コンピュータアーキテクチャ
2024

メモリシステムの高速化技法
メモリインタリーブ
キャッシュメモリ

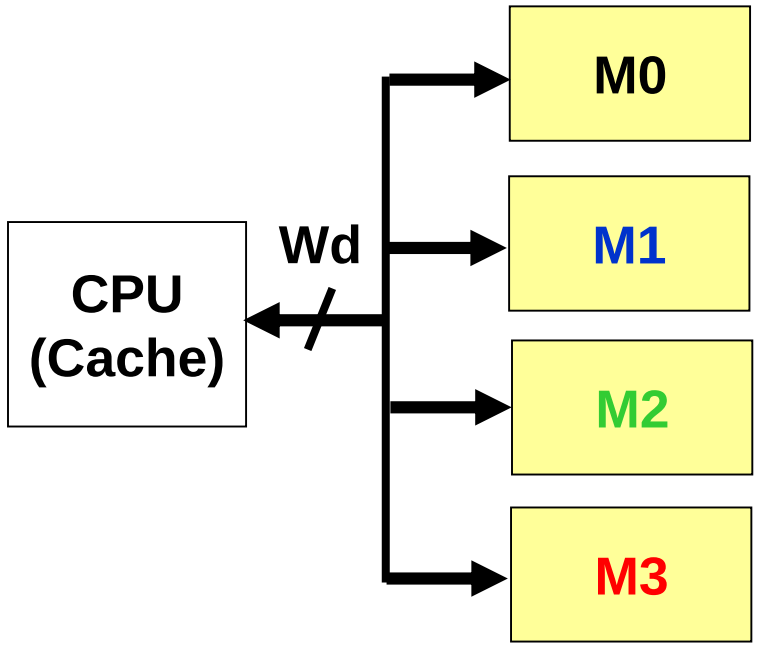
堤　利幸

CPUとメインメモリの速度差

- CPUはメインメモリよりも10(～100)倍の速い

- メインメモリの高速化技法

メモリインタリーブ



Four-way interleaved memory

キャッシュメモリ

- キャッシュメモリ
CPU(レジスタ)とメインメモリ間に置く，高速/小容量のSRAMで構成されるメモリ

- キャッシュメモリの効果
CPUから見たメインメモリの実効的なアクセス時間を短縮する効果がある.

