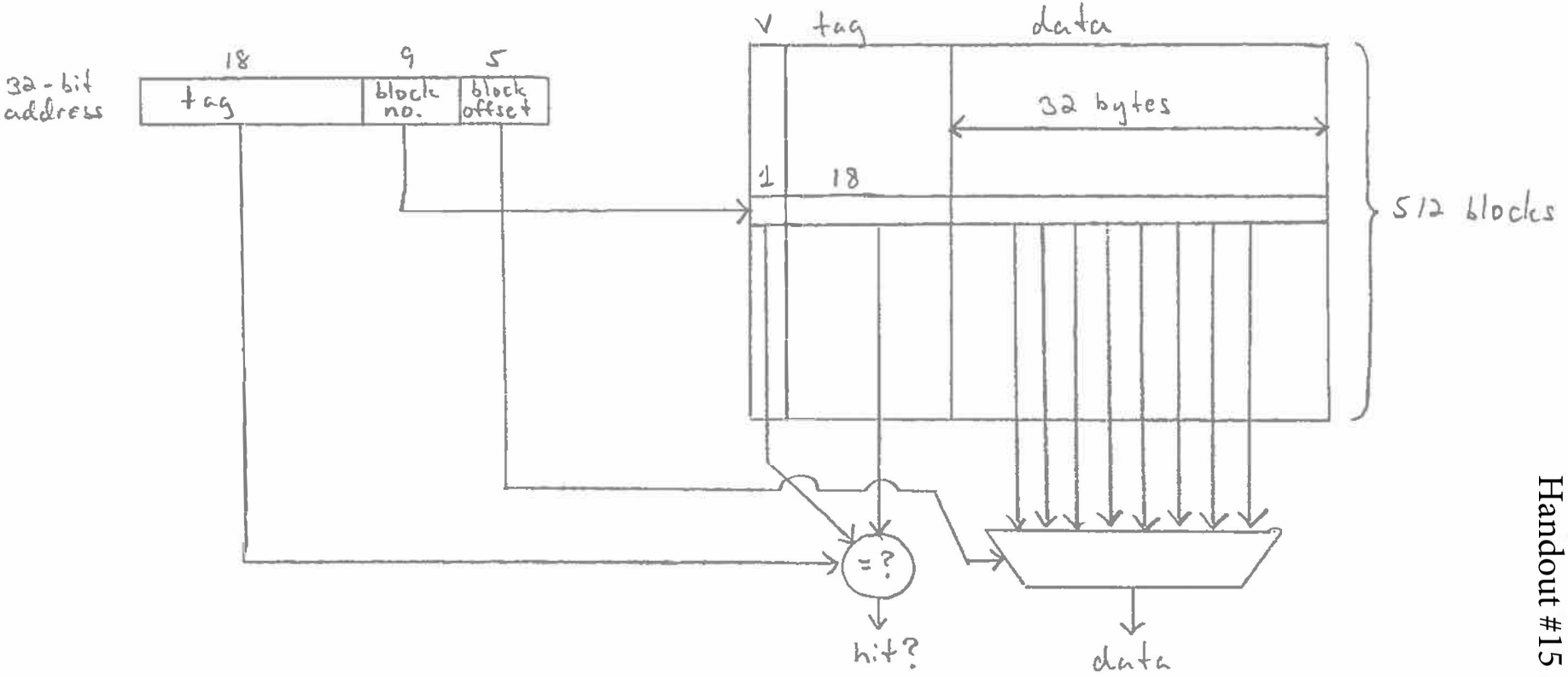
Direct-Mapped Cache Lookup

Ex: Cache Size = 16 KB
Block Size = 32 bytes
blocks = $\frac{16K}{32}$ = 512

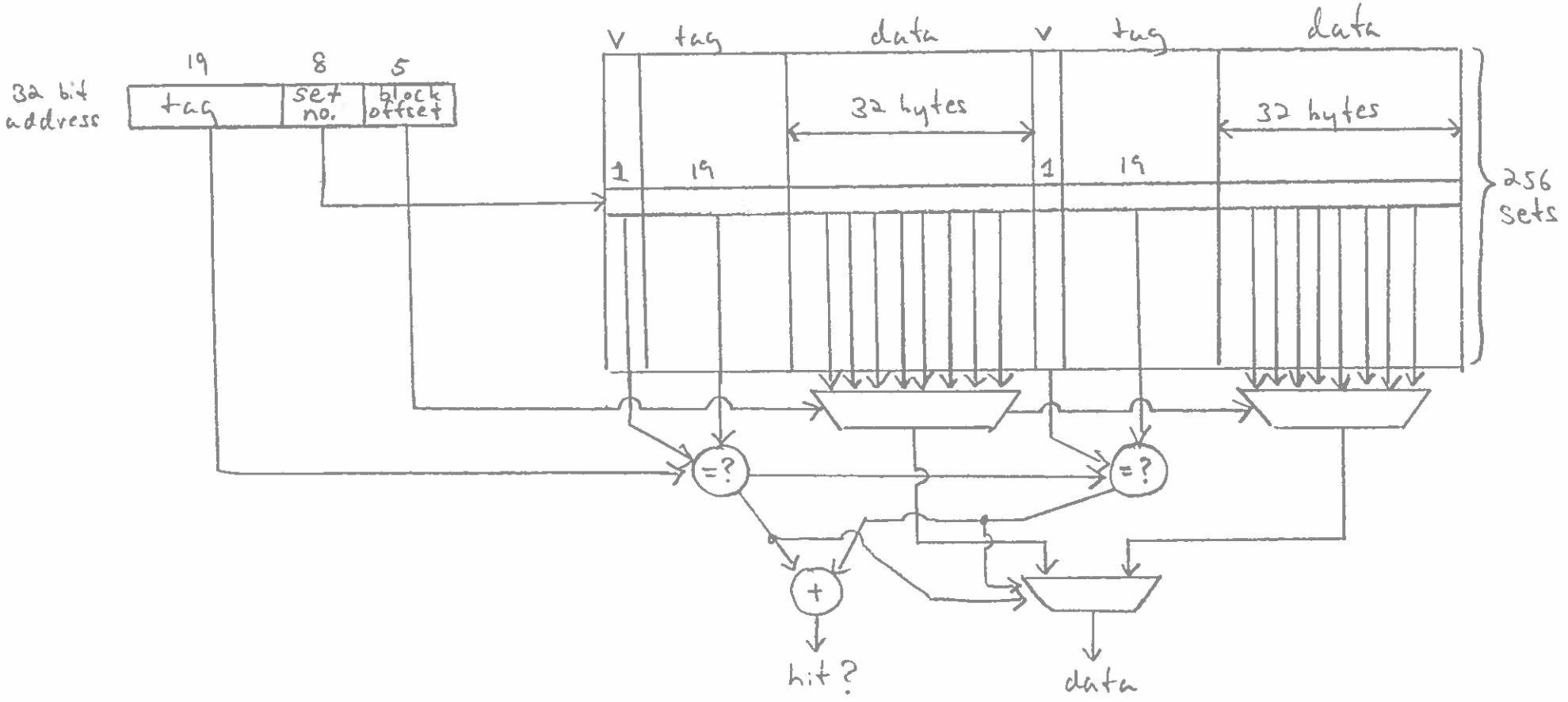


(2-way) Set Associative Cache Lookup

Ex: Cache Size = 16 KB

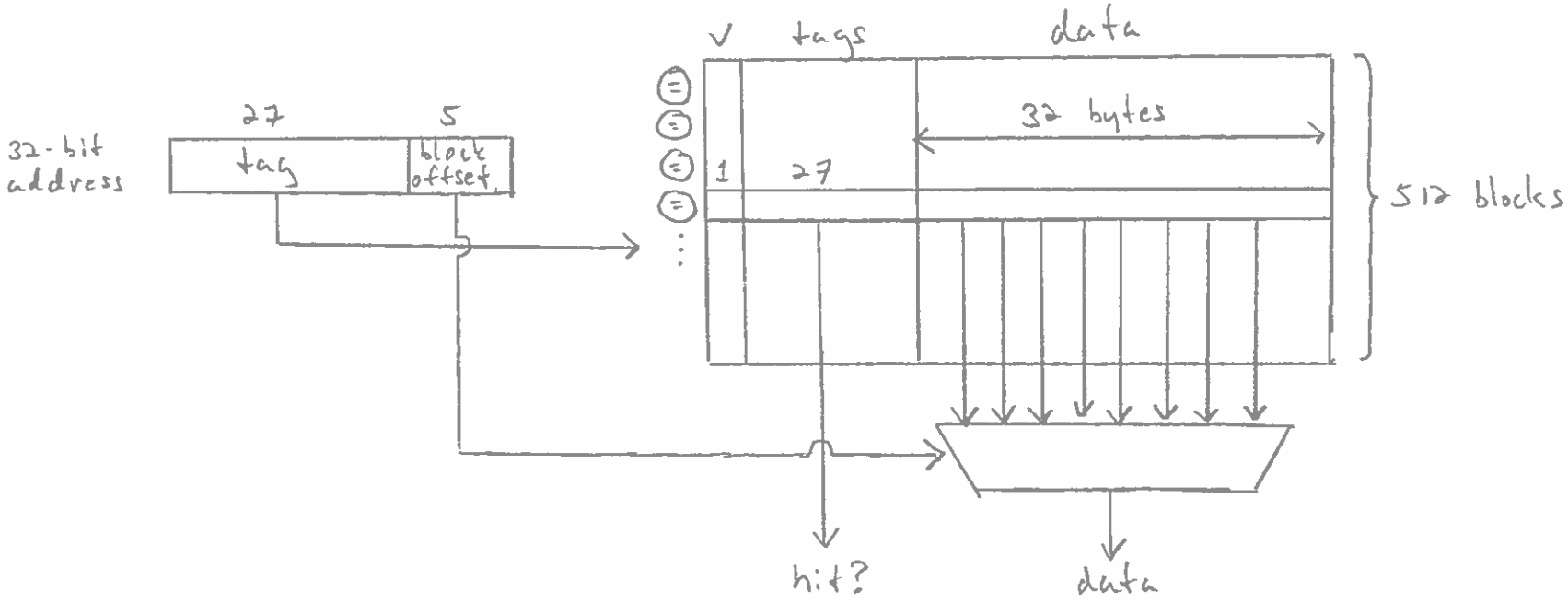
Block Size = 32 bytes

