# High Performance Computing

Cache Replacement Policy

Martin Raum

# Memory lines

Memory is subdivided in lines.

The AMD Epyc-CPU has lines size of 64 bytes.

Transfers to and from cache are not performed per-byte but per-line.

# Memory lines

Hexadecimal addresses of the first 2048 bytes split into lines and line indices:

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 00 | 08 | 10 | 18 | 20 | 28 | 30 | 38 | 40 | 48 | 50 | 58 | 60 | 68 | 70 | 78 | 0 | 1 |
| 80 | 88 | 90 | 98 | a0 | a8 | b0 | b8 | c0 | c8 | d0 | d8 | e0 | e8 | f0 | f8 | 2 | 3 |
| 100 | 108 | 110 | 118 | 120 | 128 | 130 | 138 | 140 | 148 | 150 | 158 | 160 | 168 | 170 | 178 | 4 | 5 |
| 180 | 188 | 190 | 198 | 1a0 | 1a8 | 1b0 | 1b8 | 1c0 | 1c8 | 1d0 | 1d8 | 1e0 | 1e8 | 1f0 | 1f8 | 6 | 7 |
| 200 | 208 | 210 | 218 | 220 | 228 | 230 | 238 | 240 | 248 | 250 | 258 | 260 | 268 | 270 | 278 | 8 | 9 |
| 280 | 288 | 290 | 298 | 2a0 | 2a8 | 2b0 | 2b8 | 2c0 | 2c8 | 2d0 | 2d8 | 2e0 | 2e8 | 2f0 | 2f8 | a | b |
| 300 | 308 | 310 | 318 | 320 | 328 | 330 | 338 | 340 | 348 | 350 | 358 | 360 | 368 | 370 | 378 | c | d |
| 380 | 388 | 390 | 398 | 3a0 | 3a8 | 3b0 | 3b8 | 3c0 | 3c8 | 3d0 | 3d8 | 3e0 | 3e8 | 3f0 | 3f8 | e | f |
| 400 | 408 | 410 | 418 | 420 | 428 | 430 | 438 | 440 | 448 | 450 | 458 | 460 | 468 | 470 | 478 | 10 | 11 |
| 480 | 488 | 490 | 498 | 4a0 | 4a8 | 4b0 | 4b8 | 4c0 | 4c8 | 4d0 | 4d8 | 4e0 | 4e8 | 4f0 | 4f8 | 12 | 13 |
| 500 | 508 | 510 | 518 | 520 | 528 | 530 | 538 | 540 | 548 | 550 | 558 | 560 | 568 | 570 | 578 | 14 | 15 |
| 580 | 588 | 590 | 598 | 5a0 | 5a8 | 5b0 | 5b8 | 5c0 | 5c8 | 5d0 | 5d8 | 5e0 | 5e8 | 5f0 | 5f8 | 16 | 17 |
| 600 | 608 | 610 | 618 | 620 | 628 | 630 | 638 | 640 | 648 | 650 | 658 | 660 | 668 | 670 | 678 | 18 | 19 |
| 680 | 688 | 690 | 698 | 6a0 | 6a8 | 6b0 | 6b8 | 6c0 | 6c8 | 6d0 | 6d8 | 6e0 | 6e8 | 6f0 | 6f8 | 1a | 1b |
| 700 | 708 | 710 | 718 | 720 | 728 | 730 | 738 | 740 | 748 | 750 | 758 | 760 | 768 | 770 | 778 | 1c | 1d |
| 780 | 788 | 790 | 798 | 7a0 | 7a8 | 7b0 | 7b8 | 7c0 | 7c8 | 7d0 | 7d8 | 7e0 | 7e8 | 7f0 | 7f8 | 1e | 1f |

# Cache associativity

Cache is also split into lines of the same size as memory.

Memory lines cannot be stored to cache in arbitrary positions.

$n$-way associative cache requires that

$$\text{index}_{\text{memory}} \equiv \text{index}_{\text{cache}} \pmod{n}.$$

# Cache associativity

$m$-way $n$-set associative cache of size $s$ means that memory:

- works with cache lines of size $s/(n \cdot m)$,
- divides memory into lines of that size,
- circularly assigns set indices to memory lines: $0 \ldots n-1$,
- cache can hold $m$ lines of each set index.

This is not the same as holding $m \cdot n$ lines in total, due to "conflict misses".

# Example of associativities

Associativities on one AMD Epyc-CPU are

**L1I** 4-way associative, 256 sets.

**L1D** 8-way associative, 64 sets.

**L2** 8-way associative, 1024 sets.

**L3** 16-way associative, 8192 sets.

The 32KiB L1D cache is 8-associative: each memory line has 8 possible positions in cache, but the cache has 512 lines in total.

# Cache associativity for 4-way associative 8-set

Memory lines:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | a | b | c | d | e | f |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 1a | 1b | 1c | 1d | 1e | 1f |
| 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 2a | 2b | 2c | 2d | 2e | 2f |
| 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 3a | 3b | 3c | 3d | 3e | 3f |
| 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 4a | 4b | 4c | 4d | 4e | 4f |
| 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 5a | 5b | 5c | 5d | 5e | 5f |
| 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 6a | 6b | 6c | 6d | 6e | 6f |
| 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 7a | 7b | 7c | 7d | 7e | 7f |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Lines indices of the 4-way associative cache:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | a | b | c | d | e | f |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 1a | 1b | 1c | 1d | 1e | 1f |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

# Cache replacement policy

When new a new memory line is loaded into cache, an old one might have to be replaced.

The rules for this are referred to as a cache placement policy.

The memory lines 0, 8, 28, 38, and 78 cannot be held in cache simultaneously, since the only match cache lines 0, 8, 10, and 18.

When loading them in this order, then on loading line 78, one of the previous ones has to be replaces.