込み
ＯＳ：プロセスAのＩ／Ｏ処理し，Aを待ちへ
　プロセスBを選択し，実行へ

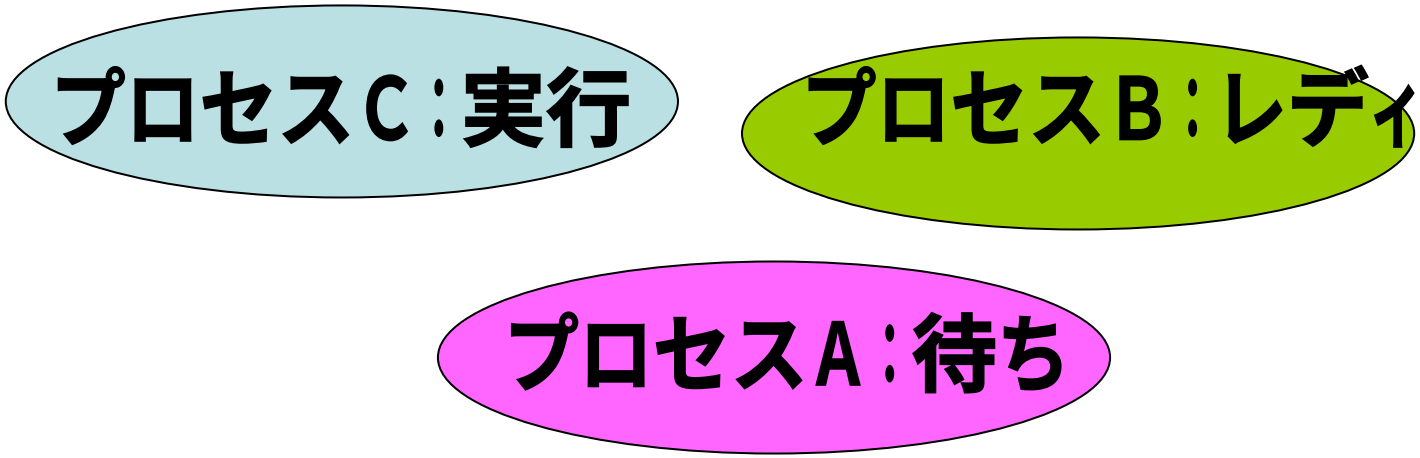


プロセスB：一定時間実行（タイムクアンタム）

１０msec

タイマー割込み

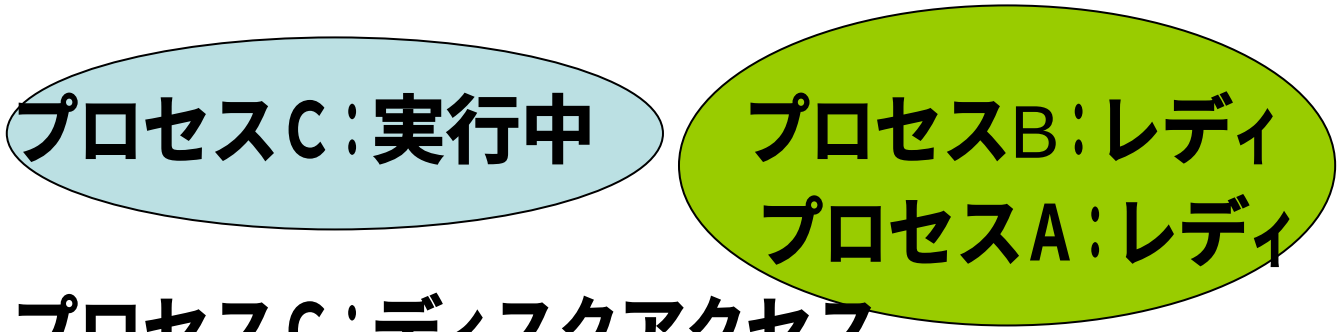
ＯＳがBをレディへ，Cを選択



ディスクからI／O割込み

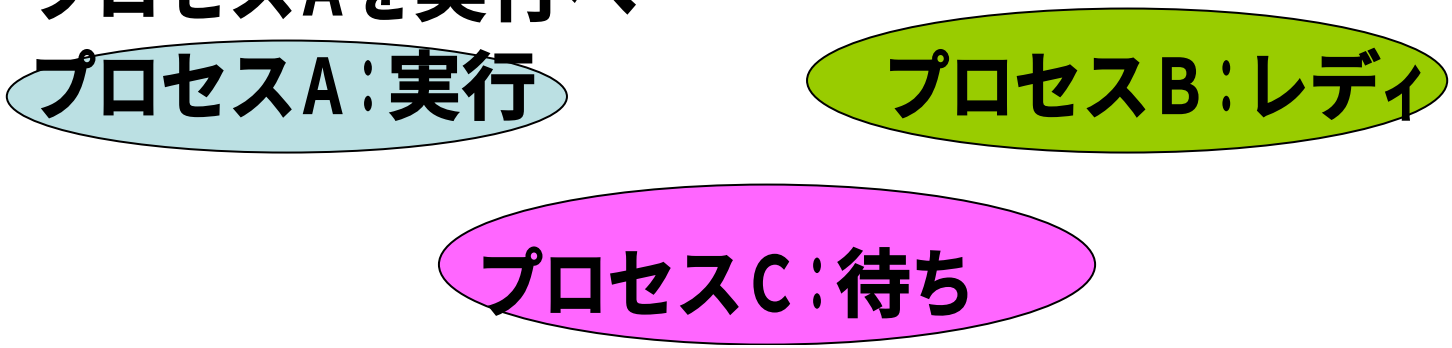
ＯＳがチェックし，プロセスAのディスク
アクセス終了を知る

ＯＳはプロセスAをレディへ

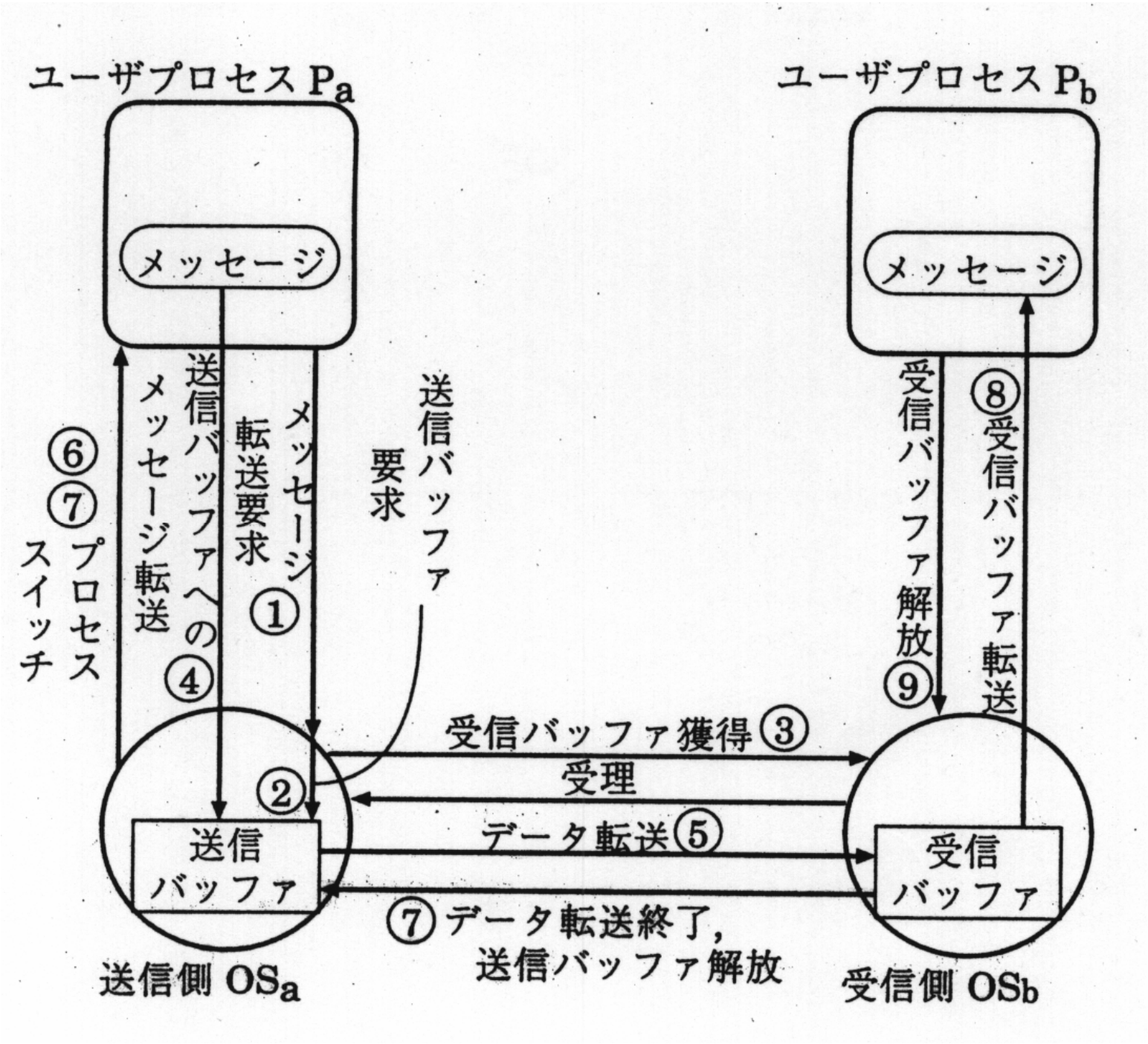


プロセスC：ディスクアクセス

ＯＳ：プロセスCを待ちへ，
プロセスAを実行へ



- プロセススケジューリング
- プロセス間通信
  - 同期：排他制御，事象待ち
  - メッセージ交換
  - メモリ空間共有，非共有


ユーザプロセス Pa
ユーザプロセス Pb
メッセージ
メッセージ
送信バッファへのメッセージ転送 ④
メッセージ転送要求 ①
送信バッファ要求
⑥⑦ プロセススイッチ
受信バッファ解放 ⑨
⑧ 受信バッファ転送
受信バッファ獲得 ③
受理
②
データ転送 ⑤
送信バッファ
受信バッファ
⑦ データ転送終了，送信バッファ解放
送信側 OSa
受信側 OSb

ユーザプロセスでは

①Paよりスーパバイザコール（ＯＳ呼出し）で

メッセージ転送要求

送信側OSaでは

②送信バッファ獲得

③受信側OSbに受信バッファ獲得依頼

④Paのメッセージを送信バッファに転送

⑤送信バッファのメッセージのパケット化。

OSbの受信バッファに相互結合網を通して転

送

⑥プロセススイッチでPaは待機状態

⑦データ転送終了。OSbより割込みを受理。

送信バッファ解放。Pa実行可状態

受信側OSbでは

受信バッファにデータ転送後

外部割込みによって起動

⑧受信バッファよりPbの所定の領域にデータ

転送

⑨受信バッファを解放

⑩Pbを実行可状態

（４）メッセージ処理の高速化

メッセージのコピーによるオーバヘッド

割込み時でのレジスタ退避などのオーバヘッド

相互結合網での各種受理信号（ACK信号など）のための通信回数の増加

高速化の基本

ユーザプロセス自身によって通信を制御

OSのオーバヘッドを削減

①通信用プロセッサの設置

通信処理：汎用マイクロプロセッサでよいか？

・通信自体がイベント起動で、処理の切換え頻繁

・データが一過性。キャッシュメモリの参照の局所性に不適合

②通信処理とデータ転送のオーバラップ

③資源の専有

④無駄なメッセージコピーの削減

⑤種々の転送モードの用意

⑥キャッシュ無効化の高速化

⑦割込み回数の削減

（５）放送機能の強化

①１対多

②多対多

（６）同期操作の高速化

バリア同期は

①バリア到達の通知

②バリア解除までに待機

ユーザレベル通信
①OSの介在の少ない方式
②Zero－Copy
②仮想記憶によるセキュリティ確保

(1)DMAを用いた方式

①固定DMA領域にコピー：AM-II,Hamlyn

②軽いアドレス変換カーネルを毎回起動、BIP,LFC

③TLBキャッシュ：VMMC-2,U-NET

(2) プログラムモードバス方式

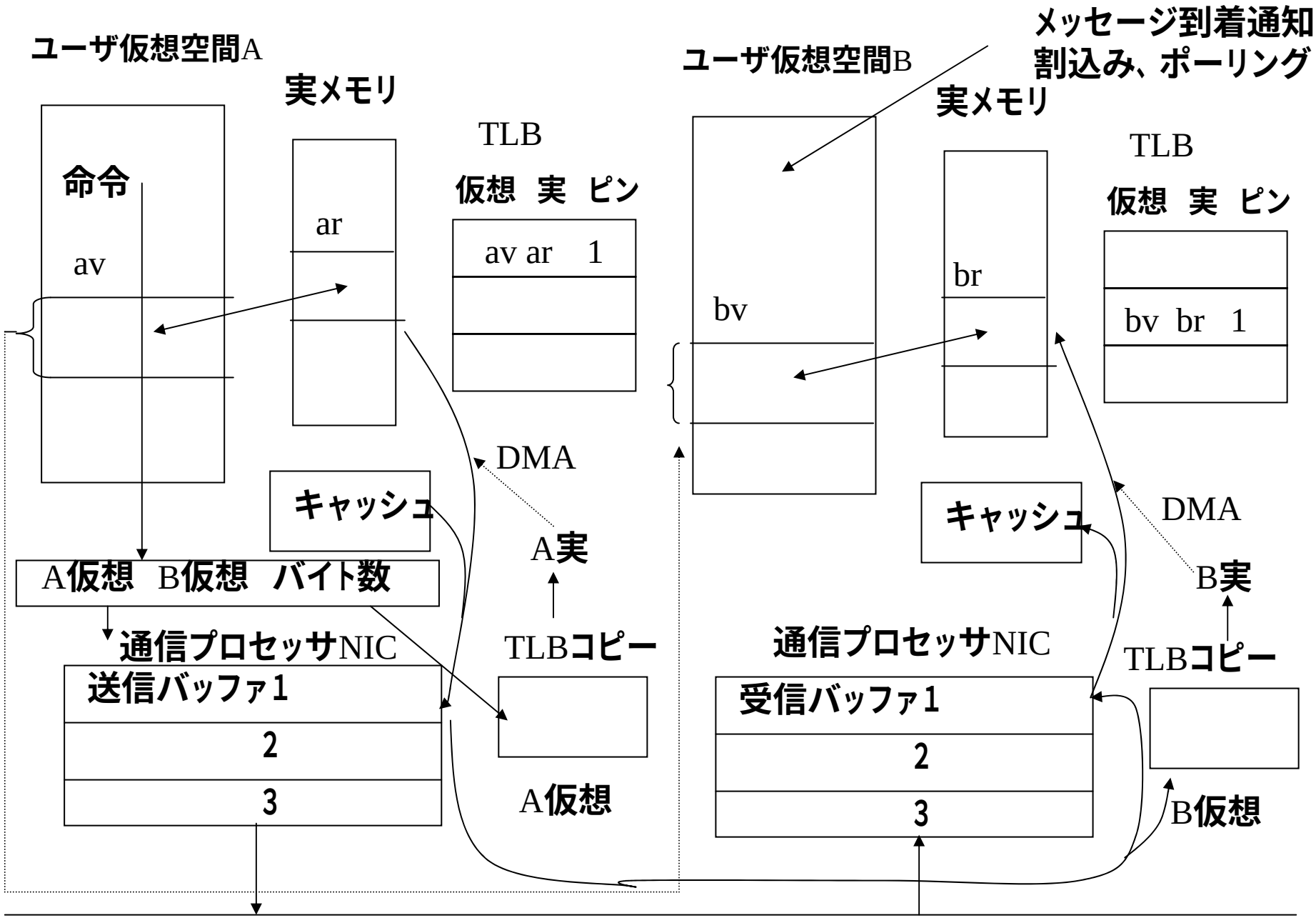
④Write Combining あり：FM,LFC,AM-II,Hamlyn,BIP

⑤同上なし

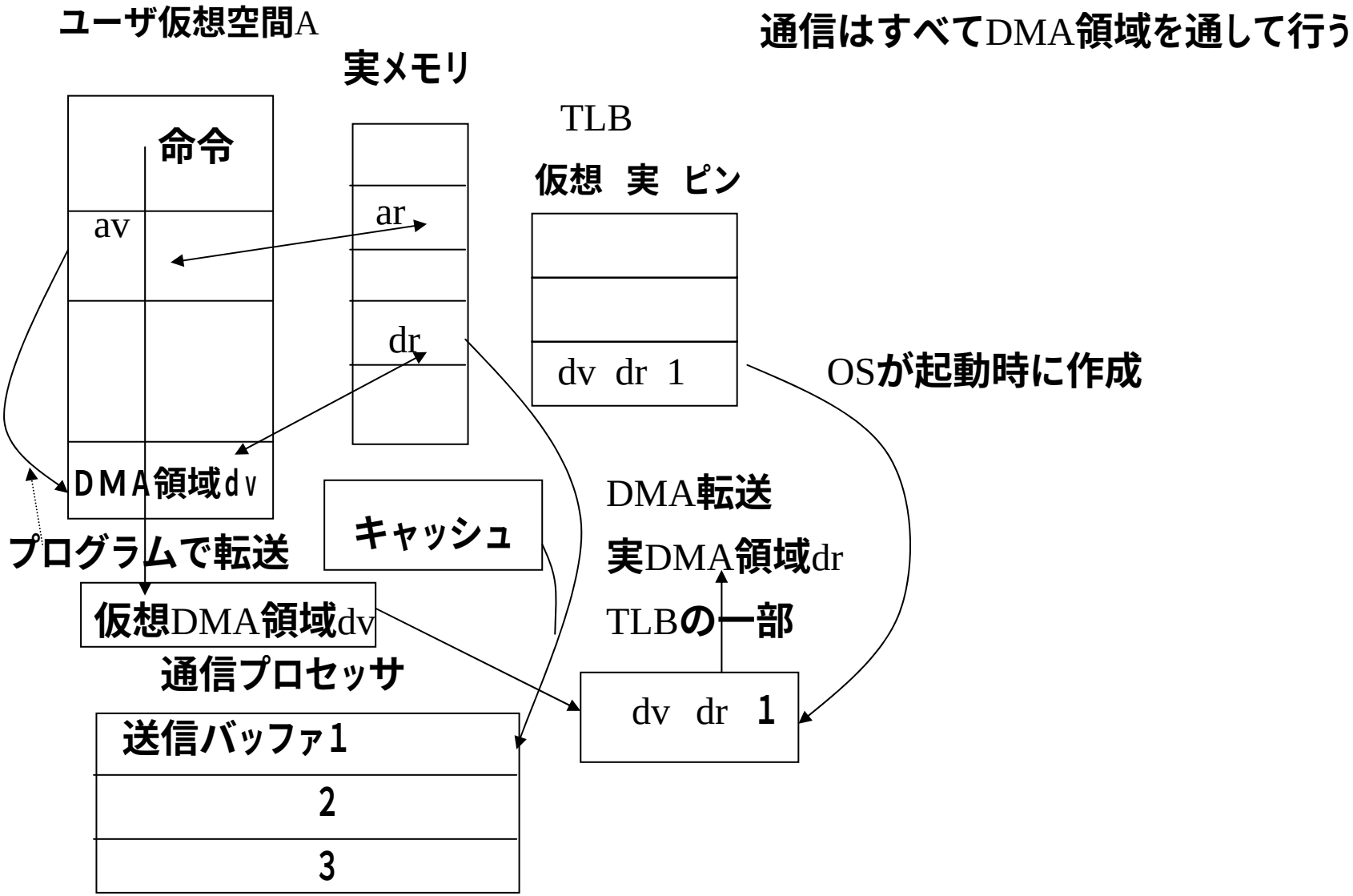
Table 1. Characteristics of 11 communication systems built for Myrinet.

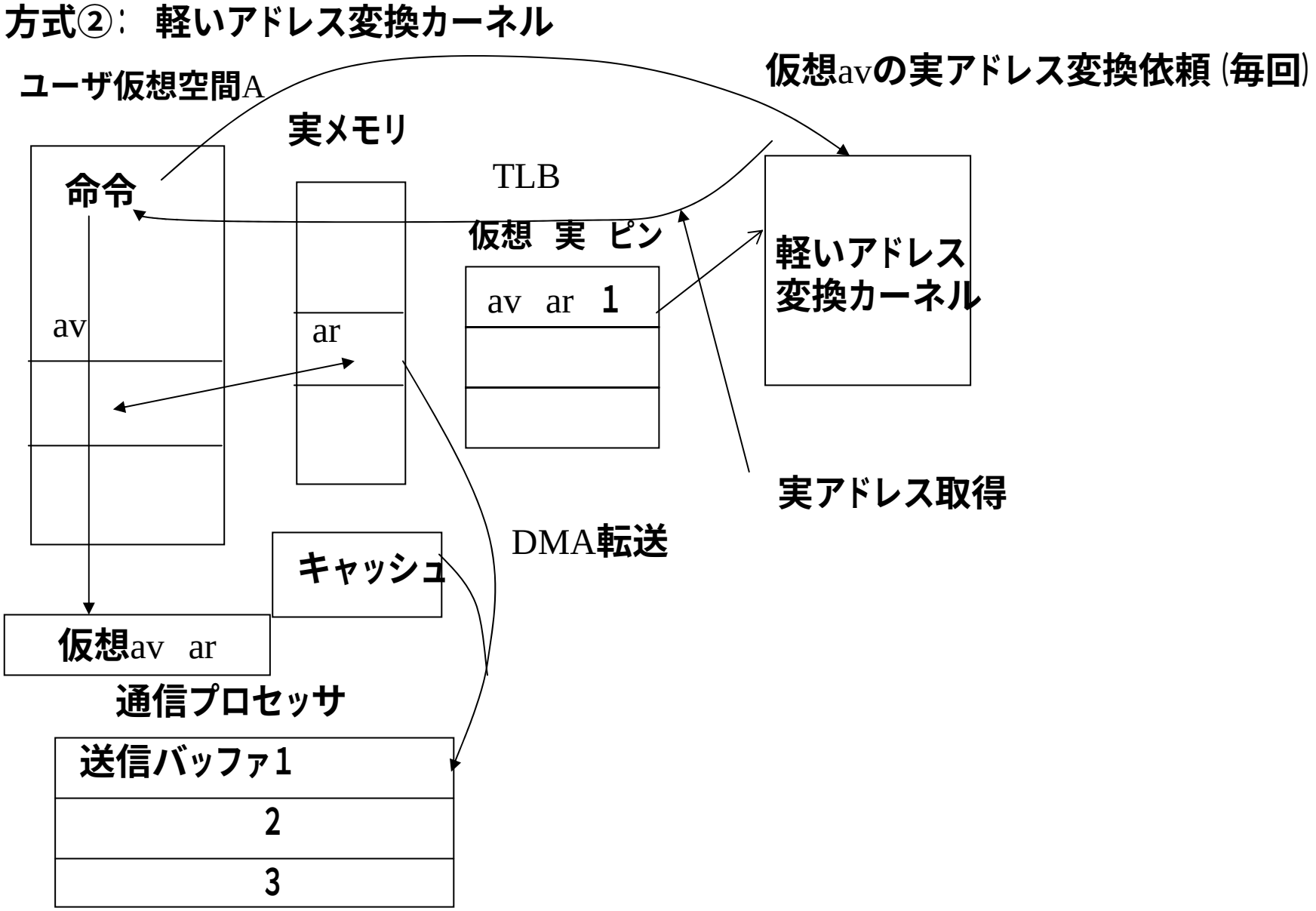
System	Data transfer (host-NI)	Translation	Protection	Control transfer	Reliability	Multicast support
AM-II[1]	PIO & DMA*	DMA areas	Yes	Polling + interrupts	Reliable, network interface: alternating bit, host: sliding window	No
FM[2]	PIO	DMA area (recv)	No	Polling	Reliable, host-level credits	No
FM/MC[3]	PIO	DMA area (recv)	No	Polling + interrupts	Reliable, unicast: host-level credits, multicast: network-interface-level credits	Yes (on network interface)
PM[4]	DMA	Software TLB* on network interface	Yes (gang scheduling)	Polling	Reliable, ACK/NACK protocol on network interface	Yes (multiple sends)
VMMC[5]	DMA	Software TLB on network interface	Yes	Polling + interrupts	Reliable, exploits hardware backpressure	No
VMMC-2[6]	DMA	UTLB* in kernel, cached on network interface	Yes	Polling + interrupts	Reliable	No
LFC[7]	PIO	User translates	No	Polling + interrupts + watchdog	Reliable, unicast: network-interface-level credits, multicast: network-interface-level credits	Yes (on network interface)
Hamlyn[8]	PIO & DMA	DMA areas	Yes	Polling + interrupts	Reliable, exploits hardware backpressure	No
Trapeze[9]	DMA	DMA to page frames	No	Polling + interrupts	Unreliable	No
BIP[10]	PIO & DMA	User translates	No	Polling	Reliable, rendezvous and backpressure	No
U-Net[11]	DMA	TLB on network interface (U-Net/MM)	Yes	Polling + interrupts	Unreliable	No

