Pertemuan ke 5

# Cache Memory




***Computer Organization***

Dosen : Eko Budi Setiawan

# Tujuan

- Menjelaskan tentang memori utama komputer
- Menjelaskan tipe dari memori, waktu dan pengontrolan
- Menjelaskan pembetulan kesalahan
- Menjelaskan cache memory termasuk didalamnya adalah fungsi pemetaan

# Memori

- Memori adalah bagian dari komputer tempat program-program dan data-data disimpan
- Istilah store atau storage untuk memori, meskipun kata storage sering digunakan untuk menunjuk ke penyimpanan disket
- Tempat informasi, dibaca dan ditulis
- Aneka ragam jenis, teknologi, organisasi, unjuk kerja dan harga

# Memori internal dan External

- Memori internal adalah memori yang dapat diakses langsung oleh prosesor
  - Register yang terdapat didalam prosesor, cache memori dan memori utama berada diluar prosesor
- Memori eksternal adalah memori yang diakses prosesor melalui piranti I/O
  - Disket dan hardisk

# Sifat Sel Memori

- Sel memori memiliki dua keadaan stabil (atau semi-stabil), yang dapat digunakan untuk merepresentasikan bilangan biner 1 atau 0
- Sel memori mempunyai kemampuan untuk ditulisi (sedikitnya satu kali)
- Sel memori mempunyai kemampuan untuk dibaca

# Karakteristik Sistem Memori

| Karakteristik | Macam/ Keterangan |
|---|---|
| Lokasi | 1. CPU<br>2. Internal (main)<br>3. External (secondary) |
| Kapasitas | 1. Ukuran word<br>2. Jumlah word |
| Satuan transfer | 1. Word<br>2. Block |
| Metode akses | 1. Sequential access<br>2. Direct access<br>3. Random access<br>4. Associative access |
| Kinerja | 1. Access time<br>2. Cycle time<br>3. Transfer rate |
| Tipe fisik | 1. Semikonduktor<br>2. Magnetik |
| Karakteristik fisik | 1. Volatile/nonvolatile<br>2. Erasable/nonerasable |

# Lokasi Memori

- Register
  - Berada didalam chip prosesor
  - Diakses langsung oleh prosesor dalam menjalankan operasinya
  - Register digunakan sebagai memori sementara dalam perhitungan maupun pengolahan data dalam prosesor
- Memori Internal
  - Berada diluar chip prosesor
  - Mengaksesannya langsung oleh prosesor
  - Dibedakan menjadi memori utama dan cache memori
- Memori Eksternal
  - Diakses oleh prosesor melalui piranti I/O
  - Dapat berupa disk maupun pita

# Kapasitas Memori

- Kapasitas memori internal maupun eksternal biasanya dinyatakan dalam bentuk byte, (1 byte = 8 bit)

  Atau dinyatakan dalam bentuk Word
- Panjang word umumnya 8, 16, 32 bit
- Memori eksternal biasanya lebih besar kapasitasnya daripada memori internal, hal ini disebabkan karena teknologi dan sifat penggunaannya yang berbeda
- Memori internal
  - Satuan transfer sama dengan jumlah saluran data yang masuk ke dan keluar dari modul memori
  - Jumlah saluran ini sering kali sama dengan panjang word, tapi dimungkinkan juga tidak sama

# Konsep Satuan Transfer

- Word, merupakan satuan “alami” organisasi memori. Ukuran word biasanya sama dengan jumlah bit yang digunakan untuk representasi bilangan dan panjang instruksi
- Addressable units, pada sejumlah sistem, adressable units adalah word. Namun terdapat sistem dengan  Atau dinyatakan dalam bentuk Word. Namun terdapat sistem dengan pengalamatan pada tingkatan byte. Pada semua kasus hubungan antara panjang A suatu alamat dan jumlah N adressable unit adalah 2A = N
- Unit of Transfer, adalah jumlah bit yang dibaca atau dituliskan ke dalam memori pada suatu saat. Pada eksternal, transfer data biasanya lebih besar dari suatu word, yang disebut dengan block

# Metode Akses

- Sequential Access
  - Memori diorganisasi unit-unit data yang disebut record
  - Akses harus dibuat dalam bentuk urutan linier yang spesifik
  - Informasi pengalamatan yang disimpan dipakai untuk memisahkan record-record dan untuk membantu proses pencarian
  - Terdapat shared read/write mechanism untuk penulisan/pembacaan memorinya
  - Pita magnetik merupakan memori yang menggunakan metode sequential access
- Direct Access
  - Pita sequential access terdapat shared read/write mechanism
  - Setiap blok dan record memiliki alamat unik berdasarkan lokasi fisiknya
  - Akses dilakukan langsung pada alamat memori
  - Disk adalah memori direct acess

# Metode Akses

- Associative Access
  - Jenis random akses yang memungkinkan pembandingan lokasi bit yang diinginkan untuk pencocokan
  - Data dicari berdasarkan isinya bukan alatamnya dalam memori
  - Contoh memori ini adalah cache memori
- Random Access
  - Setiap lokasi memori dipilih secara random dan diakses serta dialamati secara langsung
  - Contohnya adalah memori utama

# Parameter Utama

- Access Time
  - Bagi random access memory, waktu akses adalah waktu yang dibutuhkan untuk melakukan operasi baca atau tulis
  - Memori non random akses merupakan waktu yang dibutuhkan dalam melakukan mekanisme baca atau tulis pada lokasi tertentu
- Memory Cycle Time
  - Konsep ini digunakan pada random access memory
  - Terdiri dari access time ditambah dengan waktu yang diperlukan agar hilang pada saluran sinyal

