Internminne og Cache

- RAM - Random Access Memory
- CPU utfører instruksjoner raskere enn data kan hentes fra RAM
- Cache er hurtigere og mellomlagrer både instruksjoner og data
- Ordet 'cache' er fransk og betyr 'hemmelig lager'
- Statistisk: 90% av tiden utføres instruksjoner innenfor 10% av det totale minnet
- Data som brukes i RAM ligger ofte ved siden av hverandre (eksempel: array)
- MYE tid å spare på å hente mange byte av gangen og mellomlagre i cache

Minnepyramiden

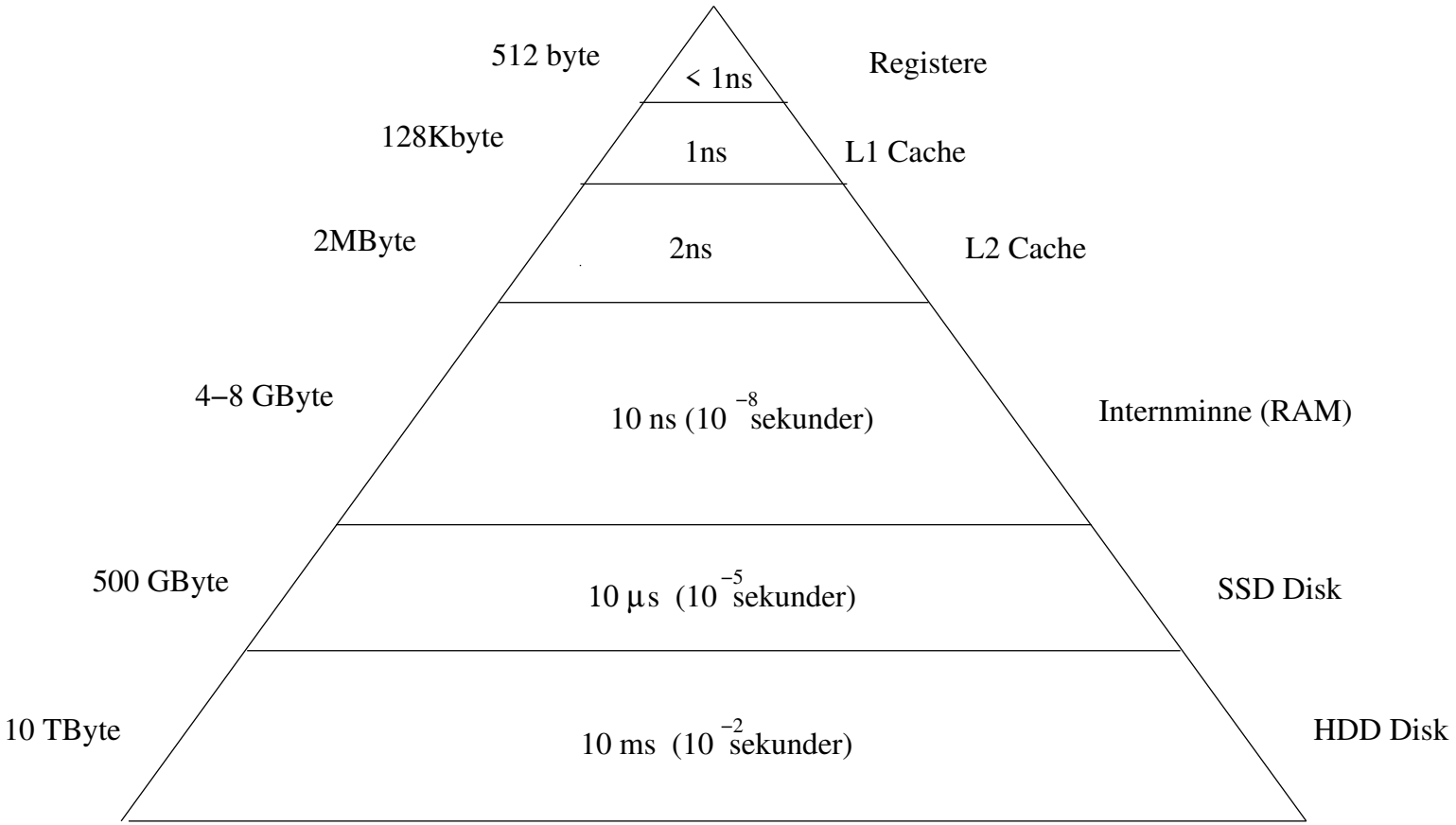

Figure: Minne-pyramiden. Størrelsen og tiden det tar å hente data øker nedover pyramiden.

