Topic V2F

Memory Hierarchy

Reading: (Section 5.1)

Memory Technology

SSD Access time: 35 to 100 microseconds = 35 to 100 thousand ns

Memory technology	Typical Access Time	$ per GB in 2008	$ per GB in 2020
SRAM			
DRAM			
Storage Drivers			

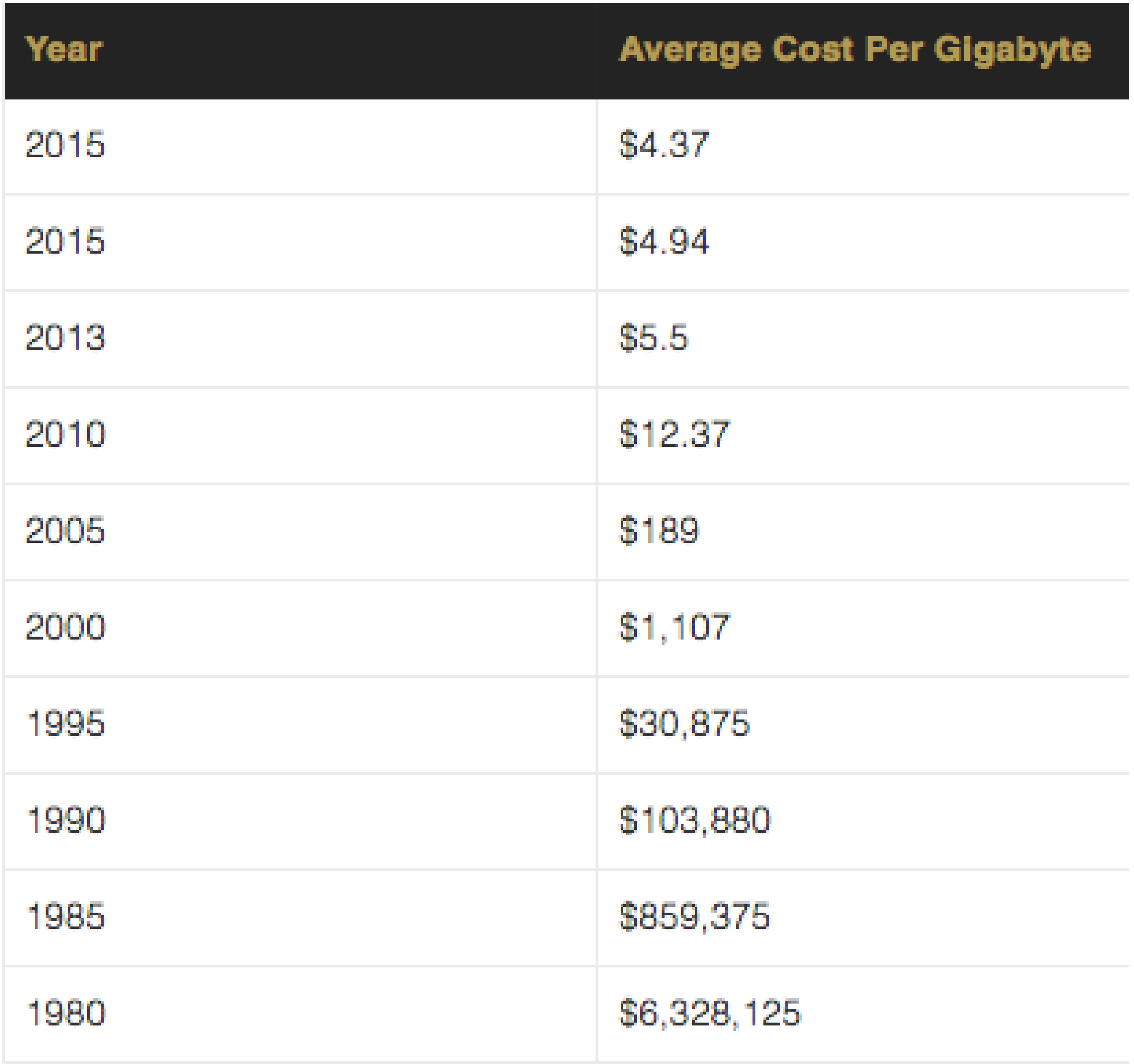
100-140 clock cycles

- Ideal memory
