Advanced Processor Architecture

Jinkyu Jeong (jinkyu@skku.edu)
Computer Systems Laboratory
Sungkyunkwan University
http://csl.skku.edu

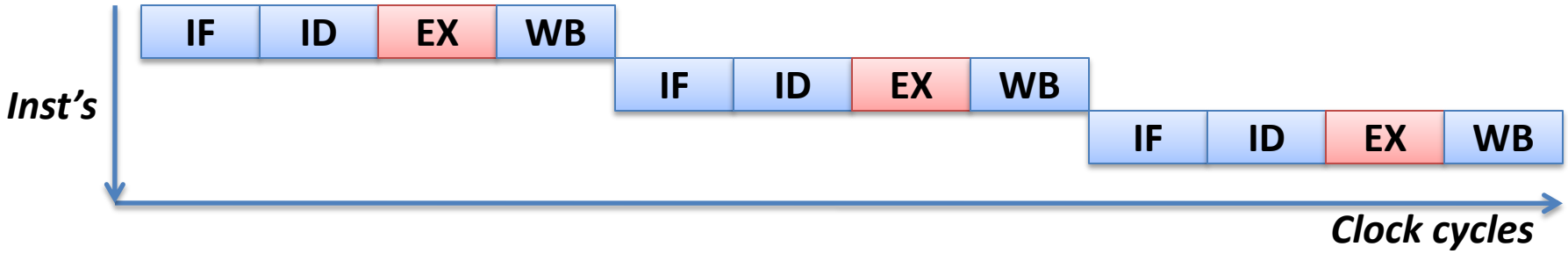
Modern Microprocessors

- More than just GHz

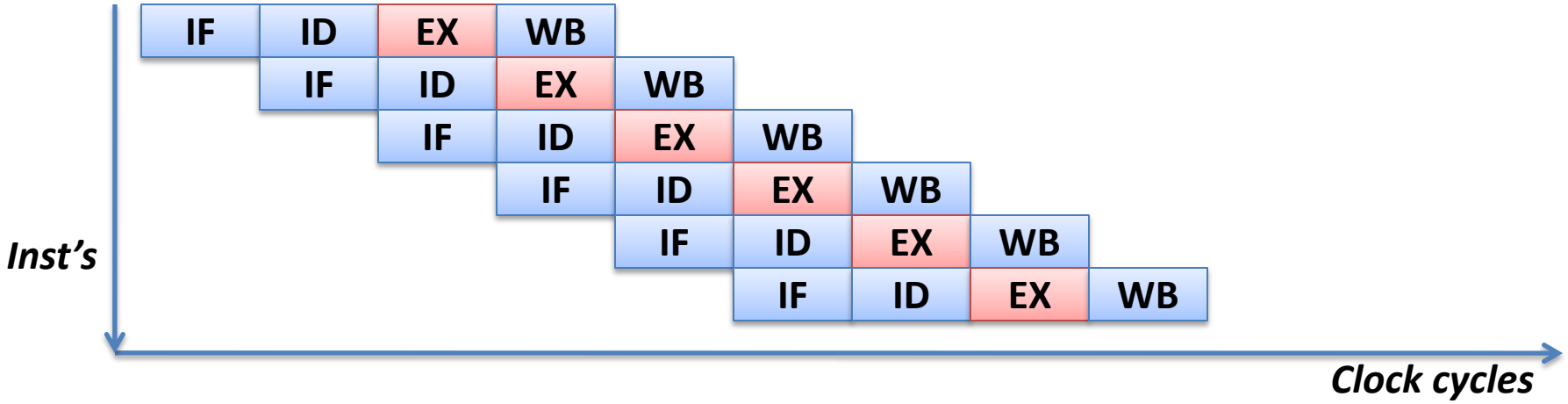
CPU	Clock Speed	SPECint2000	SPECfp2000
Athlon 64 FX-55	2.6GHz	1854	1782
Pentium 4 Extreme Edition	3.46GHz	1772	1724
Pentium 4 Prescott	3.8GHz	1671	1842
Opteron 150	2.4GHz	1655	1644
Itanium 2 9MB	1.6GHz	1590	2712
Pentium M 755	2.0GHz	1541	1088
POWER5	1.9GHz	1452	2702
SPARC64 V	1.89GHz	1345	1803
Athlon 64 3200+	2.2GHz	1080	1250
Alpha 21264C	1.25GHz	928	1019