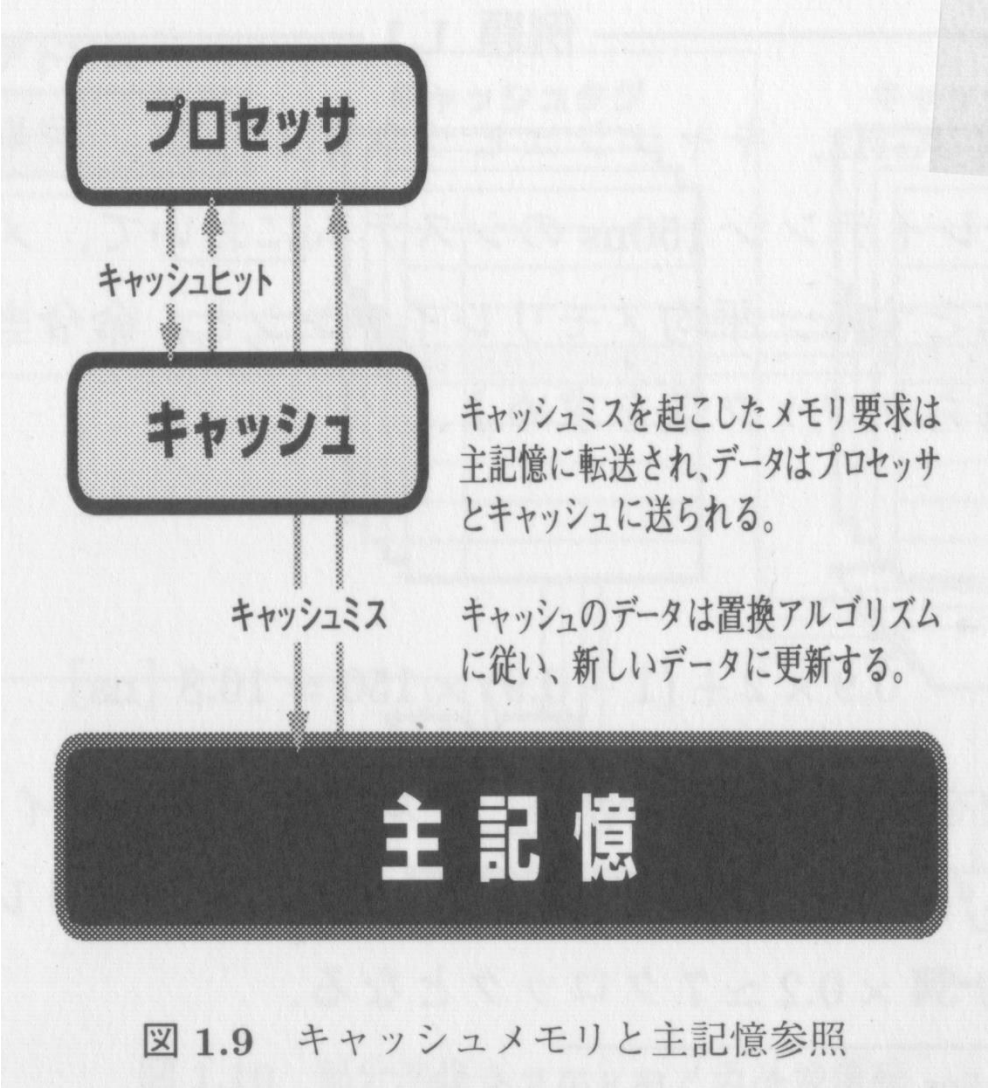
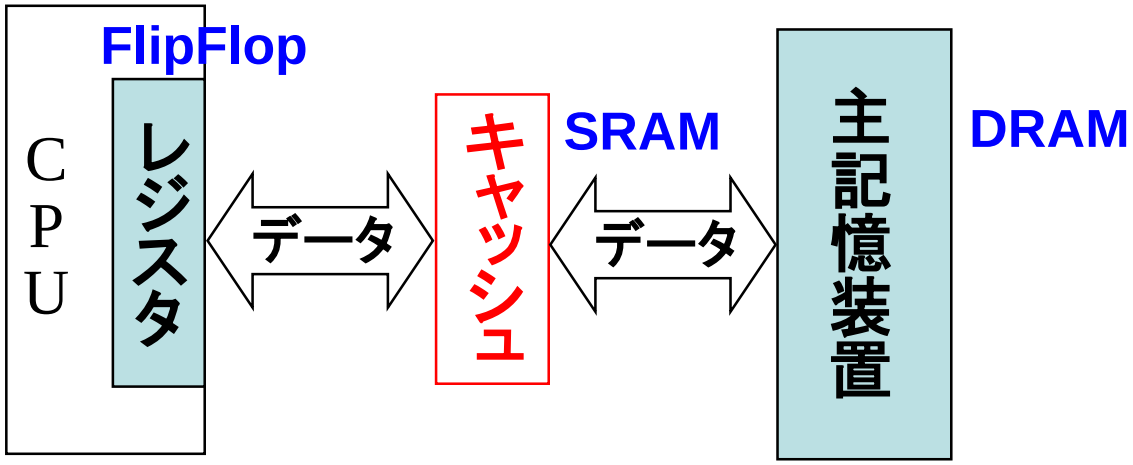


