CSE 112: Computer Organization

Instructor: Sujay Deb

Lecture 20

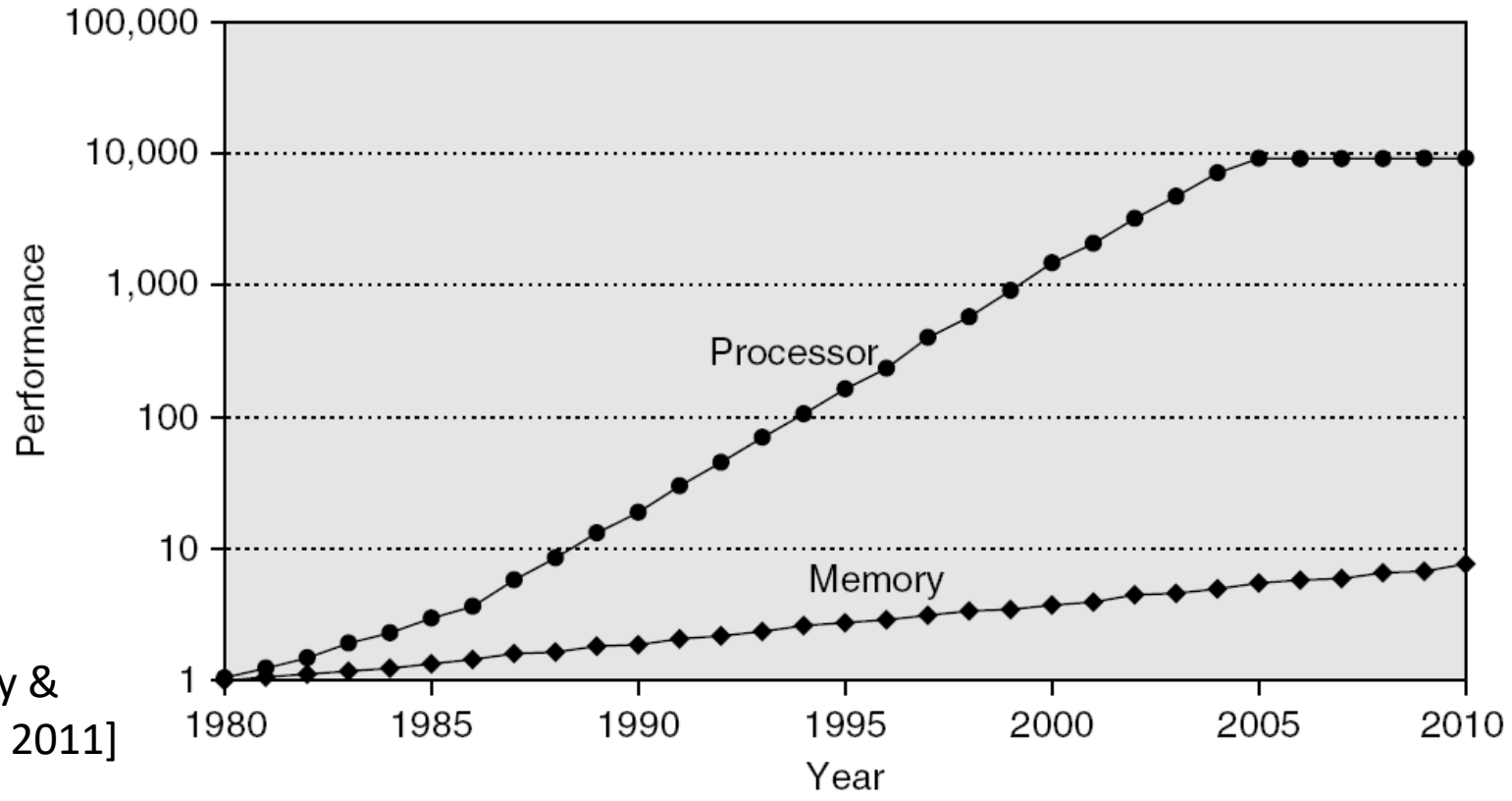
Class Interaction # 23

CSE 112: Computer Organization