DMA基本方式　転送実領域：貼付け（ピン）、通信プロセッサ　ＴＬＢアクセスの必要

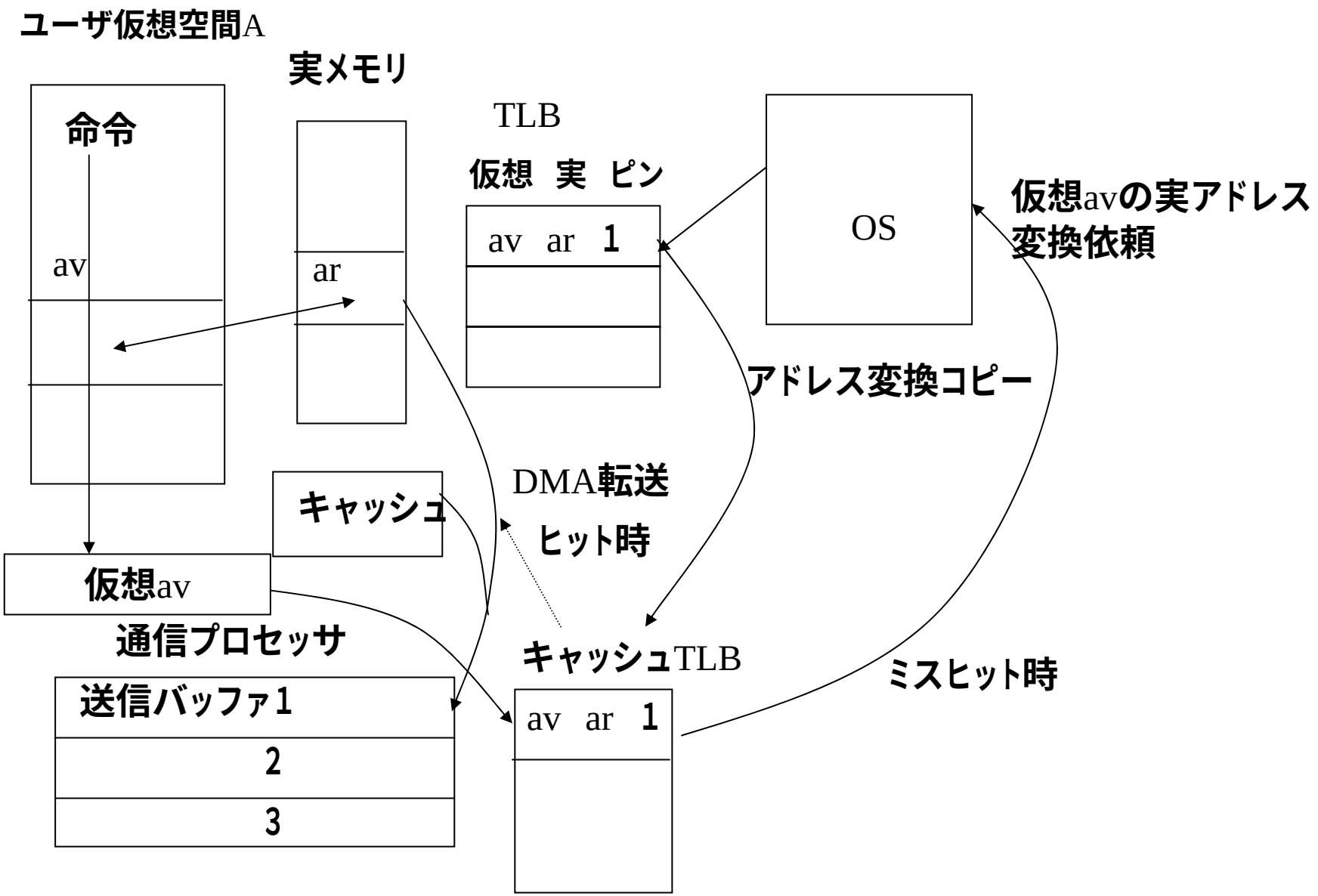


方式①：仮想空間AのDMA領域をOSに依頼して貼付け（起動時一度だけ）

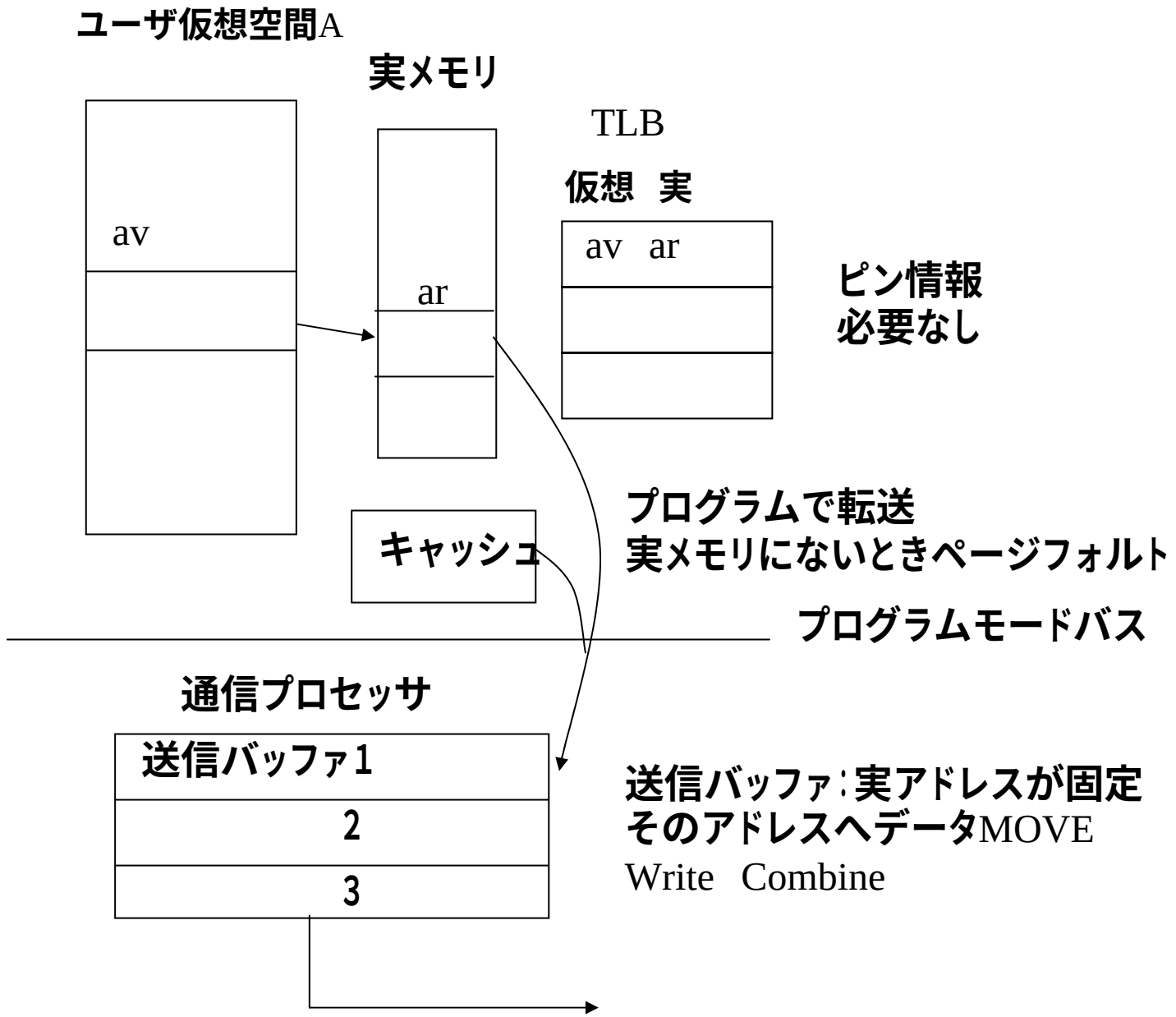


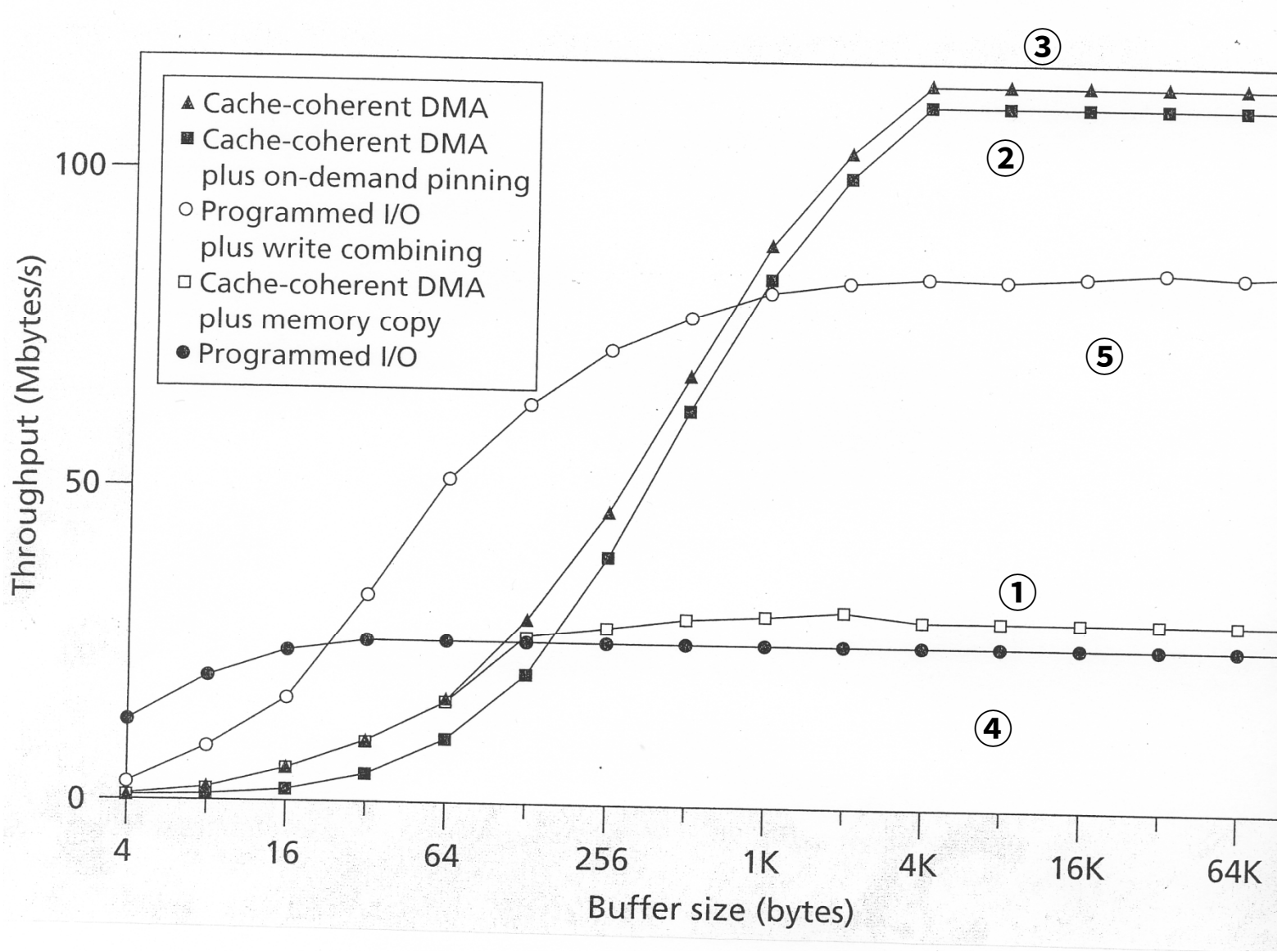

方式②： 軽いアドレス変換カーネル
ユーザ仮想空間A
仮想avの実アドレス変換依頼（毎回）
実メモリ
TLB
命令
仮想 実 ピン
av ar 1
軽いアドレス
変換カーネル
av
ar
実アドレス取得
DMA転送
キャッシュ
仮想av ar
通信プロセッサ
送信バッファ1
2
3

方式③： TLBキャッシュ



方式④、⑤　プログラムモードバス
DMAを使わない方式





A.F.Raoul et al、IEEE Computer Nov,1998,p.53

MPI Benchmark	MX/Myrinet Myricom 10G Myrinet switch	MX/Ethernet Fulcrum 10G Ethernet switch	MX/Ethernet Fujitsu 10G Ethernet switch	OpenIB with Intel MPI Mellanox InfiniBand
PingPong latency	2.4μs	2.4μs	2.8μs	4.0μs
One-way data rate (PingPong)	1204 MByte/s	1201 MByte/s	1002 MByte/s	964 MByte/s
Two-way data rate (SendRecv)	2397 MByte/s	2162 MByte/s	1762 MByte/s	1902 MByte/s

MX：Myrinet Express：メッセージパッシングソフト

Myri-10G：10 Gigabit/s,dual protocol NIC

５．５．２ＡＰ１０００の構成

メッセージ交換型マルチプロセッサの代表例

富士通のAP1000、NECのCenju-3、MITのJ-Machine

（１）プロセッサ構成

ノードプロセッサ：SPARC、通信プロセッサ（MSC）

トーラス網

①ラインセンド／バッファレシーブ機構

１０Ｂ，１０ＫＢのメッセージ転送の場合

通常の方式：112μsec/1590μsec

本機構：32.8μsec/491μsecに短縮

②ストライド転送

（２）相互結合網

T-net：各プロセッサでのメッセージ交換を行う

２次元トーラス網

リンク25MB/secの転送能力

ワームホール方式

ルーティング：次元順

S-netはバリア同期用のトリー網

B-netはホスト計算機からの放送やホスト計算機へのデータ収集

（３）後継機

AP1000+

PUT/GET機構とその専用ハードウェア（MSC+）

①PUT

②GET

キャッシュ：ストアスルー方式

通信手順

①送信プロセスMSC+にコマンド

コマンド：受信プロセッサ番号、転送先の論理アドレス、送信元の論理アドレス

①MSC+：送信データの論理アドレスを物理アドレスに変換

物理アドレスよりDMAでデータ転送

転送終了すると、指定されたメモリ領域にあるフラグを更新

③受信側MSC+：データ転送

受信データの論理アドレスを物理アドレスに変換

宛先領域にDMA転送

フラグの変更

受信プロセッサのメモリに書込み時：

キャッシュ無効化

データ転送の終了通知：

割込みオーバヘッド大

フラグを陽にチェック

通信の保護：論理（仮想）アドレス

AP1000でPUT命令実現：68μsec

AP1000+で実現すると5.1μsec

通信はすべてユーザプロセスで実現

最新のシステム例

①FUJITSU　PRIMEPOWER　HPC2500
②地球シミュレータ (NEC　SX-8)
③日立SR11000
④CRAY　X1
⑤BlueGene/L

(1) ベクトルパラレル　VS　スカラパラレル
②、④　①、③、⑤

(2) メモリ共有　VS　メッセージパッシング
①、②、③、④　⑤

(3) ネットワーク
クロスバ　VS　トーラスなど他網
①、②、③　④、⑤

TOP500LIST-June2000

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	Sandia National Laboratories United States	ASCI Red Intel	9632	1999	2379	3207
2	Lawrence Livermore National Laboratory United States	ASCI Blue-Pacific SST, IBM SP 604e IBM	5808	1999	2144	3856.5
3	Los Alamos National Laboratory United States	ASCI Blue Mountain SGI	6144	1998	1608	3072
4	IBM/Naval Oceanographic Office (NAVOCEANO) United States	SP Power3 375 MHz IBM	1336	2000	1417	2004
5	Leibniz Rechenzentrum Germany	SR8000-F1/112 Hitachi	112	2000	1035	1344
6	High Energy Accelerator Research Organization /KEK Japan	SR8000-F1/100 Hitachi	100	2000	917	1200
7	Government United States	T3E1200 Cray Inc.	1084	1998	891	1300.8
8	US Army HPC Research Center at NCS United States	T3E1200 Cray Inc.	1084	2000	891	1300.8
9	University of Tokyo Japan	SR8000/128 Hitachi	128	1999	873	1024
10	Government United States	T3E900 Cray Inc.	1324	1997	815	1191.6

TOP500LIST-June2001

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	Lawrence Livermore National Laboratory United States	ASCI White, SP Power3 375 MHz IBM	8192	2000	7226	12288
2	NERSC/LBNL United States	SP Power3 375 MHz 16 way IBM	2528	2001	2526	3792
3	Sandia National Laboratories United States	ASCI Red Intel	9632	1999	2379	3207
4	Lawrence Livermore National Laboratory United States	ASCI Blue-Pacific SST, IBM SP 604e IBM	5808	1999	2144	3856.5
5	University of Tokyo Japan	SR8000/MPP Hitachi	1152	2001	1709.1	2074
6	Los Alamos National Laboratory United States	ASCI Blue Mountain SGI	6144	1998	1608	3072
7	Naval Oceanographic Office (NAVOCEANO) United States	SP Power3 375 MHz IBM	1336	2000	1417	2004
8	Osaka University Japan	SX-5/128M8 3.2ns NEC	128	2001	1192	1280
9	National Centers for Environmental Prediction United States	SP Power3 375 MHz IBM	1104	2000	1179	1656
10	National Centers for Environmental Prediction United States	SP Power3 375 MHz IBM	1104	2001	1179	1656

TOP500LIST-June2002

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	The Earth Simulator Center Japan	Earth-Simulator NEC	5120	2002	35860	40960
2	Lawrence Livermore National Laboratory United States	ASCI White, SP Power3 375 MHz IBM	8192	2000	7226	12288
3	Pittsburgh Supercomputing Center United States	AlphaServer SC45, 1 GHz Hewlett-Packard	3016	2001	4463	6032
4	Commissariat a l'Energie Atomique (CEA) France	AlphaServer SC45, 1 GHz Hewlett-Packard	2560	2001	3980	5120
5	NERSC/LBNL United States	SP Power3 375 MHz 16 way IBM	3328	2001	3052	4992
6	Los Alamos National Laboratory United States	AlphaServer SC45, 1 GHz Hewlett-Packard	2048	2002	2916	4096
7	Sandia National Laboratories United States	ASCI Red Intel	9632	1999	2379	3207
8	Oak Ridge National Laboratory United States	pSeries 690 Turbo 1.3GHz IBM	864	2002	2310	4492.8
9	Lawrence Livermore National Laboratory United States	ASCI Blue-Pacific SST, IBM SP 604e IBM	5808	1999	2144	3856.5
10	IBM/US Army Research Laboratory (ARL) United States	pSeries 690 Turbo 1.3GHz IBM	768	2002	2050	3993.6

TOP500LIST-June2003

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	The Earth Simulator Center Japan	Earth-Simulator NEC	5120	2002	35860	40960
2	Los Alamos National Laboratory United States	ASCI Q - AlphaServer SC45, 1.25 GHz Hewlett-Packard	8192	2002	13880	20480
3	Lawrence Livermore National Laboratory United States	MCR Linux Cluster Xeon 2.4 GHz - Quadrics Linux Networx/Quadrics	2304	2002	7634	11060
4	Lawrence Livermore National Laboratory United States	ASCI White, SP Power3 375 MHz IBM	8192	2000	7304	12288
5	NERSC/LBNL United States	Seaborg - SP Power3 375 MHz 16 way IBM	6656	2002	7304	9984
6	Lawrence Livermore National Laboratory United States	xSeries Cluster Xeon 2.4 GHz - Quadrics IBM/Quadrics	1920	2003	6586	9216
7	National Aerospace Laboratory of Japan Japan	PRIMEPOWER HPC2500 (1.3 GHz) Fujitsu	2304	2002	5406	11980
8	Pacific Northwest National Laboratory United States	Cluster Platform 6000 rx2600 Itanium2 1 GHz Cluster - Quadrics Hewlett-Packard	1540	2003	4881	6160
9	Pittsburgh Supercomputing Center United States	AlphaServer SC45, 1 GHz Hewlett-Packard	3016	2001	4463	6032
10	Commissariat a l'Energie Atomique (CEA) France	AlphaServer SC45, 1 GHz Hewlett-Packard	2560	2001	3980	5120

TOP500LIST-June2004

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	The Earth Simulator Center Japan	Earth-Simulator NEC	5120	2002	35860	40960
2	Lawrence Livermore National Laboratory United States	Thunder - Intel Itanium2 Tiger4 1.4GHz - Quadrics California Digital Corporation	4096	2004	19940	22938
3	Los Alamos National Laboratory United States	ASCI Q - AlphaServer SC45, 1.25 GHz Hewlett-Packard	8192	2002	13880	20480
4	IBM - Rochester United States	BlueGene/L DD1 Prototype (0.5GHz PowerPC 440 w/Custom) IBM/ LLNL	8192	2004	11680	16384
5	NCSA United States	Tungsten - PowerEdge 1750, P4 Xeon 3.06 GHz, Myrinet Dell	2500	2003	9819	15300
6	ECMWF United Kingdom	eServer pSeries 690 (1.9 GHz Power4+) IBM	2112	2004	8955	16051
7	Institute of Physical and Chemical Res. (RIKEN) Japan	RIKEN Super Combined Cluster Fujitsu	2048	2004	8728	12534
8	IBM Thomas J. Watson Research Center United States	BlueGene/L DD2 Prototype (0.7 GHz PowerPC 440) IBM/ LLNL	4096	2004	8655	11469
9	Pacific Northwest National Laboratory United States	Mpp2 - Cluster Platform 6000 rx2600 Itanium2 1.5 GHz, Quadrics Hewlett-Packard	1936	2003	8633	11616
10	Shanghai Supercomputer Center China	Dawning 4000A, Opteron 2.2 GHz, Myrinet Dawning	2560	2004	8061	11264

TOP500LIST-June2005

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM	65536	2005	136800	183500
2	IBM Thomas J. Watson Research Center United States	BGW - eServer Blue Gene Solution IBM	40960	2005	91290	114688
3	NASA/Ames Research Center/NAS United States	Columbia - SGI Altix 1.5 GHz, Voltaire Infiniband SGI	10160	2004	51870	60960
4	The Earth Simulator Center Japan	Earth-Simulator NEC	5120	2002	35860	40960
5	Barcelona Supercomputer Center Spain	MareNostrum - JS20 Cluster, PPC 970, 2.2 GHz, Myrinet IBM	4800	2005	27910	42144
6	ASTRON/University Groningen Netherlands	Stella - eServer Blue Gene Solution IBM	12288	2005	27450	34406.4
7	Lawrence Livermore National Laboratory United States	Thunder - Intel Itanium2 Tiger4 1.4GHz - Quadrics California Digital Corporation	4096	2004	19940	22938
8	Computational Biology Research Center, AIST Japan	Blue Protein - eServer Blue Gene Solution IBM	8192	2005	18200	22937.6
9	Ecole Polytechnique Federale de Lausanne Switzerland	eServer Blue Gene Solution IBM	8192	2005	18200	22937.6
10	Sandia National Laboratories United States	Red Storm, Cray XT3, 2.0 GHz Cray Inc.	5000	2005	15250	20000

TOP500LIST-June2006

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM	131072	2005	280600	367000
2	IBM Thomas J. Watson Research Center United States	BGW - eServer Blue Gene Solution IBM	40960	2005	91290	114688
3	DOE/NNSA/LLNL United States	ASC Purple - eServer pSeries p5 575 1.9 GHz IBM	12208	2006	75760	92781
4	NASA/Ames Research Center/NAS United States	Columbia - SGI Altix 1.5 GHz, Voltaire Infiniband SGI	10160	2004	51870	60960
5	Commissariat a l'Energie Atomique (CEA) France	Tera-10 - NovaScale 5160, Itanium2 1.6 GHz, Quadrics Bull SA	8704	2006	42900	55705.6
6	Sandia National Laboratories United States	Thunderbird - PowerEdge 1850, 3.6 GHz, Infiniband Dell	9024	2006	38270	64972.8
7	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME Grid Cluster - Sun Fire X64 Cluster, Opteron 2.4/2.6 GHz, Infiniband NEC/Sun	10368	2006	38180	49868.8
8	Forschungszentrum Juelich (FZJ) Germany	JUBL - eServer Blue Gene Solution IBM	16384	2006	37330	45875
9	Sandia National Laboratories United States	Red Storm Cray XT3, 2.0 GHz Cray Inc.	10880	2005	36190	43520
10	The Earth Simulator Center Japan	Earth-Simulator NEC	5120	2002	35860	40960

TOP500LIST-June2007

Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM	131072	2005	280600	367000
2	Oak Ridge National Laboratory United States	Jaguar - Cray XT4/XT3 Cray Inc.	23016	2006	101700	119350
3	NNSA/Sandia National Laboratories United States	Red Storm - Sandia/ Cray Red Storm, Opteron 2.4 GHz dual core Cray Inc.	26544	2006	101400	127411
4	IBM Thomas J. Watson Research Center United States	BGW - eServer Blue Gene Solution IBM	40960	2005	91290	114688
5	Stony Brook/BNL, New York Center for Computional Sciences United States	New York Blue - eServer Blue Gene Solution IBM	36864	2007	82161	103219
6	DOE/NNSA/LLNL United States	ASC Purple - eServer pSeries p5 575 1.9 GHz IBM	12208	2006	75760	92781
7	Rensselaer Polytechnic Institute, Computional Center for Nanotechnology Innovations United States	eServer Blue Gene Solution IBM	32768	2007	73032	91750
8	NCSA United States	Abe - PowerEdge 1955, 2.33 GHz, Infiniband Dell	9600	2007	62680	89587.2
9	Barcelona Supercomputing Center Spain	MareNostrum - BladeCenter JS21 Cluster, PPC 970, 2.3 GHz, Myrinet IBM	10240	2006	62630	94208
10	Leibniz Rechenzentrum Germany	HLRB-II - Altix 4700 1.6 GHz SGI	9728	2007	56520	62259.2

日立スーパコンピュータ

- ベクトルパラレル
- 1982　S-810　　630MFLOPS
- 1987　S-820　　3GFLOPS
- 1992　S-3000　　32GFLOPS
- スカラパラレル
- 1995　SR2201　600GFLOPS
- 1999　SR8000　7．3TFLOPS
- 2003　SR11000　62TFLOPS
- (Power5 (1．9GHz)、16PE/ノード、121.6GFLOPS/ノード、最大512ノード、多段クロスバネット：12GB/sx2 (ノード当たり)

FUJITSU　PRIMEPOWER HPC

VPPからSPPに切り替え

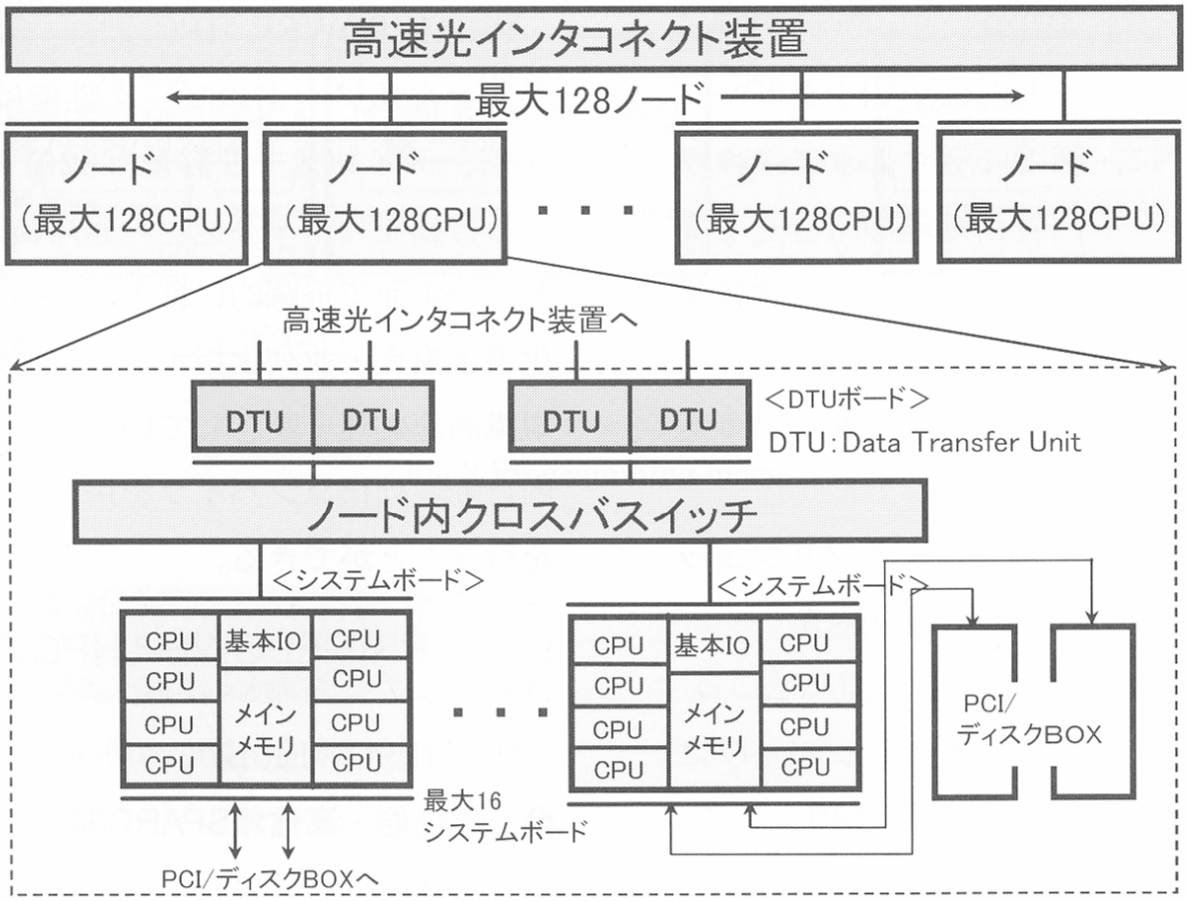


図-1　PRIMEPOWER HPCシステムの構成
Fig.1-System configuration of PRIMEPOWER HPC.

