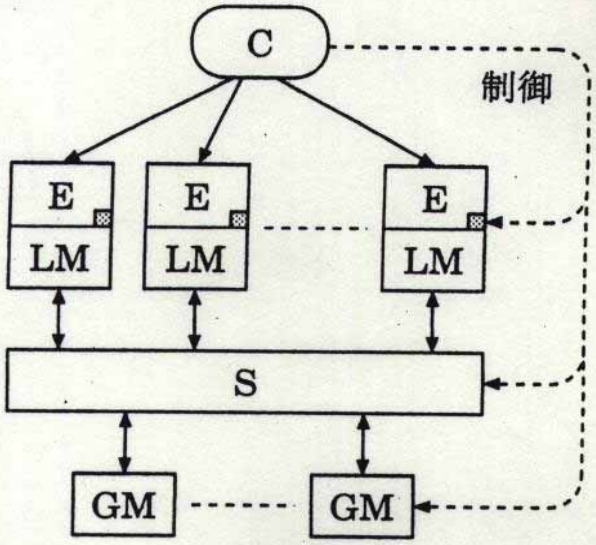
第7章　並列処理

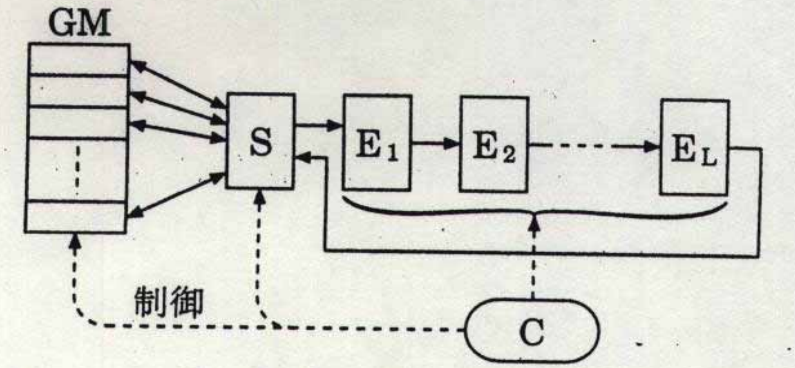
7.1　並列処理の分類

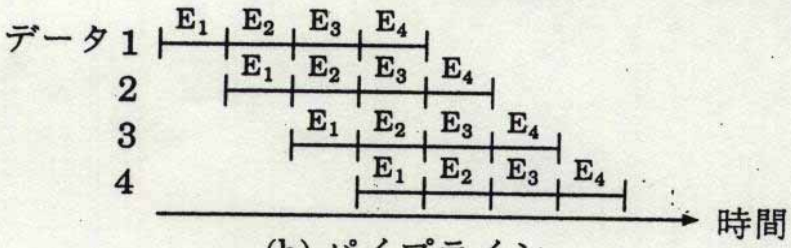
SIMD：Single Instruction
　　Multiple　Data　Stream
　①空間並列型（狭義SIMD）
　②時間並列型（パイプライン）

MIMD：Multiple　Instruction
　　Multiple　Data　Stream
　③マルチプロセッサ
　④マルチコンピュータ

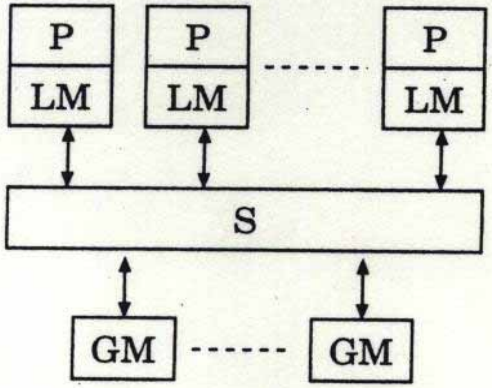


(a) SIMD

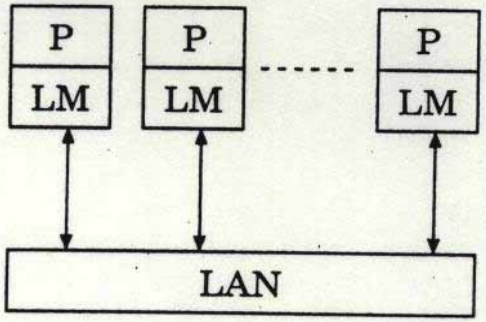




(b) パイプライン



(c) マルチプロセッサ

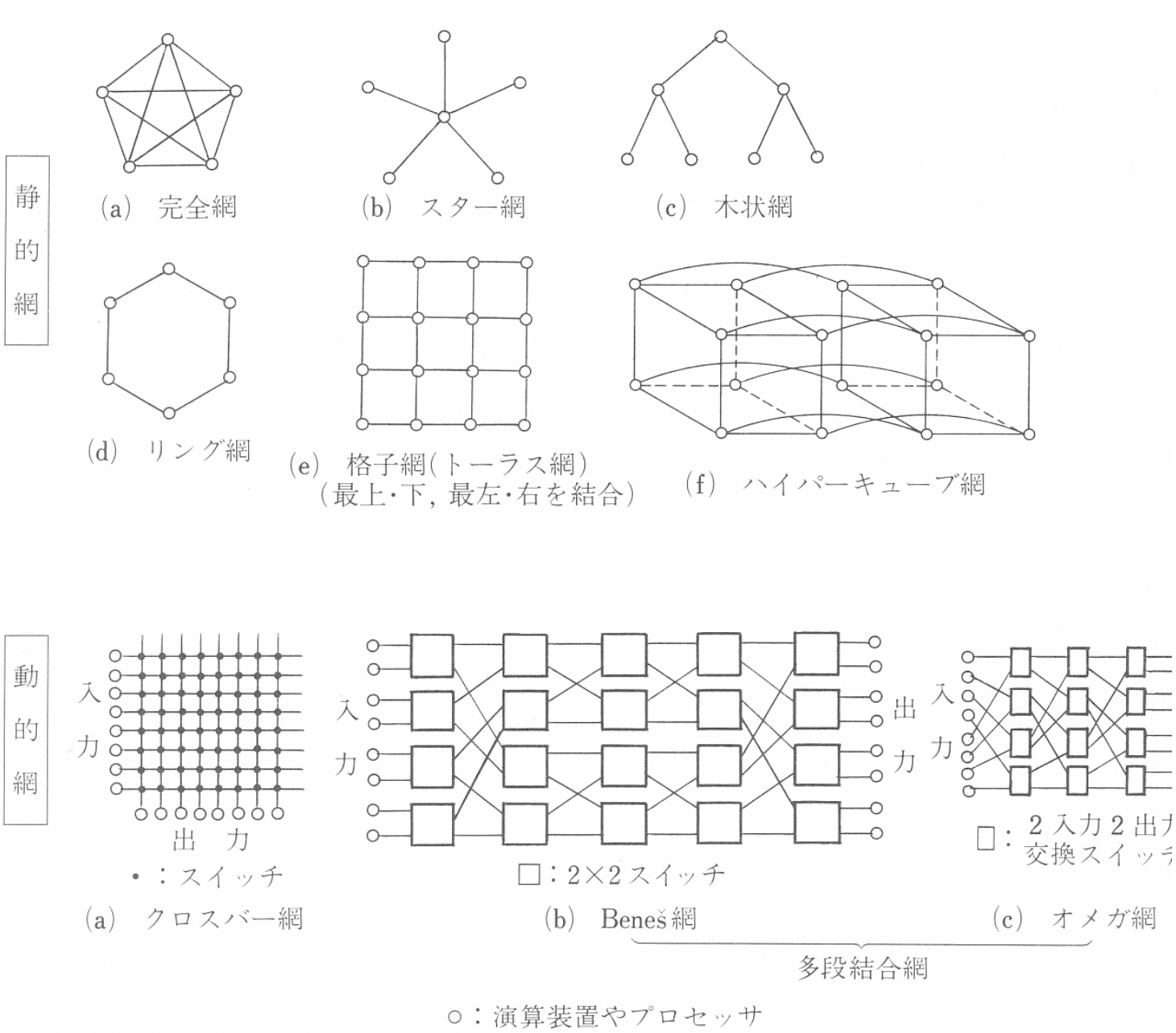


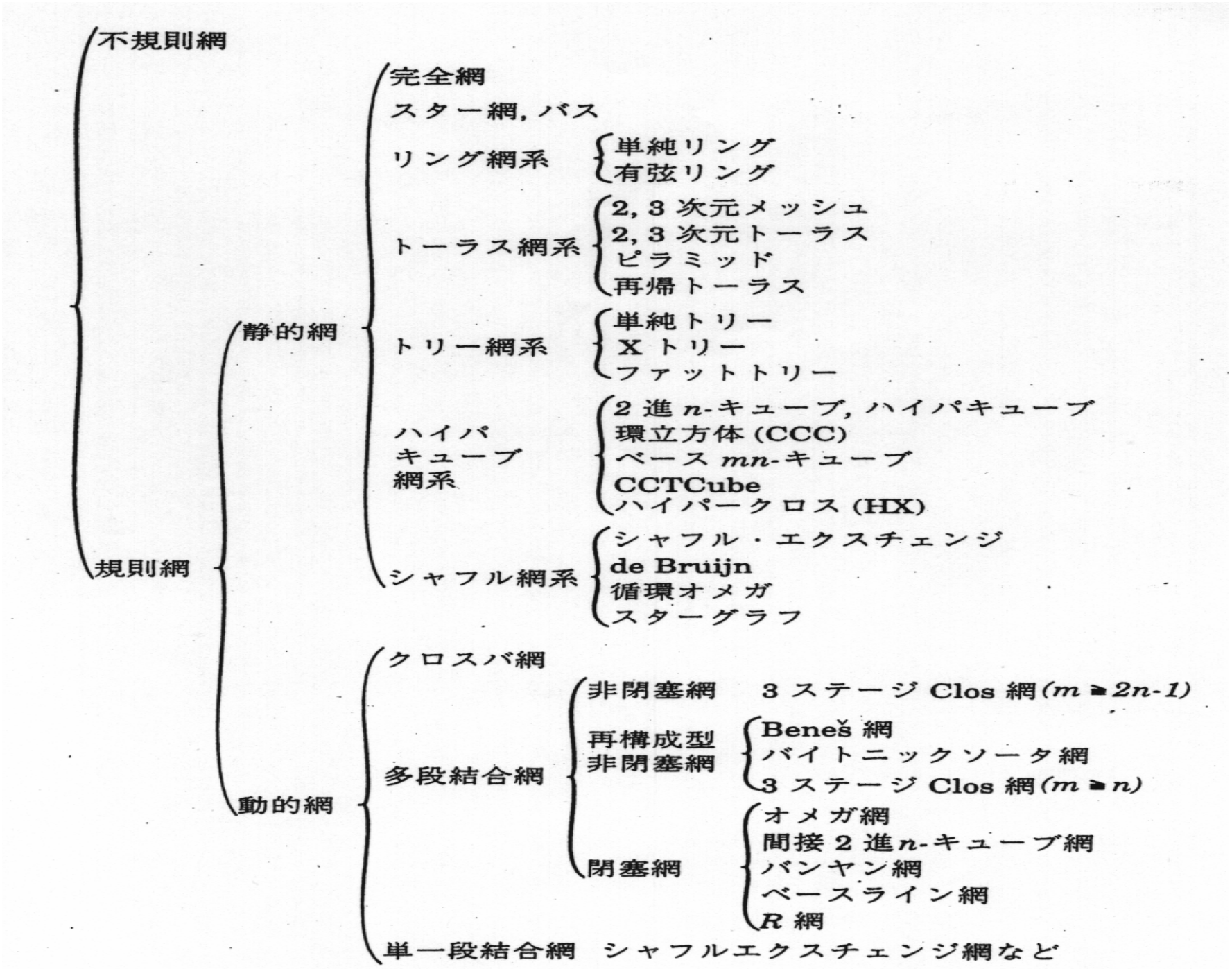
(d) マルチコンピュータ
(クラスタコンピュータ)

C：制御装置
E：演算装置
LM：ローカルメモリ
▩：抑止フラグ

GM：グローバルメモリ
P：プロセッサ
S：相互結合網

7．2相互結合網


静的網
(a) 完全網
(b) スター網
(c) 木状網
(d) リング網
(e) 格子網(トーラス網)
(最上・下，最左・右を結合)
(f) ハイパーキューブ網
動的網
入力
出力
・：スイッチ
(a) クロスバー網
□：2×2スイッチ
(b) Benes̆網
□：2入力2出力
交換スイッチ
(c) オメガ網
多段結合網
○：演算装置やプロセッサ


不規則網
規則網
静的網
完全網
スター網, バス
リング網系
単純リング
有弦リング
トーラス網系
2, 3 次元メッシュ
2, 3 次元トーラス
ピラミッド
再帰トーラス
トリー網系
単純トリー
X トリー
ファットトリー
ハイパ
キューブ
網系
2 進 n-キューブ, ハイパキューブ
環立方体 (CCC)
ベース mn-キューブ
CCTCube
ハイパークロス (HX)
シャフル網系
シャフル・エクスチェンジ
de Bruijn
循環オメガ
スターグラフ
動的網
クロスバ網
多段結合網
非閉塞網
3 ステージ Clos 網 (m ≥ 2n-1)
再構成型
非閉塞網
Beneš 網
バイトニックソータ網
3 ステージ Clos 網 (m ≥ n)
閉塞網
オメガ網
間接 2 進 n-キューブ網
バンヤン網
ベースライン網
R 網
単一段結合網
シャフルエクスチェンジ網など

7.2.1評価項目

⑴ **距離** (distance)
⑵ **次数** (degree)
⑶ **総スイッチ数／総リンク数**
⑷ **拡張性** (scalability)
⑸ **3次元実装の容易性**
⑹ **耐故障性** (fault tolerance)
⑺ **多様な網の埋込み能力** (embedability)
⑻ **容易な経路選択** (**ルーティング**)

7.2.2静的網
ハイパキューブ

$(a_n, a_{n-1}, , , , a_2, a_1)$

$(a_n, a_{n-1}, , , , a_2, \overline{a}_1)$

$(a_n, a_{n-1}, , , , \overline{a}_2, a_1)$

$(a_n, \overline{a}_{n-1}, , , , a_2, a_1)$

$(\overline{a}_n, a_{n-1}, , , , a_2, a_1)$

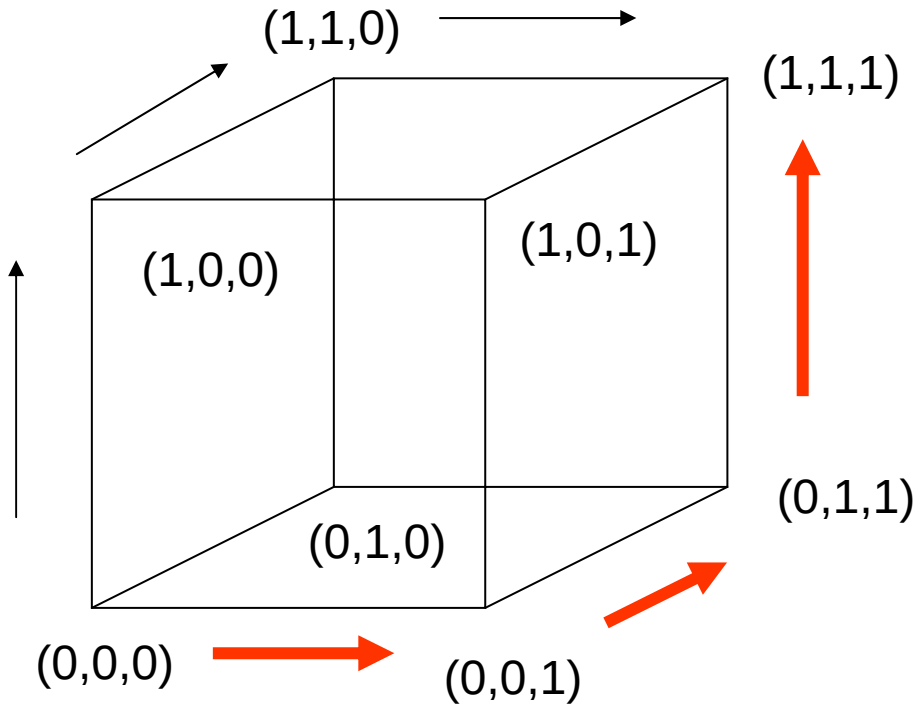
ハミング距離：1

(001)　　(110)

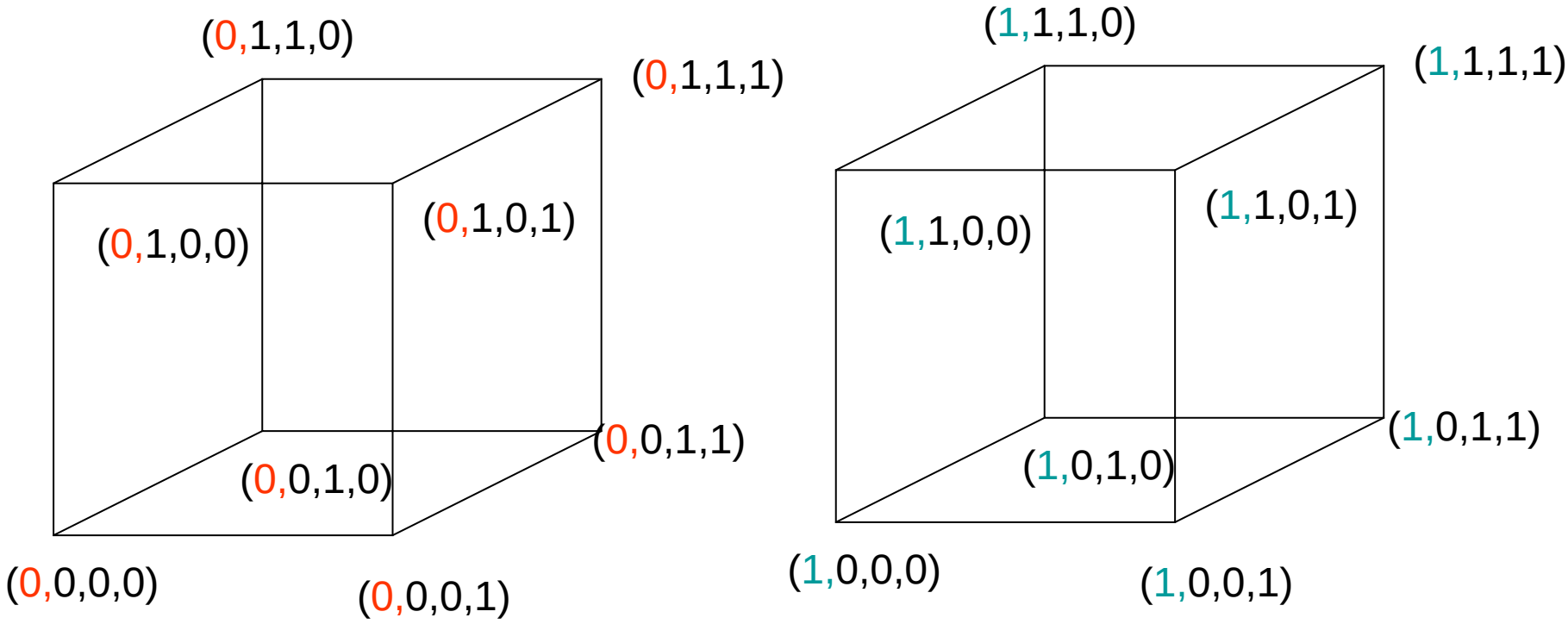
ハミング距離：3

ハミング距離：各桁の比較。異なる桁の数

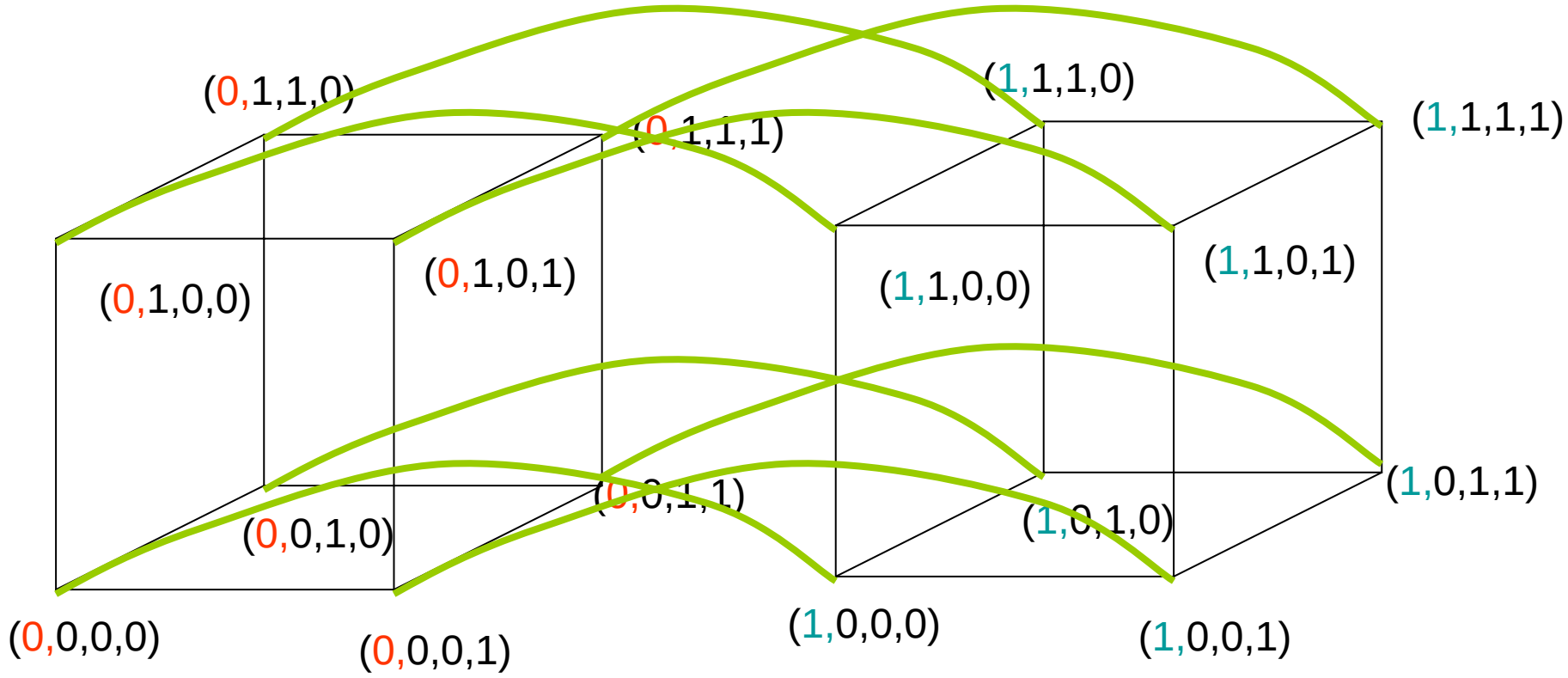
①ルーティング
下位ビットから宛先に合わせる



②拡張性


(0,1,1,0)
(0,1,1,1)
(0,1,0,0)
(0,1,0,1)
(0,0,1,1)
(0,0,1,0)
(0,0,0,0)
(0,0,0,1)
(1,1,1,0)
(1,1,1,1)
(1,1,0,0)
(1,1,0,1)
(1,0,1,1)
(1,0,1,0)
(1,0,0,0)
(1,0,0,1)

拡張
3キューブから4キューブへ



③埋め込み能力

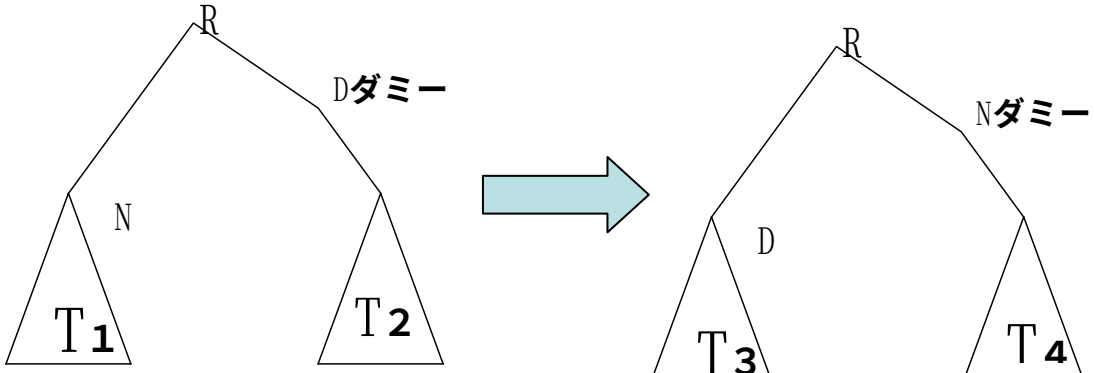
- リング
  1次元の反射2進符号
- トーラス
  2次元の反射2進符号
- トリー
  1つのダミーノードを許すと可能

反射2進符号
符号間のハミング距離＝1

	通常2進		反射2進	
	000		000	000
1	001	1	001	001
2	010	1	011	011
1	011	1	010	010
3	100	1	110	110
1	101	1	111	111
2	110	1	101	101
1	111	1	100	100

トリーの埋め込み

証明



n=k

$R:(a_n a_{n-1}...a_N...a_D...a_1)$

$D:(a_n a_{n-1}...a_N...\bar{a}_D...a_1)$で始まるダミー＋T2

$N:(a_n a_{n-1}...\bar{a}_N...a_D...a_1)$で始まるT1

$(a_n a_{n-1}...a_N...\bar{a}_D...a_1)$で始まるT3

$(a_n a_{n-1}...\bar{a}_N...a_D...a_1)$で始まるダミー＋T4

が存在する。

$R:(a_n a_{n-1}...a_N...a_D...a_1)$

$R':(a_n a_{n-1}...a_D...a_N...a_1)$

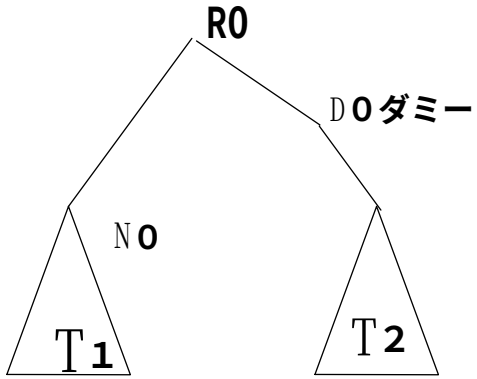
RとR'は1対１対応

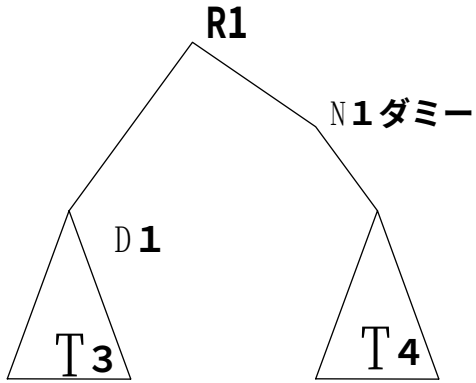
$(a_n a_{n-1}...\bar{a}_N...a_D...a_1),(a_n a_{n-1}...a_N...\bar{a}_D...a_1)$から始まるトリーに重なりがなければ

$(a_n a_{n-1}...\bar{a}_D...a_N...a_1),(a_n a_{n-1}...a_D...\bar{a}_N...a_1)$から始まるトリーに重なりがない。

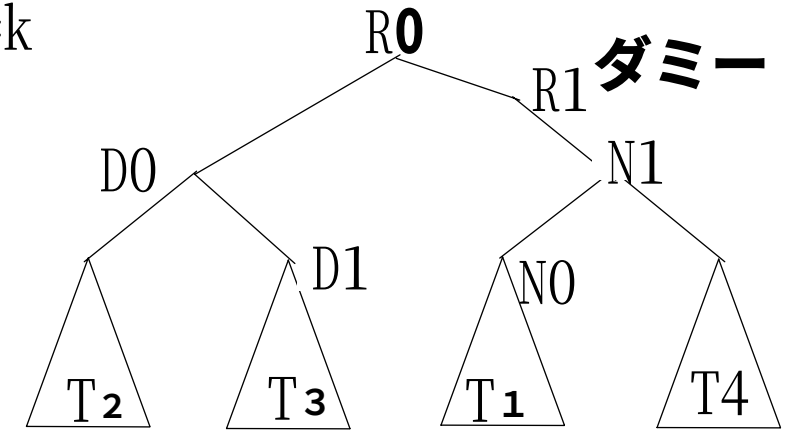
トリーの埋め込み

証明





n=k



n=k+1

7.2.3動的網

①基本通信パターン

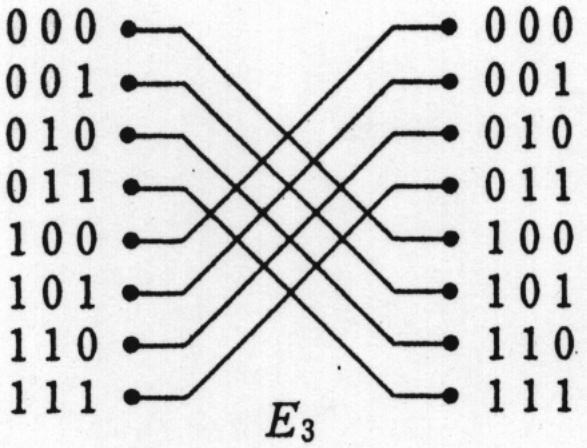
送信ノード番号Xの2進表示:$(a_n,\cdots,a_2,a_1)$

受信ノード番号Yの2進表示:$(b_n,\cdots,b_2,b_1)$

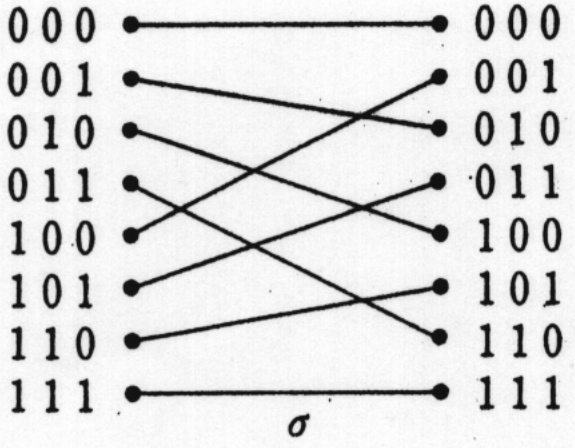
通信パターンの数（受信ノードにダブリなし）:N!

送受信ノードの対応関係:ひとつの置換

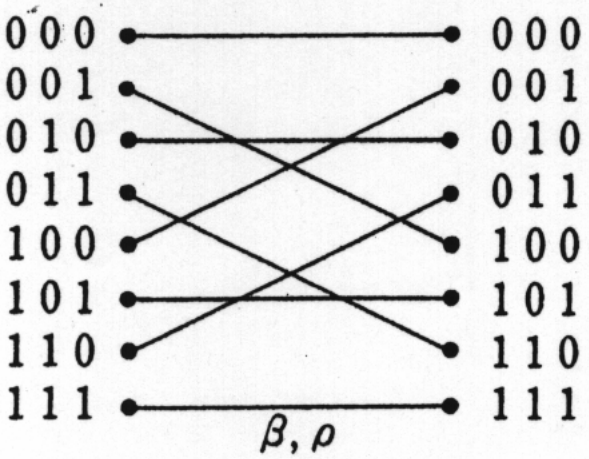
送信		受信
(1, 2, 3, 4)	＜－＞	(2, 3, 4, 1)



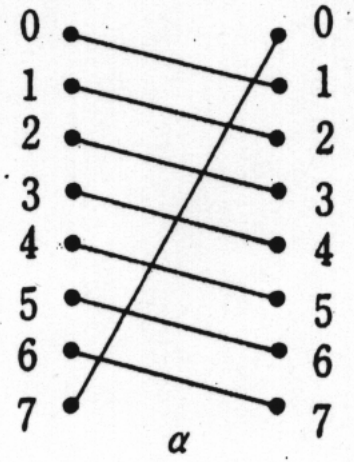
（a） エクスチェンジ置換



（b） シャフル置換



（c） バタフライ置換と
ビット逆転置換



（d） シフト置換

(1) エクスチェンジ置換

$y=E_i(x)=(a_n,\ldots,\sim a_i,\ldots,a_1)$

～aiはaiの否定を表す。

(2) シャフル置換

$y=\sigma(x)=(a_{n-1},a_{n-2},\ldots,a_1,a_n)$

k－サブシャフル σ_k 、

k－スーパシャフル σ^k

$y=\sigma_k(x)=(a_n,\ldots,a_{k+1},a_{k-1},\ldots,a_1,a_k)$

$y=\sigma^k(x)=(a_{n-1},\ldots,a_{n-k+1},a_n,a_{n-k},\ldots,a_1)$

(3) バタフライ置換

$y=\beta(x)=(a_1,a_{n-1},\ldots,a_2,a_n)$

k－サブバタフライ β_k

k－スーパバタフライ β^k

$y=\beta_k(x)=(a_n,\ldots,a_{k+1},a_1,a_{k-}1,\ldots,a_2,a_k)$

$y=\beta^k(x)=(a_{n-k+1},a_{n-1},\ldots,a_{n-k+2},a_n,a_{n-k},.,a_1)$

(4) ビット逆転置換

$y=\rho(x)=(a_1,a_2,\ldots,a_n)$

kーサブビット逆転置換　ρ_k、

kースーパビット逆転置換 ρ^k

$y=\rho_k(x)=(a_n,\ldots,a_{k+1},a_1,a_2,\ldots,a_{k-1},a_k)$

$y=\rho^k(x)=(a_{n-k+1},a_{n-k+2},\ldots,a_{n-1},a_n,a_{n-k},.,a_1)$

(5) シフト置換

$y=\alpha(x)=X+1 \mod N$

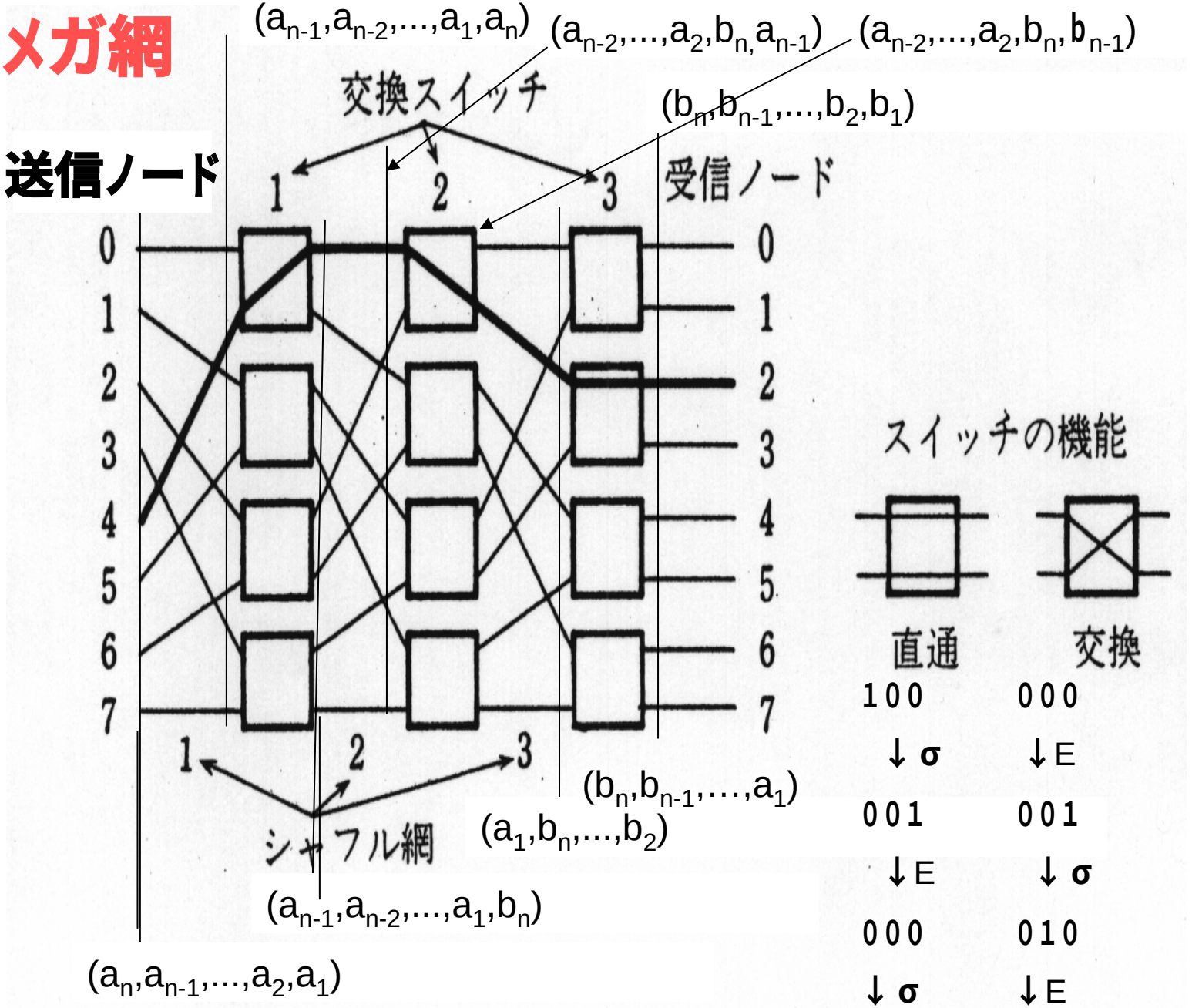
動的網の分類

①完全非閉塞網

②再構成型非閉塞網

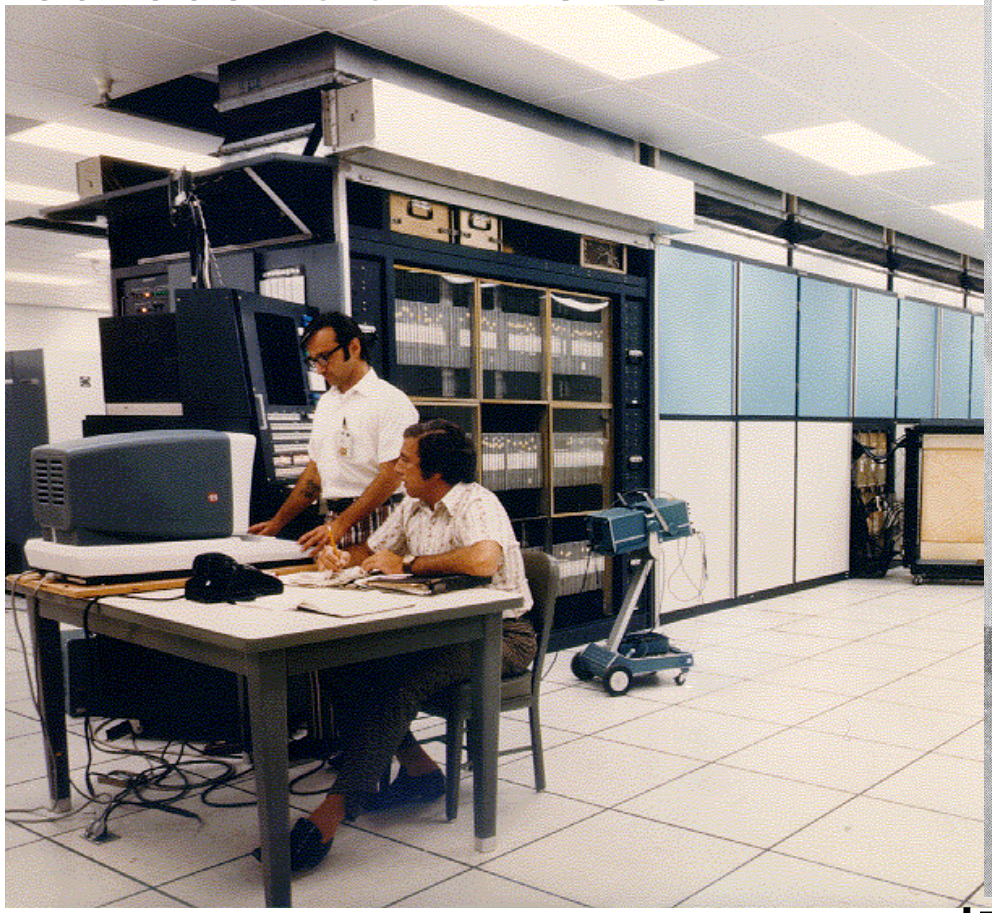
③閉塞網

②オメガ網


$(a_{n-1},a_{n-2},\ldots,a_1,a_n)$
$(a_{n-2},\ldots,a_2,b_n,a_{n-1})$
$(a_{n-2},\ldots,a_2,b_n,b_{n-1})$
$(b_n,b_{n-1},\ldots,b_2,b_1)$
交換スイッチ
送信ノード
受信ノード
1
2
3
0
1
2
3
4
5
6
7
$(b_n,b_{n-1},\ldots,a_1)$
$(a_1,b_n,\ldots,b_2)$
シャフル網
$(a_{n-1},a_{n-2},\ldots,a_1,b_n)$
$(a_n,a_{n-1},\ldots,a_2,a_1)$
スイッチの機能
直通
交換
100
↓σ
001
↓E
000
↓σ
010
000
↓E
001
↓σ
010
↓E
010

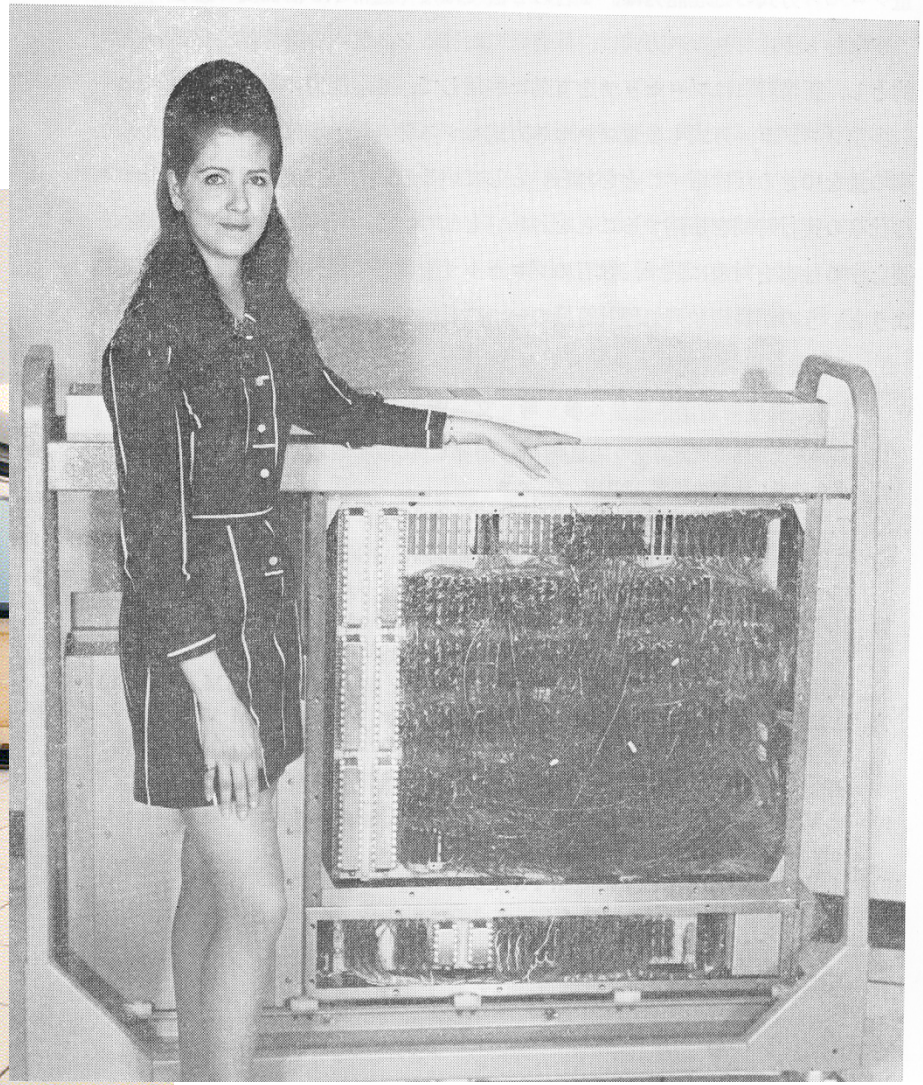
7．3　SIMD方式

パイオニア：ILLIAC　Ⅳ
１９７４　マシンサイクル
80nsec　５０MFLOPS

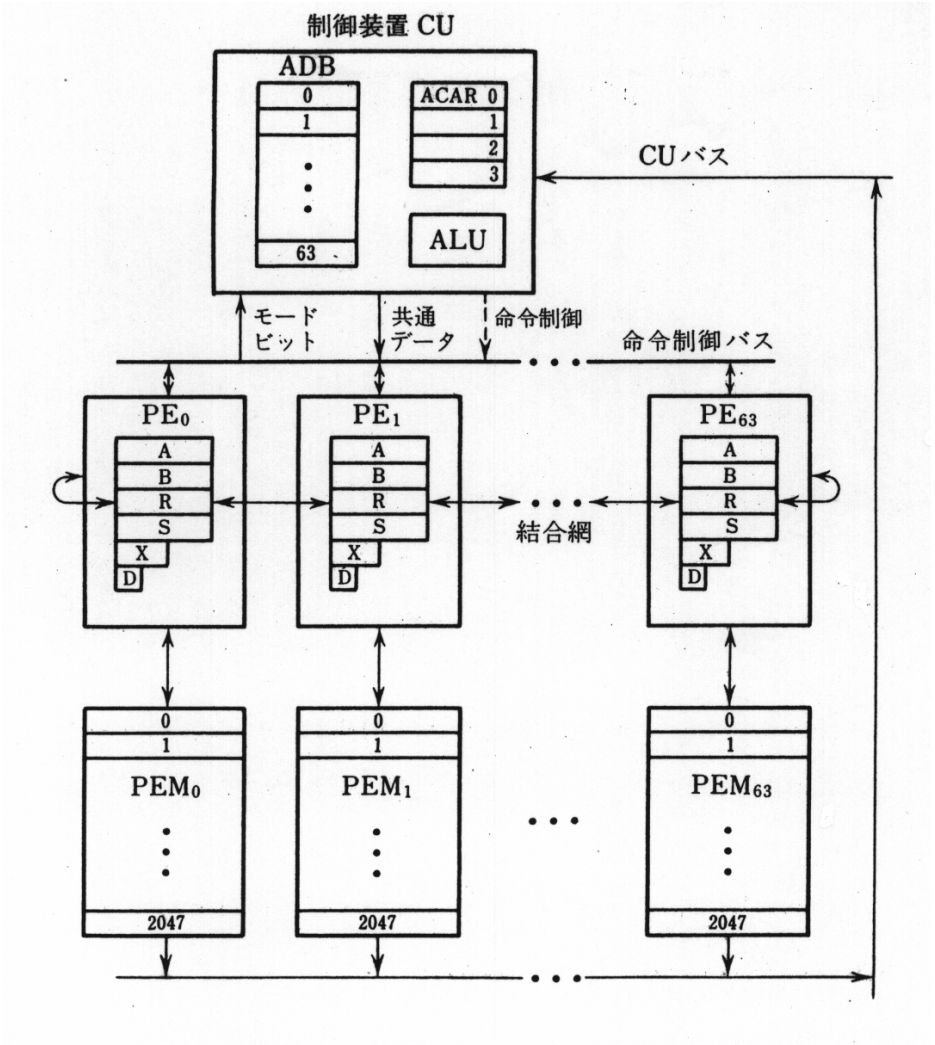


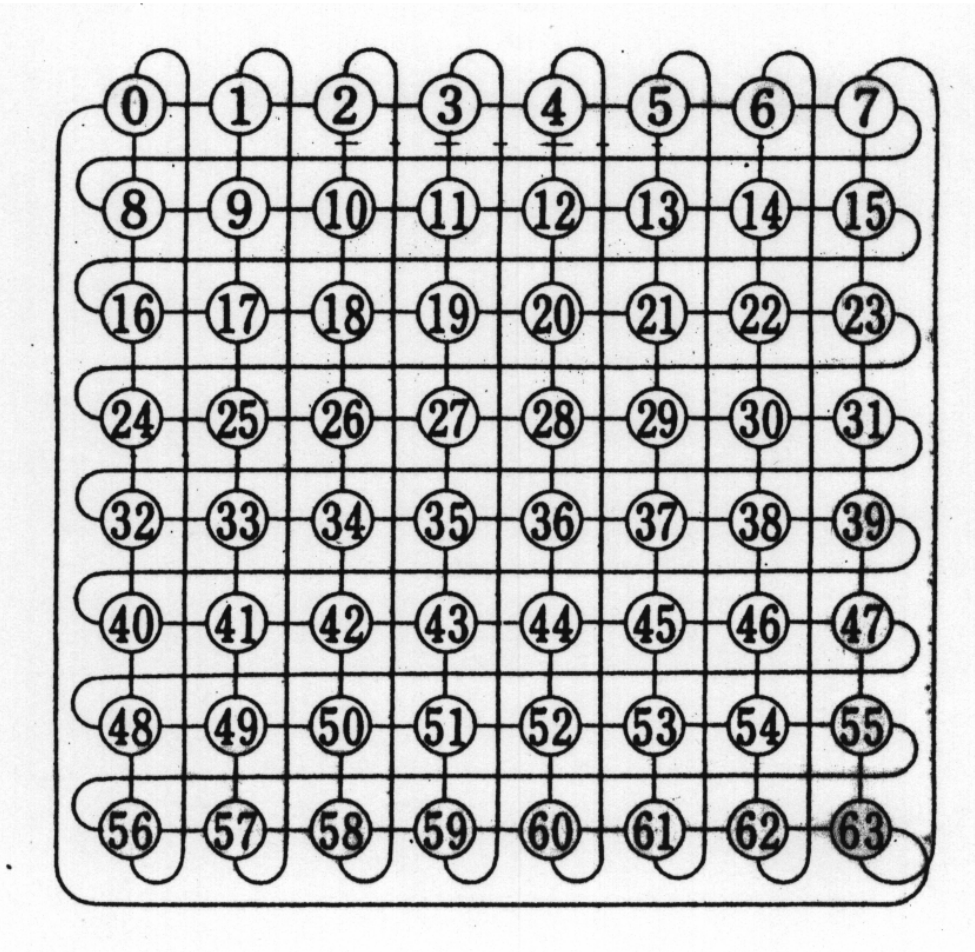
ILLIAC Ⅳ
処理装置

アレイ内の64個の処理装置はすべて同じ機能をもち，同一の構成になっている．処理装置は，演算装置・記憶装置・メモリロジックユニットなどからなり，高さ41.5インチ，長さ54インチ，幅7インチ，重さ約200ポンドである．故障が生じたときは，アレイから処理装置をはずして運搬するための専用の台車が用意されている．(p. 94, 114 参照)



加藤、苗村：並列処理計算機、オーム社、１９７６


制御装置 CU
ADB
0
1
63
ACAR 0
1
2
3
ALU
CUバス
モード
ビット
共通
データ
命令制御
命令制御バス
PE0
PE1
PE63
A
B
R
S
X
D
結合網
0
1
PEM0
PEM1
PEM63
2047


