Parallel computing platforms

Jeremy Iverson

College of Saint Benedict & Saint John's University

recap

- von Neumann architecture
  - central processing unit
  - memory
    - cache ($)
  - interconnection
- operating system
  - processes vs threads

cache performance

From Intel Performance Analysis Guide:

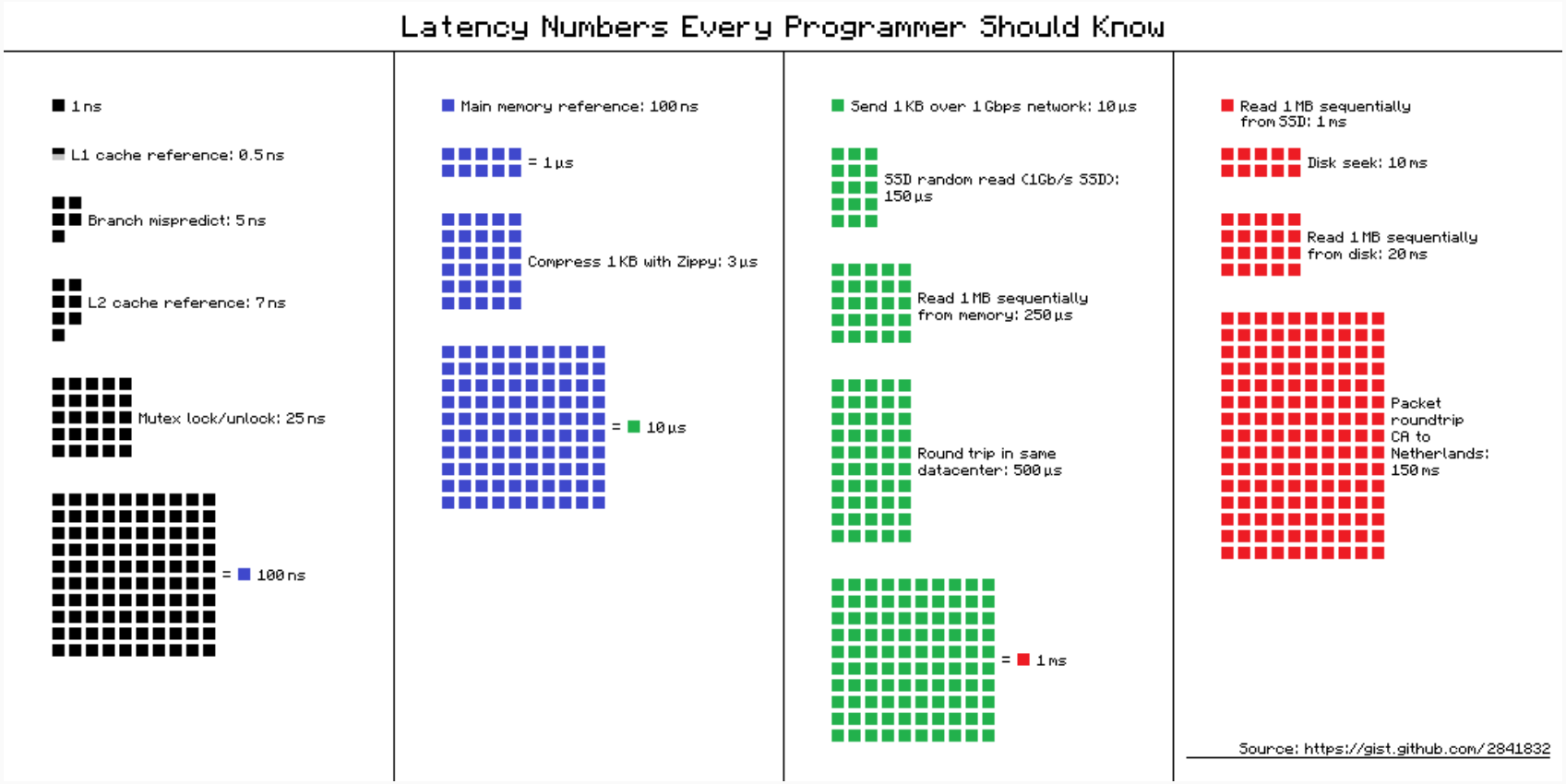
```
Core i7 Xeon 5500 Series Data Source Latency (approximate)              [Pg. 22]

local  L1 CACHE hit,                                ~4 cycles (   2.1 -  1.2 ns )
local  L2 CACHE hit,                               ~10 cycles (   5.3 -  3.0 ns )
local  L3 CACHE hit, line unshared                 ~40 cycles (  21.4 - 12.0 ns )
local  L3 CACHE hit, shared line in another core   ~65 cycles (  34.8 - 19.5 ns )
local  L3 CACHE hit, modified in another core      ~75 cycles (  40.2 - 22.5 ns )

remote L3 CACHE (Ref: Fig.1 [Pg. 5])         ~100-300 cycles ( 160.7 - 30.0 ns )

local  DRAM                                                    ~60 ns
remote DRAM                                                   ~100 ns
```