ノード：共有メモリ、スヌープ方式

SMP

512インタリーブメモリ（ノード内）

Fujitsu　Vol.53,No.6、2002,**特集　サーバ**

- Fujitsu PRIMEPOWER HPC

表-1　PRIMEPOWER HPCノード諸元

項目	諸元
CPU	SPARC64 V
CPU周波数	1.3 GHz
最大CPU数	128
アドレススヌープ性能	133 Gバイト/秒
最大メインメモリ容量	512 Gバイト
最大メインメモリインタリーブ数	512ウェイ
最大PCIスロット数	320

表-2　PRIMEPOWER HPCシステム諸元

項目	諸元
最大ノード数	128
最大CPU数	16,384
最大論理性能	85.2 TFLOPS★
最大メインメモリ容量	64 Tバイト
ノード間結合方式	クロスバ
ノード間転送性能	1ノードあたり 最大16 Gバイト/秒×2（入力/出力）

★：Tera FLoating point Operation Per Second

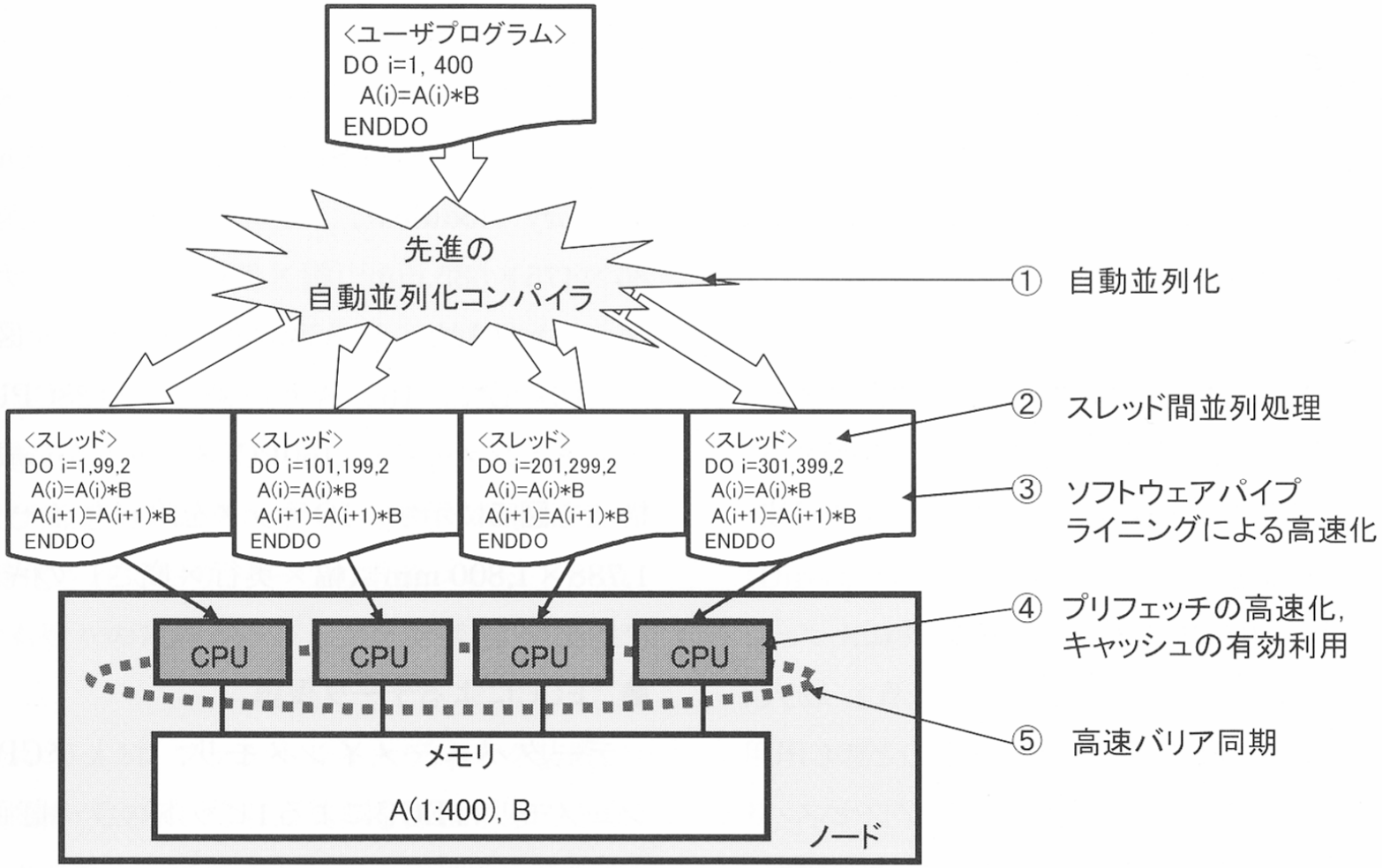


図-2　ノード内並列処理

Fig.2-Parallel execution method in node.

SPARC64 V

整数演算2台、
浮動小数点2台、
アドレス計算2台

5.2GFLOPS

命令
フェッチ

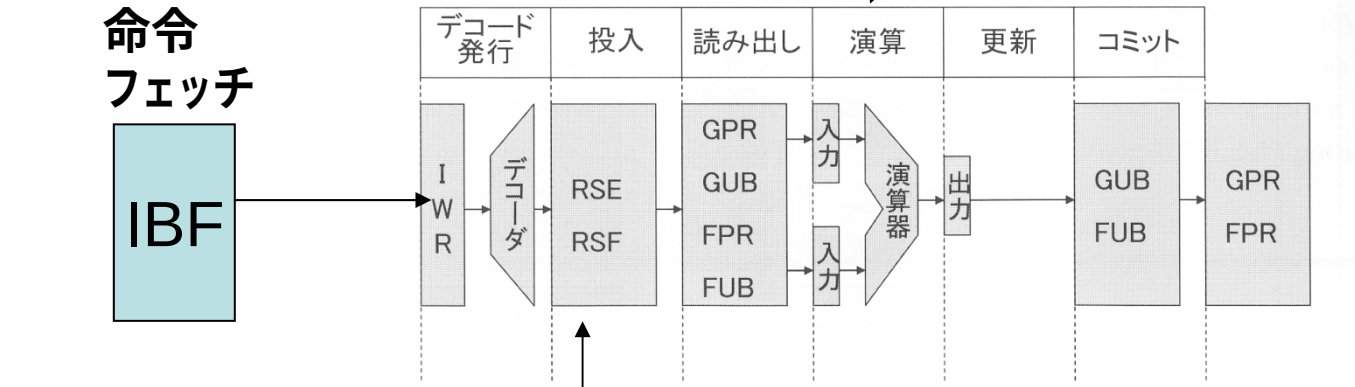


図-2　SPARC64 Vのパイプライン
Fig.2-Pipeline structure of SPARC64 V.

8命令フェッチ

4命令
発行

リオーダバッ
ファGUB,FUB

分岐
予測

リザベーションステーション

待ち行列長：16整数、16
浮動、10アドレス

正確な割込み
保障

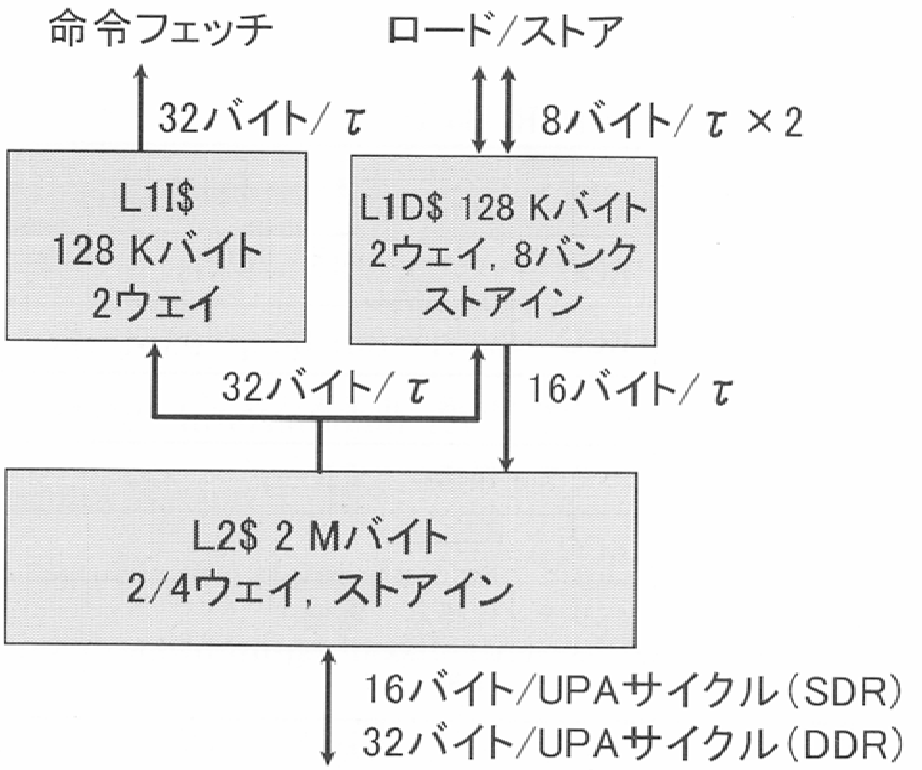


図-3　SPARC64 Vのキャッシュ
Fig.3-Cache structure of SPARC64 V.

- **デザインルール**：0.13μm
- TR**数**：19,100**万個**
- **信号ピン数**：**269**
- **チップサイズ**：17.8x15.7mm
- **動作周波数**：**1．3**GHz
- **消費電力**：**50**W

NECスーパコンピュータ

- 機種　年　サイクル単体性能　最大性能　台数
- Cray-1　1976　12.5ns　160MF　160MF　1台
- SX-1/2　1984　6ns　1.3GF　1.3GF　1台
- SX-3　1989　2.9ns　5.5GF　22GF　4台
- SX-4　1994　8ns　2GF　1TF　512台
- SX-5　1998　4ns　8GF　4TF　512台
- SX-6　2001　2ns　8GF　8TF　1024台
- (CMOSシングルチップ、8PE/1ノード、最大128ノード、0.15μm)
- SX-7　2002　1.8ns　11.4GF　23TF　2048台
- (32PE/1ノード、最大64ノード、0.15μm)
- SX-8　2004　0.5ns　16GF　65TF　4096台
  (8PE/1ノード、最大512ノード、0.09μm)
- SX-9　2007　0.3ns　102.4GF　839TF　8192台
  (3.2GHz、8-16PE/1ノード、最大512ノード、0.065μm、11層銅配線)

科学新聞
週刊
（金曜日発行）
発行所　科学新聞社

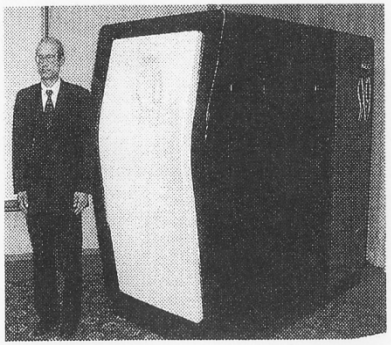
世界最速ベクトル型スパコン

単一コアで100GFLOPS超達成

SXの新モデル投入　NEC

NECは、世界最高速のベクトル型スーパーコンピュータ「SXシリーズ　モデルSX−9」を製品化して、10月25日から世界で同時発売した。これは、単一チップ当たりで100GFLOPS（1ギガフロップス＝毎秒10億回の浮動小数点演算性能）超という極めて優れた演算性能をもった、世界で最速の一チップベクトルプロセッサを実現したもの。1ノード当たり16個のCPUを搭載して演算性能1・6TFLOPS（1テラフロップス＝毎秒1兆回の浮動小数点演算性能）を、さらにそれを最大512ノード接続した、総合演算性能839TFLOPSの超大規模マルチノードシステムを実現した。

SXシリーズは、最先端科学の研究支援ツールとして、研究者にとって「使いやすさ」と「高性能」を提供するスパコンだとNECは強調する。世界中で、これまでに約1100台の累積販売台数を誇り、特に、気象・気候解析をはじめ航空宇宙、環境、流体解析などで高い評価を得ている。



SX−9（シングルノードシステム）の説明をする丸山執行役員常務

SX−9は、従来のSX−8の13ノード（1・66TFLOPS）に匹敵する性能を、1ノード（1・6TFLOPS）で実現する、驚異的な演算性能をもったスパコンである。

心臓部のプロセッサは、基本的に従来のSXベクトルアーキテクチャを継承。これに演算機の追加、ベクトルパイプライン数増強などのアーキテクチャ改良を加えた。

また、これまでの壁を破る高速化、低消費電力化など最先端技術を採用した65nm（ナノメル）CMOS、11層銅配線LSIにより3・2GHzという高周波数化を達成して、単一コアチップとしては、世界初の102・4GFLOPSという演算性能、256Gバイト／秒のメモリバンド幅を実現した。同社の丸山好一・執行役員常務は、SX−9製品発表会の席で「例えば気象予測計算分野でSX−9はスカラ型の機能より3〜4倍、物性計算分野でも4〜7倍優れている」と、ベクトル型SX−9の利点をアピールした。

さらに、設置面積と消費電力において、従来のSX−8の13ノード対し、同等の性能のSX−9の1ノードはいずれも約4分の1と小型・省エネ化している。

コストパフォーマンスでは、従来機の約6倍である。

同社の西川岳・第一コンピュータ事業本部長は「HPC（ハイ・パフォーマンス・コンピューティング）分野でも、従来の速さだけではなく、最近はECOの観点も重視されるようになってきており、世界のスパコンランキングトップ500では補完的リストとして、世界で最もエネルギー効率のよいスパコンランキングが発表されているが、SX−9は電力性能比なら世界トップ」と説明した。

レンタル価格は月額298万円（税込）で、今後3年間で700システムの販売を見込んでいる。すでに受注もあり、20ノードのSX−8Rを導入した大阪大学サイバーメディアセンターが、10ノードのSX−9を来夏に追加導入する。

Empowered by Innovation

NEC

NEC SUPERCOMPUTER
SX SERIES MODEL

SX-9

世界最速性能で、未知への扉を開く。

ペタフロップスコンピューティングを視野に入れたスーパーコンピュータSX-9登場！

果てしない宇宙の深淵を探る。生命科学の謎を解明する。

ナノテクノロジーの新たな領域を切りひらく。

スーパーコンピュータは、いまや社会の発展になくてはならない存在です。

もっと、高速に、効率的に、そしてもっと使いやすく。

そんな科学の最前線のニーズに応えて、いまハイパフォーマンスマシンが誕生しました。シングルコア性能世界一。

マルチノードではペタフロップスに迫る演算性能を実現。

圧倒的な高速性能と使いやすさを兼ね備えた先進のベクトルスーパーコンピュータSX-9。

世界最速性能*で、サイエンスの未知の扉を開きます。

※ 単一コアあたりの理論小数点演算性能（2007年10月時点）

NEC SUPERCOMPUTER
SX SERIES MODEL
SX-9

次世代オペレーティングシステム
SUPER-UX

SX-9の特長-1
容易に超高速で大規模な計算が可能

CPUはプロセッサ当り100GFLOPSを超える世界最高速の性能と256Gバイト/秒の広いメモリバンド幅を実現。シングルノードシステムでは、1.6TFLOPSを超える演算性能と最大1Tバイトの大規模共有メモリにより、容易に高速演算が可能です。さらに、マルチノードシステムでは、最大512ノードで839TFLOPSの超高速演算性能とノード当り最大128Gバイト/秒×2のノード間超高速通信により、高いシステムトータル性能が得られ、より大規模・長時間スケールの現象の計算が可能になります。

SX-9の特長-2
拡張性の高いシステムを容易に管理

従来のSXシリーズで磨きぬかれてきたスーパーコンピュータ用オペレーティングシステムSUPER-UXを提供。大規模マルチノードシステムへの対応などの機能として、リソース管理、ジョブ管理、チェックポイント・リスタートなど様々な機能を提供し、最大512ノードの構成においても容易なシステム管理が可能です。また、SX-9の性能を最大限に発揮するためのコンパイラを提供しており、総合プログラム開発環境を利用した効率的なソフトウェア開発が可能です。

SX-9の特長-3
高密度実装技術と低消費電力技術によるTCOの削減

1チップベクトルプロセッサとメモリモジュールを、高密度実装技術によりコンパクトな装置にパッケージング。最先端の低消費電力回路設計技術を用い、電力効率に優れたベクトルプロセッサの性能当たり電力をさらに改善。このため、設置環境・消費電力も大幅に改善されています。

SX-9の特長-4
SXシリーズの応用ソフトウェア資産継承

従来のSXシリーズでサポートしているさまざまなアプリケーションソフトウェアを、そのまま継承可能。SX-9の超高速演算性能を活用できます。

GFLOPS: Giga Floating-point Operations Per Second（1秒当たり10億回の浮動小数点演算）
TFLOPS: Tera Floating-point Operations Per Second（1秒当たり1兆回の浮動小数点演算）

あらゆるニーズに応える新アーキテクチャ

最先端のLSIテクノロジ、及び高密度実装技術によりSX-9は世界で初めて100GFLOPSを超えた世界最速の1チップベクトルプロセッサ、ノード演算性能1.6TFLOPS、及びメモリ容量1TバイトのⅠ大規模共有メモリアーキテクチャによるシングルノードシステム、そして最大128Gバイト/秒の超高速インタコネクトにより512ノードを接続し、839TFLOPSの総合演算性能を有する共有・分散メモリアーキテクチャの超大規模マルチノードシステムを実現。NECはSX-9を通して、あらゆるHPCニーズにお応えいたします。

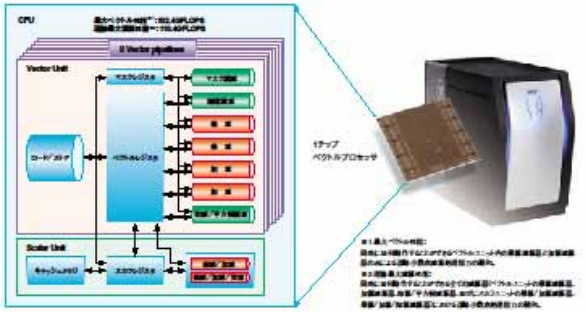
世界最速の1チップベクトルプロセッサ

● 超高速演算を可能にした強力なプロセッサ

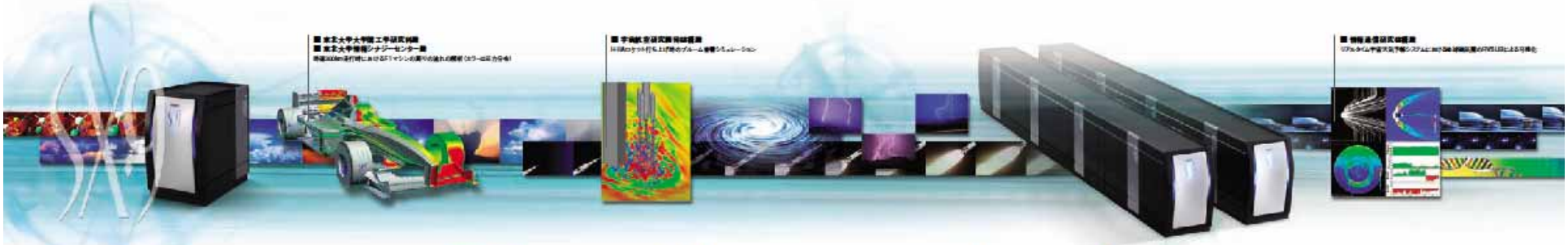
HPCにおいて並列度が増加することは、プログラミングの難しさを引き起こすだけではなく、並列処理におけるスケーラビリティが得られなくなる問題があります。「使いやすく」、かつ「高性能」なスーパーコンピュータをご提供し続けるために、NECは単一チップの演算性能にこだわり続けます。SX-9は100GFLOPS超という驚異的な演算性能を誇る世界最速1チップベクトルプロセッサを実現。これにより、演算処理において必要となる"並列度"を低く抑えることを可能とし、単に理論演算性能が高いだけではなく、実務において"使いやすい"高い実効性能を実現します。演算処理において高い実効性能を実現するためには、プロセッサの演算性能とバランスのとれたメモリバンド幅も重要な要素となります。SX-9は、1チップ当たり256Gバイト/秒という他のプロセッサの追随を許さない広いメモリバンド幅を実現しています。

● 超高速プロセッサを実現するベクトルアーキテクチャ・テクノロジ

最先端技術の結晶とも言えるスーパーコンピュータSX-9。その心臓部であるプロセッサは、基本的に従来のSXベクトルアーキテクチャを継承しつつ、演算器の追加、ベクトルパイプライン数増強などのアーキテクチャ改良を行なっています。また、これまでの"壁"を打破する高速化、低消費電力化の各種最先端テクノロジを採用した65nm CMOS、11層銅配線 LSIにより3.2GHzという高周波数化を実現。これにより、単一コアチップとしては世界初となる102.4GFLOPSの演算性能、及び256Gバイト/秒のメモリバンド幅を実現しています。



もっと使いやすく、さらに高性能に。世界最速1チップベクトルプロセッサが、未知の世界を解明する。



Hardware ハードウェア

高度なシミュレーションが求められるあらゆる分野で、その能力をいかんなく発揮するために。
NECが求めたのは、「使いやすさ」と「高性能」の両立。この課題を世界最速1チップベクトルプロセッサの開発によって解決しました。
単一チップで100GFLOPS超の演算性能と256Gバイト/秒のメモリバンド幅が生む
使いやすく高い実効性能が、科学の探求を強力に支援します。

● これまでの"壁"を軽々と超える最新テクノロジ

HPCのみならず、計算機システムが現在直面している壁、すなわち「メモリバンド幅の壁」、及び「消費電力の壁」。スーパーコンピュータには、これらの壁を打破することが求められています。SX-9はメモリアーキテクチャの改良、及び最先端テクノロジによりメモリバンド幅の壁を打破し、256Gバイト/秒という驚異的なメモリバンド幅を実現しています。また、電力効率の高い10Gbps-SerDes、Multi-Vth、クロックゲーティング、チップ内電力センサを用いた電力制御などの最先端低消費電力化技術を採用することにより、高性能かつ低消費電力のプロセッサを実現。電力効率の高いスーパーコンピューティング環境をご提供いたします。

使いやすさと高性能を併せ持つ大規模SMPノード

SX-9シングルノードシステムは、最大16CPU(ノード演算性能1.6TFLOPS)、最大1Tバイトの共有メモリを装備するAモデルと、最大8CPU(ノード演算性能800GFLOPS)、最大512GバイトのBモデルを用意し、ニーズに合わせて使いやすく、かつ高性能なSMPによるHPC環境を実現します。SX-9は従来のSXシリーズとの互換性を維持しつつ、演算性能、メモリバンド幅、メモリ容量、入出力性能などのトータルバランスを追求。実績のあるベクトル型スーパーコンピュータの資産をスムーズに継承・移行することが可能です。

スケーラブルなHPCを実現するマルチノードシステム

SX-9マルチノードシステムは、シングルノードシステムを構成要素として、片方向最大128Gバイト/秒という超高速ノード間接続装置(IXS)によりノード間を接続、最大512ノード、8192台のプロセッサによるマルチノードシステム構成が可能です。超高速な1チップベクトルプロセッサ、共有メモリアーキテクチャによるシングルノード、及び超高速ノード間接続装置による大規模な共有・分散メモリシステムによりスケーラブルなHPC環境を実現します。SX-9マルチノードシステムは、512ノードの最大構成時においても、シングルシステムイメージでの運用環境を提供。512ノード、8192台のプロセッサによる共有・分散メモリシステムは839TFLOPSを実現し、超大規模、かつ使いやすいHPCシステムをご提供いたします。

高い演算性能とバランスのとれた高い I/O性能

使いやすく、かつ高性能なシステムであるためには、演算性能とバランスのとれた高いI/O性能も重要な要素です。SX-9はシングルノードあたり最大64Gバイト/秒という超高速の総合I/O性能を有しています。

多彩なRAS機能による高信頼性

SX-9は、高集積設計により部品点数を大幅に削減。これによりハードウェアの信頼性を格段に向上させています。さらに、NECのメインフレーム開発で培われた各種高信頼性技術、高可用性技術を装載することにより、信頼性、可用性の向上を図っています。メモリには誤り検出訂正符号(ECC)を採用。回路部には二重化などによる誤り検出機能を組み込んでいます。また、各装置内にエラー箇所を検出するビルトイン診断機能(BID)を備え、速やかな故障箇所の指摘や回復・再構築処理を実行します。さらに、障害情報の自動収集、サービスセンタへの自動通報や、センタからの遠隔保守により、迅速な故障診断と容易な予防保守を実現。システムの信頼性・可用性・保守性を総合的に高めています。

周辺装置やネットワーク環境については、
SX-9ホームページをご覧ください。
http://www.nec.co.jp/hpc/sx9/product/hardware.html

パワフルで柔軟なSUPER-UX

オペレーティングシステムとしてUNIX System Vに準拠した「SUPER-UX」を提供。高速入出力機能、通信機能など、スーパーコンピュータにふさわしい機能を備えています。

● マルチプロセッサ／マルチノードサポート

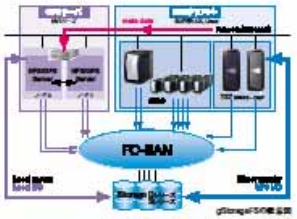
マルチプロセッサと並列処理をさらに強化してサポート。また、マルチノードシステムにおいてもシングルシステムイメージ（SSI）を実現。より使いやすく快適な利用環境を提供します。

● リソースブロック機能

システムの持つ多数のCPUや大規模なメモリ資源の分割管理が可能。特にノード内では、分割損のまったく発生しない、効率のよい資源管理を自動的に行うことができます。

● gStorageFS

gStorageFS（GFS）は、複数ノード、あるいは異機種間において、高速にファイルを共有できる機能を提供します。GFSを利用することにより、GFSクライアントでもローカルディスクに近い高性能のアクセスが可能です。ユーザはGFSを意識することなくNFSと同様に利用できます。SXシリーズ、TX7シリーズ（OS: Linux）やEM64T（OS: Linux）、SGI Altix などとの間でもファイル共有が可能です。また、GFSサーバのSStorage NVゲートウェイシリーズは、ファイルサーバとして完全な二重化構成をとっており、高信頼、高可用な運用を提供します。



強力なバッチ処理機能／ジョブスケジューリング機能

● NQSII

さまざまなシステム構成に対応可能な次世代バッチシステムです。クライアント機能、運用管理機能、ジョブ実行機能を物理的に分離することでSSIを実現。統一したユーザインタフェース、管理機能の一元化によるシンプルな運用管理、ジョブが使用可能な計算リソースの最適化が行えます。

●ジョブスケジューリング機能

ジョブのスケジュールを行うバッチスケジューラはNQSII本体とは独立に構成され、より柔軟なジョブスケジューリングを実現しています。基本的なスケジューリング機能を持つ標準スケジューラは、投入順に処理を行うFCFS（First Come First Served）方式のスケジューリングを提供します。

● JobManipulator

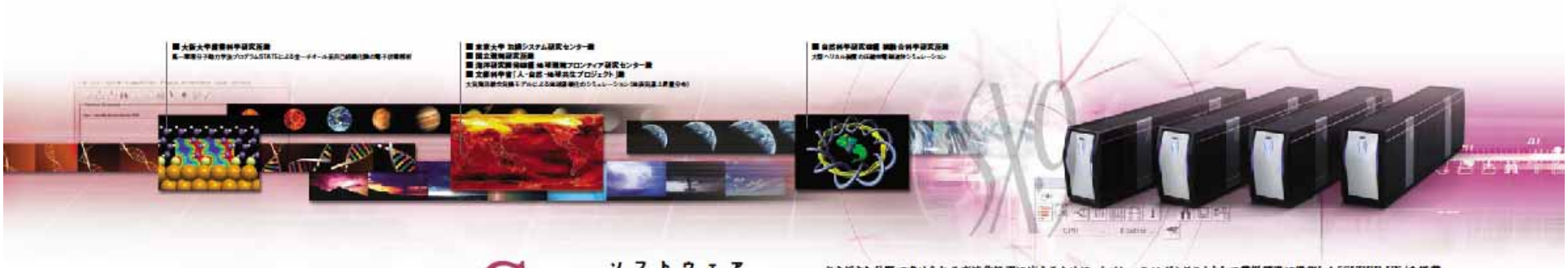
JobManipulatorは、計画的なリソース管理機能によってシステム稼動率の最大化を実現する高機能バッチスケジューラです。ジョブ実行に必要な計算リソースを計画的に割り当て、占有利用を可能にするバックフィル・スケジューリングを提供すると共に、ユーザ、グループ、および組織単位に、公平なジョブの優先度制御を実現するフェアシェアスケジューリング機能や、ジョブの実行開始時刻と必要リソース量を保証する事前予約機能をサポートしています。