  - Access time of SRAM

$$2\ GHz = 2 \square 10^{9}\ Hz$$

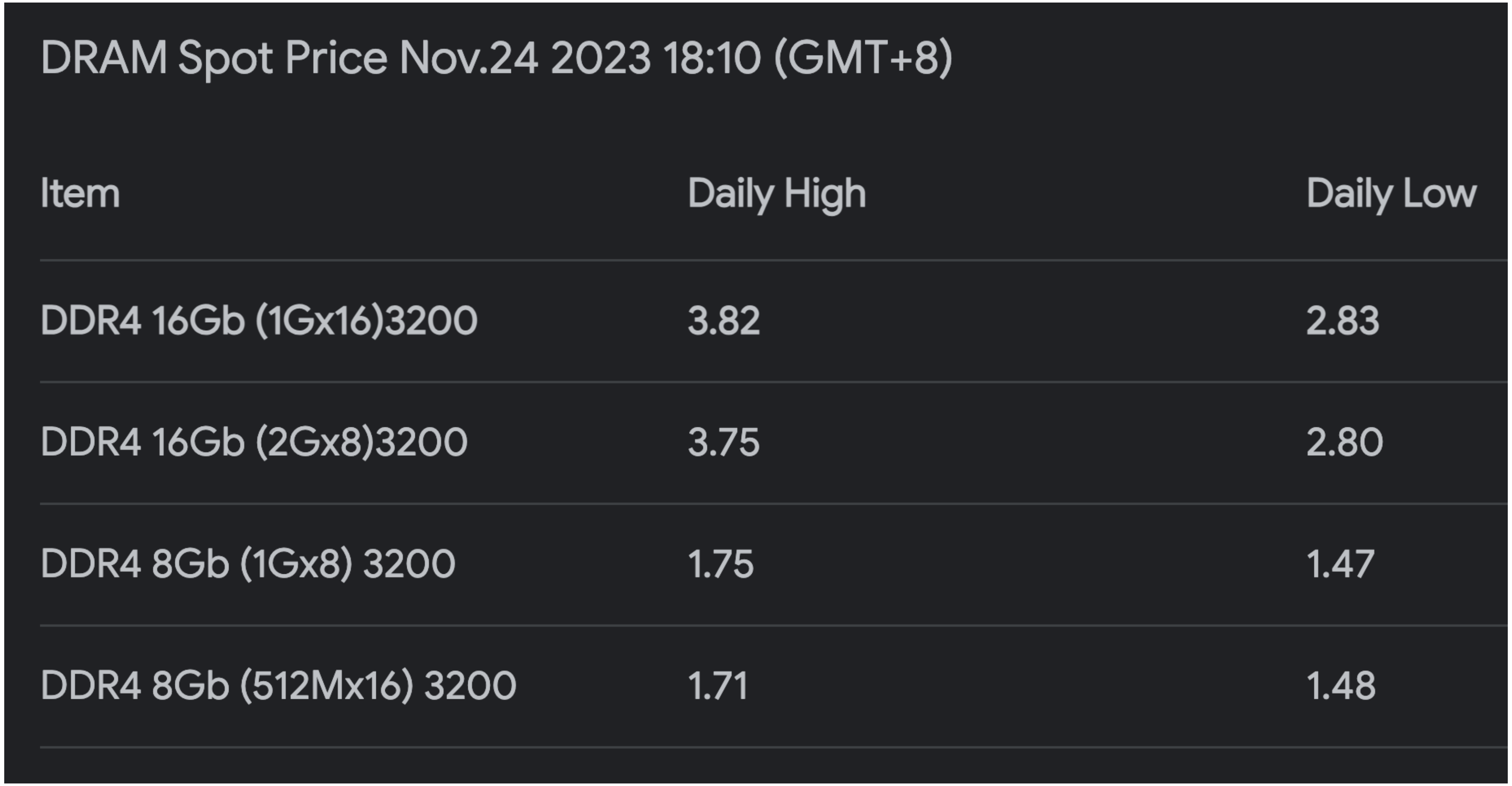
$$\square \quad \text{clock cycle} = 0.5\ ns$$

Some Historical Prices (DRAM)



Year	Average Cost Per Gigabyte
2015	$4.37
2015	$4.94
2013	$5.5
2010	$12.37
2005	$189
2000	$1,107
1995	$30,875
1990	$103,880
1985	$859,375
1980	$6,328,125