図 1.9　キャッシュメモリと主記憶参照

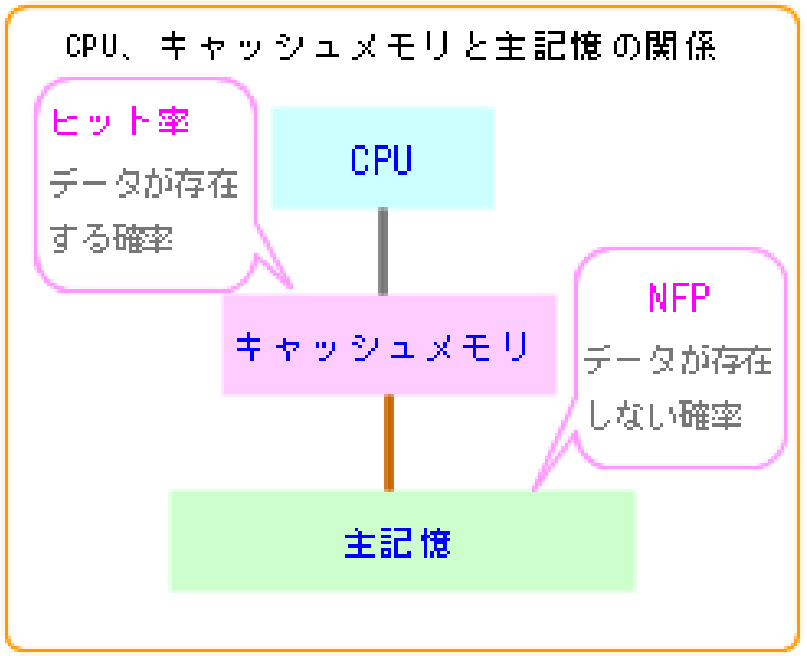
キャッシュメモリ



キャッシュメモリは，「時間の局所性」と「空間の局所性」を利用してレイテンシを引き下げる．
これらの局所性が強いほど，小容量のキャッシュにデータが存在する確率が高くなる．
「時間の局所性」は，プログラムを一度利用したデータを短時間のうちに再度利用する可能性が高いという性質．
「空間の局所性」は，プログラムを利用したデータの近傍のデータを利用する傾向が強いという性質．

キャッシュメモリの実効メモリアクセスタイム

CPUからの実効メモリアクセス時間（メモリレイテンシ)	T
キャッシュのメモリアクセス時間(メモリレイテンシ)	Tc
メインメモリのメモリアクセス時間(メモリレイテンシ)	Tm
キャッシュメモリのHit rate:	h
キャッシュメモリのMiss rate:	m = 1 -h

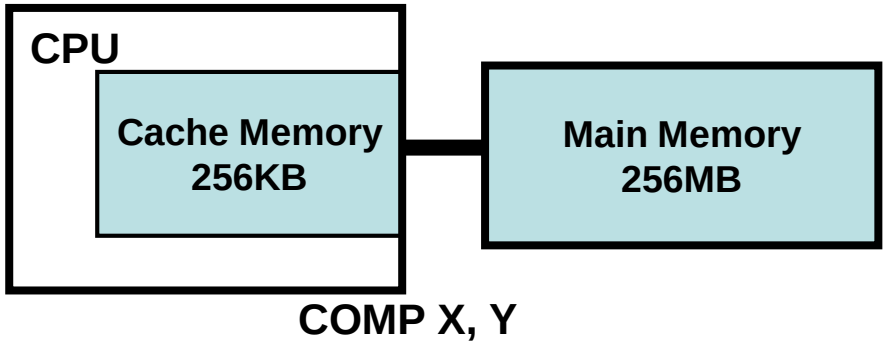


【例】

キャッシュメモリへのアクセス時間5ns,
主記憶装置へのアクセス時間が50ns,
キャッシュヒット率が0.97のとき、実効的なメモリアクセス時間 を求めよ.

キャッシュの計算

問．図に示す構成で，表に示すようなキャッシュメモリと主記憶のアクセス時間だけが異なり，ほかの条件は同じ 2 種類の COMP X と Y がある。
あるプログラムを COMP X と Y でそれぞれ実行したところ，両者の処理時間が等しかった。このとき，キャッシュメモリのヒット率は幾らか。ここで，CPU 処理以外の影響はないものとする。