Pipelining

- Sequential execution

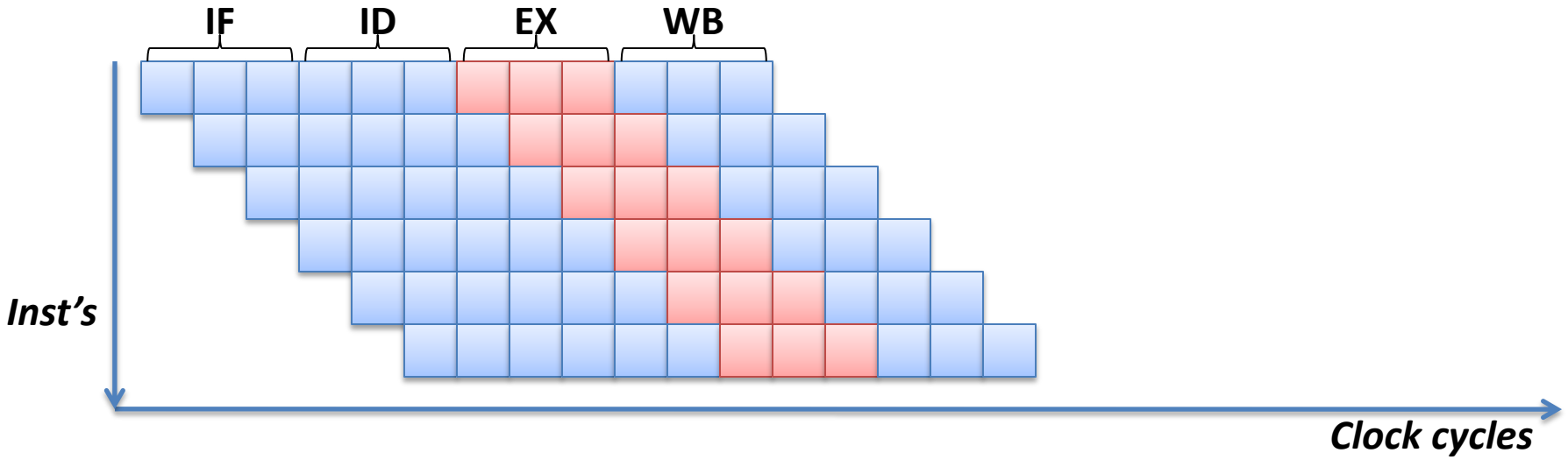


- Pipelining (RISC)



