

Caches

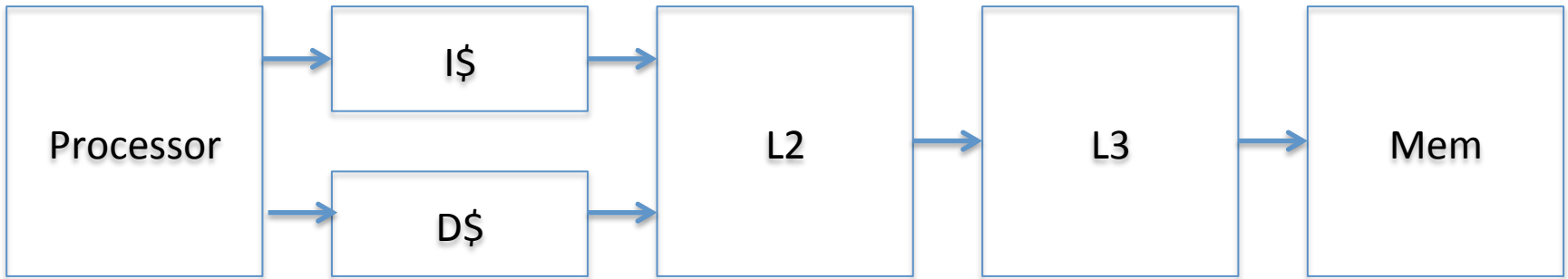
MO801

Rodolfo Azevedo

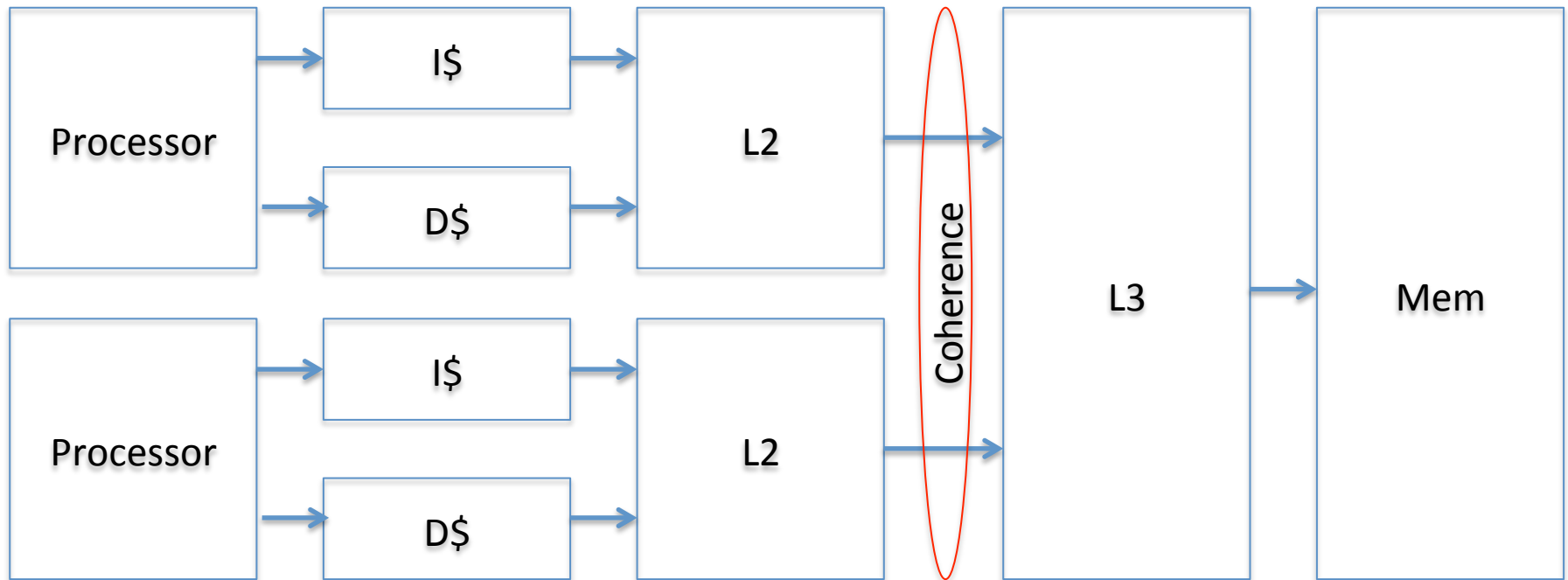
Why?

