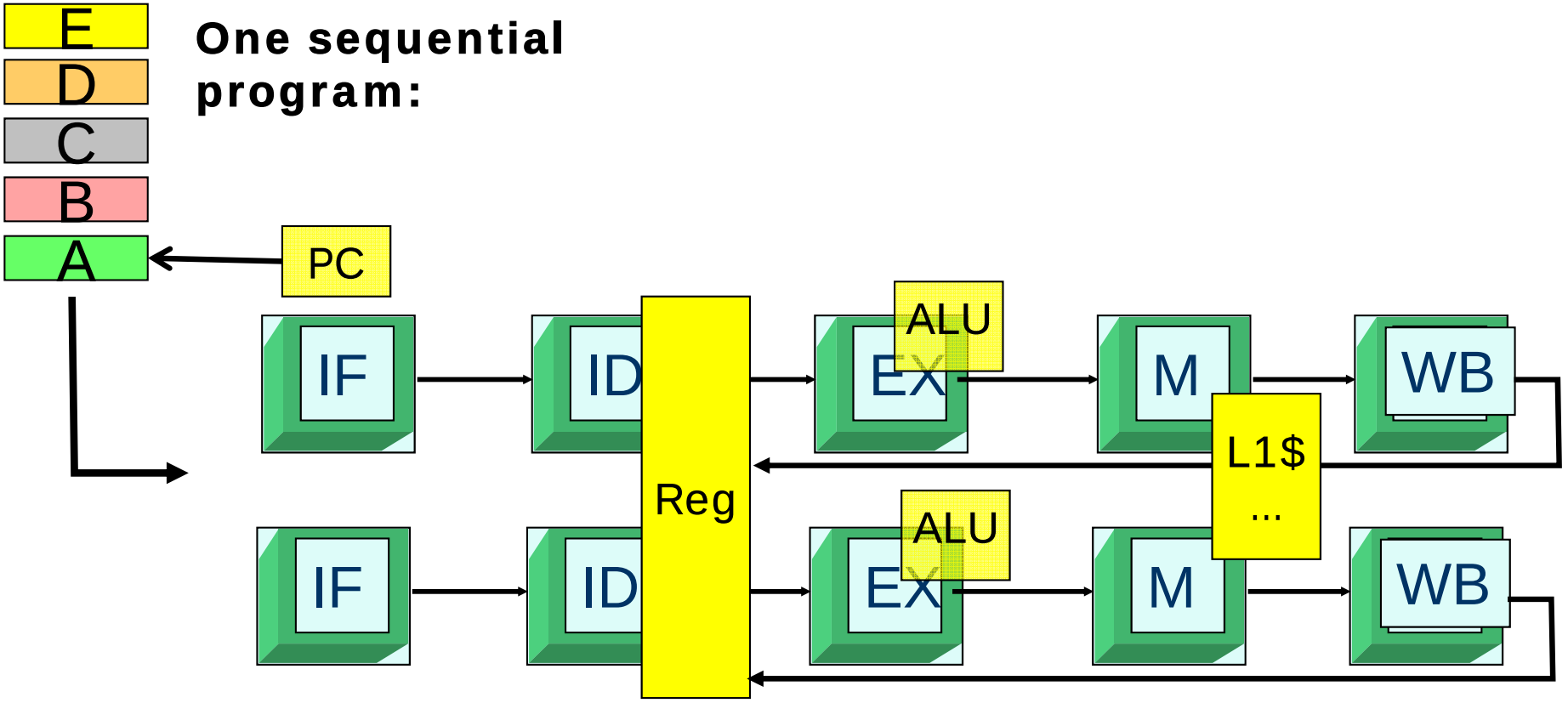