cache performance



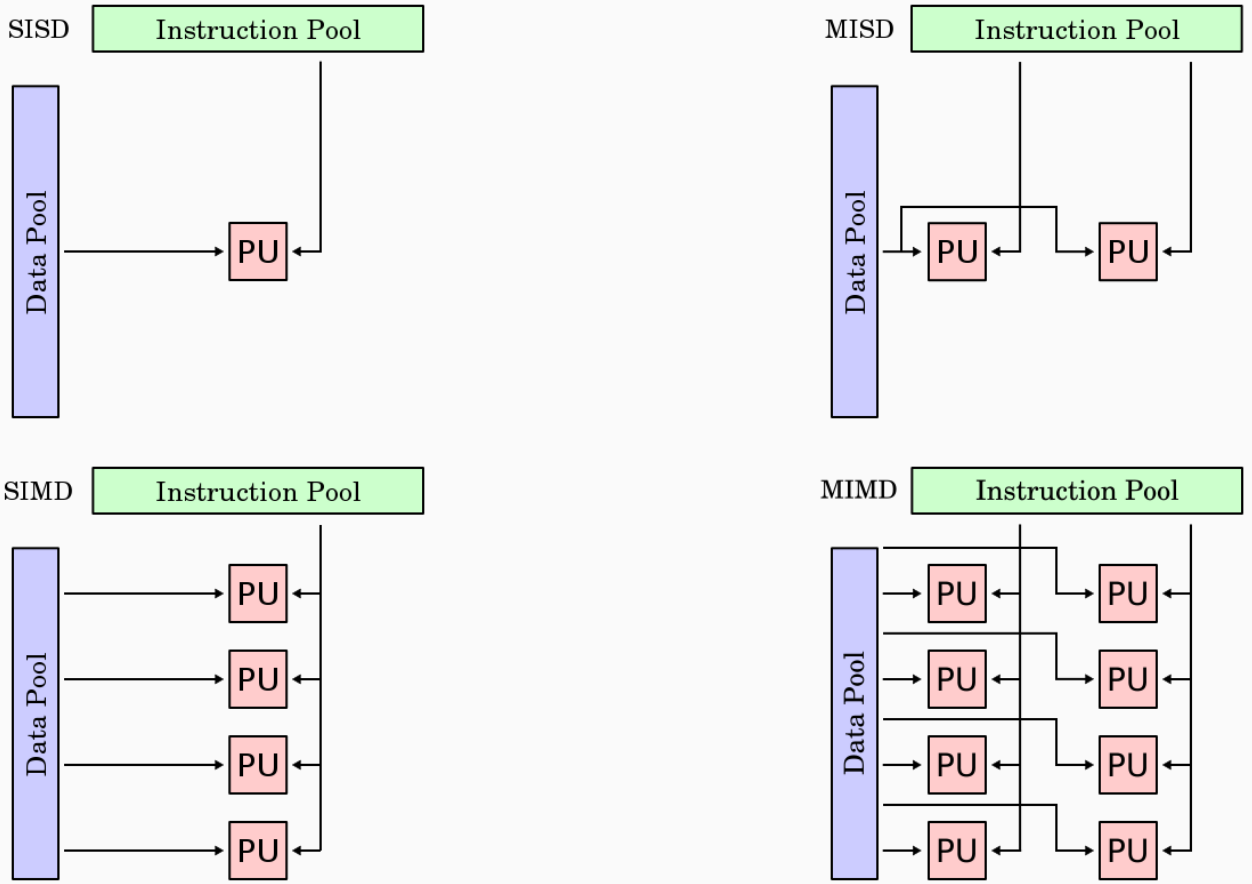
Latency Numbers Every Programmer Should Know

parallel computing platform

- logical organization
  - the user's view of the machine as it is being presented via its system software
- physical organization
  - the actual hardware architecture