$$\# \text{ Sets} = \frac{16K}{32 \cdot 2} = 256$$



Fully Associative Cache Lookup

Ex: Cache Size = 16 KB
Block Size = 32 bytes
blocks = $\frac{16\,K}{32} = 512$



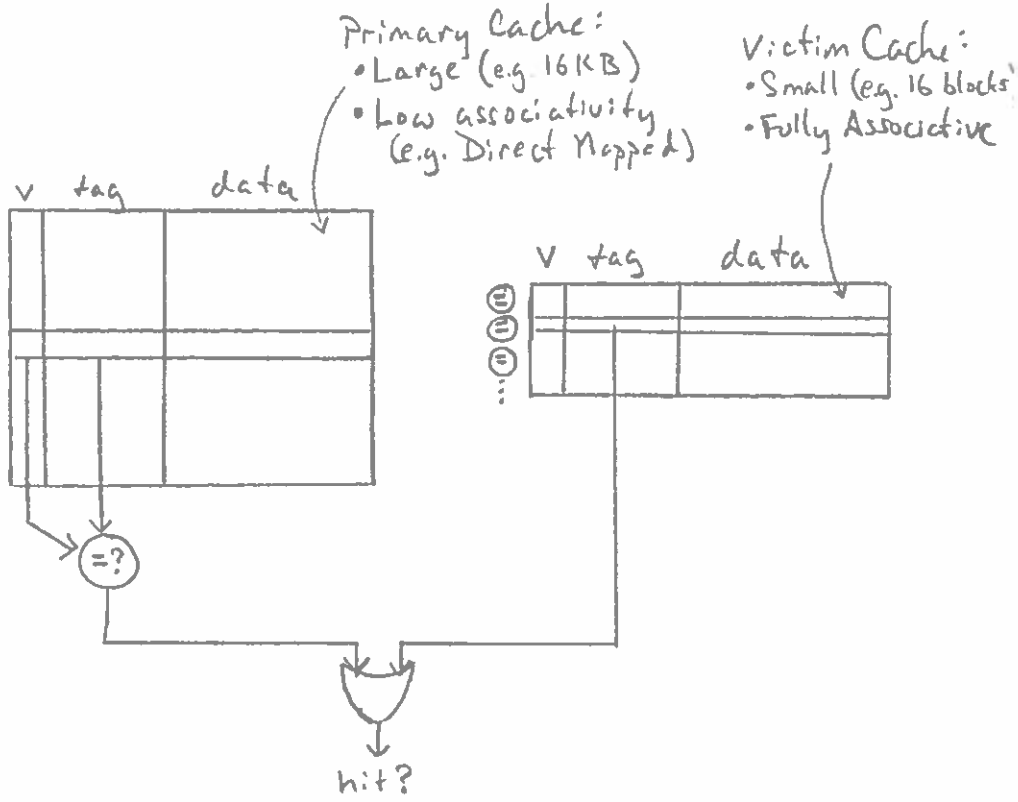
Victim Cache

Motivation:

Conflict misses usually isolated to a few sets

set	1	2	3	4	5
0	X				