# Parameter Utama

- Transfer Rate

  Kecepatan data transfer ke unit memori atau dari unit memori

  - Random access memory sama dengan 1/cycle time
  - Non-random access memory dengan perumusan :

$$TN = TA + (N/R)$$

TN = Waktu rata-rata untuk membaca atau menulis N bit

TA = Waktu akses rata-rata

N = jumlah bit

R = kecepatan transfer dalam bit

# Karakteristik Fisik

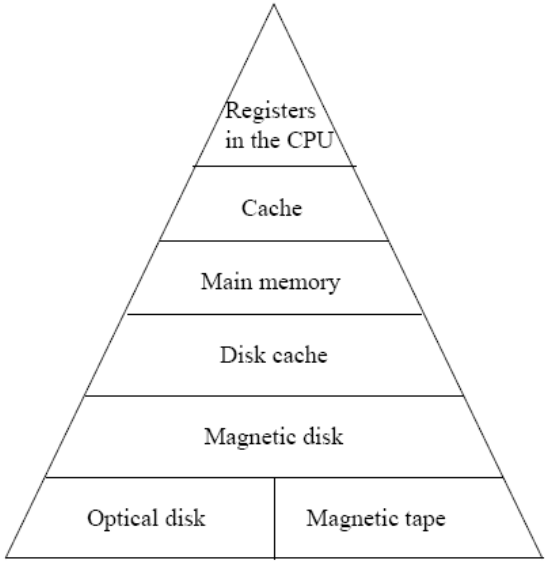
- Media Penyimpanan volatile dan non-volatile
  - Volatile memory, informasi akan hilang apabila daya listriknya dimatikan
  - Non-Volatile memory tidak akan hilang walau daya listriknya hilang
  - Memori permukaan magnetik adalah contoh non-volatile memory, sedangkan semikonduktor ada yang volatile dan non-volatile
- Media erasable dan nonerasable
  - Ada jenis memori semikonduktor yang tidak bisa dihapus kecuali dengan menghancurkan unit storage-nya, memori ini dikenal dengan ROM (Read Only Memory)

# Keandalan Memori

- Berapa Banyak ?
  - Sesuatu yang sulit dijawab, karena berapapun kapasistas memori tentu aplikasi akan menggunakannya.
- Berapa Cepat
  - Memori harus mampu mengikuti kecepatan CPU sehingga terjadi kerja antar CPU dan memori tanpa adanya waktu tunggu karena kompnen lain belum selesai prosesnya
- Berapa Mahal ?
  - Relatif. Bagi produsen selalu mencari harga produksi paling murah tanpa mengorbankan kualitasnya untuk memiliki daya saing di pasaran

# Hirarki Memori

- Registers
- L1 Cache
- L2 Cache
- Main memory
- Disk cache
- Disk
- Optical
- Tape




| Tipe memori | Teknologi | Ukuran | Waktu akses |
|---|---|---|---|
| Cache Memory | semikonduktor RAM | 128 – 512 KB | 10 ns |
| Memori Utama | semikonduktor RAM | 4 – 128 MB | 50 ns |
| Disk magnetik | Hard Disk | Gigabyte | 10 ms, 10MB/det |
| Disk Optik | CD-ROM | Gigabyte | 300ms, 600KB/det |
| Pita magnetik | Tape | 100 MB | Det -mnt, 10MB/mnt |

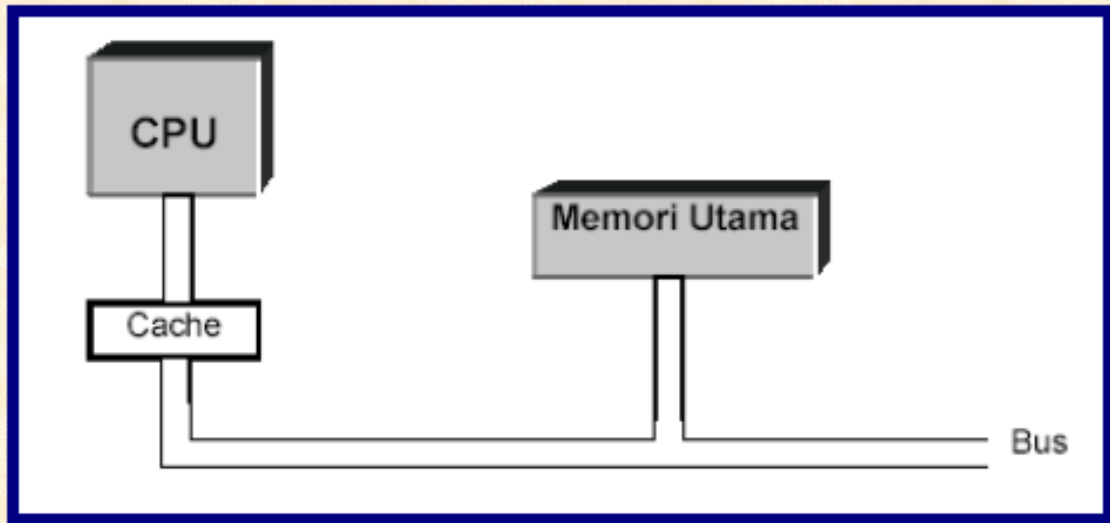
# Satuan Memori

- Satuan pokok memori adalah digit biner, yang disebut bit (binary digit)
- Bit dapat berisi sebuah angka 0 atau 1
- Memori juga dinyatakan dalam byte
  - 1 byte = 8 bit
  - Kumpulan byte dinyatakan dalam word
  - Panjang word yang umum adalah 8,16, 32 bit