柔軟で快適な運用管理を支援

● チェックポイント・リスタート機能

実行中のプログラムを任意の時点で中断させ、後に再開させることができます。これにより、緊急ジョブの最優先実行や長時間ジョブの分割実行など、柔軟な運用が可能です。マルチノードシステムではNQSIIにより、マルチノードMPIプログラムのチェックポイント・リスタート機能および実行中のジョブを他の資源の空いているノードへ移すマイグレーション機能が利用できます。

ハードウェアのポテンシャルを最大限に引き出す柔軟なシステム運用と充実したプログラム開発ツール。



Software

ソフトウェア

さまざまな分野で求められる高速化処理に応えるために。オペレーティングシステムとして業界標準に準拠した「SUPER-UX」を搭載。
大規模マルチノードシステムへの対応、効率をアップさせる運用管理など、柔軟な機能を備えています。
また、ベクトル処理、並列処理のすべてのプログラミングモデルに対応。
ライブラリ、ツールなども充実し、最適の開発・利用環境を提供します。

先進の言語処理体系とソフトウェア開発環境

SX-9の性能を最大限に引き出すためのさまざまなライブラリやツールを提供しています。

● ベクトル化・並列化の中核コンパイラ FORTRAN90／SX、C++／SX

FORTRAN90／SX は、従来のSXシリーズで実績を培ってきた高度な最適化、自動ベクトル化、自動並列化機能を装備し、SXシリーズ向けに最適設計されたコンパイラです。最新のFortran95仕様をフル実装し、SXシリーズのベクトル化・並列化の中核を担います。C++／SXコンパイラは、C言語およびC++言語のISO規格をサポートするC／C++コンパイラです。FORTRAN90／SXと共通のバックエンドをもち、高度な自動ベクトル化、自動並列化および最適化機能を提供します。

● 共有メモリ並列処理機能OpenMP

コンパイラの強力な自動並列化機能に加え、OpenMP API により、通常のプログラムから容易に並列処理を利用できます。FORTRAN 90／SX、C++／SXでOpenMP API Ver.2.5 を提供します。

● メッセージ・パッシング・インターフェース ライブラリMPI／SX、MPI2／SX

MPI／SXおよびMPI2／SXは、シングルノード内であれば共有メモリの特長を活かし、また、マルチノードシステムであればIXSのハードウェア性能を最大限に引き出すことで、低レイテンシ、高スループットのデータ転送を実現しています。MPI／SXは、MPI-1.2仕様の機能を提供し、MPI2／SXでは、MPI-2仕様の機能を提供します。

● データ・パラレル言語HPF／SX V2

HPF／SX V2は、HPF2.0基本仕様に準拠し、主要なHPF公認拡張仕様、HPF／JA拡張仕様をサポートします。

HPF：High Performance Fortran

研究者のためのパワフルなプログラム開発ツール

● GUIベースのプログラム開発環境PSUITE

SXシリーズでは、スーパーコンピュータと端末が接続されたネットワーク上で、GUIによる編集、翻訳、実行、デバッグ、チューニング作業が可能な統合プログラム開発環境PSUITEを提供します。

▲ 統合プログラム開発環境PSUITEの画面例

▲ 簡易性能解析機能（ftrace機能）出力例

● 性能解析ツールproginf、ftrace機能

FORTRAN90／SXやC++／SXが標準で備える機能で、環境変数やコンパイルオプションによって、プログラム全体やサブルーチン単位、ユーザ指定範囲の性能情報を出力することが可能です。

NEC提供 アプリケーション・ソフトウェア

● リアルタイム可視化システムRVSLIB

流体解析や構造解析などの大規模並列シミュレーション実行中に、計算途中結果をSXシリーズ上でリアルタイムに可視化するシステムです。

● 数学ライブラリ

豊富な機能を持つ数値計算ライブラリASL、統計計算ライブラリASLSTAT、ならびにBLAS／LAPACK／ScaLAPACKなどを含むライブラリ集MathKeisanを提供しています。いずれもSXシリーズ向けに最適化されています。

さらに詳しいソフトウェア開発・利用環境についてはSX-9ホームページをご覧ください。
http://www.nec.co.jp/hpc/sx9/product/software.html

HPC
High Performnce Computing

SX Series

English

トピックス
イベント情報
事例紹介

お問い合わせ

SX シリーズ
SX-9
SX-8R
SX-8i

ソフトウェア情報
NECアプリケーション
サポートソフトウェア

TX7 Series

Express5800 Series

HPC販売推進本部

TOP | 装置諸元 | 多彩な応用分野
ハードウェア | ソフトウェア | カタログダウンロード

装置諸元
Specifications

シングルノードシステム

モデルグループ	SX-9	
モデル名	A	B
中央処理装置(CPU)		
CPU台数	8~16	4~8
理論最大演算性能※1	947.2G~1,894.4GFLOPS	473.6G~947.2GFLOPS
最大ベクトル性能※2	819.2G~1,638.4GFLOPS	409.6G~819.2GFLOPS
ベクトルレジスタ	144Kバイト×8~16	144Kバイト×4~8
スカラレジスタ	64ビット×128×8~16	64ビット×128×4~8
主記憶装置(MMU)		
メモリアーキテクチャ	共有メモリ	
容量	512Gバイト、1Tバイト	256Gバイト、512Gバイト
最大転送性能	4Tバイト/秒	2Tバイト/秒
入出力機構		
最大チャンネル数	32チャンネル※3	16チャンネル
最大データ転送性能	128Gバイト/秒※3	64Gバイト/秒

マルチノードシステム

モデルグループ	SX-9
ノード数	2～512※4
中央処理装置(CPU)	
CPU台数	32～8,192
理論最大演算性能※1	3.8T～969.9TFLOPS
最大ベクトル性能※2	3.3T～838.9TFLOPS
ベクトルレジスタ	144Kバイト×32～8,192
スカラレジスタ	64ビット×128×32～8,192
主記憶装置(MMU)	
メモリアーキテクチャ	共有・分散メモリ
容量	1T～512Tバイト
最大転送性能	2,048Tバイト/秒
入出力機構	
最大チャンネル数	16,384チャンネル※3
最大データ転送性能	64Tバイト/秒※3
ノード間接続装置	
最大データ転送性能	128Gバイト/秒×2(双方向)/ノード

SX-9　**1プロセッサ**：102.4GF　**最大**839TF　8192**台**

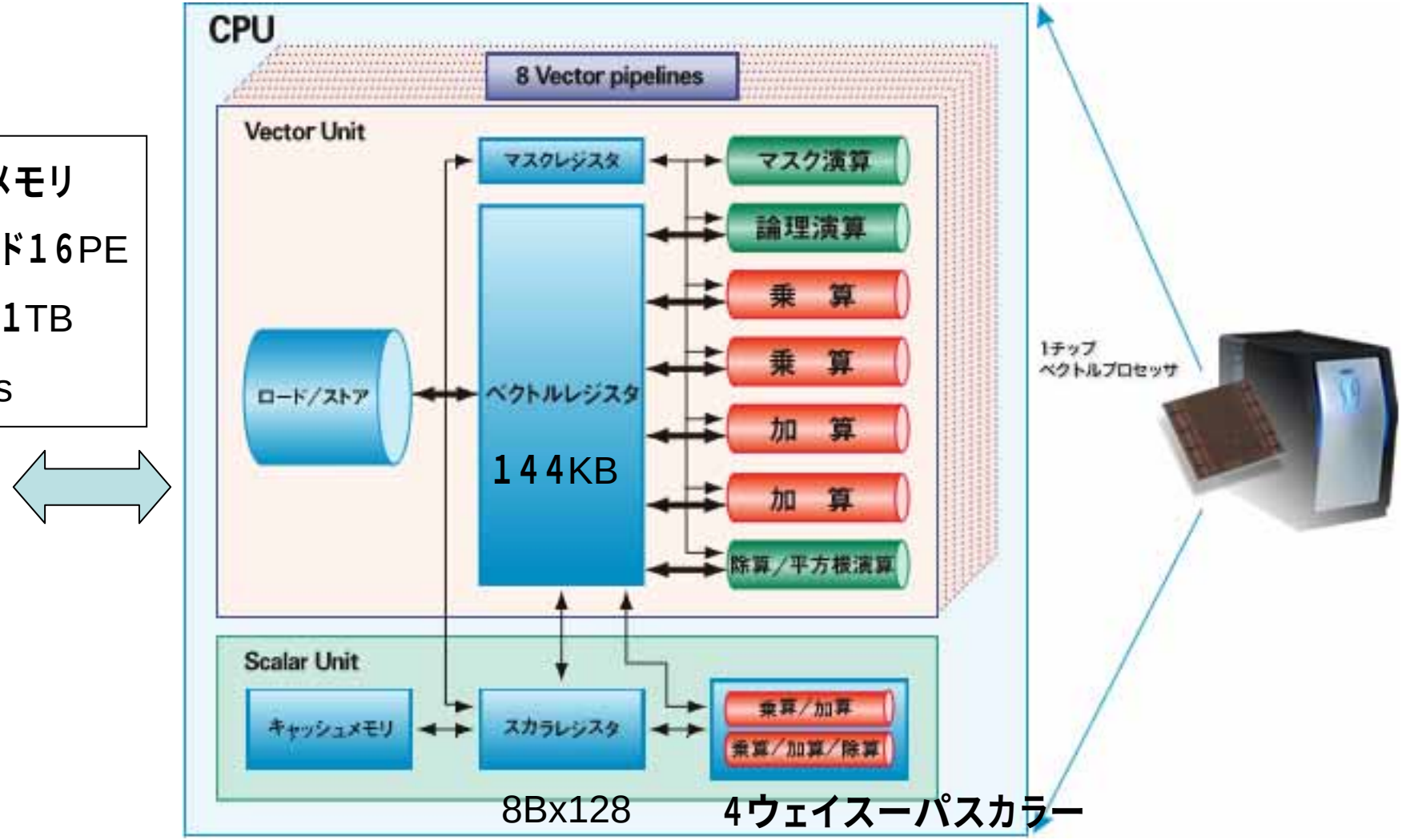
（　3.2GHz、8-16PE/**1ノード**、**最大**512**ノード**、0.065μm、**11層銅配線**）

共有メモリ

1ノード16PE

メモリ1TB

4TB/s

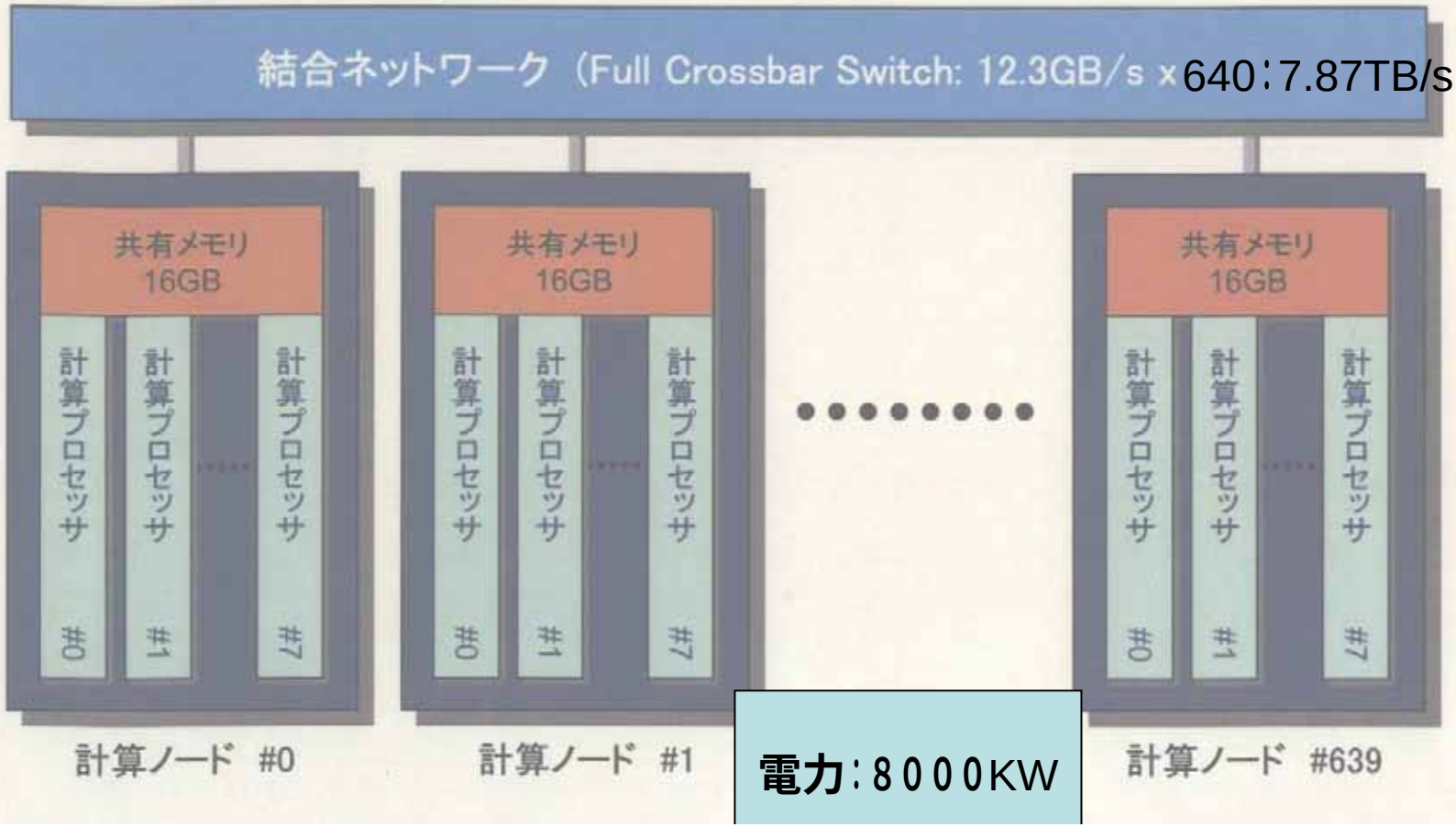


ノード間：**128GB/sx2**

地球シミュレータ

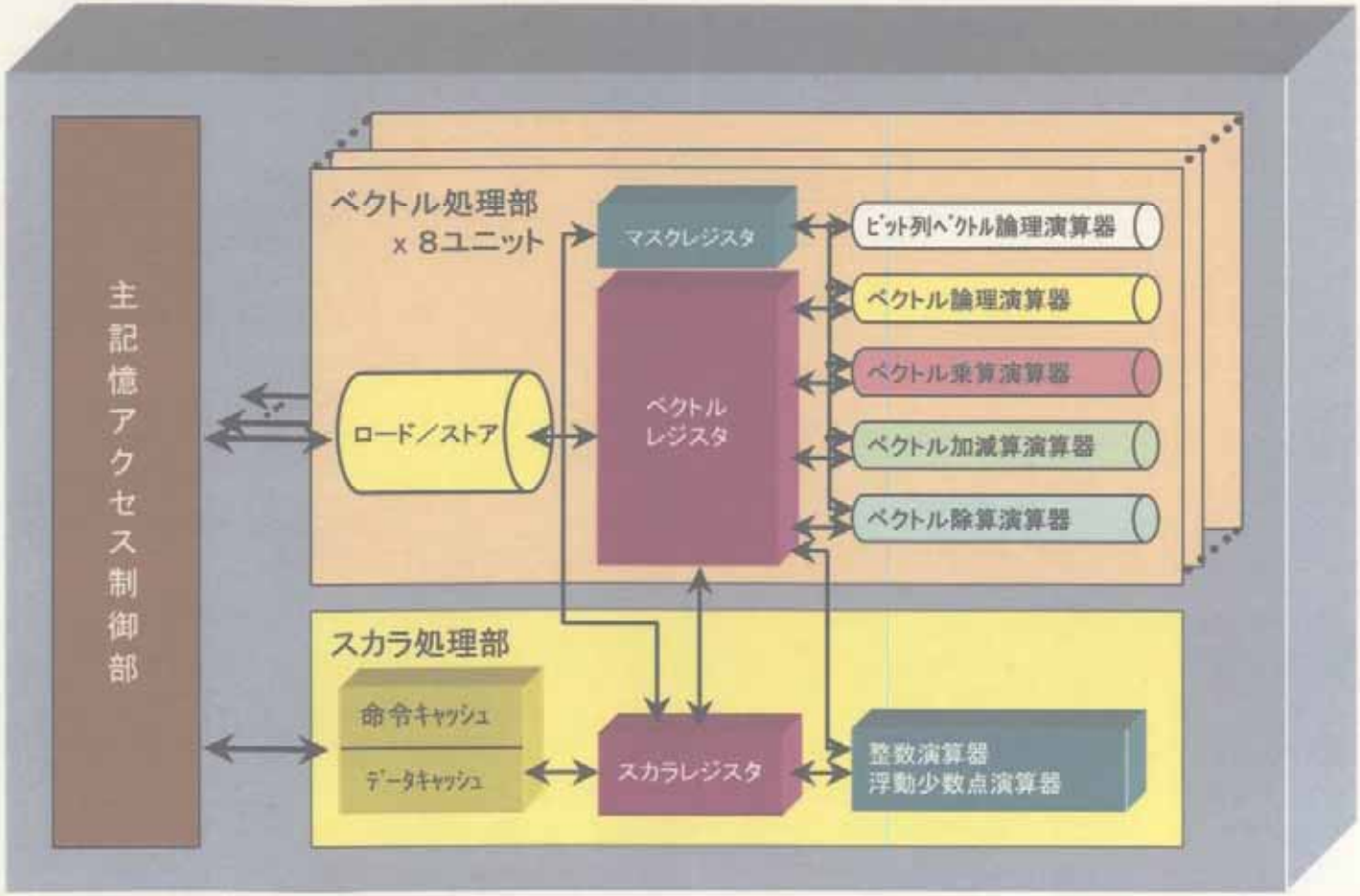
地球シミュレータの全体構成

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



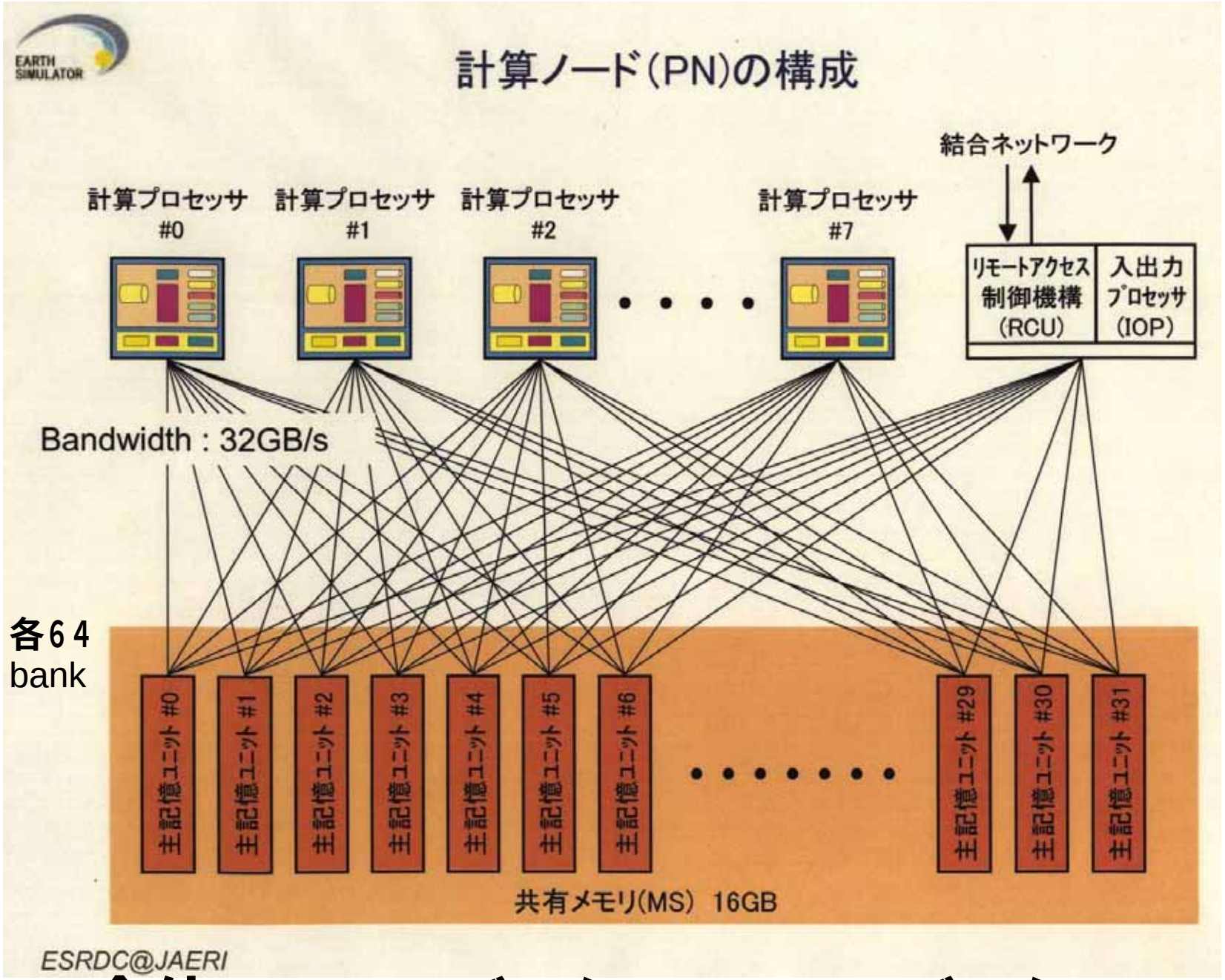
計算プロセッサ(AP)の構成

- ベクトルユニット:8セット
  - 6種のベクトルパイプライン
  - 256要素のベクトルレジスタ: 72個
  - 256ビットのマスクレジスタ: 17個
- 主記憶アクセス制御部
- スカラユニット
  - 4-ウェイ スーパースカラ
  - 64KB 命令キャッシュ
  - 64KB データキャッシュ
  - 128個の汎用レジスタ



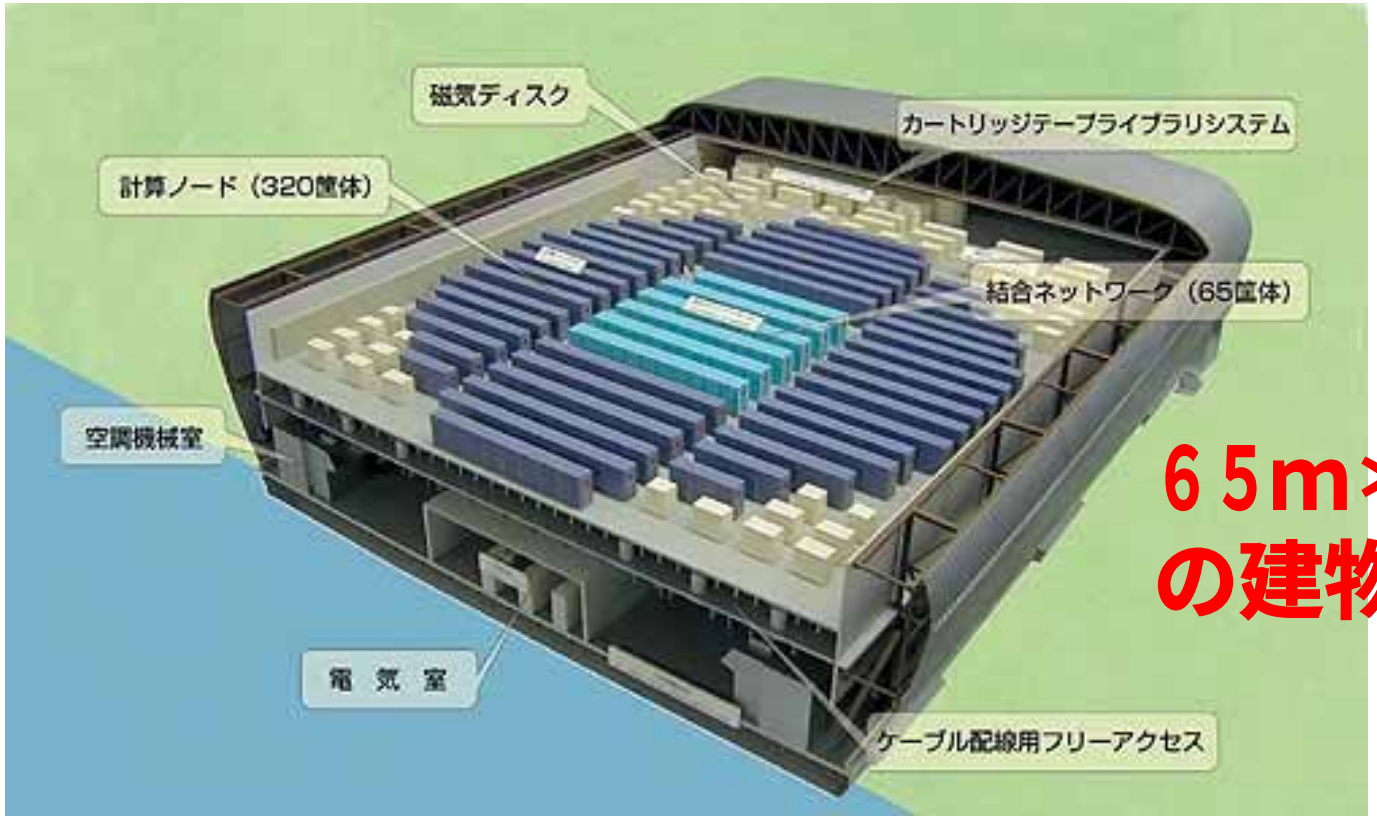
1チップLSI: 8Gflop

- 0.15μm CMOSテクノ
  + 銅配線
- 20.79mm x 20.79mm
- 5,700万トランジスタ
- 5185 ピン
- クロック周波数
  500MHz(1GHz)
- 消費電力
  135W(Typ.)