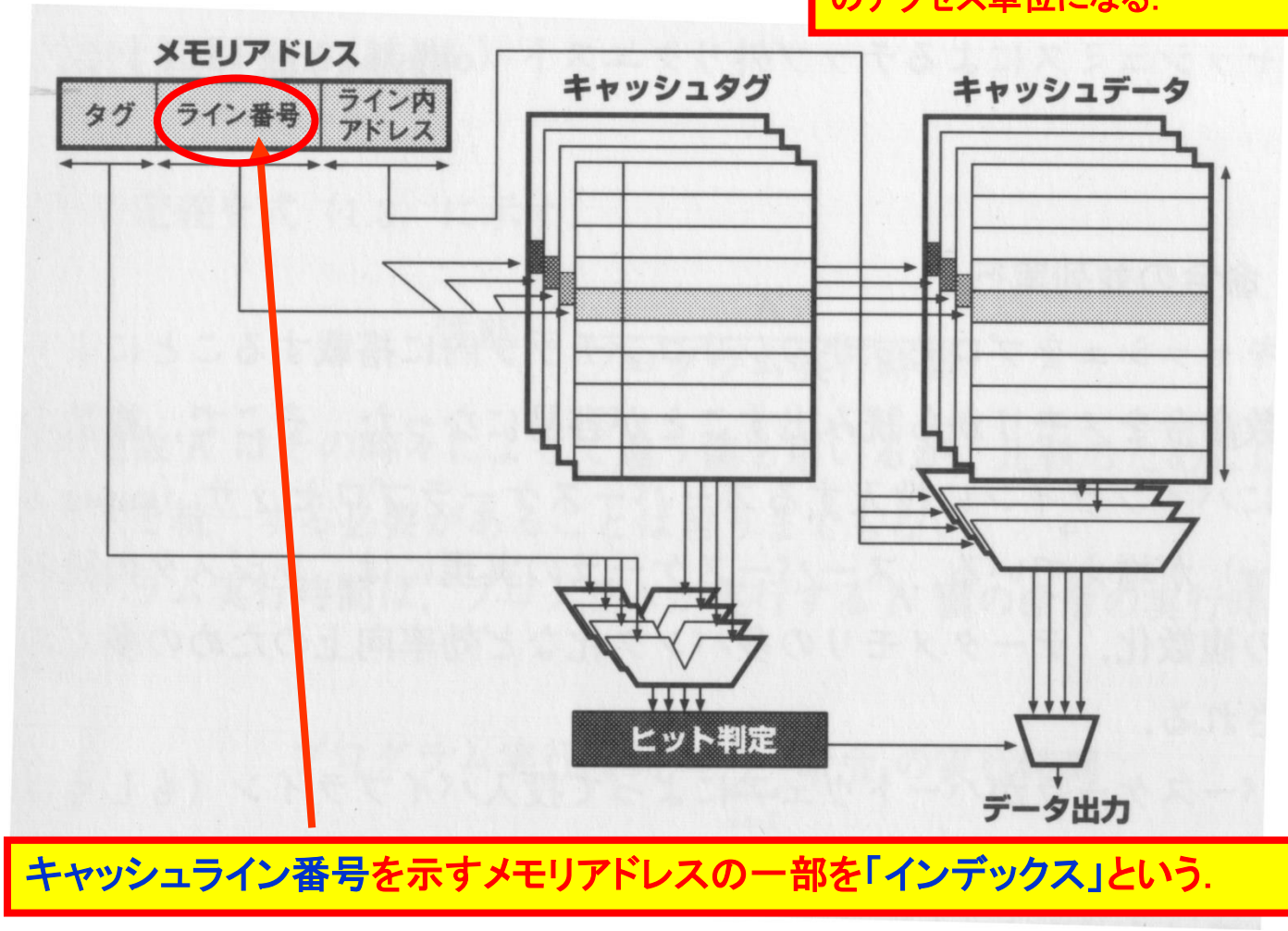
	COMP X	COMP Y
キャッシュメモリ	40 ns	20 ns
メインメモリ	400 ns	580 ns

ライトスルー方式とライトバック方式

キャッシュメモリへの書き込み方式	キャッシュメモリへの書き込み時の動作	キャッシュメモリによる高速化

キャッシュメモリの構造

キャッシュブロックはタグ領域とデータ領域のブロックに分かれる．キャッシュデータの管理の単位であるこのデータ領域のブロックを「キャッシュライン(cache line)またはキャッシュブロック(cache block)」という．
この「ライン」がキャッシュメモリとCPUまたはメインメモリのアクセス単位になる．