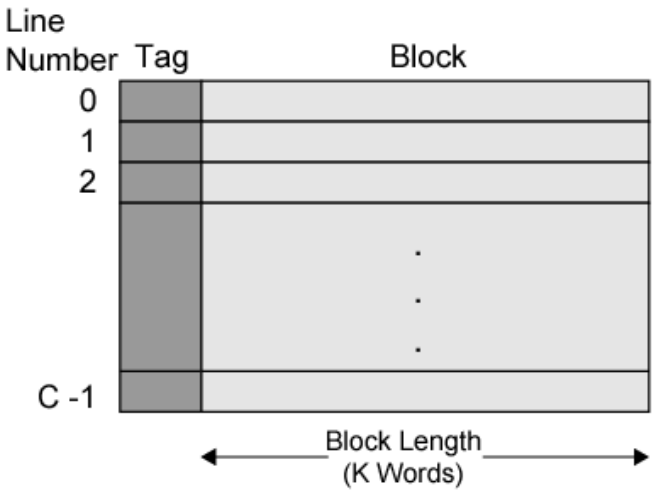
| Symbol | | Number of Bytes | |
|---|---|---|---|
| Kilo Byte | KB | 2e 10 | 1024 |
| Mega Byte | MB | 2e 20 | 1.048.576 |
| Giga Byte | GB | 2e 30 | 1.073.741.824 |
| Tera Byte | TB | 2e 40 | 1.099.511.627.776 |

# Cache Memori

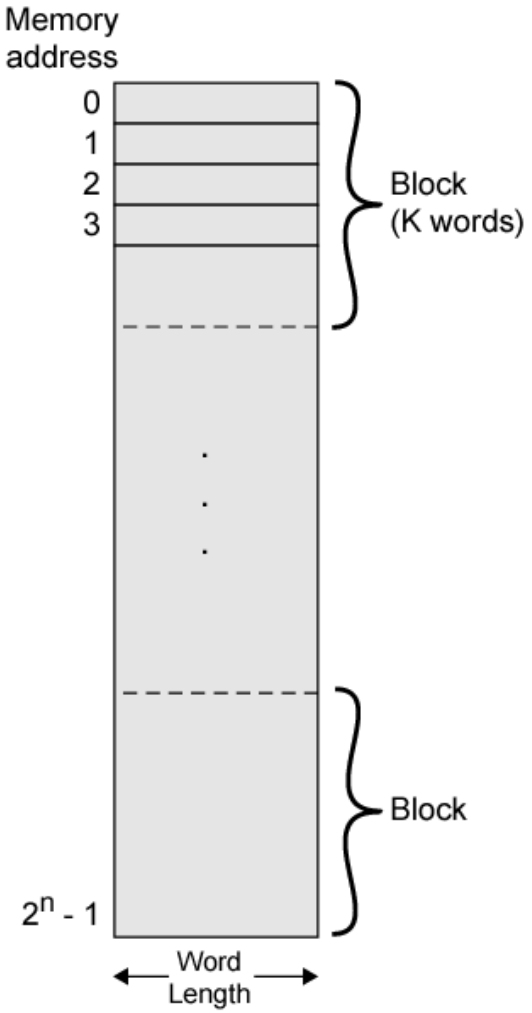
- Memori yang sangat cepat dengan ukuran yang kecil
- Terdapat diantara main memory dengan CPU
- Lokasinya terdapat di CPU
- Harga cache sangat mahal