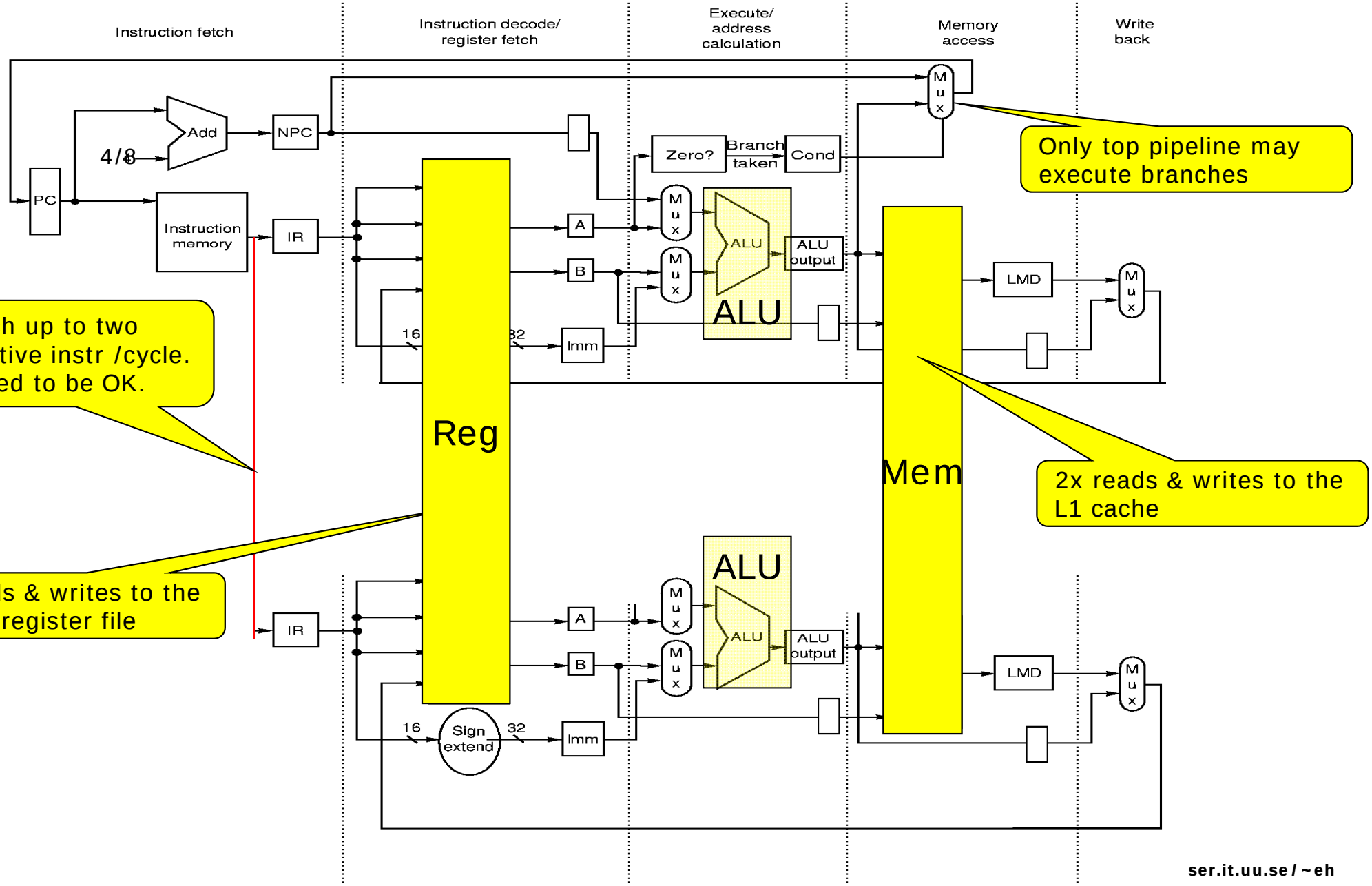
Threaded processors