Instructor: Sujay Deb

Lecture 21

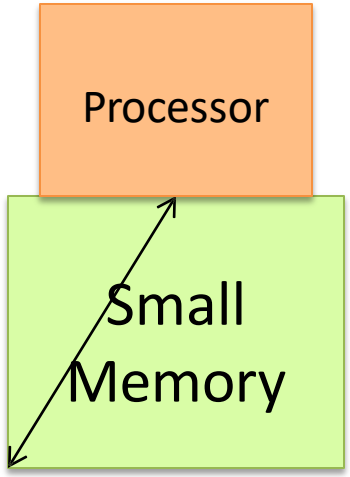
Processor-DRAM Latency Gap

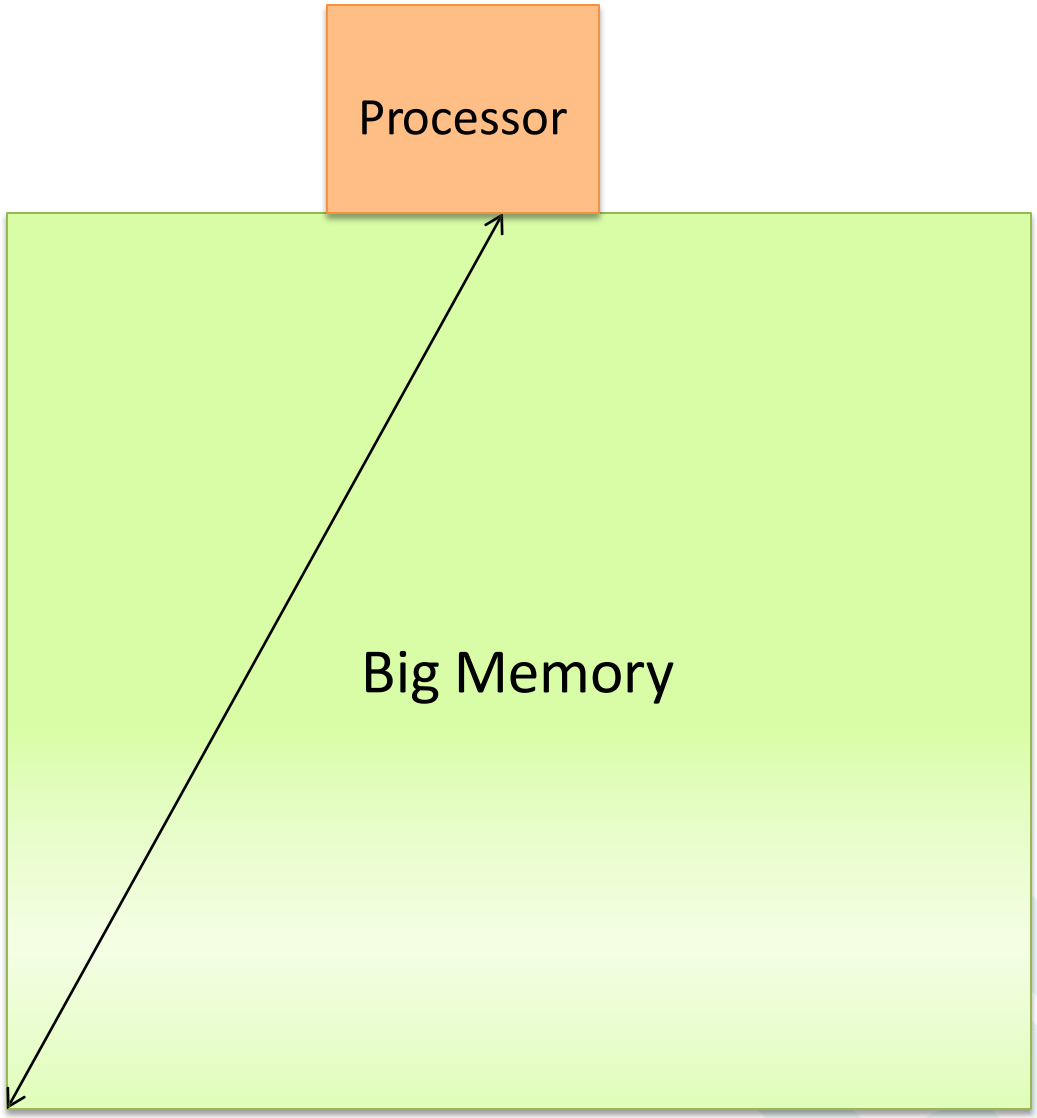


[Hennessy & Patterson 2011]