0 1 2 3 4 5 6 7
8 9 10 11 12 13 14 15
16 17 18 19 20 21 22 23
24 25 26 27 28 29 30 31
32 33 34 35 36 37 38 39
40 41 42 43 44 45 46 47
48 49 50 51 52 53 54 55
56 57 58 59 60 61 62 63

疑似トーラス網

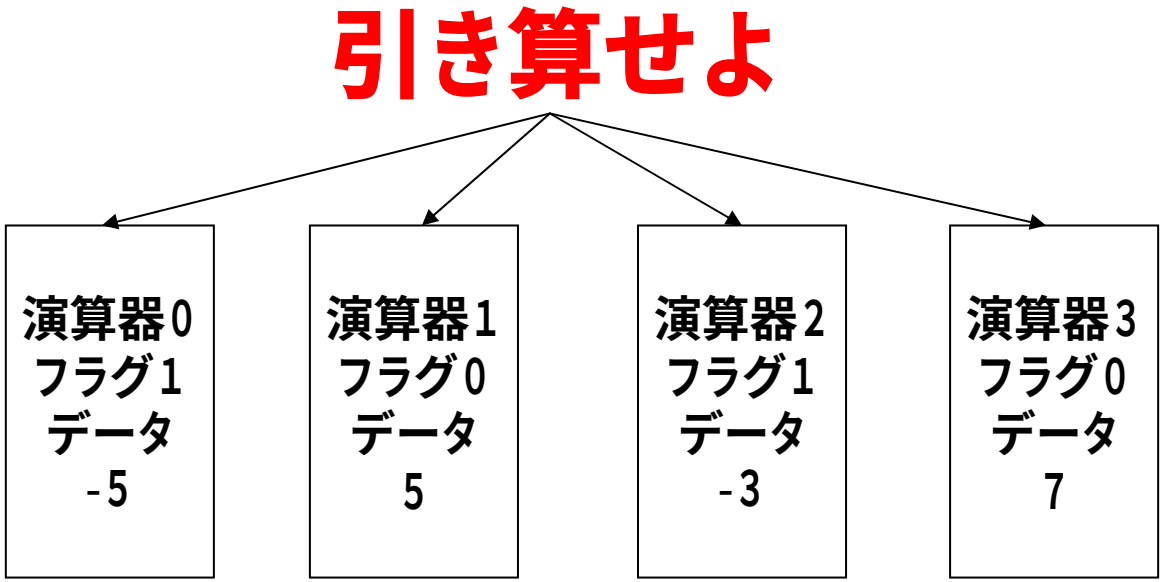
7.3.1 SIMD基本方式

SIMD演算の例：絶対値の計算

正のデータを持っているもの　抑止フラグ0

負のデータを持っているもの　抑止フラグ1

0ー(データ)の一斉実行



SIMD方式の特徴

①演算装置の簡略化

②細粒度並列処理指向

③定型処理指向

　分岐少ない方がよい

④命令同期型プログラミング

　全PE終了後次命令発行

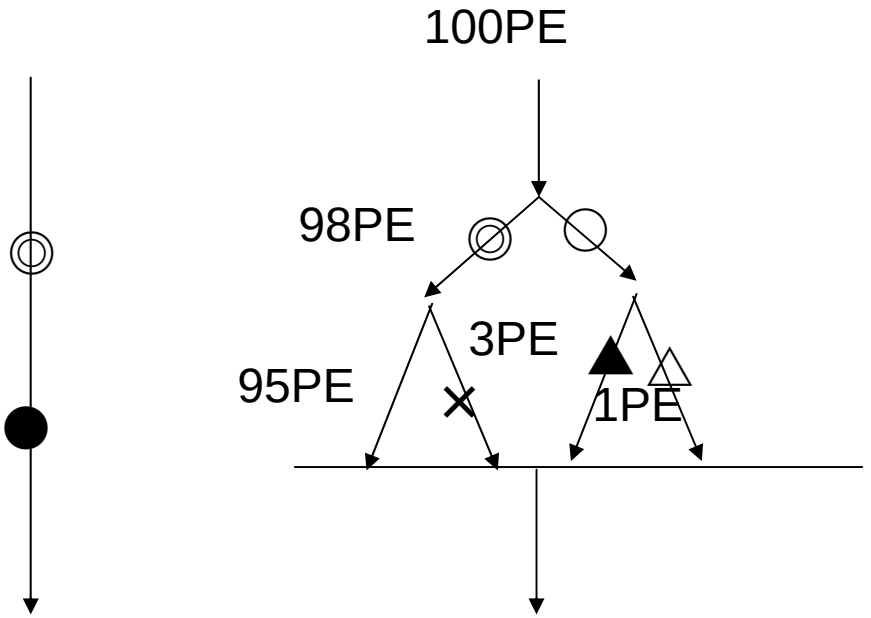
　プログラミングの制約

　　標準的なプログラミング

　　生産性が向上？

　コンパイラによる最適化

　　全PEが同時終了するよう，無資源競合



SIMDの泣き所

①非定型処理に弱い

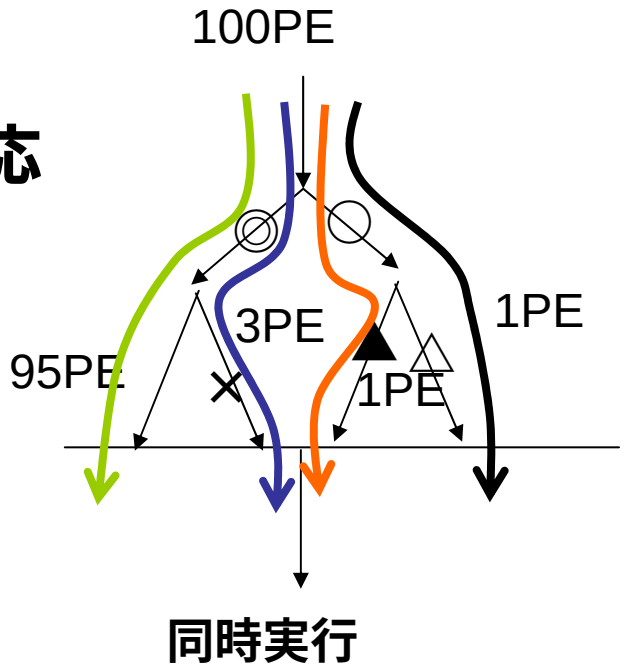
②一番遅い演算器で律速

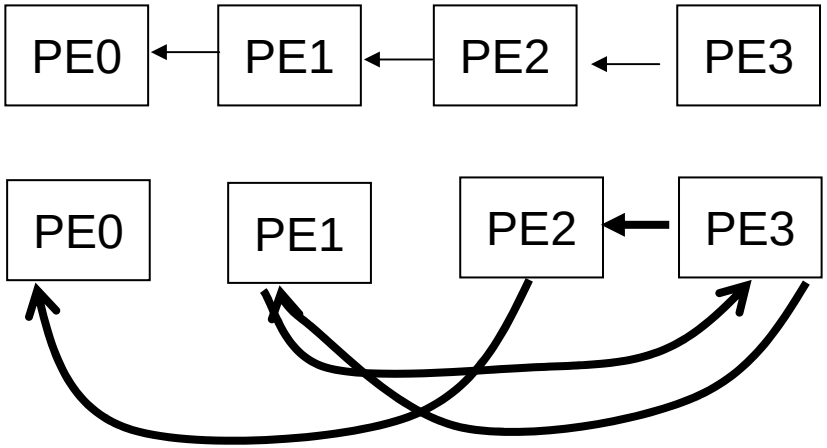
7.3.2 SIMDの柔構造化と非定型処理への対応

①通信の非同期化

②SPMD化:Single Program Multiple Data
分岐点から合流点までをPEで非同期実行

③MIMD化：マルチプロセッサ化！！
SIMDは歴史的な役割を終えたのか？





MPP

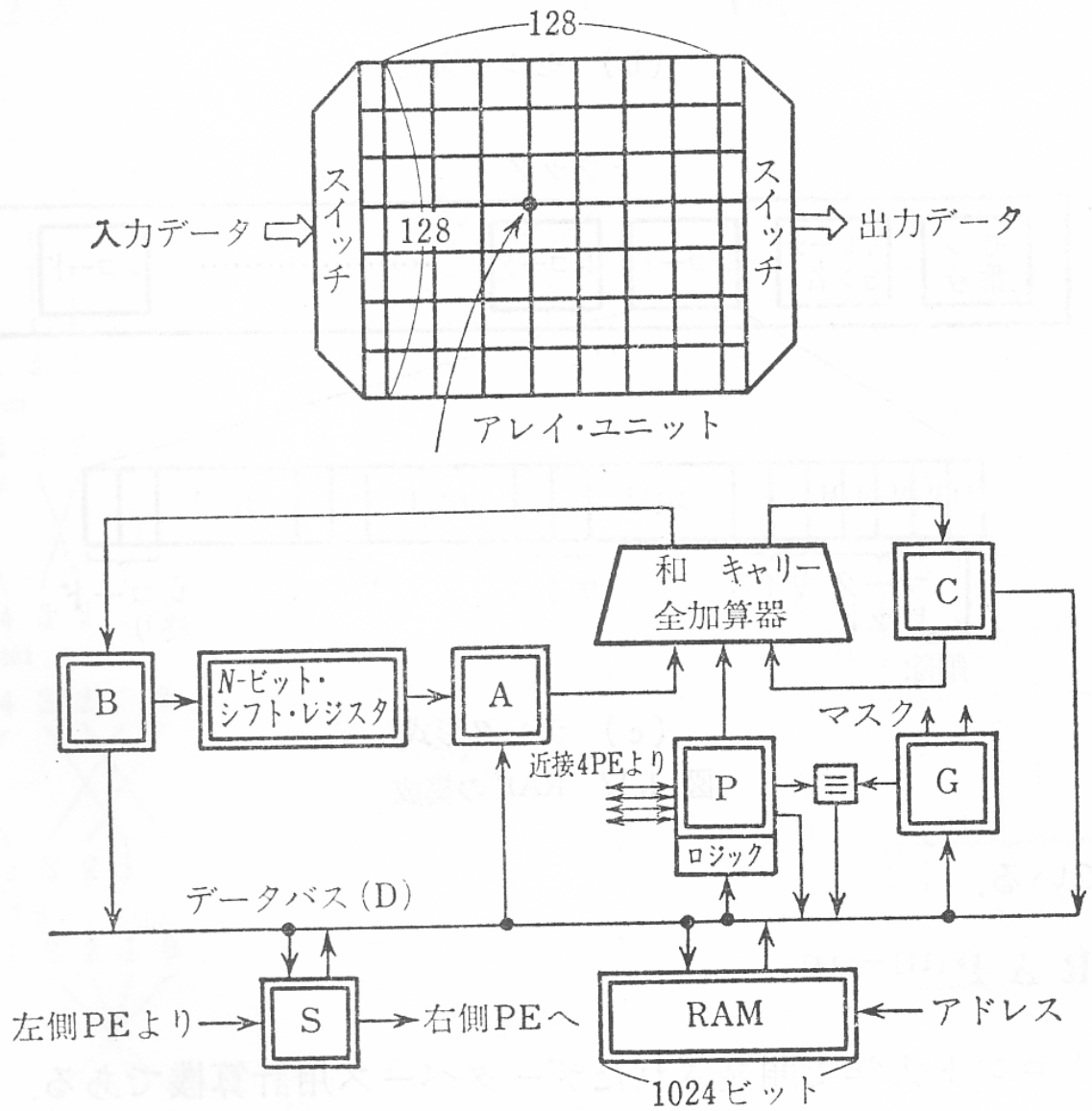


図 4.18 MPP の構成(K. E. Batcher: Design of a Massively Parallel Processor, *IEEE Trans.* C. Vol. 29, No. 9, 1980, pp. 836-840 による)

フィルタ処理：輪郭線抽出など

ラプラシアンオペレータ

$$
\begin{aligned}
\nabla^2 \varphi &= \partial^2 \varphi / \partial x^2 + \partial^2 \varphi / \partial y^2 \\
&= \varphi(I+1, J) - \varphi(I, J) - \\
&\quad (\varphi(I, J) - \varphi(I-1, J)) + \\
&\quad \varphi(I, J+1) - \varphi(I, J) - \\
&\quad (\varphi(I, J) - \varphi(I, J-1)) \\
&= \varphi(I+1, J) + \varphi(I-1, J) \\
&+ \varphi(I, J+1) + \varphi(I, J-1) \\
&- 4\varphi(I, J)
\end{aligned}
$$

	1	
1	−4	1
	1	

(a) Original photograph (b) Printout of the digital gray-level picture (c) Binary picture

Figure 3-2 Picture input and line extraction. The dark horizontal line in the upper part is due to the burn in the CRT surface of the FSS used for digitization.

金出武雄博士論文

7.4パイプライン型スーパコンピュータ

7.4.1基本方式

- 別名：ベクトルプロセッサ（コンピュータ）
- ベクトルデータの高速処理

  Z（I）＝X（I）＋Y（I）

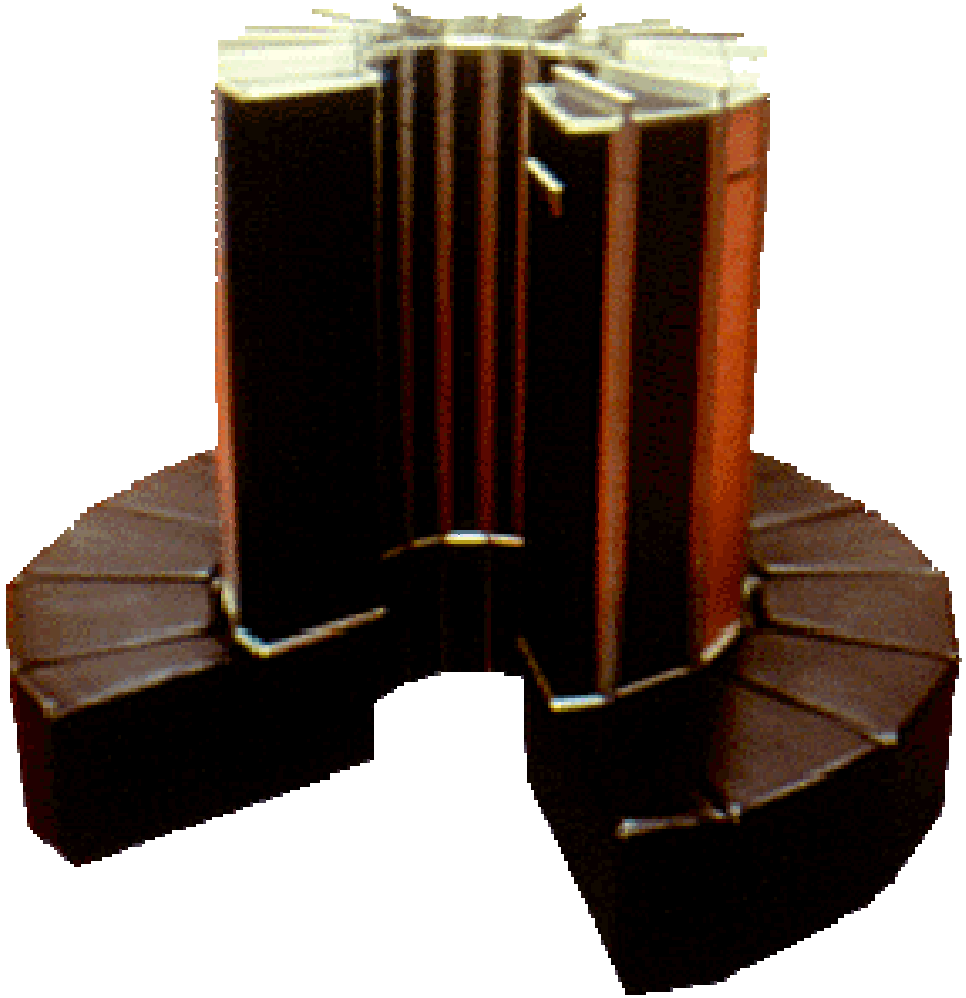
  T＝ΣX（I）＊Y（I）：　内積

  T＝MAX（X（I））, MIN（X（I））

  X（I＋1）＝A（I）＊X（I）＋B（I）：　漸化式
- 線形代数：行列、ベクトル

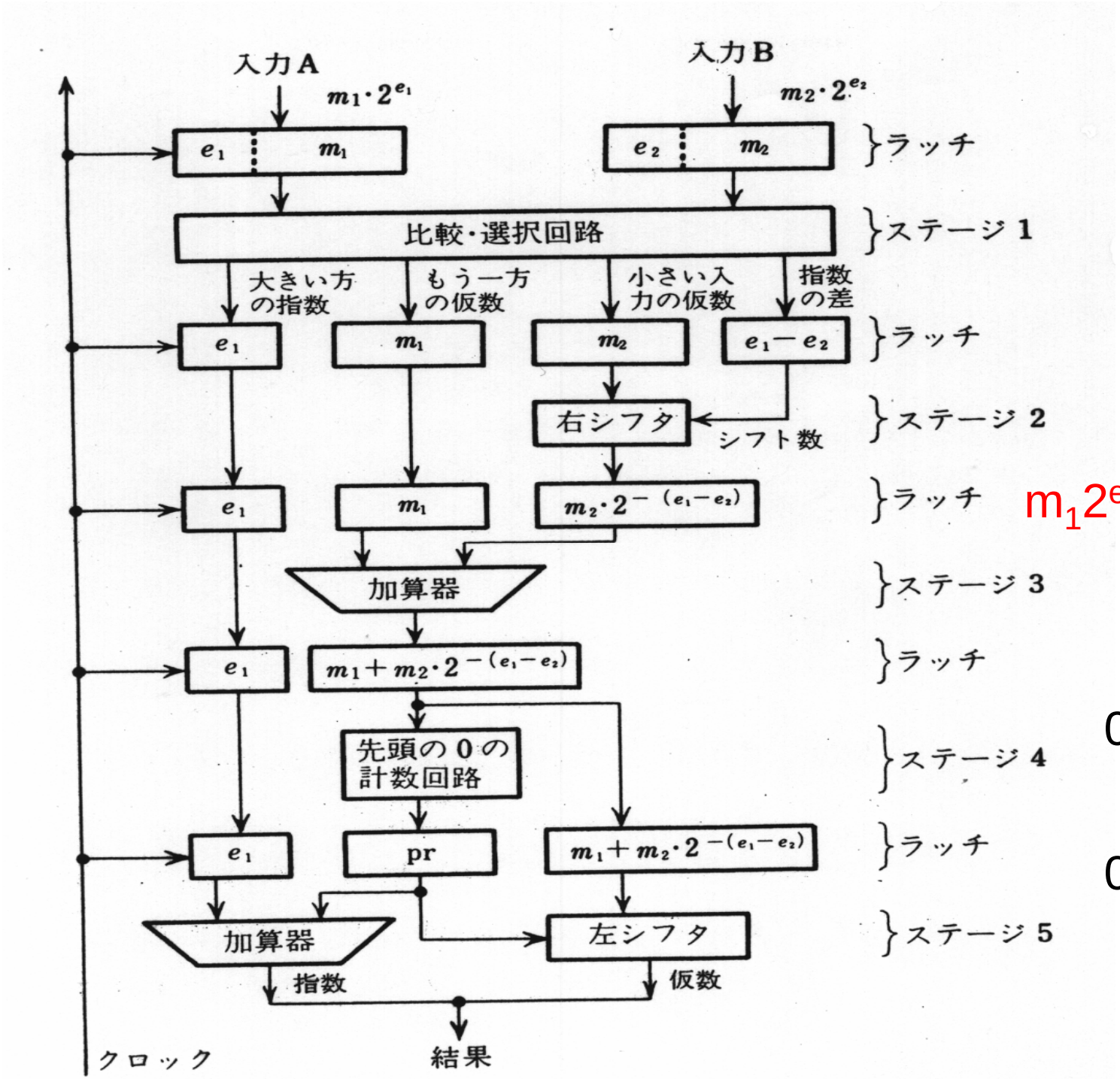
  AX、　AB、　A^T、　A^{-1}

パイオニア：Cray-1 １９７６、マシンサイクル１２．５ nsec、１６０MFLOPS



The World's Most Expensive Love-seat

CACM Vol21,No.1,１９７８,ｐｐ.６３－７２



$m_1 2^{e1} + m_2 2^{-(e1-e2)} 2^{e1}$

$0.00110..2^{10}$

$0.110\ldots\ \ 2^{8}$

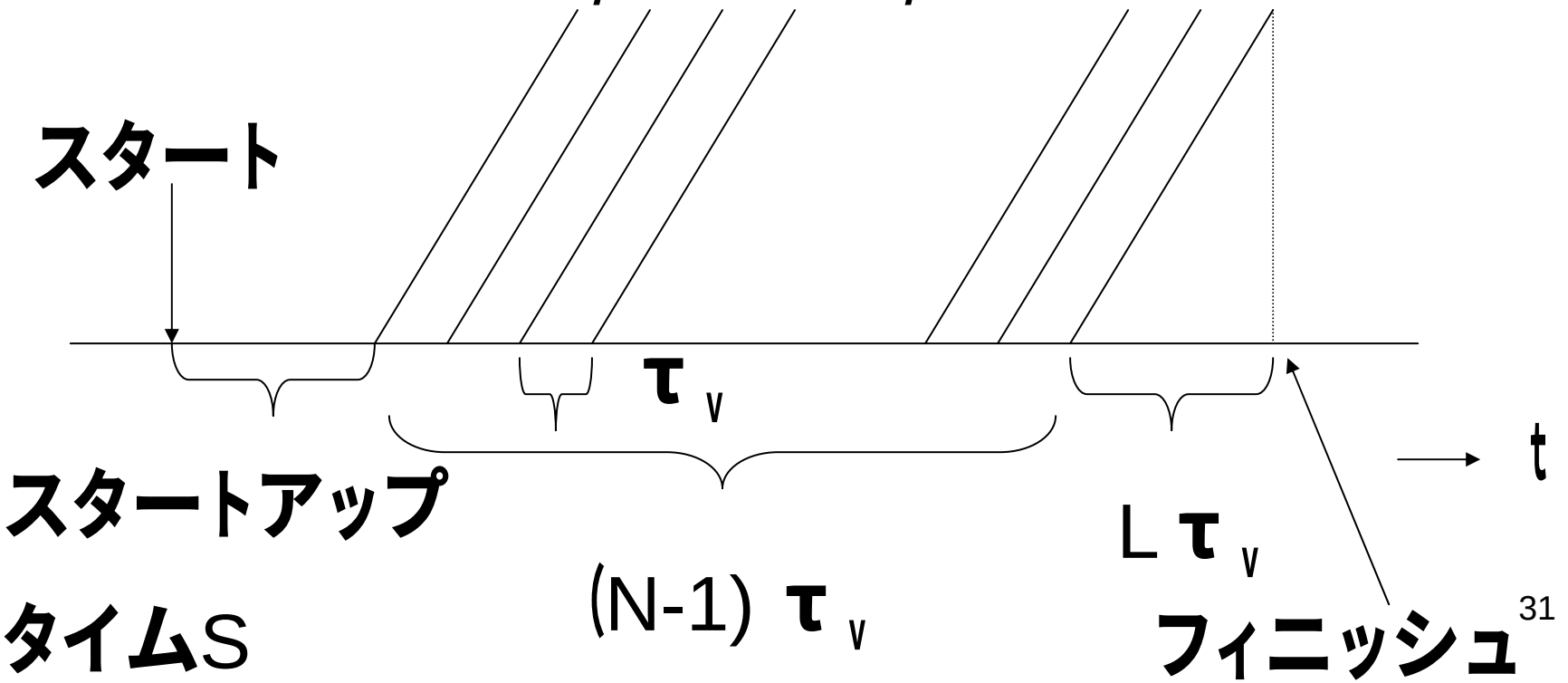
半性能長：最大性能（$1/\tau_v$）の半分の性能となるデータ要素数

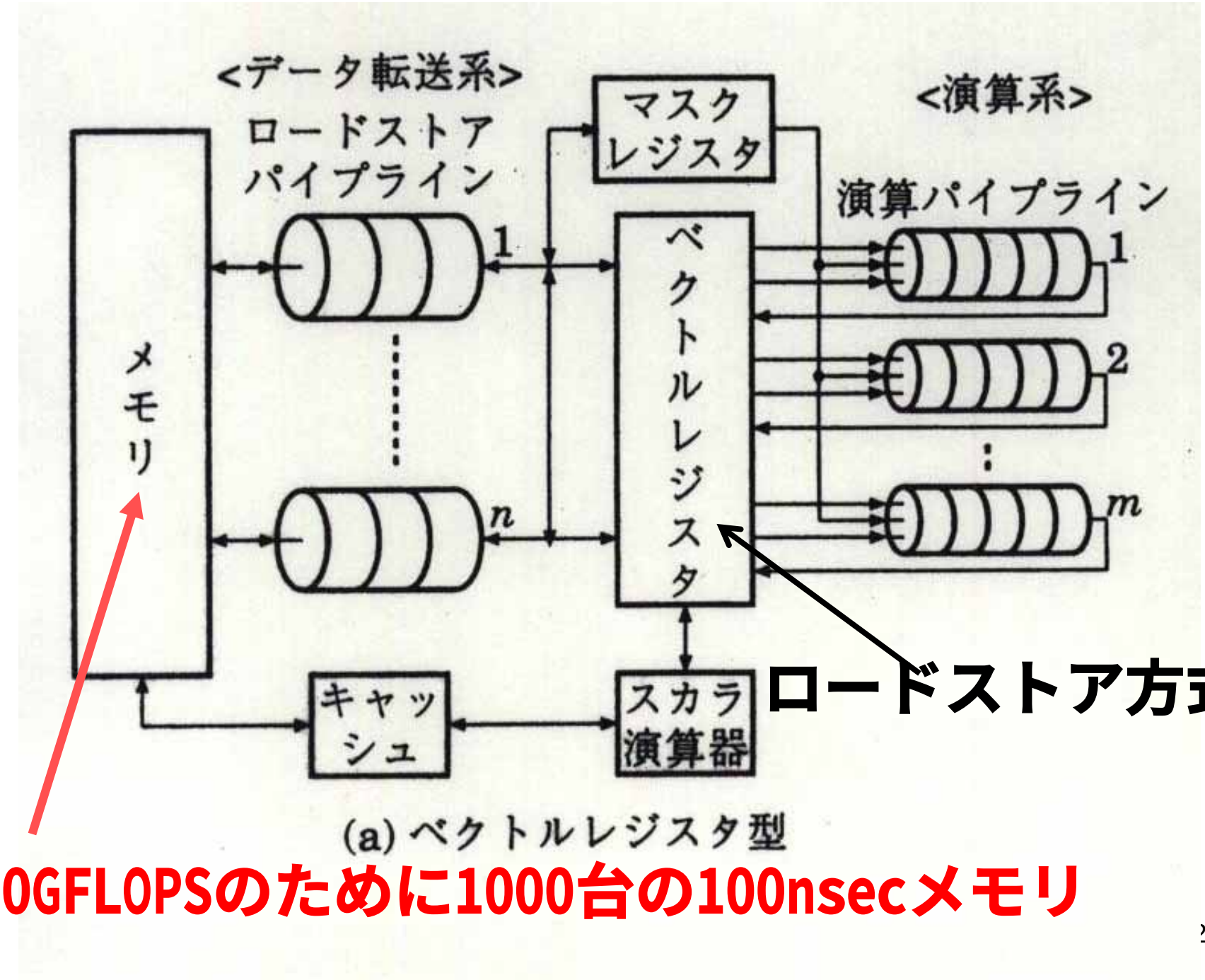
演算時間

$$T=S+(N-1+L)\tau_v$$

$$=(N+N_{1/2})\tau_v$$

$$1/(2\tau_v)=N_{1/2}/(S+(N_{1/2}-1+L)\tau_v)$$





10GFLOPSのために1000台の100nsecメモリ

ベクトル処理の高速化

Z(I)=X(I)*Y(I)

SET Vector Length

LOADV VR0 M(R2+D1)

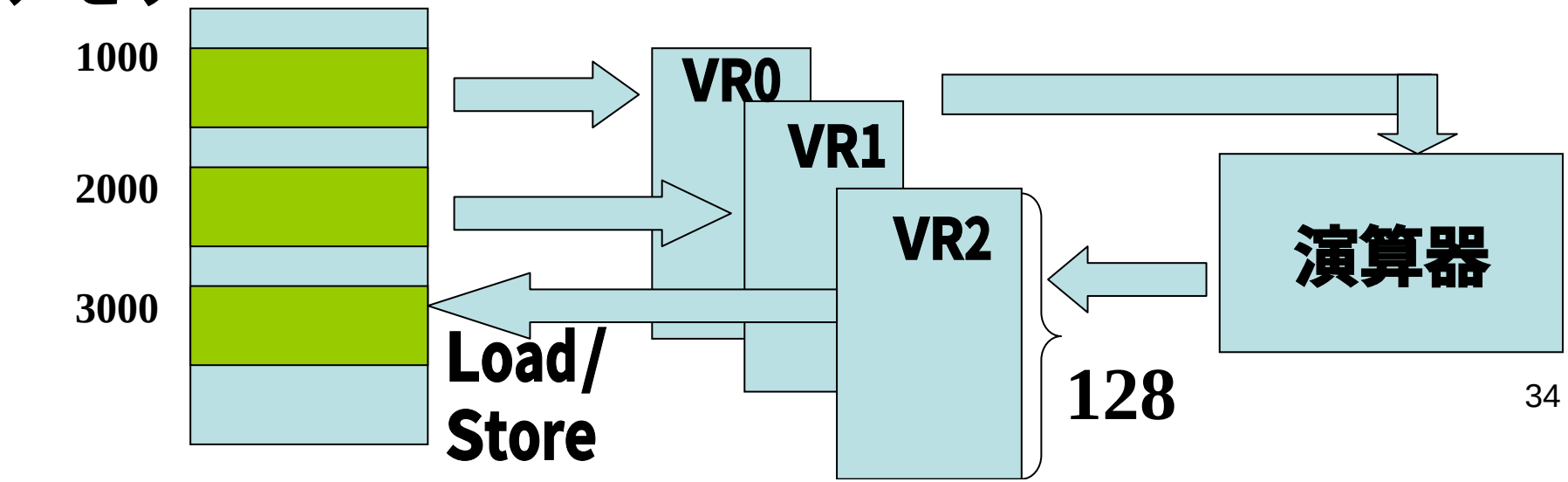
LOADV VR1 M(R2+D2)

MULFV VR2 VR1 VR0

STOREV M(R2+D3) VR2

ベクトル処理の高速化

```
Z(I)=X(I)*Y(I)
SET Vector Length   128
LOADV  VR0 M(1000)
LOADV  VR1 M(2000)
MULFV  VR2 VR1 VR0
STOREV M(3000) VR2
```

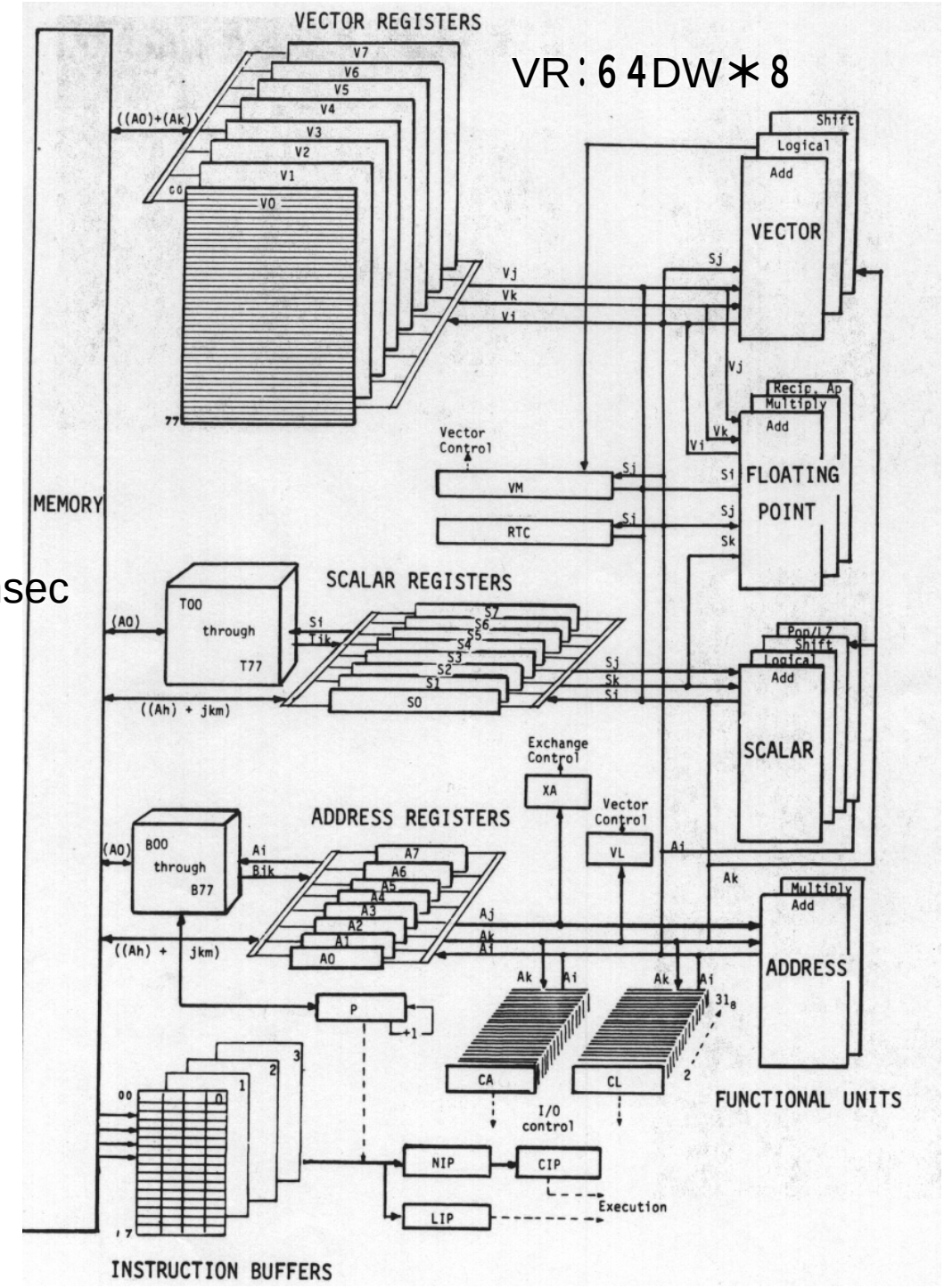


メモリ

8MB

16バンク

1DW/12．5nsec



1976年

12．5nsec

160MFLOPS

115KW

ECLロジック

ベクトルプロセッサの普及理由

①高速性：通常の計算機より1桁以上高速

②連続性：過去に作成されたFORTRANプログラムがそのまま利用

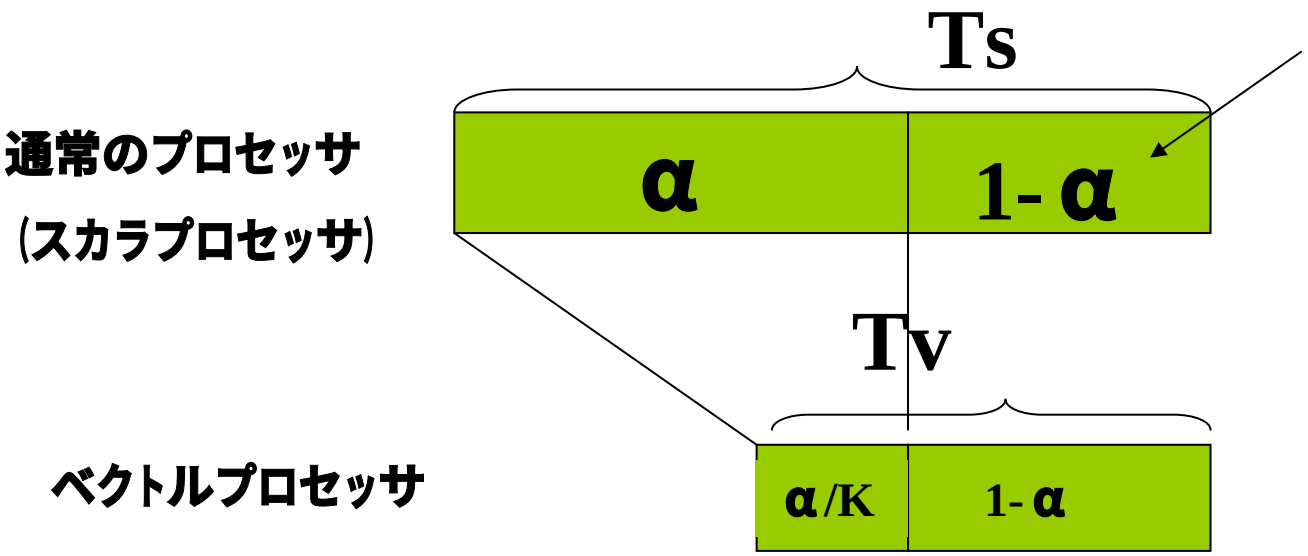
③単純性：ベクトル演算で問題が定式化できれば最大性能、　ベクトル：高等学校数学

④汎用性：多様な数値計算分野に適応

⑤単一性：各社ハードウェアシステムの構造が同一「よい」プログラムはどのスーパコンピュータでも高速実行

7.4.2ベクトルプロセッサの泣き所：アムダールの法則

①ベクトル化率（α）が高くないとだめ

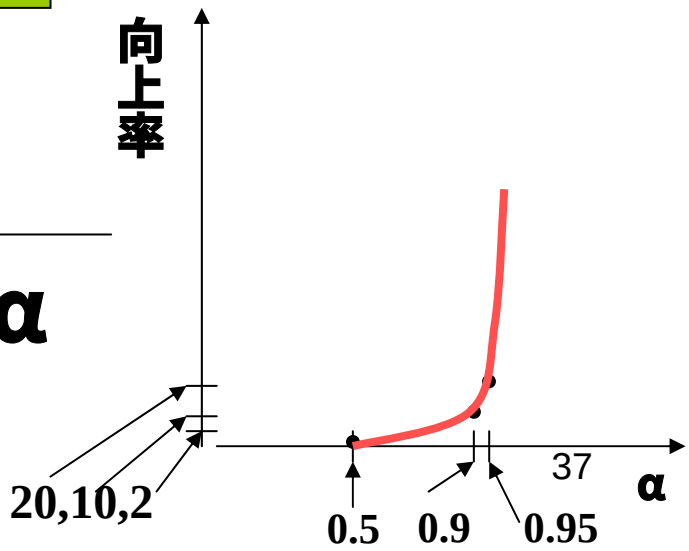


α90%で高々10倍

性能向上率

$$Ts/Tv = \frac{1}{1-\alpha+\alpha/K}$$

$$\frac{1}{1-\alpha}$$

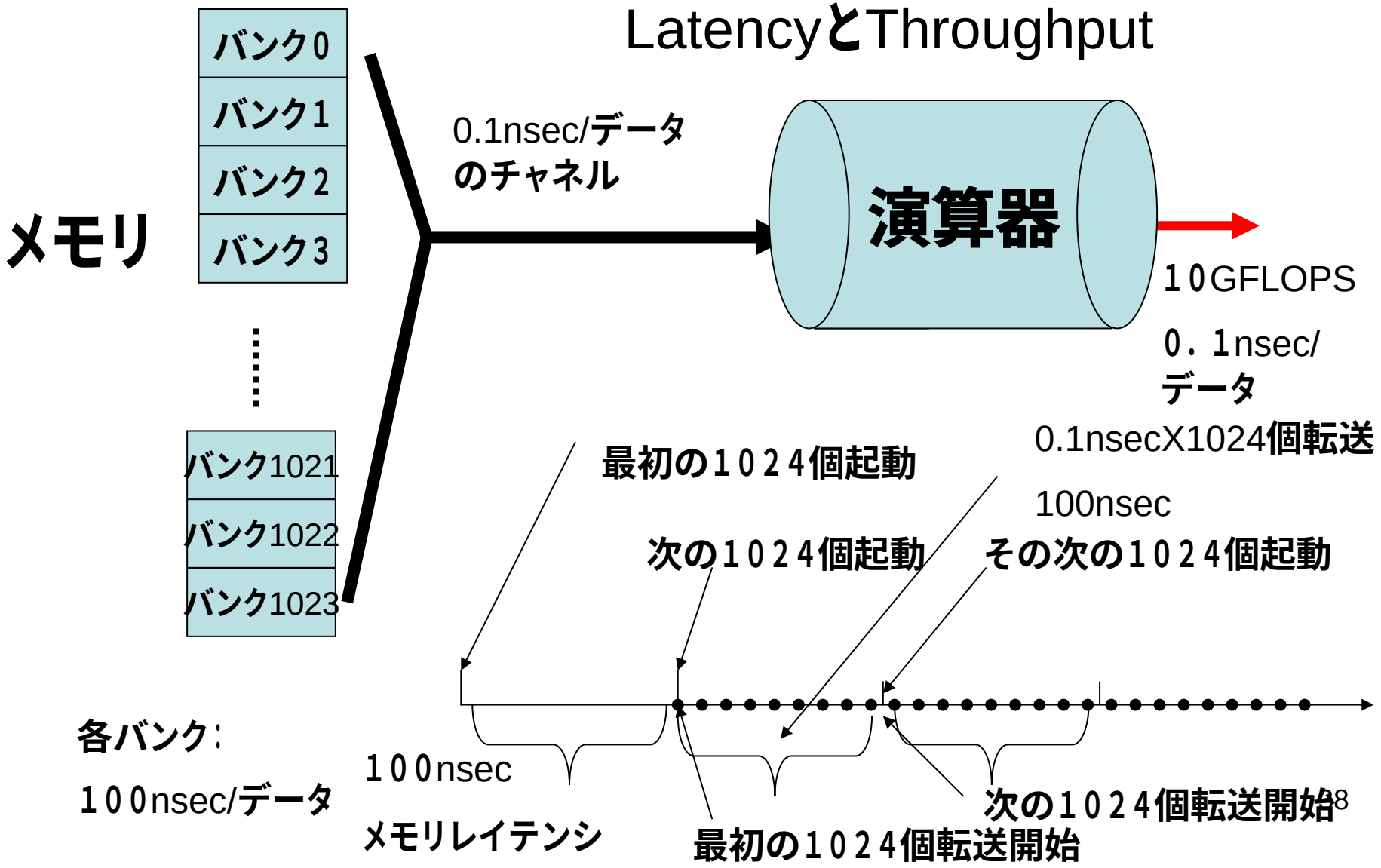


通常のプロセッサも高速の必要：二重苦

ベクトルプロセッサの泣き所②：メモリバンド幅

・１０２４バンク構成

・スカラユニット＋ベクトルユニットの開発



7.5マルチプロセッサとマルチコンピュータ

7.5.1マルチプロセッサの基本方式

(1)**メモリ共有型** VS **メッセージ交換型**

最大性能のシステム：

地球シミュレータ:5120**台のマルチプロセッサ**

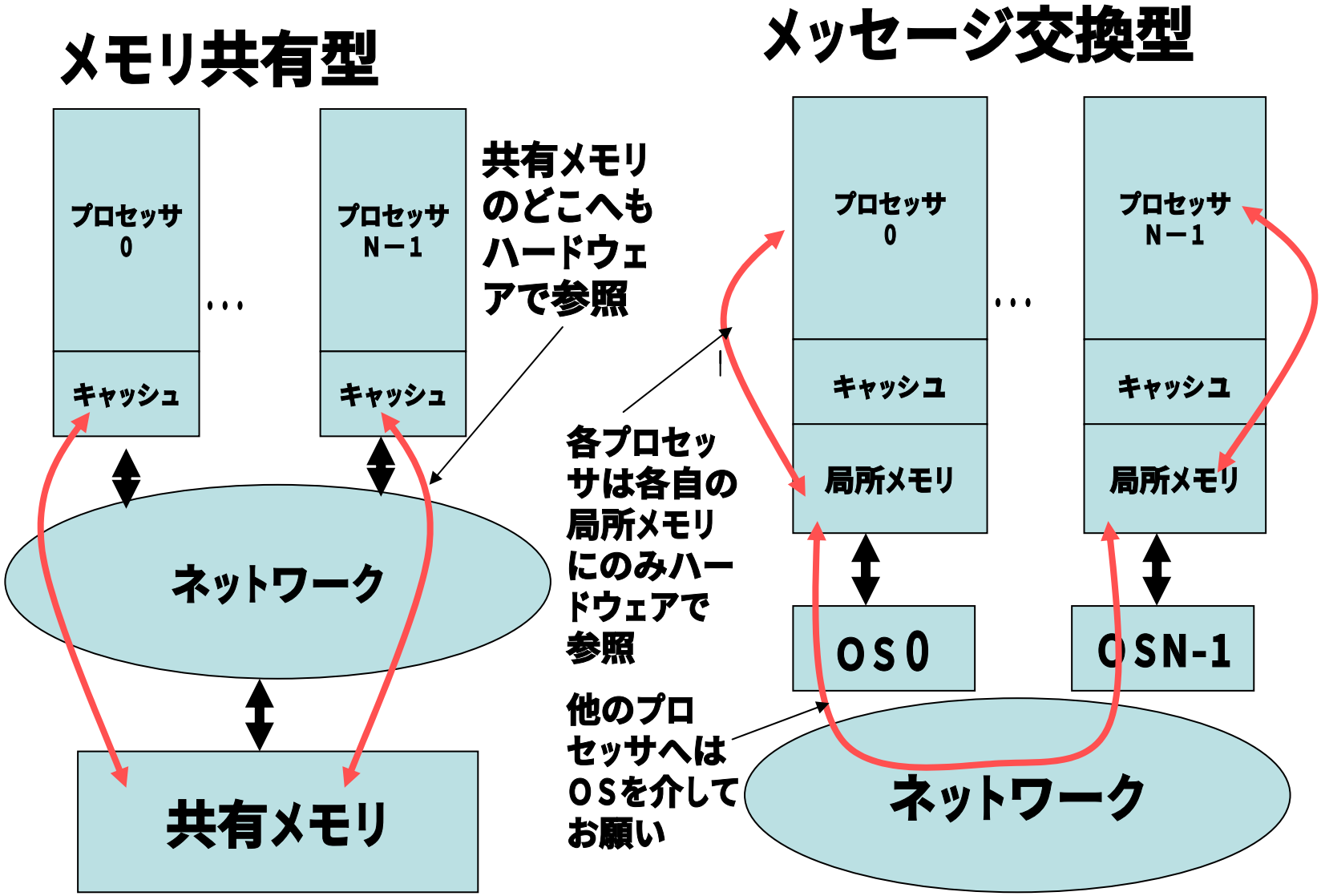
ノード：8GFLOPS**のプロセッサ**8**台、共有メモリ**

システム：640**ノード、メッセージ交換**

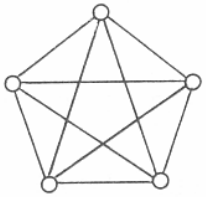
(2) **ベクトルパラレル** VS **スカラパラレル**

SX-8、CRAY X1 VS SR11000,HPC2500

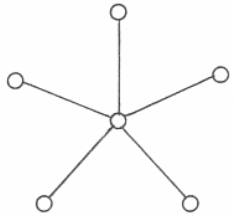
BlueGene/L


メモリ共有型
プロセッサ
0
プロセッサ
N－1
キャッシュ
キャッシュ
共有メモリ
のどこへも
ハードウェ
アで参照
ネットワーク
共有メモリ
メッセージ交換型
プロセッサ
0
プロセッサ
N－1
キャッシュ
キャッシュ
局所メモリ
局所メモリ
各プロセッ
サは各自の
局所メモリ
にのみハー
ドウェアで
参照
OS0
OSN-1
他のプロ
セッサへは
OSを介して
お願い
ネットワーク

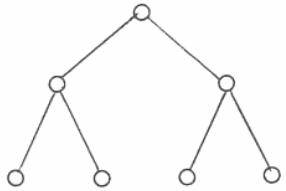
静的網



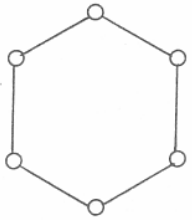
(a) 完全網



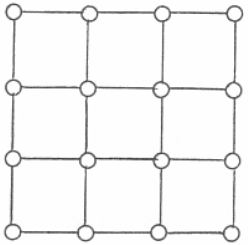
(b) スター網



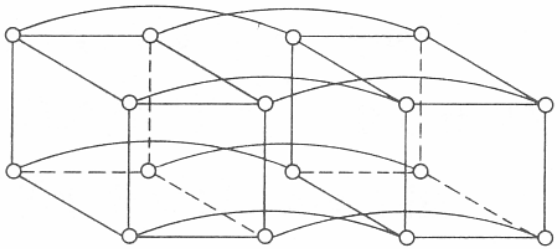
(c) 木状網



(d) リング網

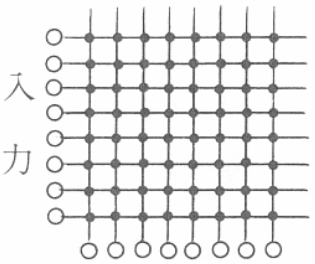


(e) 格子網(トーラス網)
(最上・下, 最左・右を結合)



