

MEMORY HIERARCHY DESIGN

unlimited amount of fast memory -> very expensive memory hierarchy.

to which takes advantage of locality,
temporal locality, spatial locality

several levels, each smaller, faster and more expensive per byte than next lower level.

dati contained in a lower level are a superset of next higher level - inclusion principle.

requirements are greater than for single core.

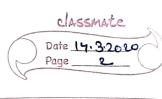
per core each clock cycle. clock rate 4.2 Crttz generate a peak of 32.8 billion 64-bit data memory refs.

in addition a peak instruction demand of about

12.8 billion 128-bit instruction reps.

total peak demand but of 409.6 981s.

this can be achieved by by multiporting and pipellining the raches. But for DRAM main memory using 2 channels is only 8x. of the demand bandwidth.



memory designers have focussed on optimizing average memory access time.

- Cache access time

- miss rate

- miss penalty.

power has become a major consideration, in highered microprocessors, but there may be 60MiB of on chip cache.

significant power.

we have leakage - static power - and we have active appower - agnormic power - when performing read and write operations

CPU is less aggressive, and power budget may

basics of memory hierarchy

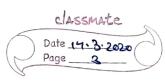
memory

storage.

be 20-50 times smaller

block (or line) spatial docality.

each block -tag.



we have direct mapped cache. fully associate rache Set associative cooke (block address) MOD (no of sets in cache). write through cache and write back, both strategies can use corite buffer. miss rate is fraction of cache accesses that result in a miss space. 3 Cs model for migses into 3 rategories - compulsory miss the very first access to a block cannot be in the cache - capacity miss. if the cache cannot contain all the blocks needed during execution of a program, capacity misses will occur. - conflict miss. · if the block placement strategy is not

fully associative, conflict misses will occur

by the when we have multithreading /

Coherence.

avg-mem access sime

= hit time + miss rate x miss penalty

BASIC OPTIMIZATIONS (6)

a large block size to reduce missrate

increases miss penalty

2 bigger caches reduce miss rode

memory + power + cost

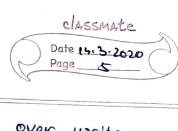
3. higher associativity to reduces misses rate.

at the cost of increased hit time + power.

potentially longer hit time of the larger cache

y multi-level caches to reduce miss penalty.

hit time, + missrate, Chit timez + missrate, miss penalty iz



5.	giving priority to read misses over write		
	to reduce miss penalty.		
	Counites can be put in write buffer)		
6.	avoiding address translation ofuning and arises		

the cache to reduce hit time

	Keywords lingle withi	Date 1.4.2020 Page 1
	energy consumption is also a considerate atteast 2 TLB out of critical	
	would prediction to reduce would time	2 .
	merging write buffers to reduce mis	s penalty.
	compiler optimization to reduce miss	rate.
y -	compiler controlled prefetching to r or miss vate.	educe miss penalty
	using HBM to extend memory hierarchi	J
	MEMORY TECHNOLOGY AND OPTIMIZATED	, ZNC
	phase changemenory - memristor	
	WAREHOUSE SCALE COMPUTERS	