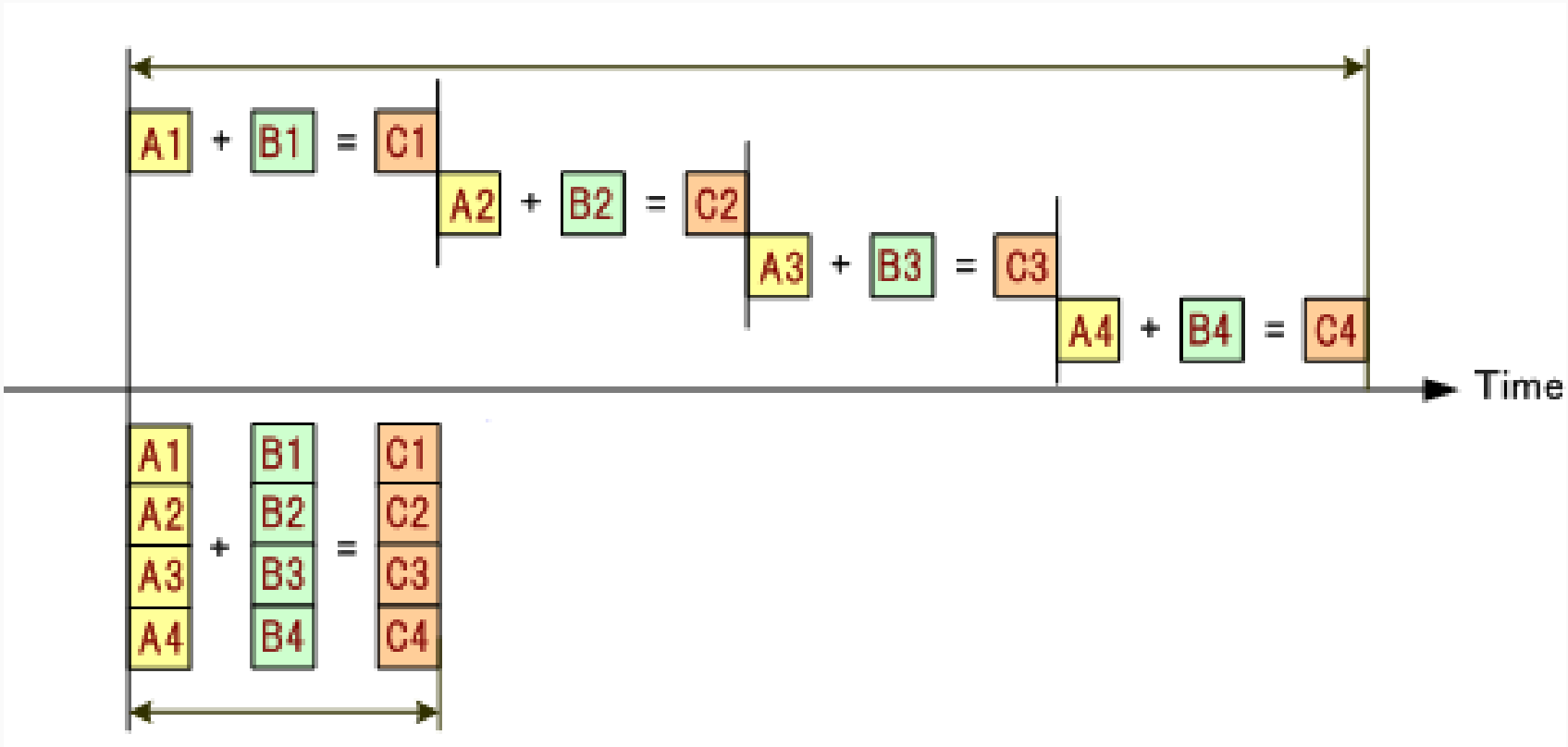
flynn's taxonomy

- based on the number of instruction streams and data streams available in the architecture