$2.97/GB

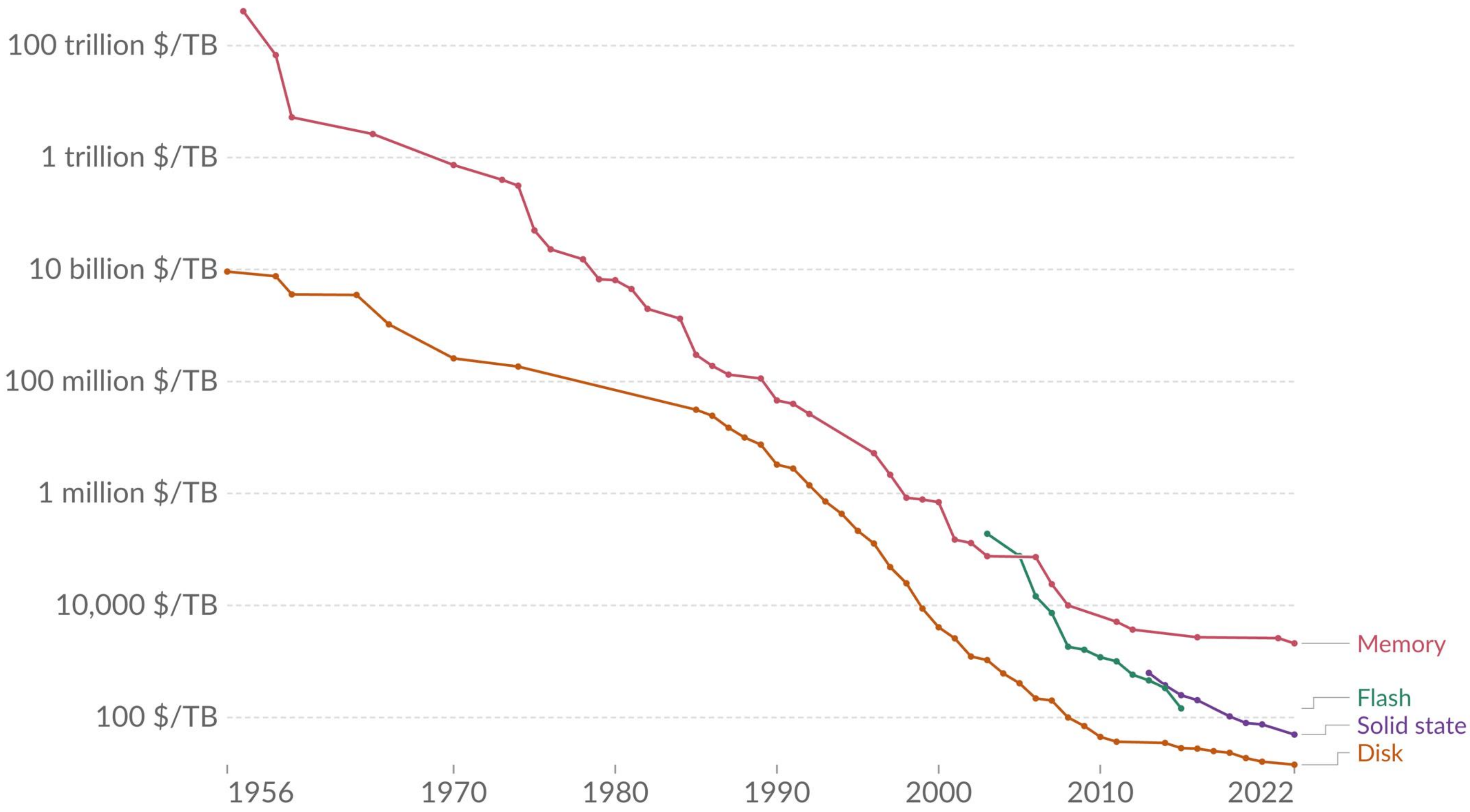


DRAM Spot Price Nov.24 2023 18:10 (GMT+8)

Item	Daily High	Daily Low
DDR4 16Gb (1Gx16)3200	3.82	2.83
DDR4 16Gb (2Gx8)3200	3.75	2.80
DDR4 8Gb (1Gx8) 3200	1.75	1.47
DDR4 8Gb (512Mx16) 3200	1.71	1.48

Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.