SRAM og DRAM

- CPU-registere og cache er laget av SRAM (Static RAM)
- SRAM består av 6 transistorer for hver bit som lagres
- Internminnett er laget av DRAM (Dynamic RAM)
- DRAM består av kun en transistor og en kapasitator(lagrer elektrisk ladning)
- DRAM er ikke like hurtig og må lades på nytt 10 ganger i sekundet
- Nyeste versjon(2020): DDR5 SDRAM (Double-Data Rate generation 5 Synchronus Dynamic RAM)

L1 og L2 Cache

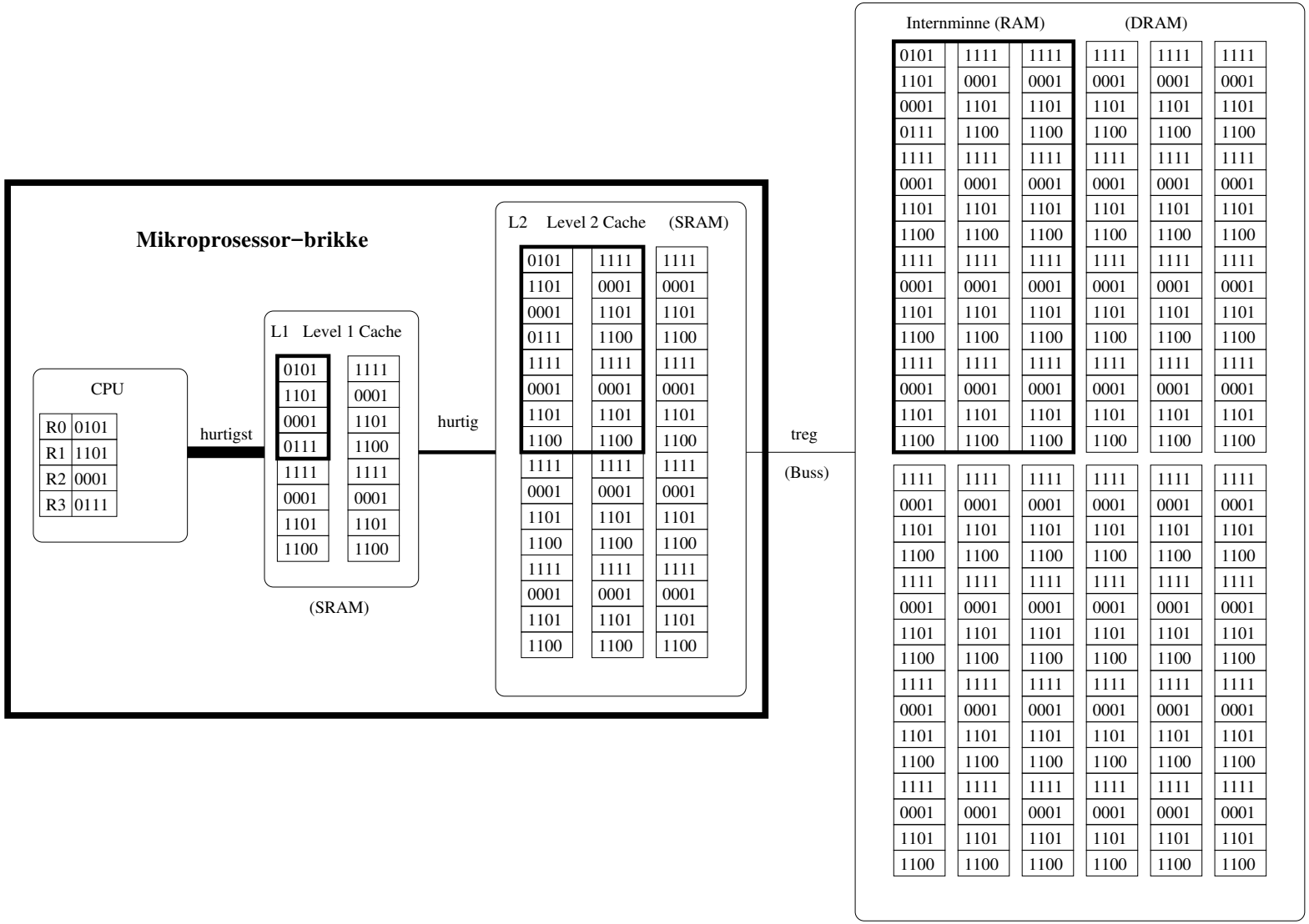

Figure: Level 1 cache (L1) ligger nærmest CPU. L2 er større, men har litt lengre aksesstid.