1	X	X			
2	X				
3	X	X	X	X	X
4	X				
5	X				
6					
7	X	X			

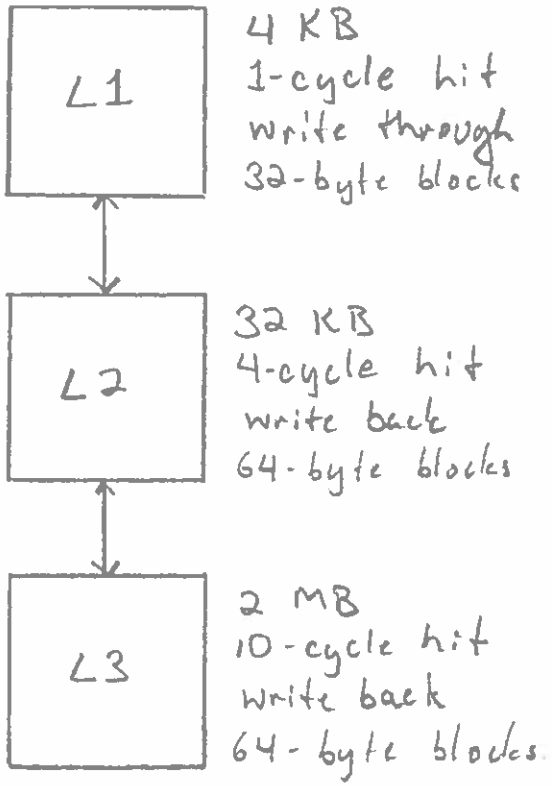
↑ number of blocks that map to each set.



- On access, check both caches (in parallel)
- Eviction from primary cache → victim cache
- Victim cache hit → primary cache
- If both caches miss, fetch from main memory → primary cache.

Multi-Level Caches

- Goal: Fast + large cache
- As you go down hierarchy:
  - Larger cache
  - Longer block size
  - Higher associativity
- Inclusion



Avg Mem. Acc. Time = $hit_{L1} + miss\ rate_{L1}(hit_{L2} + miss\ rate_{L2}(hit_{L3} + miss\ rate_{L3} \cdot miss\ penalty))$