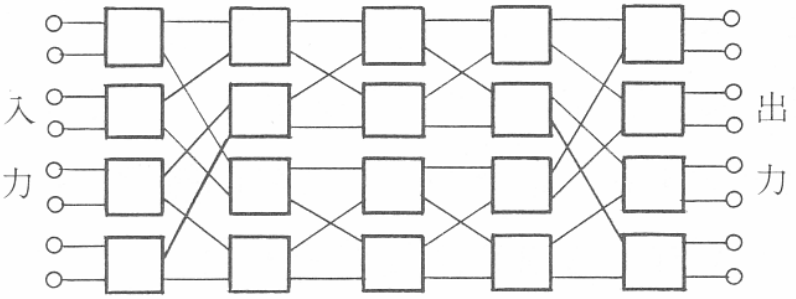
(f) ハイパーキューブ網

動的網



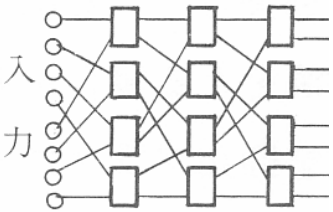
•：スイッチ

(a) クロスバー網



□：2×2スイッチ

(b) Benes̆網



□：2入力2出力
交換スイッチ

(c) オメガ網

多段結合網

○：演算装置やプロセッサ

マルチプロセッサのパイオニア：

C.mmp：カーネギメロン大、１９７１年、１６台

PACS：筑波大、１９８４年、１２８台

LINCS：阪大、１９８５年、６４台

マルチプロッサの課題

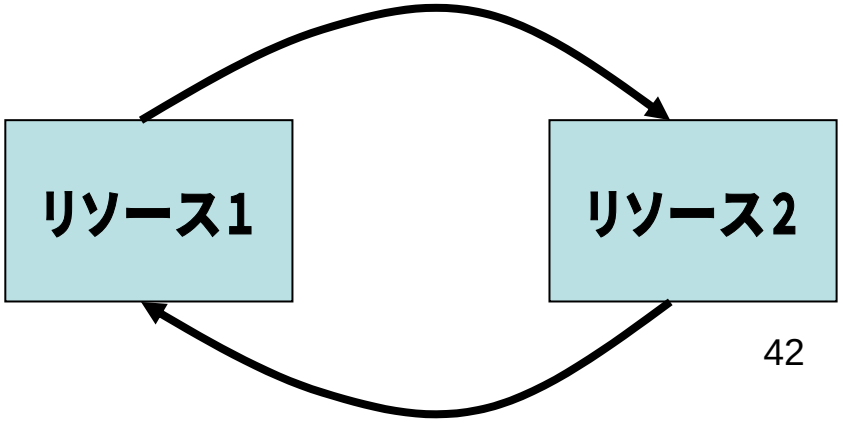
ネットワークを介した高速データ転送

プロセッサ間の高速同期機構

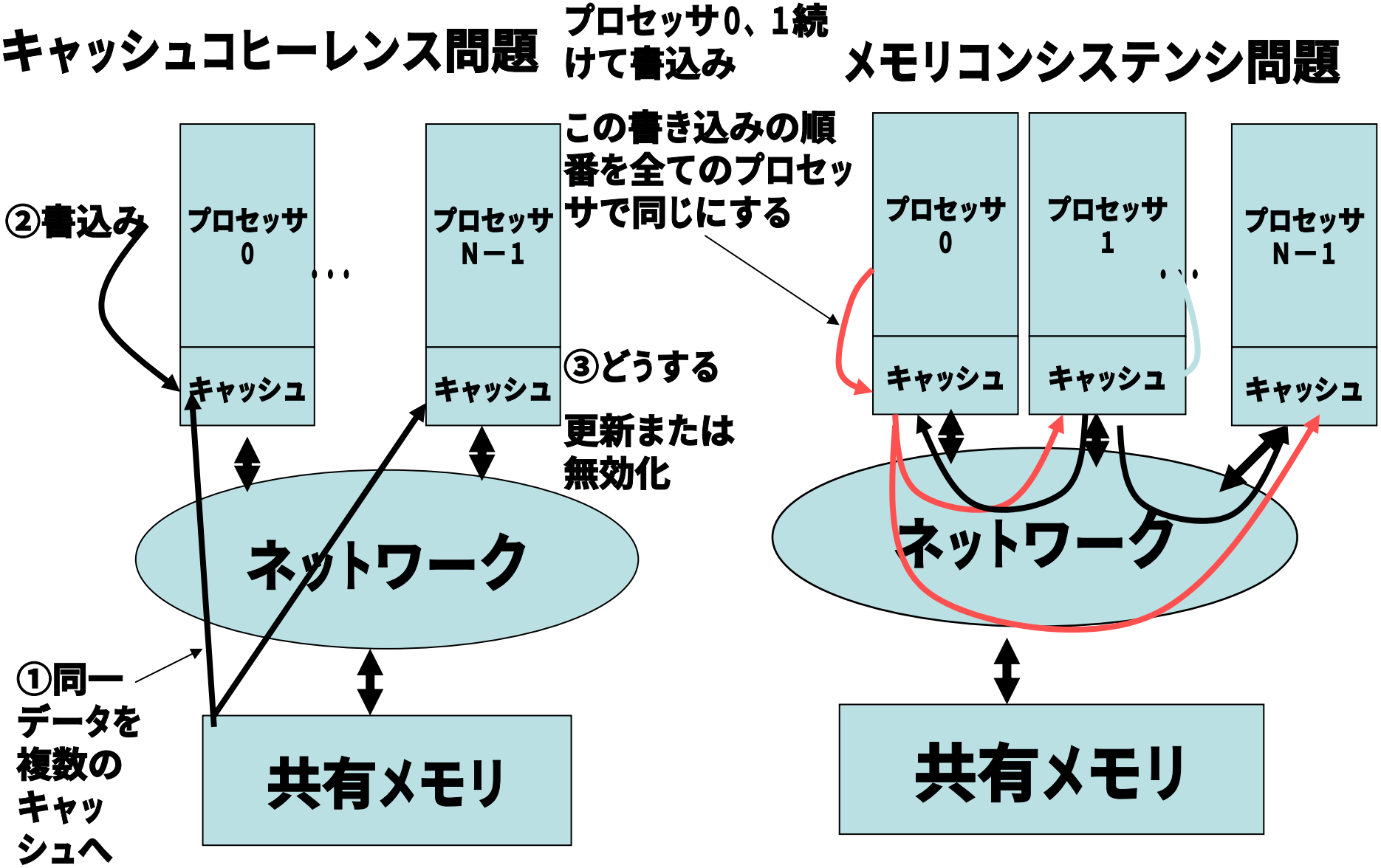
最適な処理粒度：通信時間に比べて処理時間が大きくないとだめである

プログラム分割

デッドロック対策



7.5.2メモリ共有の2つの問題



7.5.3メッセージ交換方式の問題点

ＯＳの介在で処理時間が長い：１０msec

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

数μ秒へ

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**

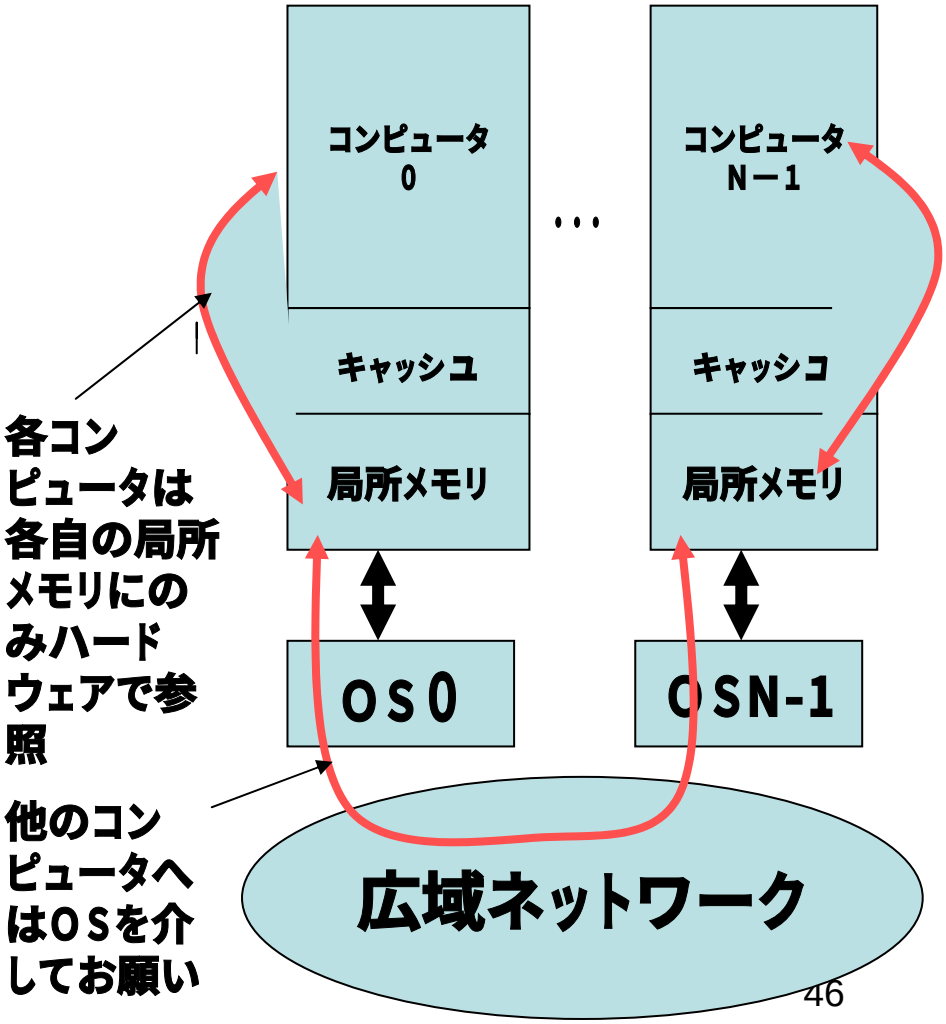
Myri-10G：**１０**Gigabit/s,dual protocol NIC

7.5.4マルチコンピュータ

- メッセージ交換型マルチプロセッサの広域版
  - 各コンピュータは他の仕事も行っている。負荷の分散
  - ネットワークの遅延大、込み具合に依存

光1000km：3.3msec

  - 安全性の確保



NoW：Network of Workstations：**ルーツ、**1990**年代**

クラスタコンピューティング

パソコンと市販の高速ネットワーク（Myrinet**など**）**で接続**

Commodity Computing（**日常品によるシステム構築**）

グリッドコンピューティング

超高速コンピュータ網形成プロジェクト（NAREGI：National Research Grid Initiative）

文部科学省：平成１５年から２０年

１００TFLOPS**の性能を目標**

7.5.5マルチプロセッサ、コンピュータの問題点

超高性能なコンピュータの実現：

マルチプロセッサやマルチコンピュータしかないが、・・・

ユーザからみた問題点

①不連続性

過去のプログラムの大幅な手直し

②複雑性

- **並列アルゴリズムの複雑性**

パイプライン方式：1次元ベクトルを意識

マルチプロセッサ：相互結合網

(ネットワーク) の2、3次元構造

効率のよいアルゴリズムを作成：困難

たとえば、1から１００までの数のすべての部分和

(1+2、1+2+3、…、1+2+・・・+100) 計算

パイプラインでは非常に簡単

マルチプロセッサでは非常に難しい

- **並列プログラミングの複雑性**

  SIMD**方式やパイプライン方式**

  1つの命令で多数のデータに対して同一演算

  1つ1つの命令は逐次実行

  集合データに対する「逐次」プログラム

  逐次プログラムは何度実行して同一結果

  決定的 (deterministic)

  デバッグも簡単

  デッドロックなし

③専用性

④多様性⇒移植性

⑤市場性

7．6最新のスーパコンピュータと地球シミュレータ

7.6.1最新のシステム例

選択肢

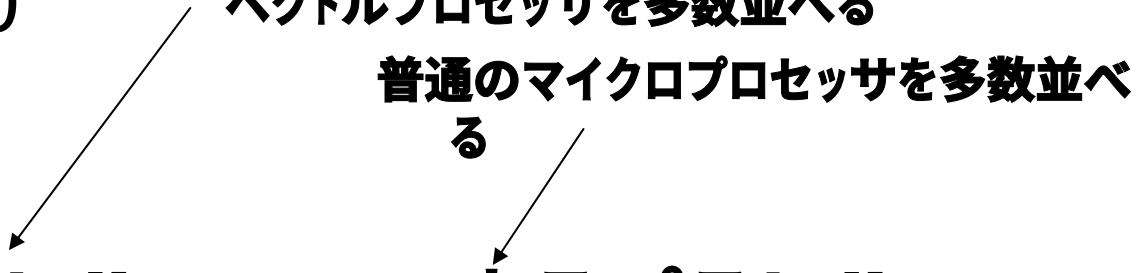
①要素プロセッサ

ベクトルプロセッサかスカラープロセッサ

②メモリ共有かメッセージ交換か

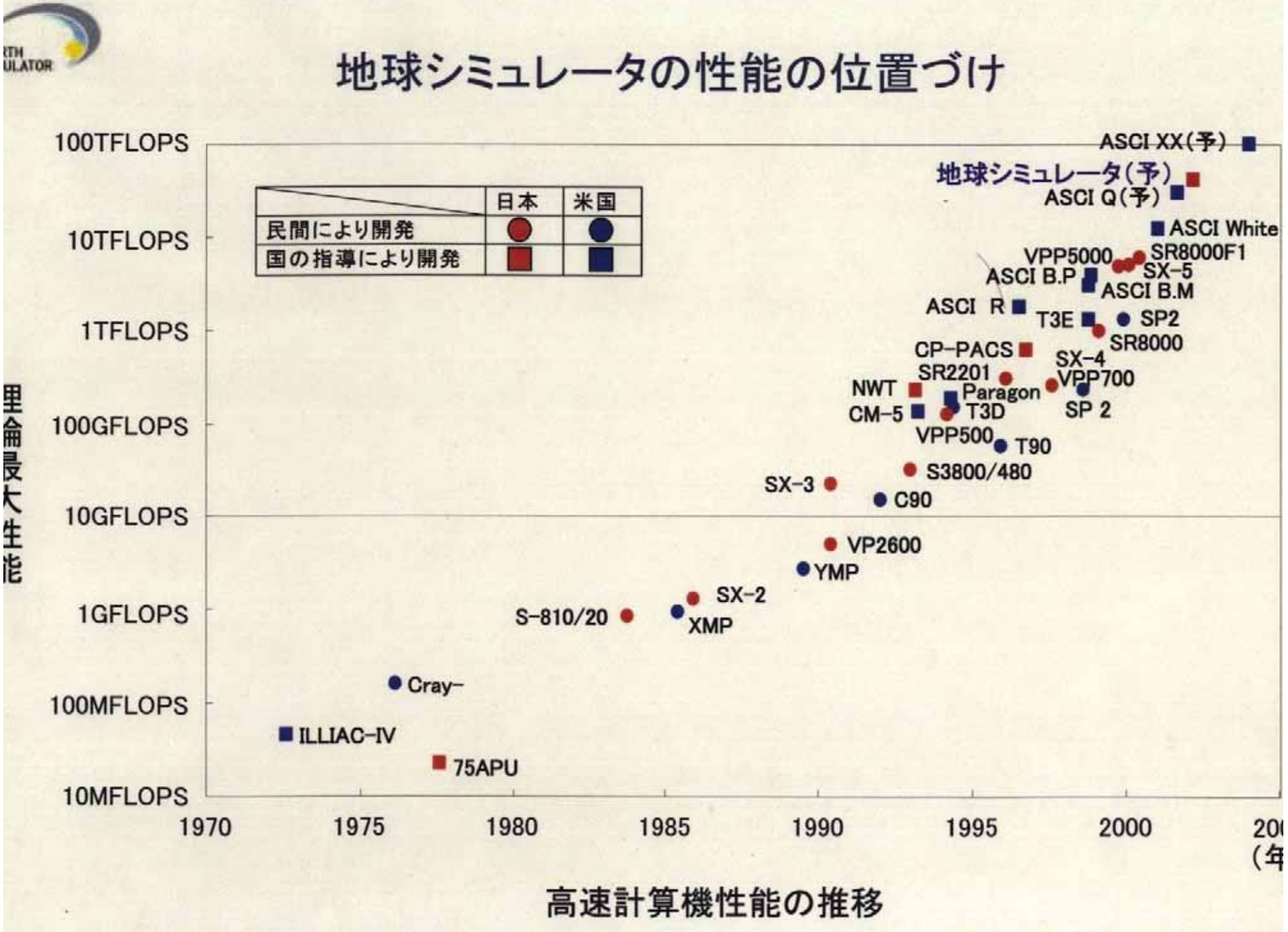
③ネットワークの構成は

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　②、④　VS　スカラパラレル　①、③、⑤
(2) メモリ共有　①、②、③、④　VS　メッセージ交換　⑤
(3) ネットワーク
クロスバ　①、②、③　VS　トーラスなど他網　④、⑤

地球シミュレータの性能の位置づけ



高速計算機性能の推移

①NEC

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）

②日立

ベクトルパラレル

1982 S-810 **630**MFLOPS

1987 S-820 **3**GFLOPS

1992 S-3000 **32**GFLOPS

スカラパラレル

1995 SR**2201** **600**GFLOPS

1999 SR8000 **7.3**TFLOPS

2003 SR11000 **62**TFLOPS

(Power5 (**1.9**GHz)、**16**PE/**ノード**、121.6GFLOPS/**ノード**、**最大512ノード**、**多段クロスバネット**：**12**GB/sx2 (**ノード当たり**)

③FUJITSU　PRIMEPOWER HPC

VPPからSPPに切り替え

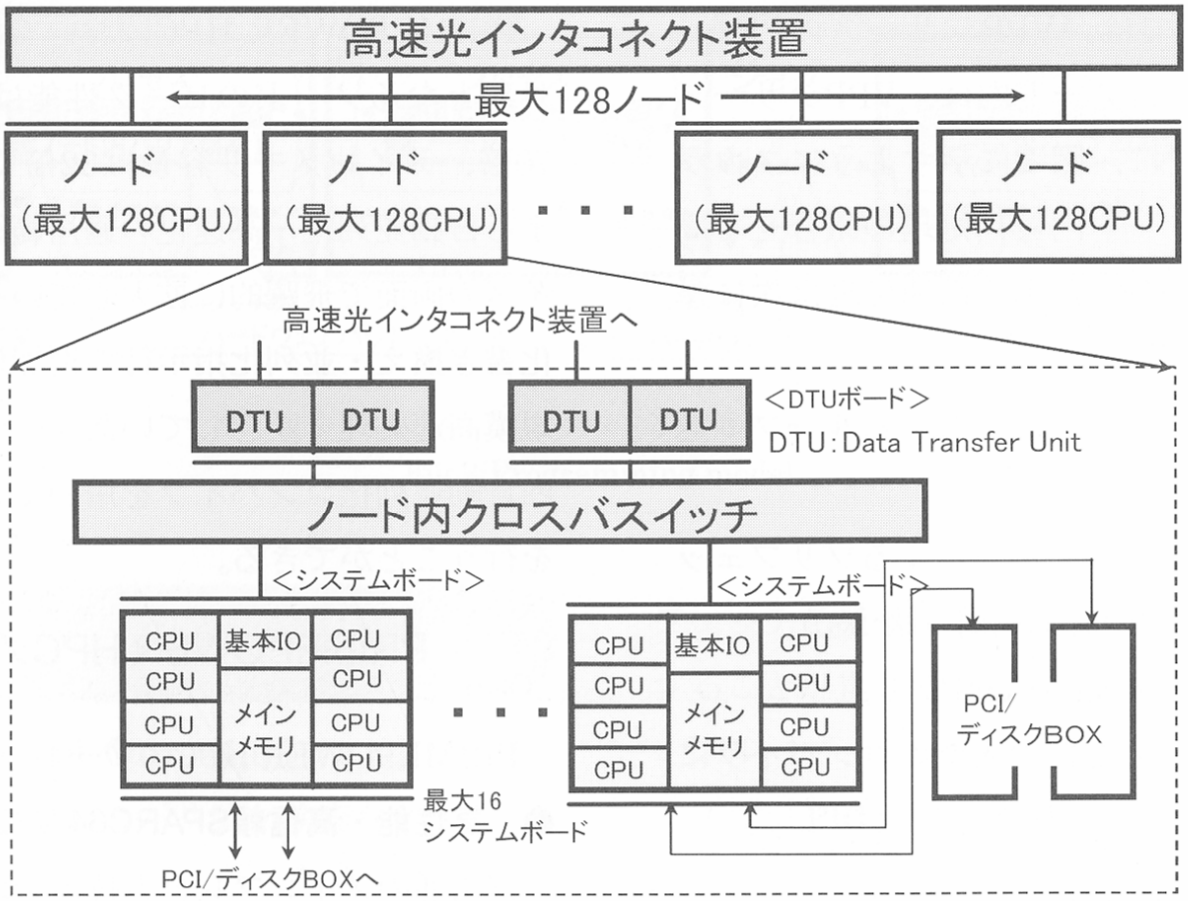


図-1　PRIMEPOWER HPCシステムの構成
Fig.1-System configuration of PRIMEPOWER HPC.

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

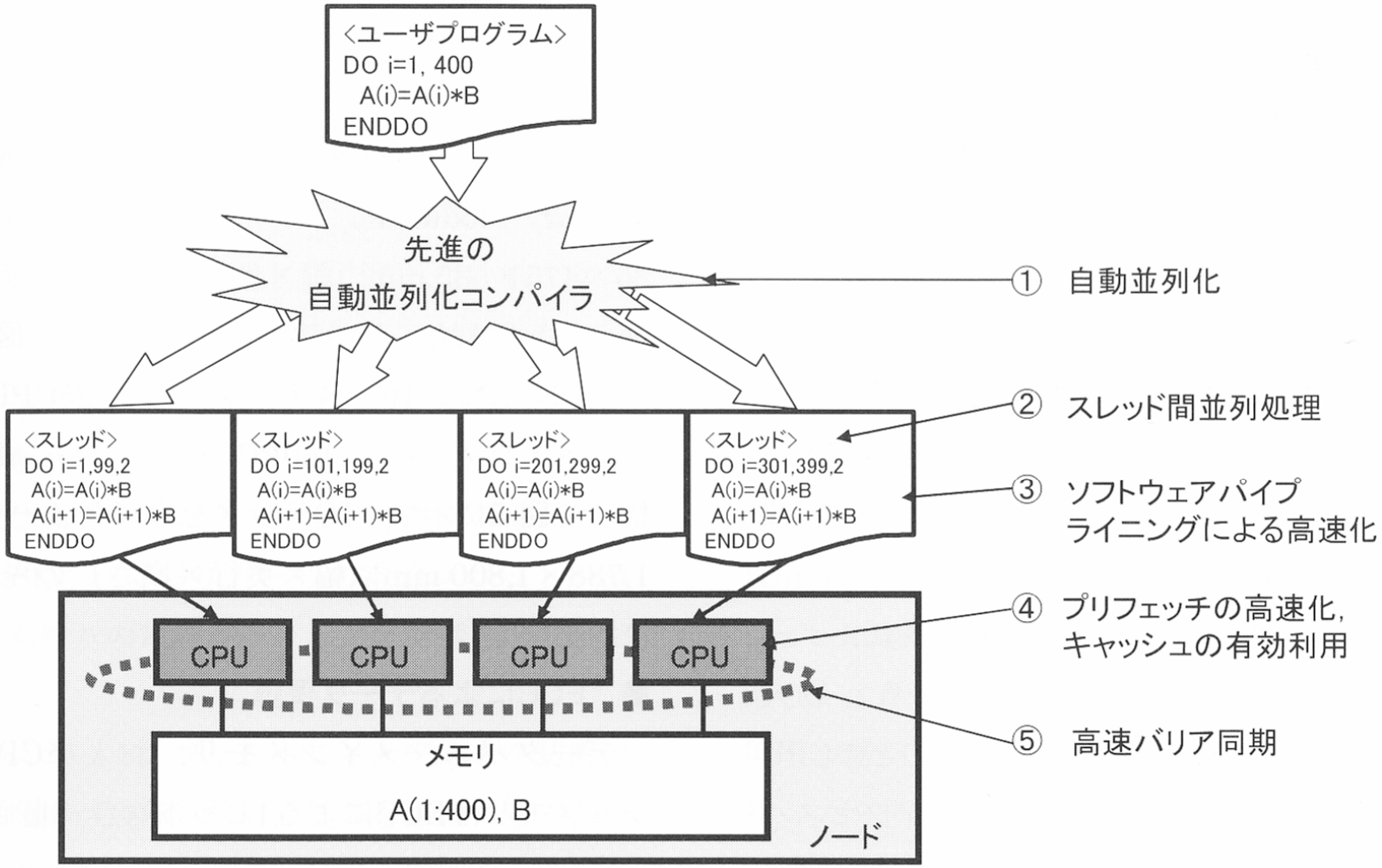


図-2　ノード内並列処理
Fig.2-Parallel execution method in node.

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

IBF

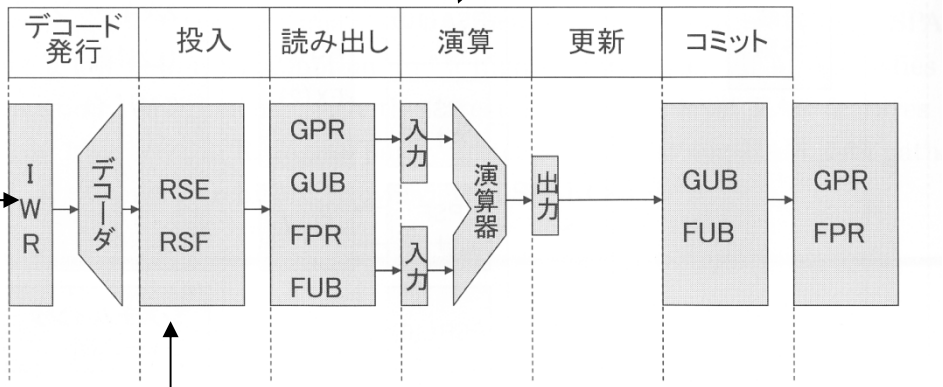


図-2　SPARC64 Vのパイプライン
Fig.2-Pipeline structure of SPARC64 V.

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

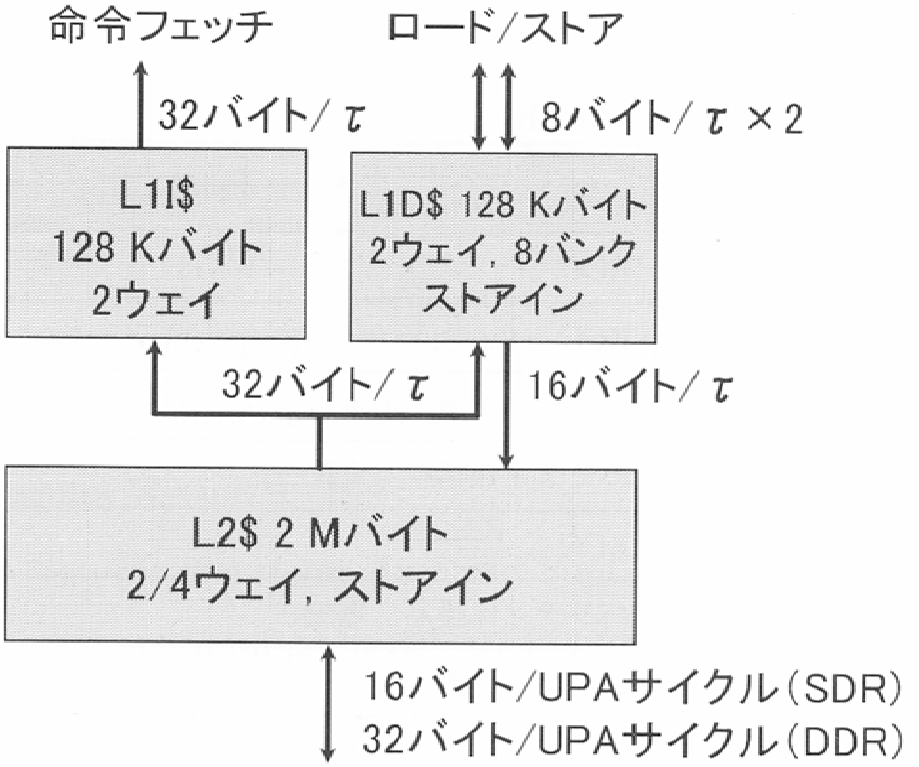


図-3　SPARC64 Vのキャッシュ
Fig.3-Cache structure of SPARC64 V.

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：1．3GHz

消費電力：50W

7.6.2 TOP500

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

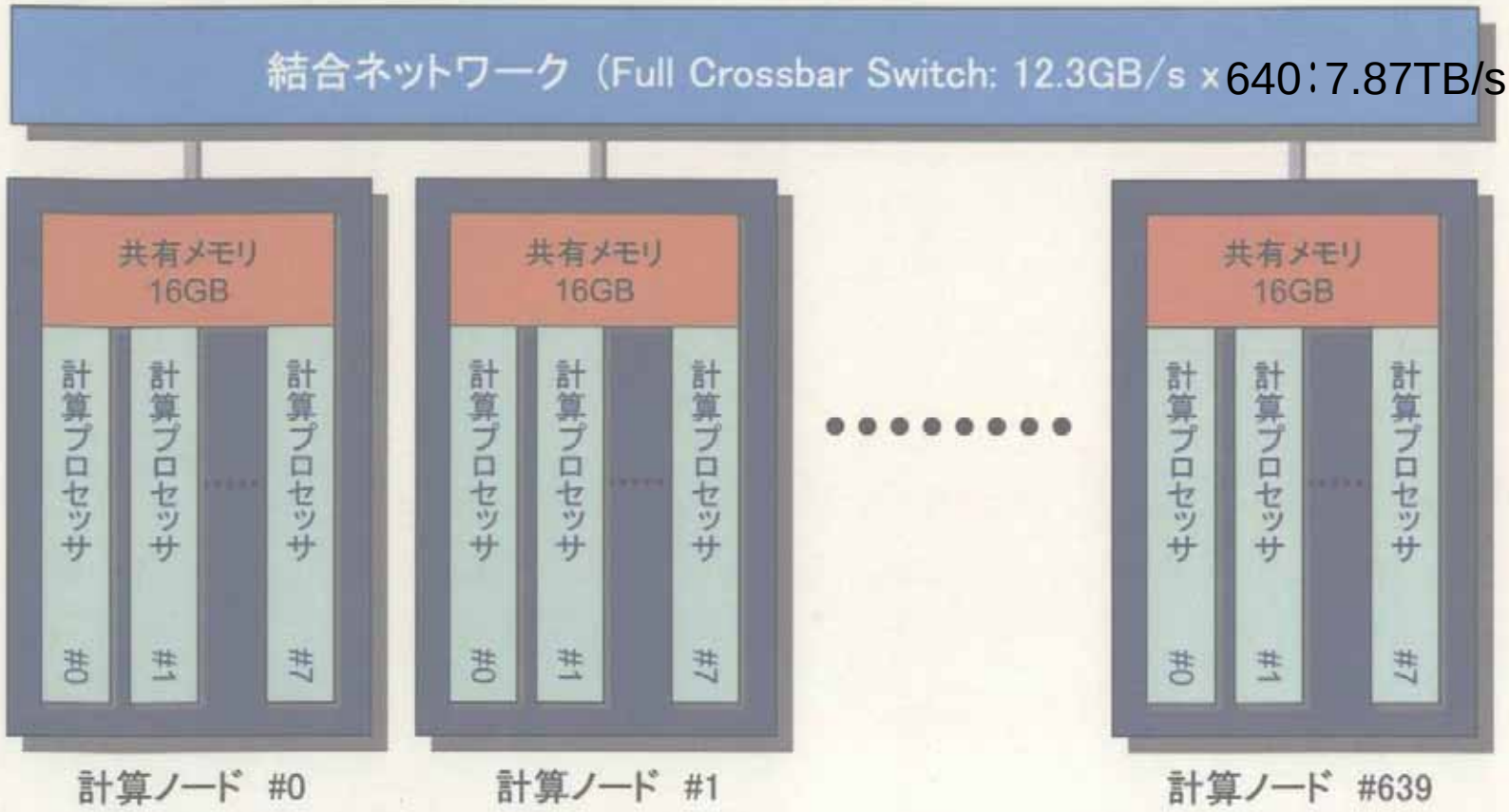
Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM	131072	2005	280600	367000
2	IBM Thomas J. Watson Research Center United States	BGW - eServer Blue Gene Solution IBM	40960	2005	91290	114688
3	DOE/NNSA/LLNL United States	ASC Purple - 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

7.6.3地球シミュレータと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）					

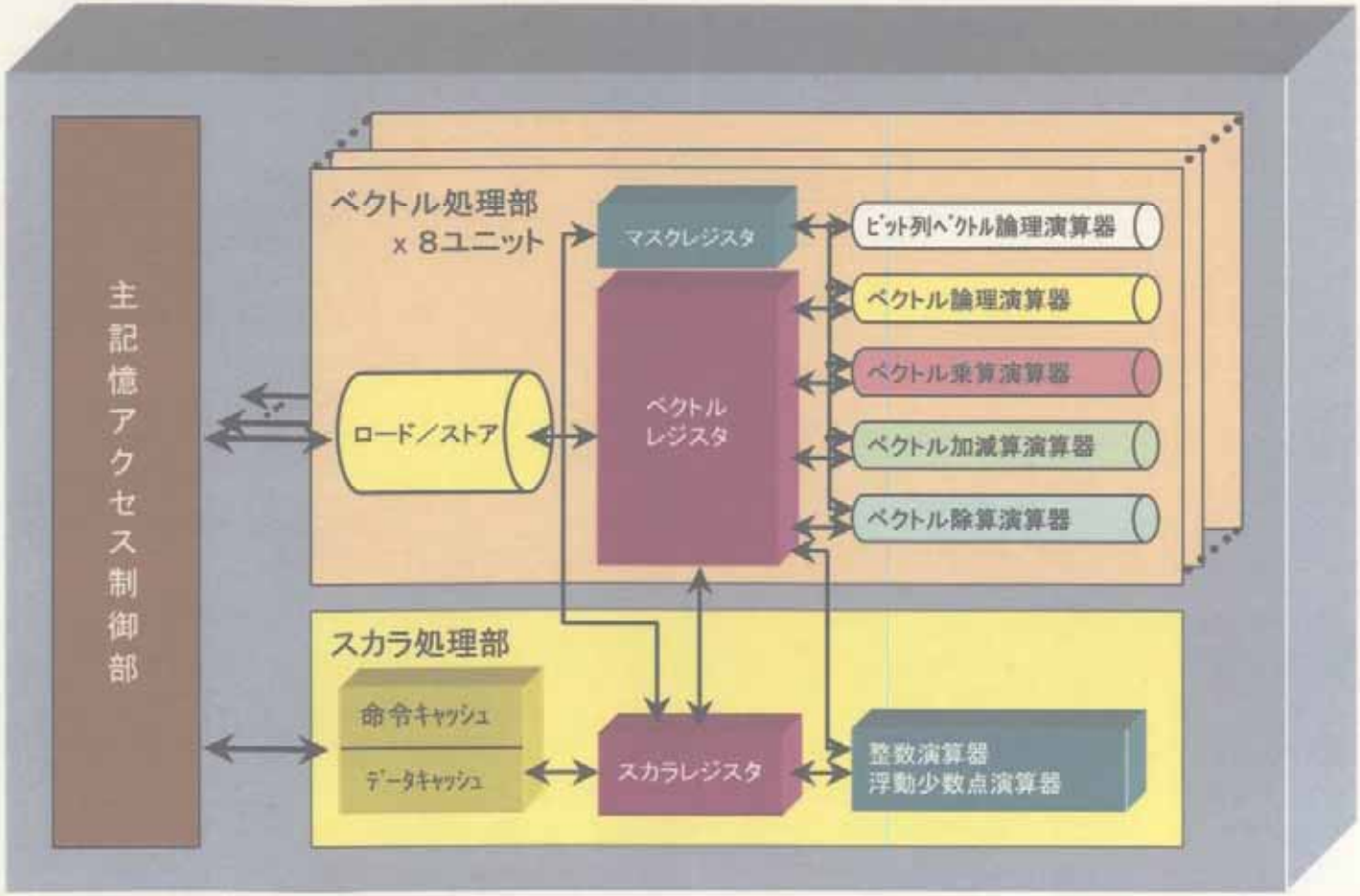
地球シミュレータの全体構成

- 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



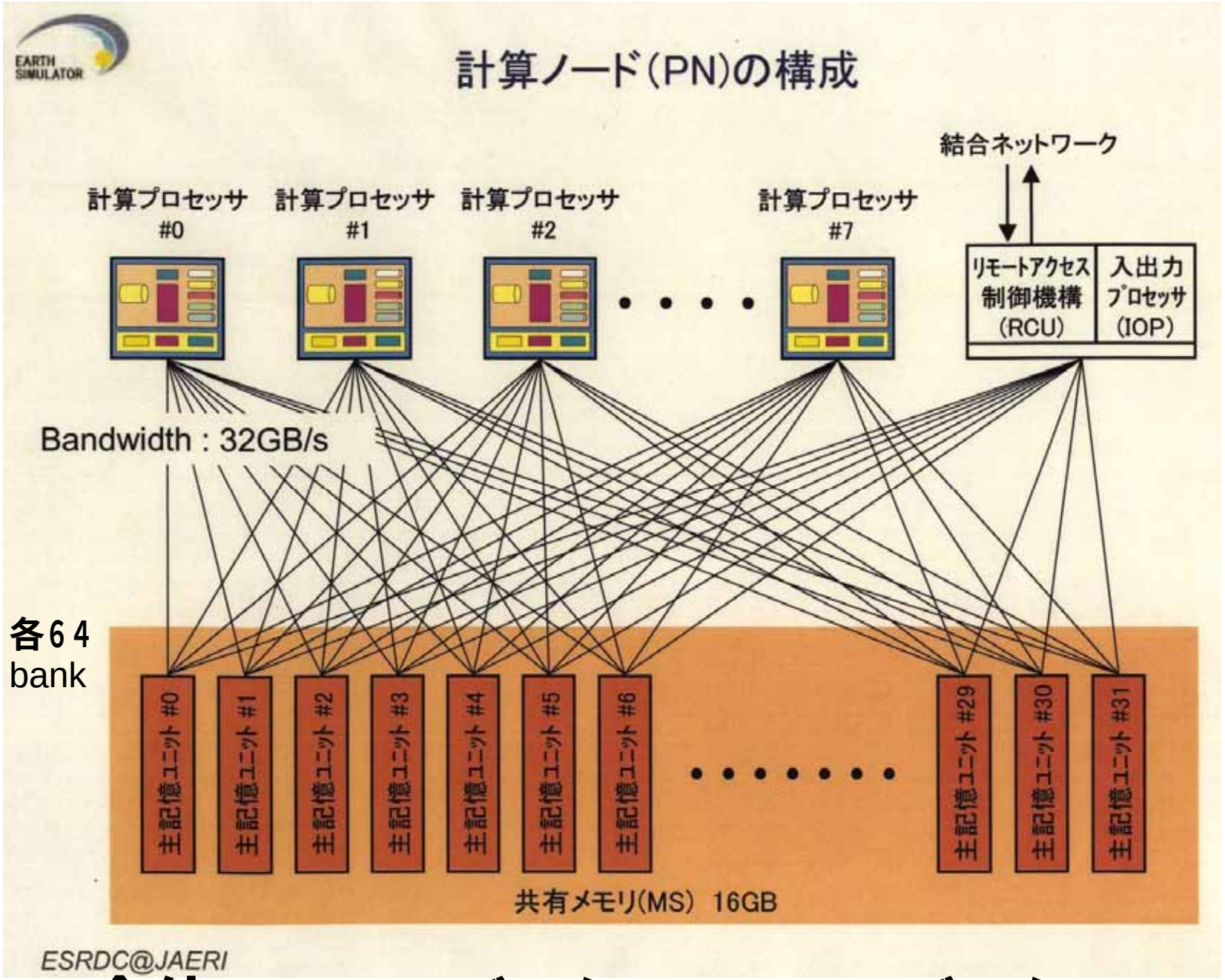
計算プロセッサ(AP)の構成

- 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



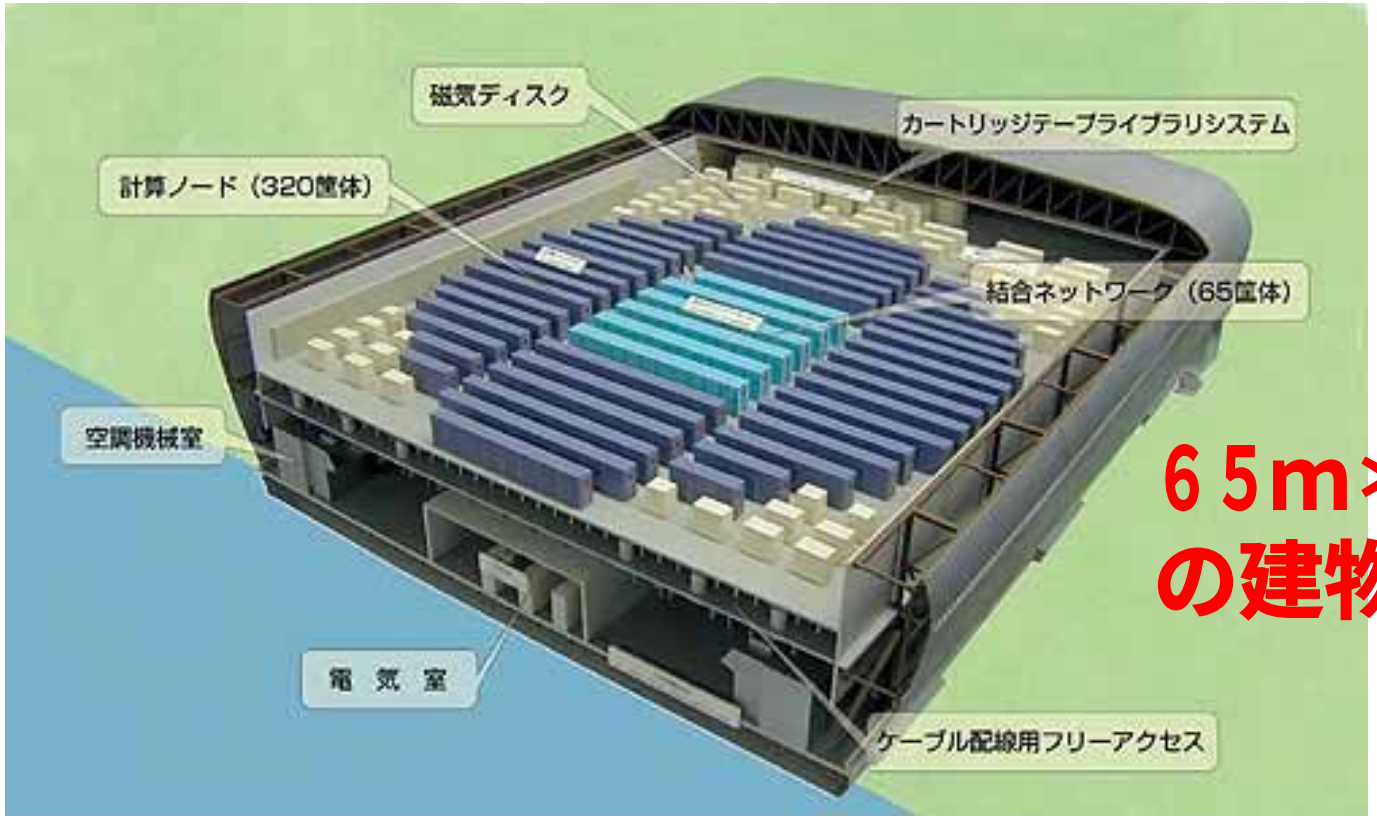
1チップLSI: 8Gflop

- 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

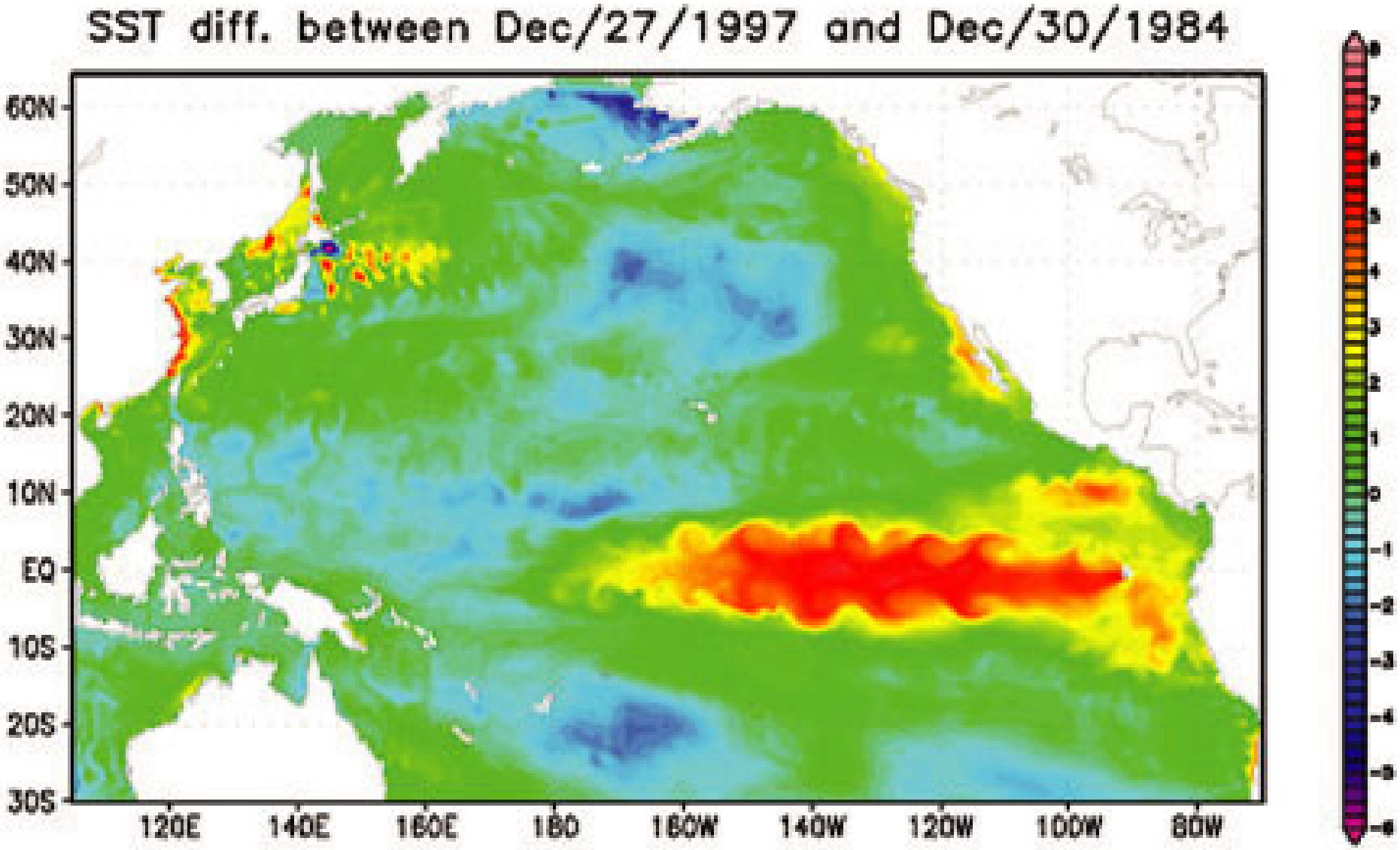
「地球シミュレータ」とは、どんなコンピュータか



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）

7.6.4 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

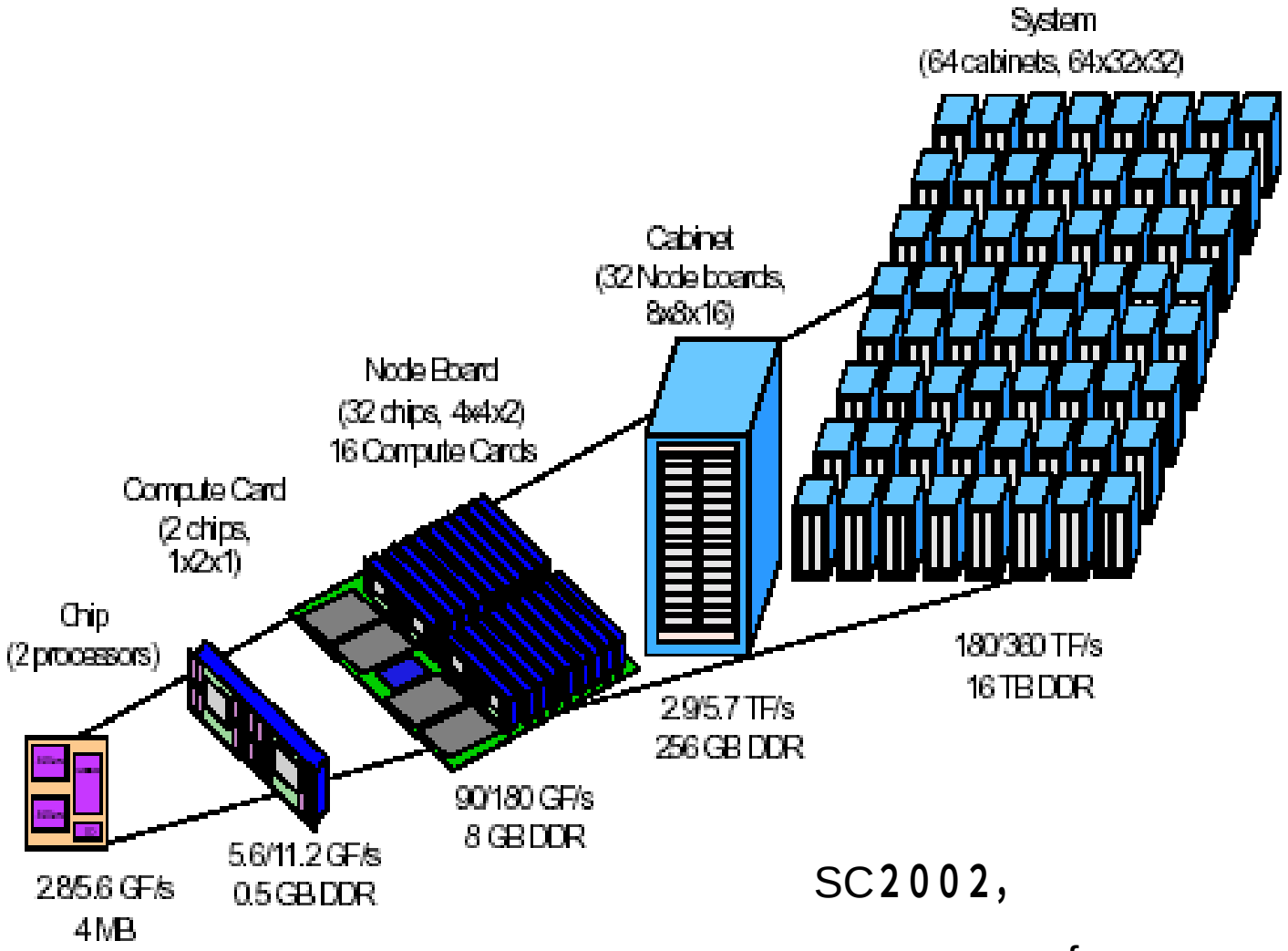
科学技術計算に利用されるスーパーコンピューター（スパコン）で、日米間の世界最速争いに拍車が掛かっている。スパコンはヒトのゲノム（全遺伝情報）を活用した医薬研究などに用途が広がってきた上、国防上も重要な役割を果たすためだ。最速マシンを日本製に奪われた米国側がナンバーワン奪回へ躍起となっている。

「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。



PowerPC４４０

２コア、内１つは
通常通信に使用

Figure 1: BlueGene/L packaging.