L1 og L2 Cache

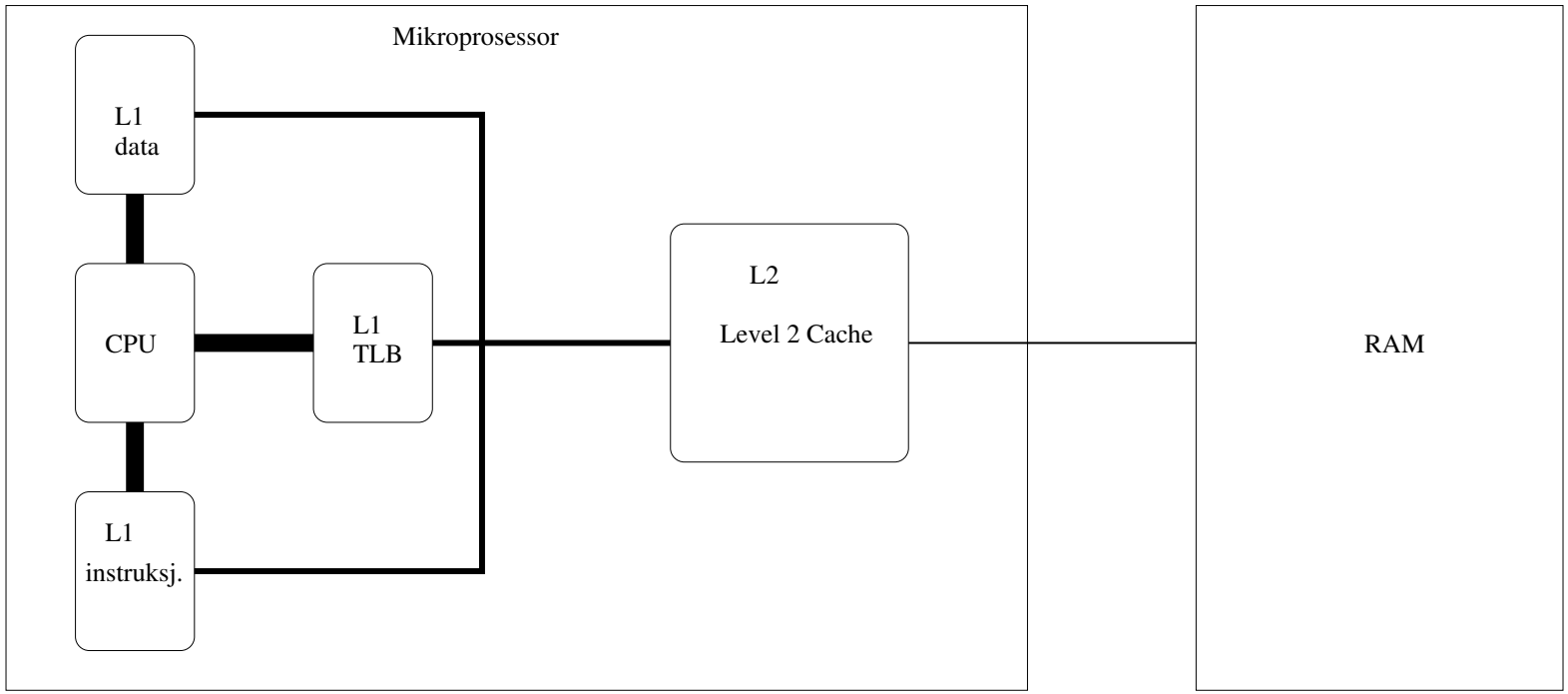

Figure: Level 1 cache (L1) bestående av tre deler. I AMD Athlon 64 er TLB i tillegg delt i to deler, en for adresser til instruksjoner og en for adresser til data.

