

Microprocessor Microarchitecture

The Past, Present, and Future of CPU Architecture



Lynn Choi

School of Electrical Engineering



高麗大學校

Computer System Laboratory

Contents



❑ Performance of Microprocessors

❑ Past: ILP Saturation

- I. Superscalar Hardware Complexity
- II. Limits of ILP
- III. Power Inefficiency

❑ Present: TLP Era

- I. Multithreading
- II. Multicore

❑ Present: Today's Microprocessor

- Intel Core 2 Quad, Sun Niagara II, and ARM Cortex A-9 MPCore

❑ Future: Looking into the Future

- I. Manycores
- II. Multiple Systems on Chip
- III. Trend – Change of Wisdoms

CPU Performance



□ T_{exe} (Execution time per program)

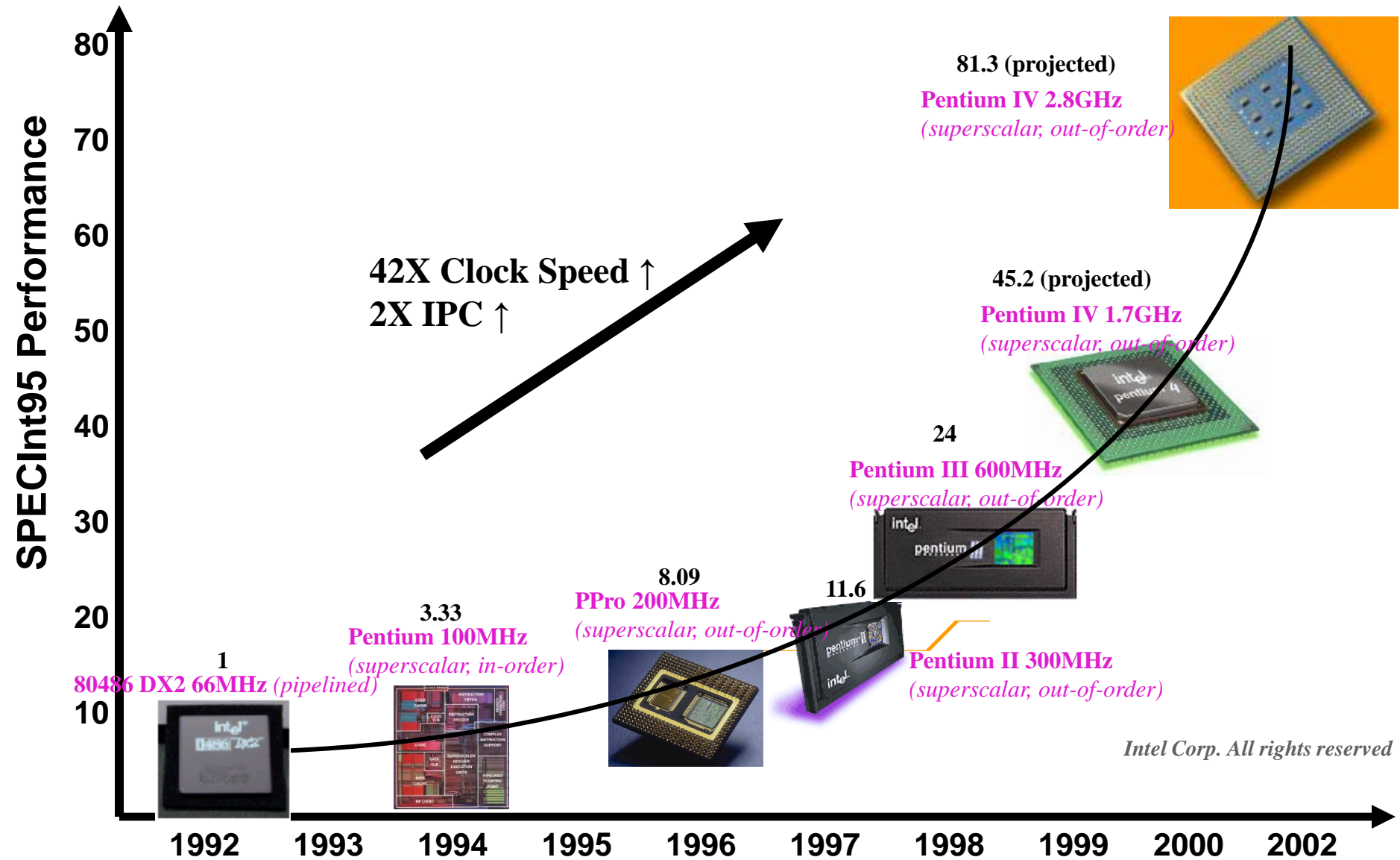
$$= NI * CPI_{\text{execution}} * T_{\text{cycle}}$$

- NI = # of instructions / program (program size)
- CPI = clock cycles / instruction
- T_{cycle} = second / clock cycle (clock cycle time)