- Caches provide transparent speedup to memory access
- Multiple cache levels: L1, L2, L3, ...
- Split L1 → L1I and L1D
 - Also called I\$ and D\$
- Unified L2, L3, ...
- In an ideal system, cache size = 0
 - Memory is fast enough to provide all required data

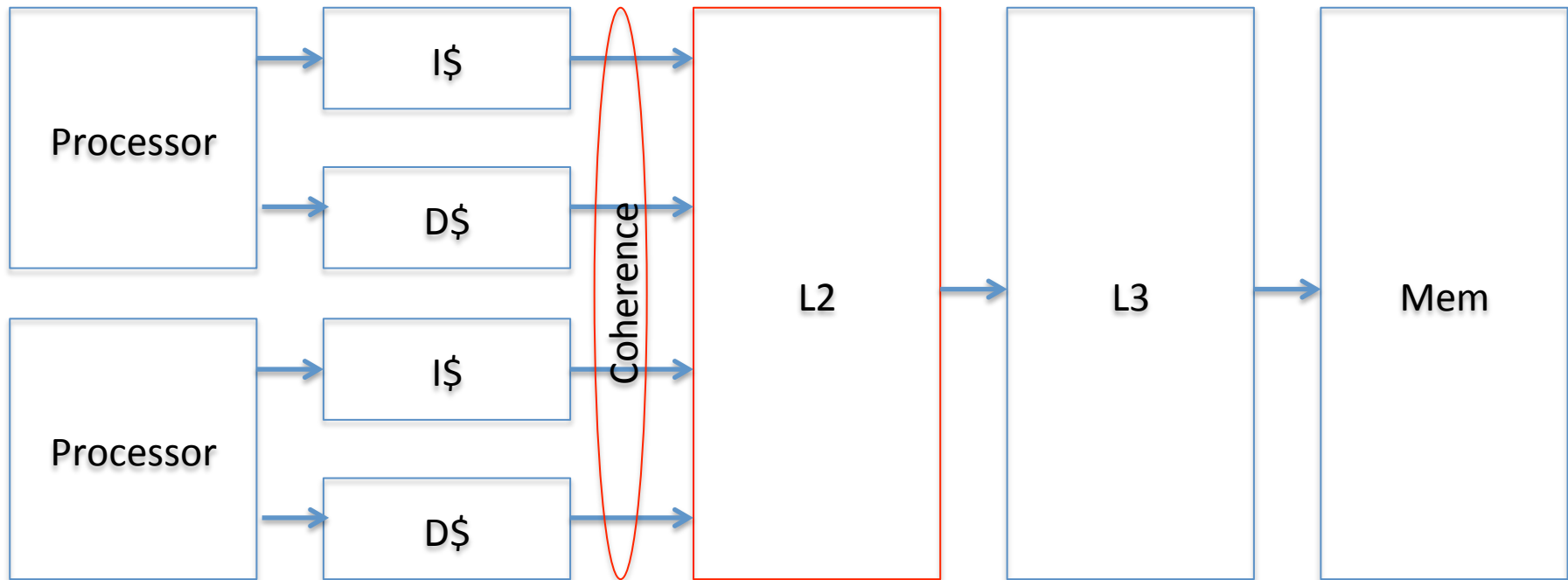
The Basics



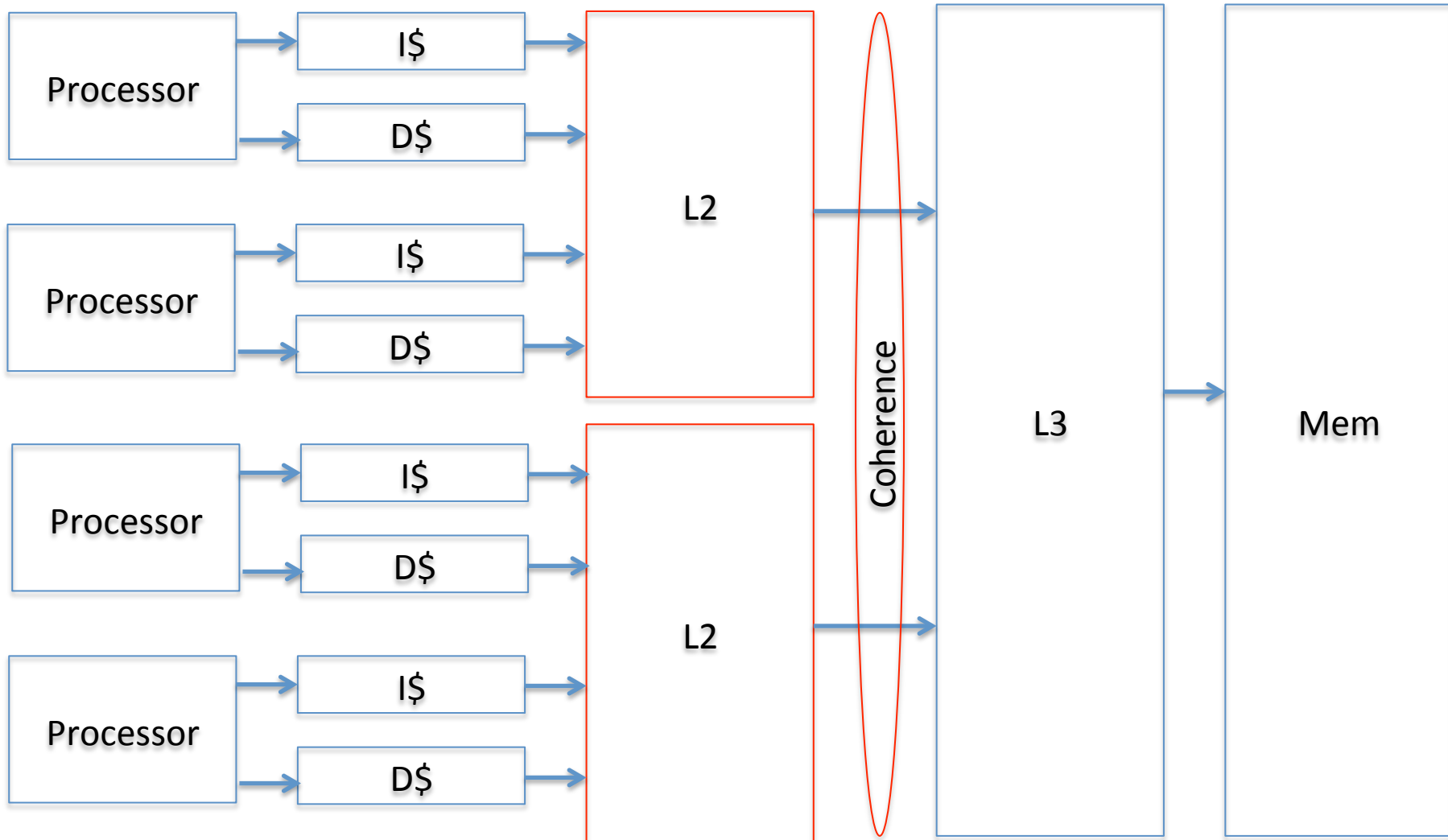
Multi-core



Multi-core



Multi-core

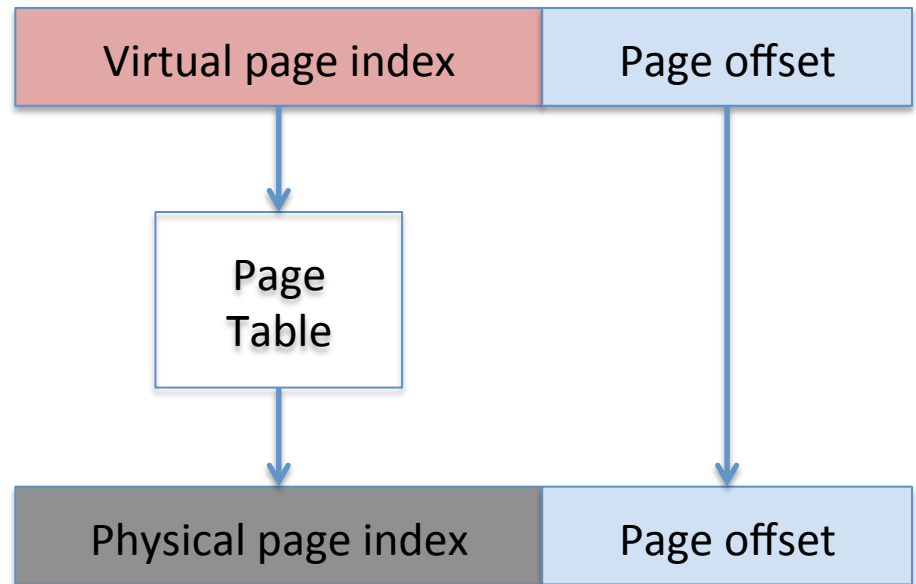