Data source: John C. McCallum (2022)

OurWorldInData.org/technological-change | CC BY

Note: For each year, the time series shows the cheapest historical price recorded until that year.

https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage

Principle of Locality

Programs access a small proportion of their address space at any time

Temporal locality

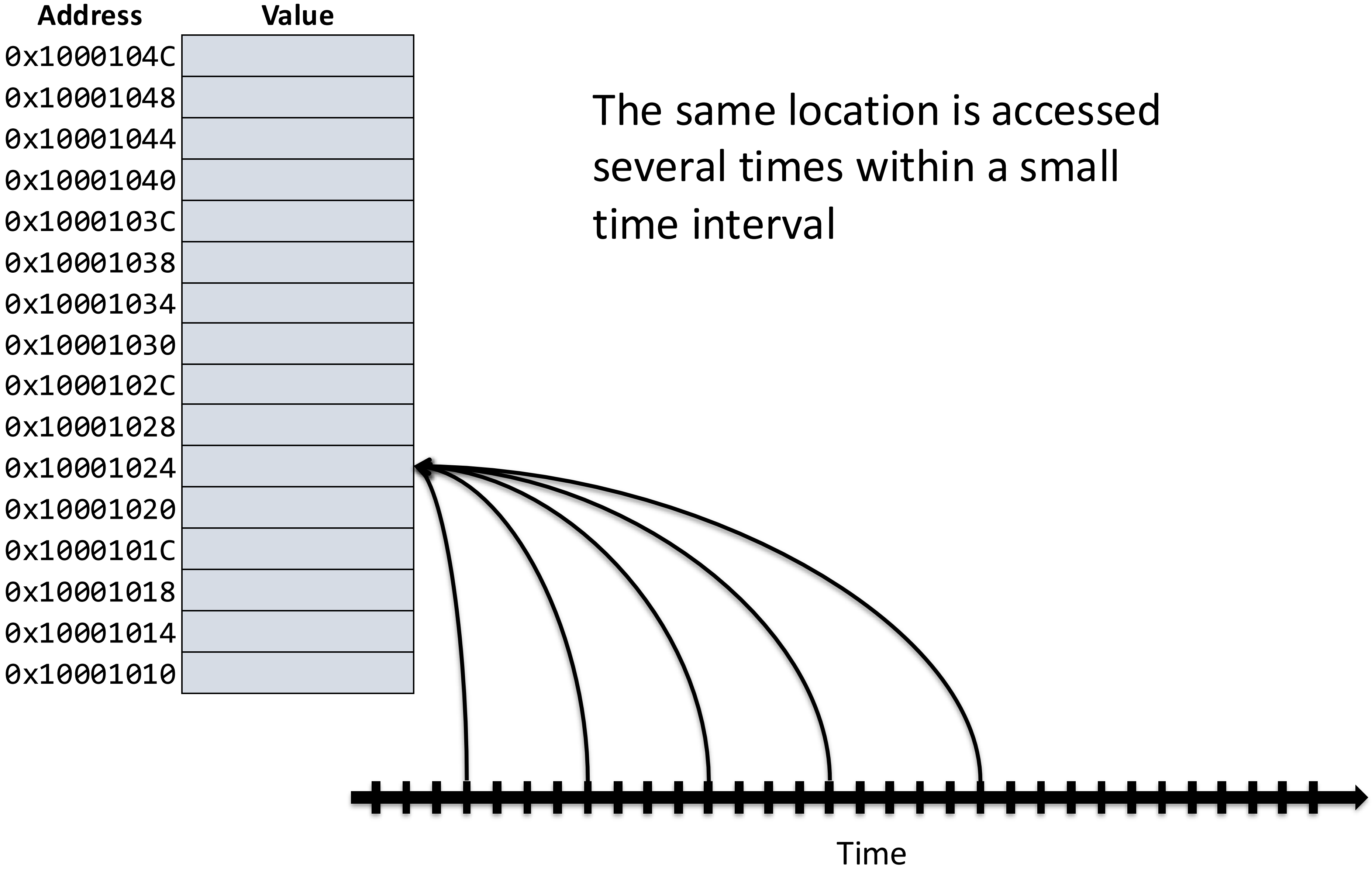
- Items accessed recently are likely to be accessed again soon
  - E.g., instructions in a loop, induction variables, webpages