SC２００２,

www．sc-conference.org/sc2002/

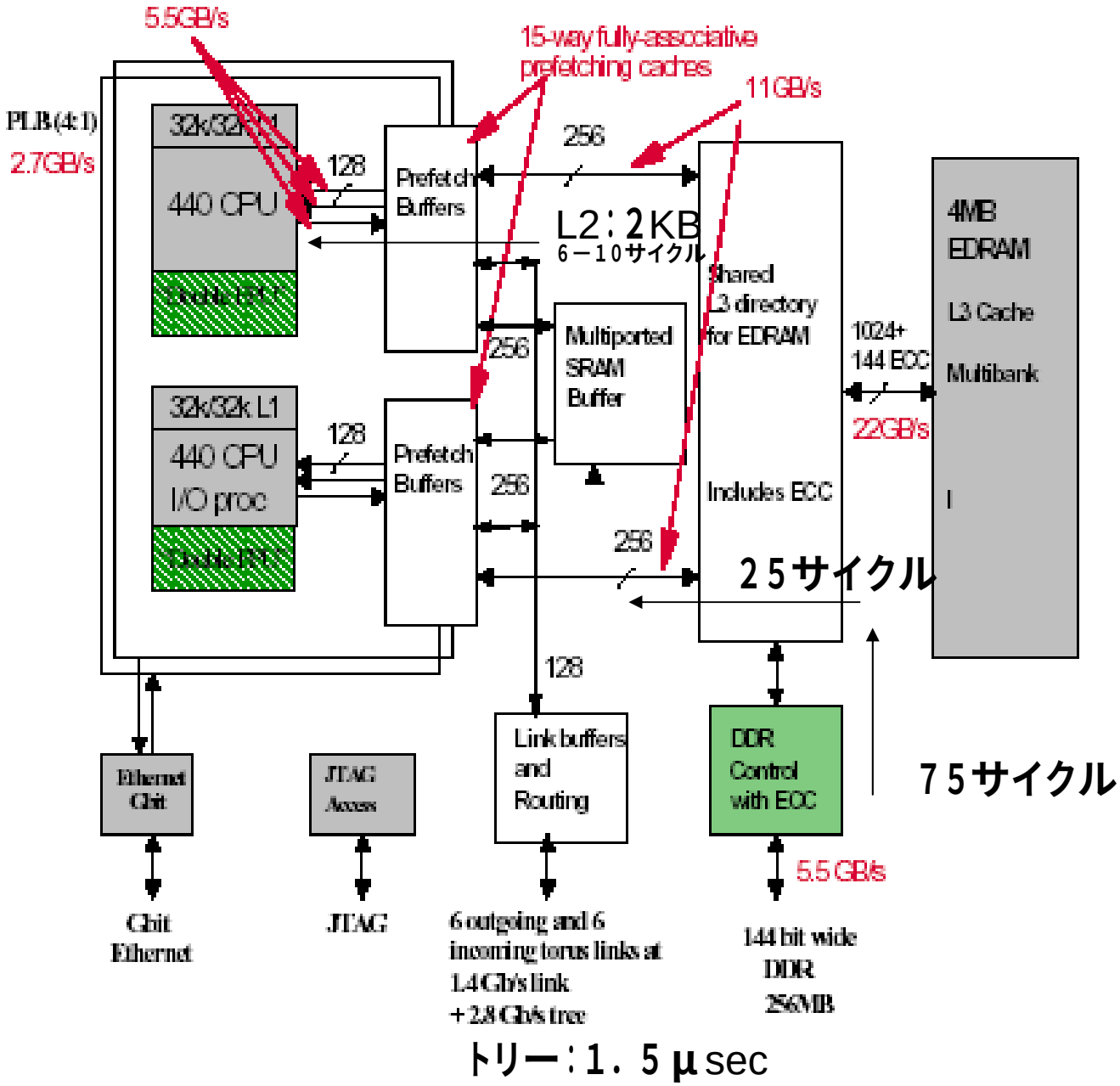


Figure 3: BlueGene/L node diagram. The bandwidths listed are targets.

7.6.5京（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
- 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

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

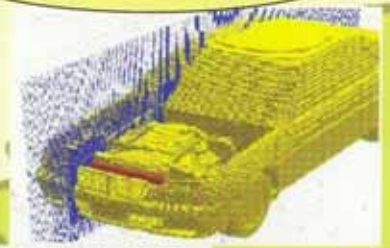
体制：国の責任で設備の整備から運用まで一体的に推進する。また装置の開発・運用を行うに当り、産学官の様々な組織から最も適したところを選択し、そのポテンシャルを活用する。

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　触媒設計

提供：（独）物質・材料研究機構

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

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

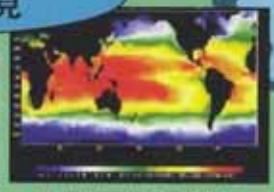
レーザー
反応解析

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

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

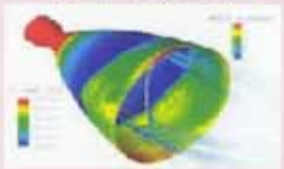
提供：国立天文台

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。

汎用京速計算機のソフトウェア開発

システムソフトウェア（※1）

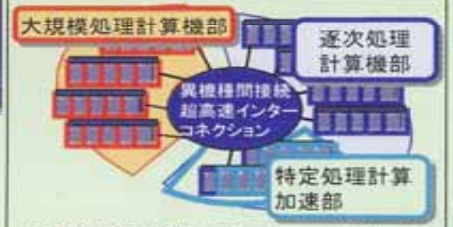
各地に散在する実験装置、データベース、スパコンを自在にどこからでも利用可能 ⇒膨大なデータの効率的利用のため、スパコンの性能を最大限活用


電子顕微鏡


超巨大データベース

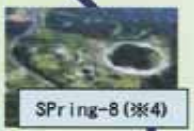
グリッドミドルウェア（※2）
（遠隔利用と外部資源接続支援）
平成15～22年度

異機種統合ソフトウェア
平成18～22年度



汎用京速計算機システム　ハードウェア


NMR（※3）


SPring-8（※4）


各地のスパコン


天文台

汎用京速計算機の研究開発体制

平成17年度後半に利用分野毎の専門家チームを発足予定

- リーダー（全体総括）
  - リーダー補佐（アプリケーション総括）
    - 4領域専門家グループ
      - ・ナノテクノロジー
      - ・ライフサイエンス
      - ・エンジニアリング
      - ・フロンティア　他
    - 利用促進協議会（仮称）
  - リーダー補佐（システムソフトウェア（OS）系）
  - サブリーダー（システムソフトウェア（OS）系総括）
    - システムソフトウェア専門家グループ

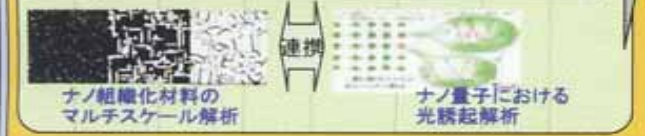
※1：ハードウェアを直接制御したり、システム使用者を支援するソフトウェア
※2：「超高速コンピュータ網形成プロジェクト（National Research Grid Initiative:NAREGI）」。平成15年度よりグリッドミドルウェアとナノシミュレーションソフトウェアの開発を進めている。
※3：Nuclear Magnetic Resonance（核磁気共鳴）。物質の構造を同定するのに用いる装置。
※4：兵庫県播磨科学公園都市にある第三世代の大型放射光施設。

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：平成15～22年度

ナノ新材料・新機能（新半導体材料等）を創出するなど最先端の知的ものづくりを実現するため、ナノ材料系全体統合シミュレーション基盤ソフトウェアの研究開発を行う。NAREGIの成果をベースに開発を行う。

- ・ナノ分野基礎科学の革新的方法論の開発
- ・電子・原子・分子から、分子複合デバイスに至る多種多様な材料系シミュレーションソフトウェアの達成、統合
- ・多種多様なアプリケーションソフトウェア間の容易なデータ交換と入出力ツール



- 化学材料
- 医薬品
- 化粧品
- 磁気ナノデバイス
- 光ナノデバイス

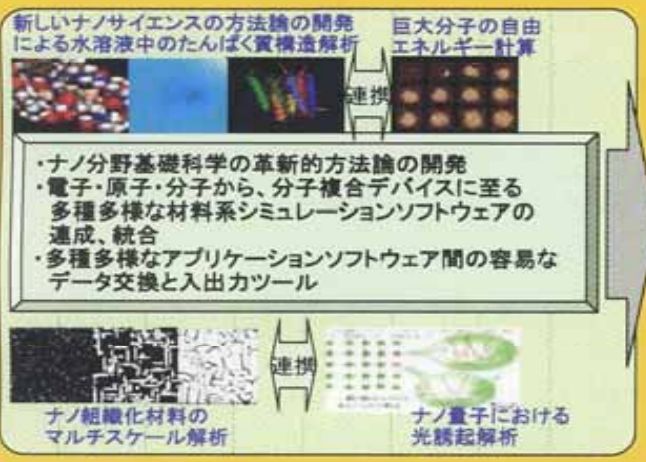
次世代生命体統合シミュレーション：平成18～24年度

テーラーメード医療・創薬などを実現するため、遺伝子レベルから細胞、循環器、人体スケールの個々の要素から全体に至るまで人間系を最適に解析可能な総合シミュレーション基盤ソフトウェアの研究開発を行う。

グランドチャレンジアプリケーション

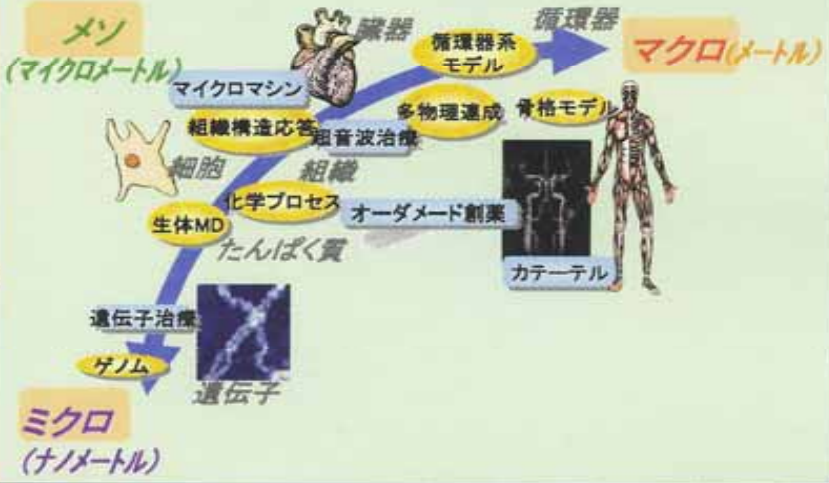
次世代ナノ統合シミュレーション：平成15～22年度

ナノ新材料・新機能（新半導体材料等）を創出するなど最先端の知的ものづくりを実現するため、ナノ材料系全体統合シミュレーション基盤ソフトウェアの研究開発を行う。NAREGIの成果をベースに開発を行う。



次世代生命体統合シミュレーション：平成18～24年度

テーラーメード医療・創薬などを実現するため、遺伝子レベルから細胞、循環器、人体スケールの個々の要素から全体に至るまで人間系を最適に解析可能な総合シミュレーション基盤ソフトウェアの研究開発を行う。



革新的シミュレーションソフトウェアの研究開発（※）：平成17年度～平成19年度

マルチスケール連成シミュレーション

研究代表者：加藤千幸教授（東大生研）

・戦略的革新シミュレーションソフトウェアの研究開発

地球シミュレータ等の超高速コンピュータ上で稼働する各種シミュレーションソフトウェアを、東大生研を中核拠点に産学官連携で普及・事業化も視野に入れて開発する。

グランドチャレンジアプリケーションに成果を展開

※：「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。

次世代高精度・高分解能シミュレーション技術の開発（※）：平成17年度～平成23年度

複数の現象が相互に影響しあうようなマルチスケール・マルチフィジックス現象のシミュレーションを実現する効率的な計算手順を確立し、複雑な工業製品の設計・試作などの先端シミュレーション技術を、我が国最先端のコンピューティング環境を駆使して開発することを目的とした研究。
現在、研究課題を選定中。（８月中旬に決定予定）

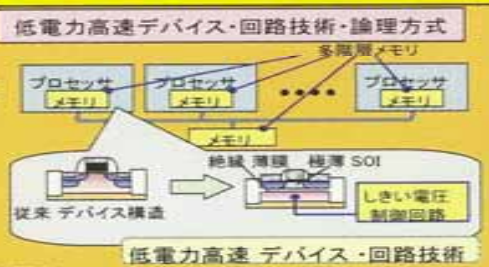
※：科学技術振興機構「戦略的創造研究推進事業」の一戦略目標下に研究領域として、「マルチスケール・マルチフィジックス現象の統合シミュレーション」を設定。

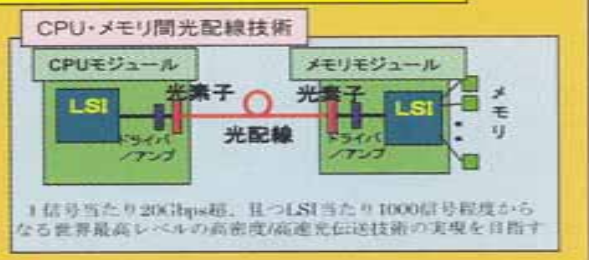
汎用京速計算機のハードウェア開発

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(※) : 平成17～19年度

①システムインターコネクト技術
（九大、富士通）
②内部結合網IP化による実行効率最適化方式
（東大、慶大、アラクサラネットワークス）
③低電力高速デバイス・回路技術・論理方式
（日立製作所、東大、筑波大）
④CPU・メモリ間光配線技術
（日本電気、東工大）





※:「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。

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(※) : 平成17～23年度

消費電力あたりの処理性能を100倍から1000倍にする超低消費電力技術の確立を目指すための基礎研究。現在、研究課題を選定中。（8月中旬に決定予定）

※: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。

汎用京速計算機の研究開発体制

平成17年度後半にプロジェクトを推進するための専門家チームを発足予定

- リーダー（全体総括）
  - リーダー補佐（ハードウェア・ネットワーク系）
  - サブリーダー（ハードウェア・ネットワーク系総括）
    - ハードウェア・ネットワーク専門家グループ

要素技術開発状況

要素技術開発成果

設計 : 平成18～19年度

利用分野チームの要求調査

利用分野での性能要求を調査し整理する。

ハードウェア仕様検討

利用分野からの性能要求調査、性能見積もり、要素技術開発状況から見た実現性等を検討し、高性能を実現するための最適なハードウェア仕様を決定

性能見積もり

利用分野毎のプログラムの性能見積もり行う

ハードウェア仕様

汎用京速計算機の研究開発体制

平成17年度後半にプロジェクトを推進するための専門家チームを発足予定

- リーダー（全体総括）
  - リーダー補佐（ハードウェア・ネットワーク系）
  - サブリーダー（ハードウェア・ネットワーク系総括）
    - ハードウェア・ネットワーク専門家グループ

要素技術開発状況

要素技術開発成果

設計：平成18～19年度

利用分野チームの要求調査

利用分野での性能要求を調査し整理する。

ハードウェア仕様検討

利用分野からの性能要求調査、性能見積もり、要素技術開発状況から見た実現性等を検討し、高性能を実現するための最適なハードウェア仕様を決定

性能見積もり

利用分野毎のプログラムの性能見積もり行う

ハードウェア仕様

実装技術：平成20～21年度

約3億個の回路をLSI、プリント基板などに織り込む

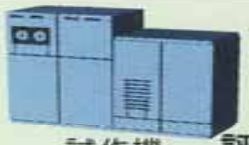
回路設計

試作

LSI

プリント基板

試作機組立



試作機

評価

試作機を組み上げ、実際のプログラム等を走行させ、システムの安定性を確認

製作開始

製作：平成21～22年度

システム全体の製作
特定処理計算加速部の完成

汎用京速計算機システム

システム強化：平成23～24年度

大規模処理計算機部と逐次処理計算機部のシステム強化

性能評価：平成22年度

Linpack（10ペタFLOPS目標）、
HPC CHALLENGEでの性能評価

総合評価：平成23～24年度

グランドチャレンジアプリケーション、「革新的シミュレーションソフトウェアの研究開発」※等での実問題を用いた総合評価

※：研究開発プロジェクト「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。

汎用京速計算機の利用促進

1．ユーザー会（利用促進協議会（仮称））の組織化

趣旨

汎用京速計算機を最大限有効に利用するためには、ユーザーにとって利用し易いシステムであることが必須。そのため、平成１７年度中に発足予定の開発チームとユーザーの密な情報交換が必要。この双方向の情報交換のために**利用促進協議会(仮称)**を組織し、汎用京速計算機の利用に関する産業界・学会からの要望を取纏め、開発・運用側に意見具申等を行い、更に普及・利用促進も図る。

なお、本利用促進協議会(仮称)は、**既に同様の目的で活動中の計算科学技術関連のユーザー会組織等を統合し、活動を強化する。**

利用想定ユーザーとしては、次の３つの目的に当てはまる機関（企業、大学、公的研究機関）である。

（1）２１世紀の基幹産業を支える最先端の知的ものづくりと社会基盤の整備

ナノテクノロジーを駆使した新材料創出、バイオテクノロジーと医用工学の融合による個人差に応じた合理的な医薬品・医療の実現、設計から製品化までの開発期間・開発コストの大幅な縮小や生産性を飛躍的に向上させる製造プロセス一貫シミュレーション、あるいは、電力、ガス等の安定した社会基盤の整備のためのシミュレーションの応用。

（2）先進的なシミュレーションを駆使した安心・安全な社会の実現

台風や豪雨、地震・津波などの自然災害の正確な予測や都市・地域スケールの災害影響評価を踏まえた、きめ細かな防災対策での施策の立案。

（3） 人類未踏のフロンティア科学技術を探求

ナノサイエンス、ライフサイエンス、環境、原子力、航空・宇宙等の幅広い科学技術のフロンティアを開拓。

参考）既存のユーザー会組織等
・戦略的基盤ソフトウェア産業応用推進協議会（共同議長：小林敏雄自動車技術研究所長、柘植綾夫総合科学技術会議議員）
・NAREGI(※1)研究グリッド産業応用協議会（会長：中村道治日立製作所副社長）
・ＩＴＢＬ(※2)共同実施者・共同研究者等

※1：「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。

※2：IT-Based Laboratory。日本原子力研究所等６つの国内研究機関により平成１２年度より開始された。計算資源、データベース等の情報資源を共有することにより仮想研究環境の構築を目指す。

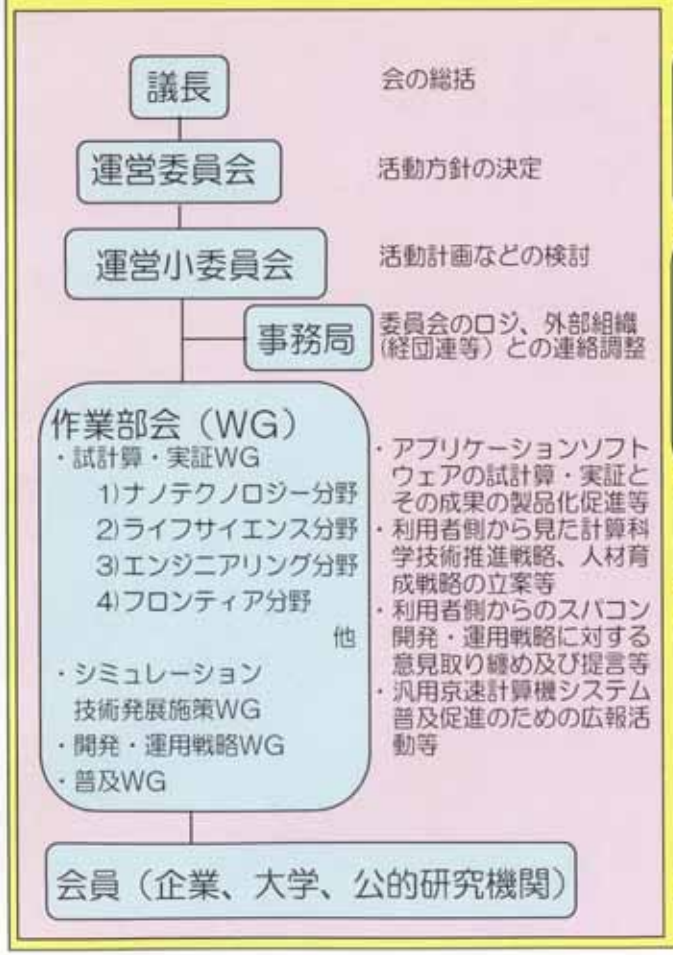
2. 利用促進協議会(仮称)の詳細について

(1) 活動内容

汎用京速計算機システムに対する産業界・学会などの利用者側（企業、大学、公的研究機関）の窓口として、開発・運用側への意見具申、普及・利用推進、情報の共有を図る。

平成１７年度中に組織化

(2) 構成と機能



(3) 利用促進協議会（仮称）参加予定機関

（平成１７年８月時点での想定メンバー１２５社、３３大学、２１研究機関）

ナノテクノロジー、バイオテクノロジー（食品、化学、医薬品など）分野　３５社

旭化成、旭硝子、味の素、出光石油化学、エーザイ、キッセイ薬品工業、キリンビール、杏林製薬、昭和電工、住友化学、住友製薬、ゾイジーン、大正製薬、大鵬薬品工業、東レ、日本たばこ産業、日立金属、富士写真フィルム、マンダム、三菱化学　他

ものづくり（自動車、電機・情報、ソフトウェアなど）分野　８１社

石川島播磨重工業、宇部興産、ＮＥＣ、川崎重工、原子燃料工業、国際電気通信基礎技術研究所、三洋電機、島津製作所、新日本製鉄、住友重機械工業、住友電装、セイコーエプソン、デンソー、東芝、トヨタ自動車、日揮、日産自動車、日立製作所、富士通、古河電工、本田技研、松下電器産業、松下電工、マツダ、三井造船、三菱重工業、三菱電機、村田製作所、リコー、横河電機　他

社会基盤の整備（建設、電力、ガス、鉄道、電話、金融など）分野　９社

関西電力、清水建設、大成建設、竹中工務店、東京ガス、東京電力、日本電信電話　他

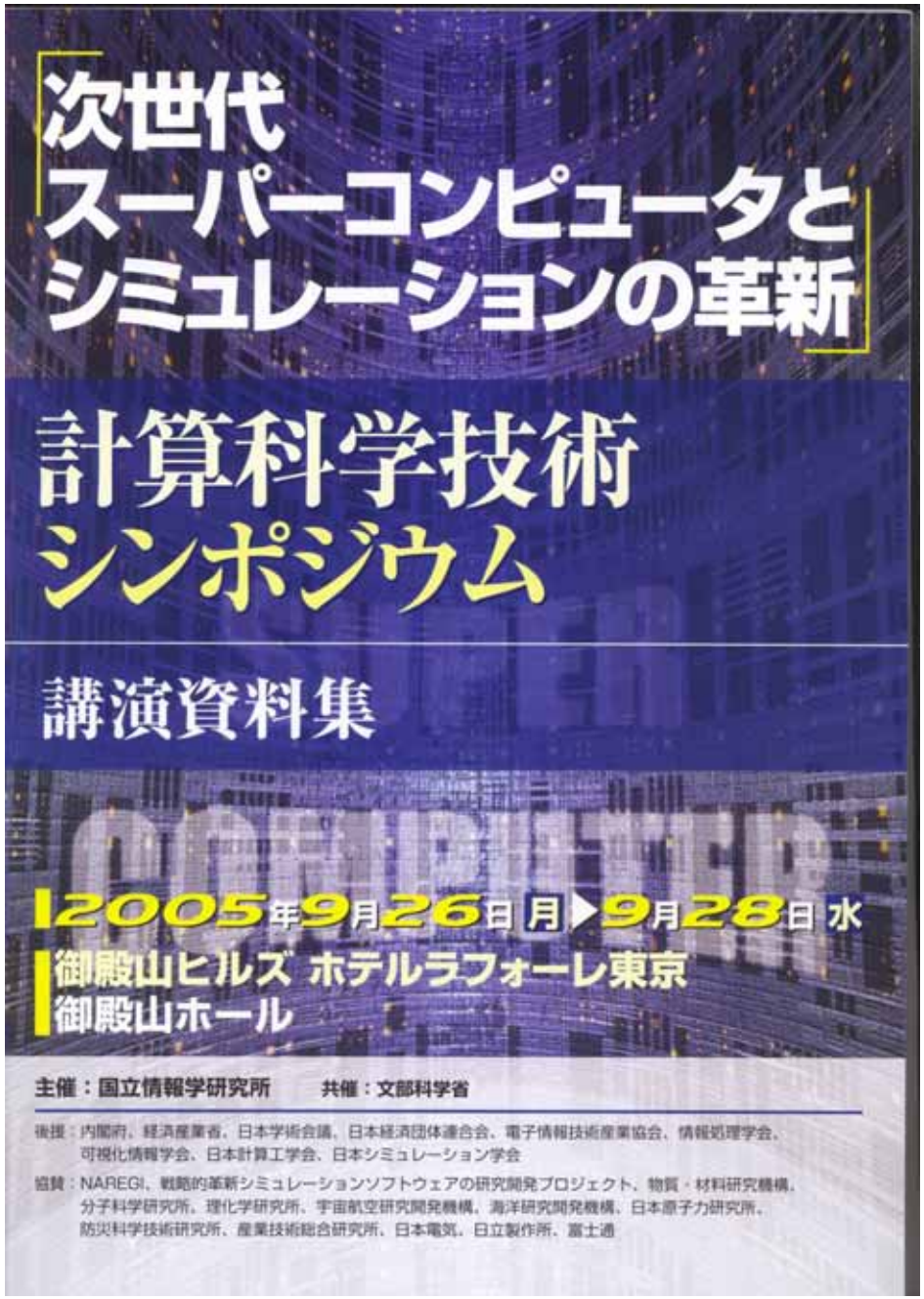
安全、安心な社会の実現とフロンティア科学技術の探求分野

３３大学　２１公的研究機関

＜大学＞大阪大、九大、京大、東工大、東大、東北大、名古屋大、北大　他

＜公的研究機関など＞宇宙航空研究開発機構、海洋研究開発機構、国立医薬品食品衛生研究所、国立環境研究所、国立情報学研究所、産業技術総合研究所、鉄道総合技術研究所、電力中央研究所、日本原子力研究所、物質・材料研究機構、分子科学研究所、防災科学技術研究所、理化学研究所　他

（五十音順）


次世代
スーパーコンピュータと
シミュレーションの革新
計算科学技術
シンポジウム
講演資料集
2005年9月26日 月 ▶ 9月28日 水
御殿山ヒルズ ホテルラフォーレ東京
御殿山ホール
主催：国立情報学研究所　共催：文部科学省
後援：内閣府、経済産業省、日本学術会議、日本経済団体連合会、電子情報技術産業協会、情報処理学会、
可視化情報学会、日本計算工学会、日本シミュレーション学会
協賛：NAREGI、戦略的革新シミュレーションソフトウェアの研究開発プロジェクト、物質・材料研究機構、
分子科学研究所、理化学研究所、宇宙航空研究開発機構、海洋研究開発機構、日本原子力研究所、
防災科学技術研究所、産業技術総合研究所、日本電気、日立製作所、富士通

7.6.6グリッドコンピューティング

高速グリッドコンピューティング環境を構築する

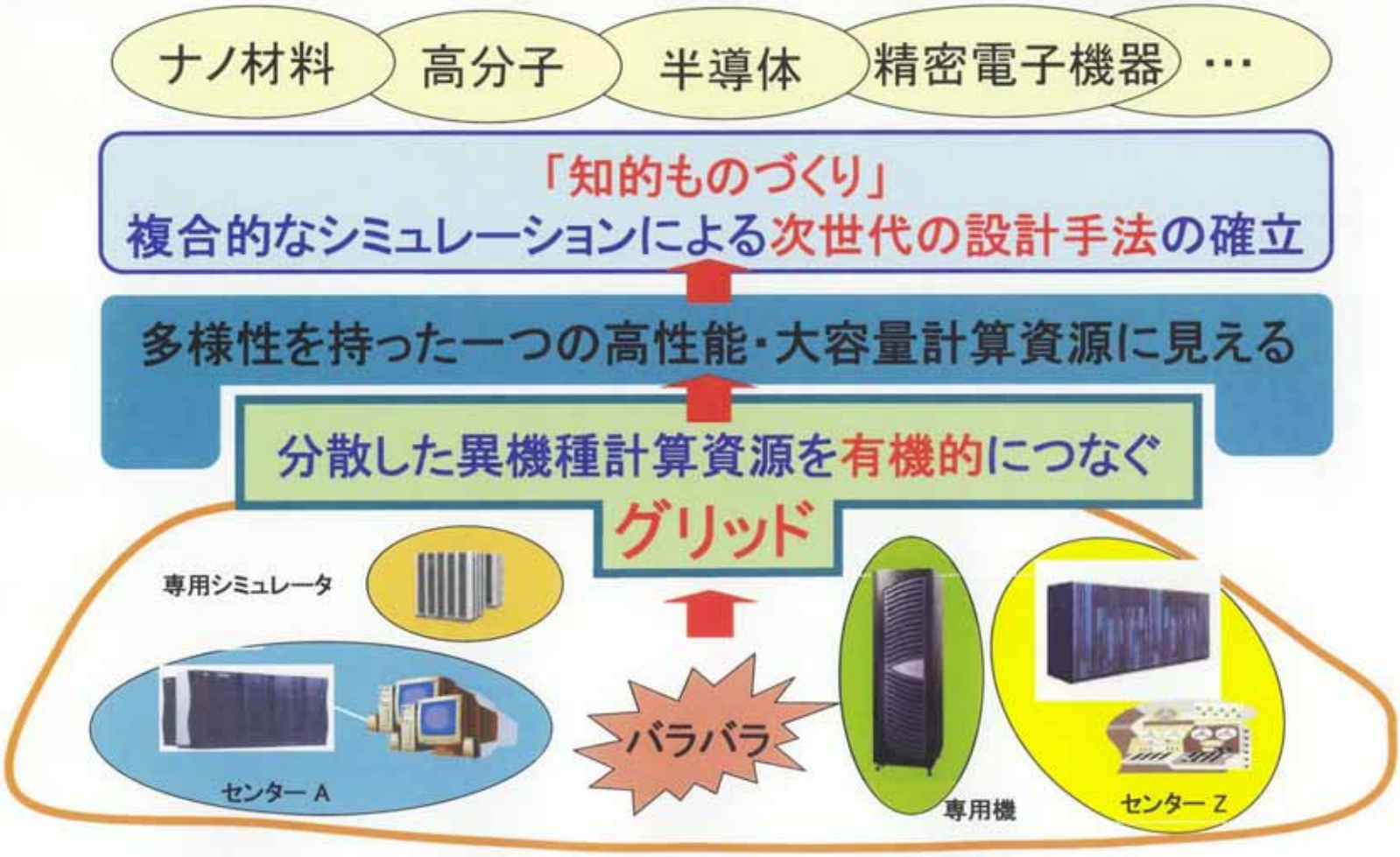
『超高速コンピュータ網形成プロジェクト（ＮＡＲＥＧＩ）』
中間成果報告会

主催：文部科学省

２００５年7月11日（月）　９：３０～１５：００

国立情報学研究所　学術総合センター

背景

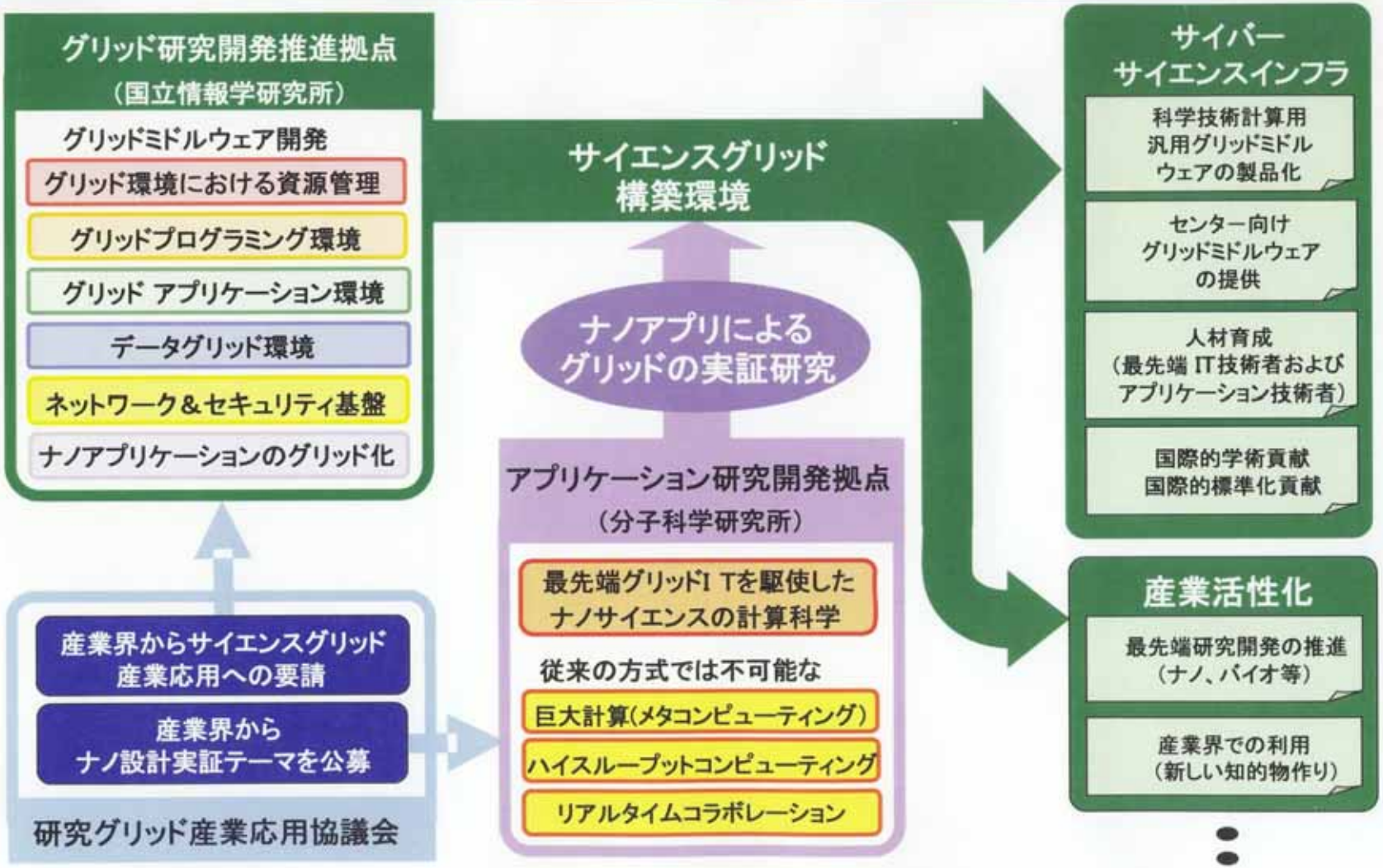

ナノ材料
高分子
半導体
精密電子機器
…
「知的ものづくり」
複合的なシミュレーションによる次世代の設計手法の確立
多様性を持った一つの高性能・大容量計算資源に見える
分散した異機種計算資源を有機的につなぐ
グリッド
専用シミュレータ
バラバラ
センター A
専用機
センター Z

達成すべき目標

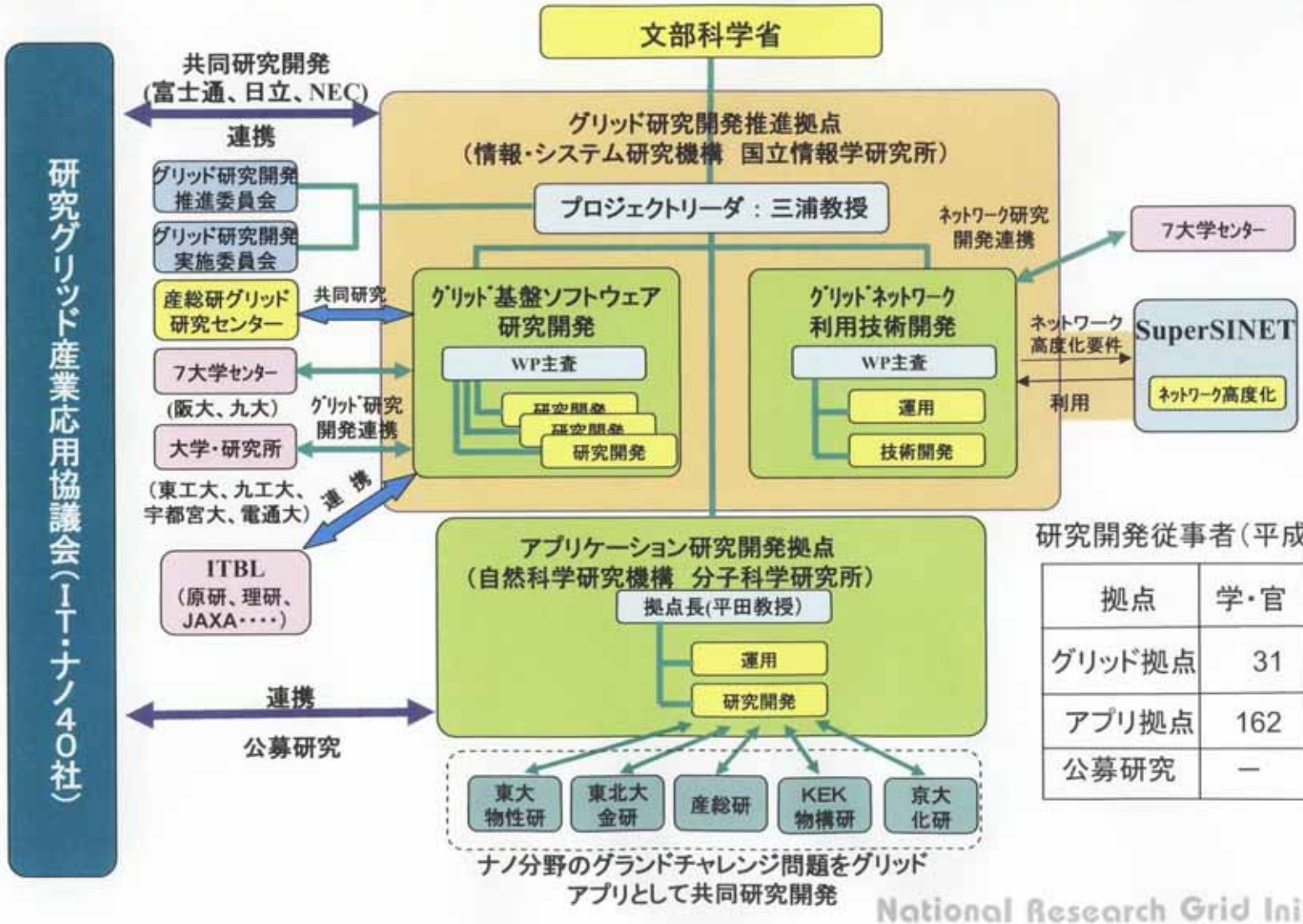
① 学術研究、産業界における研究開発を対象に計算機資源をグリッドで接続し、大規模なシミュレーション、ハイスループット処理、マルチスケール・マルチフィジックスの連成解析が可能な、100Tflops級のサイエンスグリッドを実現するグリッド基盤ソフトウェアを開発。

② 開発したグリッド基盤ソフトウェアが実用的に有効であることをナノサイエンス分野での実証研究で確認。

研究開発の取り組み方と成果の利用



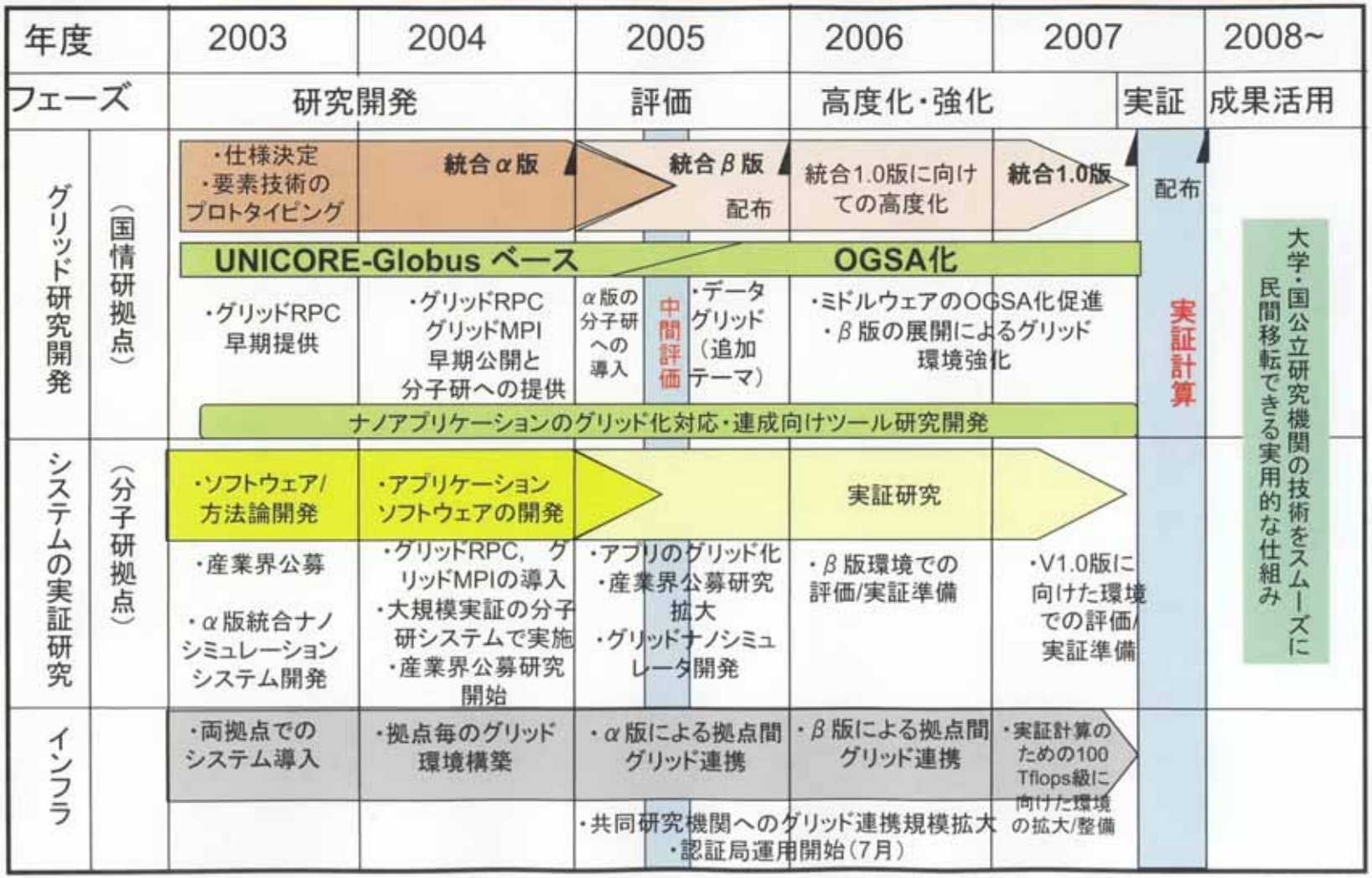
研究開発体制図



研究開発従事者（平成16年度）

拠点	学・官	産
グリッド拠点	31	110
アプリ拠点	162	5
公募研究	―	45

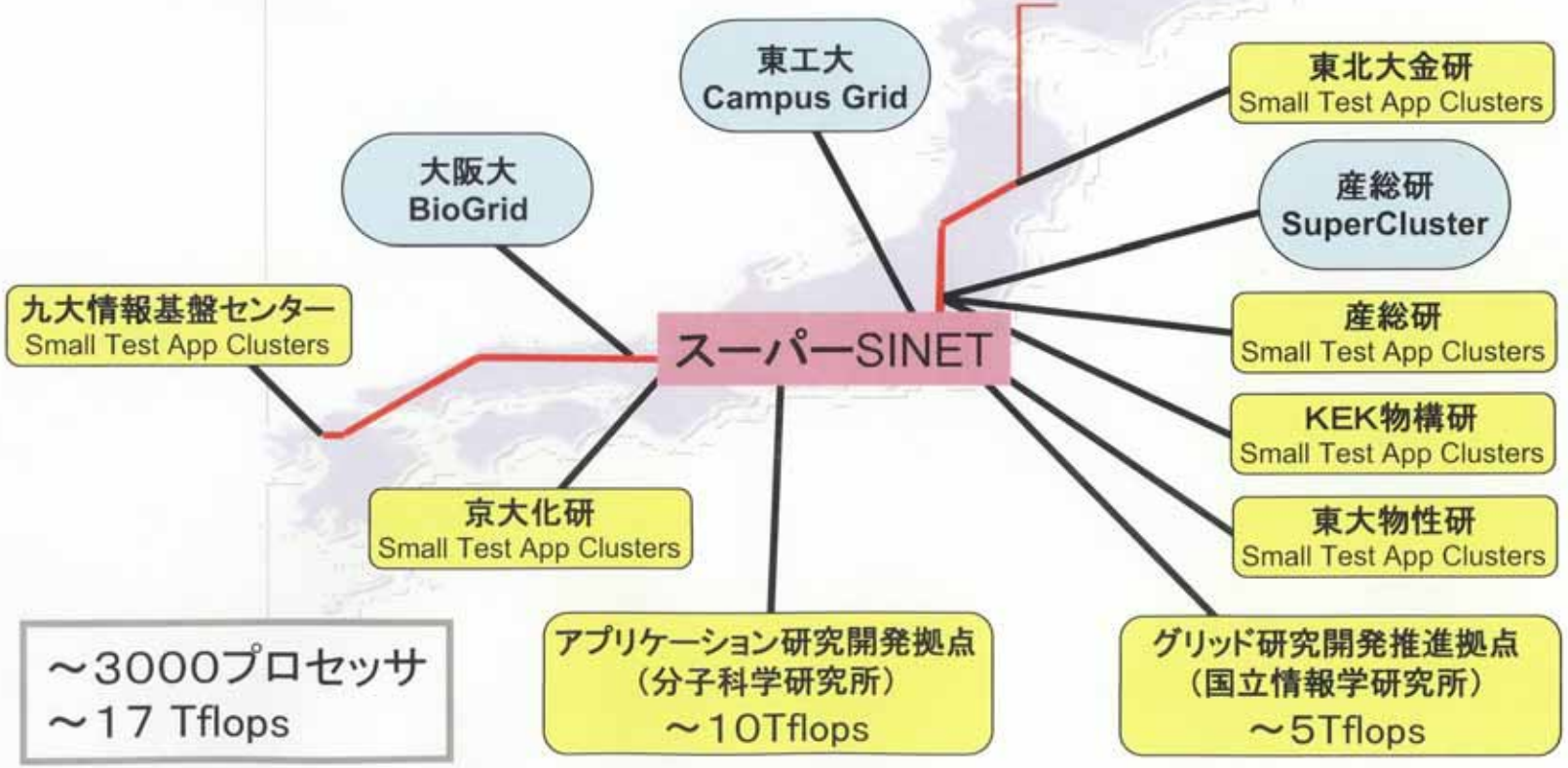
プロジェクト全体の進め方



OGSA: Open Grid Services Architecture

NAREGI 広域分散テストベッドインフラの構築

グリッドミドルウェア開発、ナノ分野のグリッド実証研究を考慮したヘテロ、広域分散環境のテストベッドを構築



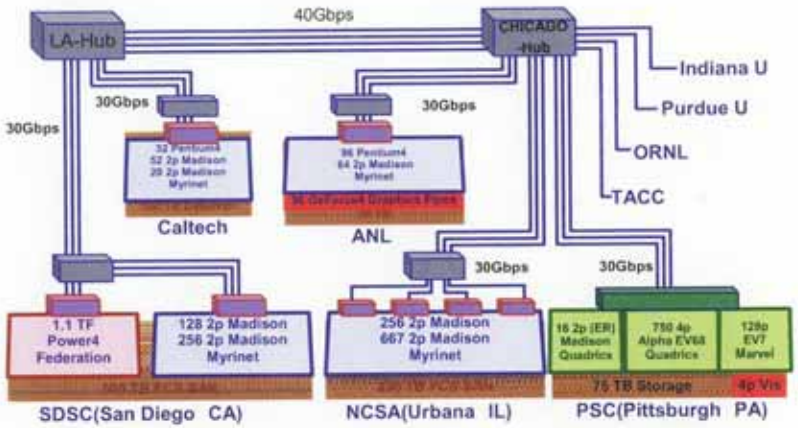
国際グリッドプロジェクトとの比較

欧州連合(EU)のEGEEプロジェクト

予算規模:70 M Euro(2004-2007)【93億円】
参加機関:>70
計算機資源: 8768 CPUs(Mid 2005)
アプリ分野:高ｴﾈﾙｷﾞｰ物理学向けデータグリッド
バイオ他
ネットワーク:GEANT−NRENs(10Gbps)
基盤ミドルウェア:g-Lite (Globus ベース)
プロジェクト開始時期:2004年4月
中心的機関: CERN(ジュネーブ)
・運用主体のプロジェクトである

EGEE:Enabling Grids for E-SciencE
CERN: European Organization for Nuclear Research

米国テラグリッド(Extensible Tera-scale Facility)



予算規模:98M$(2001-2003総計)【109億円】
参加機関:>9
計算機資源: 40Tflops
・Teragrid固有:20Tflops
・NSFセンター等接続分:20Tflops
アプリ分野:サイエンス一般
ネットワーク:専用線 (40Gbps)
基盤ミドルウェア:NMIパッケージ (Globus ベース)
プロジェクト開始時期: 2002年4月
運用開始時期:2004年10月
中心的機関: National Science Foundation(NSF)
・運用主体のプロジェクトである