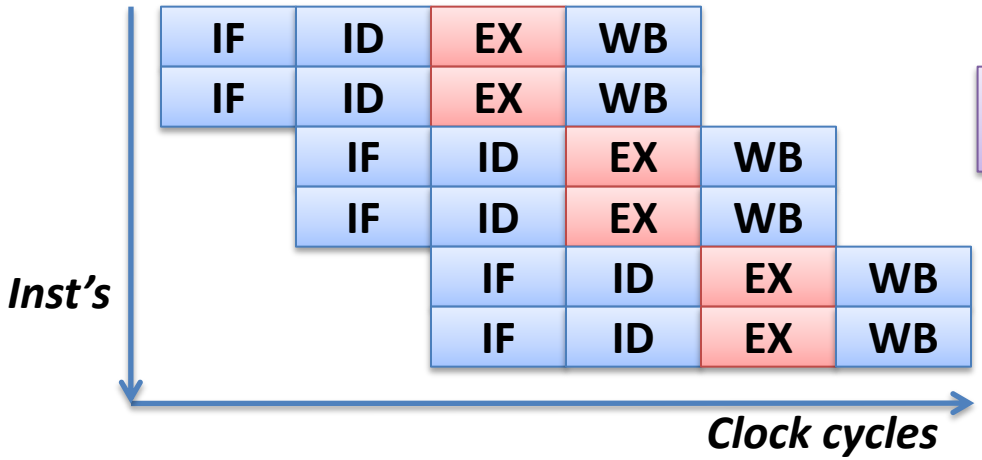
Superpipelining

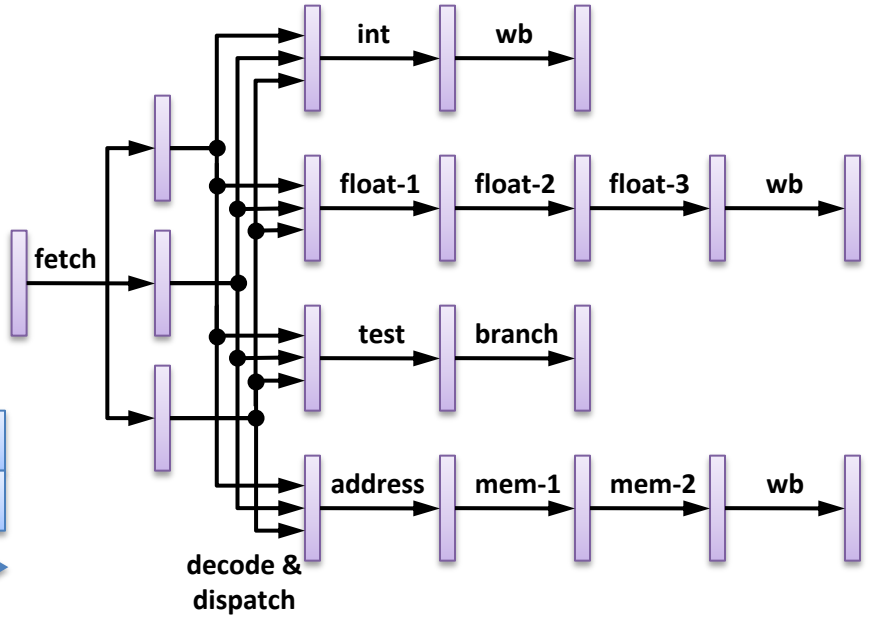
- Superpipelining
  - Subdivide each pipeline stage
  - Higher clock speed
  - 10-15 in Athlon, 12+ in Pentium Pro/II/III, 14 in UltraSparc-III, 16-25 in PowerPC G5, 20+ in Pentium 4



Superscalar

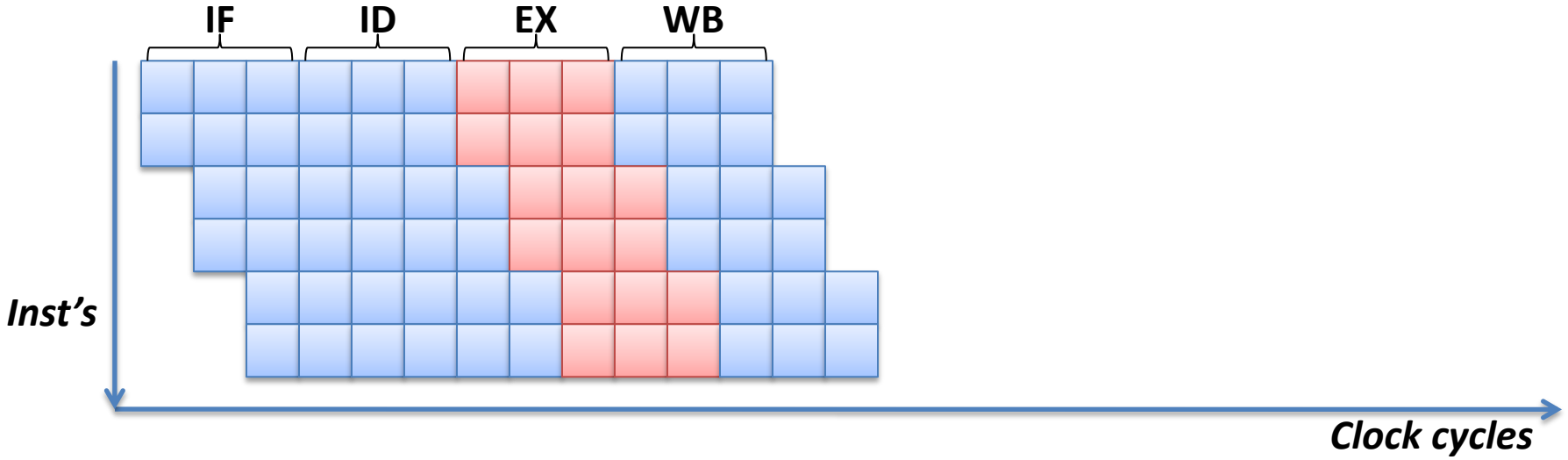
- Superscalar
  - The execution stage has a bunch of different functional units
  - Execute multiple instructions in parallel
  - Pentium: 2-way superscalar





Superpipelined Superscalar

- Superpipelining + Superscalar
  - 2-way: MIPS R5000
  - 3-way: PowerPC G3/G4, Pentium Pro/II/III/M/4, Athlon
  - 4-way: UltraSparc, MIPS R10000, PowerPC G4e, Alpha 21164 & 21264, Core 2 Duo
  - 5-issue: PowerPC G5