Spatial locality

- Items near those accessed recently are likely to be accessed soon

Temporal Locality

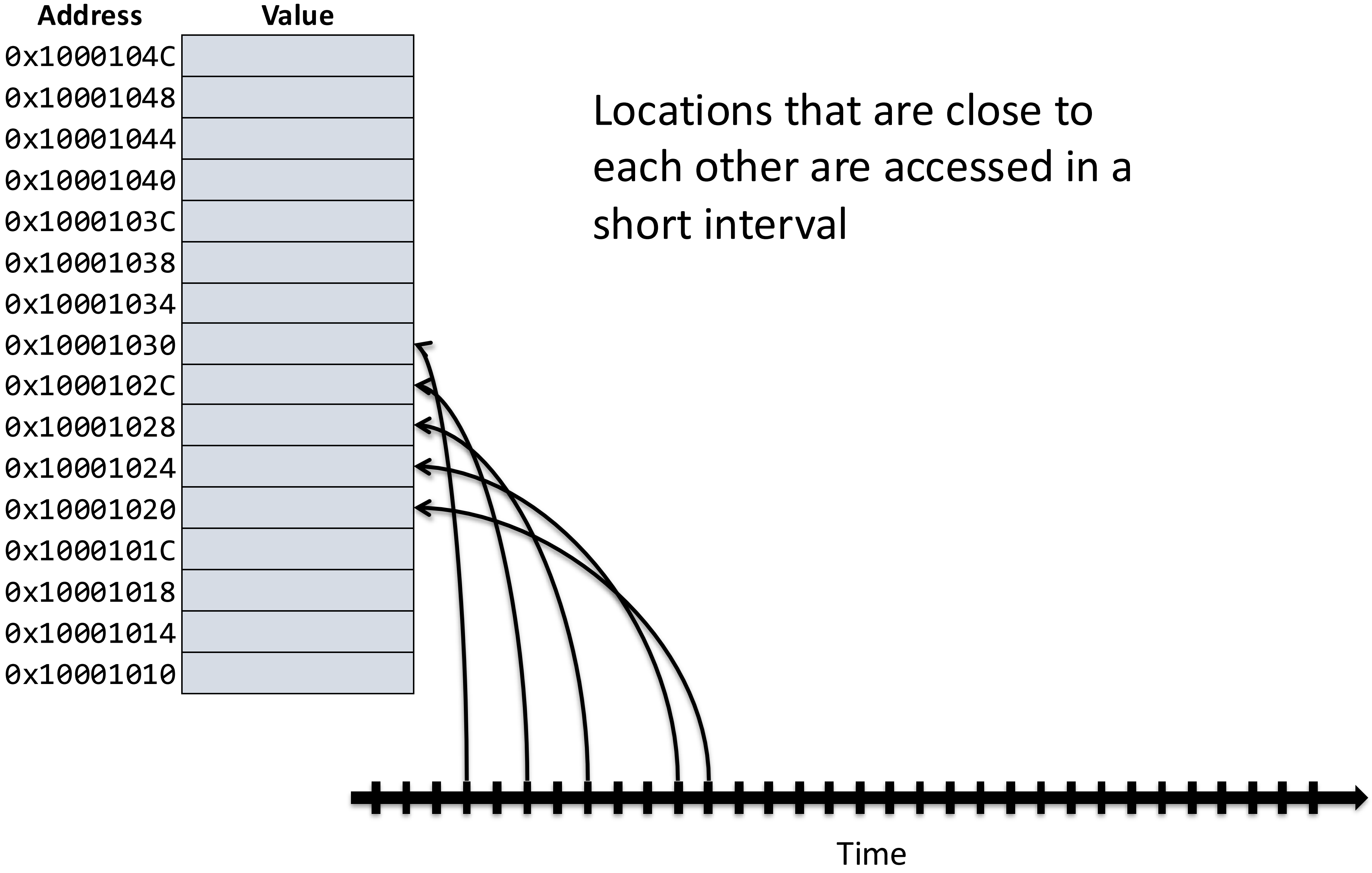
The same location is accessed several times within a small time interval



Address	Value
0x1000104C	
0x10001048	
0x10001044	
0x10001040	
0x1000103C	
0x10001038	
0x10001034	
0x10001030	
0x1000102C	
0x10001028	
0x10001024	
0x10001020	
0x1000101C	
0x10001018	
0x10001014	
0x10001010	