Address Translation

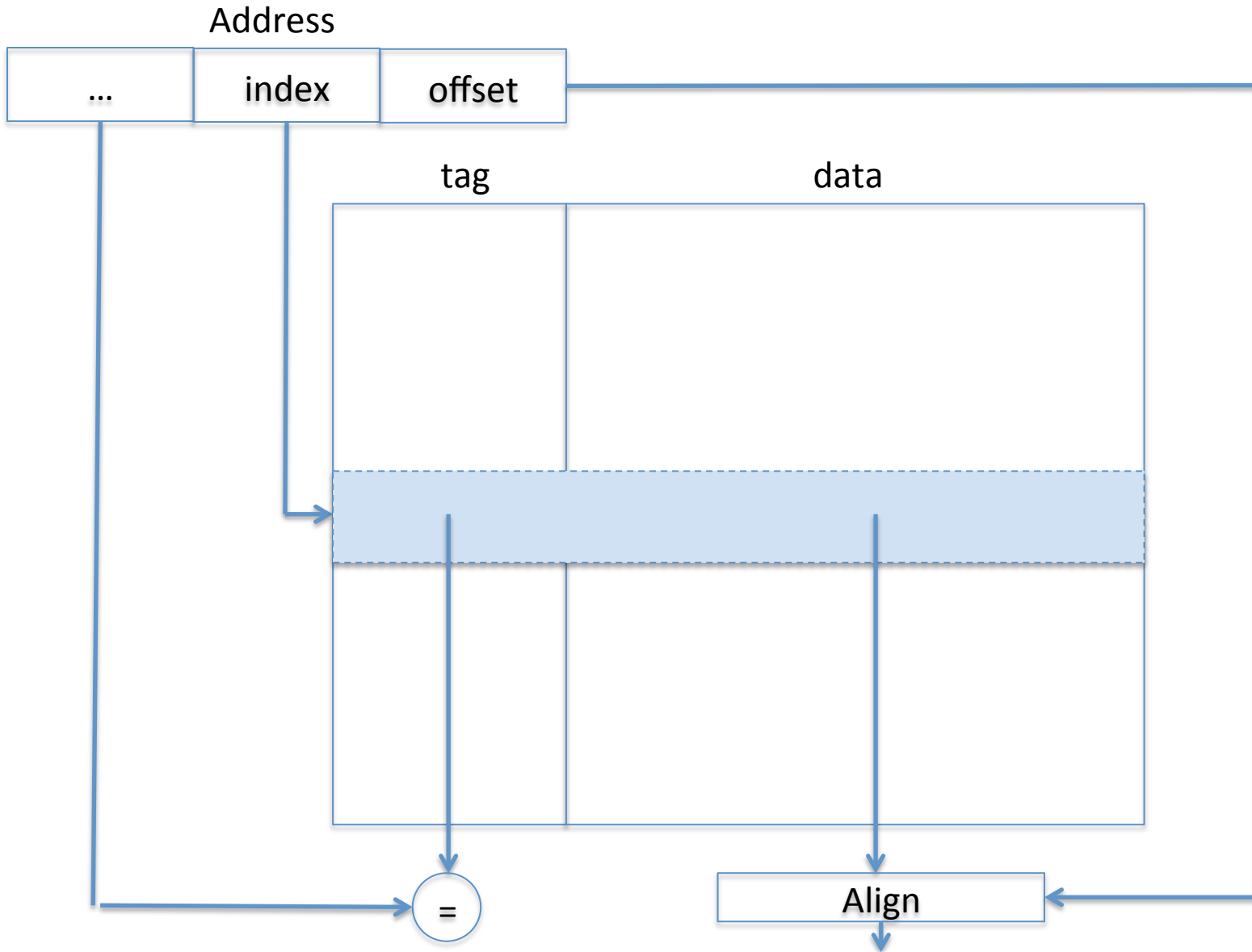
Concepts

- Physical address
- Virtual Address
- Page
- TLB
- Virtual aliasing
- Translation order
 - Before cache
 - After cache