Multitasking og Multiprocessing

- **Multitasking/Multiprogramming** Software(OS) brukes til å fordele tid fra samme CPU mellom flere prosesser
- **Multiprocessing** To eller flere CPU'er i samme computersystem kjører flere prosesser virkelig samtidig, på samme tidspunkt
- **Symmetric Multiprocessing** SMP, to eller flere prosessorer deler samme internminne og kjører flere prosesser virkelig samtidig, på samme tidspunkt
- **Multi Core Multiprocessing** To eller flere prosessorer på samme brikke deler cache og databus og kjører flere prosesser samtidig. Regnes også som SMP

Multiprosessor og Multicore

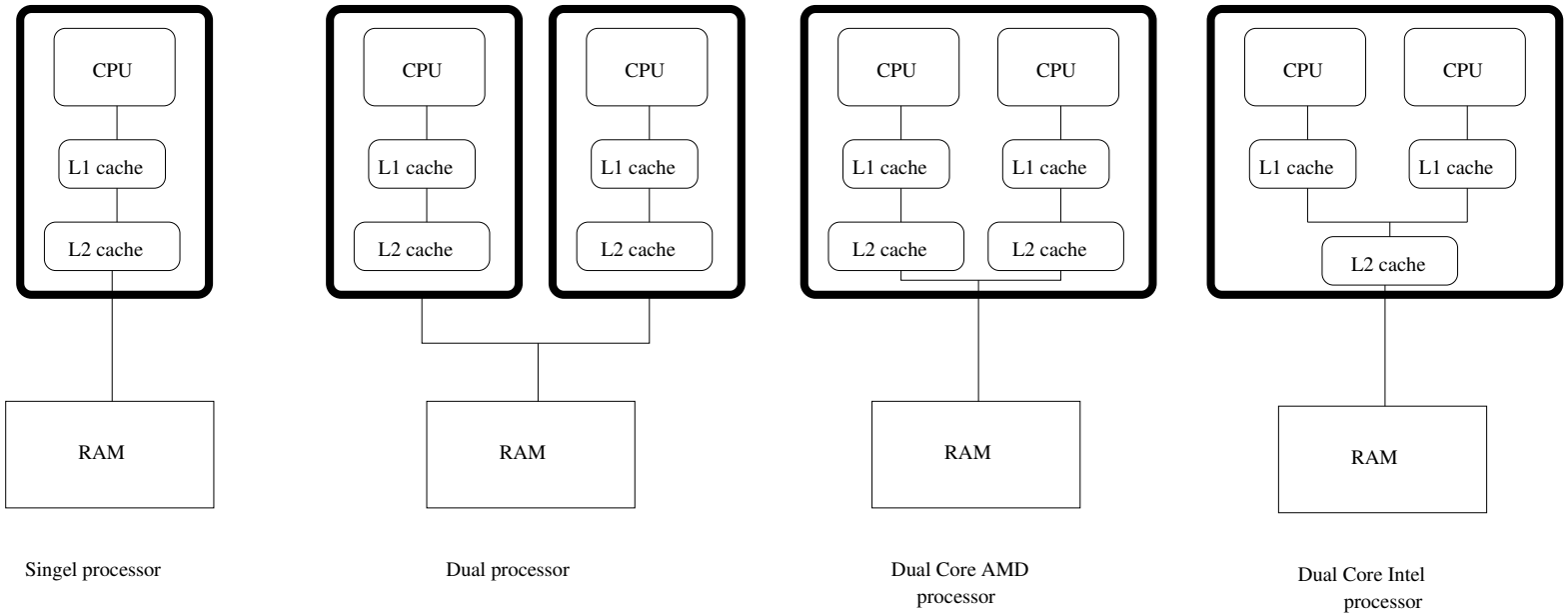

Figure: Single og dual prosessor og dual core prosessorer. Den tykke linjen markerer grensen for brikken/chip'en