全体で2048バンク、24nsec/バンク

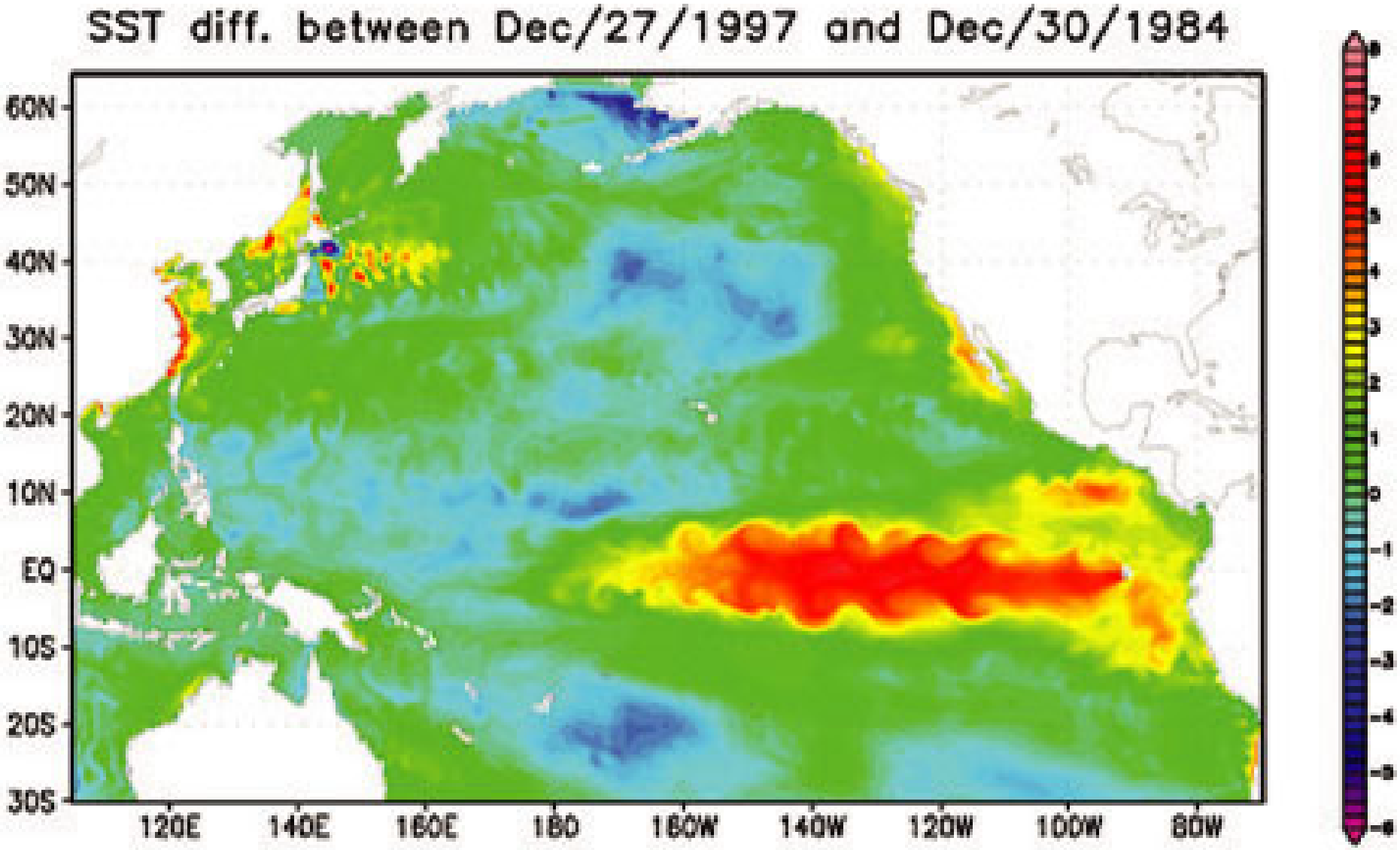
「地球シミュレータ」とは、どんなコンピュータか



６５ｍ＊５０ｍの建物

- ８台のスーパーコンピュータからなる計算ノードを、高速のネットワークで６４０台つないだものです。
  （総計５１２０個のスーパーコンピュータから構成）
- 完成時には世界最速のコンピュータになると予想されます。
- 平成１４年３月からの利用開始を目指して開発中です。

地球を１０km四方に分割



地球シミュレータ施設（航空写真）

計算ノード・結合ネットワーク間ケーブル敷設作業
(平成13年2月～平成13年5月)

計算ノード・結合ネットワーク間ケーブル敷設完了
（平成13年5月）

地球シミュレータ　設置完了（平成14年1月）

BlueGene/L：IBM

- 2005年稼動予定
- 65，536プロセッサ
- 360　TFLOPS
- メッセージパッシング
- 3D-トーラス：64x32x32
  適応ルーティング，仮想チャネル4本
- ブロードキャスト，リダクション：トリー

表 2.5　ASCI　プラットフォームの概要

名称	Red	Blue Pacific	Blue Mountain	White	T30
設置研究所	Sandia	Lawrence Livermore	Los Alamos	Lawrence Livermore	Los Alamos
メーカ	Intel	IBM	SGI	IBM	未定
使用 MPU	9,536×Pentium Ⅱ Xeon	5,856×Power PC	Origin2000 MIPS R10000	8,192×Power 3-Ⅱ	未定
目標性能	1.8Tflops メモリ 606GB Disk 容量 40TB	3.1Tflops メモリ 2.6TB Disk 容量 75TB	3.1Tflops メモリ 2.5TB Disk 容量 75TB	10.2Tflops メモリ 2.5TB Disk 容量 75TB	30 + Tflops
実績	3.2Tflops ('99/10 月)	3.9Tflops ('98/10 月)	3.1Tflops ('98 年)	― ― ―	― ― ―

(注意)　性能はピーク性能値である。

京都新聞　2003年（平成15年）11月23日

日米スパコン計算最速でしのぎ

用途拡大で躍起

毎秒35兆回→360兆回…1000兆回へ

科学技術計算に利用されるスーパーコンピューター（スパコン）で、日米間の世界最速争いに拍車が掛かっている。スパコンはヒトのゲノム（全遺伝情報）を活用した医薬研究などに用途が広がってきた上、国防上も重要な役割を果たすためだ。最速マシンを日本製に奪われた米国側がナンバーワン奪回へ躍起となっている。

「日本は科学技術計算で新時代を切り開いた」とで称賛されるべきだが、米国が新時代に遅れてはならない」―。エーブラハム米エネルギー長官は今月十日、ワシントンでの講演で世界最速を誇る日本の「地球シミュレータ」（ES）を引き合いにし、最速の座を取り戻すと強調した。

二〇〇二年に完成したESは、海洋科学技術センターなどが開発し、NECが製造に当たった。平和利用を目的とする用途は気候変動予測、地殻変動の解明などで、実測値で毎秒三五・八六〇兆回の計算能力を持っている。それまでの最速は、核兵器の備蓄管理などを目的に米国で用いられるIBM製マシンの毎秒七・二二六兆回だっただけに、「衝撃的な数字」（米ニューヨーク・タイムズ紙）だ。

米国の科学者らが今月中旬にまとめた最新のスパコン上位五百機リストでも、ESは登場以来の首位を堅持、二位の米ヒューレット・パッカード製マシンの三倍近い能力となっている。

ただ、IBMは遺伝子からつくられる、たんぱく質の構造解析などに利用するため、毎秒一〇〇〇兆回規模の計算能力を備えたスパコン開発計画を一九九九年から推進中。〇一年には計画の一環として、米エネルギー省と共同で「ブルー・ジーン（青い遺伝子）／L」と呼ぶスパコン開発を打ち出している。目指す能力は毎秒三百六十兆回。既に小型試作機が完成し、IBMは最新ランキングで試作機が七十三位に入ったとアピール。「〇五年の完成時にはトップになる」と首位奪還を予告している。

8キャビネット
45TFLOPS
電力216KW

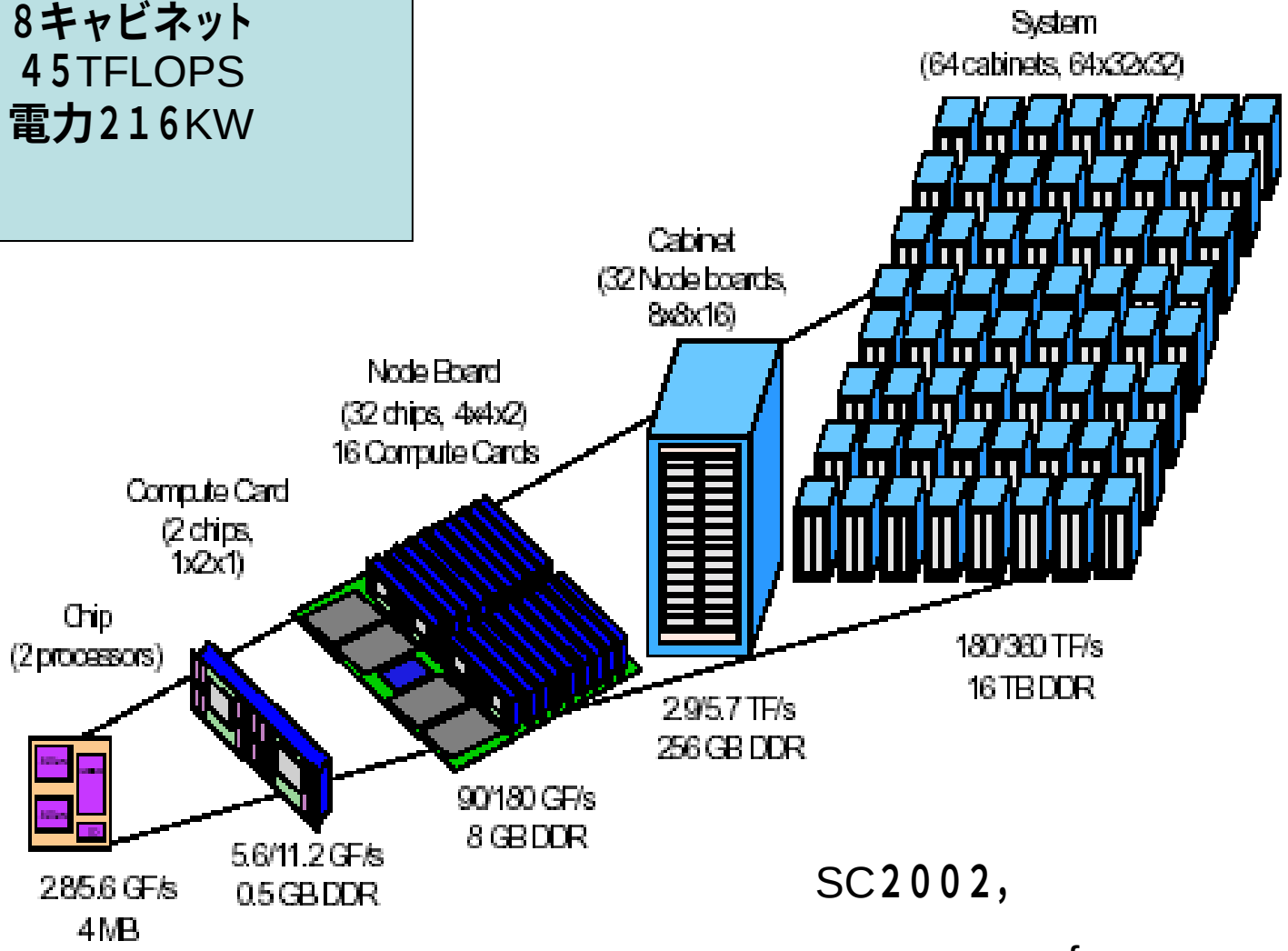


Figure 1: BlueGene/L packaging.

PowerPC440

2コア、内1つは
通常通信に使用

SC2002,

www. sc-conference.org/sc2002/

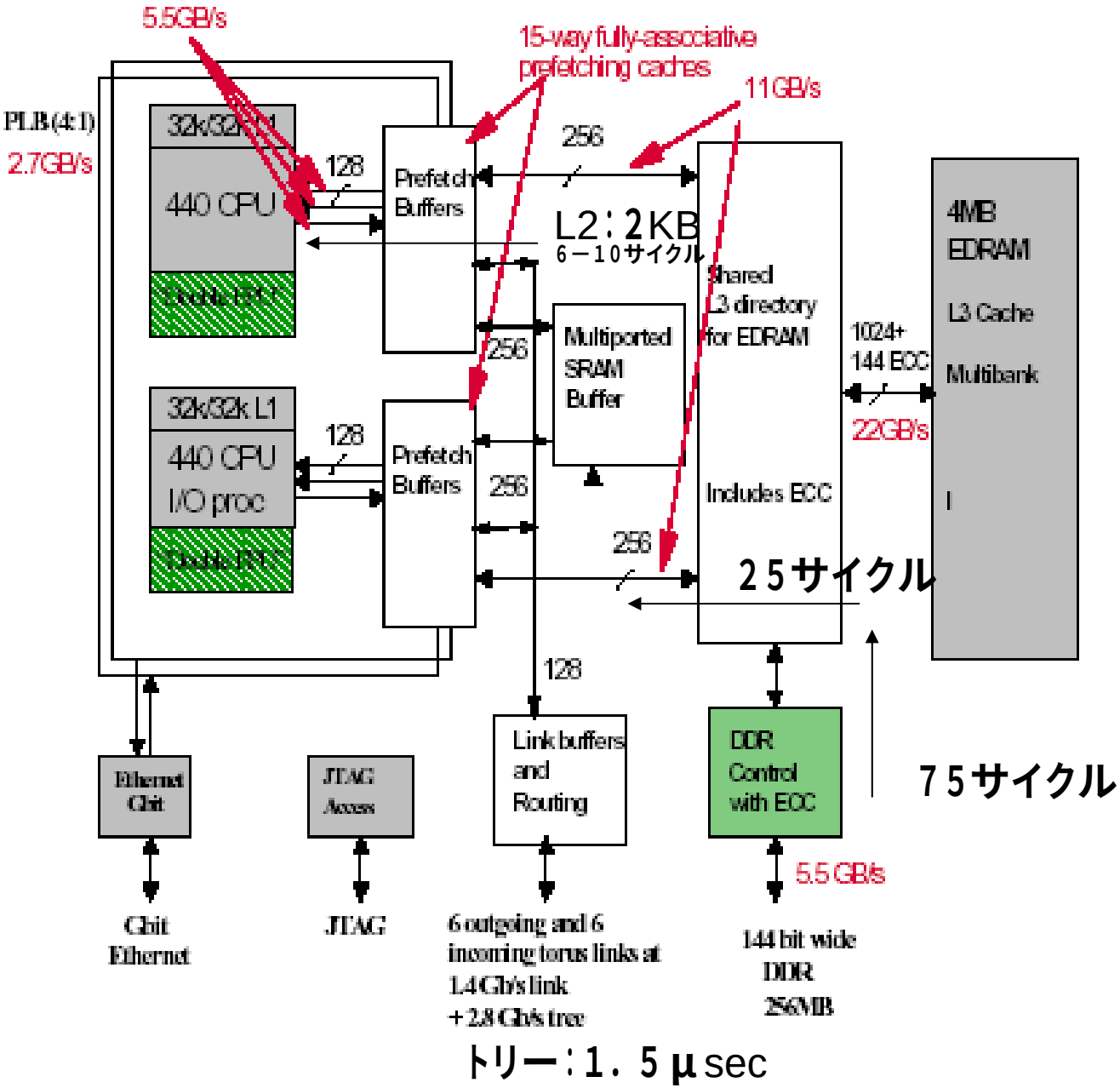


Figure 3: BlueGene/L node diagram. The bandwidths listed are targets.

目指せ世界最速タイトル奪還

次世代スパコン開発へ

1秒1京回

文部科学省は二十五日までに、最大演算速度が一〇ペタフロップス（一秒間に一京回＝一京は一兆の一万倍）の次世代スーパーコンピューター「京速計算機システム」の設計、開発に着手することを決めた。

二〇〇六年度概算要求に研究費数十億円を盛り込む。総事業費は八百億―一千億円に上る見込み。一〇年完成を目指し、米国のスパコンが持つ世界最速のタイトル奪還に挑む。

京速計算機が完成すれば、この約七十三倍の処理速度を持つことになる。

日本が〇二年から運用を開始した「地球シミュレータ」は、〇四年までの二年半、三五・九テラフロップスで世界最速を誇っていたが、〇五年六月現在で世界第四位に後退している。

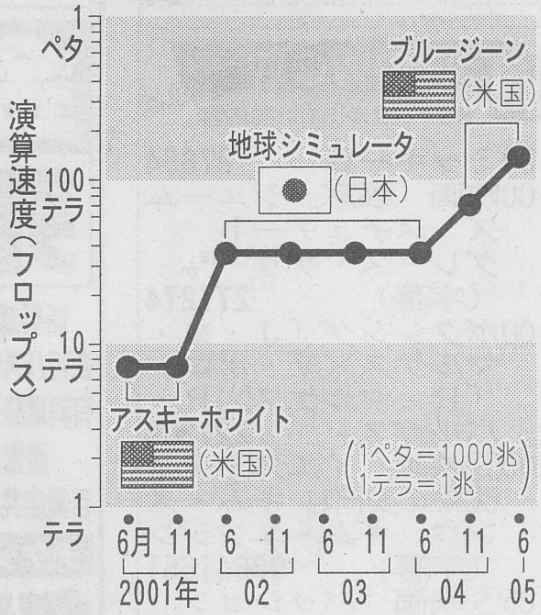
米国も一〇年をめどに数ペタフロップスのスパコン開発を目指しており、競争は激化している。

また、京速計算機システムの運用を担う「先端計算科学技術センター（仮称）」の設立方法や建設場所について調査研究も始める。

現在の世界最速スパコンは、米ローレンスリバモア国立研究所のスパコン「ブルージーン」で、一三六・八テラフロップス（一秒間に約百三十六兆八千億回）。

フロップス　コンピューターの計算方法の一つである「浮動小数点演算」を1秒間に何回実行できるかを示す単位。コンピューターの処理速度の目安として使われる。例えば、1ギガフロップスなら1秒間に10億回、1テラフロップスなら1秒間に1兆回の浮動小数点演算能力を持つ。

世界最高速スパコンの変遷
（独マンハイム大などが年2回調査、対数グラフ）



2005.7.25
京都新聞

文科省

次世代スーパコンピュータ
(京 (けい) 速コンピュータ)

地球シミュレータの２５０倍の性能：１０ＰＦＬＯＰＳ

プロセッサはどうする

ネットワークはどうする

省電力はどうする：３０MW？

使いやすいソフトウェアはどうする

アプリケーションはどうする

お金は：１１００億円 (5年間)

開始：２００６

完成：２０１２

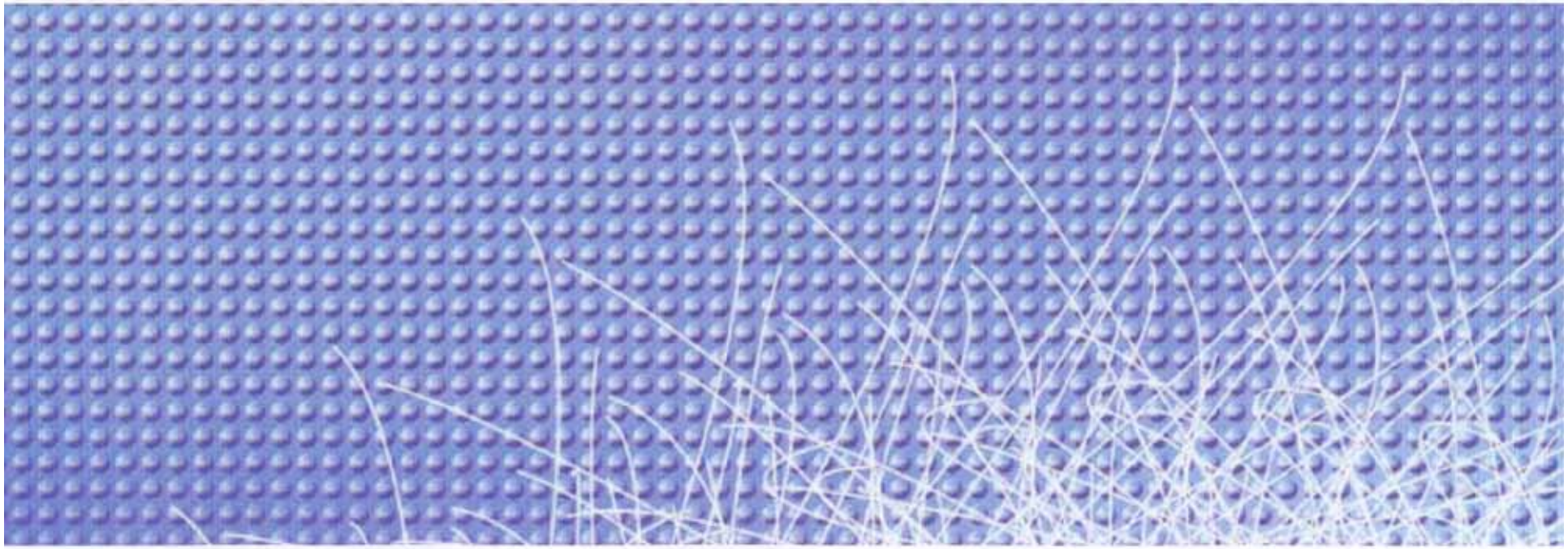
場所：神戸ポートアイダンド、2007.3.決定

体制：理化学研究所統括

富士通＋日立＋NEC

次世代スーパーコンピュータの開発と利用

独立行政法人 理化学研究所

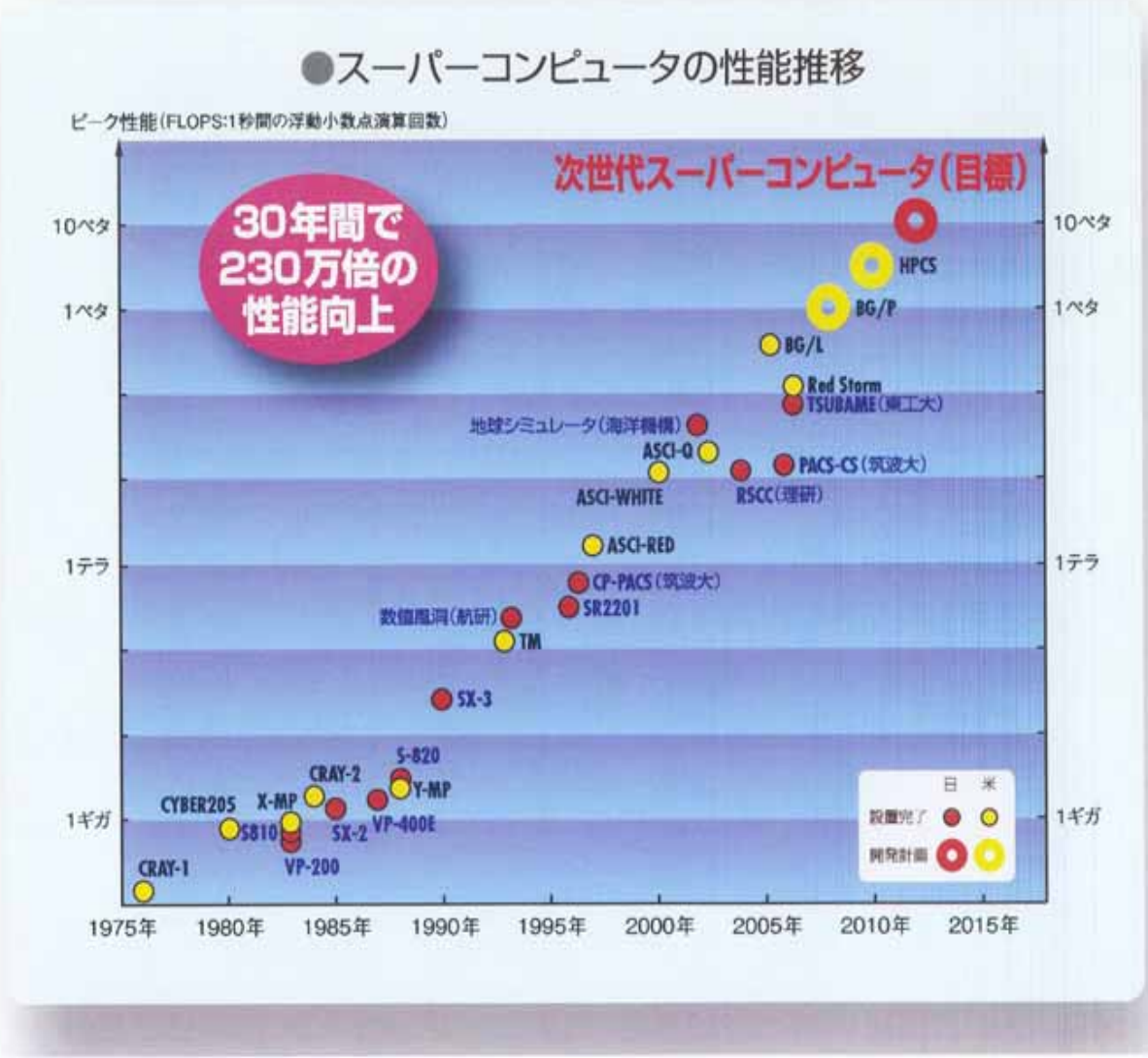


広範な分野での利活用
-次世代スーパーコンピュータが拓く世界-

スーパーコンピュータによるシミュレーションは、実験、理論と並ぶ研究開発の第３の手法として益々重要になっています。スーパーコンピュータは、自動車や飛行機の設計・製作のための構造解析や流体解析、天気予報のための気象シミュレーションなど様々な分野で使われており、今後の科学技術と産業の発展に不可欠です。

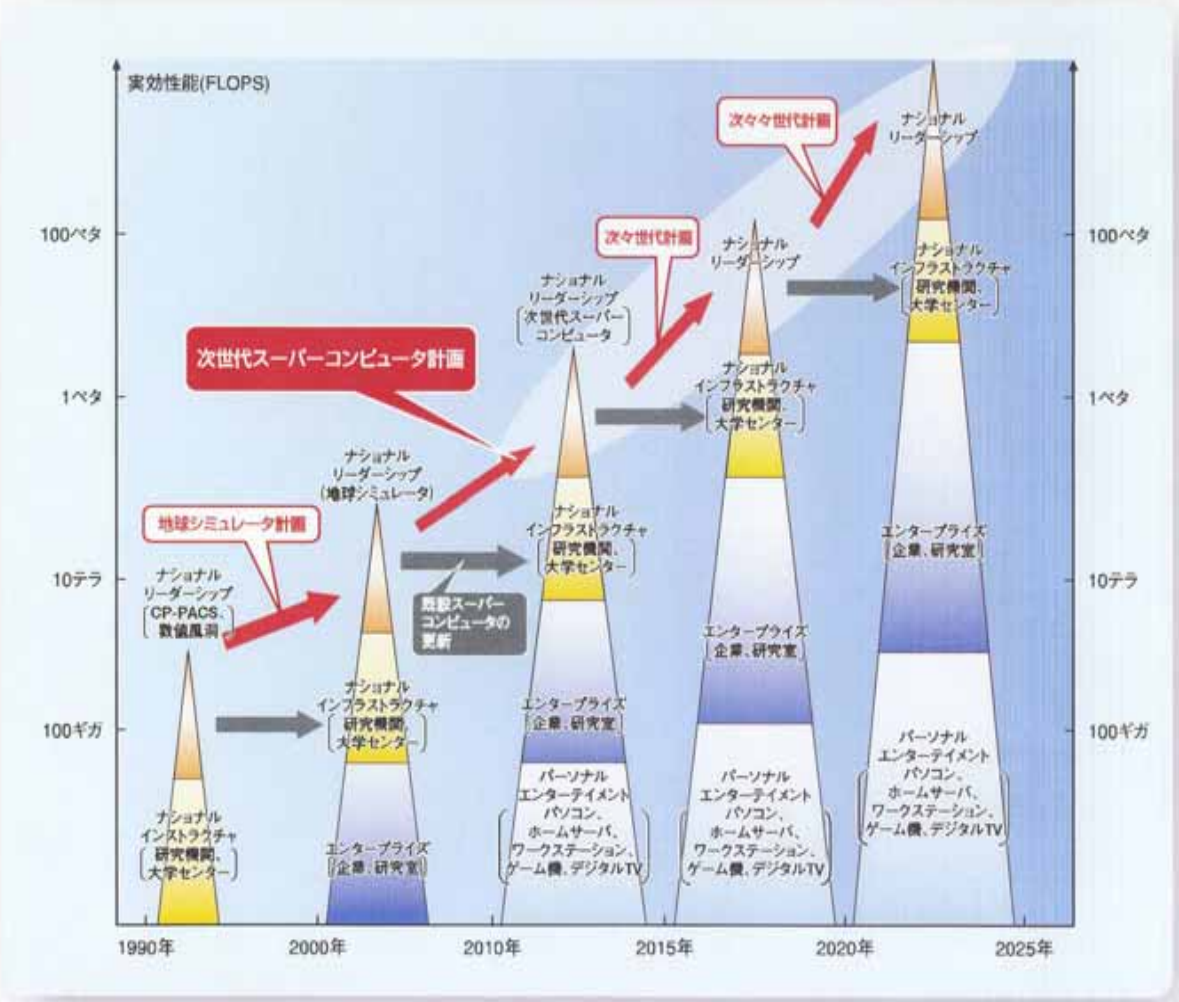
スーパーコンピュータの性能は急速に進歩しています。

この30年間でスーパーコンピュータの計算能力は200万倍以上にもなっています。スーパーコンピュータの開発には半導体技術、光通信技術、ネットワーク技術、品質管理技術など、エレクトロニクスに関する総合的かつ高度な技術力が必要です。これらは家電や情報通信など多くの分野で国民生活と産業に不可欠な技術の基盤となります。