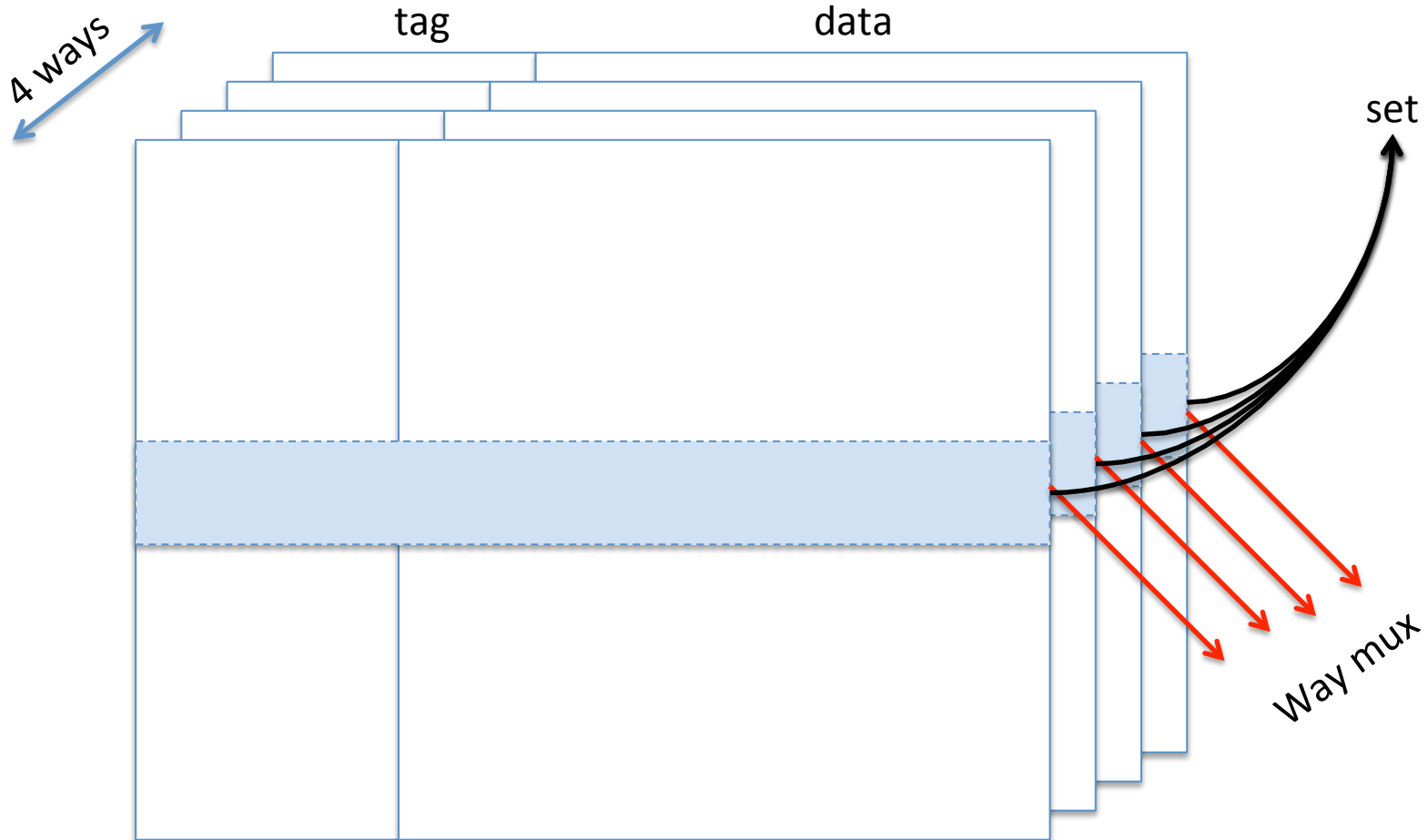
Mapping



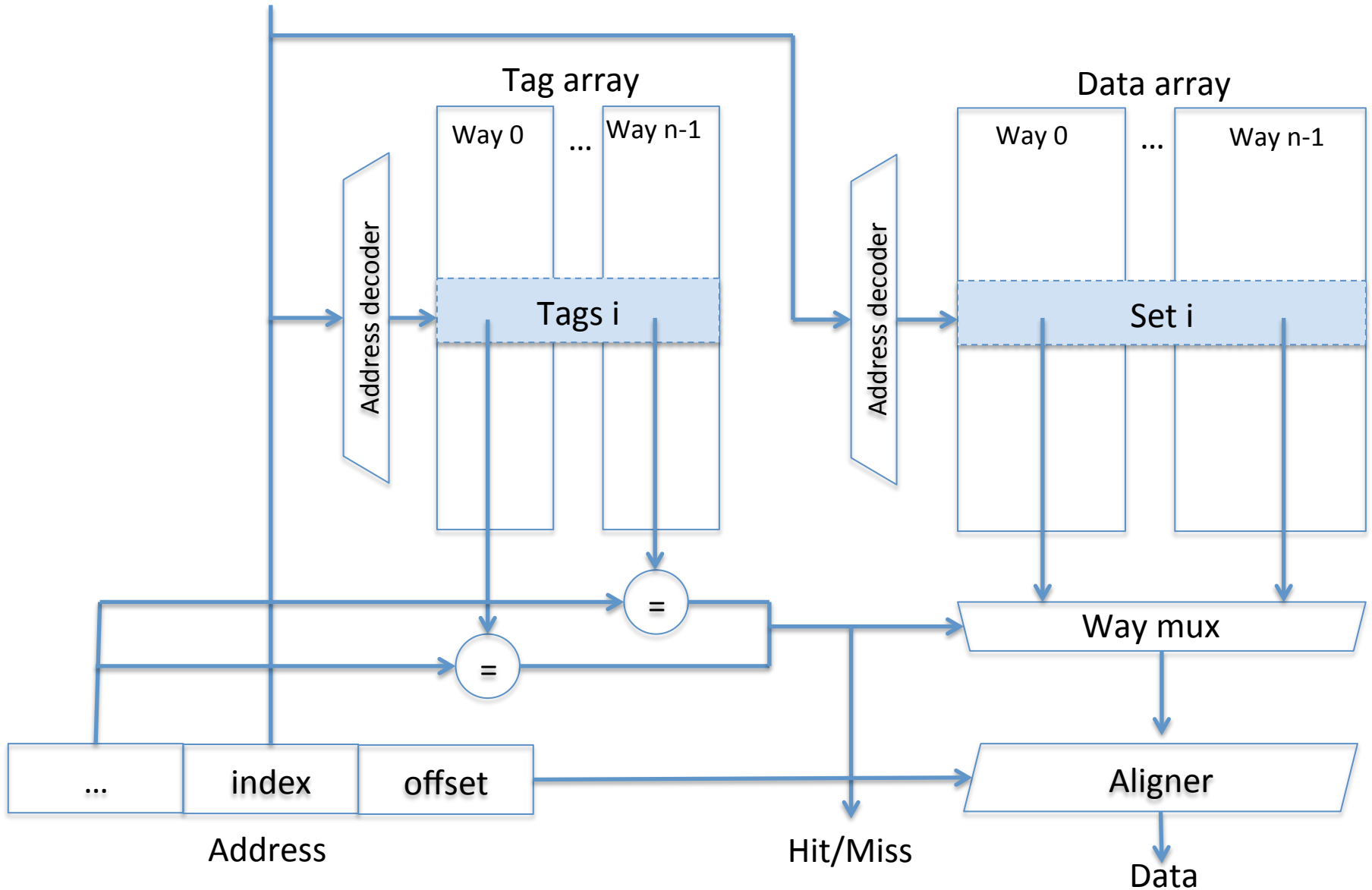
Internal Structure



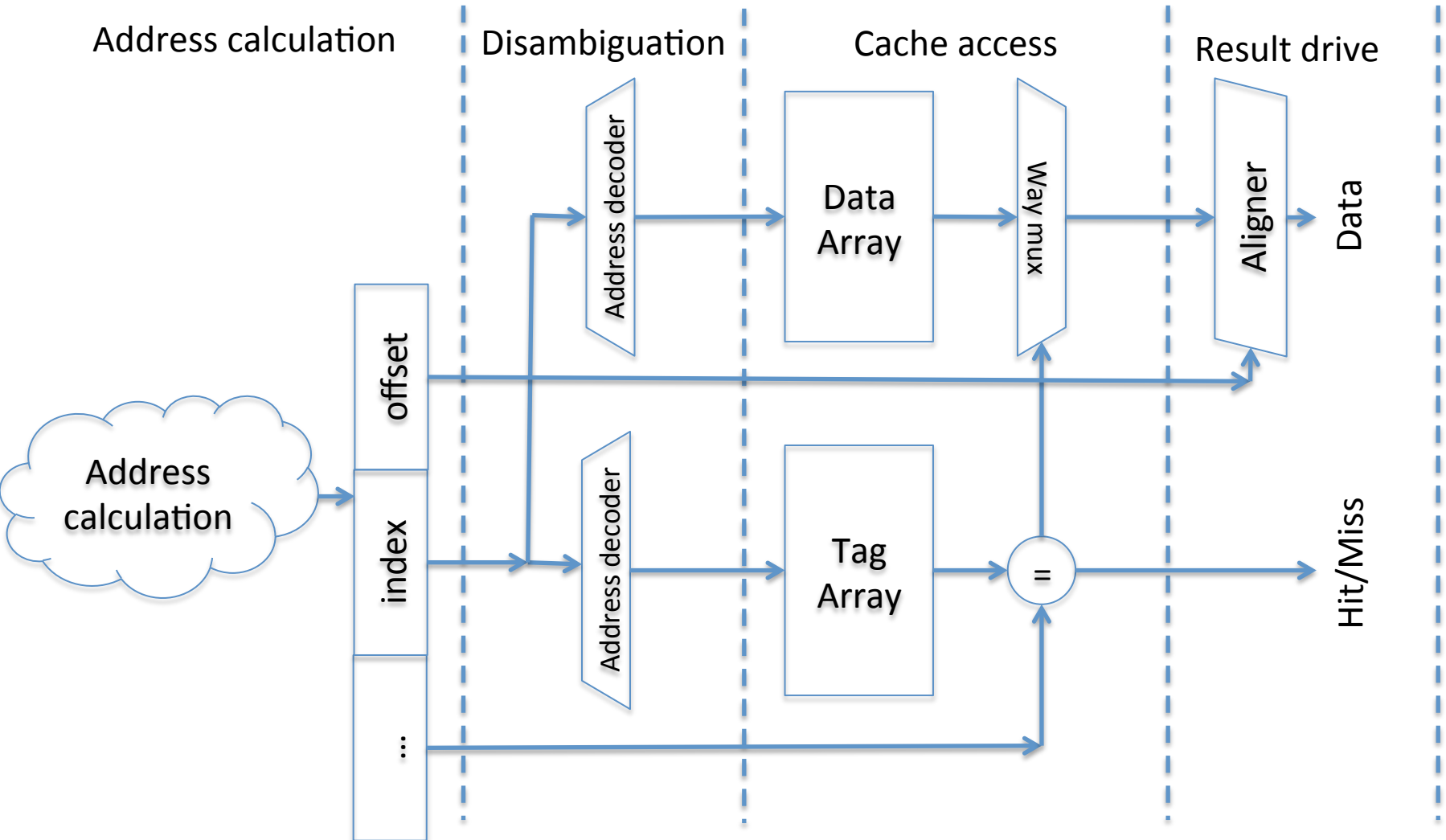
Multi ways



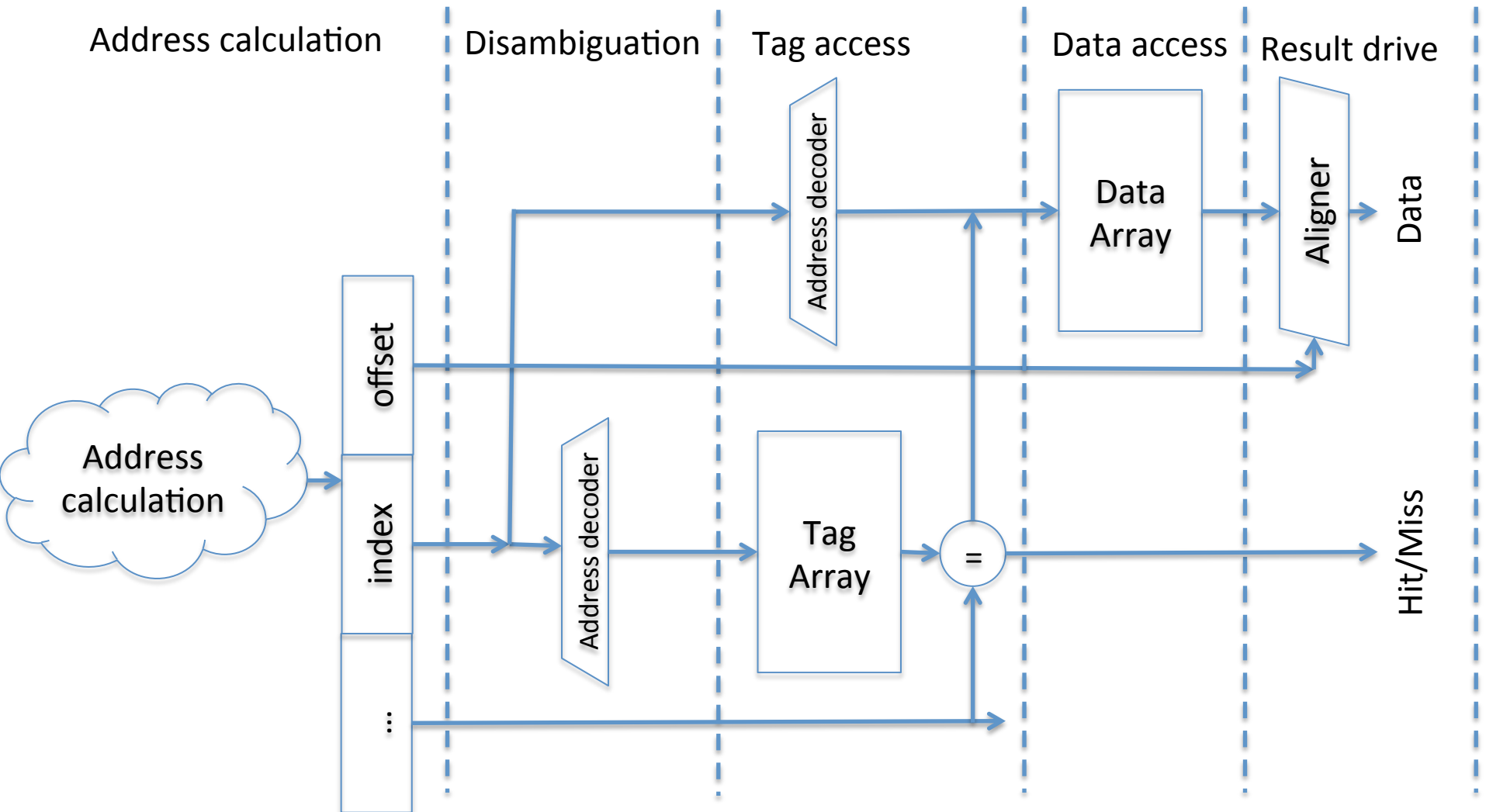
Real organization



Pipeline for Parallel Access



Pipeline for Serial Access



Lockup-free caches

- Can hold more than one ongoing access
 - Miss Status Holding Registers (MSHR)
- While fulfilling one miss, execute another access (Hit on Miss)
 - Primary miss: first miss to a cache block
 - Secondary miss: subsequent miss to a missing block
 - Structural-stall miss: extra misses without enough hardware resources

Other concepts

- Multiport
- Multibank

Instruction cache considerations

- Multiport vs single port
- Lookup free vs blocking
- Associativity