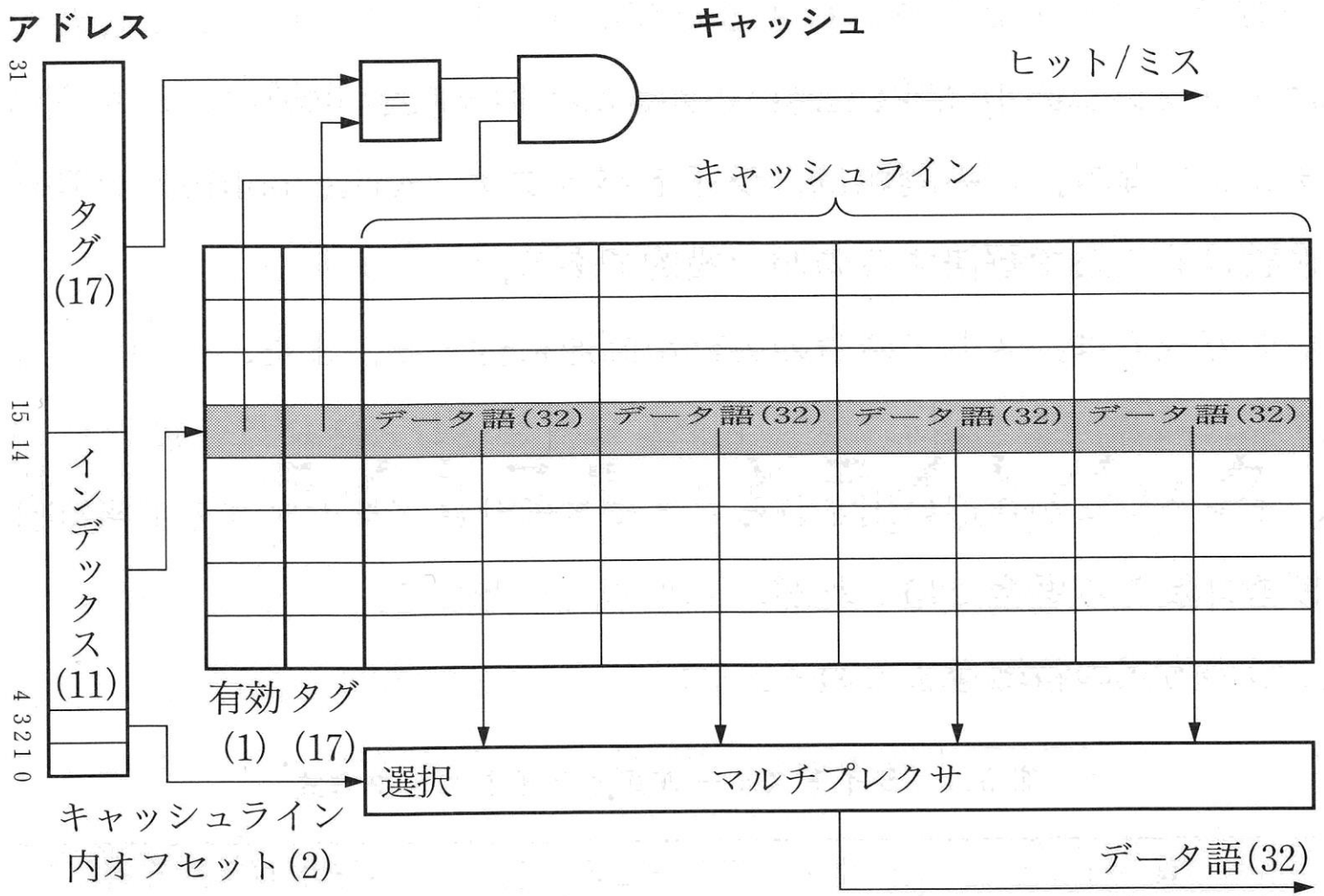
キャッシュライン番号を示すメモリアドレスの一部を「インデックス」という．

キャッシュメモリのマッピング方式

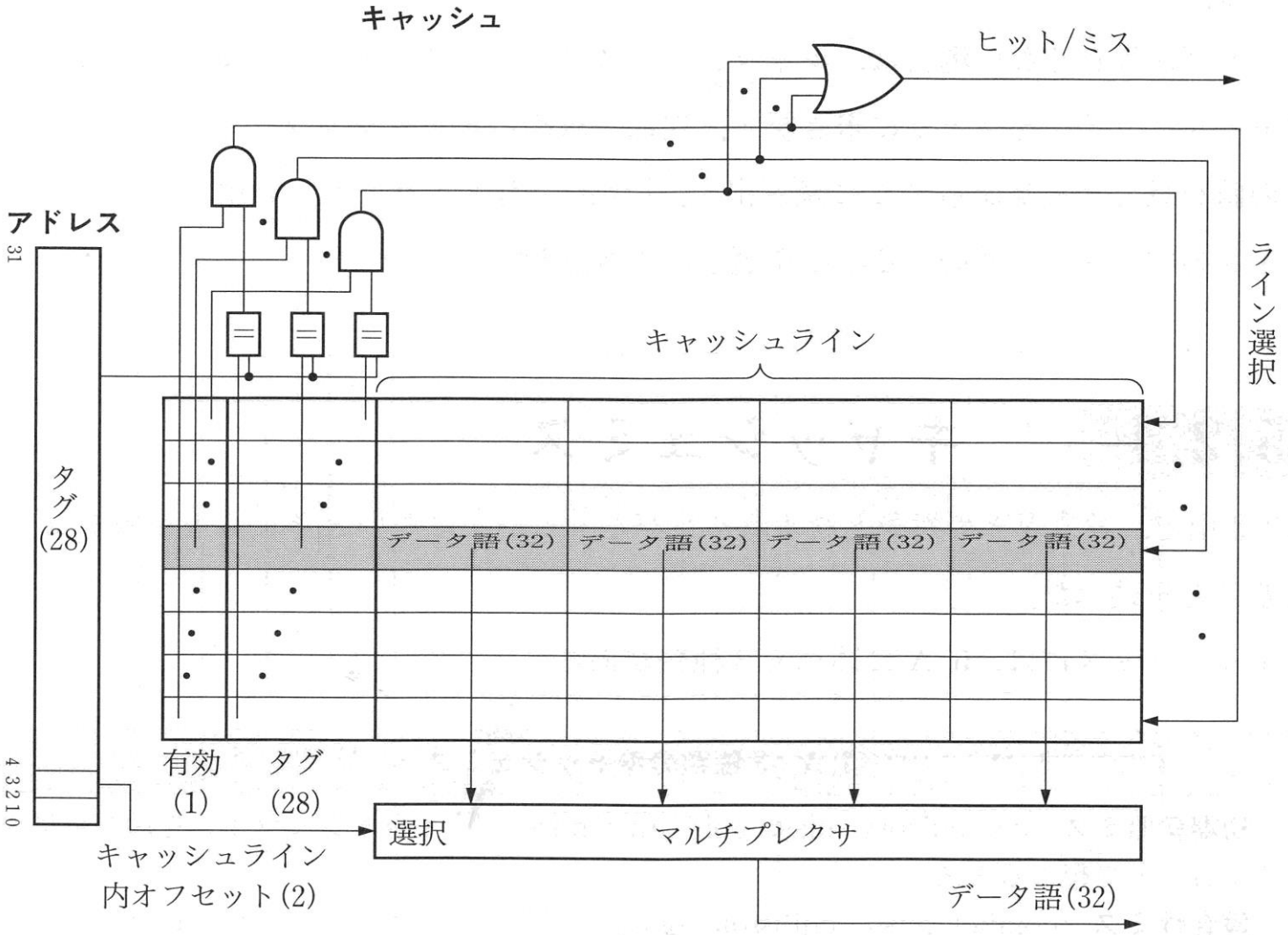
マッピングとは，メインメモリのブロックとキャッシュメモリのブロックをどのように対応させるかを決めること.
マッピング方式には3つの方式がある.

- ダイレクトマッピング方式 (direct mapping，直接マッピング)
- セットアソシアティブ方式 (set associative，部分連想マッピング)
- フルアソシアティブ方式 (full associative，完全連想マッピング)