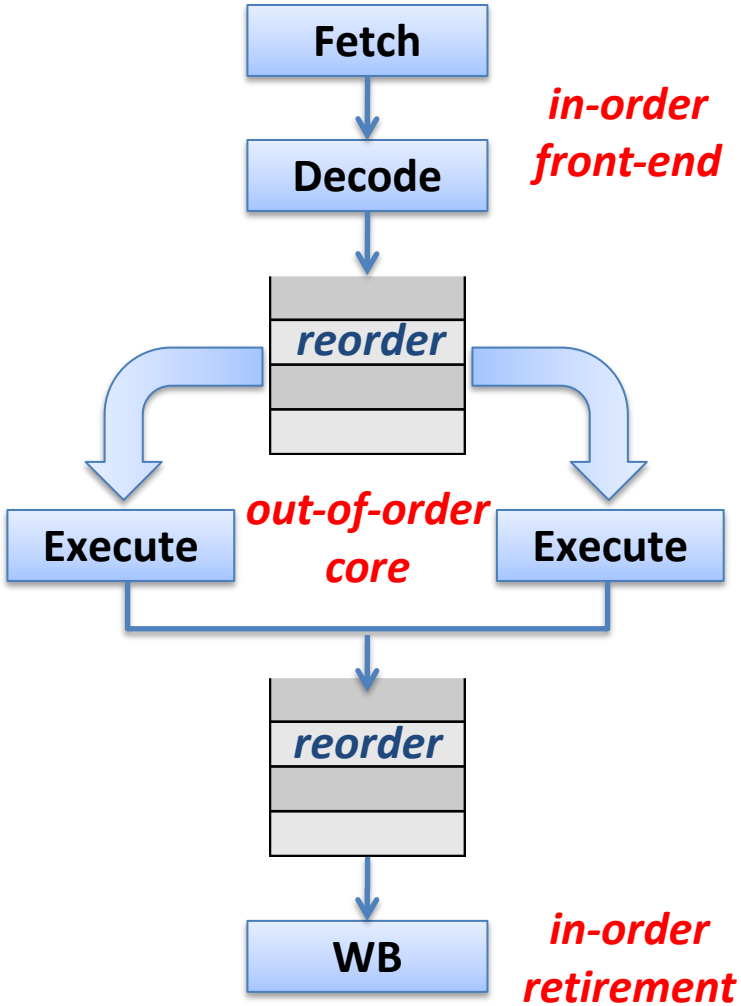


Tackling Instruction Dependencies

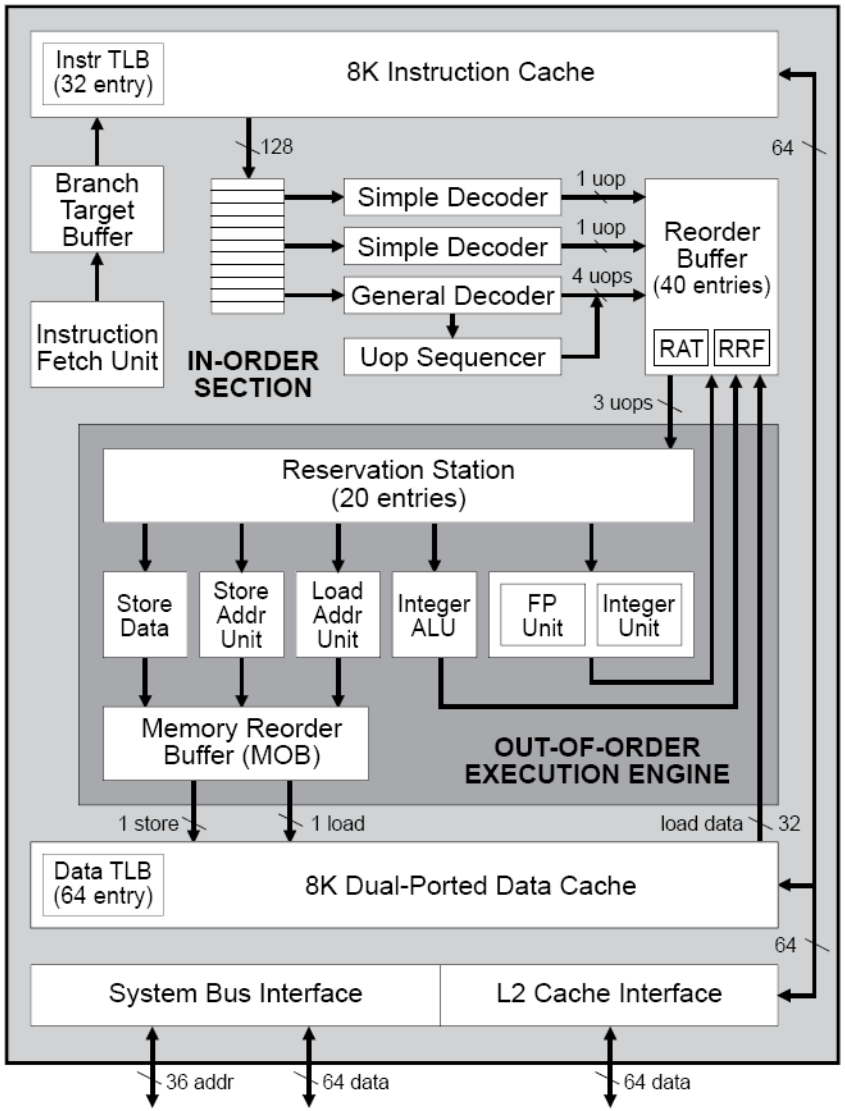
- Branch prediction + speculative execution
  - Mispredict penalty: 10 – 15 cycles in Pentium Pro/II/III
- Instruction scheduling
  - In-order execution + compiler optimization
    - Rearrange the instructions at compile time
    - Compiler can see further down the program than the hardware
    - SuperSparc, HyperSparc, UltraSparc, Alpha 21064 & 21164
  - Out-of-order execution