NMI:National Middleware Initiative

National Research Grid Initiative

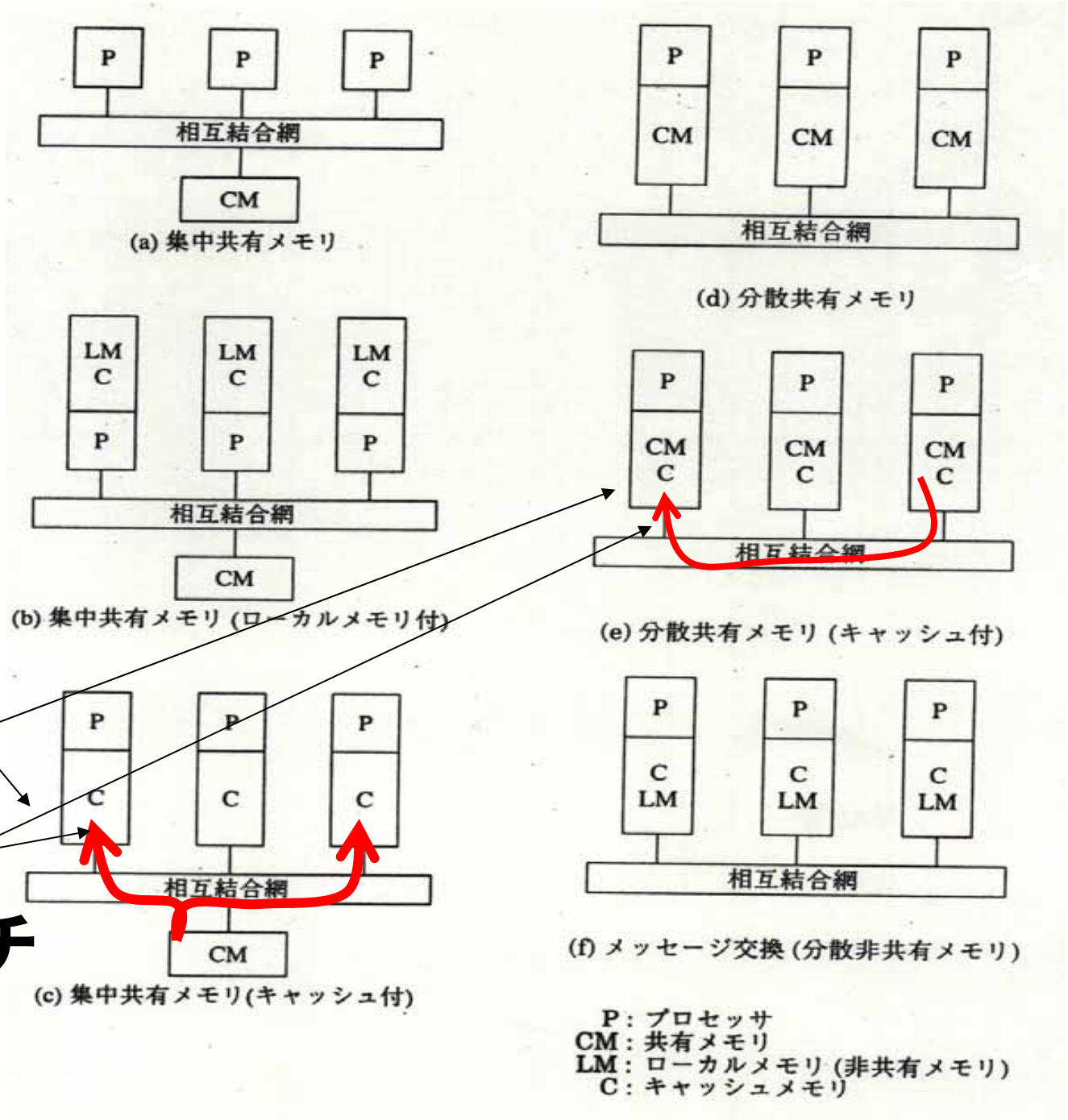
7．7メモリ共有型マルチプロセッサ

重要問題

①キャッシュコヒーレンス問題

②メモリコンシステンシ問題

種々のマルチプロセッサ



キャッシュ
コヒーレンス問題

書込み

7.7.1キャッシュコヒーレンスの分類

(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

7.4.1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

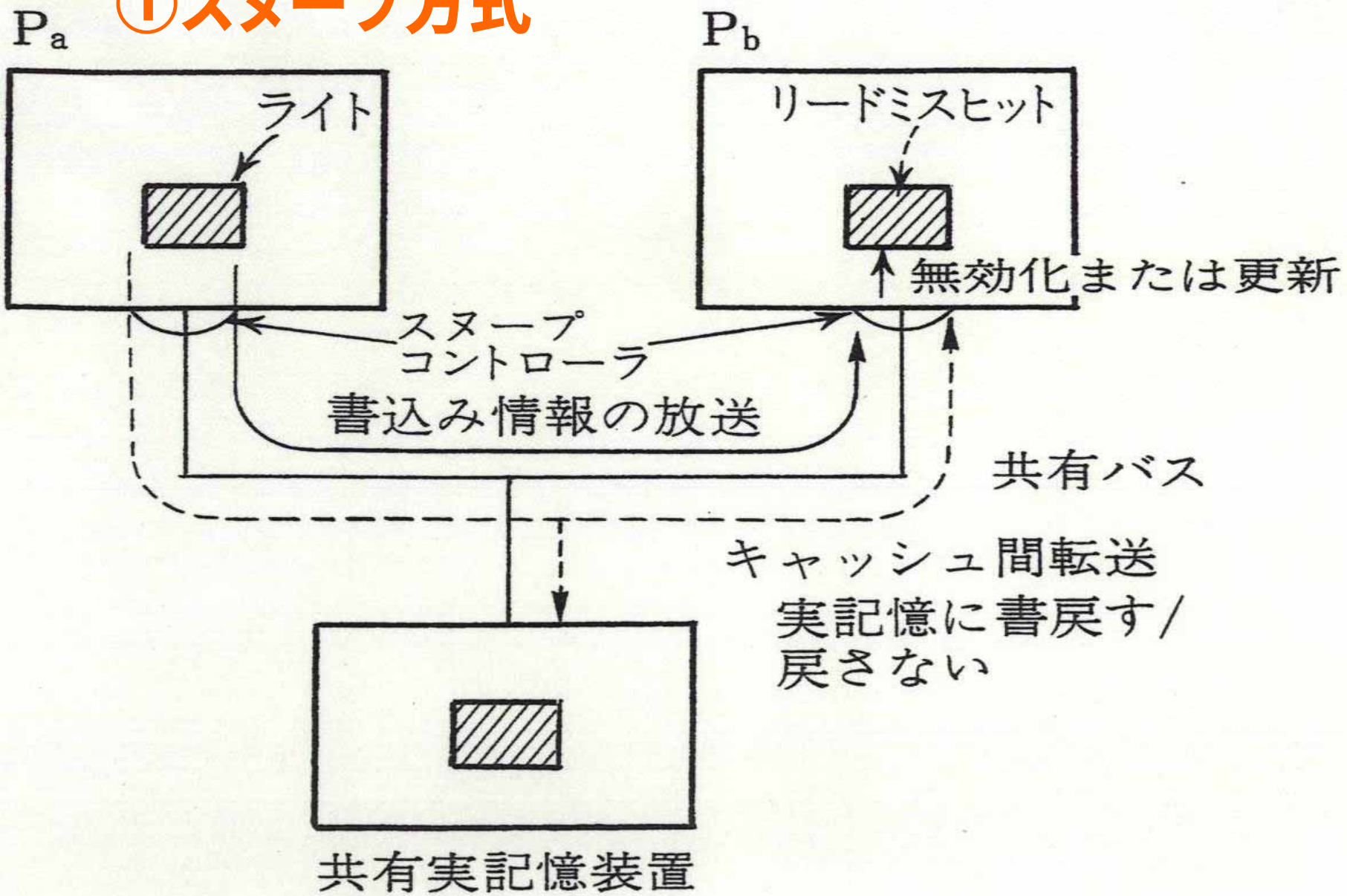
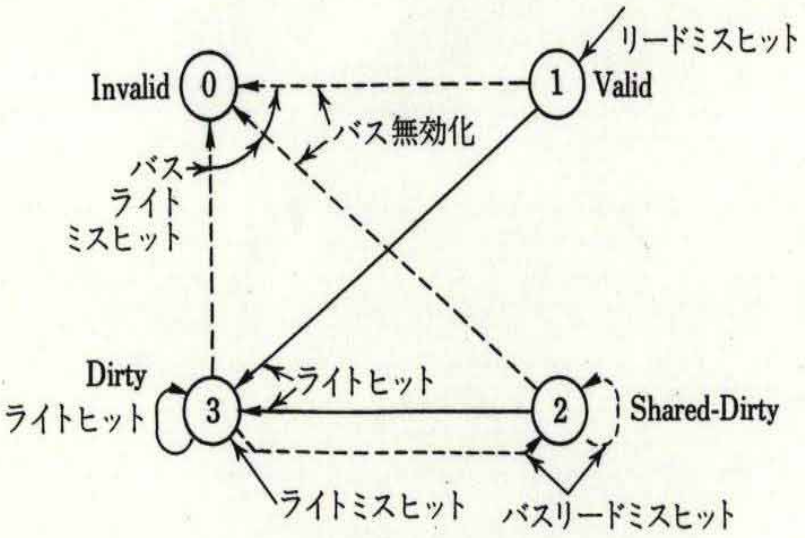


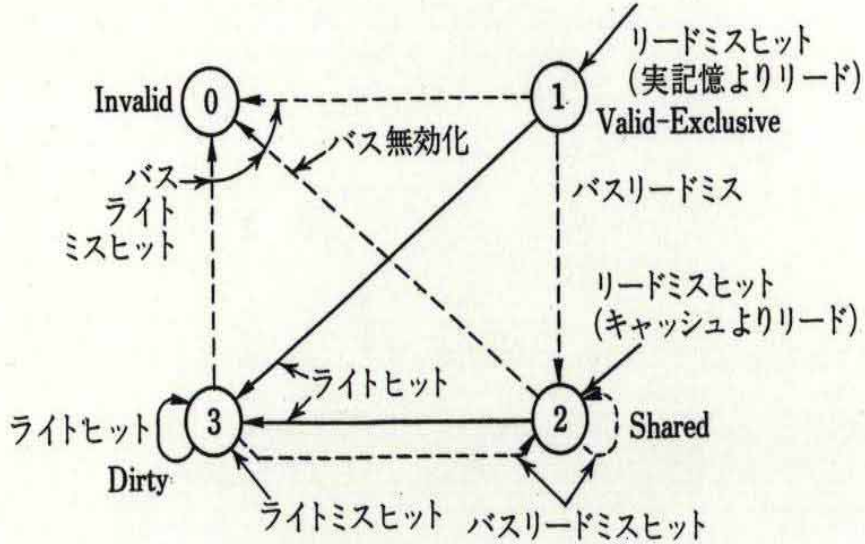
図 4.24　スヌープキャッシュの基本方式

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)

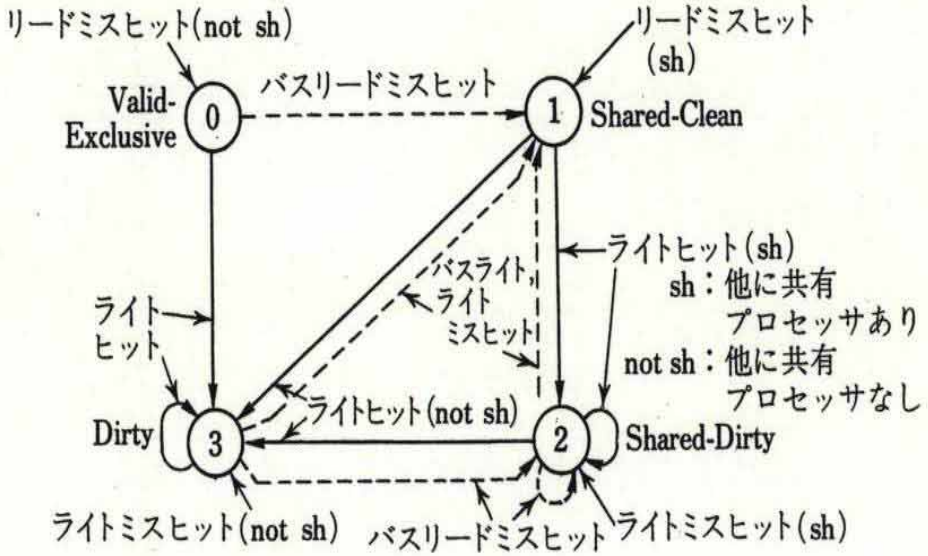


(a) バークレイプロトコル

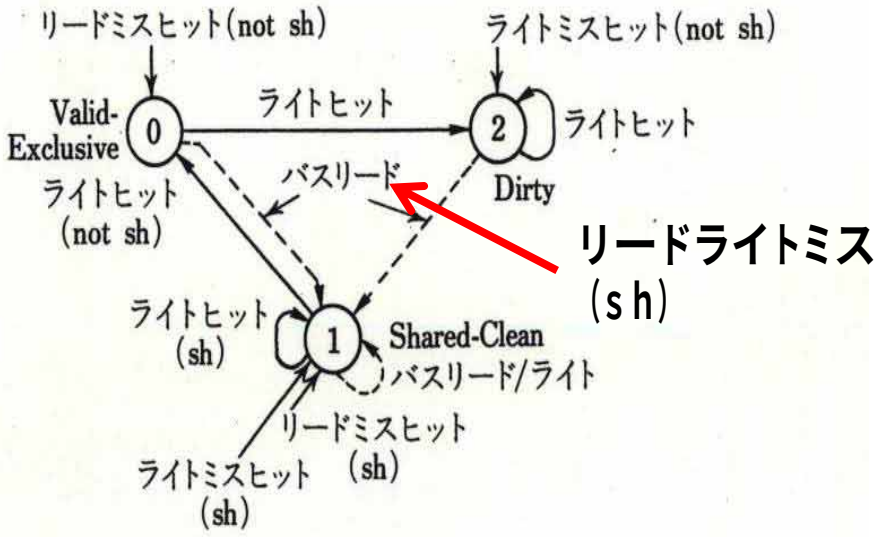


(b) イリノイプロトコル

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

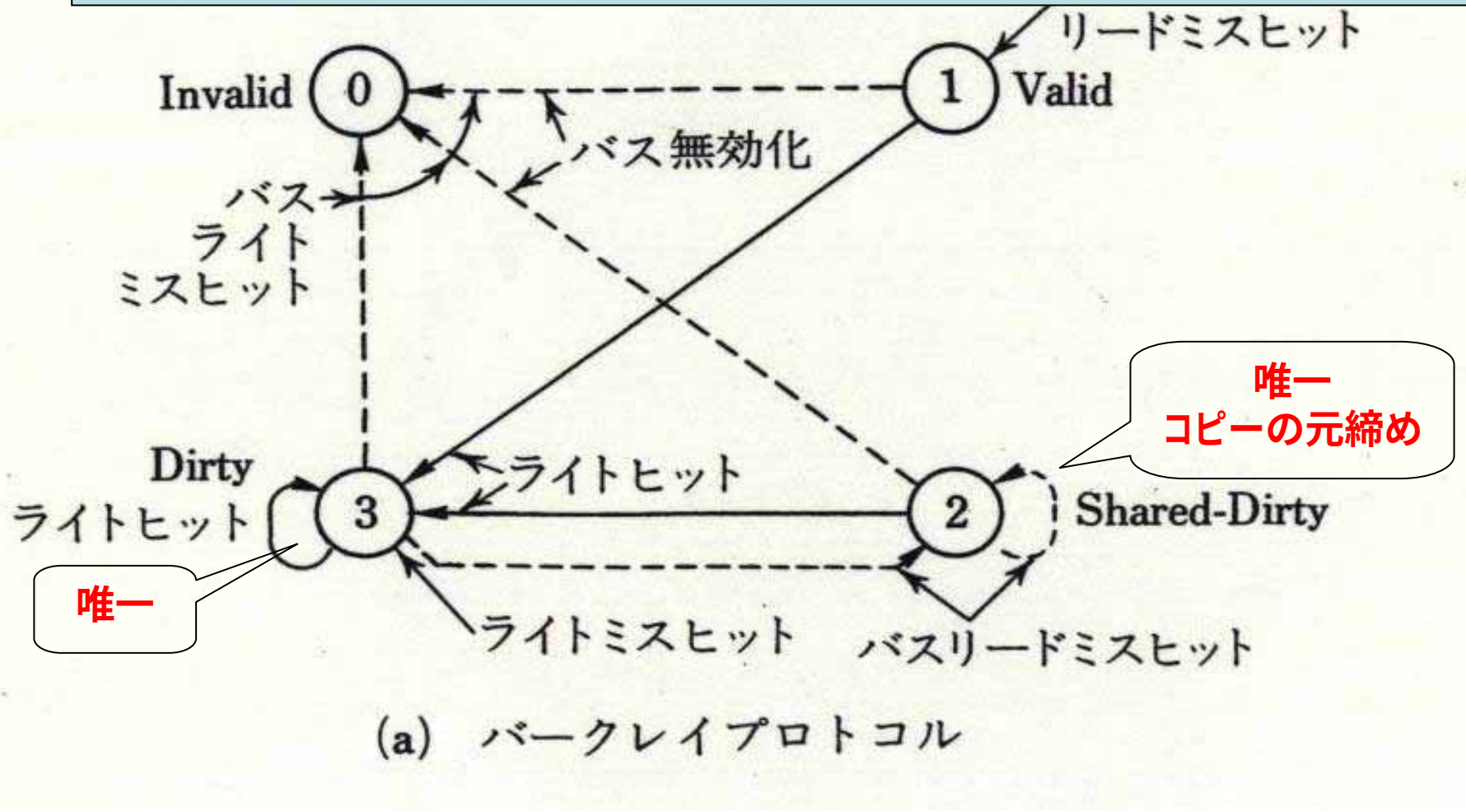
DEC Firefly
Multiprocessor

図 4.25 種々のプロトコル[13)]

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

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

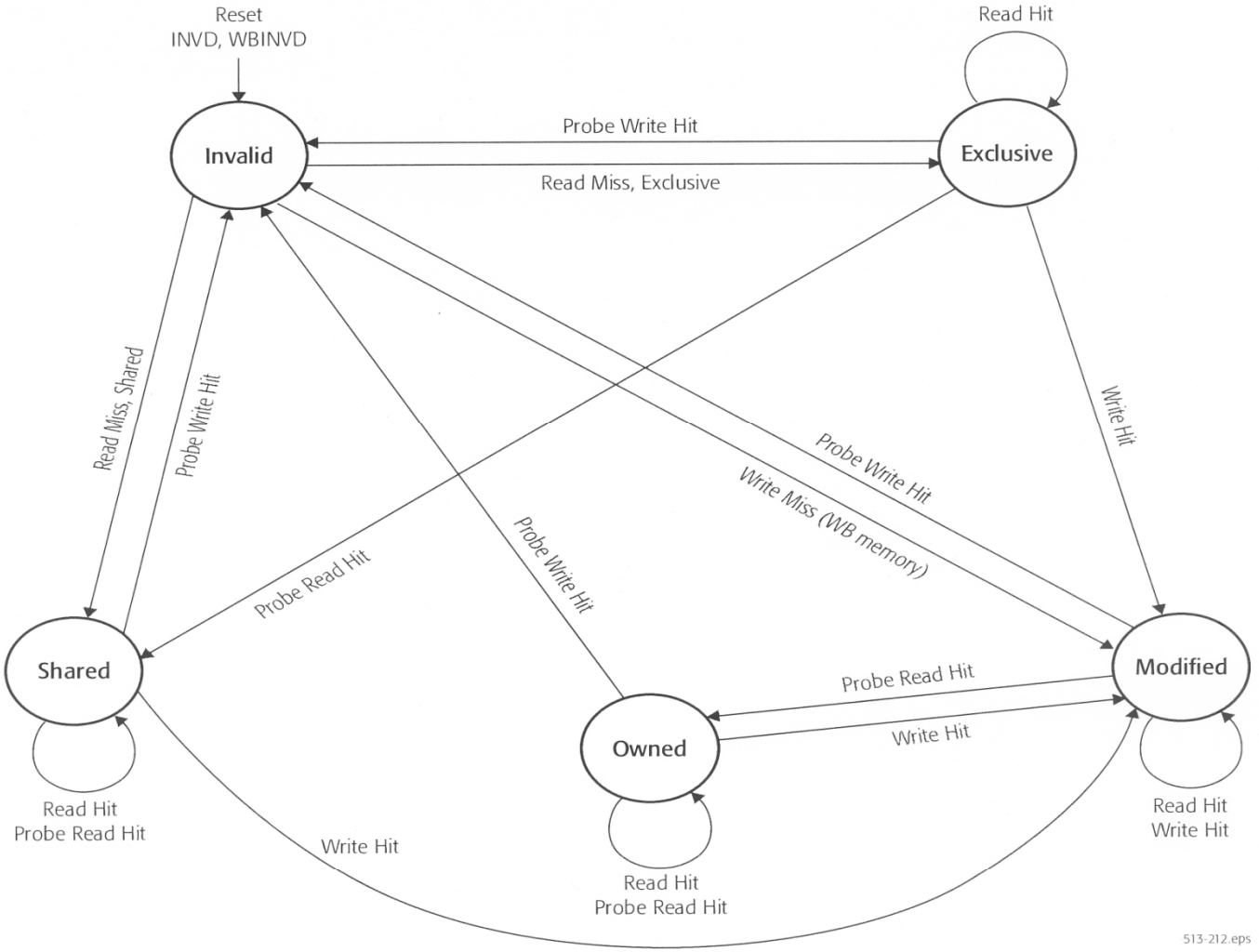


Figure 7-2. MOESI State Transitions

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

②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

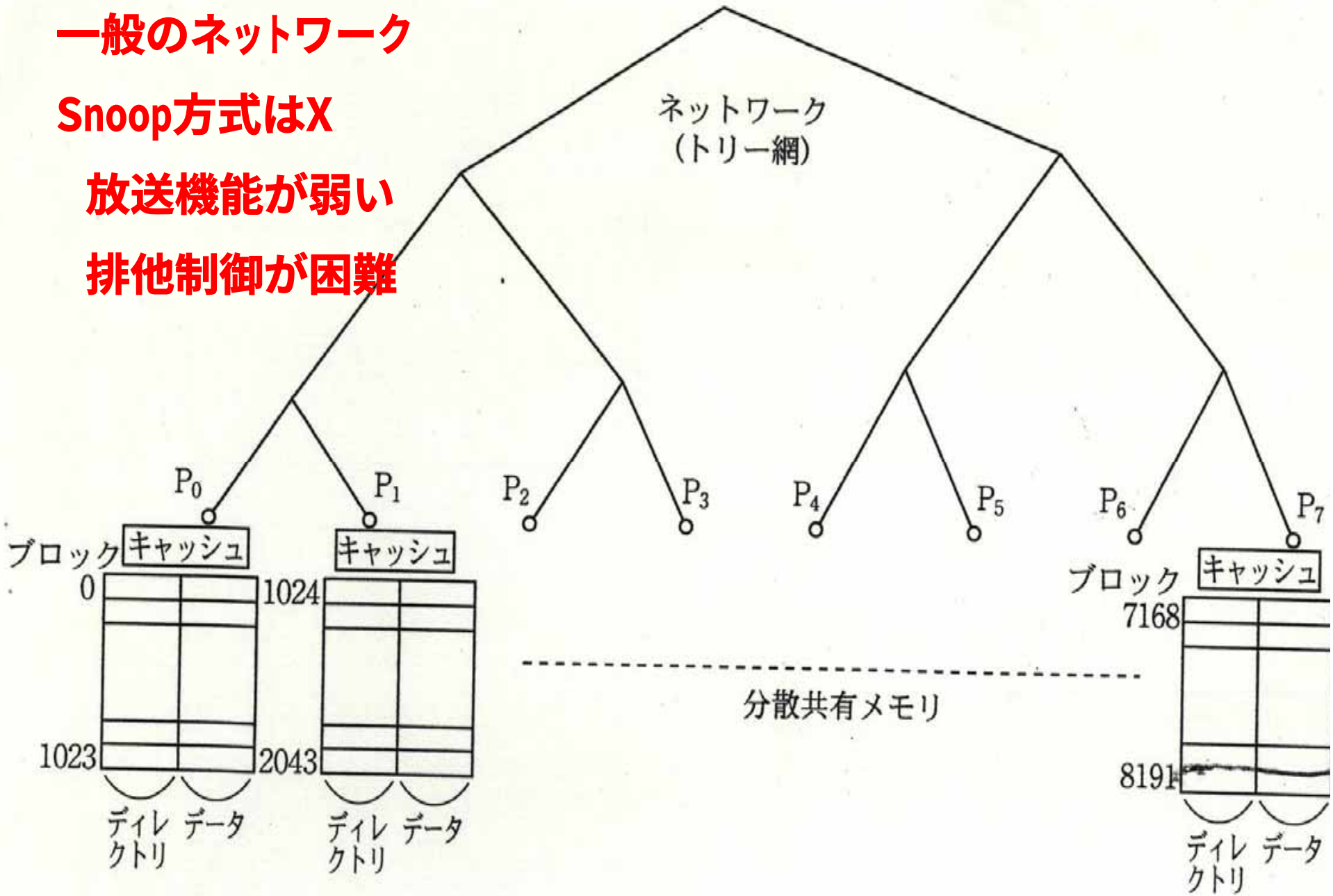


図 4.26 ディレクトリ方式

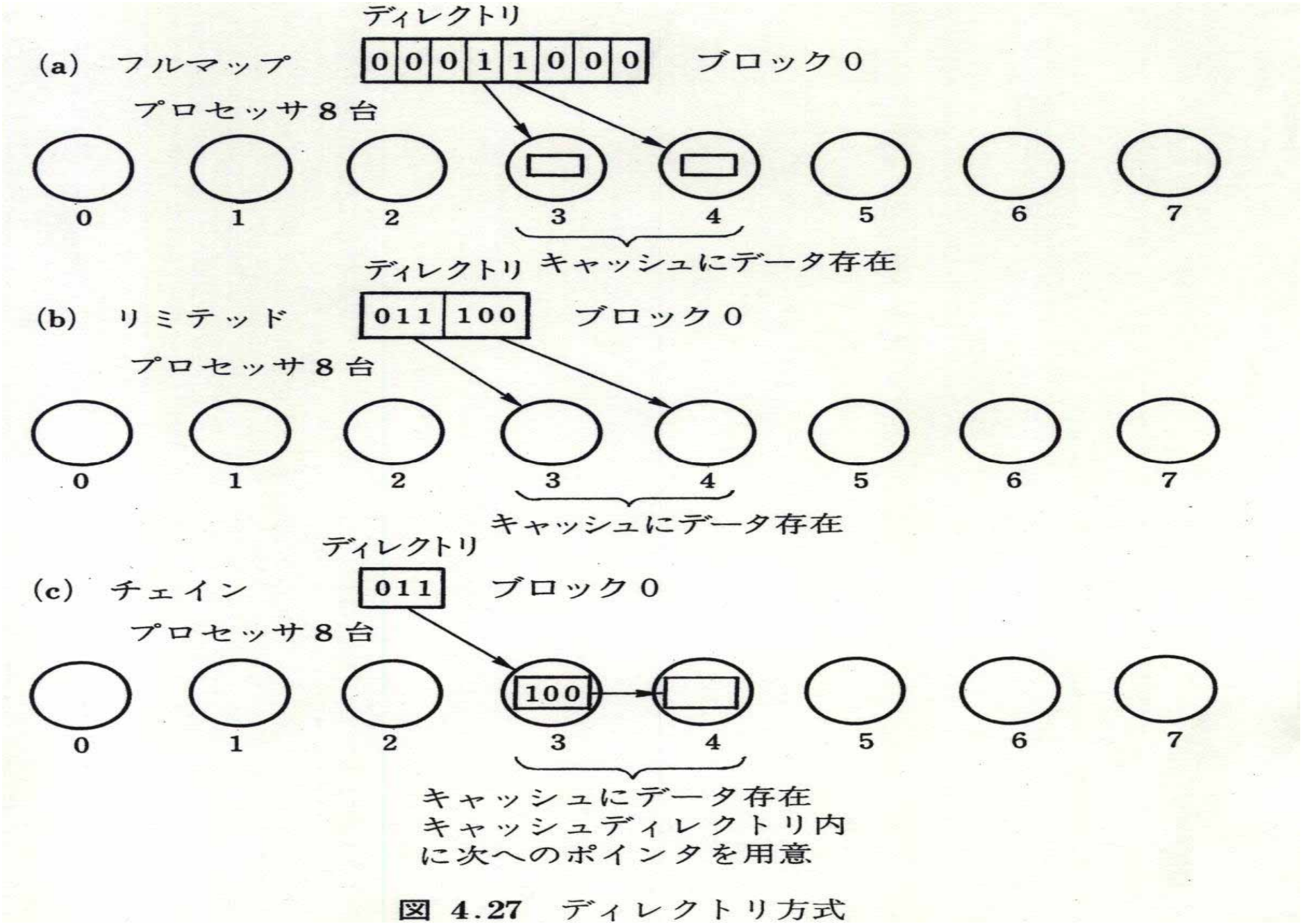
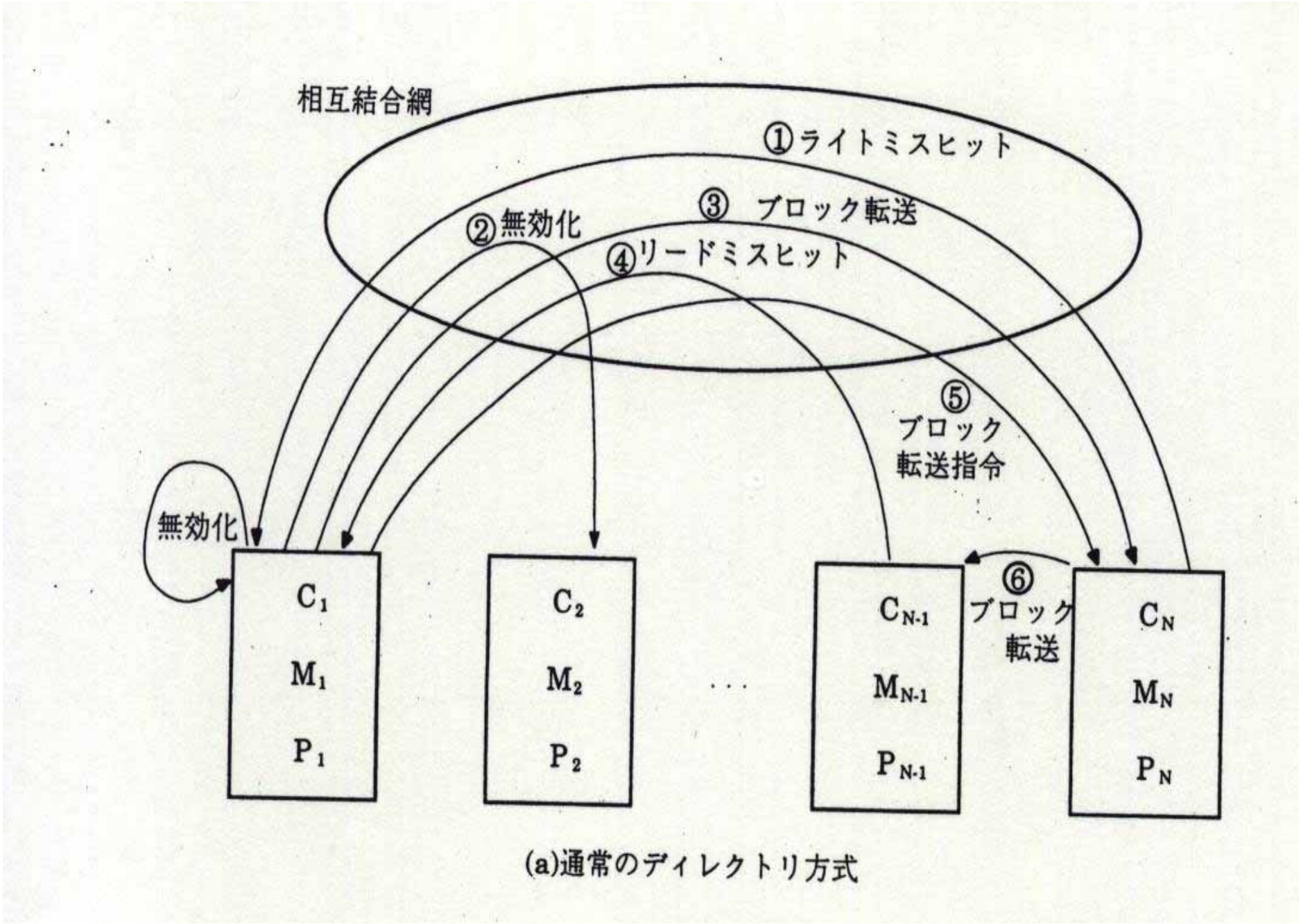
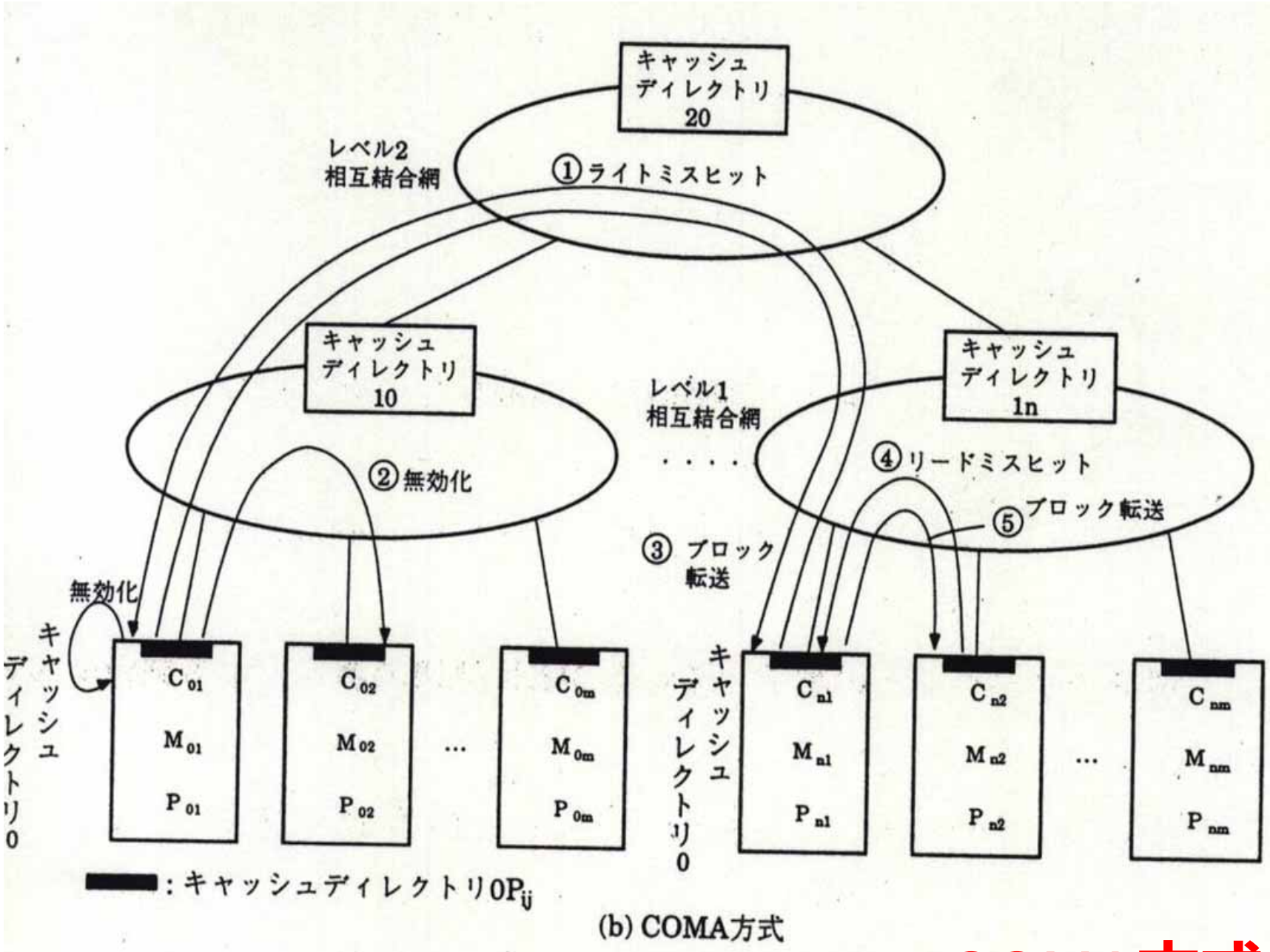


図 4.27 ディレクトリ方式



(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

③具体例

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／古い値を使ってしまう

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

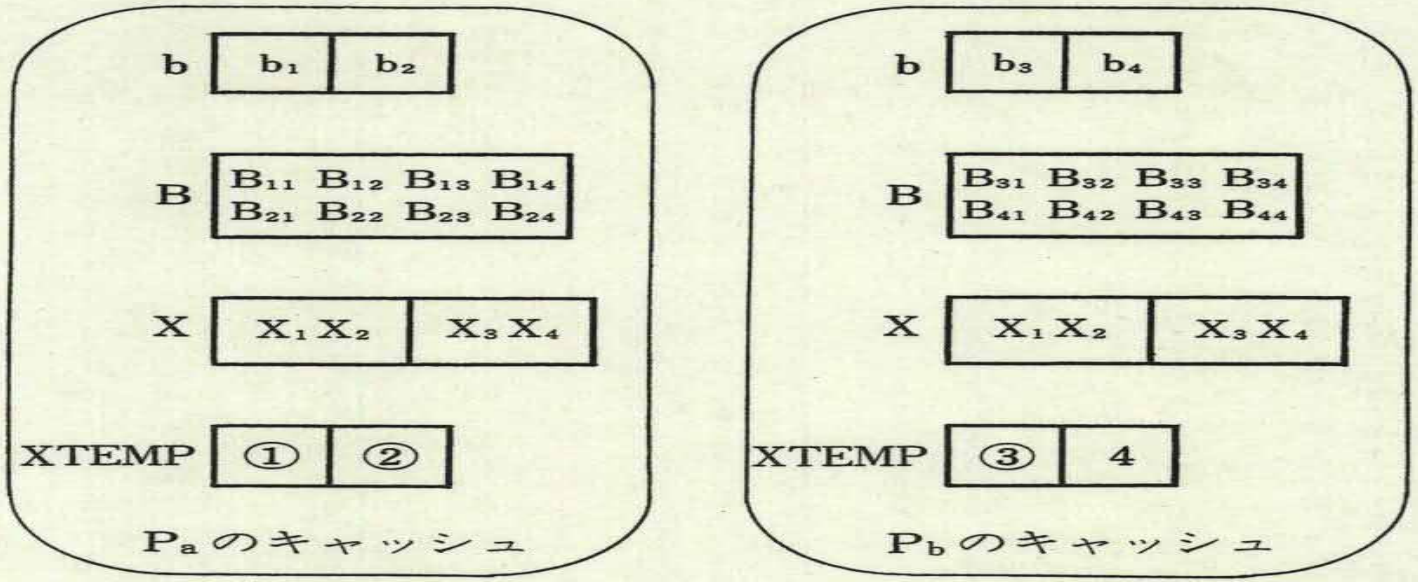
同期操作

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キャッシュ無効化命令：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

7.8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？

7.8.1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

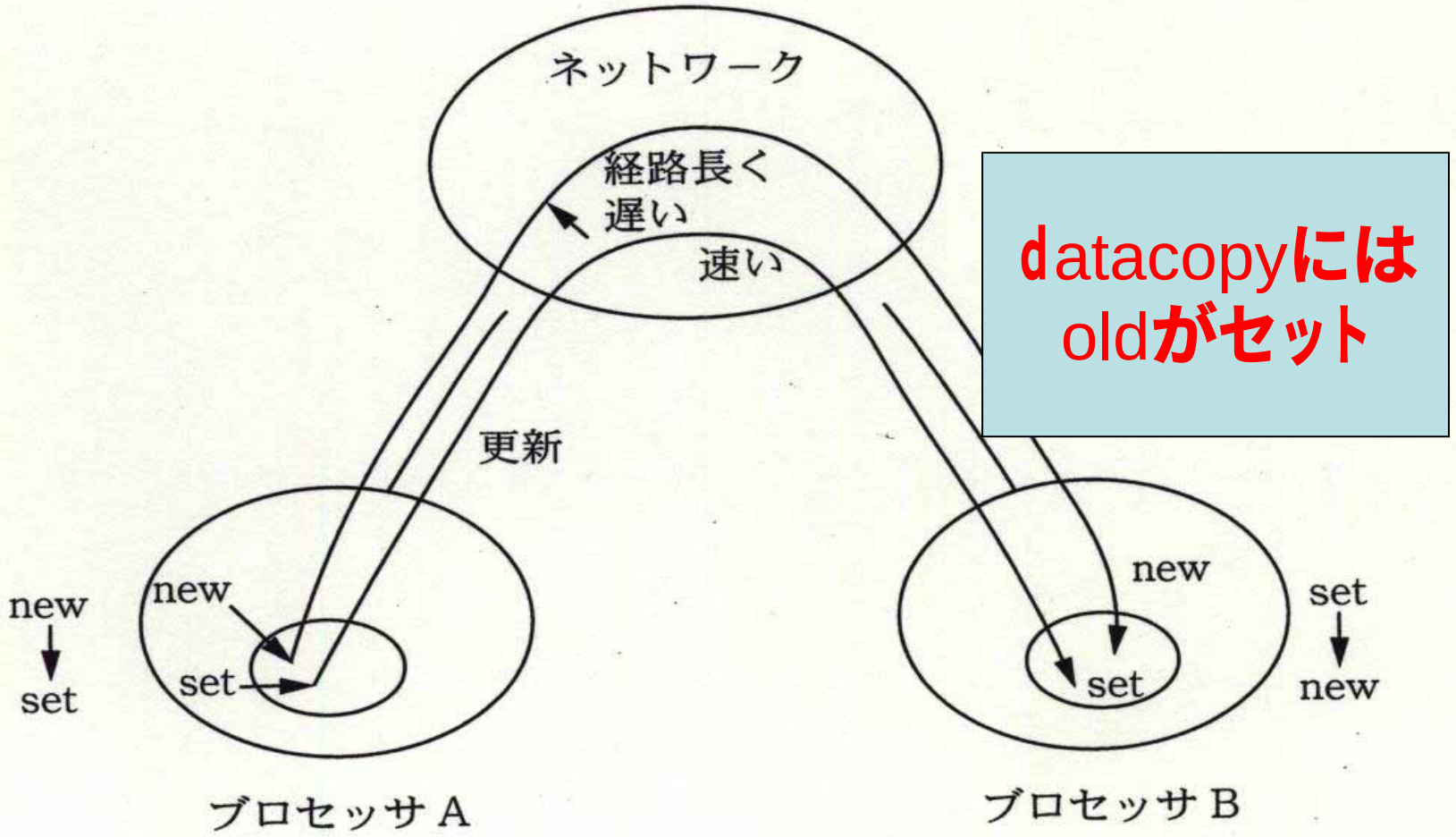
① data＝new；

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

7.8.2逐次 (sequential) コンシステンシ

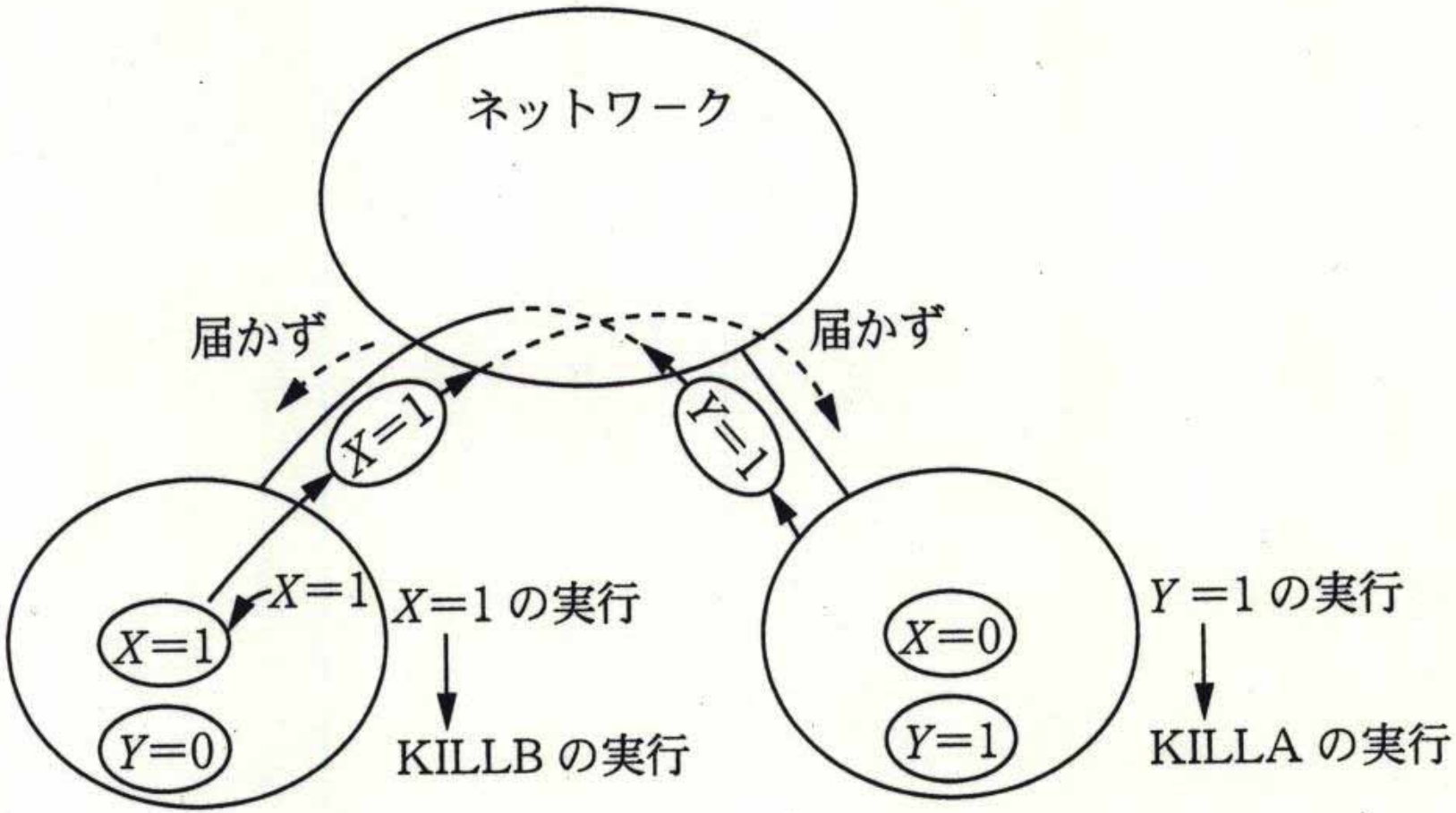
X＝Y＝0

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

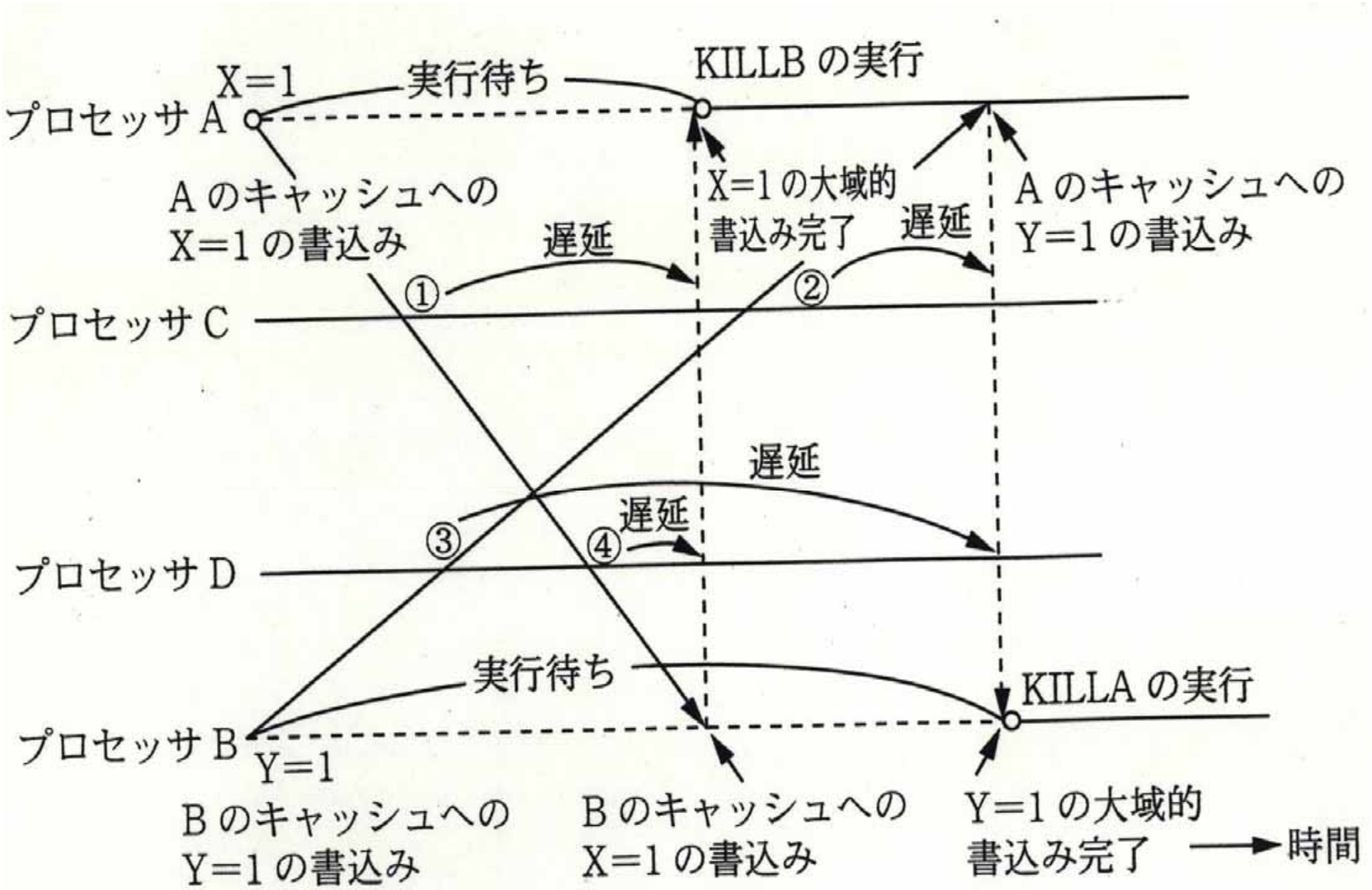


図 4.30 逐次コンシステンシの実現

(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

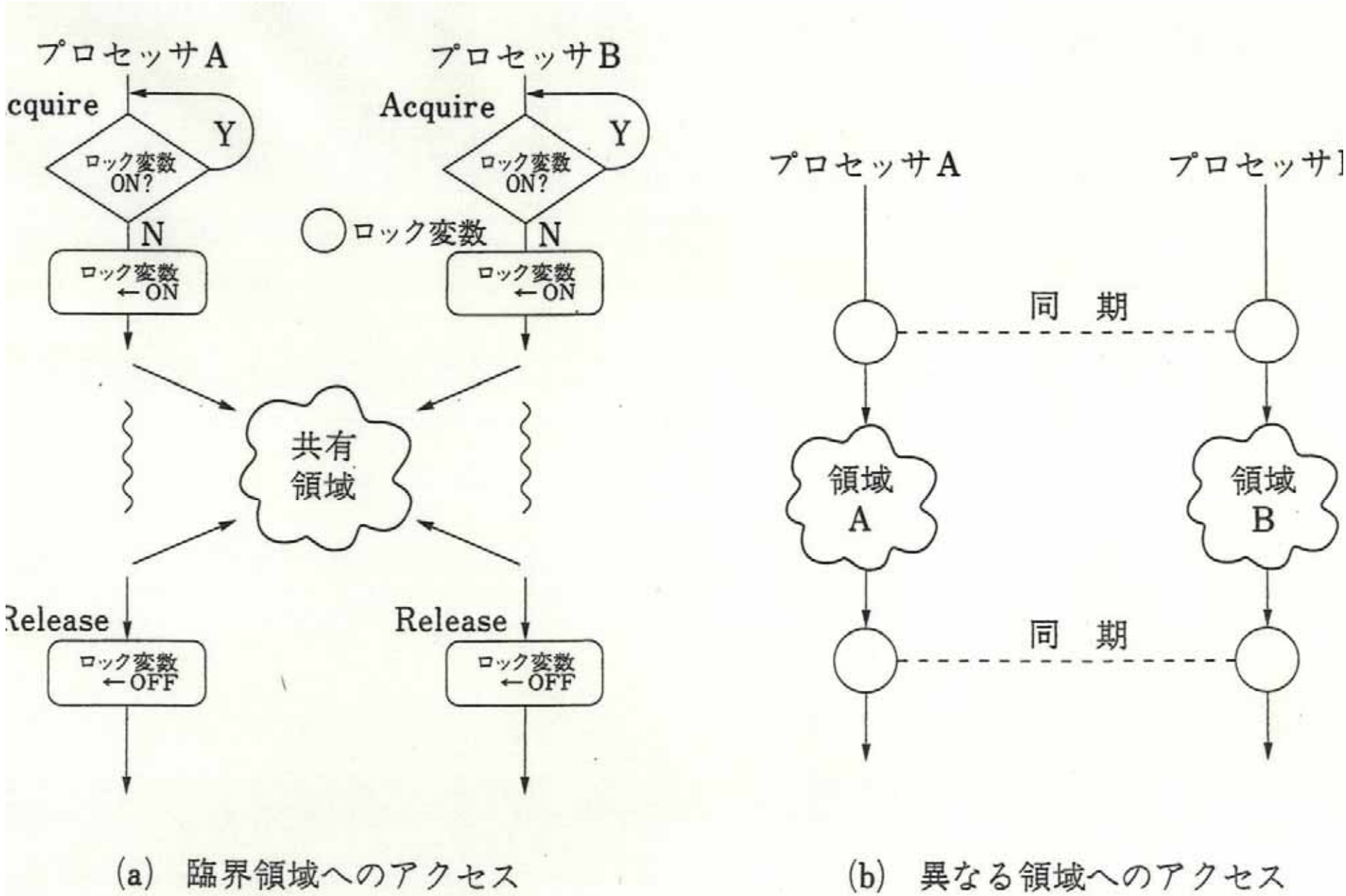
Acquire同期命令

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

7．9　メッセージパッシング型マルチプロセッサ

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

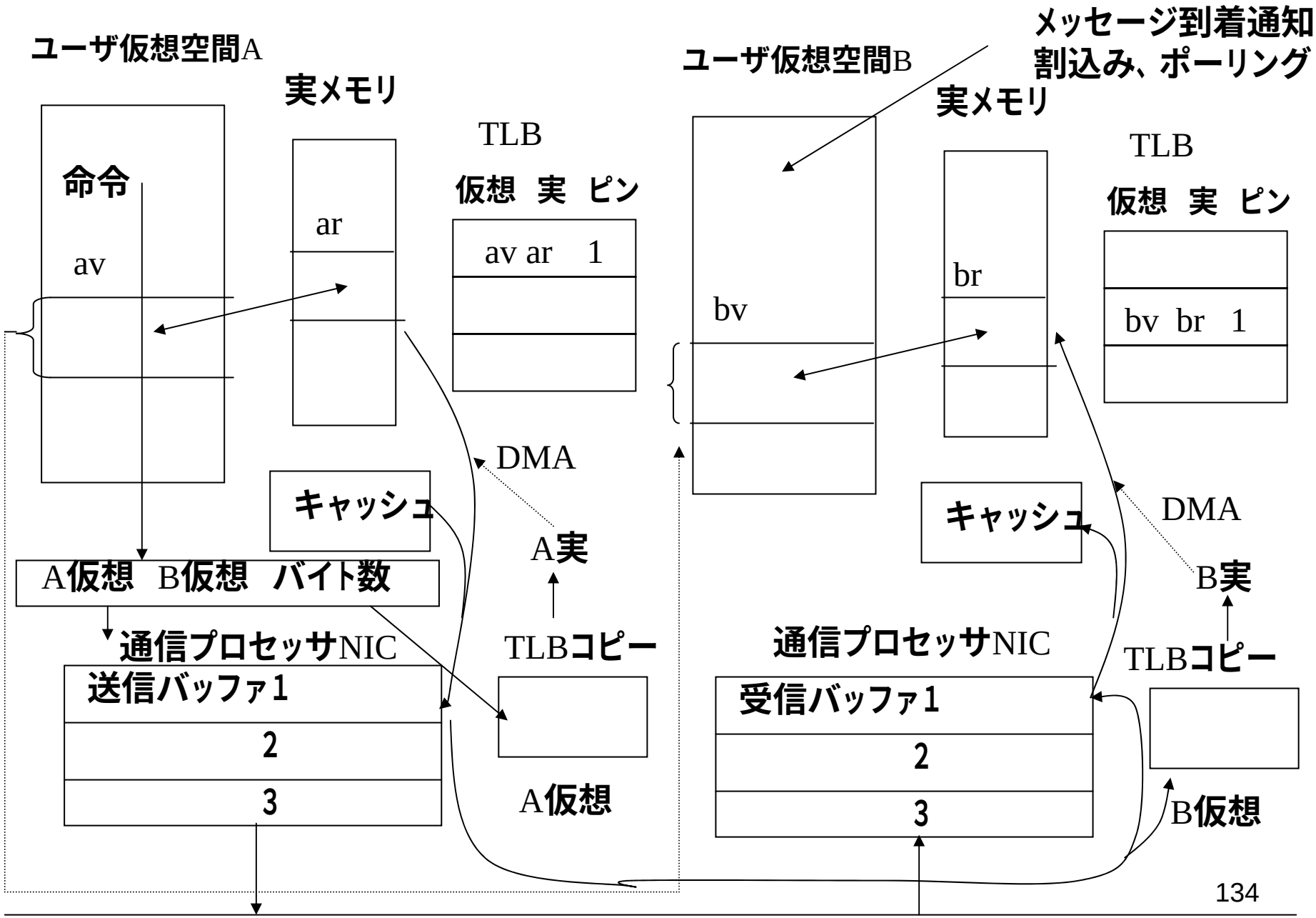
④Write Combining あり：FM,LFC,AM-II,Hamlyn,BIP