# Cache / Main Memory Structure

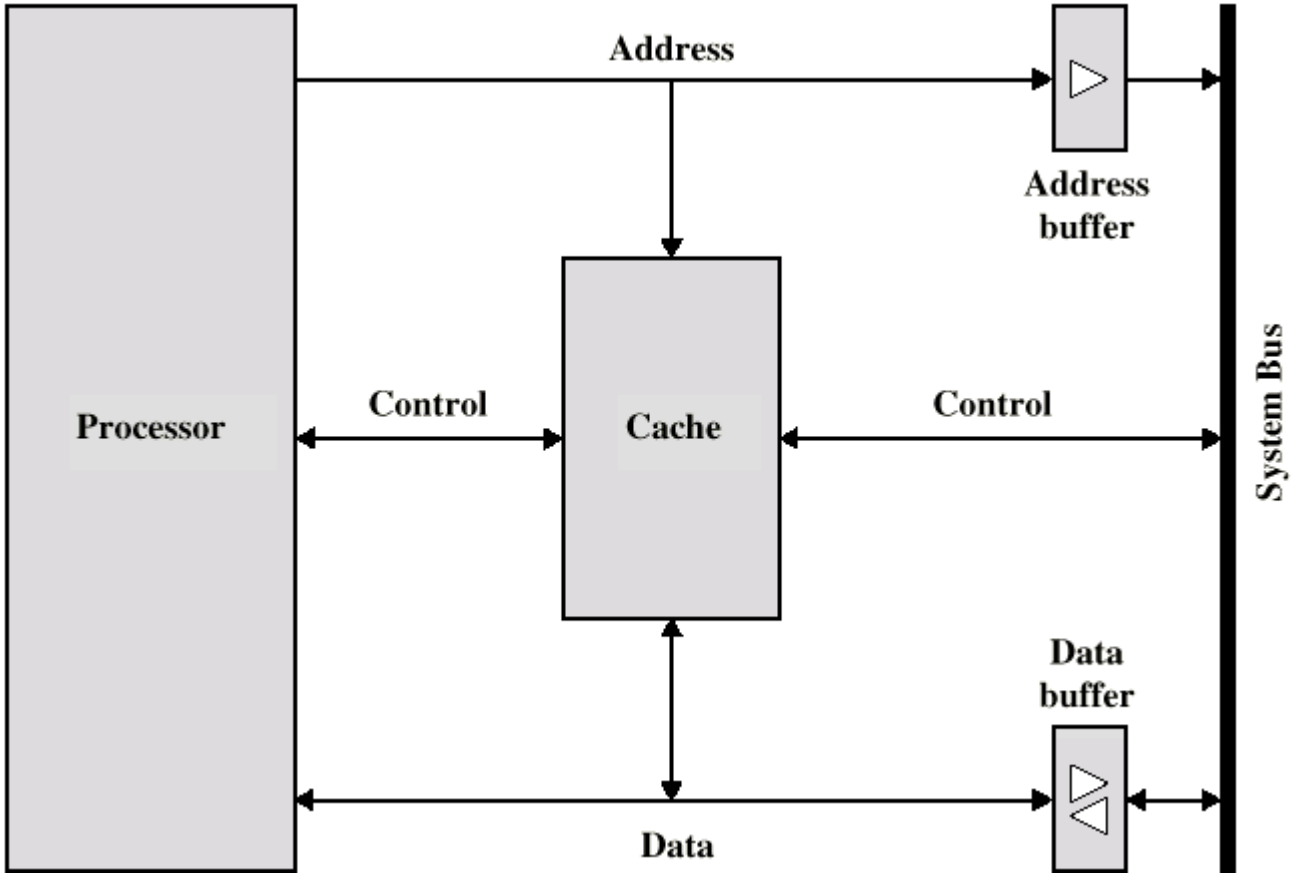



(a) Cache




(b) Main memory

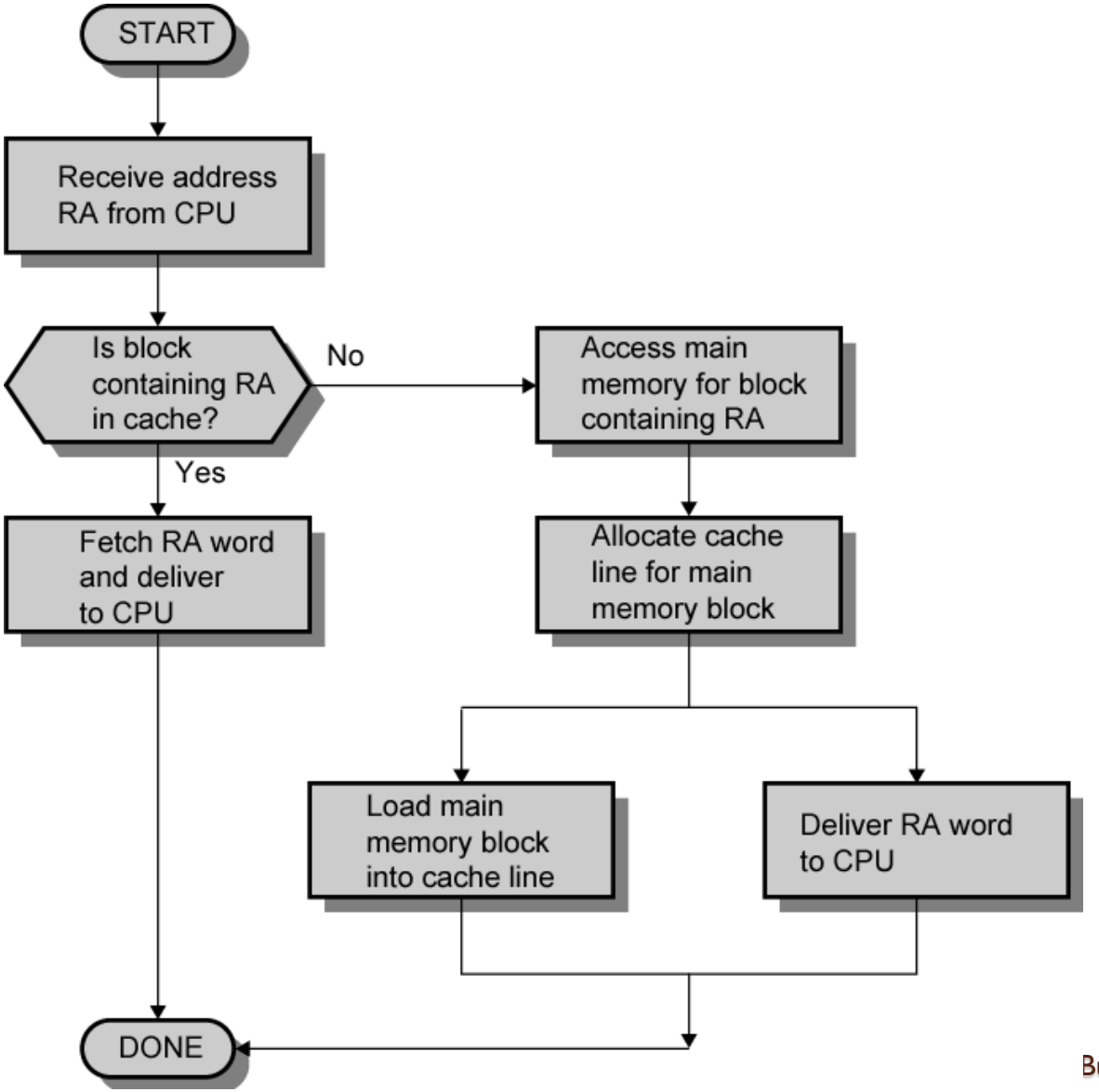
# Organisasi Cache Memori



# Comparison Cache Size

| Processor | Type | Year of Introduction | L1 cachea | L2 cache | L3 cache |
|---|---|---|---|---|---|
| IBM 360/85 | Mainframe | 1968 | 16 to 32 KB | — | — |
| PDP-11/70 | Minicomputer | 1975 | 1 KB | — | — |
| VAX 11/780 | Minicomputer | 1978 | 16 KB | — | — |
| IBM 3033 | Mainframe | 1978 | 64 KB | — | — |
| IBM 3090 | Mainframe | 1985 | 128 to 256 KB | — | — |
| Intel 80486 | PC | 1989 | 8 KB | — | — |
| Pentium | PC | 1993 | 8 KB/8 KB | 256 to 512 KB | — |
| PowerPC 601 | PC | 1993 | 32 KB | — | — |
| PowerPC 620 | PC | 1996 | 32 KB/32 KB | — | — |
| PowerPC G4 | PC/server | 1999 | 32 KB/32 KB | 256 KB to 1 MB | 2 MB |
| IBM S/390 G4 | Mainframe | 1997 | 32 KB | 256 KB | 2 MB |
| IBM S/390 G6 | Mainframe | 1999 | 256 KB | 8 MB | — |
| Pentium 4 | PC/server | 2000 | 8 KB/8 KB | 256 KB | — |
| IBM SP | High-end server/ supercomputer | 2000 | 64 KB/32 KB | 8 MB | — |
| CRAY MTAb | Supercomputer | 2000 | 8 KB | 2 MB | — |
| Itanium | PC/server | 2001 | 16 KB/16 KB | 96 KB | 4 MB |
| SGI Origin 2001 | High-end server | 2001 | 32 KB/32 KB | 4 MB | — |
| Itanium 2 | PC/server | 2002 | 32 KB | 256 KB | 6 MB |