- Four-issue 2 GHz superscalar accessing 100 ns DRAM could execute 800 instructions during the time for one memory access!
- Long latencies mean large bandwidth-delay products which can be difficult to saturate, meaning bandwidth is wasted

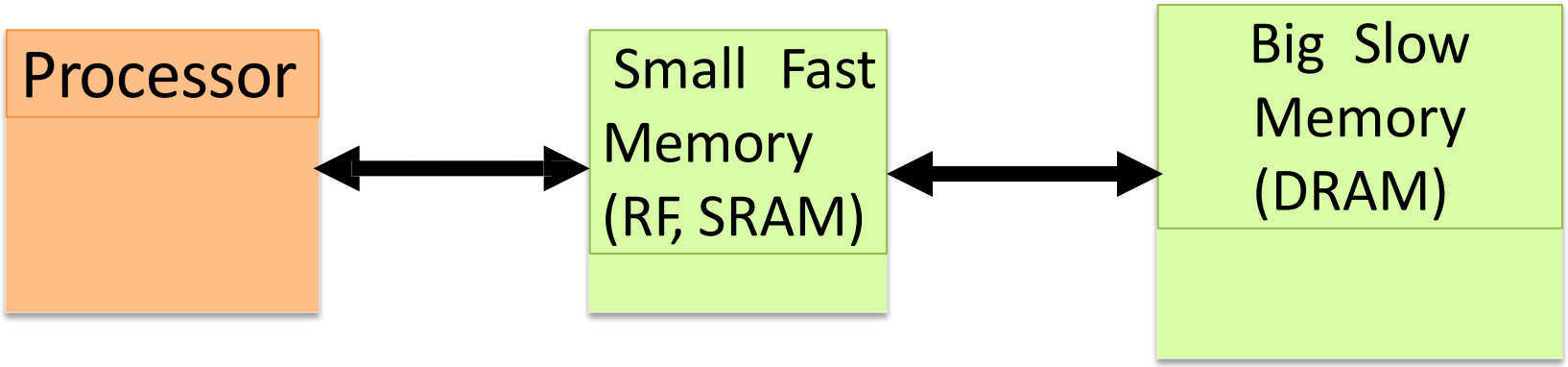
Physical Size Affects Latency





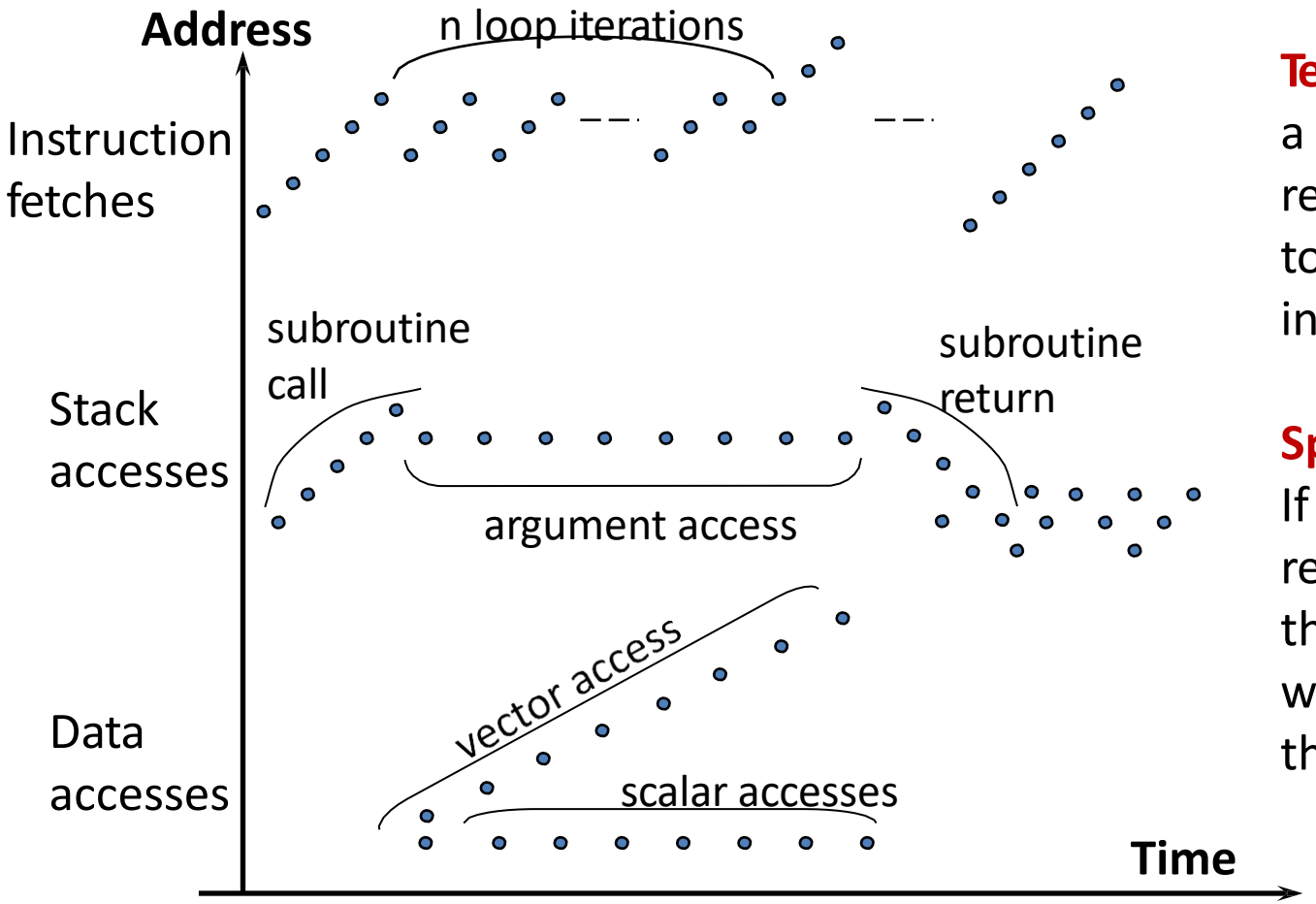
- Signals have further to travel
- Fan out to more locations

