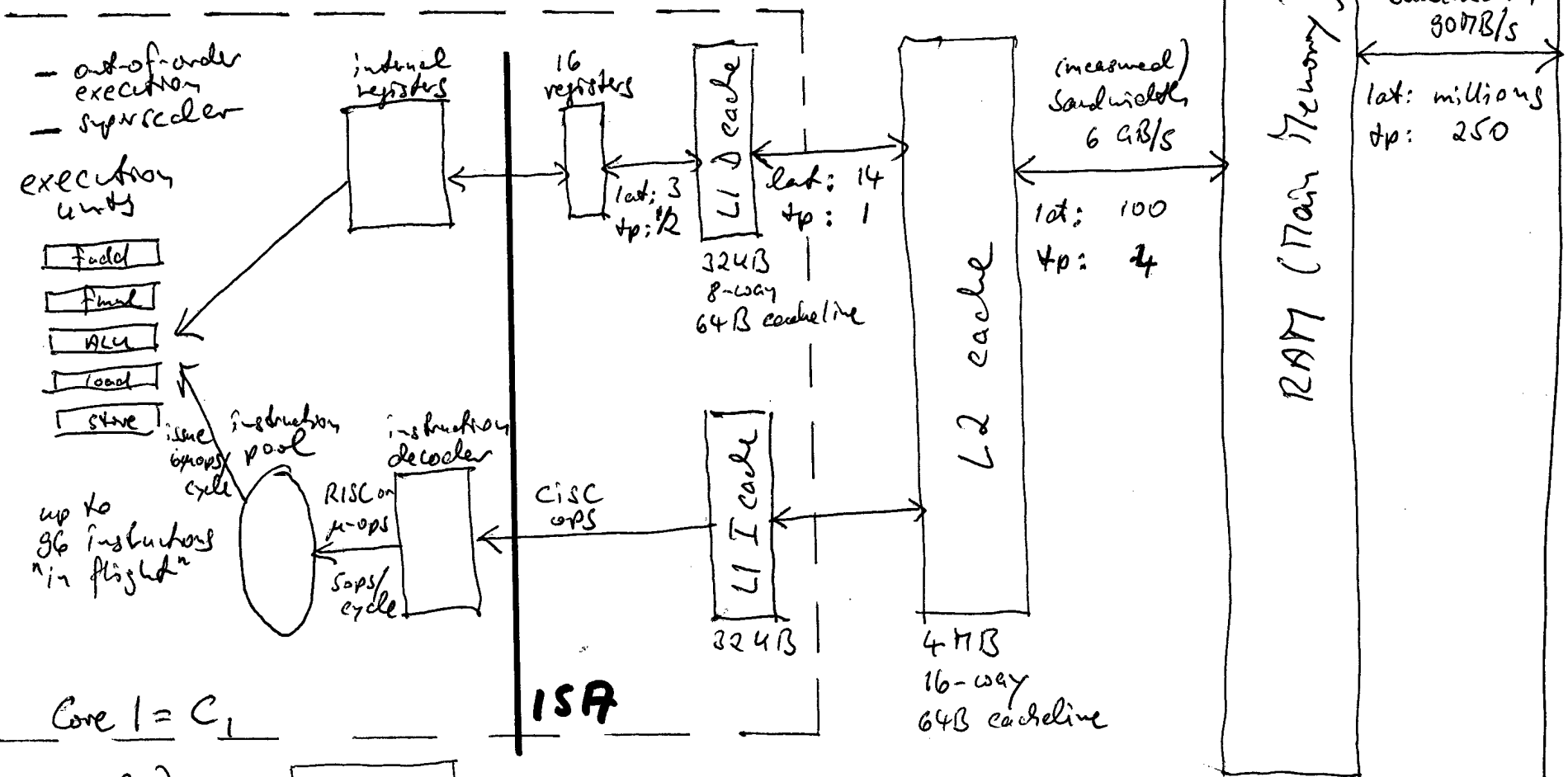
Microarchitecture (Example: Core)

we measure latency and throughput in (not in Bytes per second as usual)

cycles per double

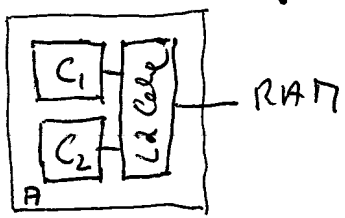
~~doubles per cycle~~

1 double = 8 bytes

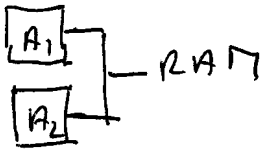


Core 1 = C_1

Core 2 Duo:



2 x Core 2 Duo:
(Core 2 Quad)



4 cores not on one die yet!