| IBM POWER5 | High-end server | 2003 | 64 KB | 1.9 MB | 36 MB |
| CRAY XD-1 | Supercomputer | 2004 | 64 KB/64 KB | 1MB | — |

# Cache Read Operation - Flowchart



# Cache Design

- Kapasitas
- Ukuran Blok
- Mapping Function
  - Direct Mapping
  - Assosiative Mapping
  - Set Assosiative Mapping
- Replacement Algorithm
  - Least Recently Used (LRU)
  - FIFO
  - Least Frequently Used (LFU)
  - Random
- Write Policy [write through, write back, write once]
- Jumlah Cache [ single/dua level, Unified atau split]

# Kapasitas & Ukuran Blok Cache

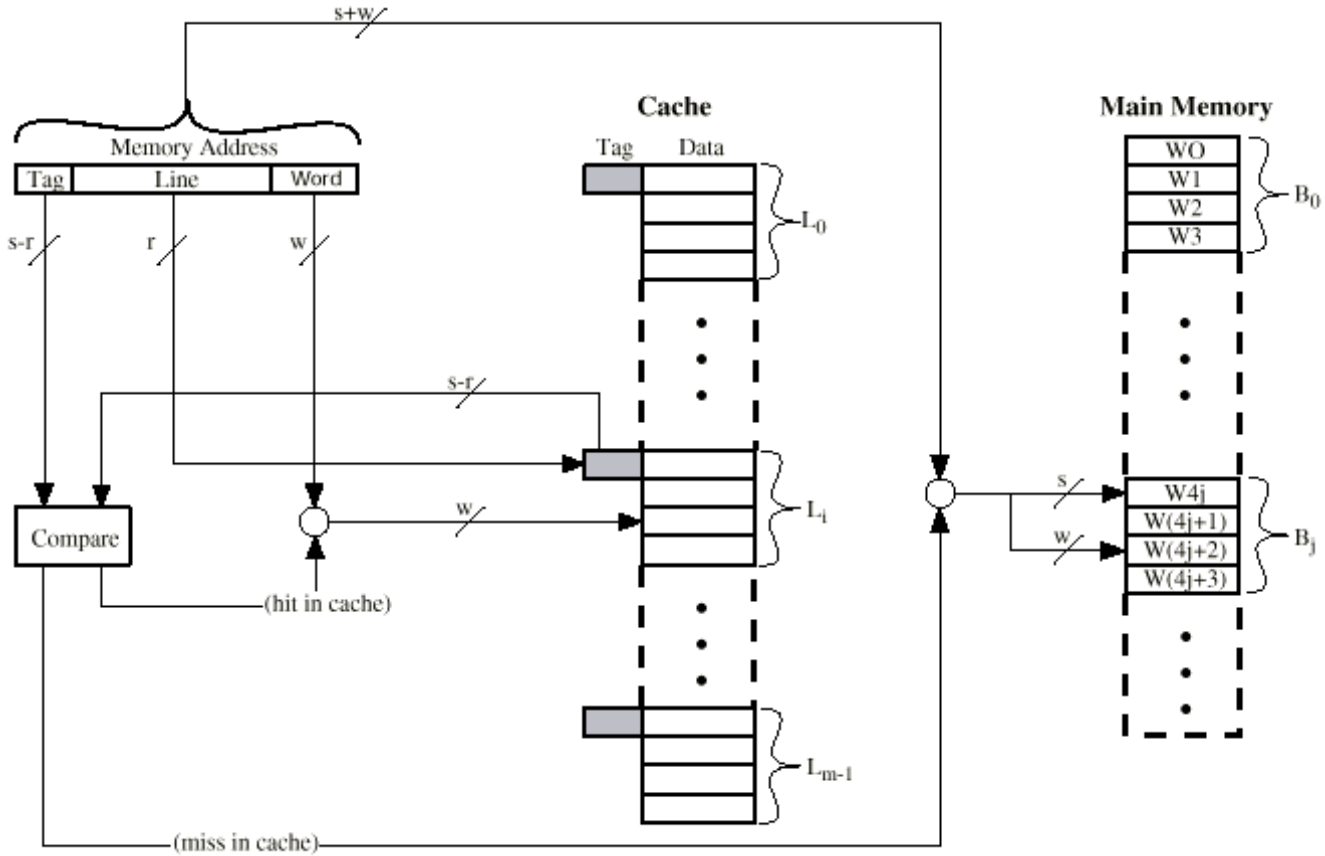
- Kapasitas Cache
  - AMD mengeluarkan prosesor K5 dan K6 dengan cache yang besar (1MB), kinerjanya tidak bagus
  - Intel mengeluarkan prosesor tanpa cache untuk alasan harga yang murah, yaitu seri Intel Celeron pada tahun 1998, kinerjanya sangat buruk terutama untuk operasi data besar, floating point, 3D processing
  - Sejumlah penelitian telah menganjurkan bahwa ukuran cache antara 1KB dan 512 KB akan lebih optimum
- Ukuran Blok Cache
  - Hubungan antara ukuran blok dan hit ratio sangat rumit untuk dirumuskan, tergantung pada karakteristik lokalitas programnya dan tidak terdapat nilai optimum yang pasti telah ditemukan
  - Ukuran antara 4 hingga 8 satuan yang dapat dialamati (word atau byte) cukup beralasan untuk mendekati nilai optimum