Memory Hierarchy



- Capacity: Register << SRAM << DRAM
- Latency: Register << SRAM << DRAM
- Bandwidth: on-chip >> off-chip
- On a data access:
  - if data is in fast memory -> low-latency access to SRAM
  - if data is not in fast memory -> long-latency access to DRAM
- Memory hierarchies only work if the small, fast memory actually stores data that is reused by the processor

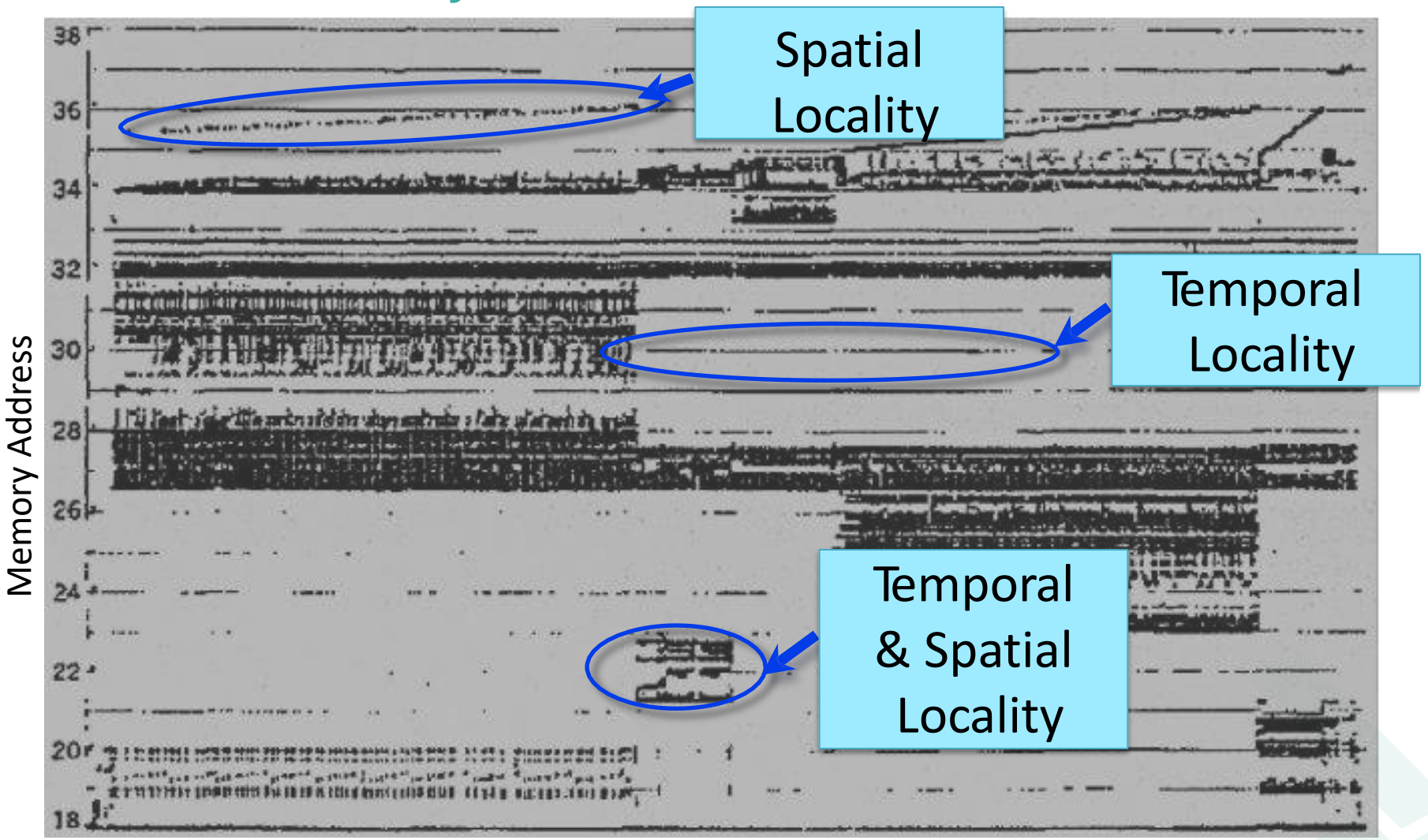
Common And Predictable Memory Reference Patterns