継続的な最先端・高性能スーパーコンピュータの開発ビジョン

これまで、国家プロジェクトで開発された最先端のスーパーコンピュータシステム（ナショナルリーダーシップシステム）の技術は、全国の計算機システムに展開されてきました。更に、スーパーコンピュータを将来に亘って継続的に開発することにより、我が国の研究開発基盤の強化と技術の維持向上に大きく貢献することが期待されます。



国の総合科学技術会議は、我が国の科学技術及び産業の競争力の基盤となる「次世代スーパーコンピュータの開発・利用」プロジェクトを「国家基幹技術」と位置づけ、集中的に投資して推進することとしています。

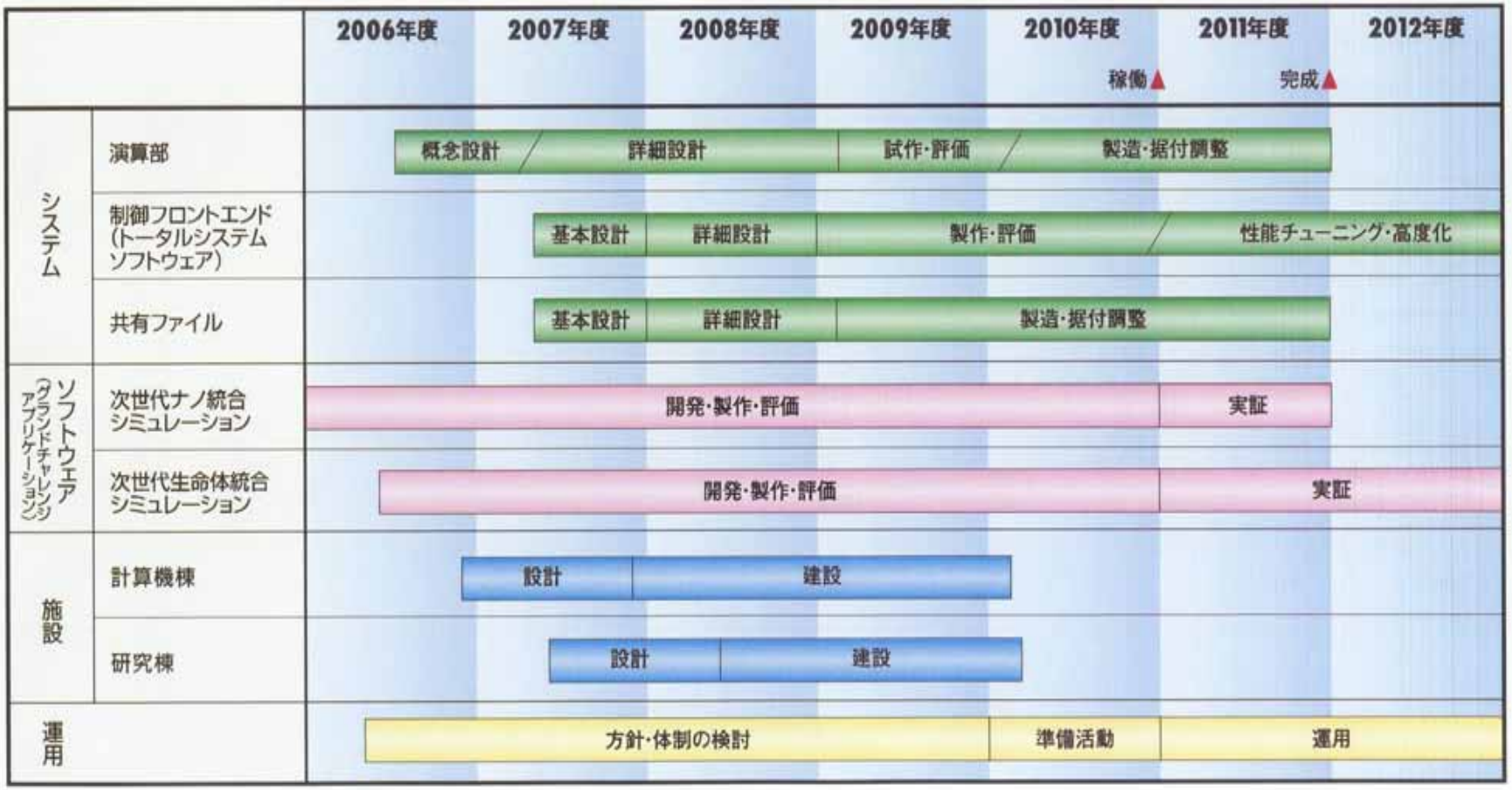
プロジェクトの概要

理論、実験と並び、現代の科学技術の方法として確固たる地位を築きつつあるスーパーコンピューティングを更に発展させるため、長期的な国家戦略を持って取り組むべき重要技術（国家基幹技術）である「次世代スーパーコンピュータ」を2012年の完成を目指して開発します。
今後とも我が国が科学技術・学術研究、産業、医・薬など広範な分野で世界をリードし続けるべく、

①世界最先端・最高性能の「次世代スーパーコンピュータ(注)」の開発・整備 (注) 10ペタFLOPS級
②次世代スーパーコンピュータを最大限利活用するためのソフトウェア（＝グランドチャレンジ・アプリケーション）の開発・普及
③次世代スーパーコンピュータの共同利用と学術情報ネットワークを介した全国のスーパーコンピュータの重層的な利用環境の整備（＝サイバー・サイエンス・インフラストラクチャの整備）
④次世代スーパーコンピュータを中核とする世界最高水準のスーパーコンピューティング研究教育の拠点の形成

を文部科学省のイニシアティブにより、開発主体である理化学研究所を中心に産学官の密接な連携の下、一体的に推進します。

【開発スケジュール(予定)】


2006年度
2007年度
2008年度
2009年度
2010年度
2011年度
2012年度
稼働
完成
システム
演算部
概念設計
詳細設計
試作・評価
製造・据付調整
制御フロントエンド(トータルシステムソフトウェア)
基本設計
詳細設計
製作・評価
性能チューニング・高度化
共有ファイル
基本設計
詳細設計
製造・据付調整
ソフトウェア(グランドチャレンジアプリケーション)
次世代ナノ統合シミュレーション
開発・製作・評価
実証
次世代生命体統合シミュレーション
開発・製作・評価
実証
施設
計算機棟
設計
建設
研究棟
設計
建設
運用
方針・体制の検討
準備活動
運用

次世代スーパーコンピュータの開発・整備

理化学研究所は、プロジェクトの中核機関として、スーパーコンピュータ開発をリードする最高水準の汎用システムである「次世代スーパーコンピュータ」を開発・整備します。

【システム構成概要】

【世界最速のシステム】

●1秒間に1京(ケイ=10の16乗)回の計算性能=10ペタFLOPS級

【汎用システム】

●科学技術・産業で用いられる多様なアプリケーションやこれまで不可能だった複雑かつ大規模なシミュレーションが実行可能

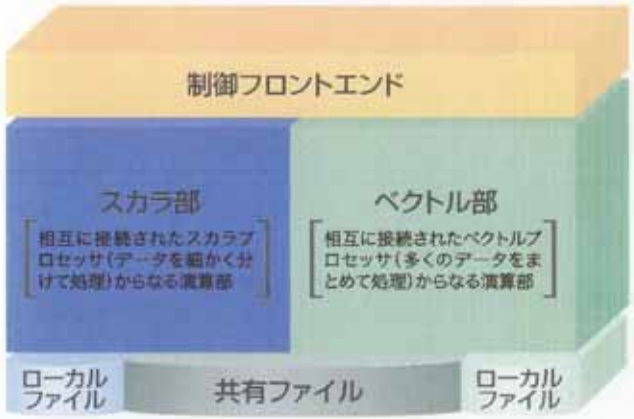
【革新的なシステム】

●先端技術の積極的導入により、画期的な省電力、省スペースを実現。理化学研究所とメーカー3社(富士通、日本電気、日立製作所)との共同により、日本の技術の総力を結集して開発

【システムの基本的な構成】

多様なシミュレーションの実行に最適な計算環境を提供する複合汎用スーパーコンピュータシステム

- シミュレーションの特性に応じた最適な演算部で計算を実行
- スカラ部とベクトル部の併用により、従来困難だった複雑かつ大規模なシミュレーションも実行可能



【建設地】

- 兵庫県神戸市中央区港島南町7丁目（ポートアイランド第2期内）
  ポートアイランド南駅より徒歩約1分（JR新神戸駅から約25分、神戸空港から約10分）

ライフサイエンス分野

次世代生命体統合シミュレーションソフトウェアの研究開発

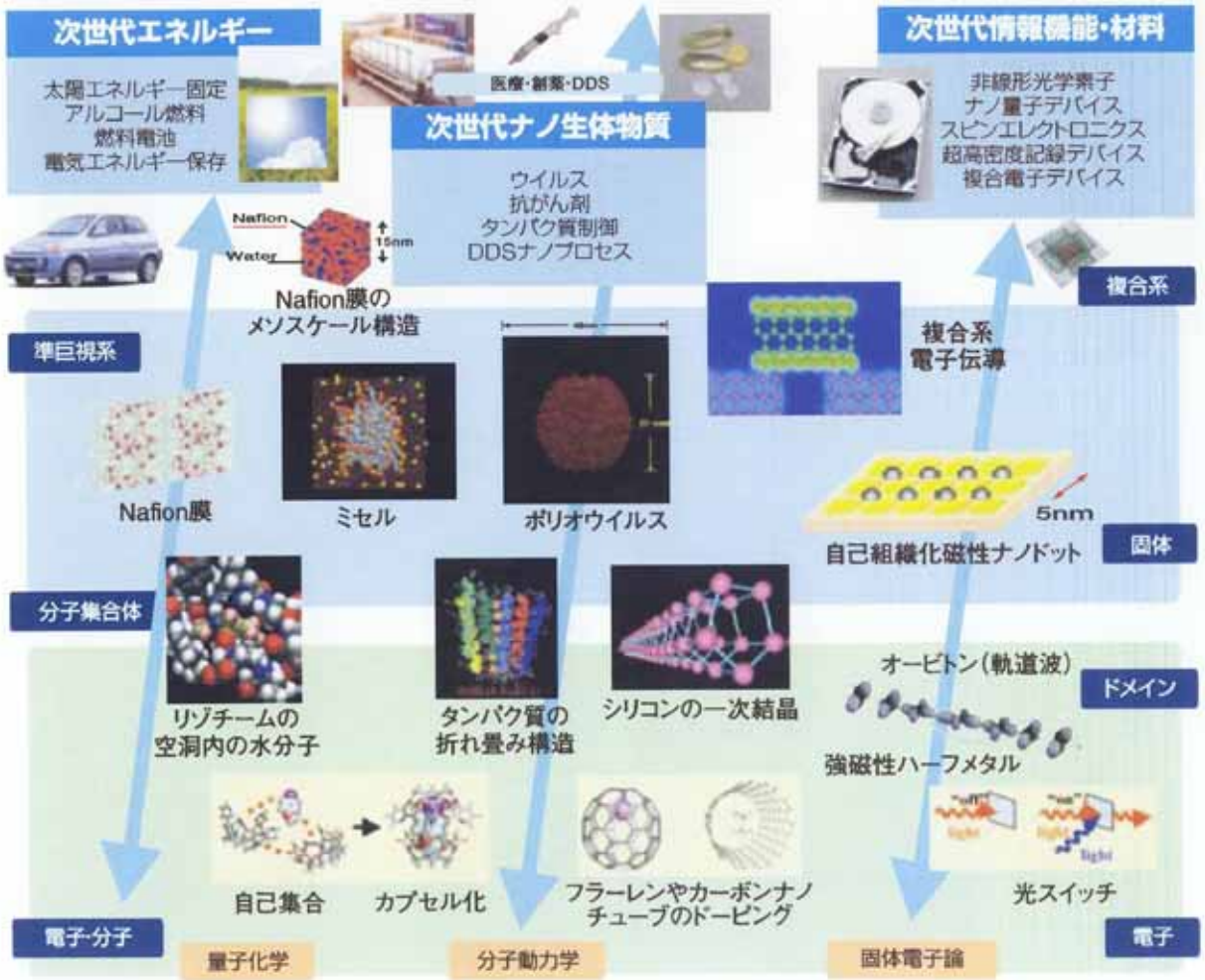
【概要と目標】

理化学研究所は、次世代スーパーコンピュータの性能を十分に発揮させ、分子から全身まで生体内で起こる種々の現象を統合的に理解するためのシミュレーションソフトウェアの研究開発を進めます。シミュレーションによる予測を基礎とした研究手法をライフサイエンスの分野に提供し、生命現象の統合的な理解による科学的進展、産業応用等を通じたヘルスサイエンスへの貢献を目指します。

ナノテクノロジー分野

次世代ナノ統合シミュレーションソフトウェアの研究開発

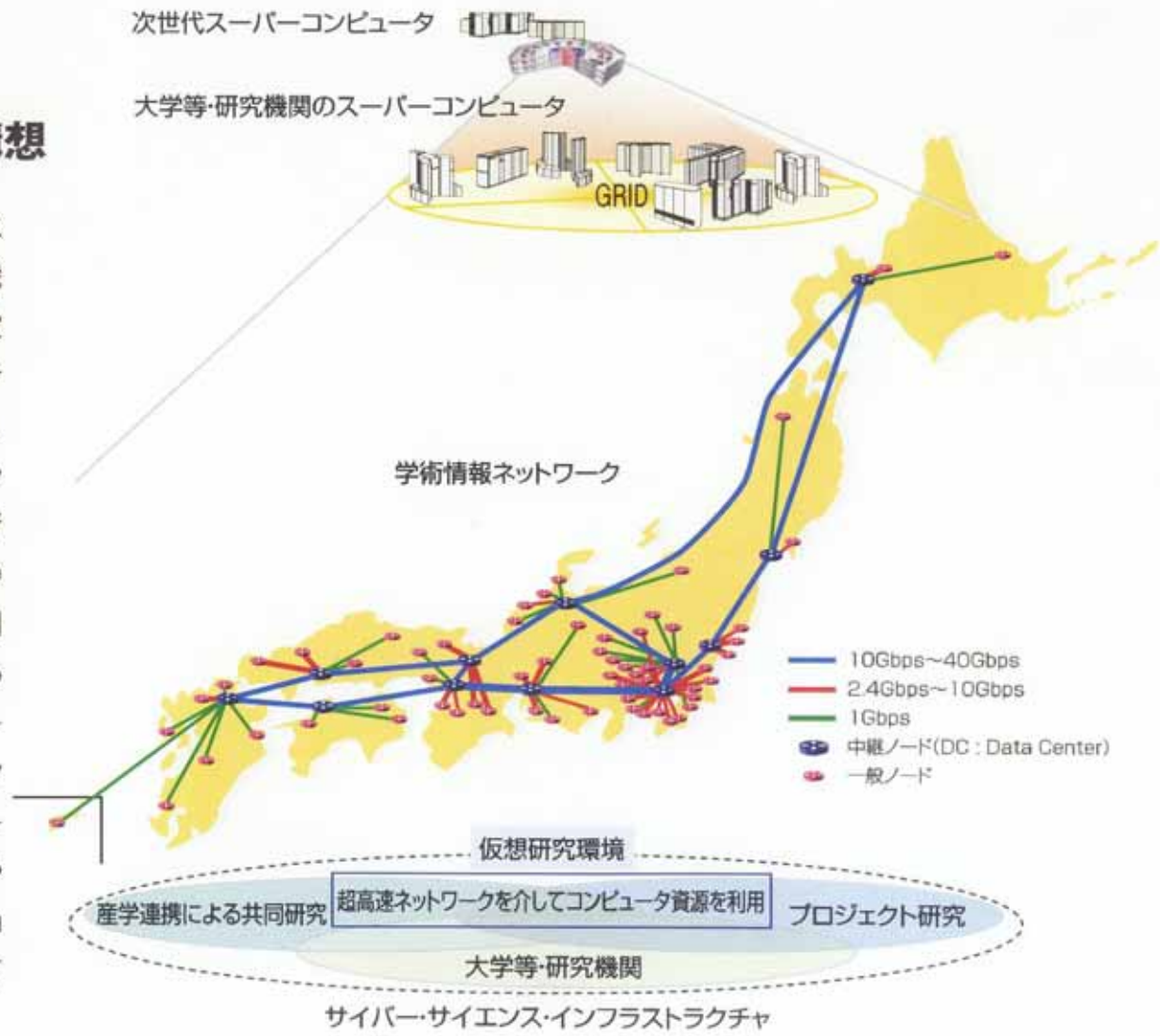
分子科学研究所は、最先端の知的ものづくりを実現する新材料開発のためのアプリケーション・ソフトウェアの開発を行う拠点です。開発されたソフトウェアが次世代スーパーコンピュータで十分に性能を発揮するよう、理化学研究所のハードウェア開発と密接な連携を図っています。



利用環境の整備

サイバー・サイエンス・インフラストラクチャ構想

サイバー・サイエンス・インフラストラクチャ構想とは、大学や研究機関が有しているコンピュータ等の設備、基盤的ソフトウェア、学術コンテンツ及び学術データベース、人材、研究グループそのものを超高速ネットワーク上で共有する最先端学術情報基盤であり、大学等との連携の下、国立情報学研究所や全国共同利用情報基盤センター等が中核となって推進しています。次世代スーパーコンピュータは、サイバー・サイエンス・インフラストラクチャ上で、スーパーコンピュータを保有する大学や研究機関等と連携しながら、ナショナルリーダーシップシステムとしての機能を発揮していきます。



資料３

「最先端・高性能汎用スーパーコンピュータの開発利用」プロジェクトの実現に向けて（案）

文部科学省研究振興局２００５年１０月２６日発表

開発主体：理化学研究所

平成１８年度-平成２２年度

１１５４億円（平成１８年度４０億円）

米国：２００９年1PFLOPSコンピュータ開発予定

科学新聞２００５年１１月４日

平成１７年８月１０日

文部科学省研究振興局

文部科学省ホームページより

最先端・高性能汎用スーパーコンピュータの開発利用（案）

目的：世界最先端・最高性能のスーパーコンピュータ「汎用京速計算機」システムの開発・整備及び利用技術の開発・普及

趣旨及び効果：理論、実験と並び、現代の科学技術の方法として確固たる地位を築きつつあるスーパーコンピューティング（シミュレーション（数値計算）やデータマイニング、解析等）について、今後とも我が国が世界をリードし続けるため、

（1）スーパーコンピュータを最大限利活用するためのソフトウェア等の開発・普及

（2）世界最先端・最高性能の汎用京速(注)計算機システムの開発・整備　　(注)京速＝10ペタFLOPS

（3）上記（2）を中核とする世界最高水準のスーパーコンピューティング研究教育拠点（COE）「先端計算科学技術センター（仮称）」の形成により研究水準向上と世界をリードする創造的人材の育成を総合的に推進。

世界最高性能の科学技術計算環境を実現し、複雑で多様な現象の系全体のシミュレーションや高度なデータマイニング、解析等を、幅広い分野で行い、「知的ものつくり」や「科学的未来設計」を実問題で可能とし、先端的スーパーコンピューティングにおける国際的なリーダシップを確立。科学技術・学術や産業の競争力強化、安全・安心な社会の構築に貢献。

また、世界の英知を結集し、世界水準の人材育成を行い、シミュレーションにおける我が国の国際的な地位を確立する。

概要：平成18年度は、世界最先端・最高性能の汎用京速計算機システムの開発・整備の前提であるシステム全般の設計・研究開発等に着手する。

1. ソフトウェア（OS、ミドルウェア、アプリケーションソフトウェア）等の設計・研究開発
2. ハードウェア（計算機システム及び超高速インターコネクション）の設計・研究開発
3. 「先端計算科学技術センター（仮称）」の形成に関する調査研究

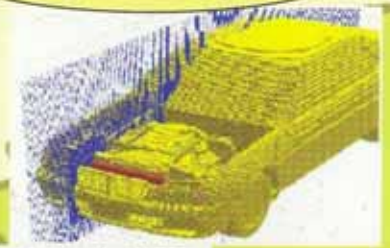
体制：国の責任で設備の整備から運用まで一体的に推進する。また装置の開発・運用を行うに当り、産学官の様々な組織から最も適したところを選択し、そのポテンシャルを活用する。

事業期間：平成18年度～24年度

科学技術・学術の発展と産業競争力強化に貢献
（以下を例に、様々な科学技術・学術・産業分野を対象）

広汎な分野での利活用 － 次世代スパコンが拓く世界 －

ものづくり



自動車開発

提供：日産自動車（株）

ナノテクノロジー

物質設計

提供：（独）物質・材料研究機構

触媒設計

提供：（独）物質・材料研究機構

防災

津波被害予測

提供：東北大学

雲の解析

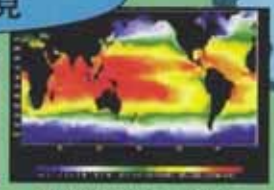
提供：気象研究所

ライフサイエンス

提供：東京大学 他

地球環境

エルニーニョ現象の影響予測



提供：（独）海洋研究開発機構

原子力

提供：日本原子力研究所

原子炉丸ごと解析

提供：日本原子力研究所

レーザー反応解析

天文・宇宙物理

銀河形成解明

提供：（独）理化学研究所

惑星形成解

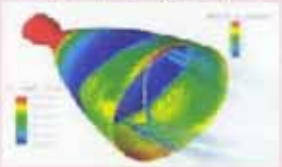
提供：国立天文台

オーロラ発生解明

提供：（独）海洋研究開発機構

航空・宇宙

ロケットエンジン設計



提供：（独）宇宙航空研究開発機構

航空機開発

提供：（独）宇宙航空研究開発機構

先端計算科学技術センター（仮称）

汎用京速計算機が目指すグランドチャレンジ（例）

世界最高水準の科学技術創造立国を実現するため、国際競争力を支える新産業創造等の政策目標の実現をも視野に入れ、ナノテクノロジー／ライフサイエンス分野を革新する汎用京速計算機のグランドチャレンジを明示して戦略的に研究開発を進める。

＜ナノテクノロジー分野アプリケーション＞

次世代ナノ統合シミュレーション

電子・原子・分子から、ナノスケールの分子複合デバイスに至るまで、ナノ材料を丸ごと解析することにより、次世代ナノ材料（新半導体材料等）の創出などの実現を目指す。

＜ライフサイエンス分野アプリケーション＞

次世代生命体統合シミュレーション

遺伝子から全身の血流まで、人体丸ごと解析することにより、テーラーメード医療や創薬などの実現を目指す。

研究開発スケジュール（案）

年度		平成17年度	平成18年度	平成19年度	平成20年度	平成21年度	平成22年度	平成23年度	平成24年度
開発項目＼評価等		★研究開発チーム発足	計画本格化判断★（設計仕様、開発体制、立地・運用方針等）			研究開発状況評価★（システム性能・機能等）		COE形成、運用評価★（利用状況、研究成果、人材育成状況等）	
ソフトウェア	システムソフトウェア	NAREGI[※4]（平成15年度より）	異機種統合ソフトウェア設計・製作			異機種統合ソフトウェア評価			
			グリッドミドルウェア設計・製作			グリッドミドルウェア評価			
	グランドチャレンジアプリケーション	（※4）	次世代ナノ統合シミュレーション設計・製作			次世代ナノ統合シミュレーション評価			
			次世代生命体統合シミュレーション設計・製作					次世代生命体統合シミュレーション評価	
	革新的シミュレーションソフトウェアの研究開発[※1]				革新アプリケーション評価				
	次世代高精度・高分解能シミュレーション技術の開発[※3]								
ハードウェア	要素技術開発：将来のスーパーコンピューティングのための要素技術の研究開発[※1]								
	要素技術開発：通信・演算情報量の爆発的増大に備える超低消費電力技術の創出[※2]								
	大規模処理計算機部		設計		実装技術設計・評価		製作	システム強化	
	逐次処理計算機部		設計		実装技術設計・評価		製作	システム強化	
	特定処理計算加速部		設計		実装技術設計・評価		製作		
	異機種間接続超高速インターコネクション		設計		実装技術設計・評価		製作		
	遠隔可視化装置					実装設計・評価	製作		
その他	ファイルシステム				設計		製作	システム強化	
	立地調査、建屋建設、付帯設備整備等		検討	設計	建設	付帯設備整備			

：「最先端・高性能汎用スーパーコンピュータの開発利用」以外のプロジェクトを示す。　：プロジェクト部分に該当。

※1：「次世代IT基盤構築のための研究開発」の研究開発領域の一つ。

※2：科学技術振興機構「戦略的創造研究推進事業」の一戦略目標下の研究領域として、「情報システムの超低消費電力化を目指した技術革新と統合化技術」を設定。

※3：科学技術振興機構「戦略的創造研究推進事業」の一戦略目標下の研究領域として、「マルチスケール・マルチフィジックス現象の統合シミュレーション」を設定。

※4：「超高速コンピュータ網形成プロジェクト(National Research Grid Initiative)」。平成15年度よりグリッドミドルウェアとナノシミュレーションソフトウェアの開発を進めている。