# Mapping (Pemetaan)

- Cache mempunyai kapasitas yang kecil dibandingkan memori utama
- Aturan blok-blok mana yang diletakkan dalam cache
- Terdapat tiga metode, yaitu pemetaan langsung, pemetaan asosiatif, dan pemetaan asosiatif set

# Direct Mapping

- Teknik paling sederhana,yaitu teknik ini memetakan blok memori utama hanya kesebuah saluran cache saja



# Direct Mapping

- i = j modulus m dan m = 2r

  dimana :

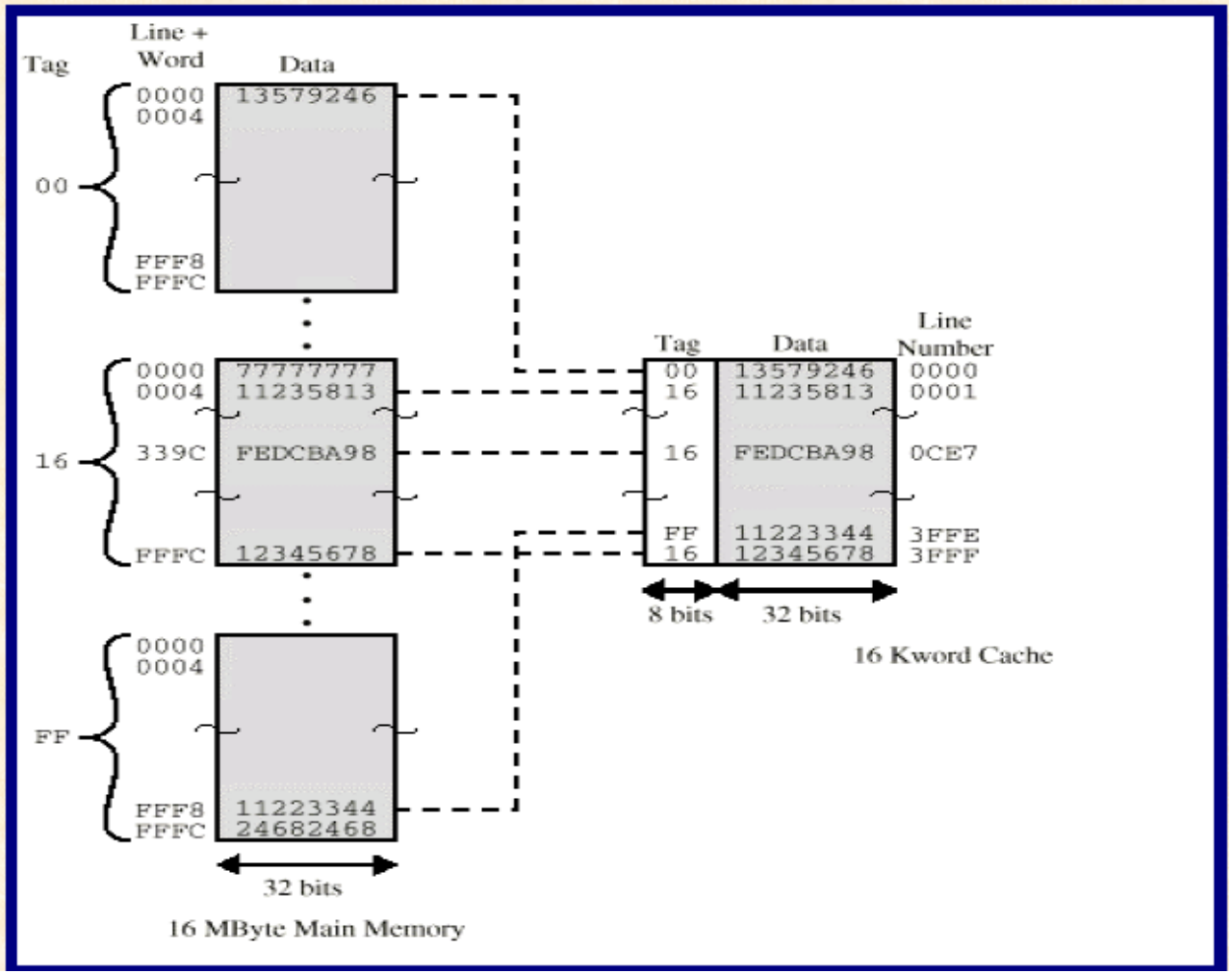
  i = nomor saluran cache

  j = nomor blok memori utama

  m = jumlah saluran yang terdapat dalam cache

| Saluran Cache | Blok-blok memori utama |
|---|---|
| 0 | 0,m, . . . . ., 2S - m |
| 1 | 1, (m=1), . . . . ., 2S – (m +1) |
| | |
| M-1 | (m-1), (2m-1), . . . ., 2S - 1 |