Erik Hagersten
Uppsala University

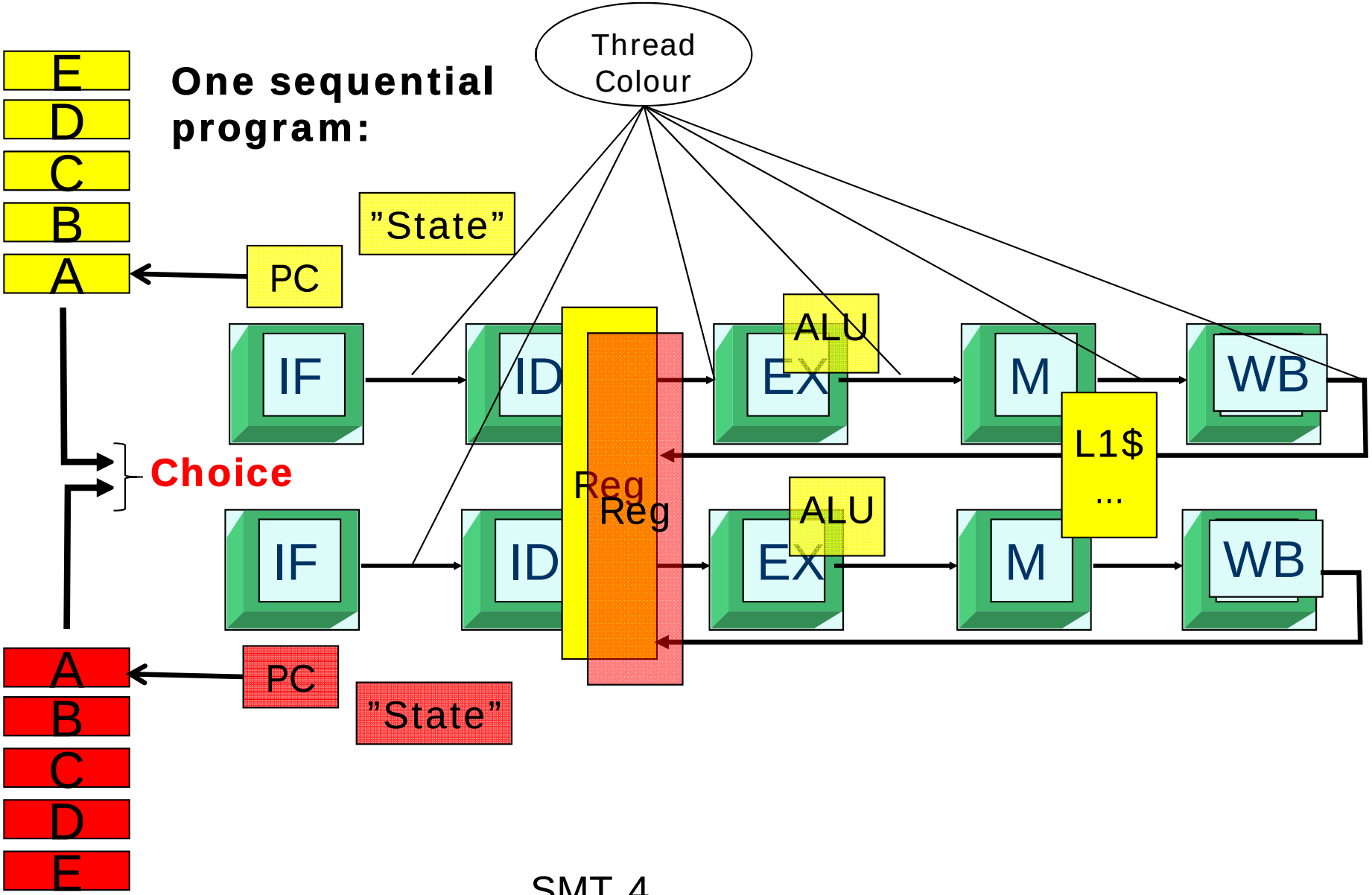
A 5-stage 2-way superscalar pipeline



A 2-way superscalar 5-stage pipeline Need 2x ILP!



A 5-stage 2-way superscalar pipeline, Multithreaded 2-ways



UPPSALA
UNIVERSITET

What resources are replicated (for each thread) in a typical multithreaded architecture?

- ❑ Registers
- ❑ Cache
- ❑ PC
- ❑ ALU

AVDARK
2013

UPPSALA UNIVERSITET

Choosing between different threads

- Fixed interleaving (Xeon Phi, HEP, ...)
  - Each of N threads executes one instruction every N:th cycles
  - If thread is not ready to go during its slot → bubble
- Hardware-controlled thread scheduling
  - E.g., hardware keeps track of which threads are ready to go (Niagra-1)
  - E.g., picks next thread to execute based on hardware priority scheme (~Hyperthreading)
  - I-count: Chose the thread with least Instr in-flight
  - Course-grained: Run one thread until it ”blocks”

UPPSALA UNIVERSITET

HyperThreading (Pentium 4) Simultaneous Multithreading (SMT)

- First commercial SMT design (2-way SMT)
- Shared resources: Caches, execution units, branch predictors, …
- Area overhead due to hyperthreading ~ 5%
- 1.01 (INT), 1.07 (FP) SPECrate performance boost ☹
- 1.20 avg. improvement for SPEC mixes
- Not a great hit. Revived for Nehalem 2008.
- Intel Atom (in-order x86 core) has two-way SMT multithreading

AVDARK 2013

UPPSALA
UNIVERSITET

Fixed interleaving scheduling of instructions between N threads:

- Runs one thread until it ”blocks” and then switches
- Lets the threads take turns and run every N:th cycle
- Selects threads that are “ready” to run

AVDARK
2013