かってデータフローコンピュータという魅惑的な研究があった

１９７０年代後半から８０年代前半

スーパスカラ方式のルーツ

特徴

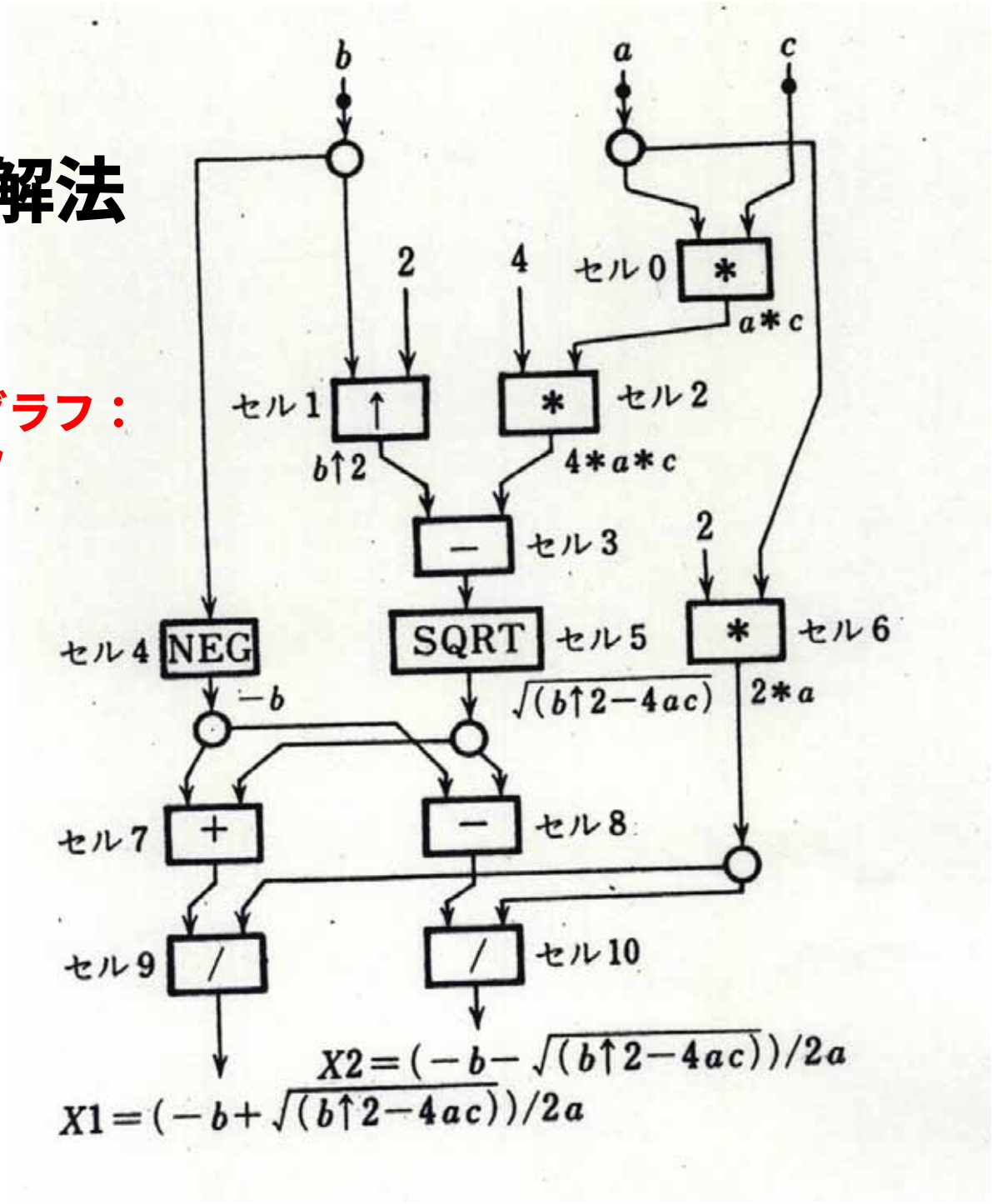
・プログラムカウンタがない

・プログラムに内在するすべての並列性を実行時に取り出せる

VLIW方式のルーツ：水平型マイクロプログラム

２次方程式の解法

データ依存を示すグラフ：
データフローグラフ



データ駆動 ⟷ コントロール駆動

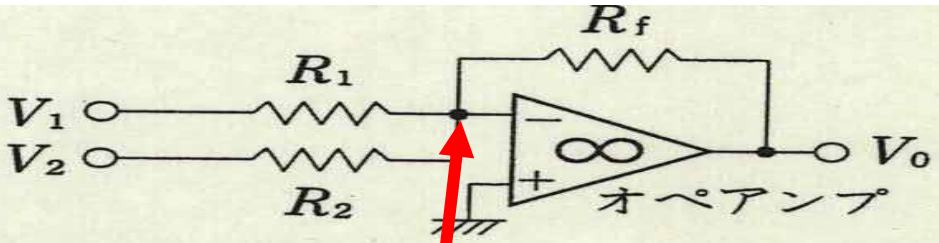
データが揃ったセル　プログラム

↓　カウンタで指された

発火　命令のみ実行

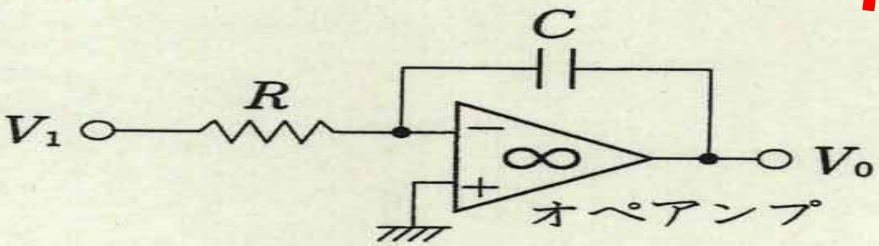
↓

結果を次のセルに渡す



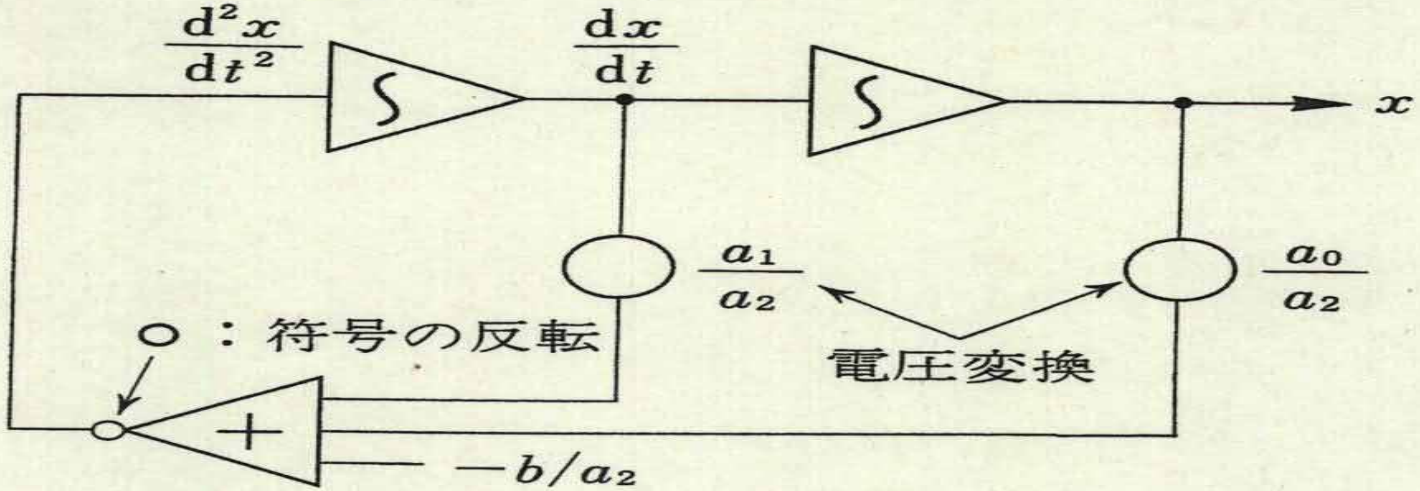
(a) 加算器 $V_0 = -R_f\left(\frac{V_1}{R_1} + \frac{V_2}{R_2}\right)$

仮想接地：0 V

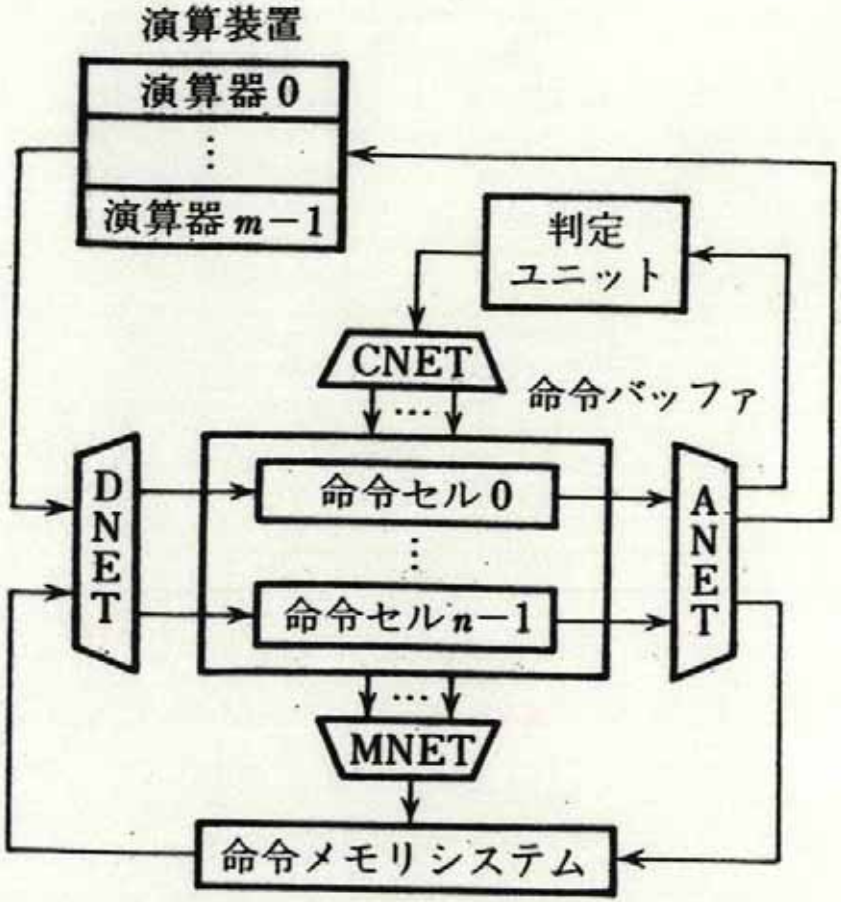


(b) 積分器 $V_0 = -\frac{Q}{C} = -\frac{1}{CR}\int V_1\,dt$

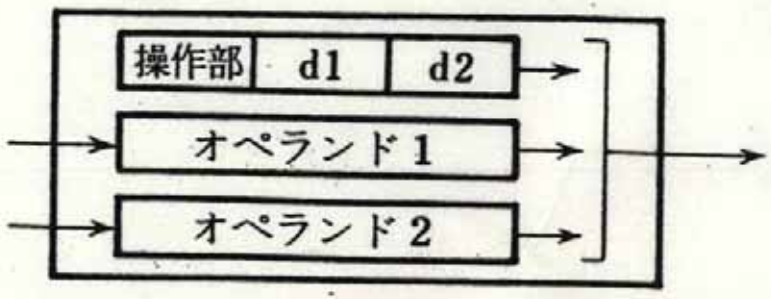
$$a_2\frac{d^2x}{dt^2} + a_1\frac{dx}{dt} + a_0 x = b$$



(c) 常微分方程式を解く回路



（a） ハードウェア構造．ANET：アービトレーション
ネットワーク，DNET：ディストリビューション
ネットワーク，CNET：制御ネットワーク，MNET
：メモリコマンドネットワーク



（b） 各命令セルの構造．d1,d2：デス
ティネーションセルを指定．

図 5. 23　データ駆動型計算機の構成（J. B. Dennis, *et al.*: 'A Preliminary Architecture for a Basic Data-Flow Processor', *Proc. of 2nd Annual Int. Symp. on Computer Architecture*, pp, 126-132 (1975) による）

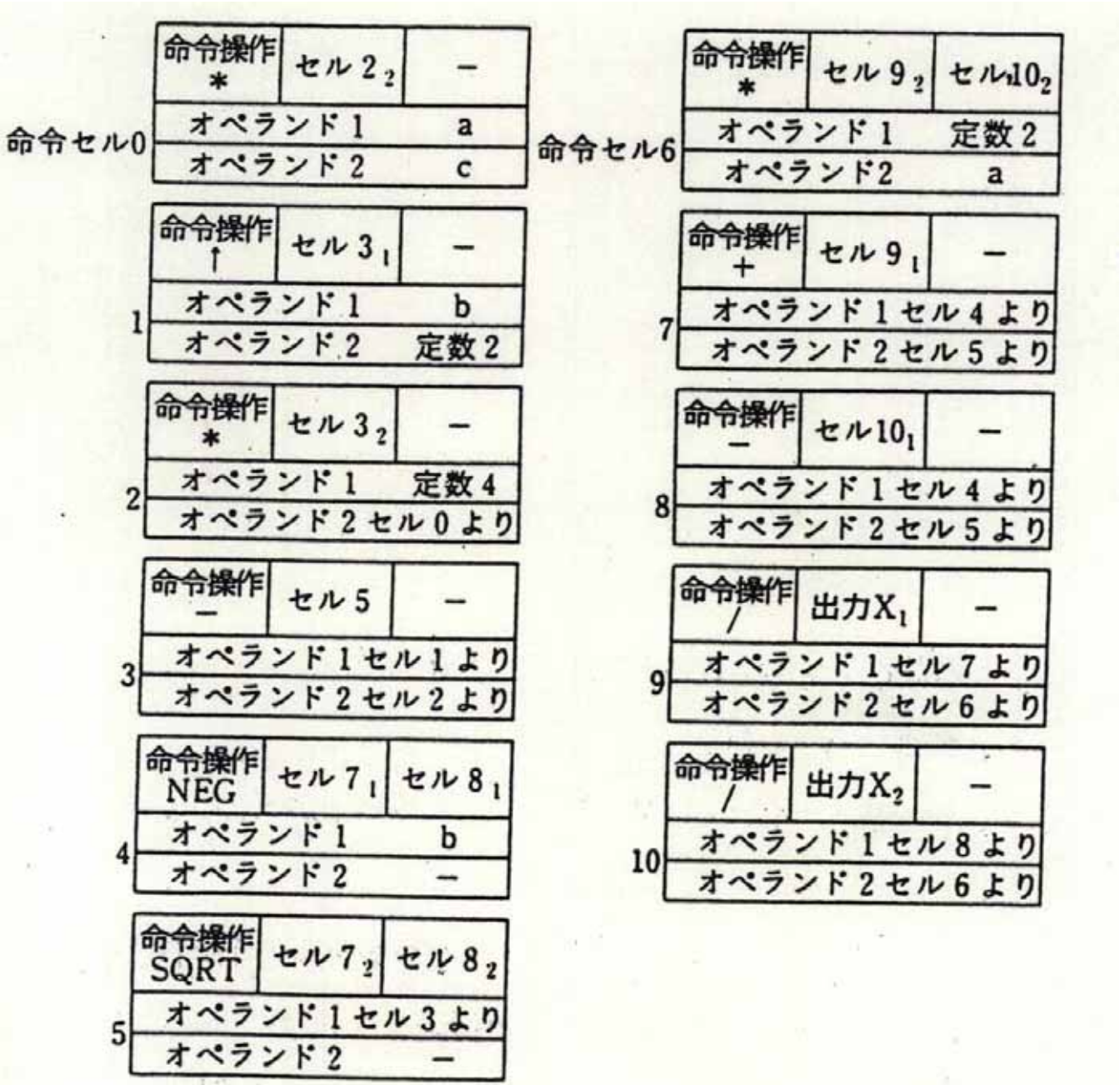
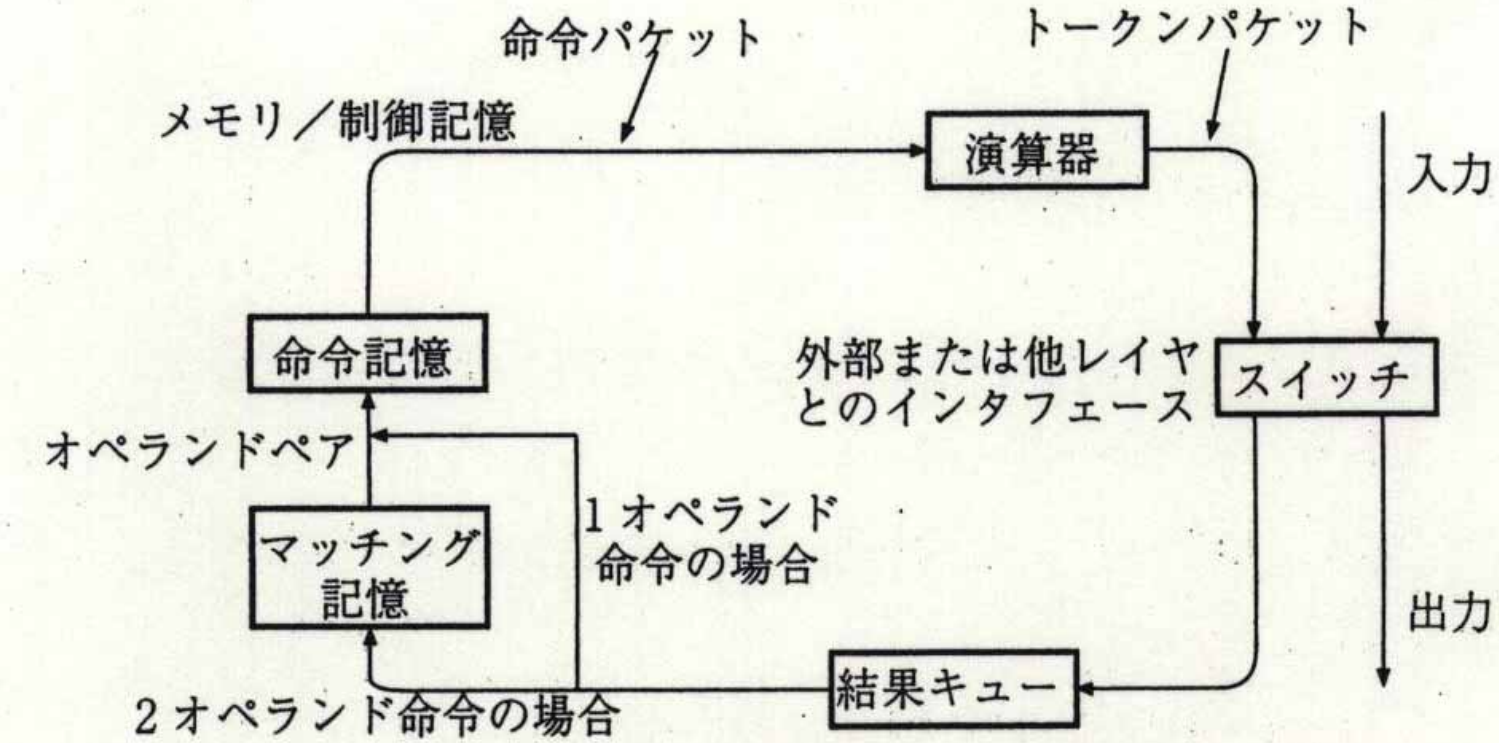
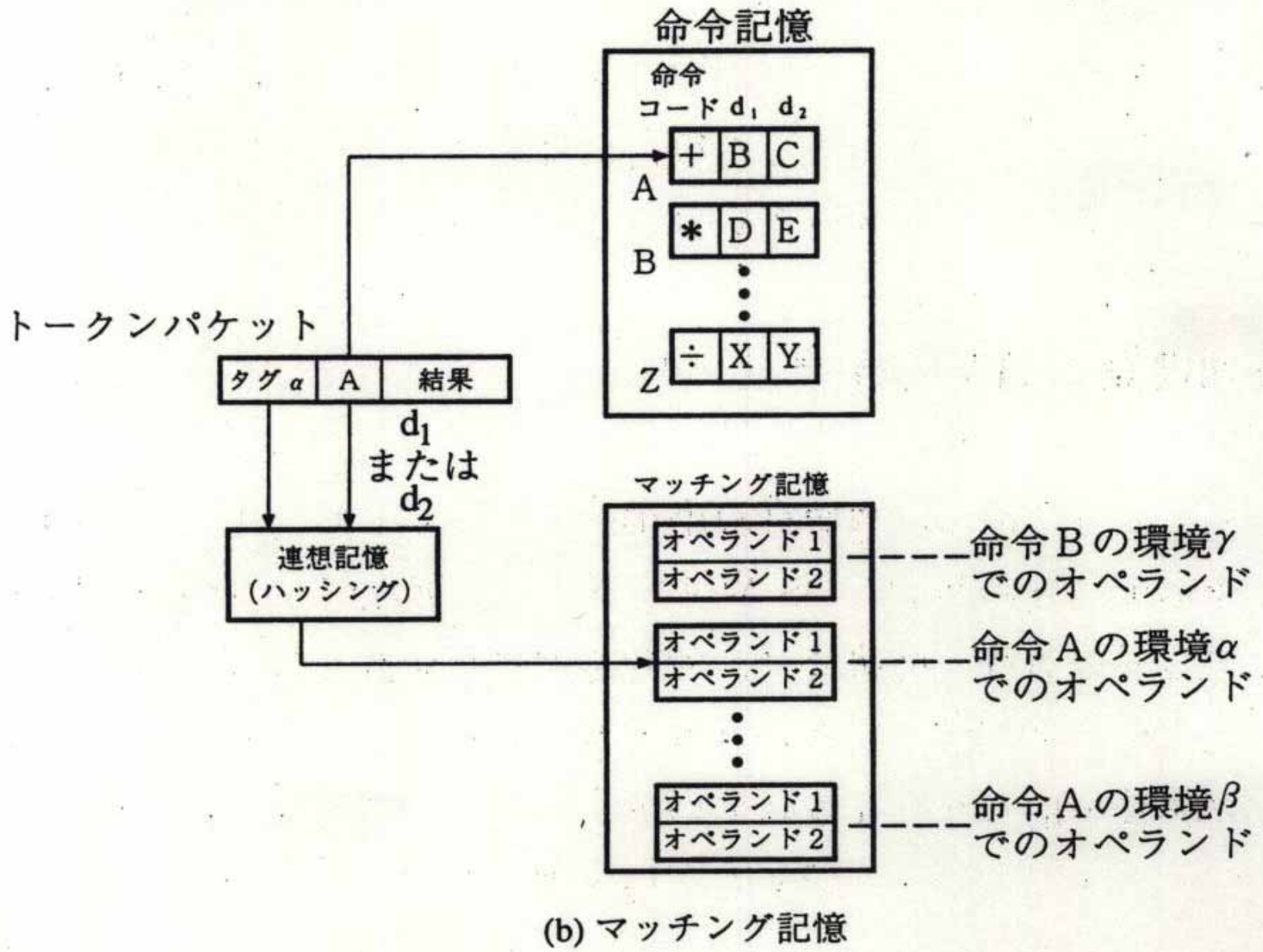


図 5. 24　2 次方程式の根を求めるデータフロープログラムに対するパケット群



(a) 循環パイプラインの構造



(b) マッチング記憶

図 5. 25　循環パイプライン型データ駆動コンピュータ

（４）データ駆動方式のマルチスレッド化

従来のデータ駆動方式の欠点

①細粒度並列処理

②メモリバンド幅

③マッチング記憶

④構造体メモリ

データ駆動方式：処理と通信を統合化

①粗粒度化

②スレッド起動・切換えの高速化

- スレッドごとの専有レジスタ
- スレッド起動の高速化
- スレッドの多重実行

５．７並列言語と
コンパイラ

５.７.１並列言語

（１）逐次言語

（２）逐次言語の並列処理拡張

①ループ並列化の記述

```
DOALL   I=1,N
A(I+1)=A(I)+1.0
END   DOALL
```

②ベクトル演算／配列演算の記述

A、B、Cがベクトルの場合、

C=A+Bの表現

③プロセス生成・終了、同期操作の指定

FORK-JOINなどのプロセスの生成や終了の明示

④データ分割の指定

ベクトルや配列に対してそのデータ分割法を明示

⑤通信文の明示的利用

Send/Receiveなどプロセス間での通信を

明示

①HPFのアプローチ

Fortranプログラム内でデータ分割をユーザで指定

所有者計算規則（owner computes rule）による目的並列プログラムを生成

SEND/RECEIVE文の挿入

メッセージ交換型マルチプロセッサを対象

Fortran-D、HPF（High Performance Fortran）、Vienna Fortranなど

②MPIのアプローチ

メッセージ通信のプロトコルの標準化

これらを利用したメッセージ交換

各プログラム（プロセス）：Ｃなど通常の逐次言語で記述

PVM、MPIやLindaなどが代表例

ネットワーク上のワークステーションにも適用

手軽な並列処理の利用形態

（３）並列処理言語

データフロー言語のような関数型言語

メッセージ交換モデルをベースとしたOccam

５．７．２ＨＰＦの概要

1992年：HPFF（High Performance Fortran Forum）

で標準化

ユーザが与えるコンパイラへの指示：

ディレクティブ

ディレクティブ：!HPF$で表示

①プロセッサ構成指定

```
!HPF$ PROCESSORS P(16)
```

!HPF$ PROCESSORS R(4,4)

「仮想」プロセッサ

②データ分割指定

配列データの分割：各次元方向に、

BLOCK**分割と**CYCLIC**分割**

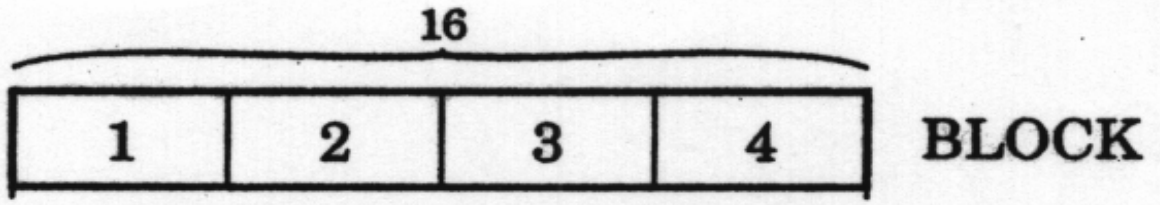
③データ相互配置指定

!HPF$ ALIGN A(I,J) with B(J,I)

B(J,I)**が(**BLOCK,BLOCK**)：データ分割、**

Ａ：Ｂの転置行列

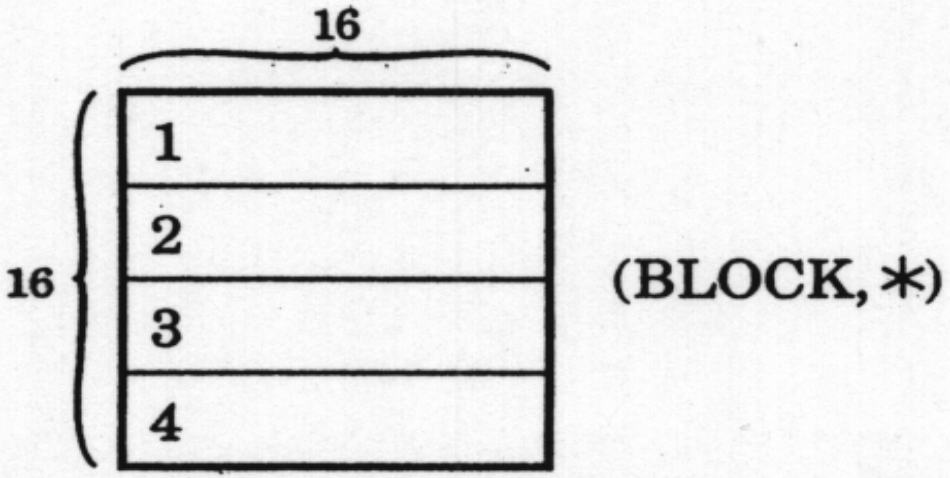
Ｂの転置元とＡへの転置先が同一



(i) !HPF$ DISTRIBUTE A(BLOCK) ONTO P(4)