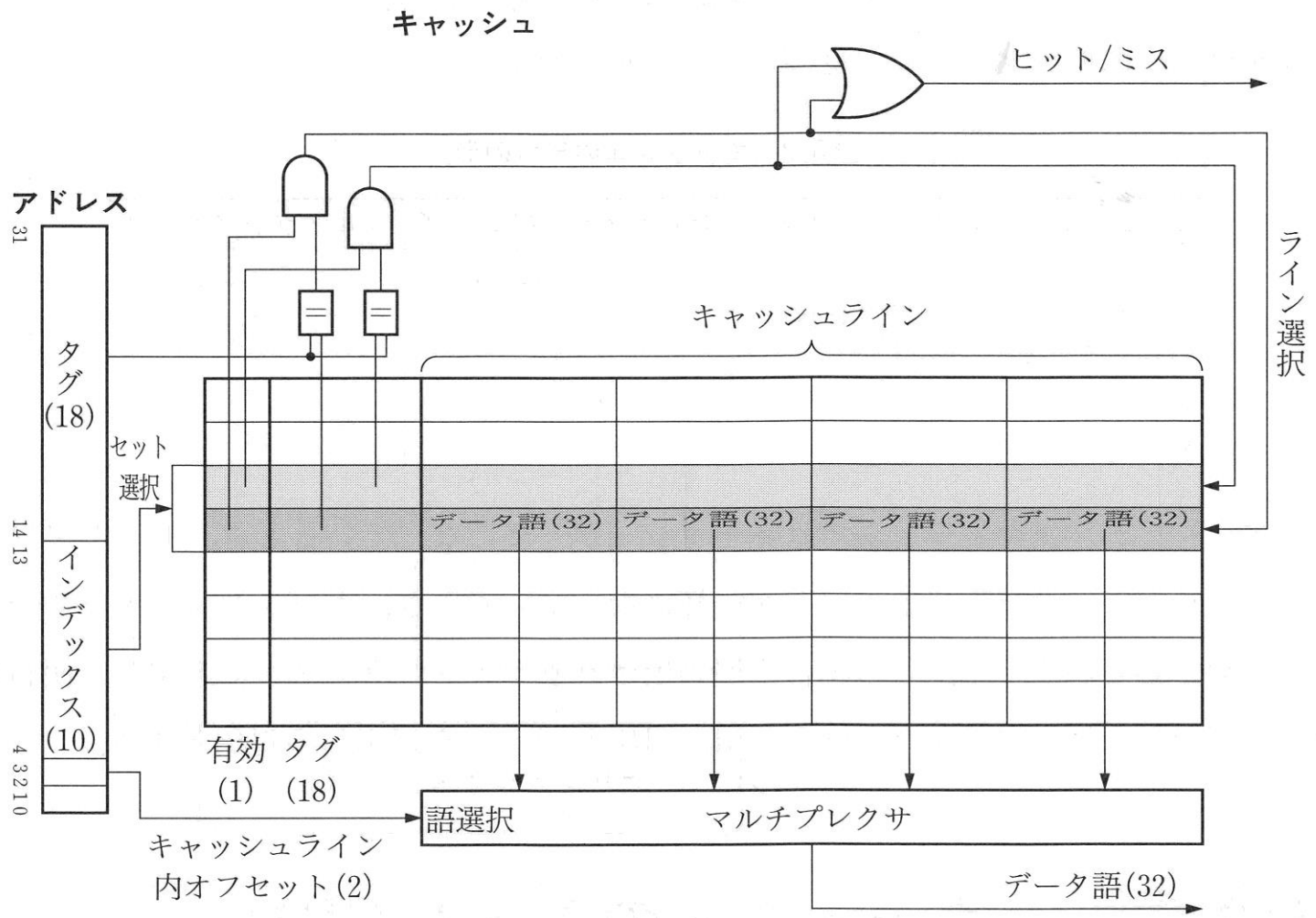
ダイレクトマッピング方式キャッシュメモリ



フルアソシアティブ方式キャッシュメモリ



セットアソシアティブ方式キャッシュメモリ



キャッシュメモリのマッピング方式

L＝ S x A　　L:総ライン数，S＝セット数，A:ウェイ数（＝連想度）

マッピング方式名	セット数	ウェイ数（＝連想度）	概要	ハードウェア規模	ヒット判定時間	ヒット率
ダイレクトマッピング方式（1ウェイセットアソシアティブ方式）			● キャッシュメモリの特定のブロックと対応したもの ● つまり，予め決められた位置にしか置けない． ● ウェイ数が1のキャッシュ ● キャッシュメモリのブロック番号＝メインメモリのブロック番号 (mod キャッシュメモリのブロック数)			
セットアソシアティブ方式			● キャッシュメモリの複数のブロックの集合と対応したもの ● つまり，決められた範囲内では任意の位置に置ける． ● ウェイ数が2以上のキャッシュ ● キャッシュメモリのブロック番号＝メインメモリのブロック番号 (mod キャッシュメモリのセット数)			
フルアソシアティブ方式			● キャッシュメモリの任意のブロックと対応したもの ● つまり，任意の位置における． ● セット数が1のキャッシュ			