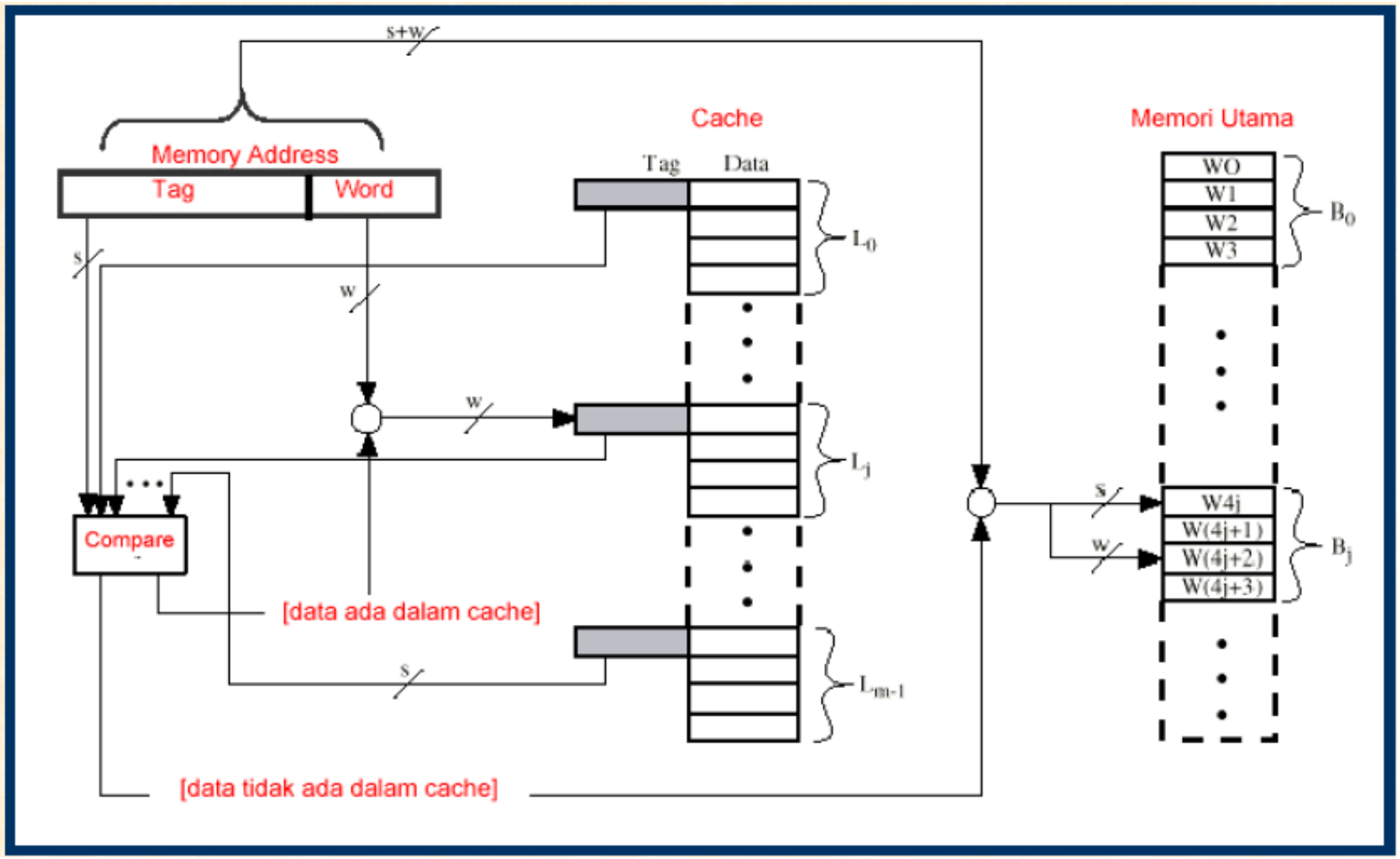
# Direct Mapping



# Pemetaan Assosiatif

- Mengatasi kekurangan pemetaan langsung
- Tiap blok memori utama dapat dimuat ke sembarang saluran cache
- Alamat memori utama diinterpretasikan dalam field tag dan field word oleh kontrol logika cache
- Tag secara unik mengidentifikasi sebuah blok memori utama
- Mekanisme utuk mengetahui suatu blok dalam cache dengan memeriksa setiap tag saluran cache oleh kontrol logika cache
- Fleksibilitas dalam penggantian blok baru yang ditempatkan dalam cache
- Kelebihan : Algoritma penggantian dirancang untuk memaksimalkan hit ratio, yang pada pemetaan langsung terdapat kelemahan
- Kekurangan : Kompleksitas rangkaian sehingga mahal secara ekonomi

# Pemetaan Assosiatif



# Pemetaan Assosiatif Set

- Menggabungkan kelebihan yang ada pada pemetaan langsung dan pemetaan asosiatif
- Memori cache dibagi dalam bentuk set-set
- Alamat memori utama diinterpretasikan dalam tiga field, yaitu : Field tag, field set, field word
- Setiap blok memori utama dapat dimuat dalam sembarang saluran cache
- Cache dibagi dalam v buah set, yang masing-masing terdiri dari k saluran