Flynn's taxonomy by Cburnett / CC BY 3.0 / presenting the four together

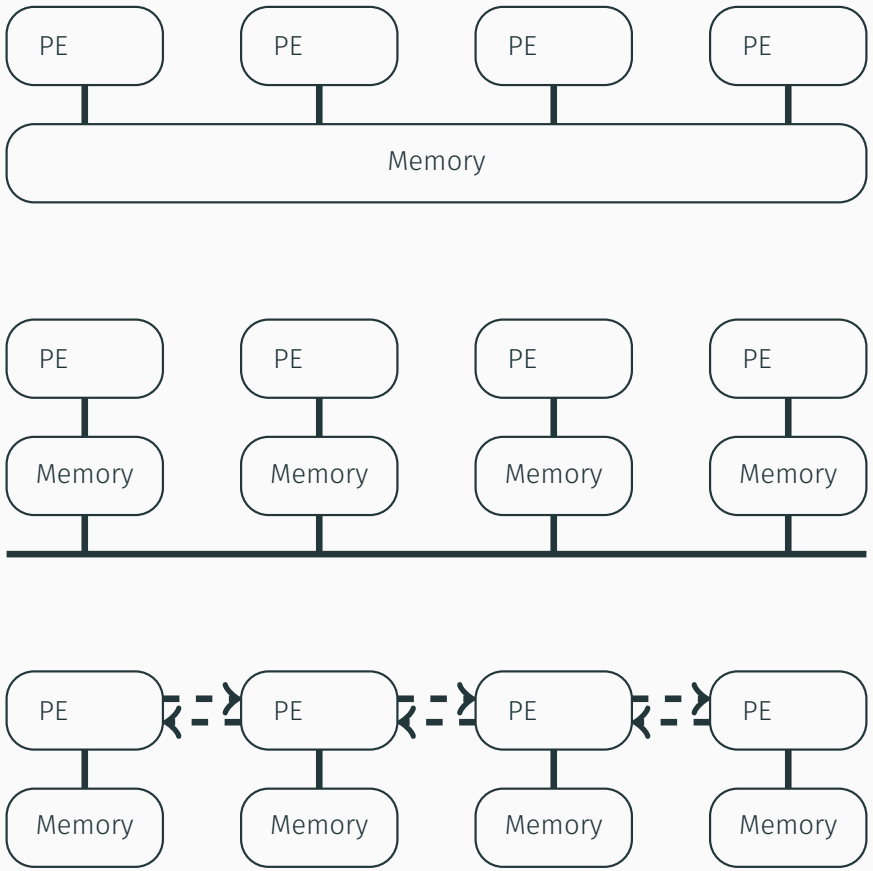
simd



SIMD / cropped from original

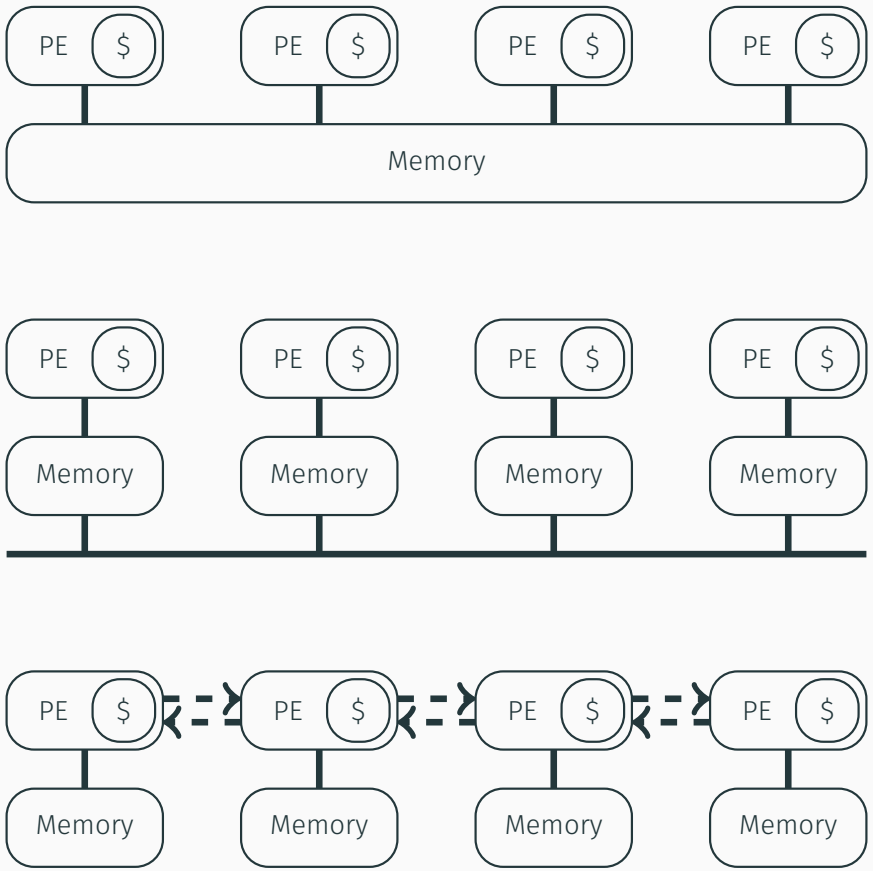
communication models

- shared-address space
  - UMA / NUMA / ccNUMA
- message-passing



communication models

- shared-address space