    - Reorder instruction execution sequence in hardware at run time
    - Register renaming reduces the dependency further
    - MIPS R10000, Alpha 21264, POWER/PowerPC, Pentium Pro, Pentium 4, Core 2 Duo, Core i7, …

Intel Pentium Pro

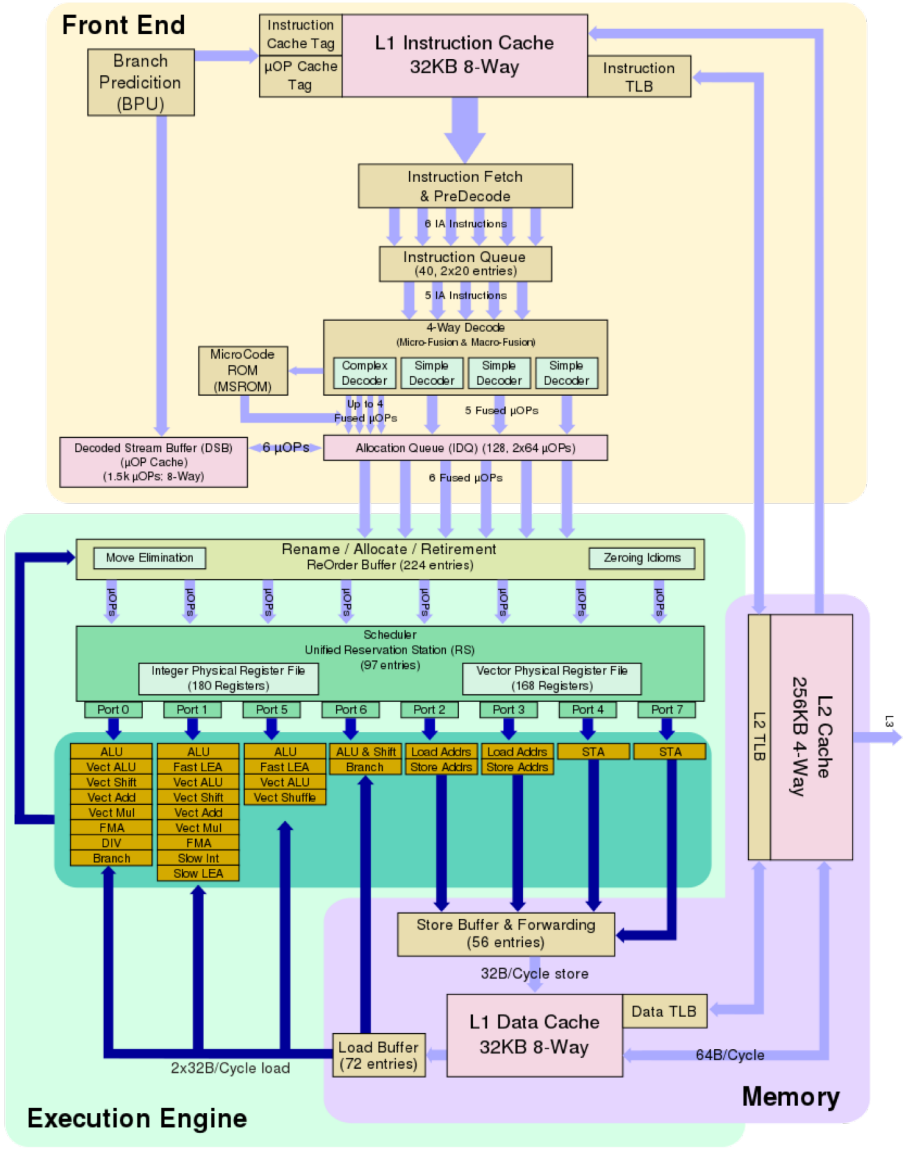
- In-order front-end
  - Multiple branch prediction
  - Micro-operations
  - Register renaming
- Out-of-order execution core
  - 3-way superscalar
  - Multiple execution units
  - Dataflow analysis
  - Speculative execution
- In-order retirement
  - Precise faulting semantics