⑤同上なし

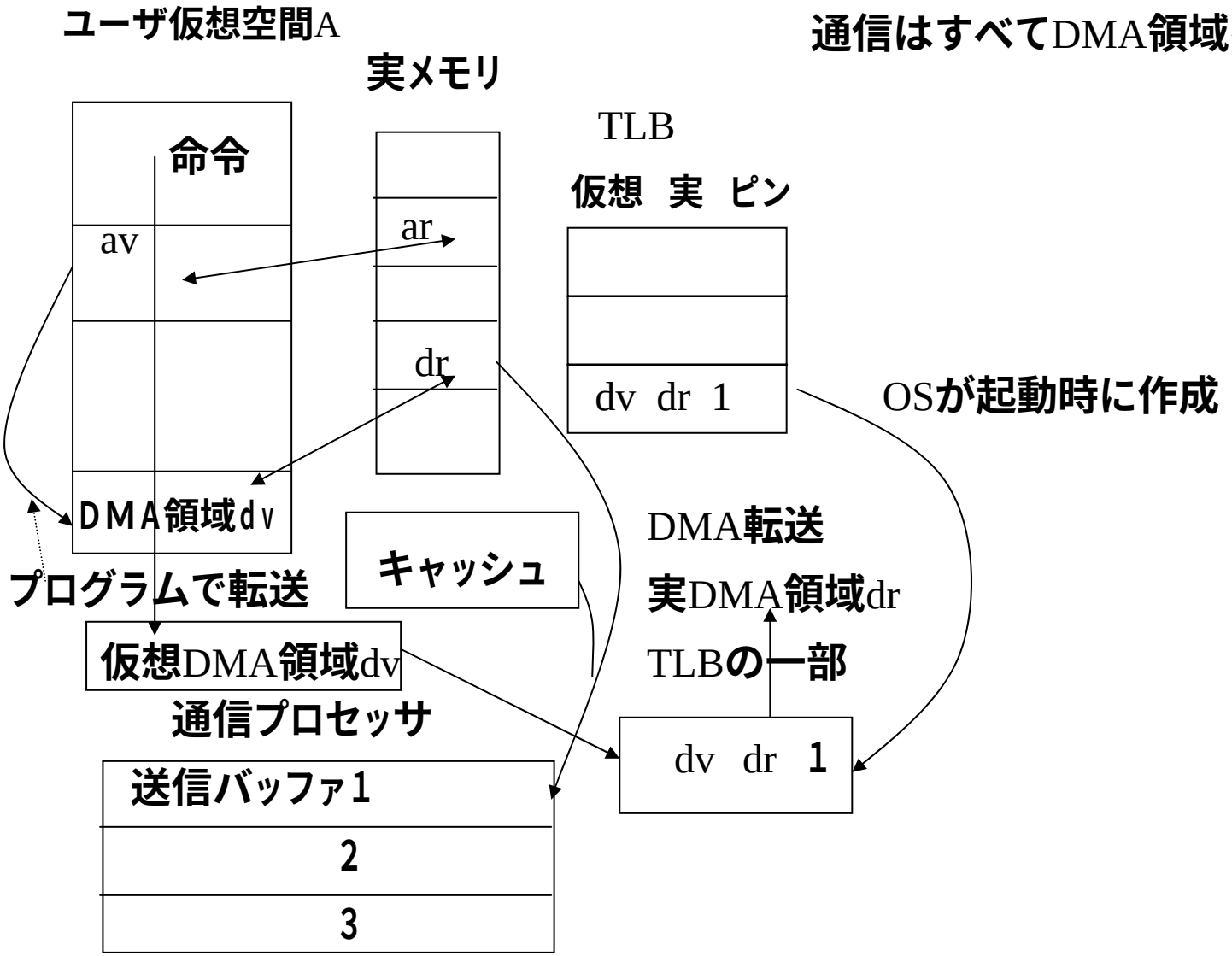
Table 1. Characteristics of 11 communication systems built for Myrinet.

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

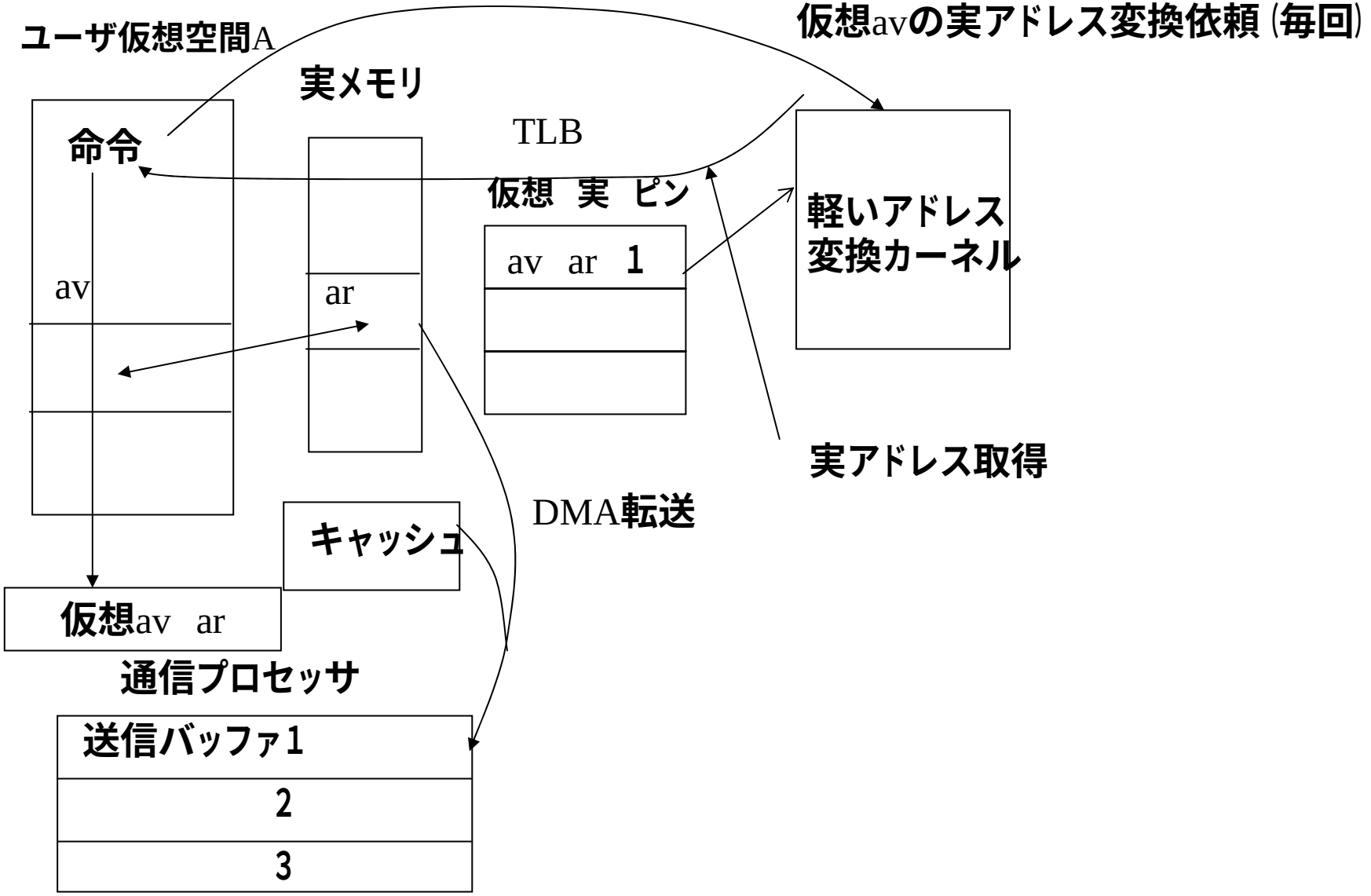
DMA基本方式　転送実領域：貼付け（ピン）、通信プロセッサ　ＴＬＢアクセスの必要



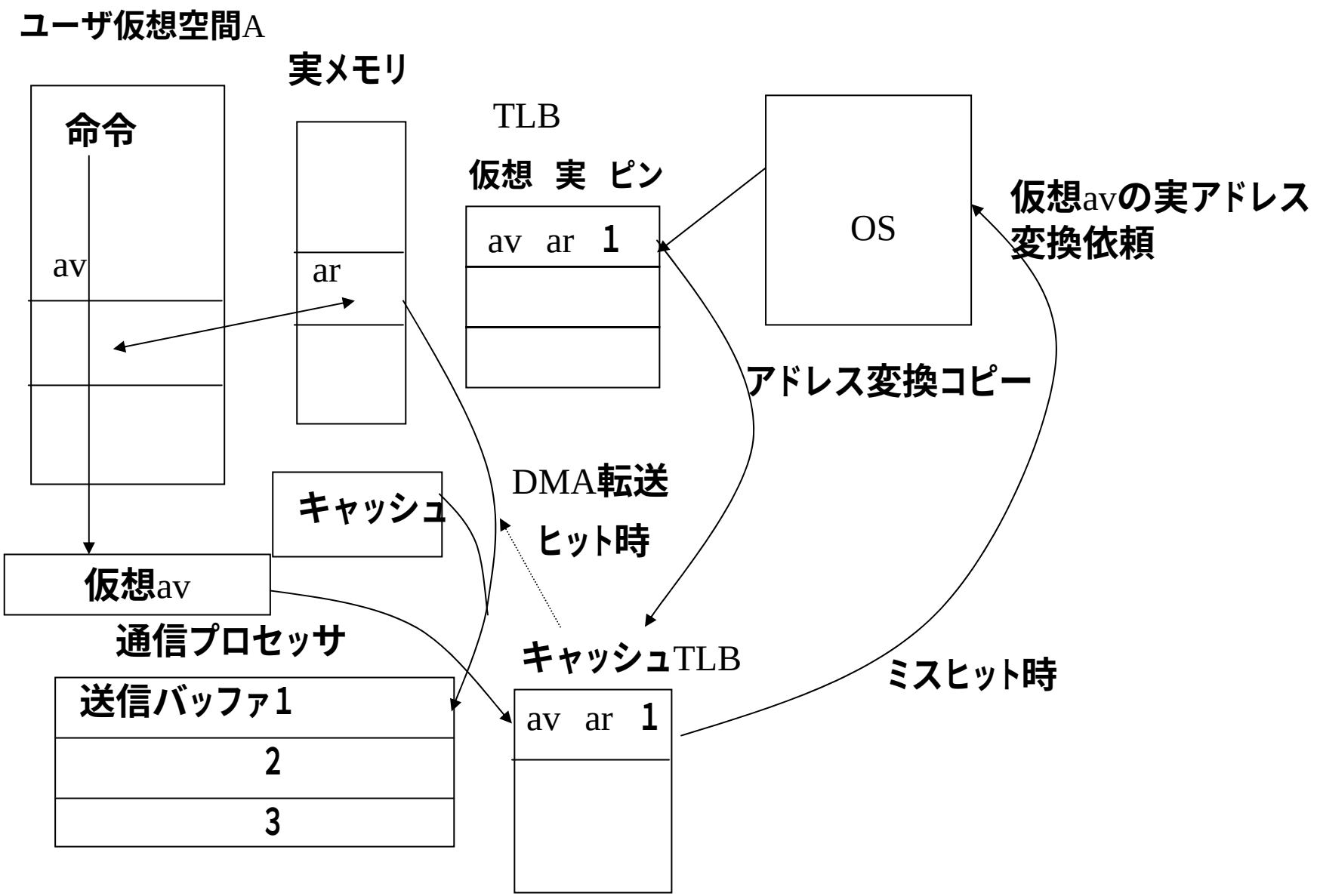
方式①：仮想空間AのDMA領域をOSに依頼して貼付け（起動時一度だけ）



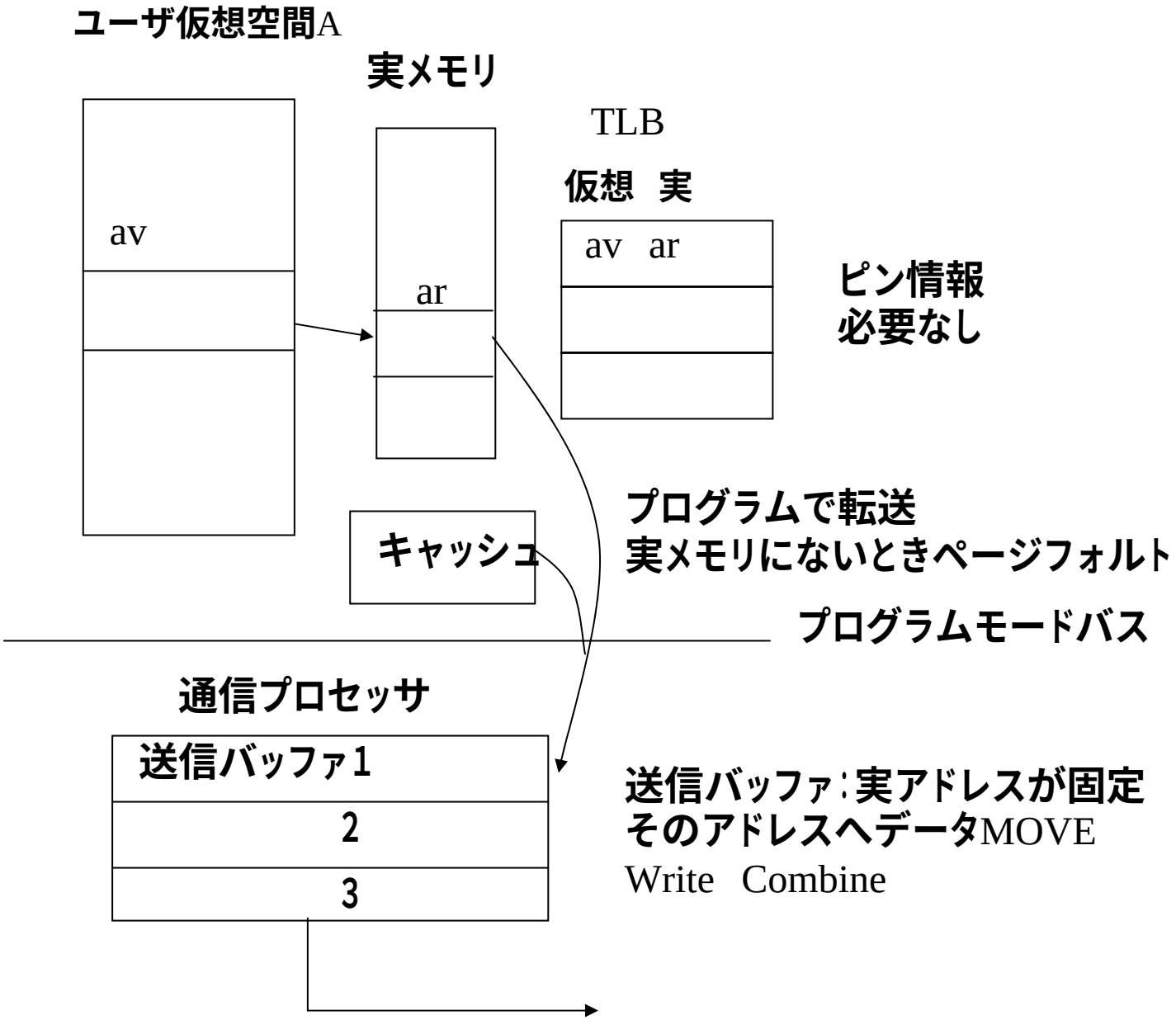
方式②： 軽いアドレス変換カーネル

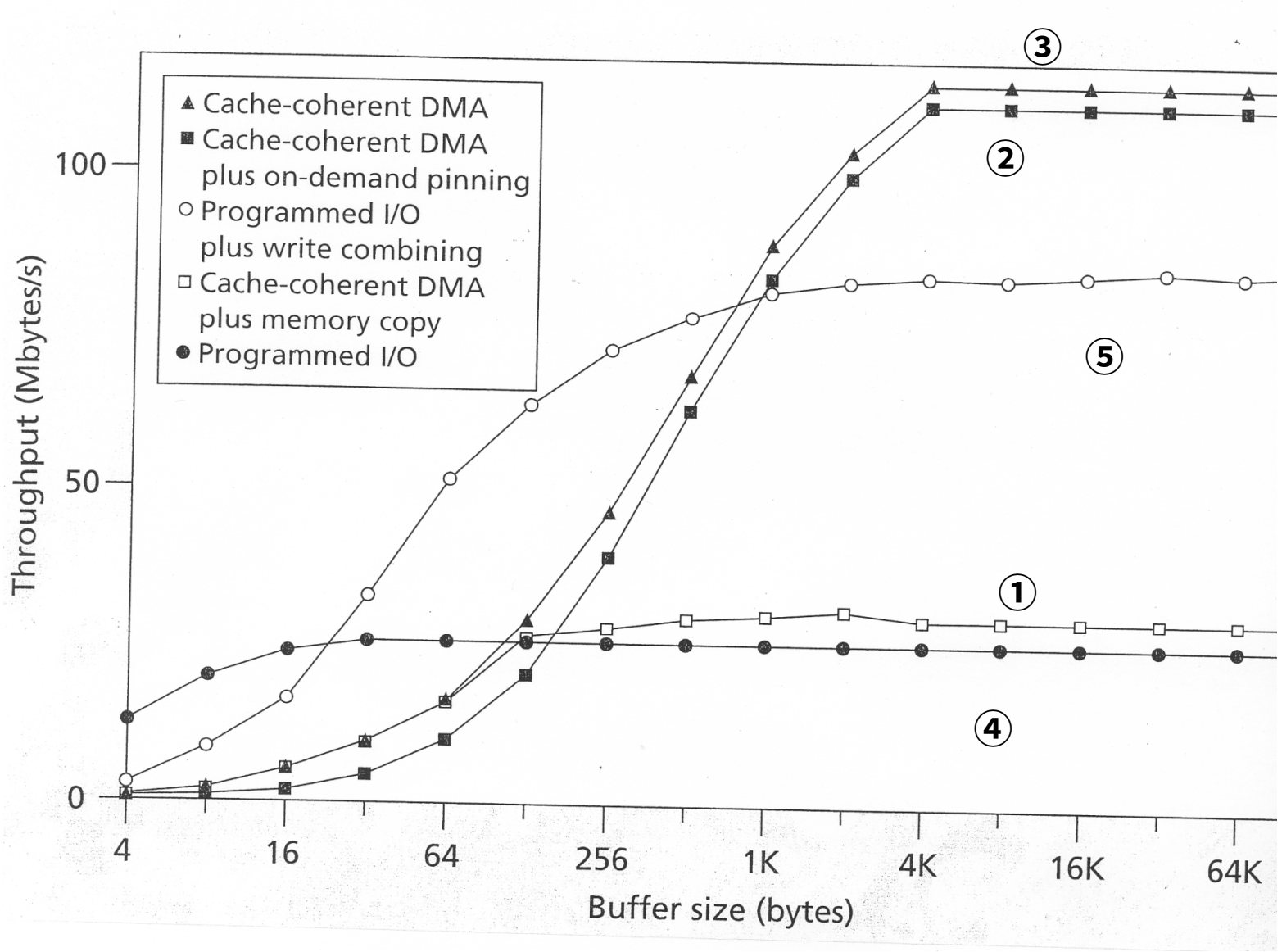


方式③： TLBキャッシュ



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

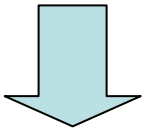
7．10　マルチコア型プロセッサ

7.10.1大域並列の利用

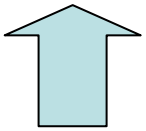
- パイプライン：時間並列
- 乱実行、投機実行による時間並列の高速化
- 局所並列：スーパスカラ、VLIW

プログラムカウンタの近傍にある命令の並列実行

- 非常に複雑な構造



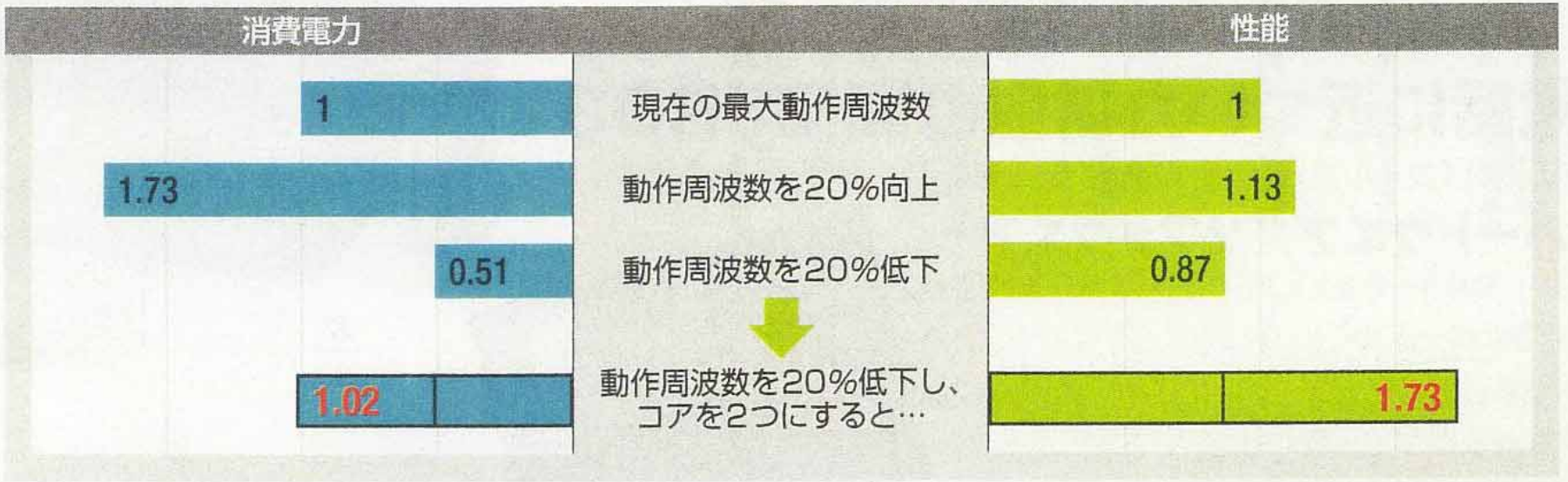
- 大域並列：マルチコア型プロセッサ



- 周波数向上が困難
  - 深いパイプラインで性能向上：小さくなった
  - 電力の問題：消費電力$\propto f^3$
  - リーク電流の増大

●Core 2 Duoの設計思想

現在の動作周波数での性能と消費電力を1としたときの相対値で表示



インテルが開発者会議で発表した値。現在の動作周波数からさらに20%周波数を上げても、性能は1.13倍にしかならないが、消費電力は1.73倍と大幅に増えてしまう。逆に20%周波数を下げると、性能は0.87倍になるが、消費電力は0.51倍とほぼ半減する。それならば、動作周波数を落としてコアを複数にすることで消費電力と性能のバランスをとろうというのがCore 2 Duoの設計思想だ

日経栄パソコン2006.8.14

①深いパイプラインの限界

周波数向上:40%/年

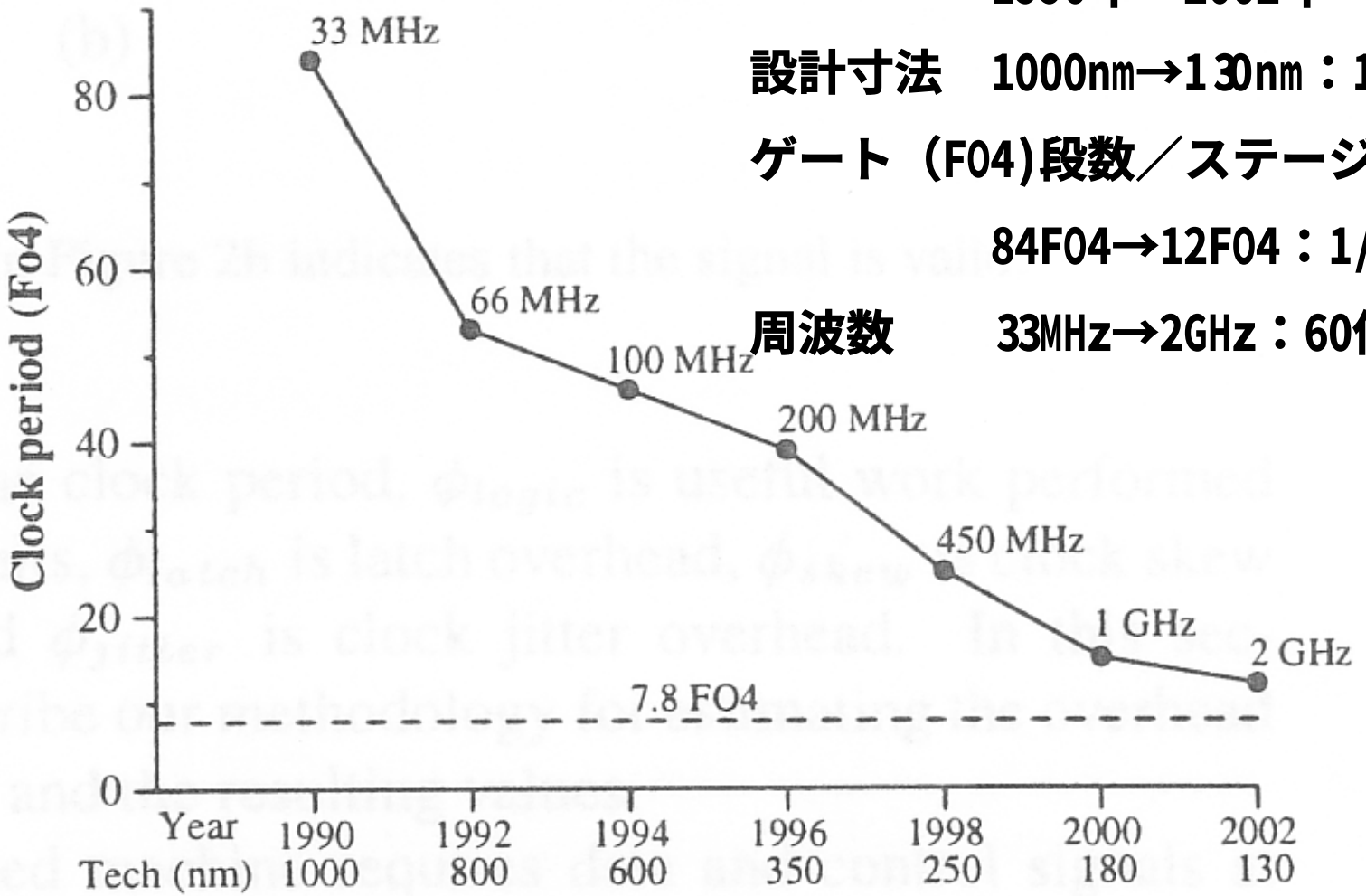
1990年　2002年

設計寸法　1000nm→130nm：1/8

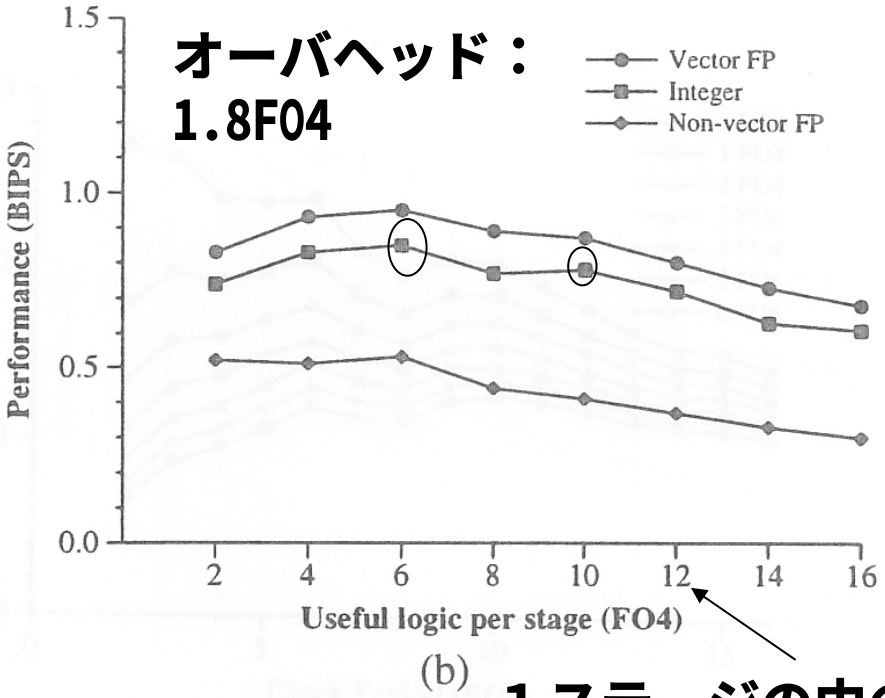
ゲート（FO4)段数／ステージ

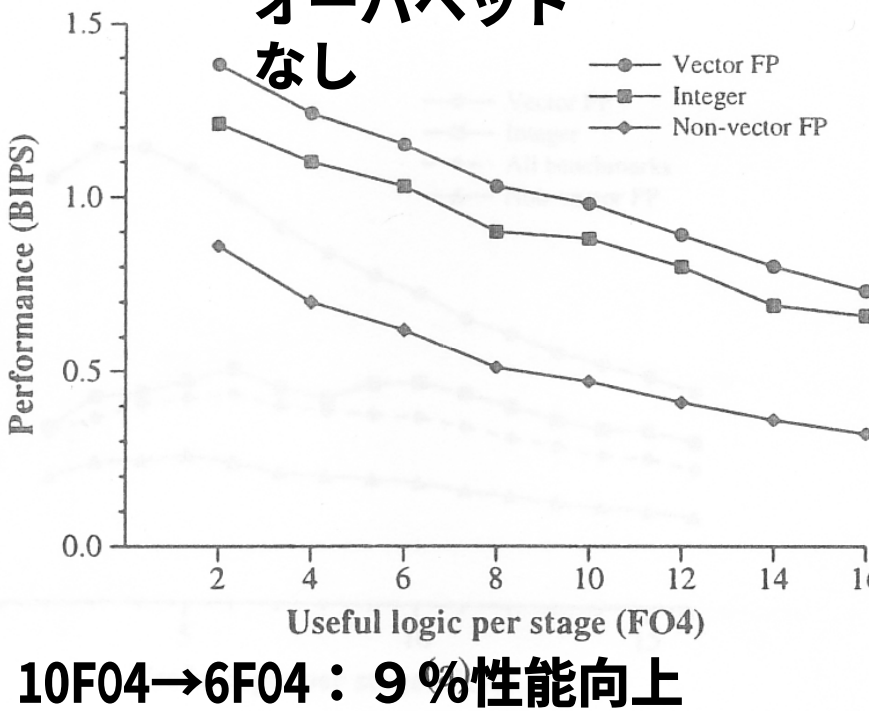
84FO4→12FO4：1/7

周波数　33MHz→2GHz：60倍



N.Joppi et.al.,ISCA,2002

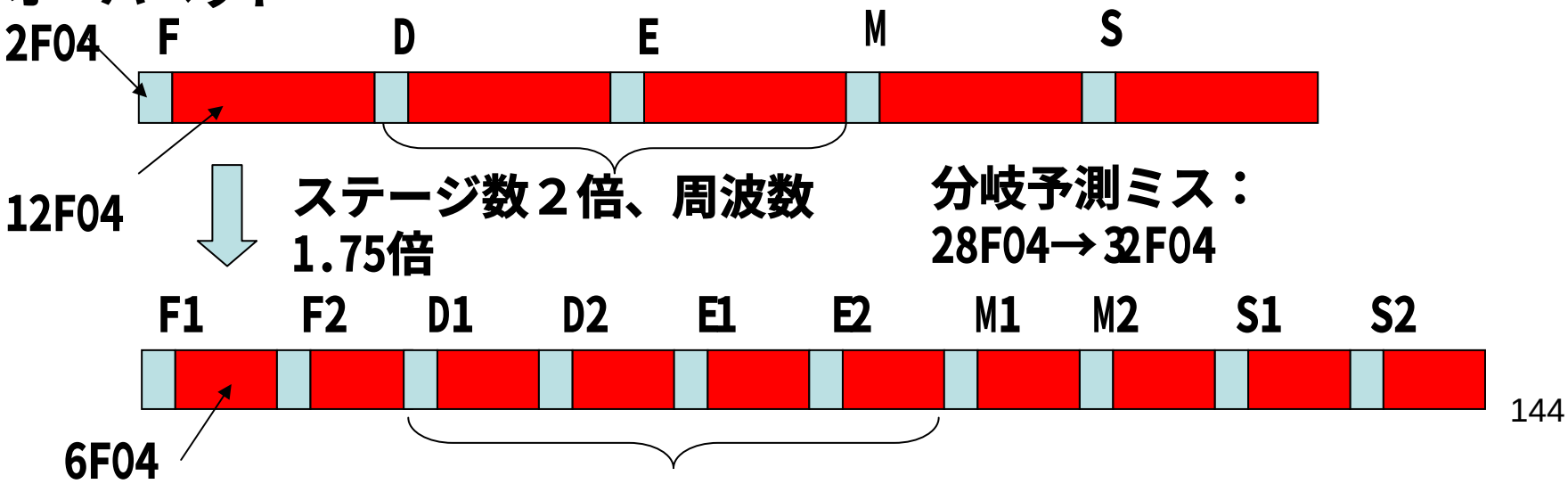




１ステージの中の
有効ゲート数

10F04→6F04：９％性能向上

11.8F04→7.8F04：周波数1.5倍



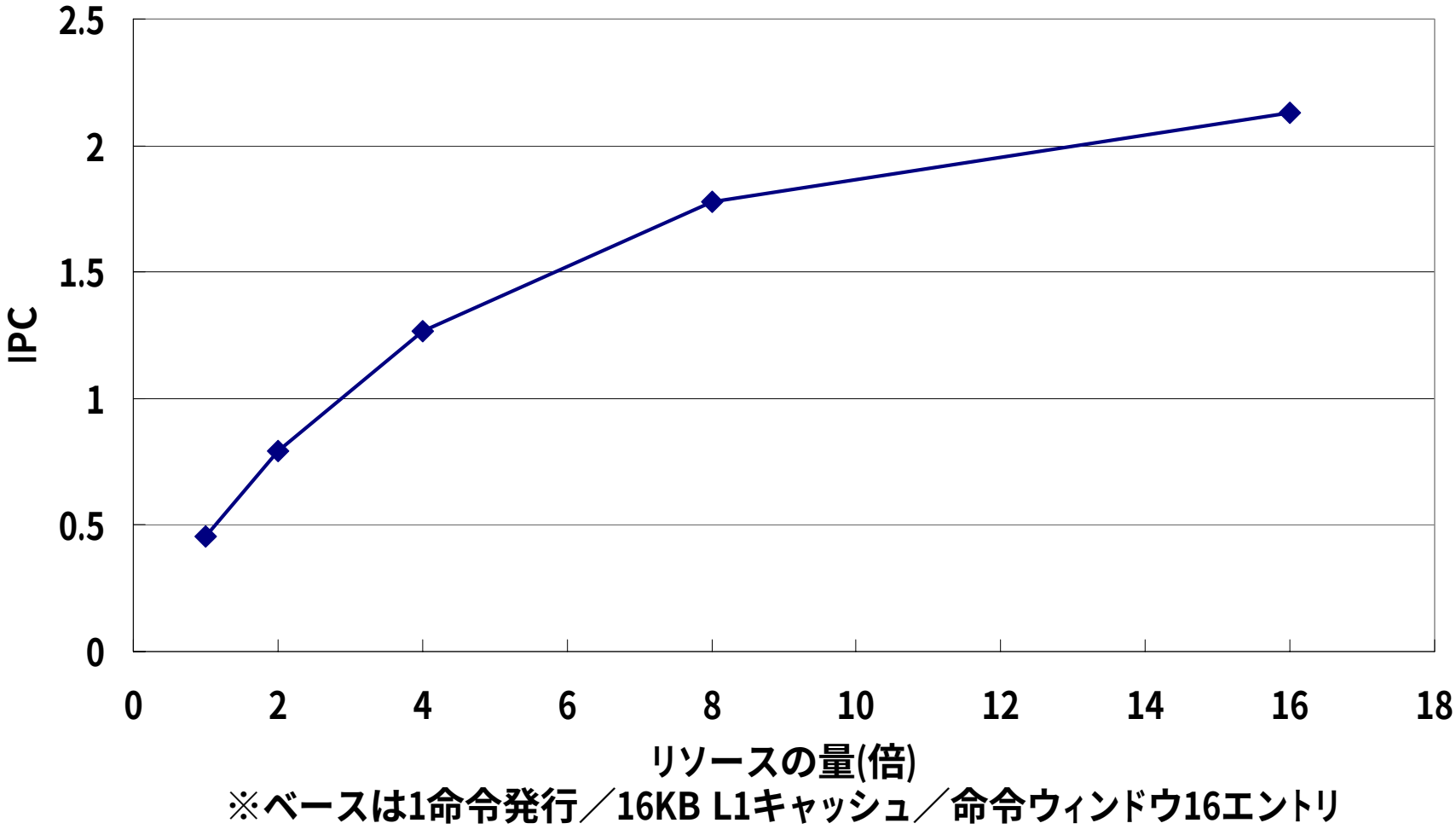
②IPC向上率の鈍化

ベースアーキテクチャ

命令セット：SimpleScalar PISA (MIPS R10000**とほぼ同じ**)
発行幅：4(**フェッチ、デコード、発行、ライトバック、コミット幅**)
命令ウィンドウ：64**エントリ**
機能ユニット：**全ての機能ユニットを発行幅と同じだけ持つ**
L1**命令キャッシュ**：64KB/32B line/2-way/**ヒット・レイテンシ**2**サイクル**
L1**データ・キャッシュ**：64KB/32B line/2-way/**ヒット・レイテンシ**2**サイクル**
L2**キャッシュ**：2MB/64B line/4-way/**ヒット・レイテンシ**16**サイクル**
メモリ・アクセス：128**サイクル**
gshare：8K**エントリ**

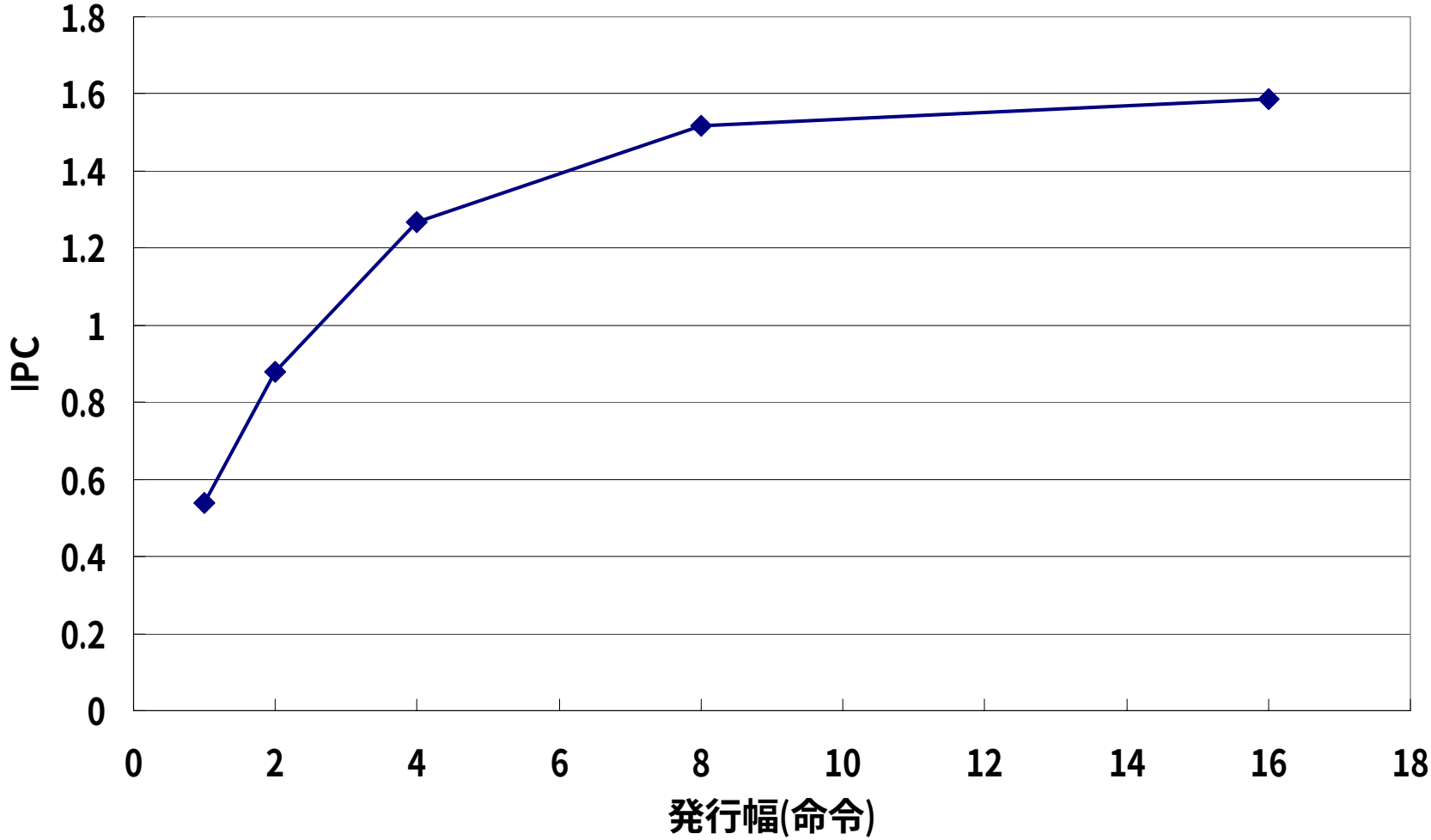
BTB：2K**エントリ**/4-way
分岐予測ミス・ペナルティ：10**サイクル**