Temporal Locality: If a location is referenced it is likely to be referenced again in the near future

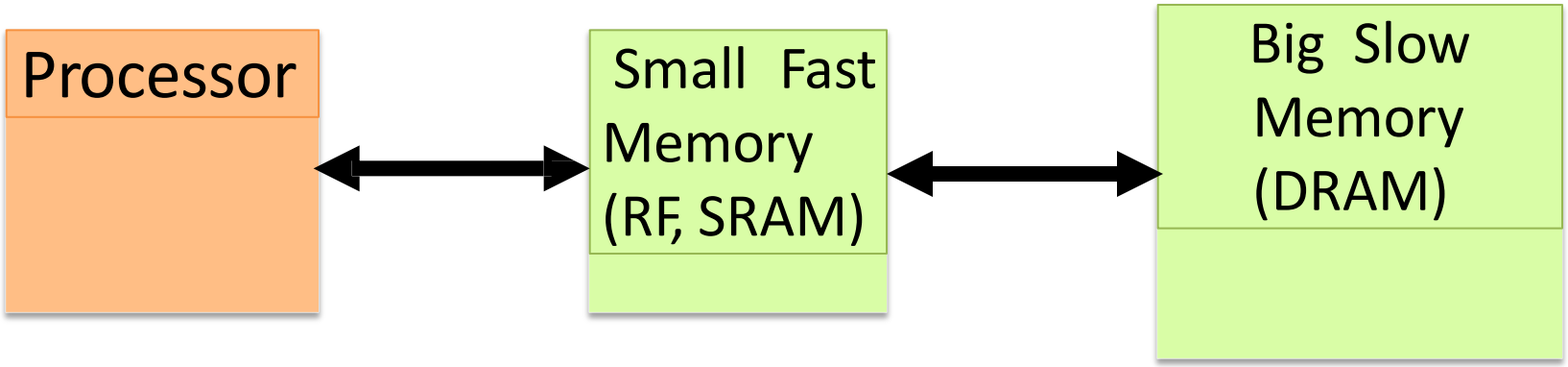
Spatial Locality: If a location is referenced it is likely that locations near it will be referenced in the near future

Real Memory Reference Patterns



[From Donald J. Hatfield, Jeanette Gerald: Program Restructuring for Virtual Memory. IBM Systems Journal 10(3): 168-192 (1971)]

Caches Exploit Both Types of Locality



- Exploit temporal locality by remembering the contents of recently accessed locations
- Exploit spatial locality by fetching blocks of data around recently accessed locations

Class Interaction # 24