Spatial Locality

Locations that are close to each other are accessed in a short interval



Address	Value
0x1000104C	
0x10001048	
0x10001044	
0x10001040	
0x1000103C	
0x10001038	
0x10001034	
0x10001030	
0x1000102C	
0x10001028	
0x10001024	
0x10001020	
0x1000101C	
0x10001018	
0x10001014	
0x10001010	

Taking Advantage of Locality

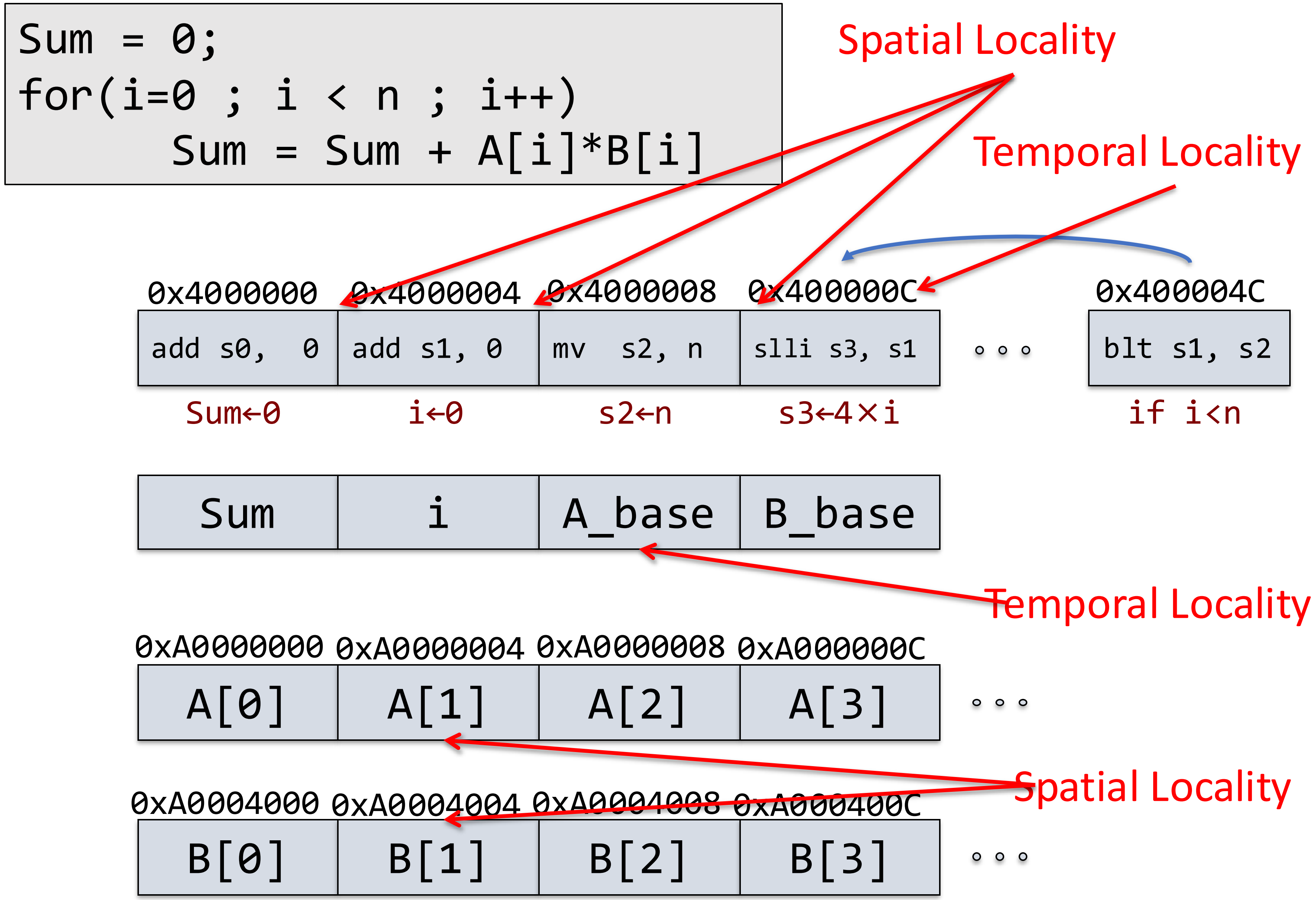
Memory hierarchy

Everything (code and data) is stored on Solid State Drives (SSDs)