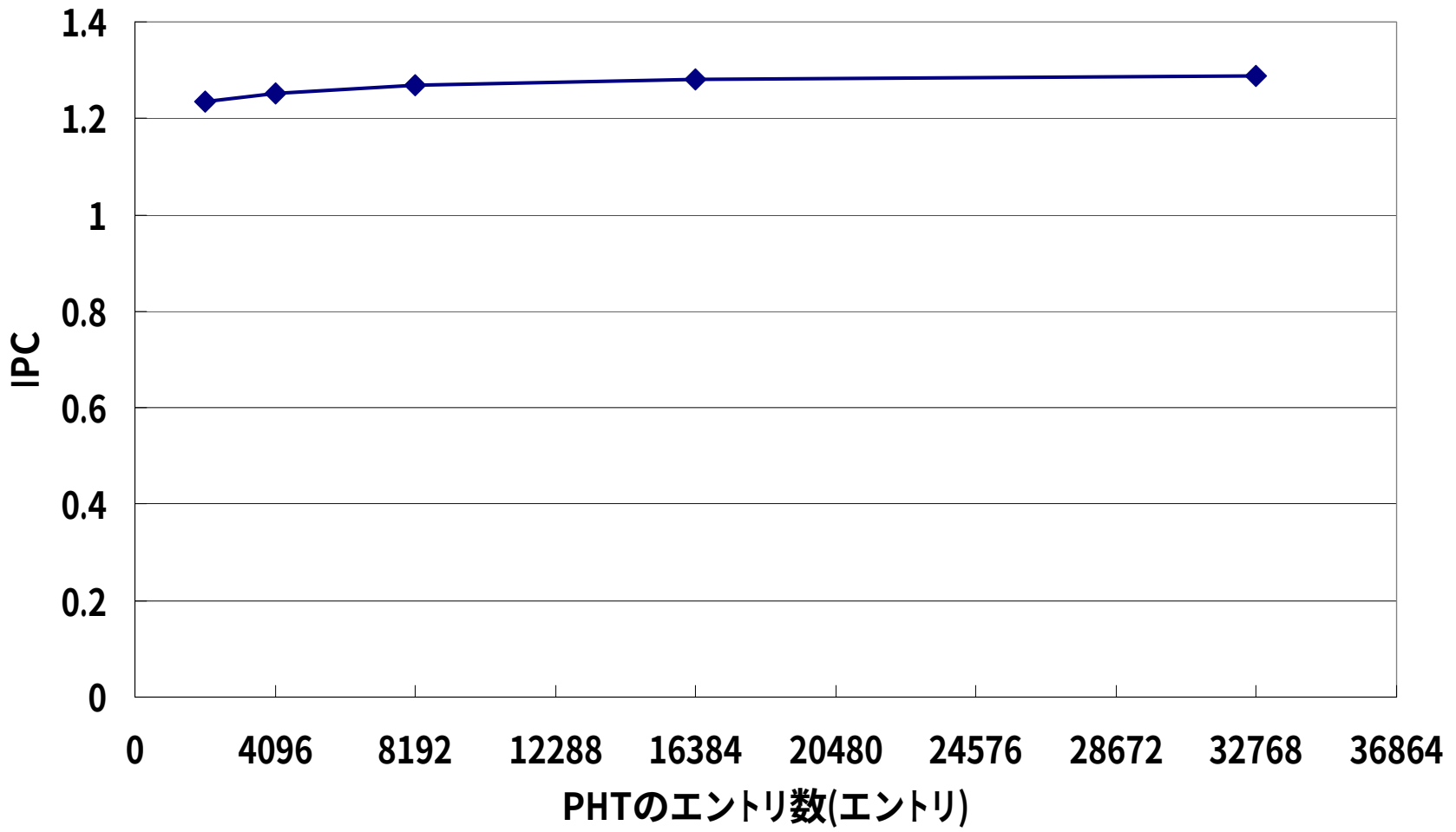
全リソース(L1キャッシュ、PHT、命令ウィドウ、発行幅)の変更



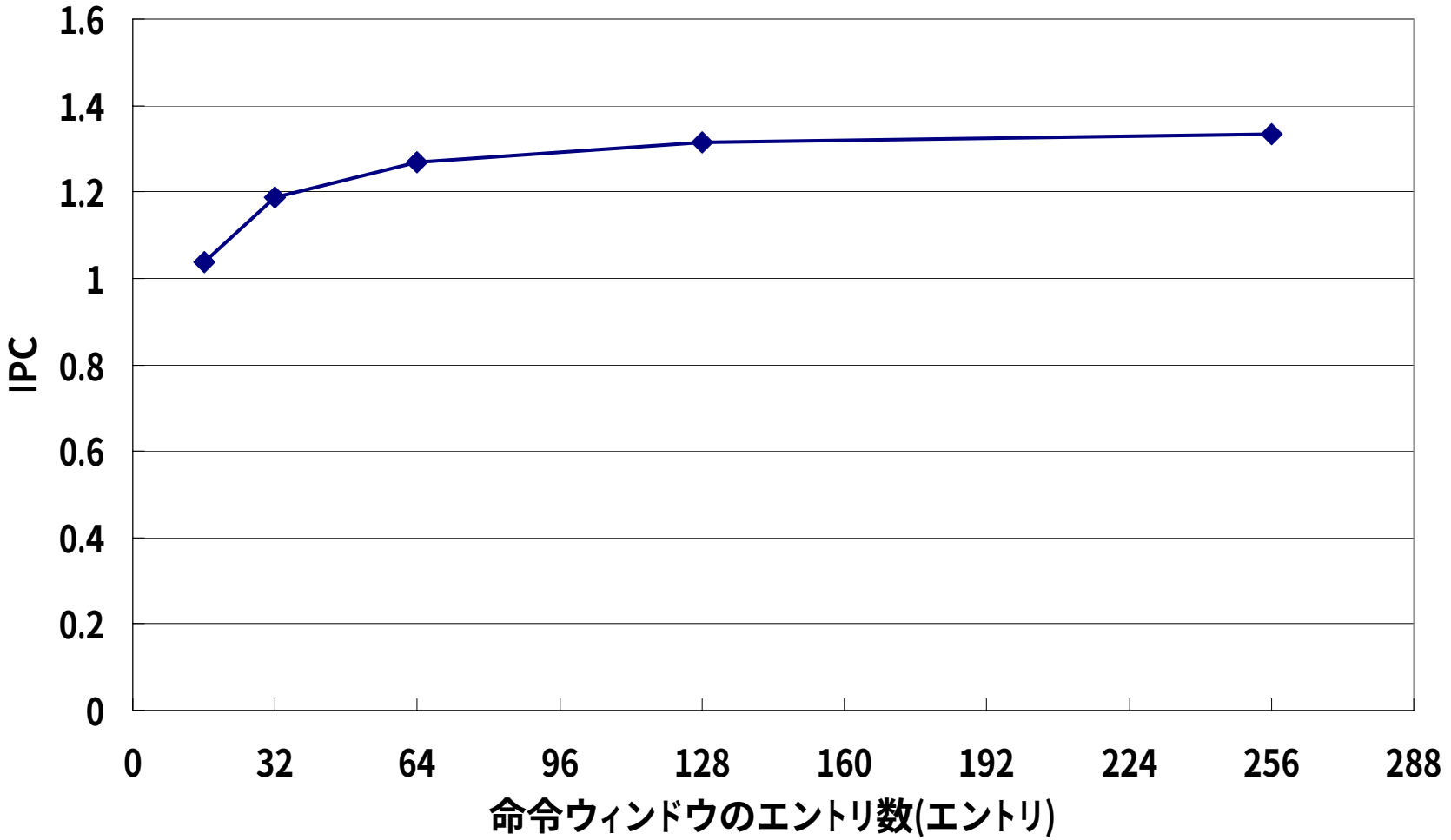
※ベースは1命令発行／16KB L1キャッシュ／命令ウィンドウ16エントリ

発行幅のみ変更

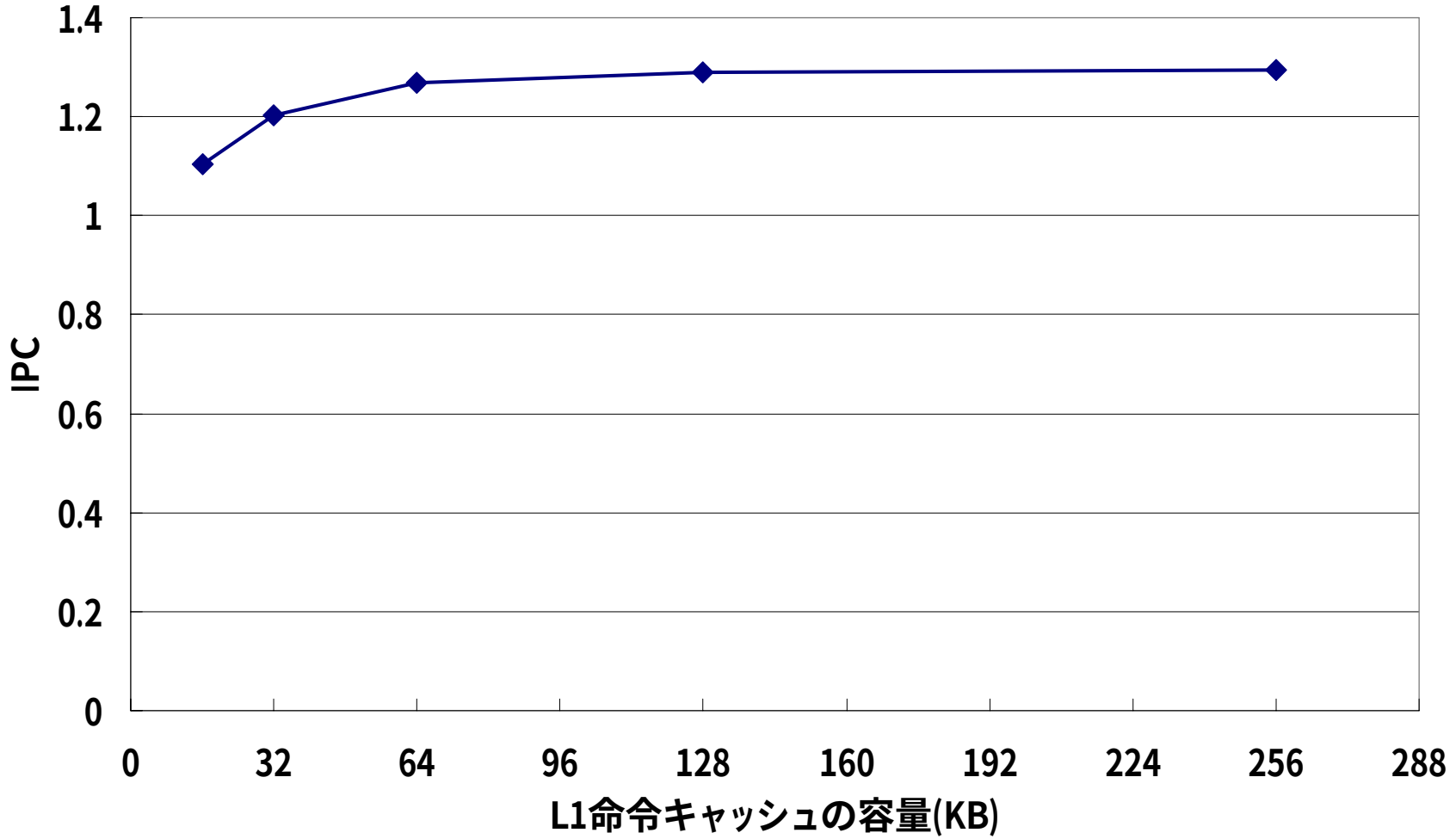



gshareのPHTのエントリ数のみ変更
1.4
1.2
1
0.8
0.6
0.4
0.2
0
IPC
0
4096
8192
12288
16384
20480
24576
28672
32768
36864
PHTのエントリ数(エントリ)

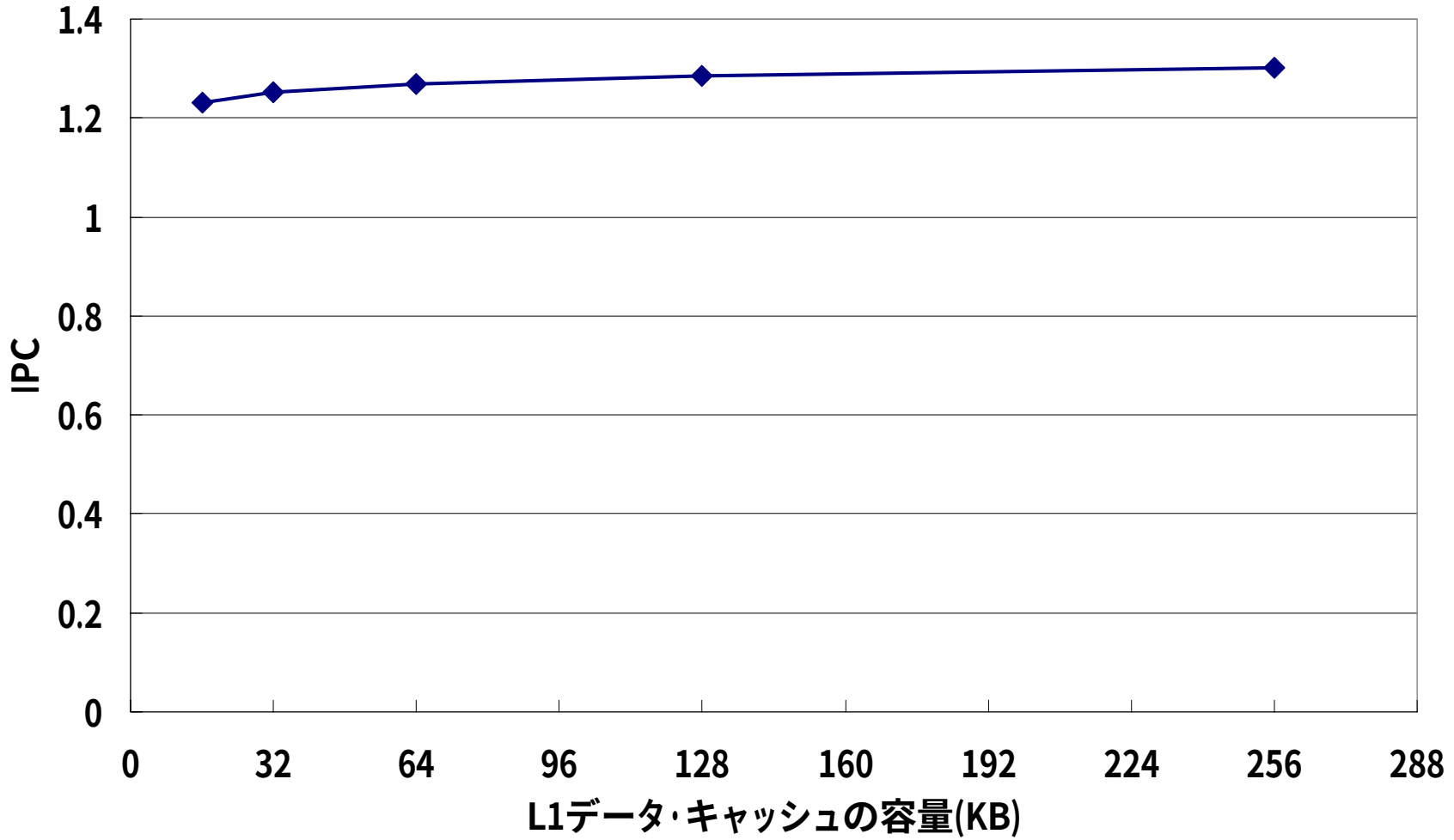
命令ウィンドウのエントリ数のみ変更



L1命令キャッシュの容量のみ変更



L1データ･キャッシュの容量のみ変更



③省電力化

CMOSの電力消費

・動的

回路がON、OFFするとき　$\alpha f C V^2$

・漏れ電流　$V I_{leak}$

・貫通電流　$\alpha f t_{st} I_{short} V$

α：ゲート動作率、f：周波数、C：ゲート総容量、

V:電源電圧、t_{st}：スイッチング時間

$P = \alpha f C_L V^2 + V I_{leak} + \alpha f t_{st} I_{short} V$

$F_{max} \propto (V - V_{threshold})^2 / V \quad V$

$I_{leak} \propto \exp(-qV_{threshold}/kT)$

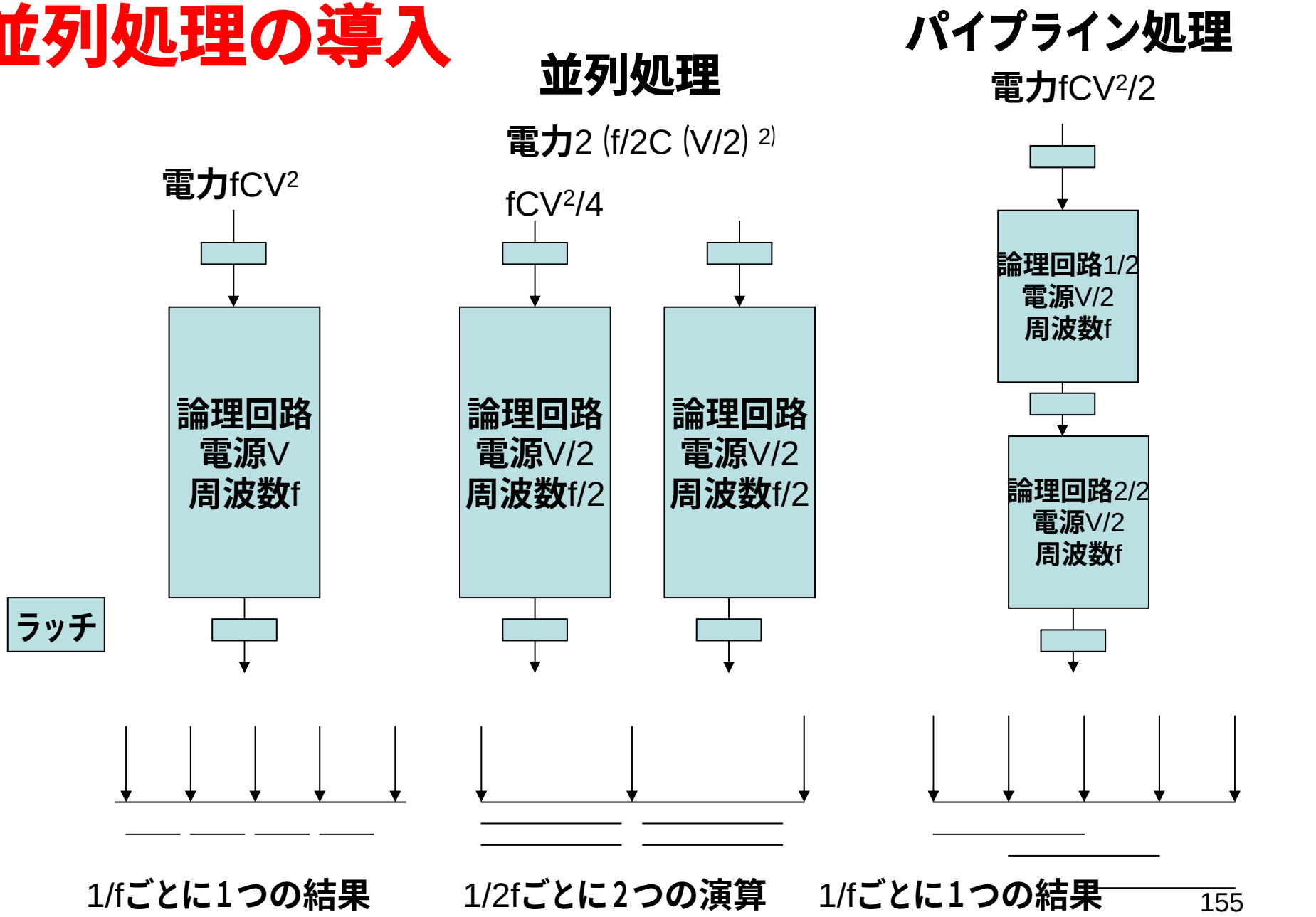
T.Mudge: Power: A First-Class Architectural Design Constraint, IEEE Computer, pp.52-58, April 2001

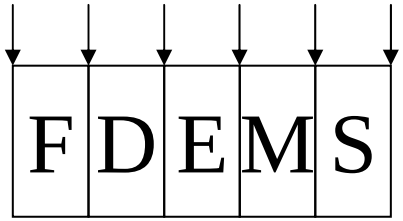
基本的な考え方

- スイッチング回数を少なくする
- 動作をしない（と予想される）回路には
  - クロック供給しない
  - 電源を供給しない
- 電源電圧を制御して、必要十分な処理速度で実行
- 電源電圧、周波数を落として並列処理、パイプラインで行う
- 基盤バイアス印加による閾値制御：
  - リーク電流削減（スブスレッショルドリーク電流）
    - 高速部分：低閾値、
    - 低速部分や待機時：高閾値（バイアス印加）

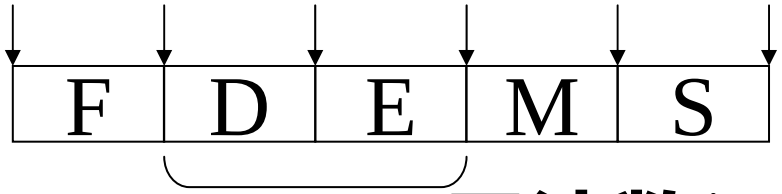
- **デバイスレベル**
  低電源電圧化、低ゲート容量化、基盤バイアス制御
- **回路レベル**
  パストランジスタ論理
  ゲート付きクロック
  グリッチの削減
  動的電源遮断
  非同期回路
- **アーキテクチャレベル**
  データパスの最適化：必要な演算幅の決定など
  並列処理、パイプライン処理
  キャッシュメモリ
  バス: アドレスの反射2進符号化、データ圧縮
- **OS、コンパイラ、アルゴリズムレベル**
  符号化
  ビット変化の少ないコード生成
  動的電源電圧制御
  動的周波数制御

並列処理の導入


並列処理
パイプライン処理
電力fCV2
電力2 (f/2C (V/2) 2)
電力fCV2/2
fCV2/4
論理回路
電源V
周波数f
論理回路
電源V/2
周波数f/2
論理回路
電源V/2
周波数f/2
論理回路1/2
電源V/2
周波数f
論理回路2/2
電源V/2
周波数f
ラッチ
1/fごとに1つの結果
1/2fごとに2つの演算
1/fごとに1つの結果



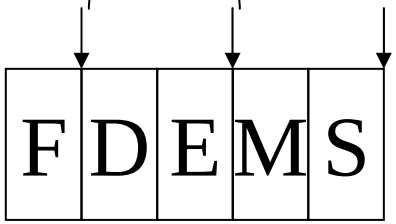
周波数f、電源V：電力：fV^2



ボルテージスケーリング

周波数f/2、電源V/2：電力：1/8

分岐ミスのとき



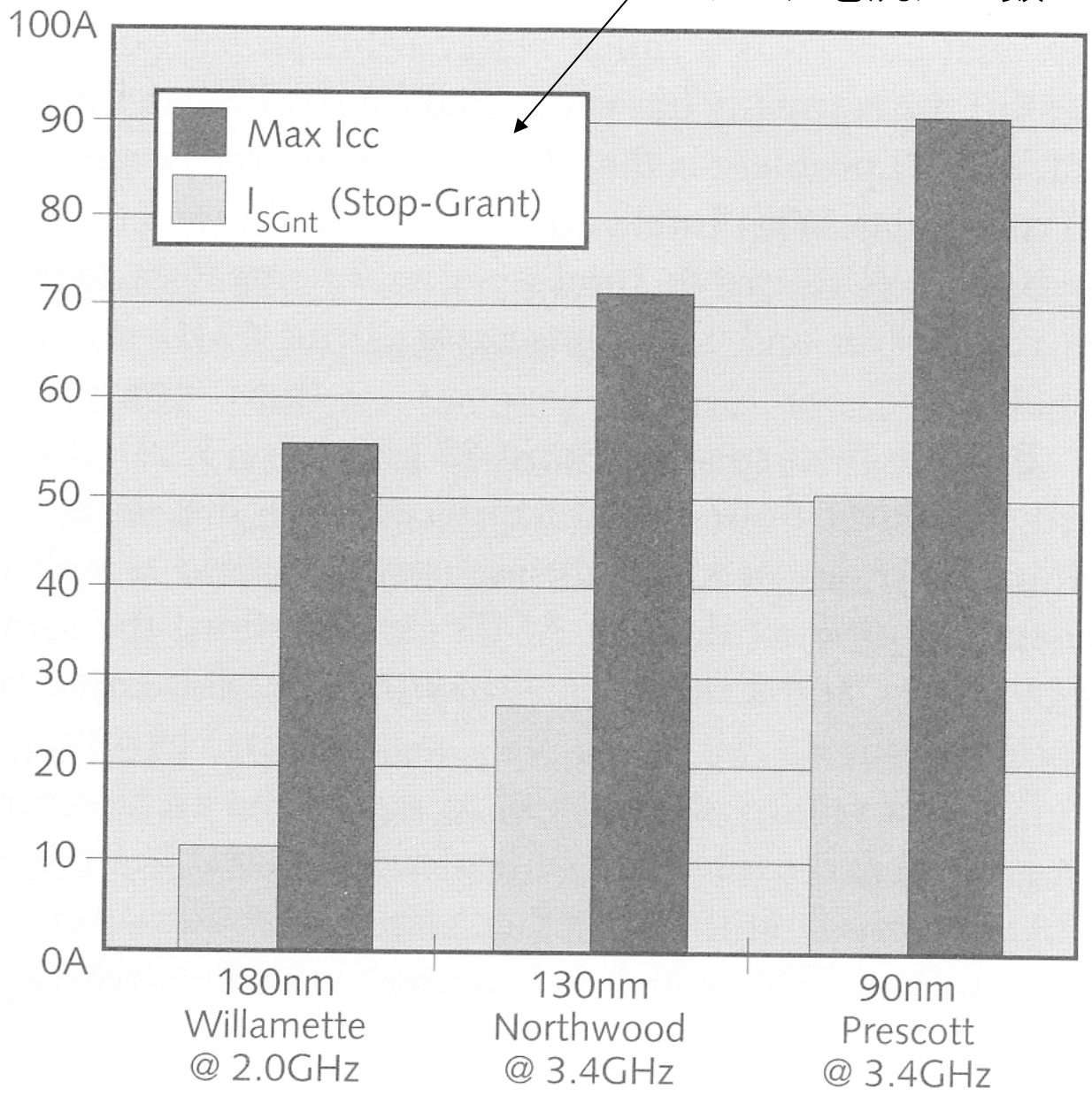
パイプラインステージ統合

周波数f/2、電源V：電力：1/2

低電源電圧化が困難なとき、有効（リーク電流）

アイドル状態での電力消費

リーク電流、TR数の増大

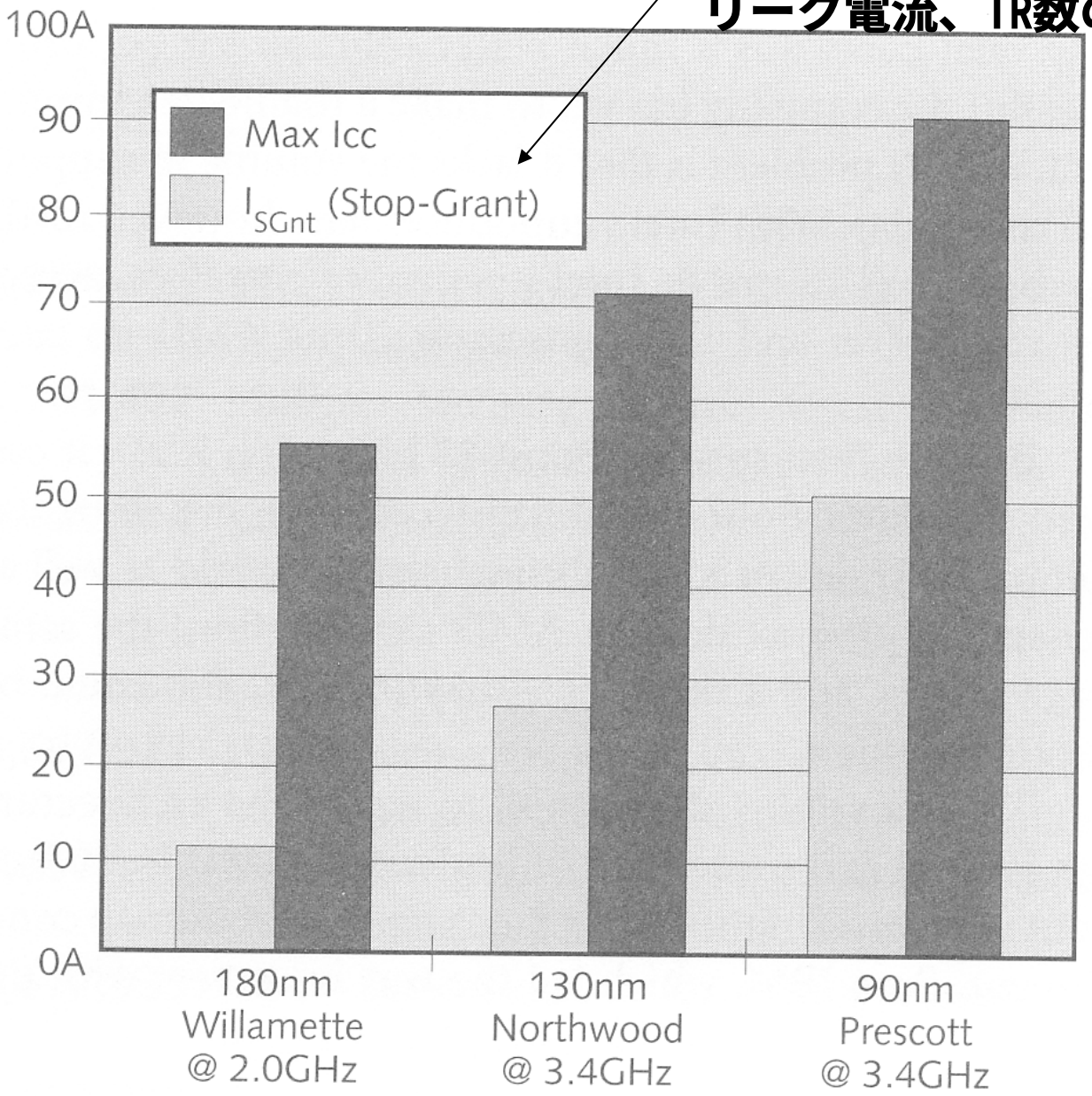


Microprocessor Report

May 2004

アイドル状態での電力消費

リーク電流、TR数の増大



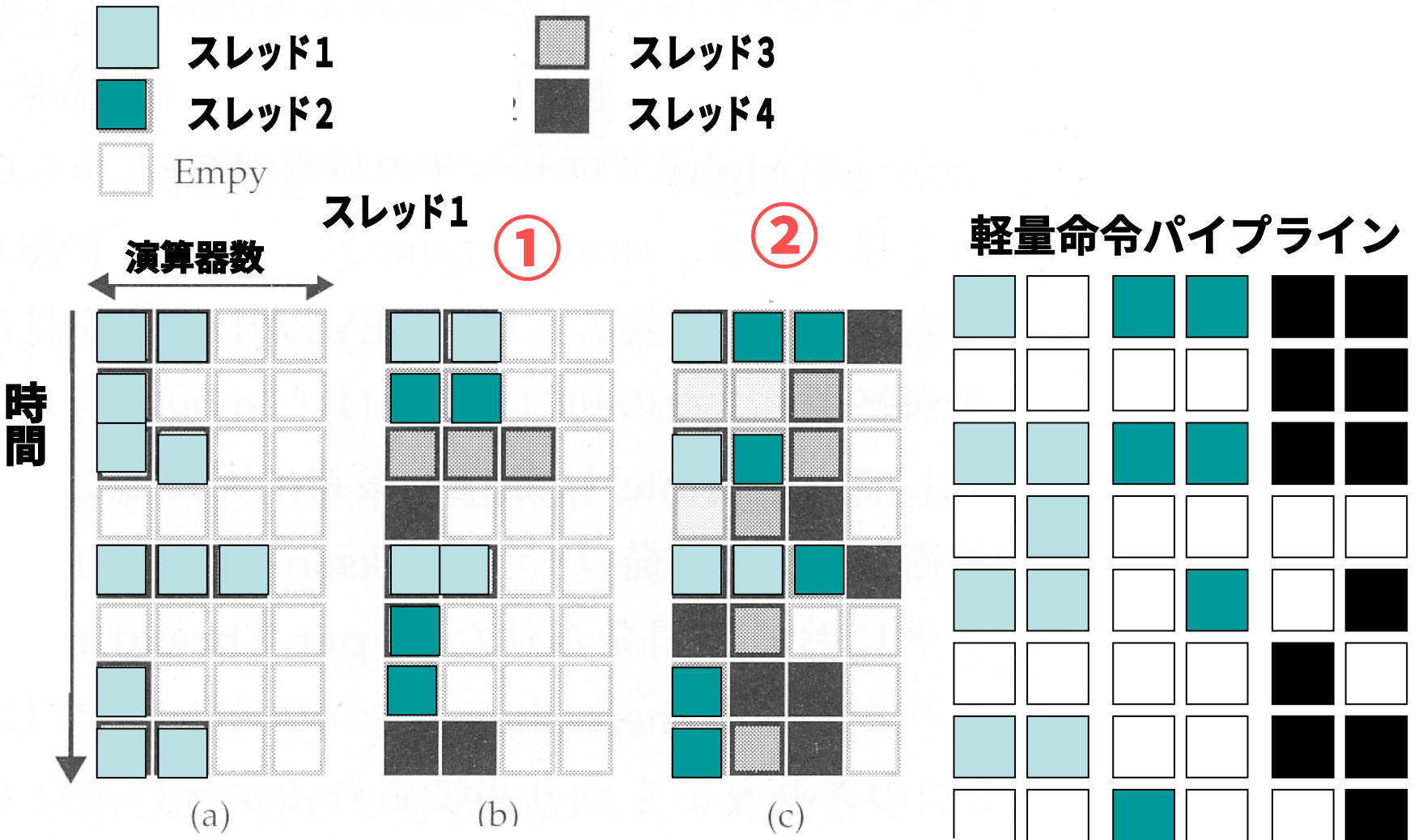
Microprocessor Report

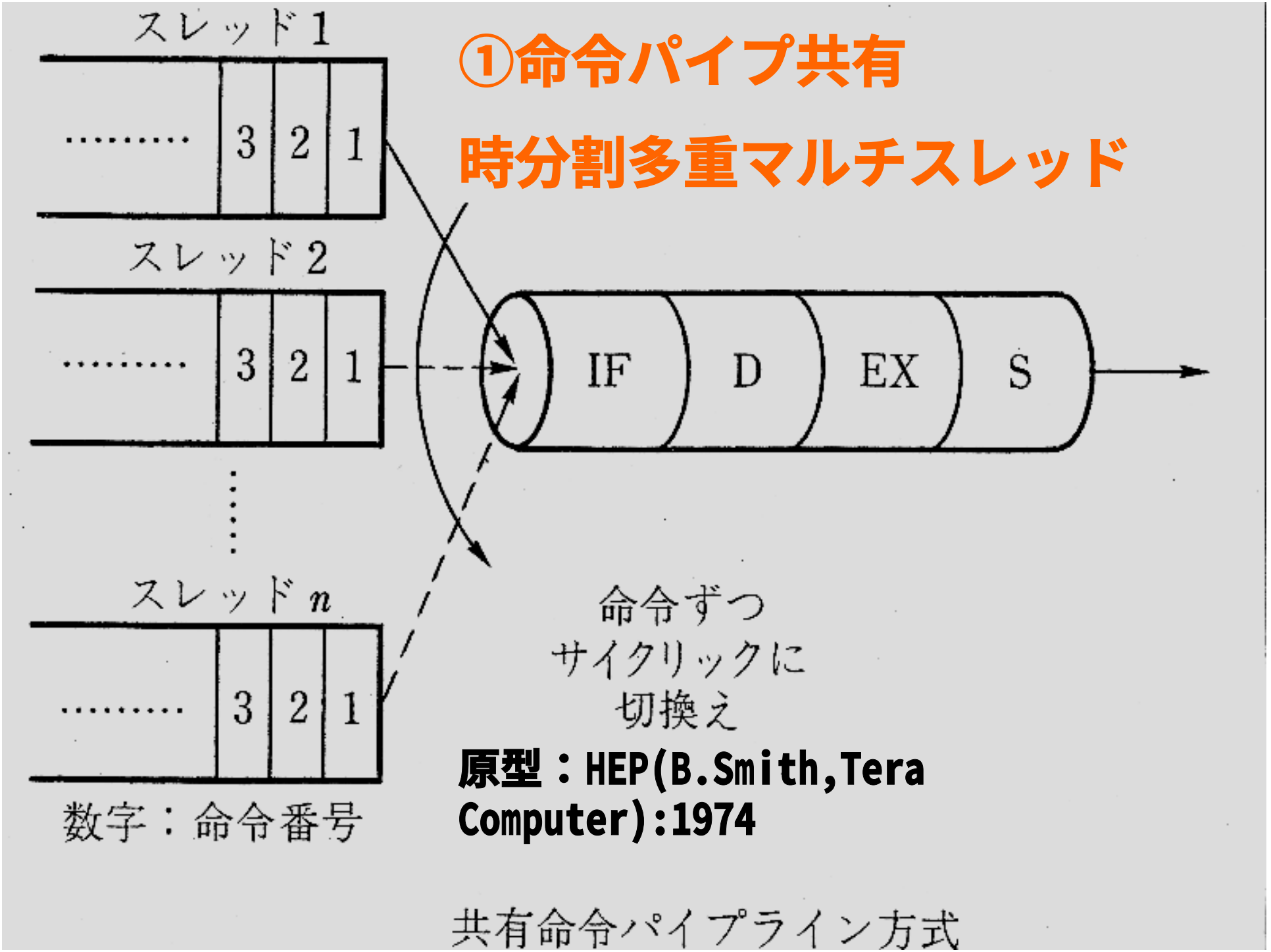
May 2004

Product	Pentium 4	Pentium 4	Pentium 4	Pentium 4	Pentium 4	Pentium 4 EE	Pentium 4 EE	Pentium 4	Pentium 4
Die Code Name	Willamette	Willamette	Willamette	Northwood	Northwood	Gallatin	Gallatin	Prescott	Prescott
Process	180nm	180nm	180nm	130nm	130nm	130nm	130nm	90nm	90nm
Frequency	1.80GHz	2.0GHz	2.0GHz	2.0GHz	3.4GHz	3.2GHz	3.2GHz	3.2GHz	3.4GHz
VID	1.75V	1.75V	1.75V	1.50V	1.55V	1.6V	1.6V	1.4V (max)	1.4V (max)
Icc Max	52.7A	55.0A	57.4A	44.3A	71.6A	67.4A	77.7A	78A	91A
Isg	11.1A	11.3A	16.7A	27A	27A	32A	35A	40A	50A
TDP (C)	66.7W	71.8W	75.3W	52.4W	89W	92W	102.9W	89W	103W
L2/L3 Cache Size	256K L2	256K L2	256K L2	512K L2	512K L2	512K L2/1M L3	512K L2/1M L3	1M L2	1M L2
Transistors	42Mil	42Mil	42Mil	55Mil	55Mil	178Mil	178Mil	125Mil	125Mil
Package	423pin PGA	423pin PGA	478 PGA	478 PGA	478 PGA	478 PGA	478 PGA	478 PGA	478 PGA

7.10.2マルチコア型プロセッサの分類

①命令パイプ共有時分割多重マルチスレッド
（粗粒度、細粒度）
Temporal　Multi-Threading（TMT)
②命令パイプ共有同時多重マルチスレッド
Simultaneous　Multi-Threading(SMT)
③命令パイプ多重マルチスレッド
④CMP（**チップマルチプロセッサ**）


スレッド1
スレッド2
スレッド3
スレッド4
Empy
スレッド1
①
②
軽量命令パイプライン
演算器数
時間
(a)
(b)
(c)
スーパスカラ
命令パイプ共有
時分割多重マ
ルチスレッド
命令パイプ共有
同時多重マルチ
スレッド(SMT)
スレッド１ スレッド２ スレッド３
命令パイプ多重
マルチスレッド
／CMP ③、④


スレッド1
3
2
1
スレッド2
3
2
1
スレッドn
3
2
1
数字：命令番号
IF
D
EX
S
命令ずつ
サイクリックに
切換え
①命令パイプ共有
時分割多重マルチスレッド
原型：HEP(B.Smith,Tera Computer):1974
共有命令パイプライン方式

②SMT　Simultaneous Multithreading

命令パイプ共有同時多重マルチスレッド

基本命令パイプライン：1つ
　＋複数のレジスタ、PCなど
多数のスレッドの実行
Intel Hyper Threading
　15-25%性能向上、
　　チップサイズ5%増
　　2スレッド実行

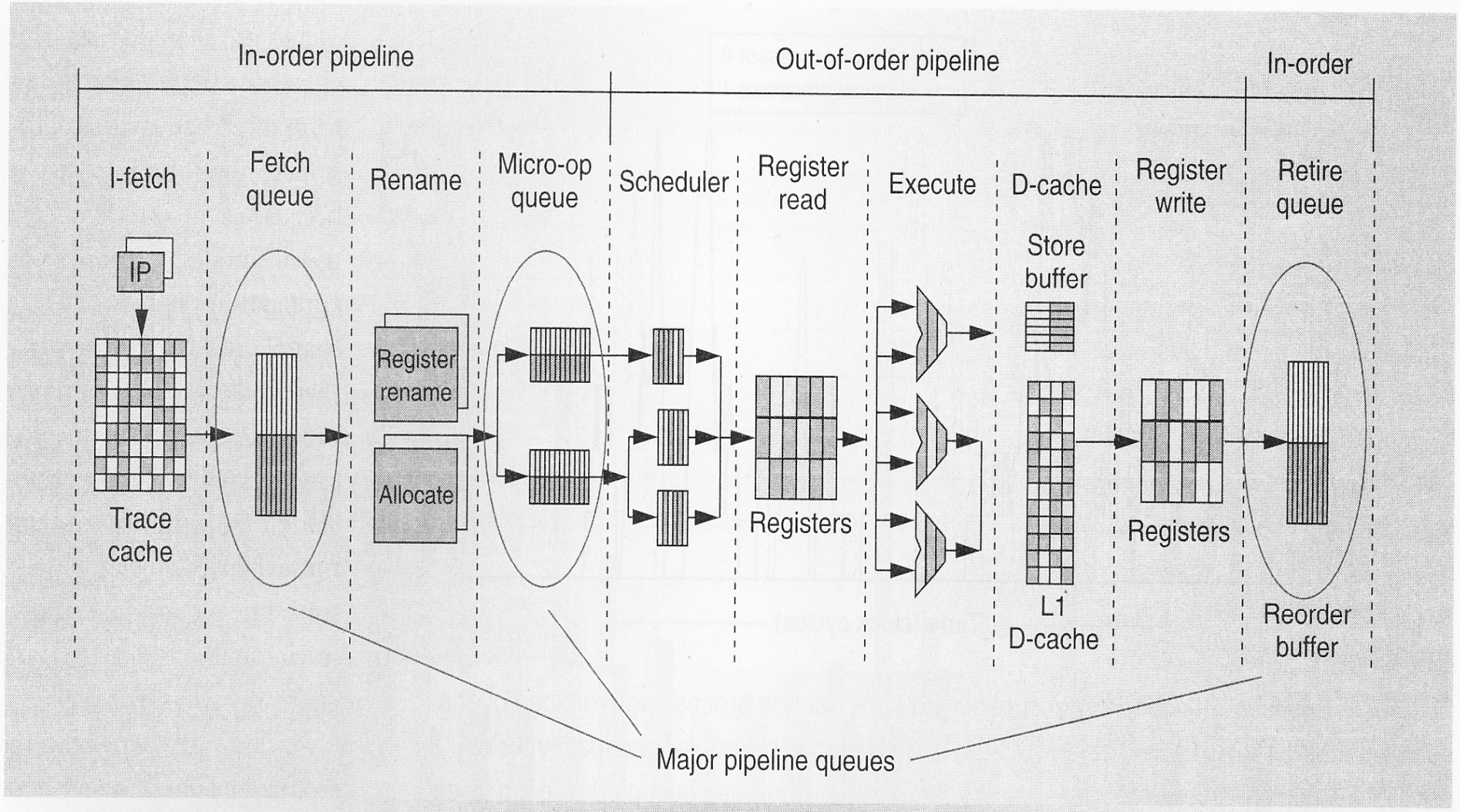
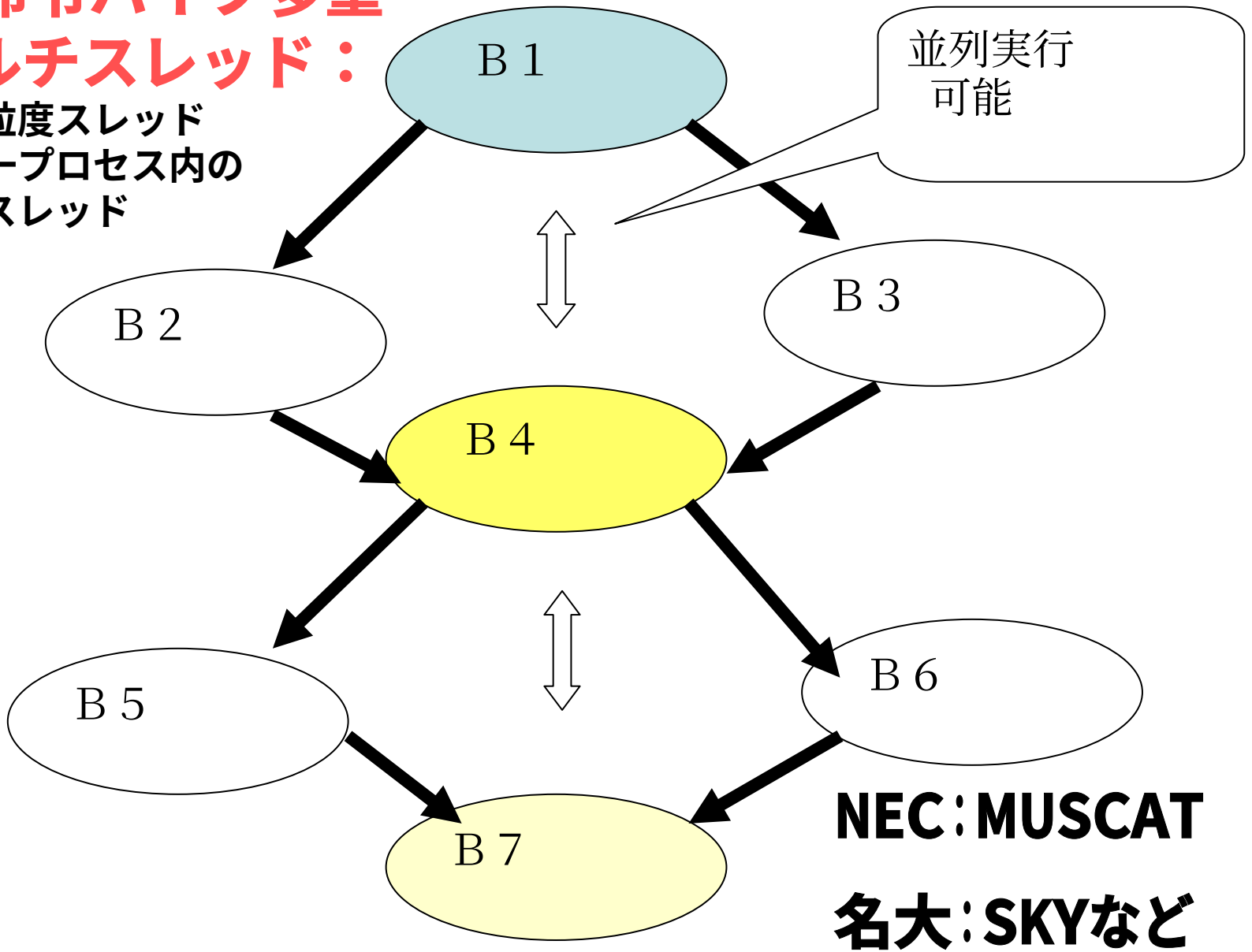


Figure 4. In this view of a Netburst microarchitecture's execution pipeline, the light and dark areas indicate the resource utilization of the two software threads running on the two logical processors.