P6 Microarchitecture

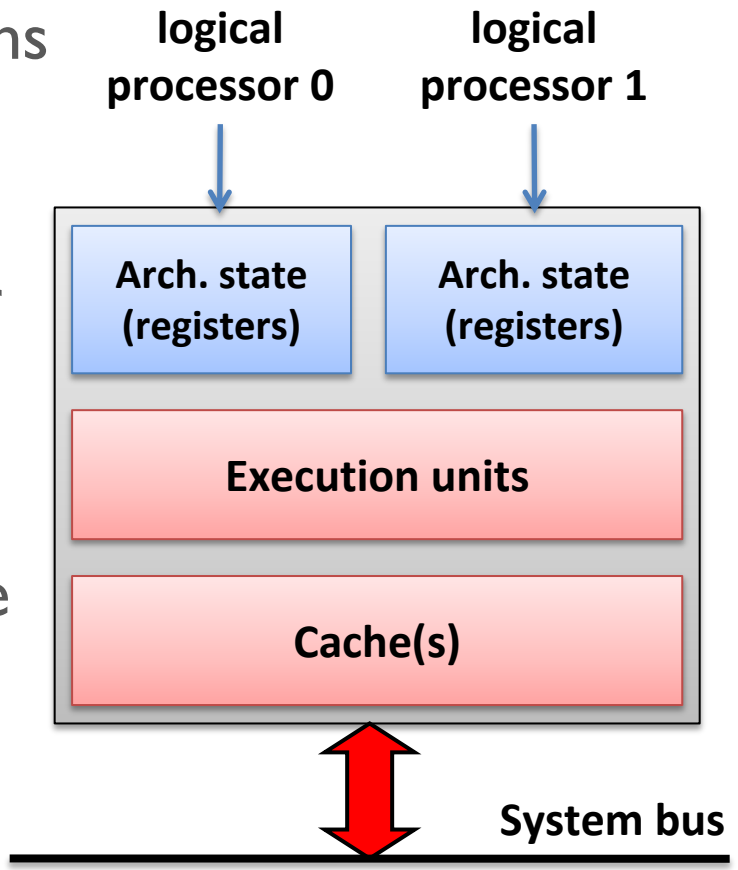


Skylake Microarchitecture



Hyper-Threading

- Simultaneous multithreading technology (SMT)
  - Utilizes thread-level parallelism
  - Fill pipelines with the instructions from multiple threads running at the same time
  - An SMT processor appears as if it were multiple independent processors
  - Uses processor resources more effectively
  - Cost: <5% in added die area