Memory hierarchy:

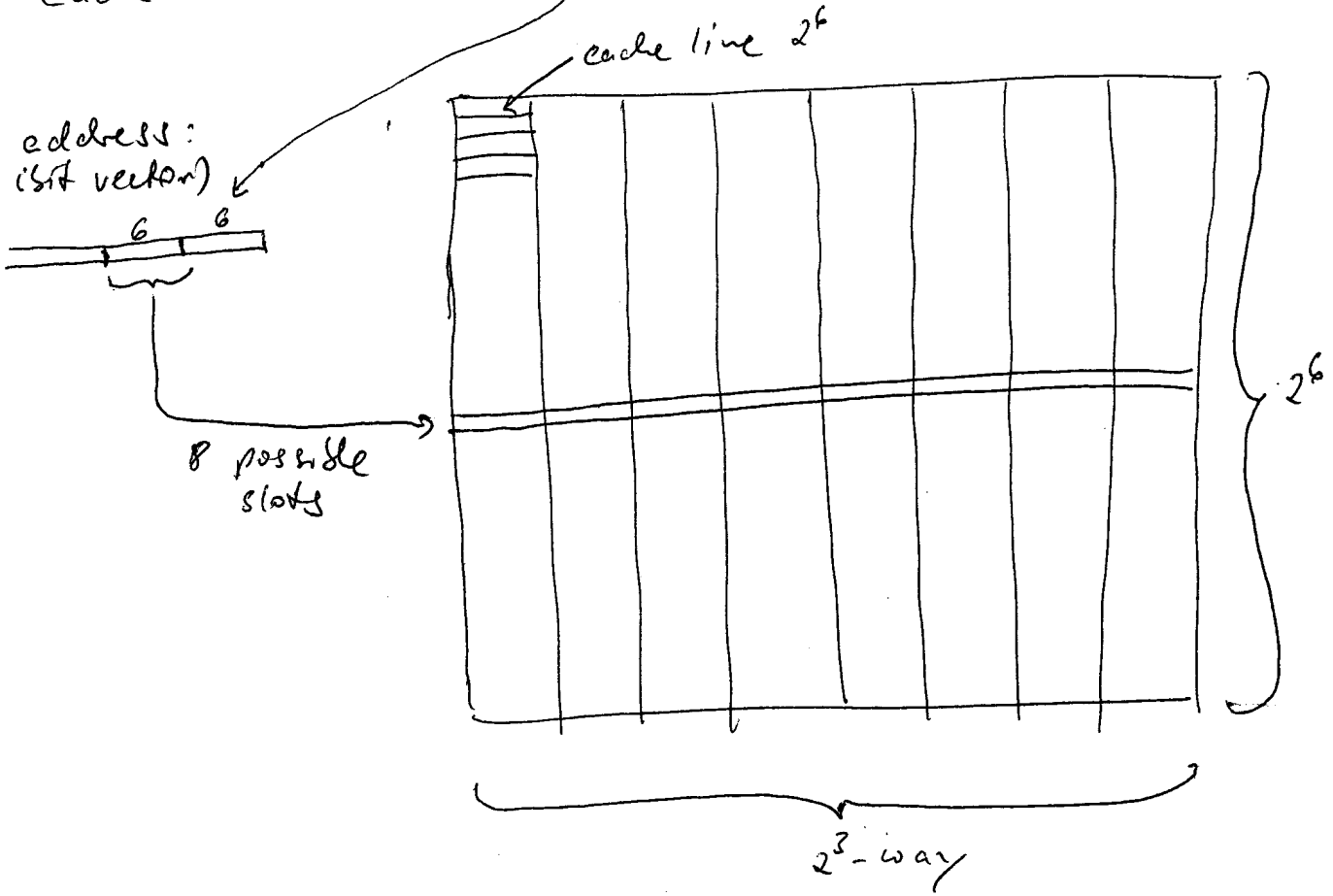
Registers
Caches
Main memory
Hard disk

Cache structure (Example L1 D-cache Core)

Size: $32\,\text{kB} = 2^{15}\,\text{B} = 2^{12}$ doubles

Associativity: 8-way $= 2^3$-way

Cache line: $64\,\text{B} = 2^6\,\text{B} = 2^3$ doubles



- memory is addressed Byte wise
- cache line: smallest amount of data brought into cache
- cache misses (CM):
  - compulsory: CM because data is accessed the first time
  - capacity: CM because cache is full
  - conflict: CM because data is loaded into occupied slot

Example: Load vector of length n into empty cache

$\Rightarrow \frac{n}{8}$ (compulsory) misses