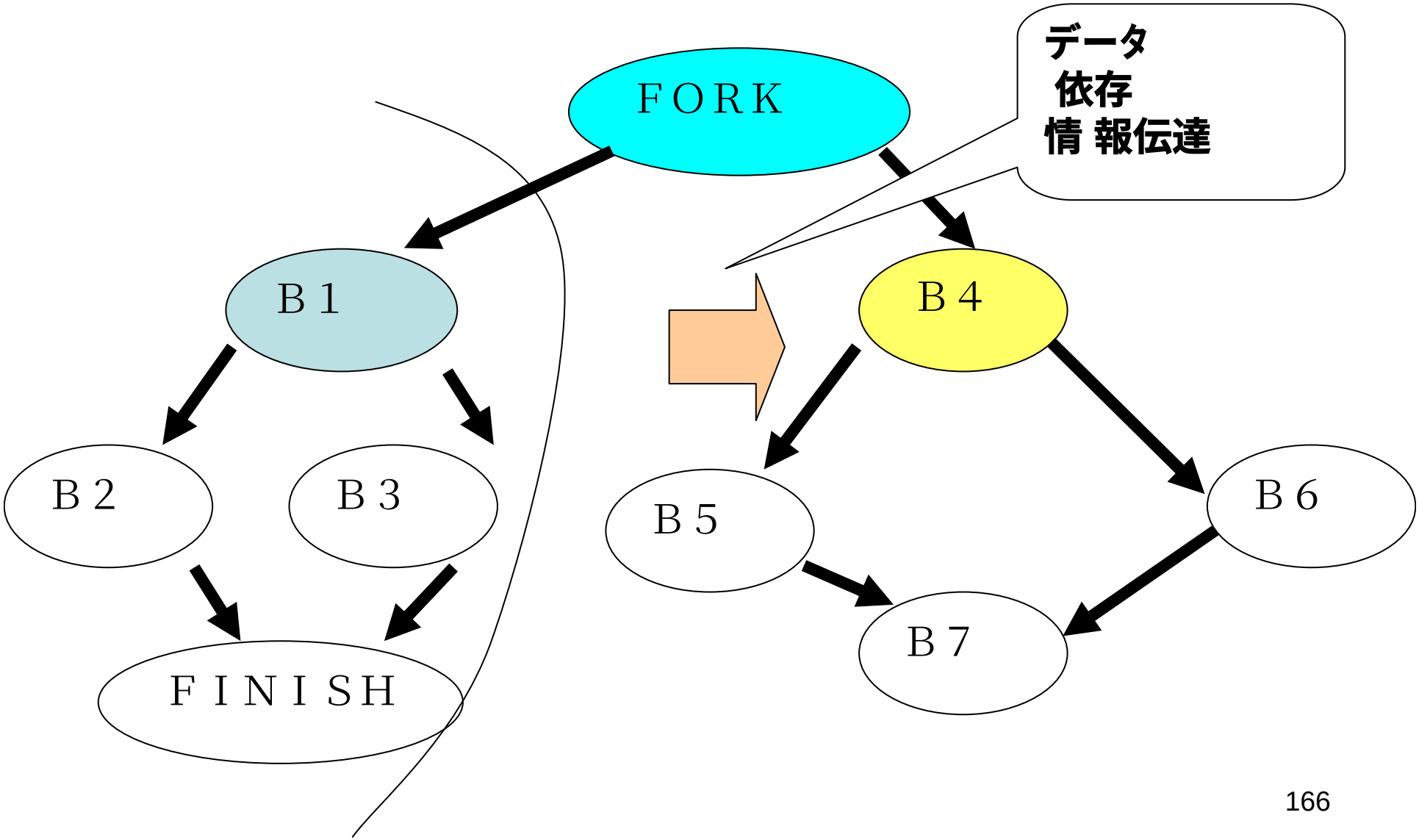
③命令パイプ多重
マルチスレッド：

細粒度スレッド
単一プロセス内の
スレッド

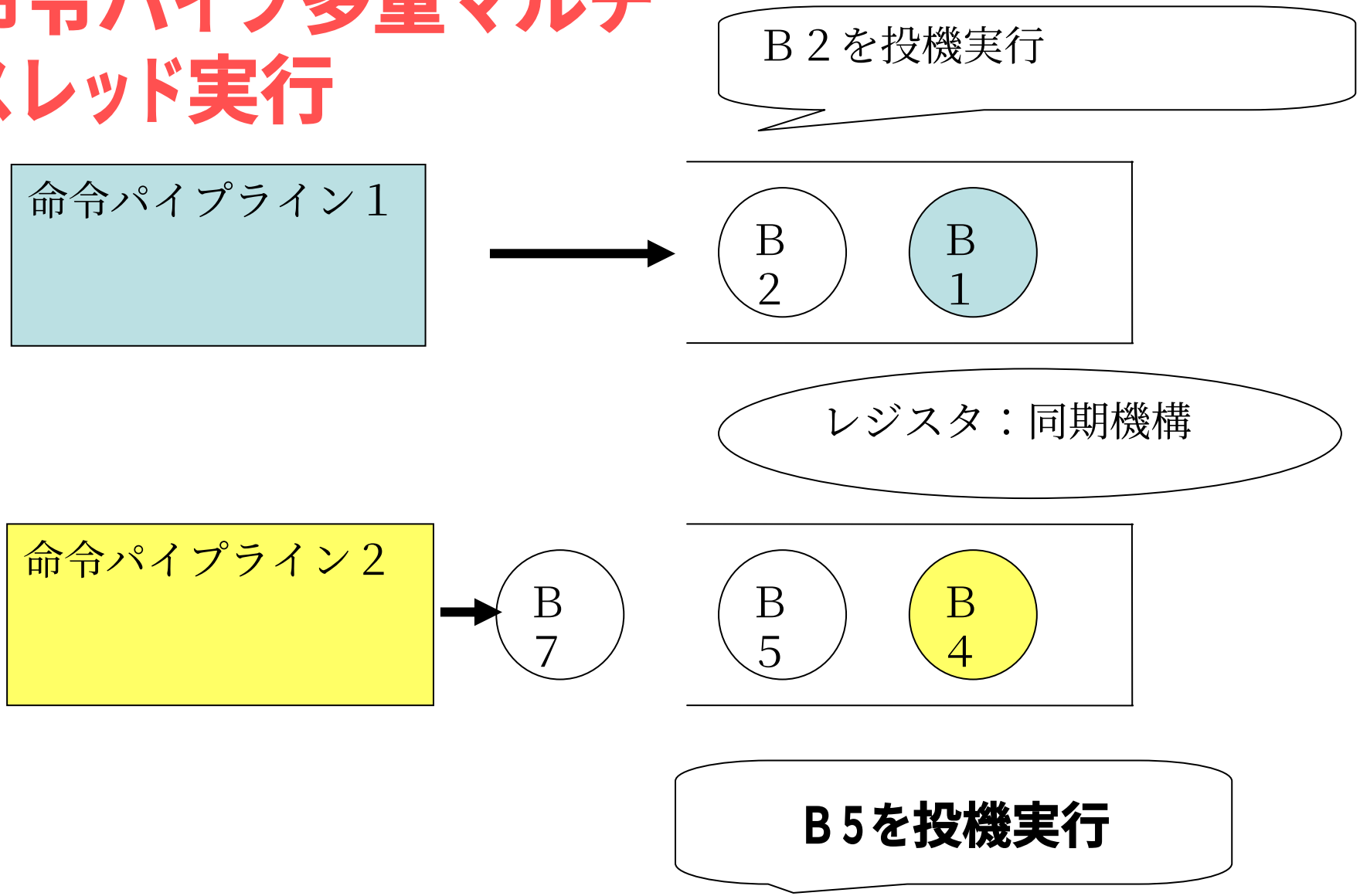


NEC：MUSCAT

名大：SKYなど


ＦＯＲＫ
データ
依存
情 報伝達
Ｂ１
Ｂ２
Ｂ３
Ｂ４
Ｂ５
Ｂ６
Ｂ７
ＦＩＮＩＳＨ

命令パイプ多重マルチスレッド実行



NEC　MUSCAT

表 2: シミュレーションパラメタ

項目	パラメタ
各 PE の構造	
パイプライン	IF, ID, Issue/Reg, EX, WB, Graduate Issue/Reg から WB まで out-of-order 実行
命令ウィンドウ	PE 毎 32 命令 (整数 / ロードストア各 16)
演算リソース	ALU 2×PE 数 L/S パイプ 1×PE 数
ロード / ストア	3 サイクルレイテンシ
スーパスカラ度	4 命令同時デコード / 終了
分岐投機	4 分岐まで仮実行
分岐履歴	2048entry 4 状態 (PE 間共有)
キャッシュ	
キャッシュ方式	命令／データ分離
キャッシュ容量	各 32Kbyte (64*byte* × 512*entry*)
マッピング方式	4Way Set Associative LRU 追い出し

鳥居、近藤、本村、西、小長谷：On Chip Multiprocessor **指向制御並列アーキテクチャ**MUSCAT**の提案**、JSPP**97**、pp.229-236、1997

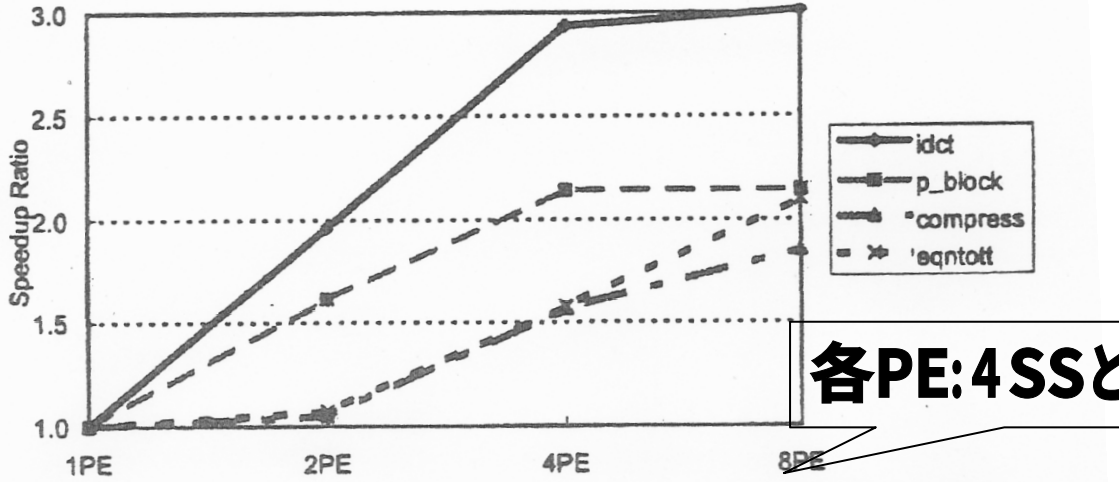


図 9: MUSCAT の性能向上率

各PE:4SSと等価

8PEと32SS
ハード規模同じ

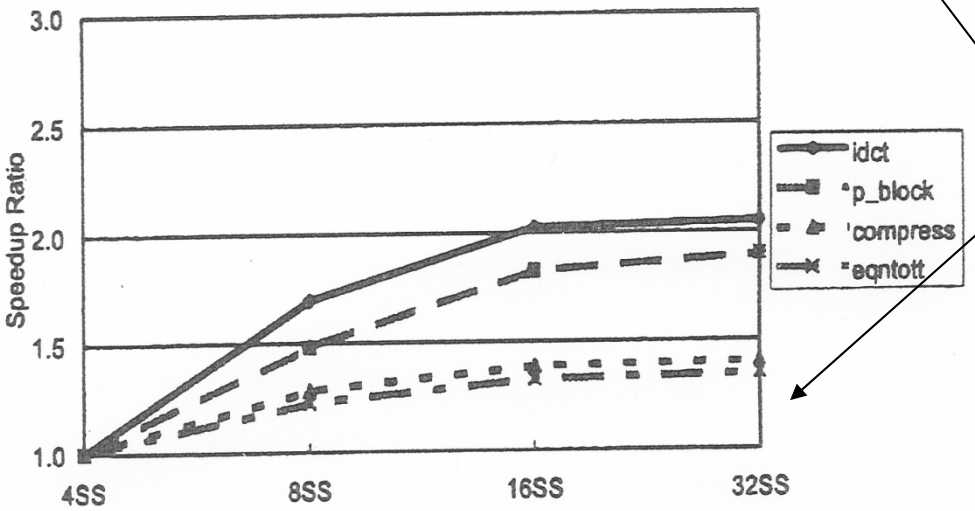


図 11: スーパスカラ強化モデルの性能向上率

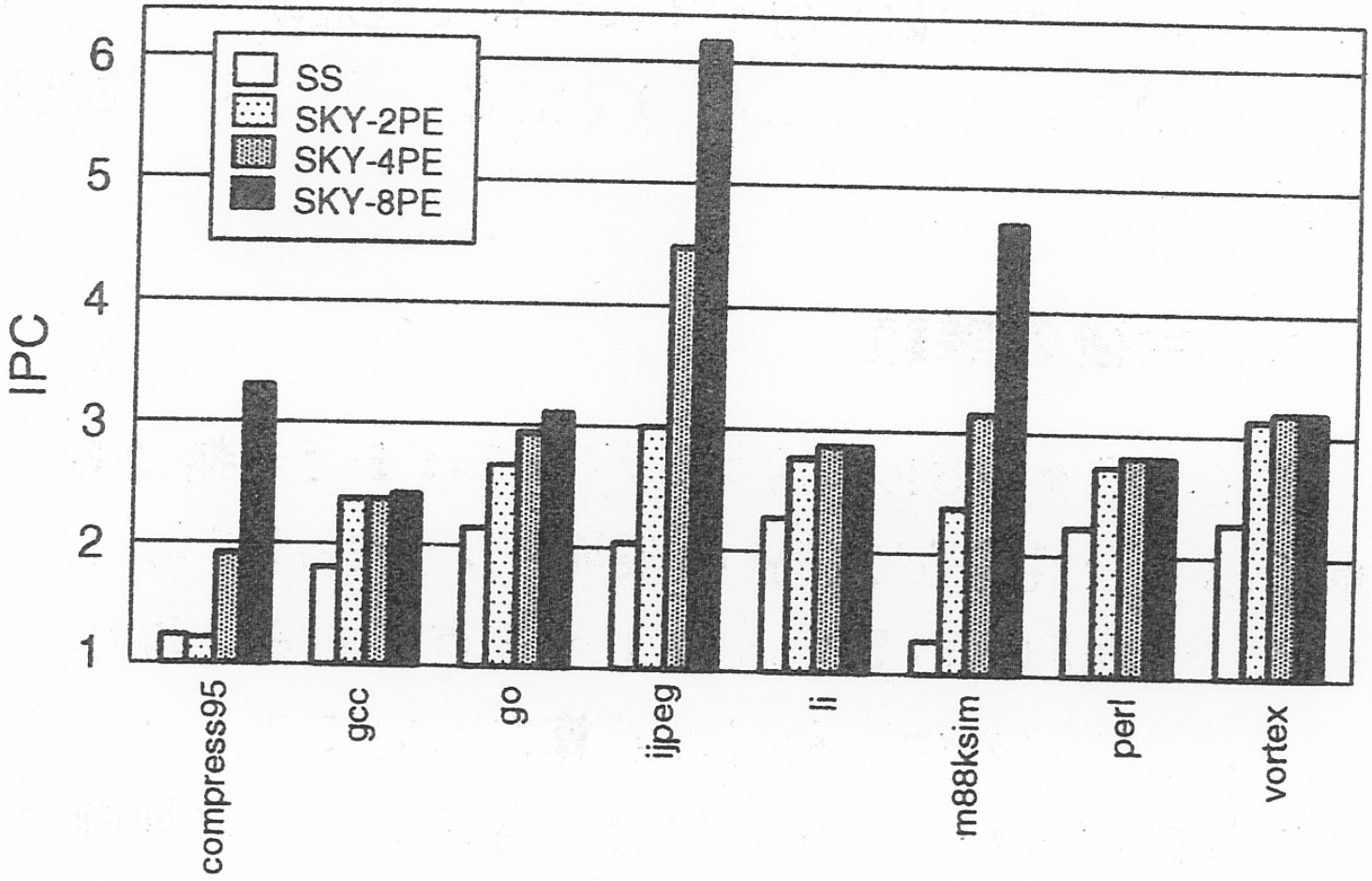


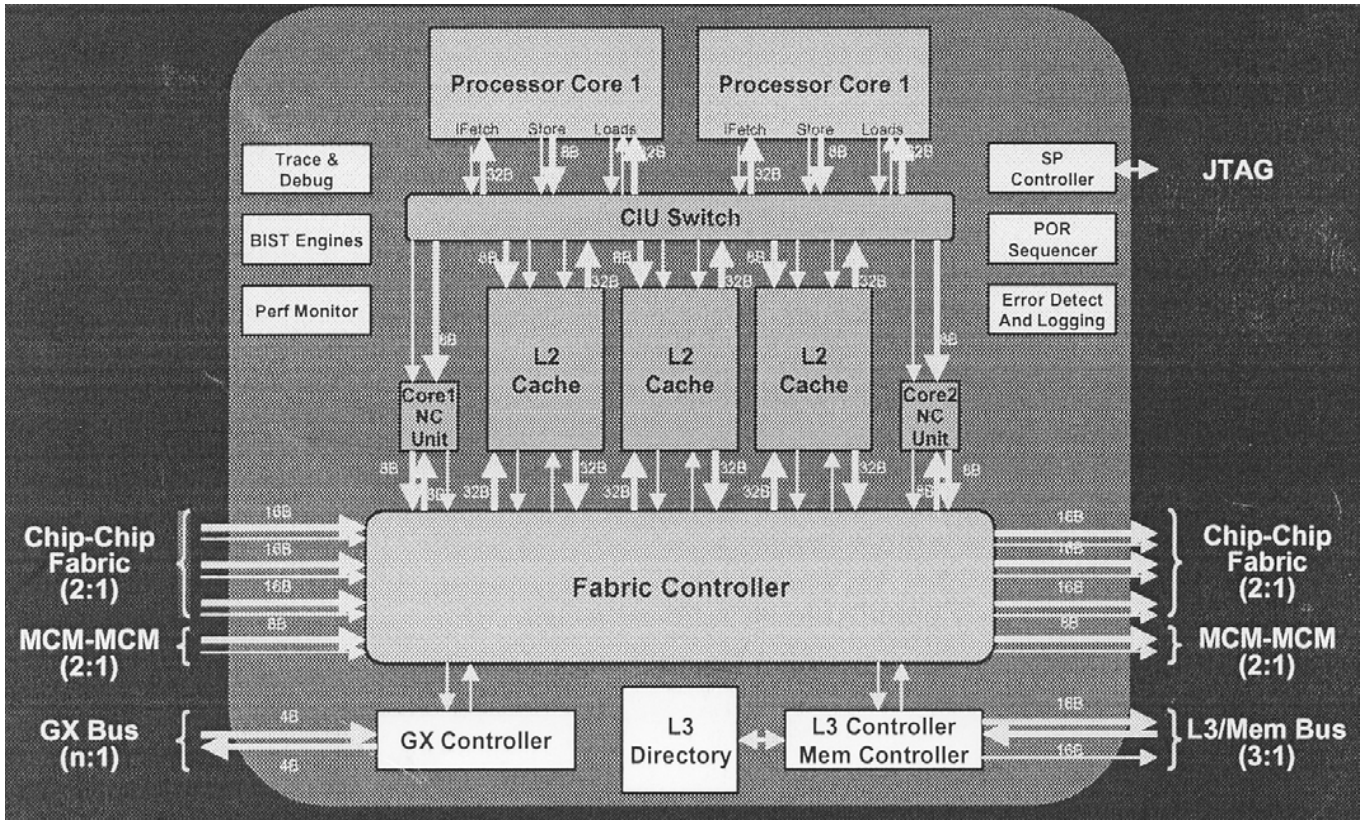
図6　SKYの台数効果

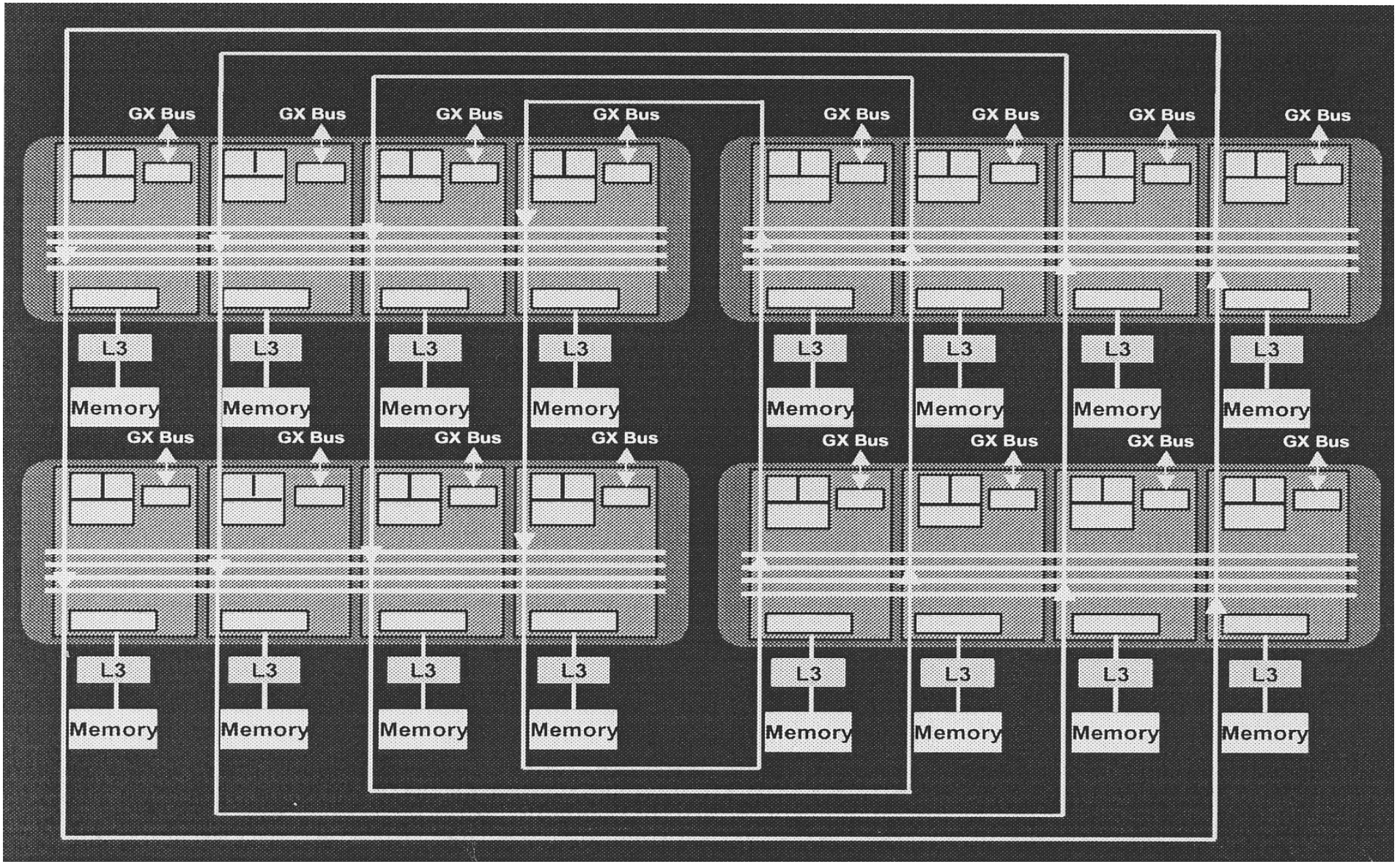
SS:8多重

小林、岩田、安藤、島田：非数値計算プログラムのスレッド間命令レベル並列を利用するプロセッサアーキテクチャSKY、JSPP98、pp。87－94、1998

④オンチップマルチプロセッサ　CMP

- IBM POWER4：2台のSMP、スヌープキャッシュ
- MCM：8台までのSMP、最大構成：4MCM（32台）
- L1（I:64KBx2,D:32KBx2）：2状態、L2（1．5MB）7状態、L3（32MB）：5状態




GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
L3
L3
L3
L3
L3
L3
L3
L3
Memory
Memory
Memory
Memory
Memory
Memory
Memory
Memory
GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
GX Bus
L3
L3
L3
L3
L3
L3
L3
L3
Memory
Memory
Memory
Memory
Memory
Memory
Memory
Memory

Pentium EE840

デュアルコアプロセッサ → **４個のプロセッサ**

各プロセッサ：ハイパースレッデイング

９０nmデザインルール

3.2GHz

各プロセッサ

- L1**データキャッシュ：１６**KB
- L1**トレースキャッシュ１２**KB
- L2**キャッシュ：１**MB
- SSE3

Intel Core 2 Duo 2006 7 27発表

デュアルコア

14段パイプライン

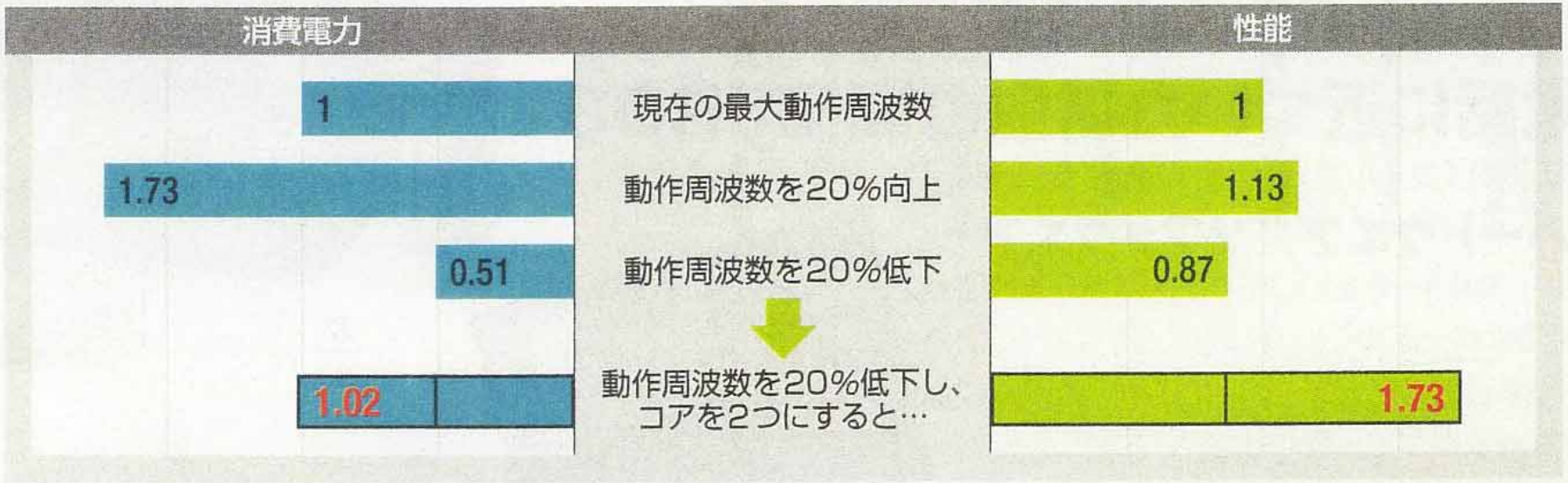
65nmデザインルール、29,100万Tr、L2：4MB

●Coreマイクロアーキテクチャーが採用した技術

技術名	概要
ワイド・ダイナミック・エグゼキューション	1クロック当たりの命令実行数を向上させる技術。同時実行可能な命令数を4つに増やした。また、連続して使われることの多い命令を合体させ、まとめて実行する機能（マクロフュージョン）も搭載
インテリジェント・パワー機能	CPUの各部分の消費電力をリアルタイムで管理し、あまり利用されていない部分の電力を最小限に抑える
アドバンスト・スマート・キャッシュ	2つのコアが、1つの2次キャッシュを共有する。一方のコアが使ったデータをもう一方のコアが使う場合、別途メモリーにアクセスする必要がなくなるなどのメリットがある
スマート・メモリー・アクセス	データの書き出しと読み込みの命令が続いた場合、書き出しの命令を実行する前に、次の命令が利用するデータを読み込む。またデータの利用パターンを分析・予測することで、必要とされるであろうデータを事前にキャッシュに取り込んでおくことも可能
アドバンスト・デジタル・メディア・ブースト	画像や暗号の処理など高度な数値計算で用いられる、SIMD拡張命令（SSE）の処理効率を、従来の倍に向上させた

●Core 2 Duoの設計思想

現在の動作周波数での性能と消費電力を1としたときの相対値で表示



インテルが開発者会議で発表した値。現在の動作周波数からさらに20%周波数を上げても、性能は1.13倍にしかならないが、消費電力は1.73倍と大幅に増えてしまう。逆に20%周波数を下げると、性能は0.87倍になるが、消費電力は0.51倍とほぼ半減する。それならば、動作周波数を落としてコアを複数にすることで消費電力と性能のバランスをとろうというのがCore 2 Duoの設計思想だ

日経パソコン2006．8．14

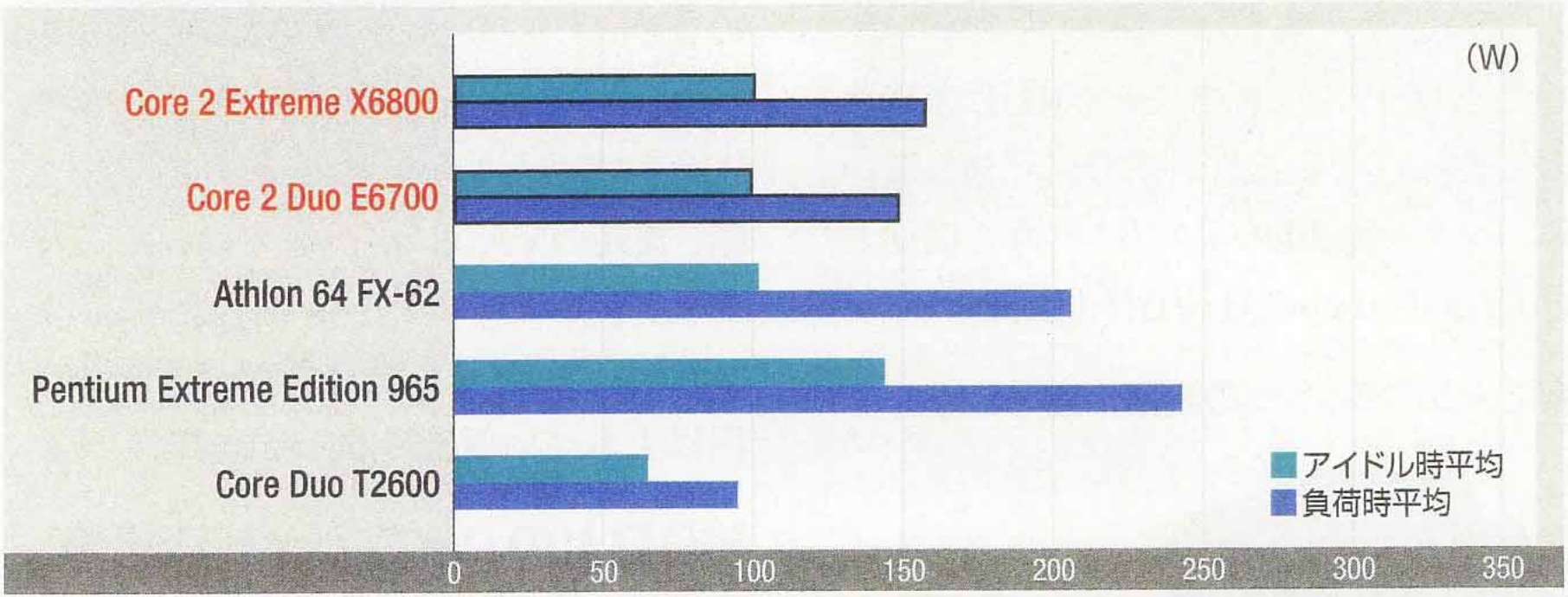
●従来版デュアルコアCPUを引き離す性能

Core 2 Extreme X6800およびCore 2 Duo E6700と、従来のハイエンドデスクトップ向けデュアルコアCPU「Pentium Extreme Edition 955」(3.46GHz)の性能を比較した。Core 2系の2製品はPentium Extreme Editionの性能を大きく上回った。測定には「PCMark05」を利用。いずれも、マザーボードはインテルの「D975XBX」を使用し、2GBのメモリー(DDR2-667)を搭載して測定した

●AMDのハイエンドCPUと比較しても優位

米AMDのハイエンドCPU「Athlon 64 FX-62」(2.8GHz)とCore 2 Extreme、Core 2 Duoの性能を比較した。Core 2系の2製品はAthlonの性能を上回った。測定には「PCMark05」を利用。マザーボードは、Athlon 64 FX-62は「M2N32-SLI Deluxe」(アスーステック・コンピュータ製：AMD提供の評価キット)、Core 2系はインテルの「D975XBX」を使用した。いずれも、1GBのメモリー(DDR2-800)を搭載して測定した

●消費電力の測定結果



Core 2系の2製品はノートパソコン向けの「Core Duo」よりも消費電力は大きいが、Pentium Extreme EditionやAthlon 64 FX-62よりも低い。特に、負荷時の差が大きいのが分かる

データ提供「日経WinPC」編集部

日経パソコン2006.8.14

Itanium2-Montecito

2個のItanium2プロセッサ、2スレッド

時分割多重マルチスレッド命令パイプ共有 (TMT)

各クロセッサ

- L1**キャッシュ**16KB、**1サイクル**
- L2**キャッシュ**1MB、**5サイクル**
- L3**キャッシュ**12MB、**14サイクル**

スレッドスイッチ：15サイクル (粗粒度TMT)

- L3**キャッシュミス、タイムアウトなど5つの事象**

1.72**億**TR

消費電力：100W

周波数：1.8GHz

IEEE Micro
Vol25 No2 2005

SPARC：Niagara

1多重で6ステージのSPARCプロセッサ　8台

各プロセッサ：4スレッド

時分割多重マルチスレッド命令パイプ共有

(細粒度TMT)

L1**キャッシュ：8**KB、Write　Through

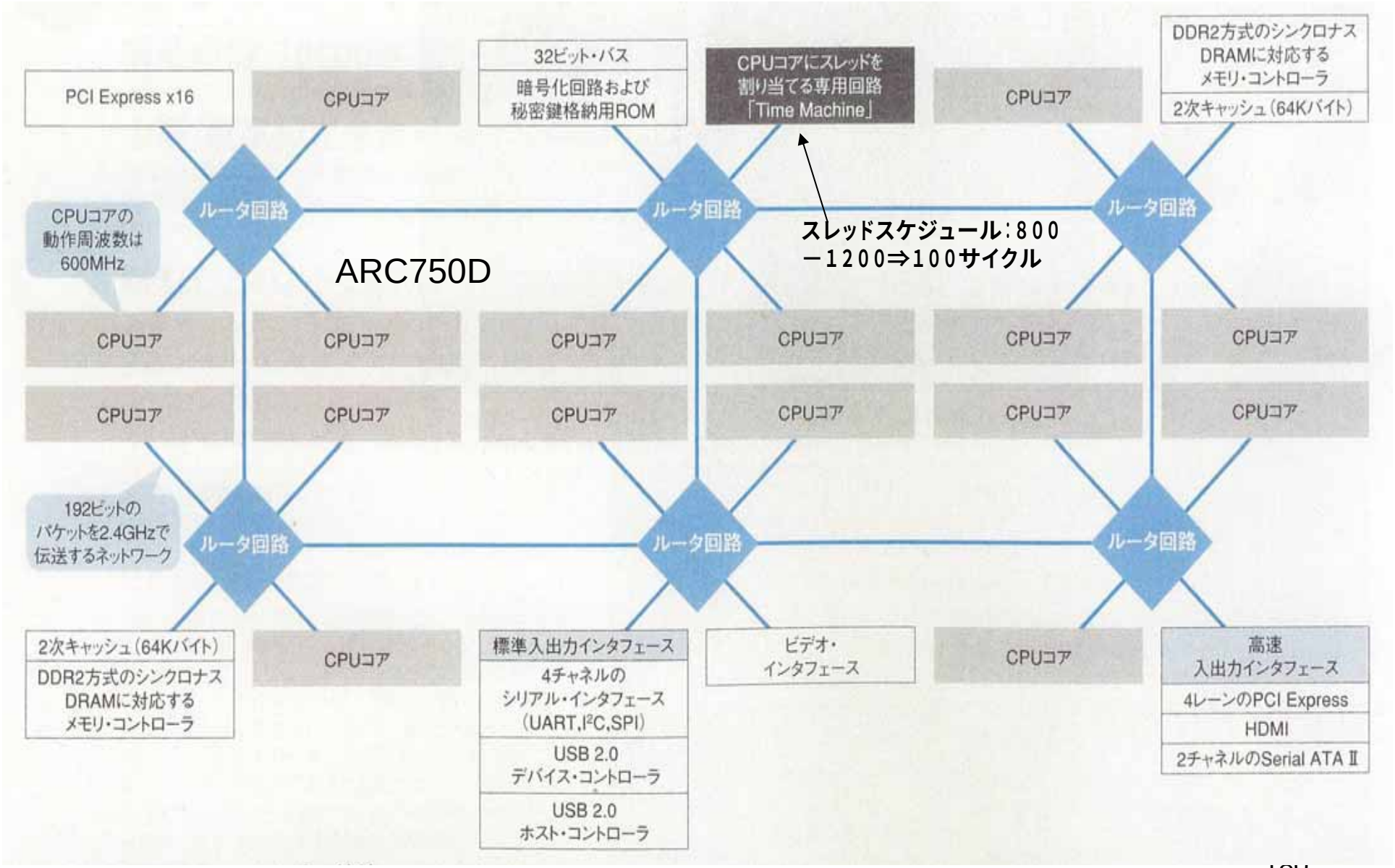
L2**キャッシュ：3**MB、**4バンク、クロスバSWで共有**

対応するバンクにL1キャッシュディレクトリのシャドウ

命令レベル並列が小さく、

スレッドレベル並列の大きなWEBやデータベース応用向き

ボストンサーキット社gCORE16：14GIPS、20W、共有メモリ



A.Kurland etal、**日経エレクトロニクス、2006．9．25**、pp.179-186

7.10.3マルチコアの相互結合網

評価項目

①スループット

②レイテンシ

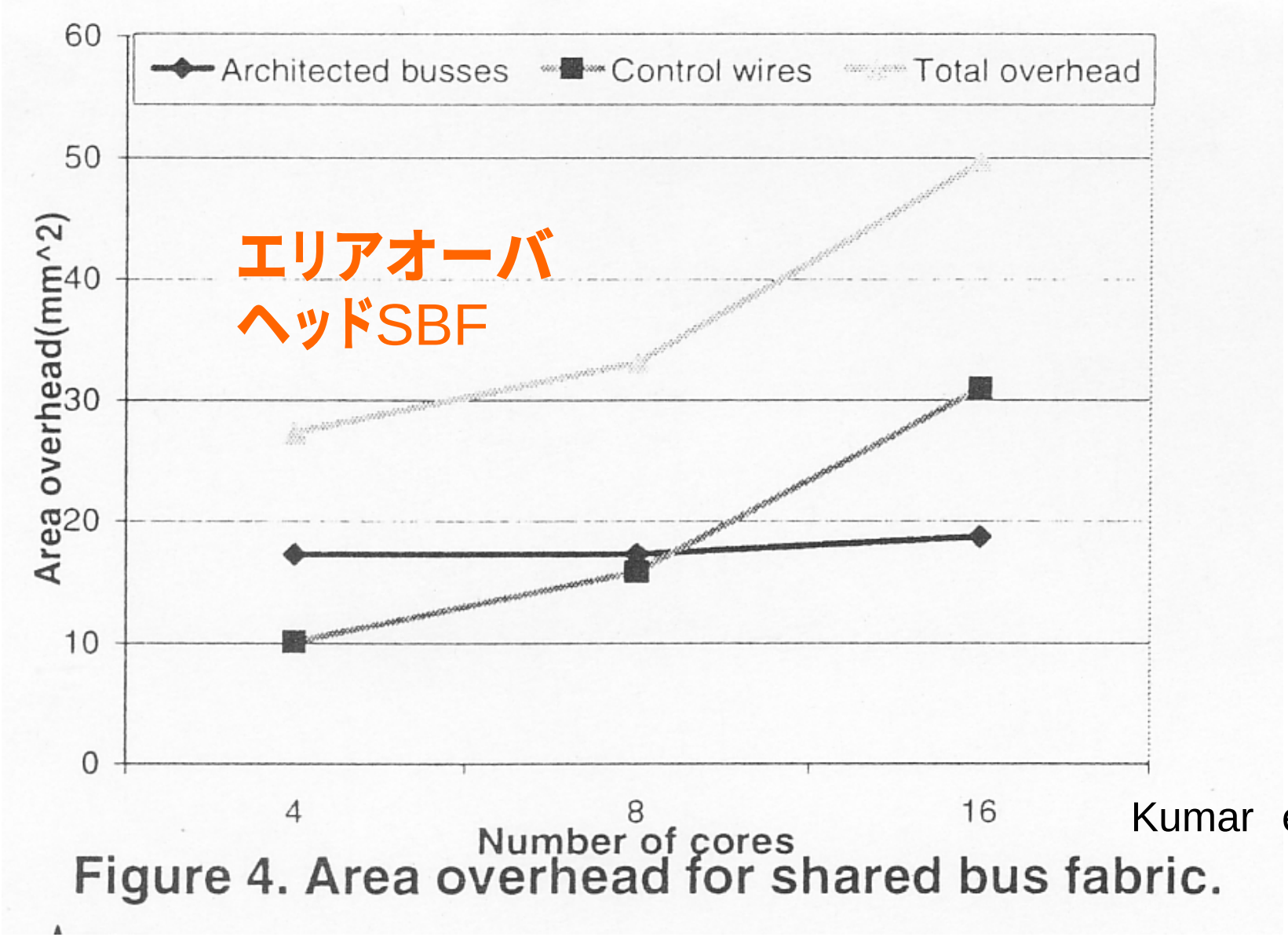
③消費電力：　スイッチ＋配線

④面積

⑤配線層数

・P.P.Pande etal: Performance Evaluation and Design Trade-offs for Network-on-Chip Interconnect Architectures, IEEE Trans. Computers, Vol54, No.8,pp.1025-1040,2005

・R.Kumar etal: Interconnections in Multi-core Architectures: Understanding Mechanisms, Overhead, and Scaling, pp.408-419, ISCA, 2005



Kumar etal

Figure 4. Area overhead for shared bus fabric.

１６コアで配線面積５０mm^2 ⇔65nm**デザイン、ダイサイズ**：400mm^2、**１コア**：**１０**mm^2、L2**キャッシュ**：0.125MB／mm^2**を仮定**

ロジック(SBF**の下で隠蔽**)： **５．６**mm^2(**４コア**)、**８．６**mm^2(**８コア**)、**１７．９４**mm^2(**１６コア**)

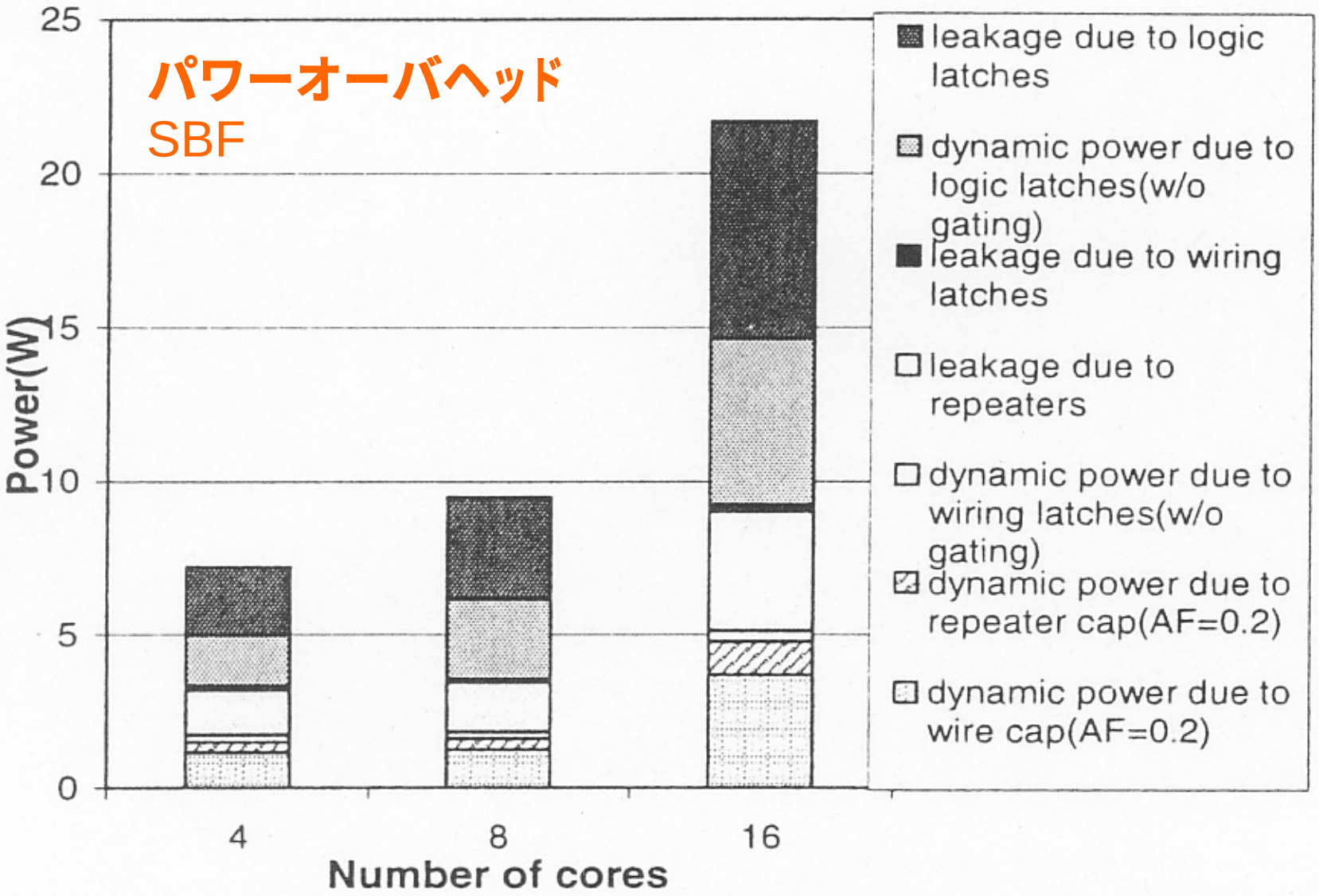


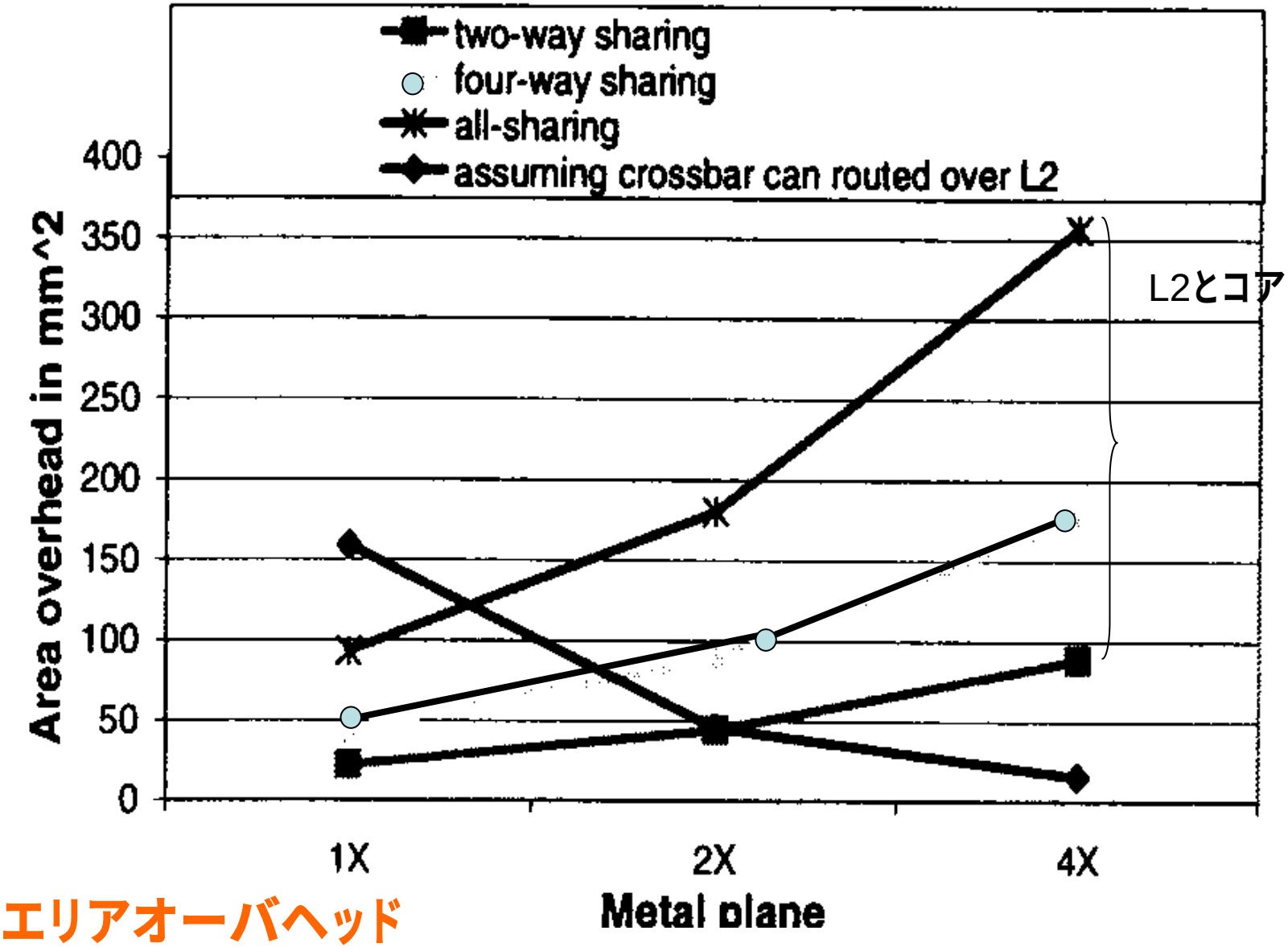
Figure 5. Power overhead for shared bus fabric.

1コア (Power4) **：10W**

オーバヘッド：2コア分に相当 (**22mm² 16コアの場合**)

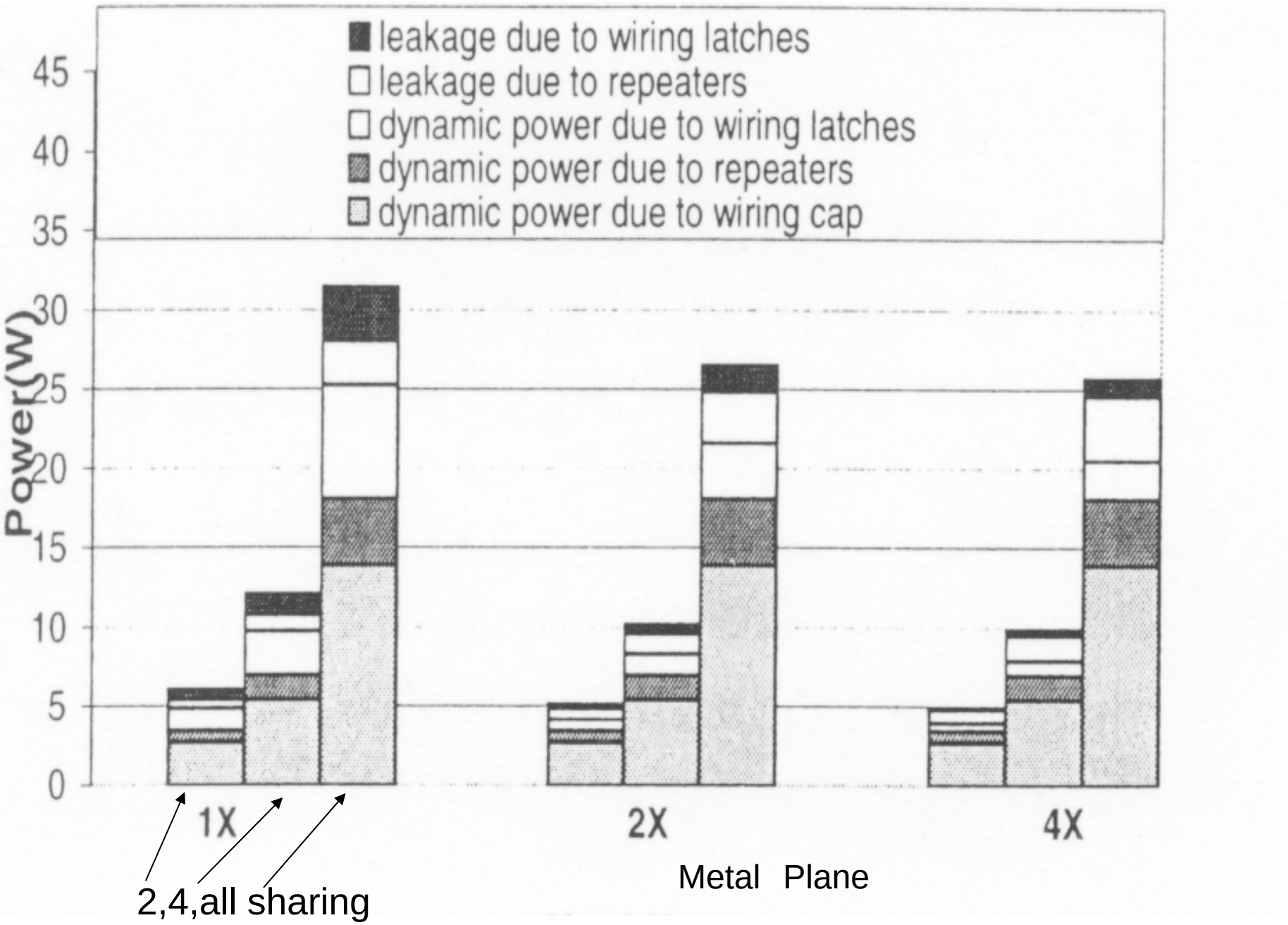
8コアX8L2キャッシュバンクのクロスバスイッチ

メタルプレーン：1X：0.5μm,2X：1.0μm,4X:2.0μm，8X:4.0μmピッチ

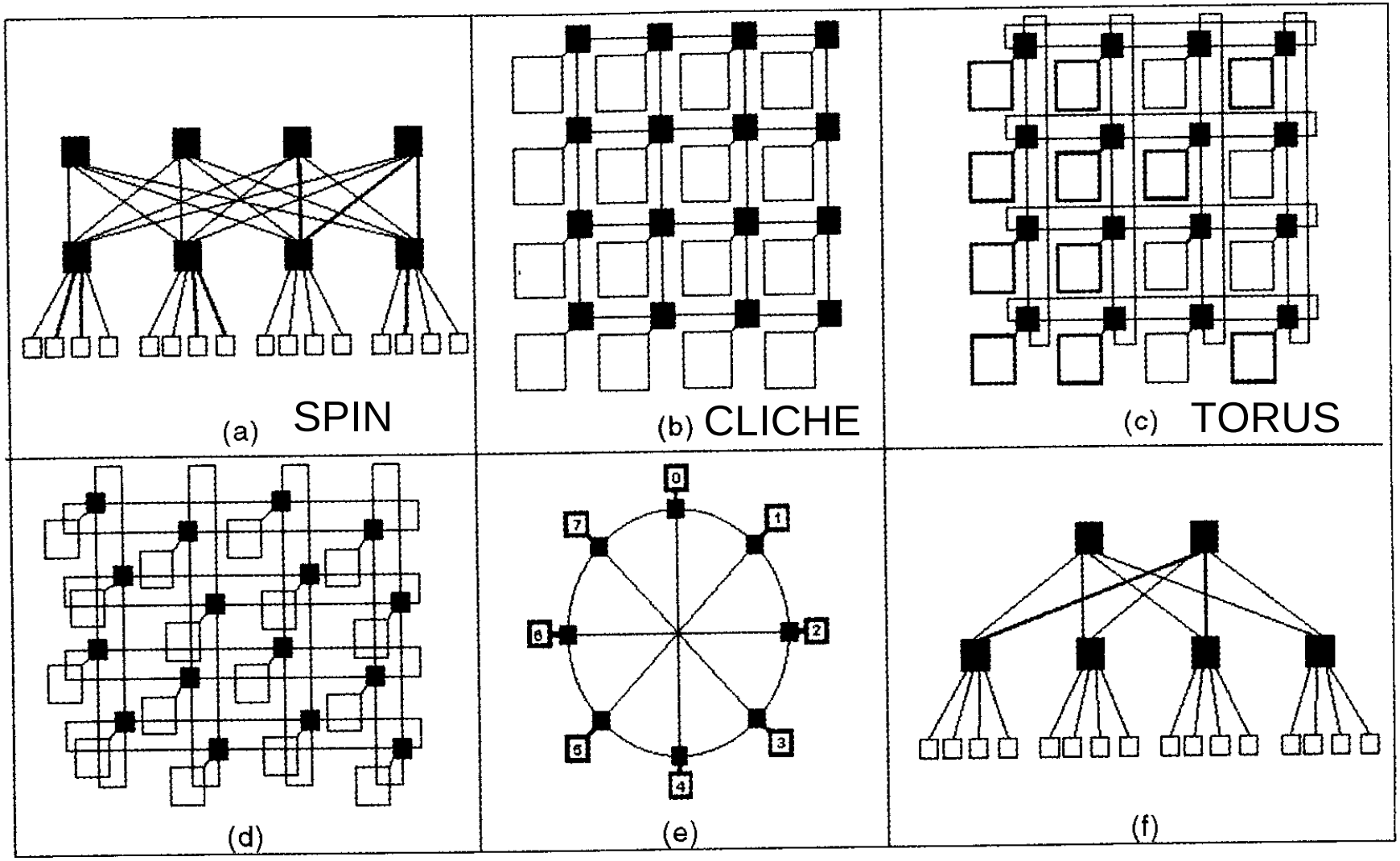


L2とコア同一面

エリアオーバヘッド



クロスバスイッチのパワーオーバヘッド



Folded TORUS OCTAGON BFT Pande etal

SPIN:Scalable,Programmable,Integrated Network, CLICHÉ: Chip-Level Integration of Communicating Heterogeneous Elements, BFT:Butterfly Fat Tree