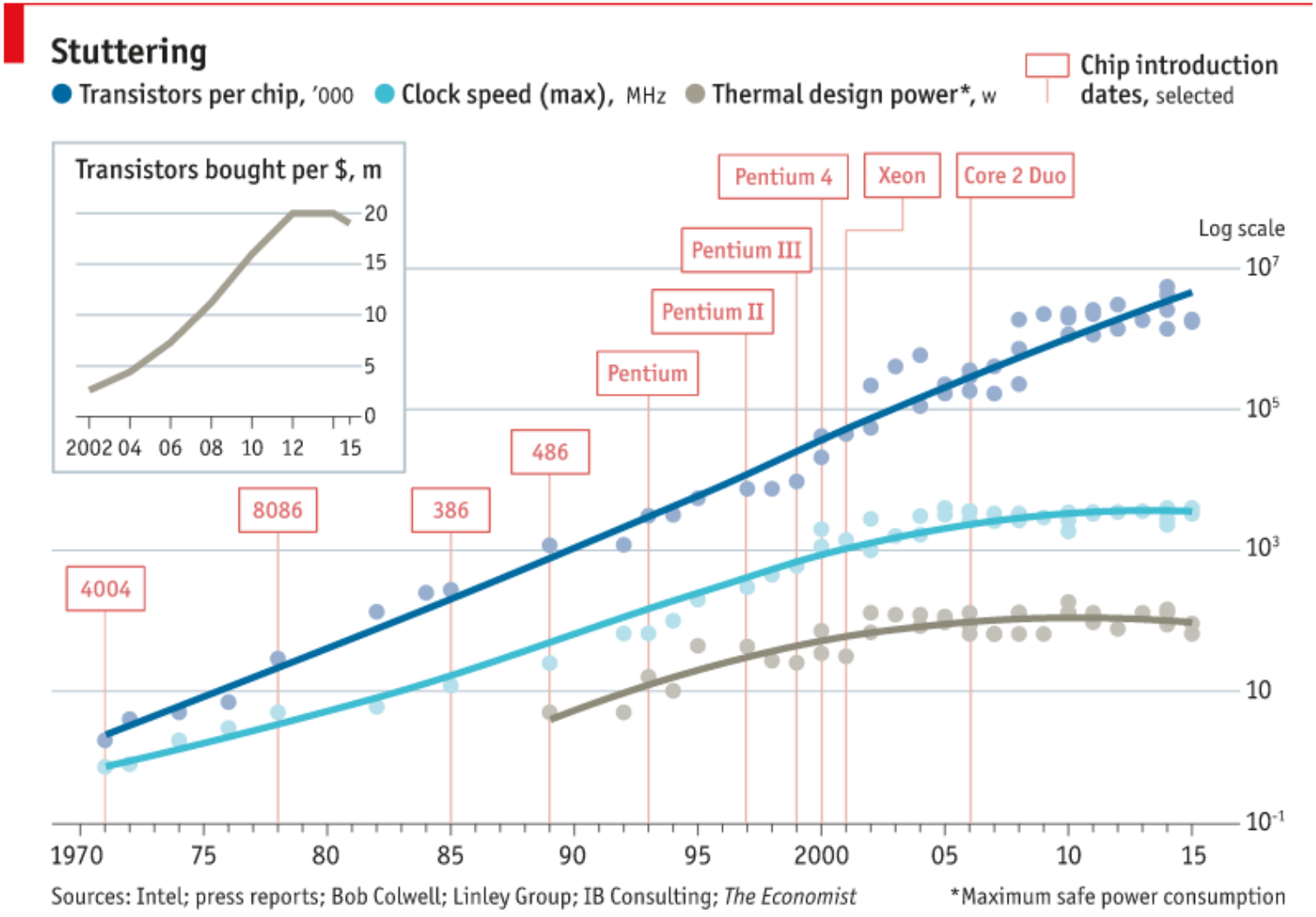


Multi-core

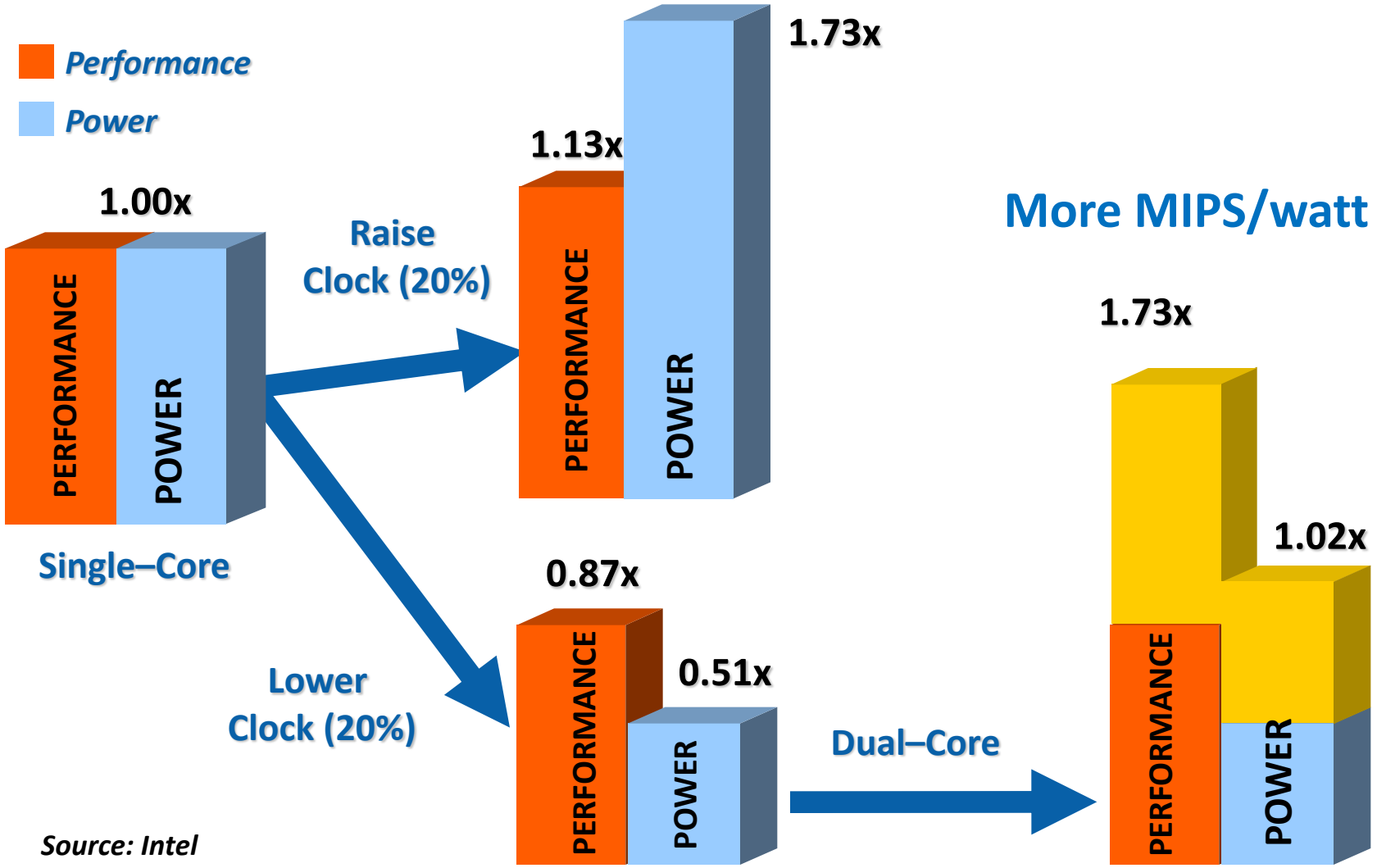
- Put two or more processor cores onto a single chip
  - Previously called CMP (Chip Multiprocessor)

- Examples
  - AMD Opteron: dual-core (Apr. 2005)
  - AMD dual-core Athlon 64 X2: dual-core (May 2005)
  - Intel Core Duo, Core 2 Duo: dual-core
  - Sun UltraSparc T1: eight-core, 32 threads (Nov. 2005)
  - Intel Xeon X7460: six-core (Sep. 2008)
  - Intel Xeon E7-8890 v4: 24-core (Jun. 2016)

CPU Trends


Stuttering
Transistors per chip, '000
Clock speed (max), MHz
Thermal design power*, w
Chip introduction dates, selected
Transistors bought per $, m
20
15
10
5
0
2002 04 06 08 10 12 15
Pentium 4
Xeon
Core 2 Duo
Pentium III
Pentium II
Pentium
486
8086
386
4004
Log scale
10^7
10^5
10^3
10
10^{-1}
1970 75 80 85 90 95 2000 05 10 15
Sources: Intel; press reports; Bob Colwell; Linley Group; IB Consulting; The Economist
*Maximum safe power consumption

Why Multi-core?

- Memory wall
  - CPU 55%/year, Memory 10%/year (1986 – 2000)
  - Caches show diminishing returns
- ILP(Instruction Level Parallelism) wall
  - Control dependency
  - Data dependency
- Power wall
  - Dynamic power $\propto$ Frequency^3
  - Static power $\propto$ Frequency
  - Total power $\propto$ The number of cores

Single-core vs. Multi-core