  - UMA / NUMA / ccNUMA
- message-passing



cache coherence

- update
  - increases communication on the bus
- invalidate
  - increases idling time

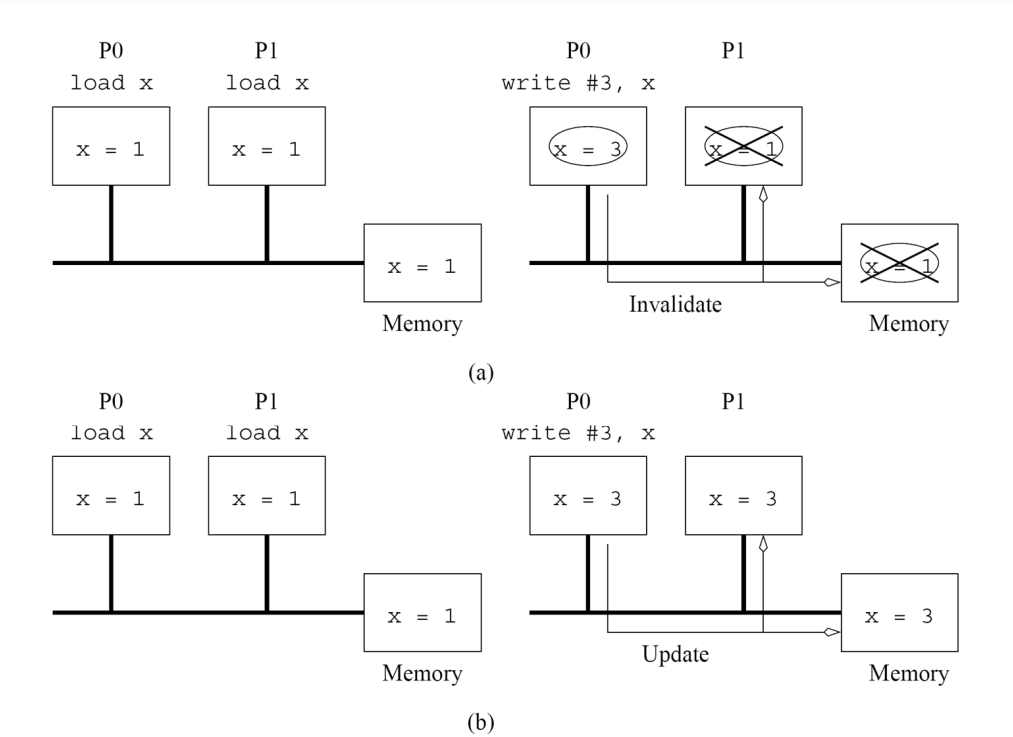


Figure 2.21 Cache coherence in multiprocessor systems: (a) Invalidate protocol; (b) Update protocol for shared variables.

false sharing