$$m = v \times k$$

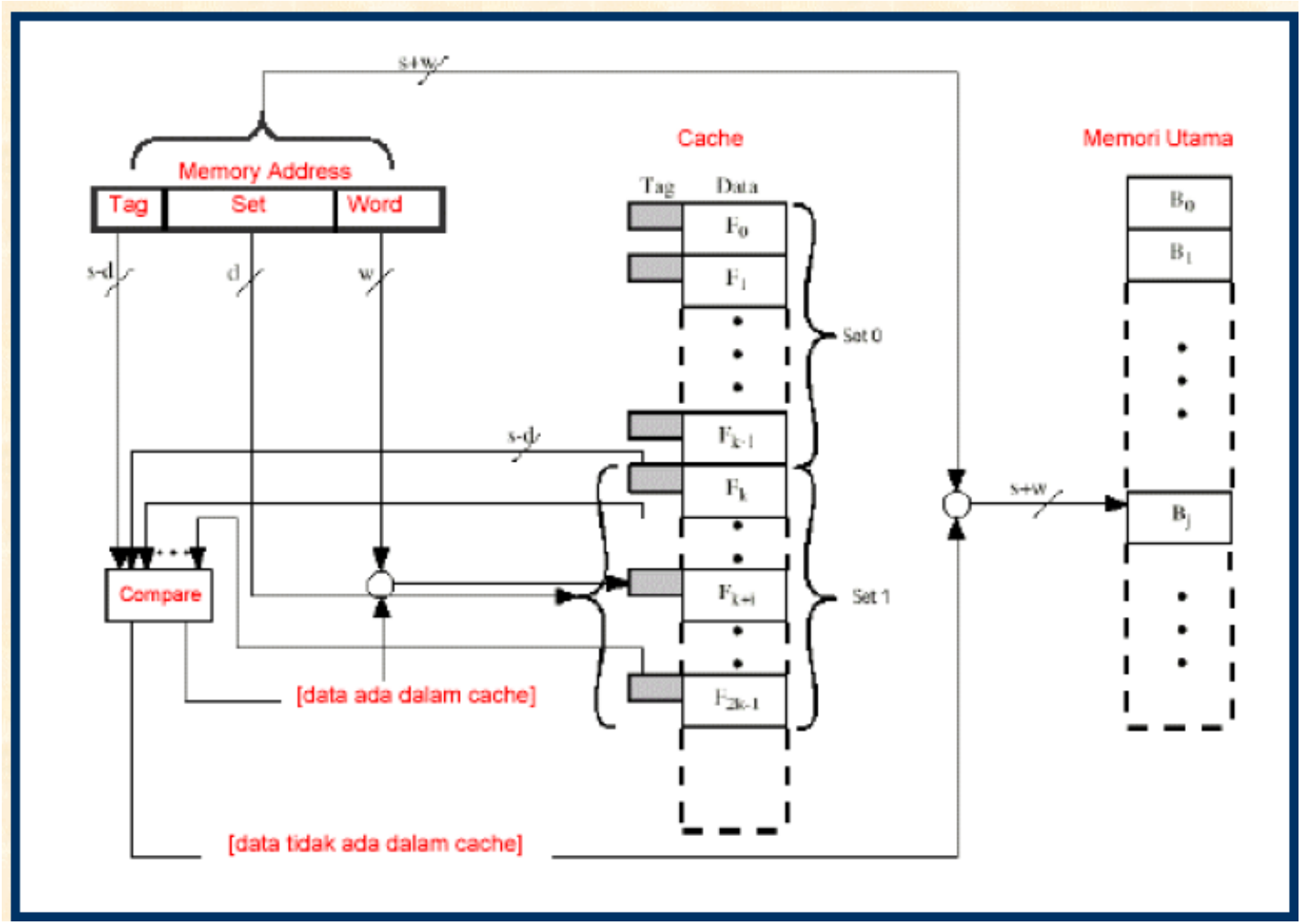
i = j modulus v dan v =2d dimana :

i = nomor set cache

j = nomor blok memori utama

m = jumlah saluran pada cache

# Pemetaan Assosiatif Set



# Algoritma Penggantian

- Algoritma Least Recently Used (RLU), yaitu mengganti blok data yang terlama berada dalam cache dan tidak memiliki referensi. (EFEKTIF)
- Algoritma First in First Out (FIFO), yaitu mengganti blok data yang awal masuk
- Algoritma Least Frequently Used (LFU) adalah mengganti blok data yang mempunyai referensi paling sedikit
- Algoritma Random, yaitu penggantian tidak berdasarkan pemakaian datanya, melainkan berdasarkan sot dari beberapa slot kandidat secara acak

# Write Policy – Mengapa ?

- Apabila suatu data telah diletakkan pada cache maka sebelum ada penggantian harus dicek apakah data tersebut telah mengalami perubahan
- Apabila telah berubah maka data pada memori utama harus diupdate
- Masalah penulisan ini sangat kompleks, apalagi memori utama dapat diakses langsung oleh modul I/O, yang memungkinkan data pada memori utama berubah, lalu bagaimana dengan data yang telah dikirim pada cache ?
- Tentunya perbedaan ini menjadikan data tidak valid

# Write Policy – "write through & back"

- Write Through
  - Operasi penulisan melibatkan data pada memori utama dan sekaligus pada cache memori sehingga data selalu valid
  - Kekurangan teknik ini adalah
    - Lalu lintas data ke memori utama dan cache sangat tinggi
    - Mengurangi kinerja sistem, bisa terjadi hang
- Write Back
  - Teknik meminimasi penulisan dengan cara penulisan pada cache saja
  - Pada saat akan terjadi penggantian blok data cache maka baru diadakan penulisan pada memori utama
  - Masalah : manakala data di memori utama belum di-update telah diakses modul I/O sehingga data dimemori utama tidak valid

# Cache

- Cache Data
- Cache instruksi yang disebut unified cache
  - Keuntungan unified cache :
    - Hit rate yang tinggi karena telah dibedakan antara informasi data dan informasi instruksi
    - Hanya sebuah cache saja yang perlu dirancang dan diimpelemtasikan
- Split cache
  - Mesin-mesin superscalar seperti Pentium dan Power PC
  - Menekankan pada paralel proses dan perkiraan-perkiraan eksekusi yang akan terjadi
- Kelebihan utama split cache
  - Mengurangi persaingan antara prosesor instruksi dan unit eksekusi untuk mendapatkan cache, hal ini sangat utama bagi perancangan prosesor-prosesor pilening