Copy recently accessed (and nearby) items from disk to smaller DRAM memory

- Main memory

Copy more recently accessed (and nearby) items from DRAM to smaller SRAM memory



```
Sum = 0;
for(i=0 ; i < n ; i++)
      Sum = Sum + A[i]*B[i]
```

Spatial Locality

Temporal Locality

0x4000000 0x4000004 0x4000008 0x400000C 0x400004C

add s0, 0 | add s1, 0 | mv s2, n | slli s3, s1 | ∘ ∘ ∘ | blt s1, s2

Sum←0 i←0 s2←n s3←4×i if i<n

Sum | i | A_base | B_base

Temporal Locality

0xA0000000 0xA0000004 0xA0000008 0xA000000C

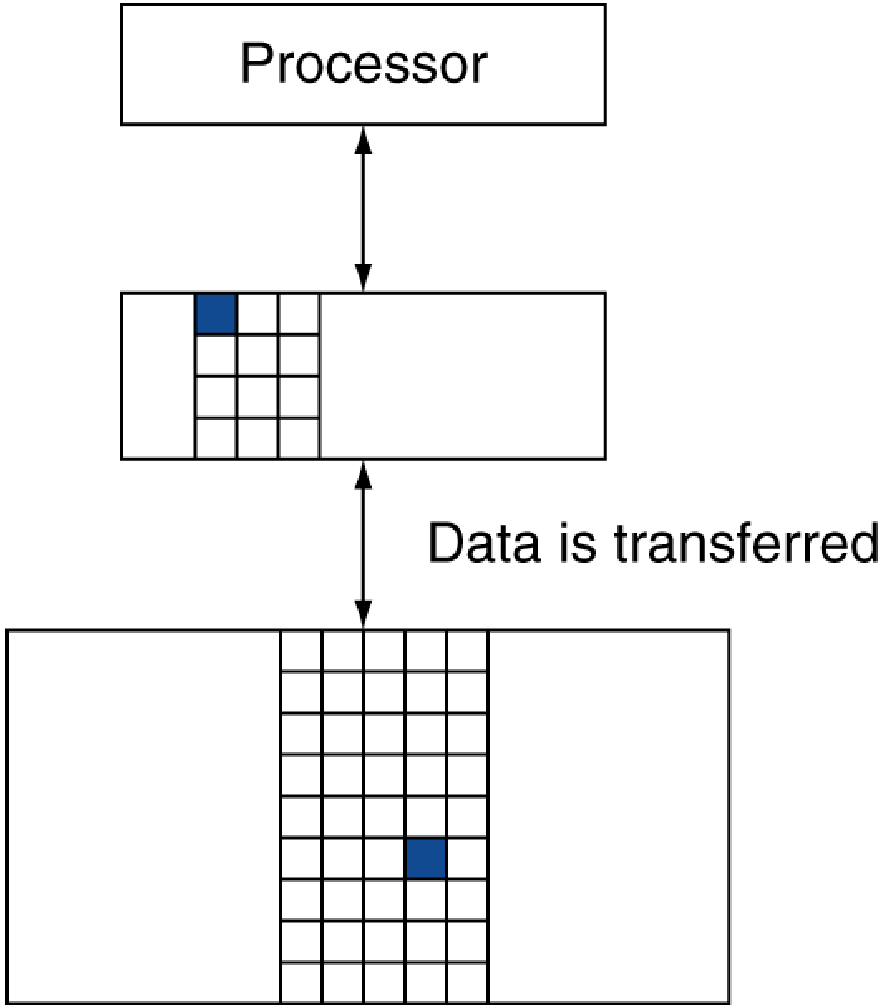
A[0] | A[1] | A[2] | A[3] | ∘ ∘ ∘

Spatial Locality

0xA0004000 0xA0004004 0xA0004008 0xA000400C

B[0] | B[1] | B[2] | B[3] | ∘ ∘ ∘

Memory Hierarchy Levels



Block (aka line): unit of copying

- May be multiple words

If data is present in upper level

- Hit: access satisfied by upper level
  - Hit ratio: hits/accesses

If data is absent in upper level

- Miss: block copied from lower level
  - Time taken: miss penalty
  - Miss ratio: misses/accesses
    = 1 – hit ratio