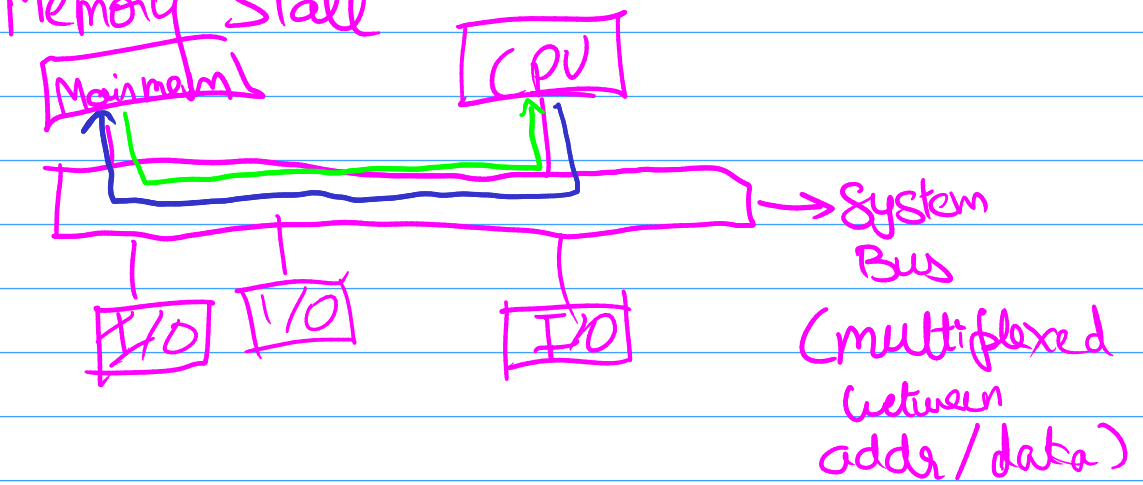
Memory Stall



(multiplexed between addr / data)

All these cant access @ once

Bus arbitration → Technique to share the bus amongst different components

When I/O is done (mainly to main memory through I/O controller & mem controller)

↓

DMA (not through some CPU)

Assumption till now

memory access time is negligible:

→ Operands are in regs

→ Operands are immediate

memory stall is zero

Not true, you can have delays in I/O sw due to CPU relinquishing main memory
⇒ CPU needs to wait.

Time required: Time to acquire system bus (α)
+
memory access time (locate memory cell) (β)

To either W/R one byte of data → $\alpha+\beta$

⇒ k bytes of data = $k(\alpha+\beta)$ [which is quite a bit]

if $\alpha = 0.1ns$
$\beta = 3ns$

500 bytes ⇒ $500 \times 3.1ns$ which isn't

[$\alpha \ll \beta$ (CPU decides when to relinquish the bus, I/O must request CPU)
if ⇒ CPU normally has access, only in the worst case (as we're considering), α has some value.]

⇒ Write/Read k bytes of data @ a time
⇒ Time = $\alpha + k\beta$
↳ access system bus just once

A few questions

① Why do I read k bytes of data?

(i) To reduce the system bus latency

*(ii) Locality (??)

② Where do I keep 'k' bytes of data?

Ans: Local memory in CPU, a.k.a cache.

Why is cache faster?

(i) Technology it uses

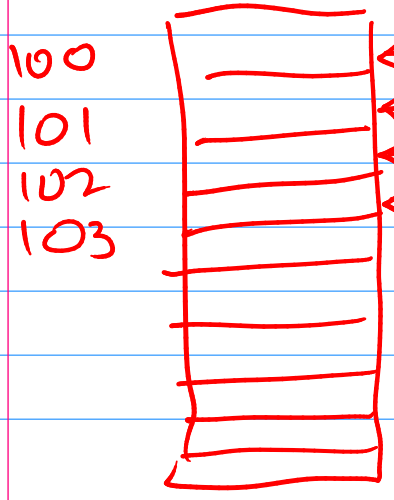
(ii) Closer to CPU

Locality of Reference: Spatial & Temporal

Spatial LOR

Access to memory locations occuring close together in space

Temporal LOR: Access to the memory location occuring close together in time



eg: two instructs next to each other

eg: jmp label

x → t

y → t+1

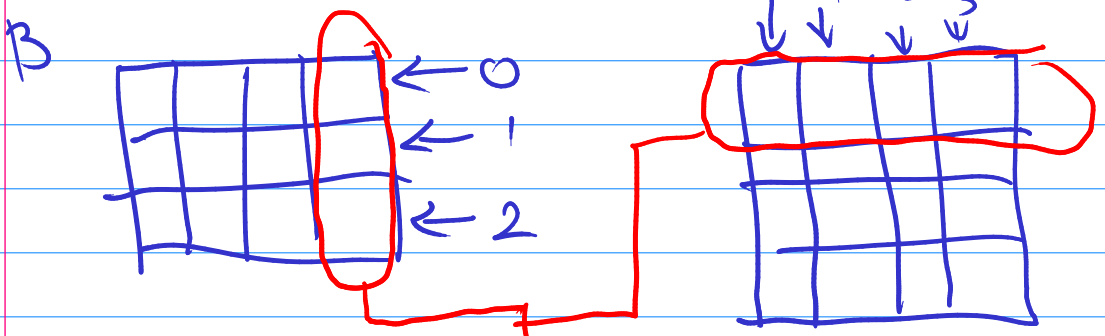
x may not be y-4 but they are temporally close

```
for(i=0; i<l; i++)
  for(j=0; j<m; j++)
    for(k=0; k<n; k++)
      A[i][j] = B[i][k] x C[k][j]
```

Row major

Spatial locality

C[0][0] then C[1][0] C[2][0] C[3][0]

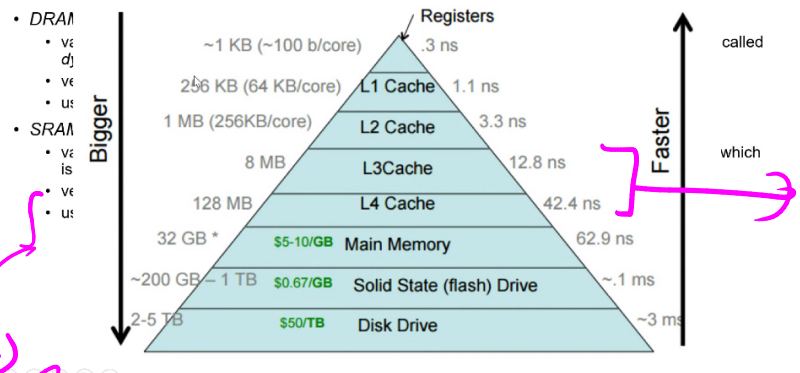


done @ RUNTIME by OS & H/W ← Bring the entire row / column @ once

* ⇒ This is why we should read k bytes of data

$\alpha + k\beta$, on caching we replace β with γ

Memory Hierarchy



some arch's only

Animation (ol, will try to take ss of this

Memory Wall
Performance Gap

The Memory Wall

Performance Gap