Intel Core og AMD K10

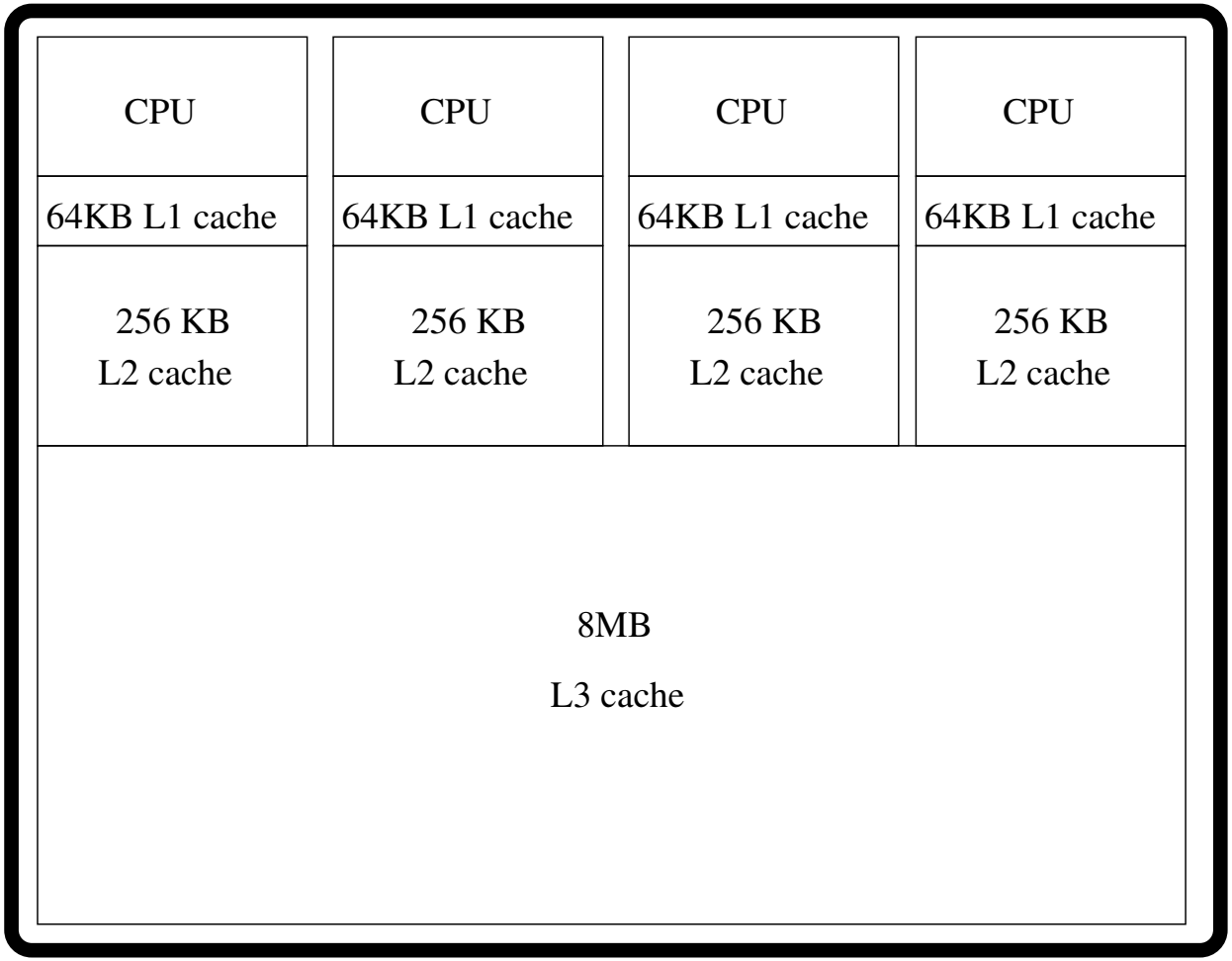

Figure: Intel Core i7 og AMD Opteron K10. K10 kjernen har 512KB L2 cache og i7 kjernene har hyperthreading, ellers er de i store trekke relativt like.

Hyperthreading

- Mange av Intels prosessorer som Pentium 4 og Xeon har såkalt hyperthreading teknologi
- En single core CPU kan inneholde to prosesser samtidig
- Hardware switcher selv mellom de to prosessene i løpet av nanosekunder
- Egne registere for hver prosess, men deler felles ALU
- Multitasking styres av OS
- Hyperthreading styres av hardware, betraktes som 2 CPUer av OS
- Hyperthreading er et intel-konsept; generell betegnelse: SMT (Simultaneous multithreading)
- AMD Zen-prosessorer har implementert SMT.

Hvorfor kan ikke en prosess utnytte to CPU-er?

Maskininstruksjoner som regner ut Fibonacci-rekken 1 1 2 3 4 5 13 21 34 55 etc:

```
1. mov 1, %ax     # %ax = 1
2. mov 1, %bx     # %bx = 1
3. add %ax, %bx   # %bx = %bx + %ax
4. add %bx, %ax   # %ax = %ax + %bx
5. jmp 3
```

Hvordan kan to prosessorer utnyttes?

Paralelliserbar kode

$$S = 1 + 2 + 3 + 4 + \cdots + 2000 = \sum_{i=1}^{2000} i$$

- En CPU kan regne ut $\sum_{i=1}^{1000} i$
- En annen CPU kan regne ut $\sum_{i=1001}^{2000} i$
- Men operativsystemet ser bare maskininstruksjoner og aner ikke hva som foregår!
- Programmereren må selv sørge for parallellisering, to prosesser eller to tråder
- Tråder (threads) kan kjøre uavhengig på hver sin CPU, men programmereren må fordele jobben på trådene
- Vi skal studere tråder i detalje senere i kurset