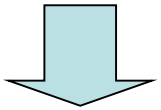
３つのＣ　キャッシュミスの原因

この3つのCを減少さえるにはどのようにすればよいのだろうか？

①　初期参照ミスを減少されるには，

②　容量ミスを減少させるには，

③　競合性ミスを減少させるには，



キャッシュラインの置換，追い出し

キャッシュリードのときに，キャッシュミスが発生した場合や，
キャッシュライトのときに，キャッシュの該当セットが一杯になっている場合は，
CPUの演算の実行を一時止めて，
メインメモリへのアクセスして，その近辺（ライン分）のデータを，キャッシュメモリ
にコピーしてから，実行を再開する．

キャッシュミス時どのラインを置換するか，キャッシュライト時どのラインを追い出すかは
次の方法で決められる．

ランダム	
FIFO (First-In-First-Out)	
LRU (Least-Recently-Used)	

分割キャッシュ

ユニファイドキャッシュ (Unified Cache)

- 命令とデータを別々に扱わない単一キャッシュ
- プリンストンアーキテクチャともいう.

セパレートキャッシュ (Separated Cahe)

- 命令用キャッシュとデータ用キャッシュを分離したキャッシュ
- ハーバードアーキテクチャともいう.

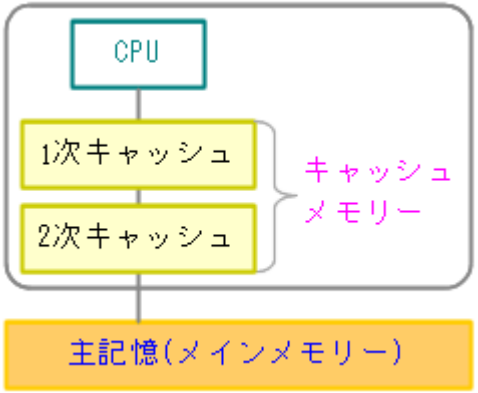
  ハーバード大で開発されたMarkIIが命令とデータを別々に記憶装置に持っていたことに由来する.

- セパレートキャッシュの利点

多階層キャッシュ

- マイクロプロセッサのクロック周波数の向上により，
  CPUとメインメモリのアクセス時間のギャップはますます大きくなっている．
- LSIの集積度の向上により，CPU内に高速小容量のキャッシュをCPU内に
  搭載できるようになっている．

- CPU内に1次キャッシュを，CPU外に2次キャッシュを備える多階層のキャッシュ
  構成法が実現されている．



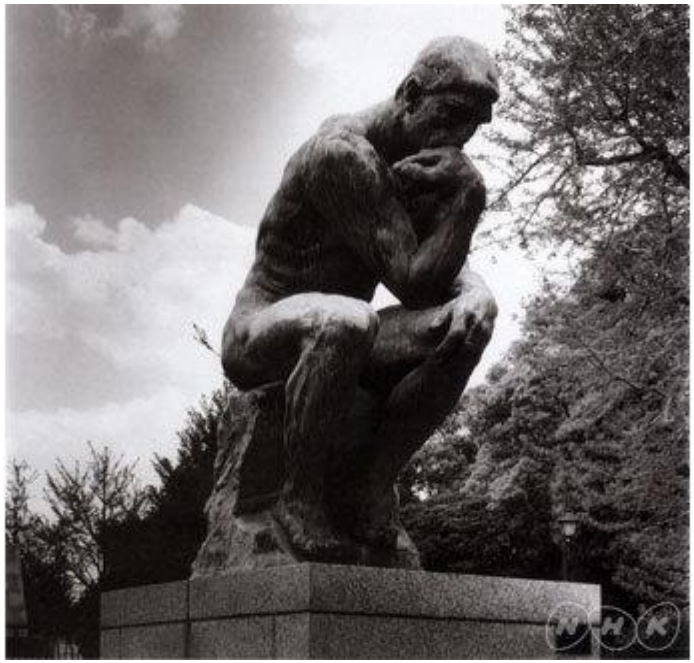
参考文献

電子情報通信レクチャーシリーズ　電子情報通信学会編
コンピュータアーキテクチャ
坂井修一
コロナ社

IT Text 情報処理学会　編集
コンピュータアーキテクチャ
内田啓一郎，小柳　滋
Ohmsha

コンピュータ設計の基礎知識
 ハードウェア・アーキテクチャ・コンパイラの設計と実装
清水尚彦
共立出版株式会社

“Computer Architecture and Implemnetation”
Harvey G. Cragon
CAMBRIDGE UNIVERSITY PRESS


NHK