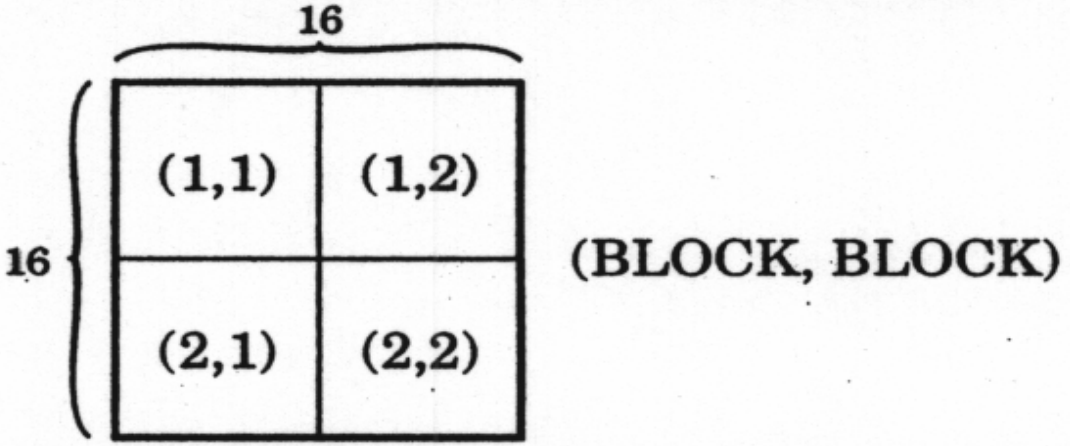
(ii) !HPF$ DISTRIBUTE A(CYCLIC) ONTO P(4)

(a) 1 次元の例

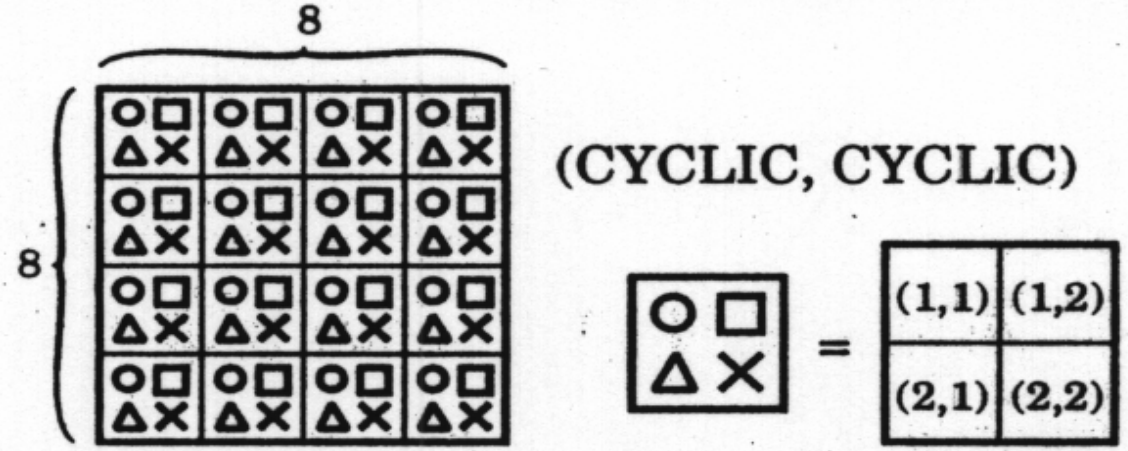


(i) !HPF$ DISTRIBUTE A(BLOCK, *) ONTO P(4)

(ii) !HPF$ DISTRIBUTE A(*, BLOCK) ONTO P(4)



(iii) !HPF$ DISTRIBUTE A(BLOCK, BLOCK) ONTO R(2,2)



(iv) !HPF$ DISTRIBUTE A(CYCLIC, CYCLIC) ONTO R(2,2)

(b) 2 次元の例

図 5. 26　データ分割の例

プロセッサに配置

!HPF$ ALIGN A(I) WITH B(I+1)

B(I+1)とA(I)が同一プロセッサに

配置される。

５．７．３並列化コンパイラ

（１）ベクトルコンパイラとの差異

ベクトル化コンパイラでは、

①データ依存関係の解析

②ループ変換

データ依存関係にサイクル構造

ベクトル化できないループ

ループ変換

ループ分割、スカラエクスパンジョン、

ループ交換、ストリップマイニング、

ウェーブフロント

③マクロ演算機構

総和演算／内積演算

サイクリックリダクション

④多重ループの高速化

マルチプロセッサとベクトルプロセッサの

根本的差異

ベクトルプロセッサ

①集中メモリ

②少数のパイプライン演算器

マルチプロセッサ

分散メモリ方式

①データの局所的配置

②処理と通信のオーバラップ実行

③一括通信

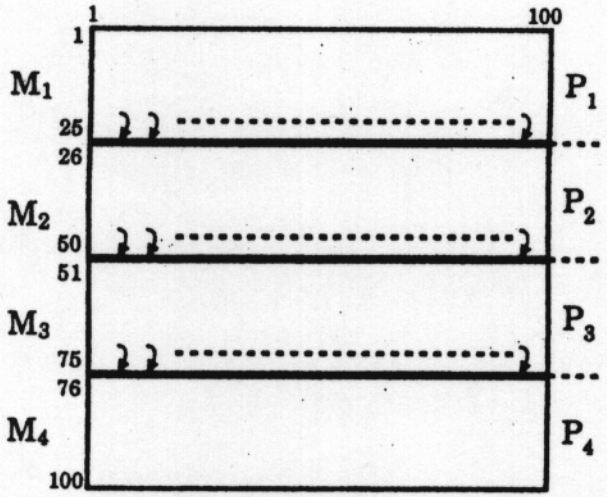
（２）データ分散配置

```
      DO 10 I=2,100
      DO 20 J=1,100
      A(I,J)=A(I-1,J)+1.0
   20 CONTINUE
   10 CONTINUE
```

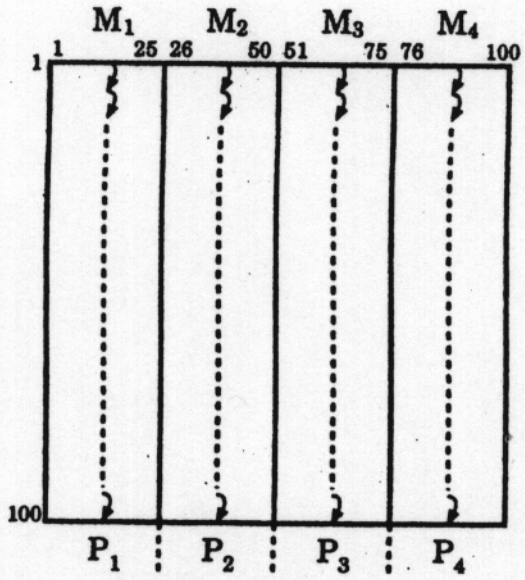
（３）所有者計算規則

プロセッサPi**のメモリ**Mi**に対する代入文（すなわち代入文での左辺のデータ格納先が**Mi**である）の計算（代入文の右辺の計算）はプロセッサ**Pi**でなされる。**

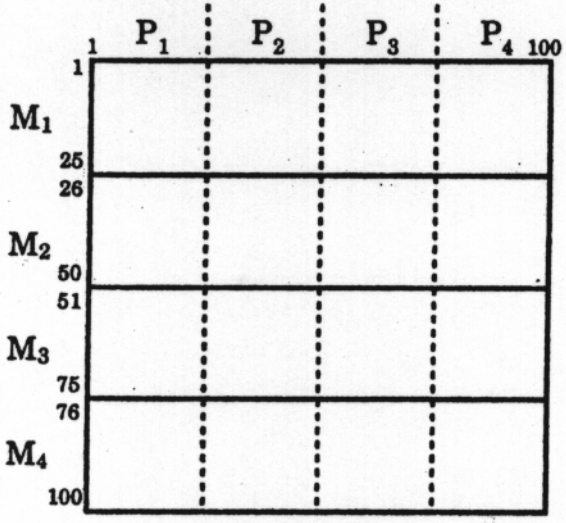
A=B+C**の場合**



(a) プロセッサ行分割/メモリ行分割



(b) プロセッサ列分割/メモリ列分割



(c) プロセッサ列分割/メモリ行分割

```
      DO 10 I=1,16
      DO 20 J=1,16
      A(I,J)=B(I,J)+C(I、J)
20 CONTINUE
10 CONTINUE
```

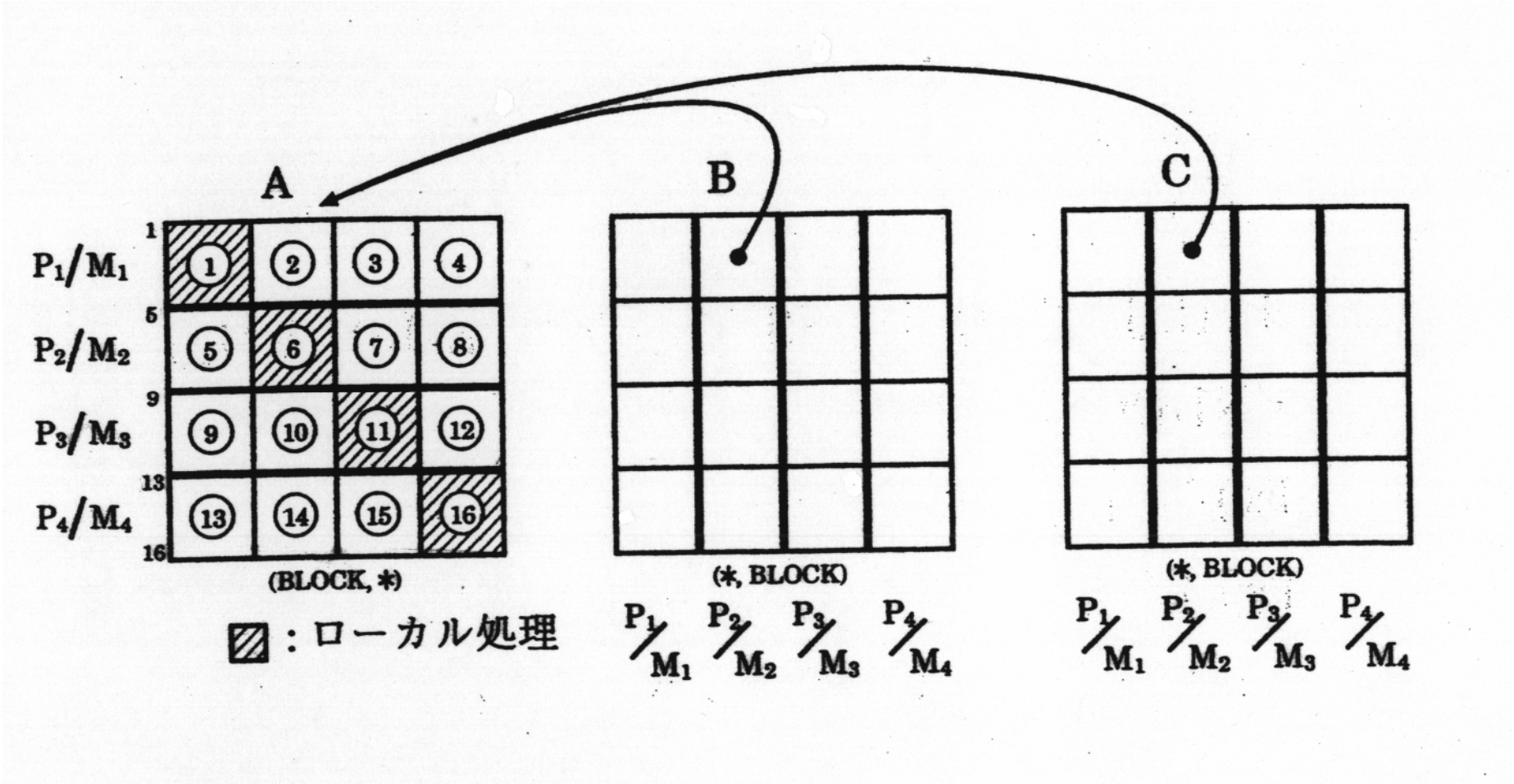
A:(BLOCK,*)、B、C:(*,BLOCK)で分割

単純な並列化コンパイラ：

プロセッサP1**のプログラム**

・B(1:16,1:4)、C(1:16,1:4)**の**P2、P3,P4**への**SEND

```
      DO J=1,4
```


A
B
C
1
5
9
13
16
P1/M1
P2/M2
P3/M3
P4/M4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
(BLOCK, *)
(*, BLOCK)
(*, BLOCK)
：ローカル処理
P1/M1
P2/M2
P3/M3
P4/M4
P1/M1
P2/M2
P3/M3
P4/M4

```
DO I=5,8
SEND P2 B(I,J)、C(I,J)
ENDDO
DO J=1,4
DO I=9,12
SEND P3 B(I,J)、C(I,J)
ENDDO
DO J=1,4
DO I=13,16
SEND P4 B(I,J)、C(I,J)
```

ENDDO

・B(1:4、5:16)、C(1:4、5:16)のP2,P3,P4からのRECEIVE

```
DO I=1,4
DO J=5,8
T1(I,J),T2(I,J)=RECEIVE P2 B(I,J)、C(I,J)
ENDDO
DO I=1,4
DO J=9,12
T1(I,J),T2(I,J)=RECEIVE P3 B(I,J)、C(I,J)
ENDDO
```

```
DO I=1,4
DO J=13,16
T1(I,J),T2(I,J)=RECEIVE P4 B(I,J)、C(I,J)
ENDDO
```

・自ﾃﾞｰﾀでの計算

```
DO I=1,4
DO J=1,4
A(I,J)=B(I,J)+C(I、J)
ENDDO
```

・RECEIVEﾃﾞｰﾀによる計算

```
DO I=1,4
DO J=5,16
A(I,J)=T1(I,J)+T2(I、J)
ENDDO
```

（４）通信の最適化

①通信ベクトル化（message vectorization）

```
DOALL
A(I,J)=A(I-1,J)+(I+1,J)+
        A(I,J-1)+A(I,J+1)-4*A(I,J)
END DOALL
```

②通信合体法（message coalescing）

③通信集団化法（message aggregation）

④通信と処理のオーバラップ

⑤検査官／執行官方式

たとえば、

A(I)=B(b(I))+C(I)

収束演算の場合：

b(I)**がしばらくの間変更されないで使用**

される場合

検査官の処理は１度でよく、毎回データ

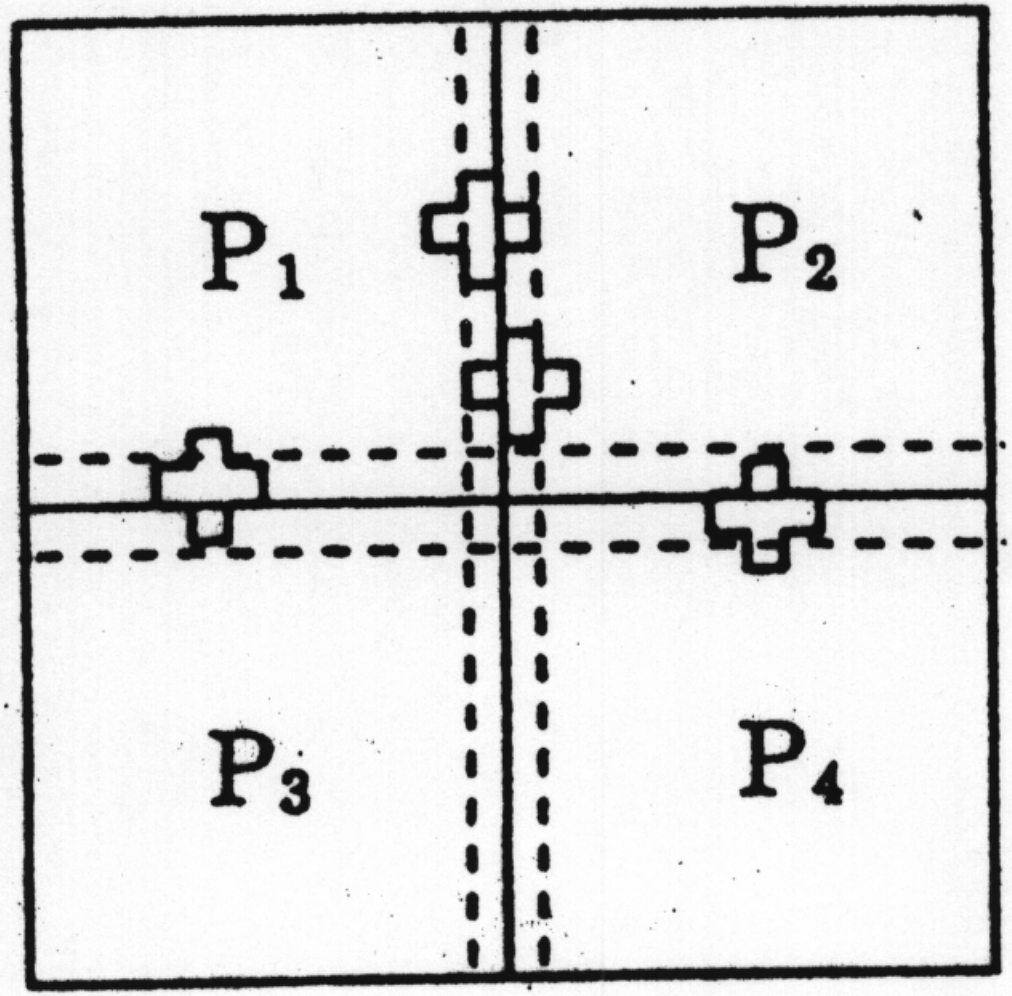
要求側からリード要求を発行する方式
より効率よい

（５）今後の課題

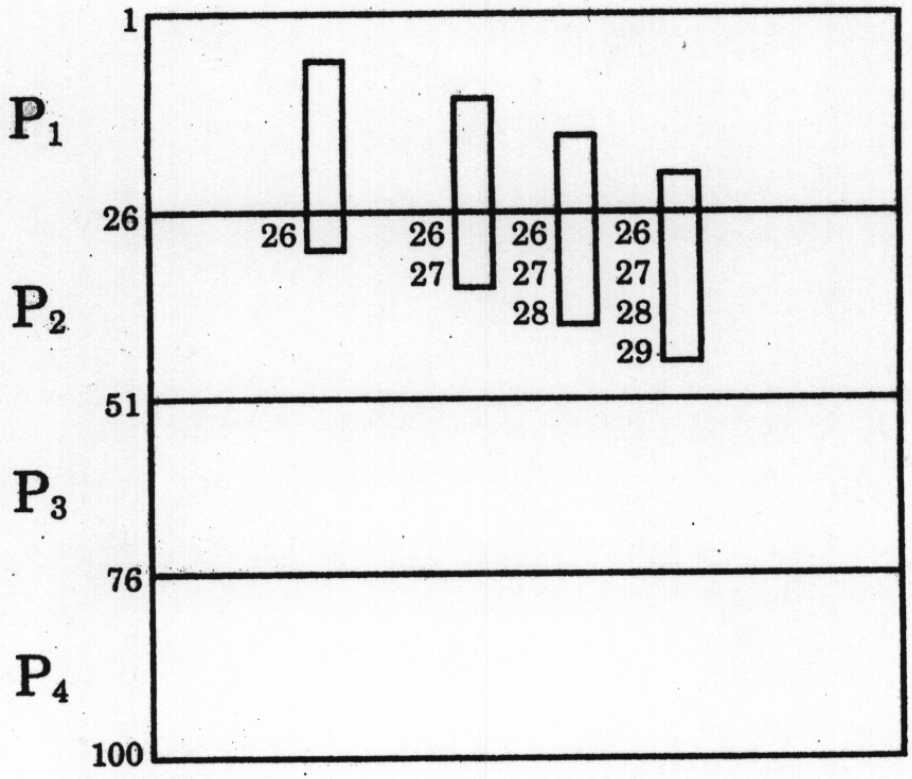
並列化コンパイラ：研究初期段階

人工知能の格好の例題

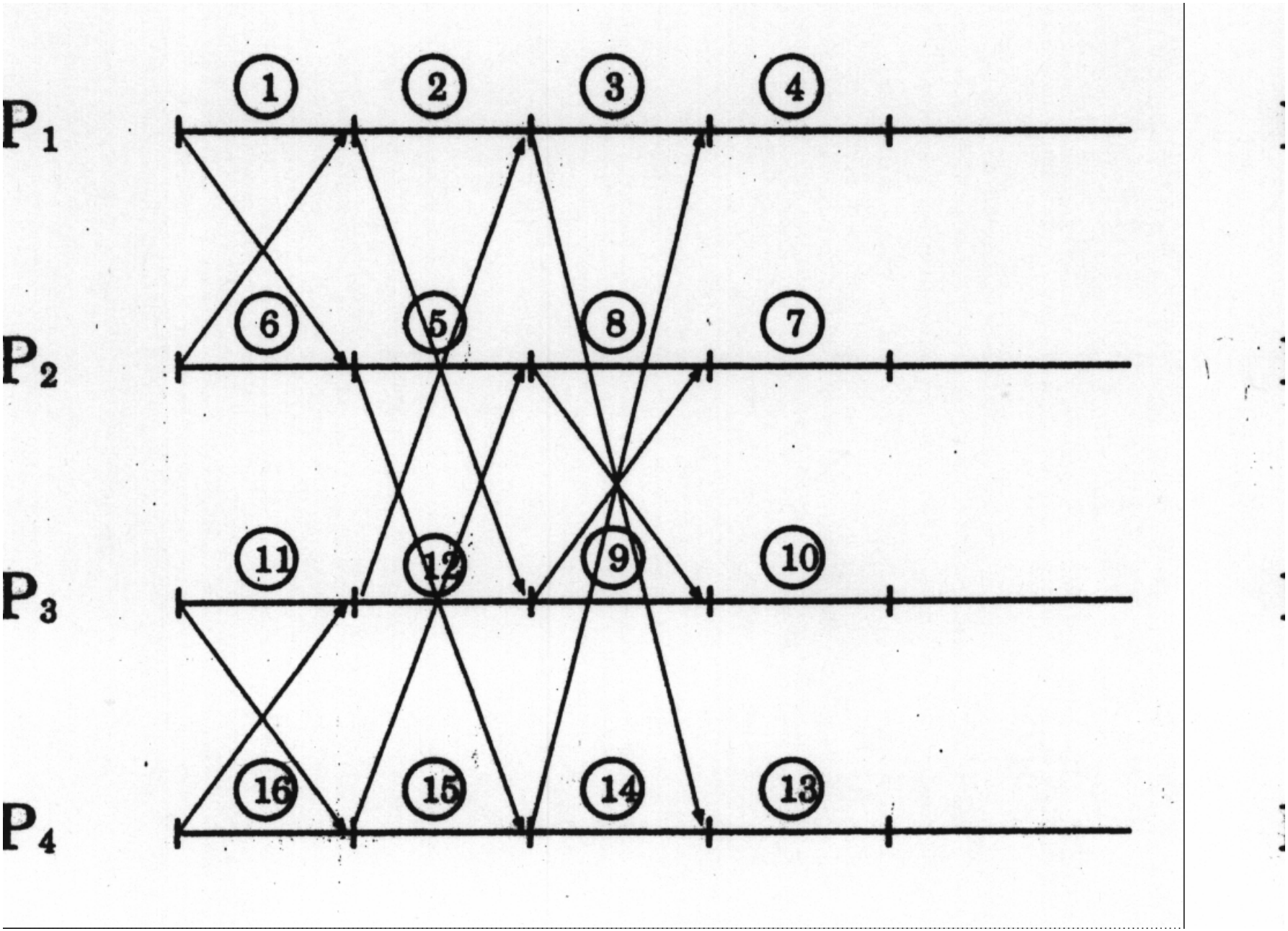
データベースの構築などによる知識の蓄積と例題による解法

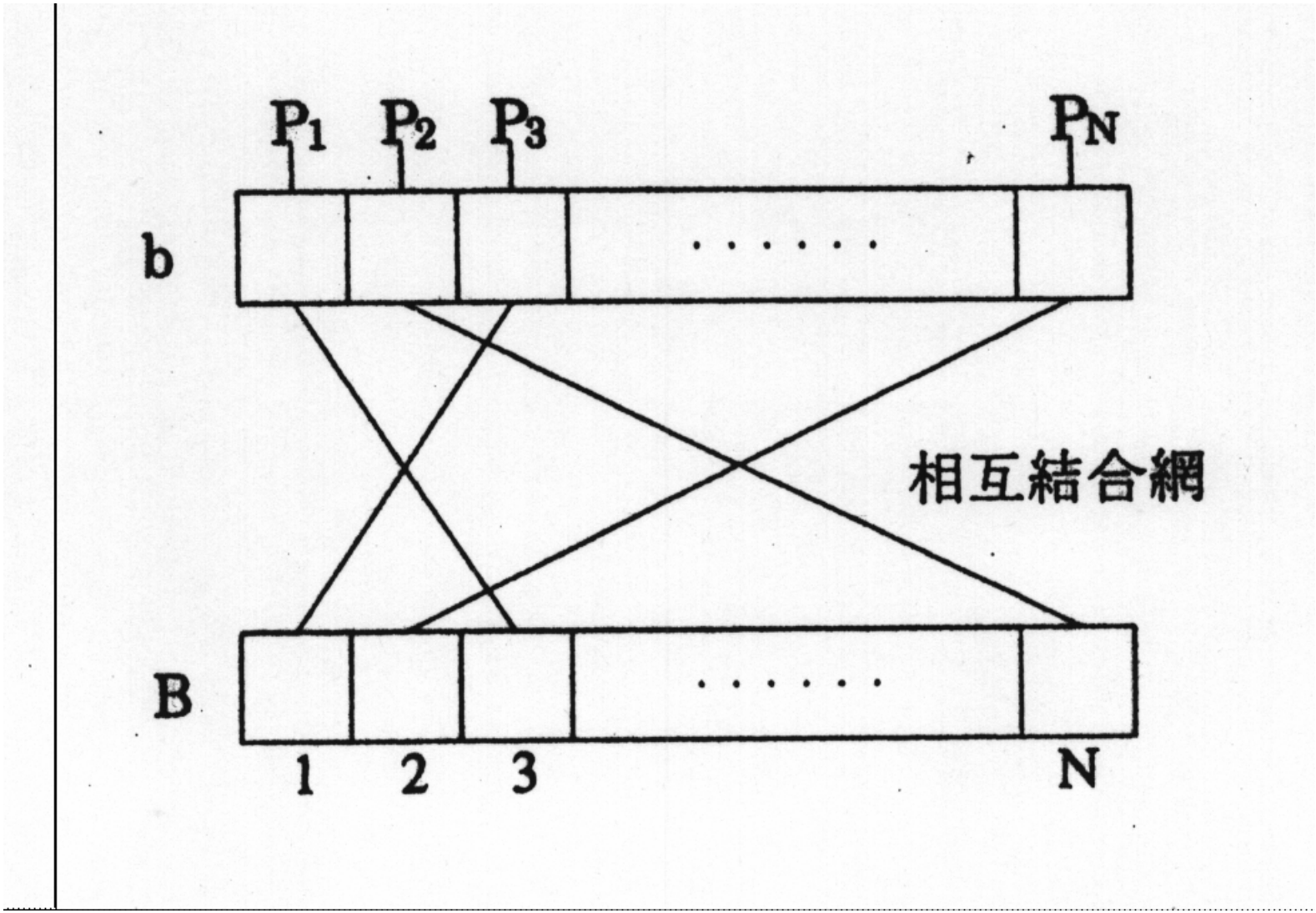


：ラプラス演算




```
      DO  10  J = 1 , N
      DO  20   I = 1 , N-4
      A ( I , J ) = B ( I+1 , J ) + B ( I+2 , J ) +
                    B ( I+3 , J ) + B ( I+4 , J )
   20 CONTINUE
   10 CONTINUE
```


P1
P2
P3
P4
1
2
3
4
6
5
8
7
11
12
9
10
16
15
14
13


P1
P2
P3
PN
b
相互結合網
B
1
2
3
N

５．８マルチプロセッサのその他の課題

マルチプロセッサのスケジューリング、
負荷分散、
デバッグなどの研究課題