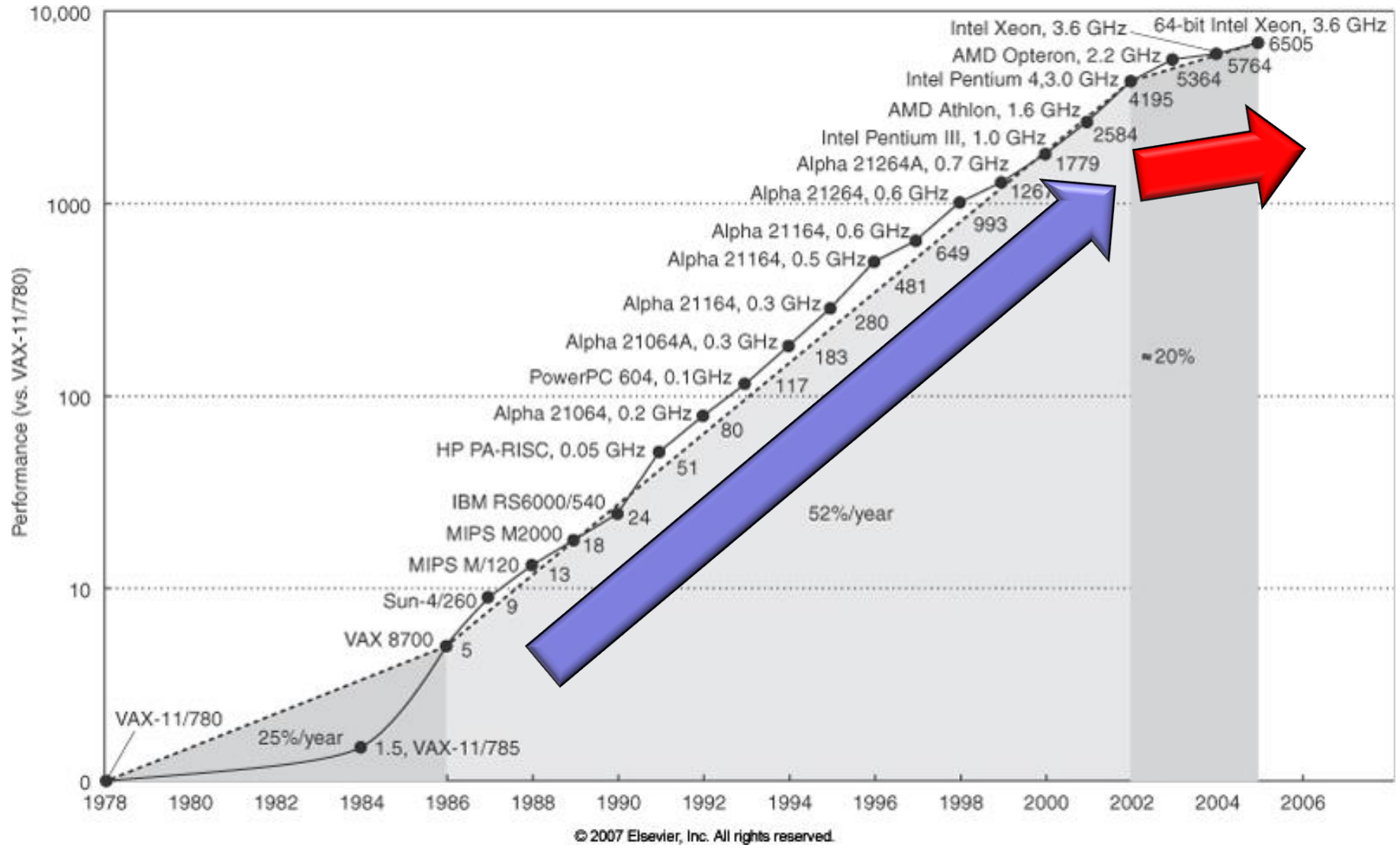
□ *To increase performance*

- Decrease NI (*or program size*)
 - Instruction set architecture (CISC vs. RISC), compilers
- Decrease CPI (*or increase IPC*)
 - Instruction-level parallelism (Superscalar, VLIW)
- Decrease T_{cycle} (*or increase clock speed*)
 - Pipelining, process technology

Advances in Intel Microprocessors



Microprocessor Performance Curve



ILP Saturation I – Hardware Complexity

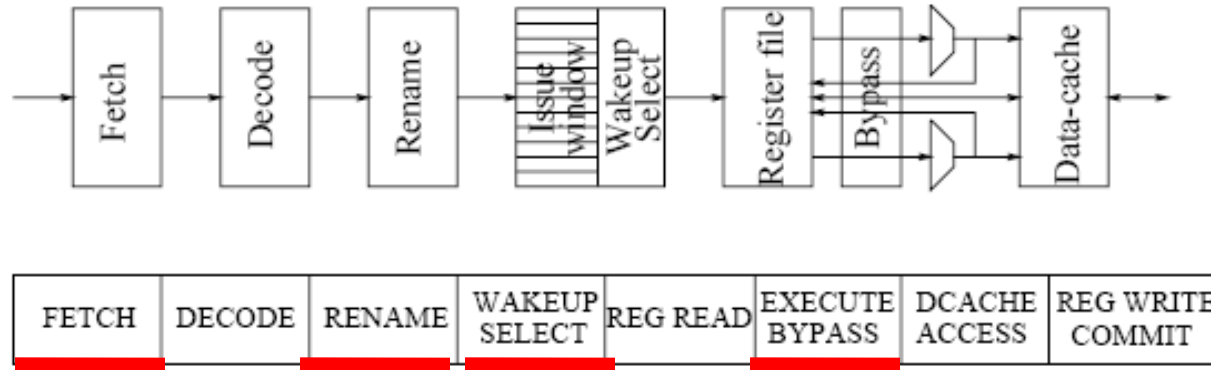


Figure 1: Baseline superscalar model.

IEEE All rights reserved

❑ Superscalar hardware is not scalable in terms of issue width!

- *Limited instruction fetch bandwidth*
- *Renaming complexity \propto issue width²*
- *Wakeup & selection logic \propto instruction window²*
- *Bypass logic complexity \propto # of FUs²*
- *Also, on-chip wire delays, # register and memory access ports, etc.*

❑ Higher IPC implies lowering the Clock Speed!

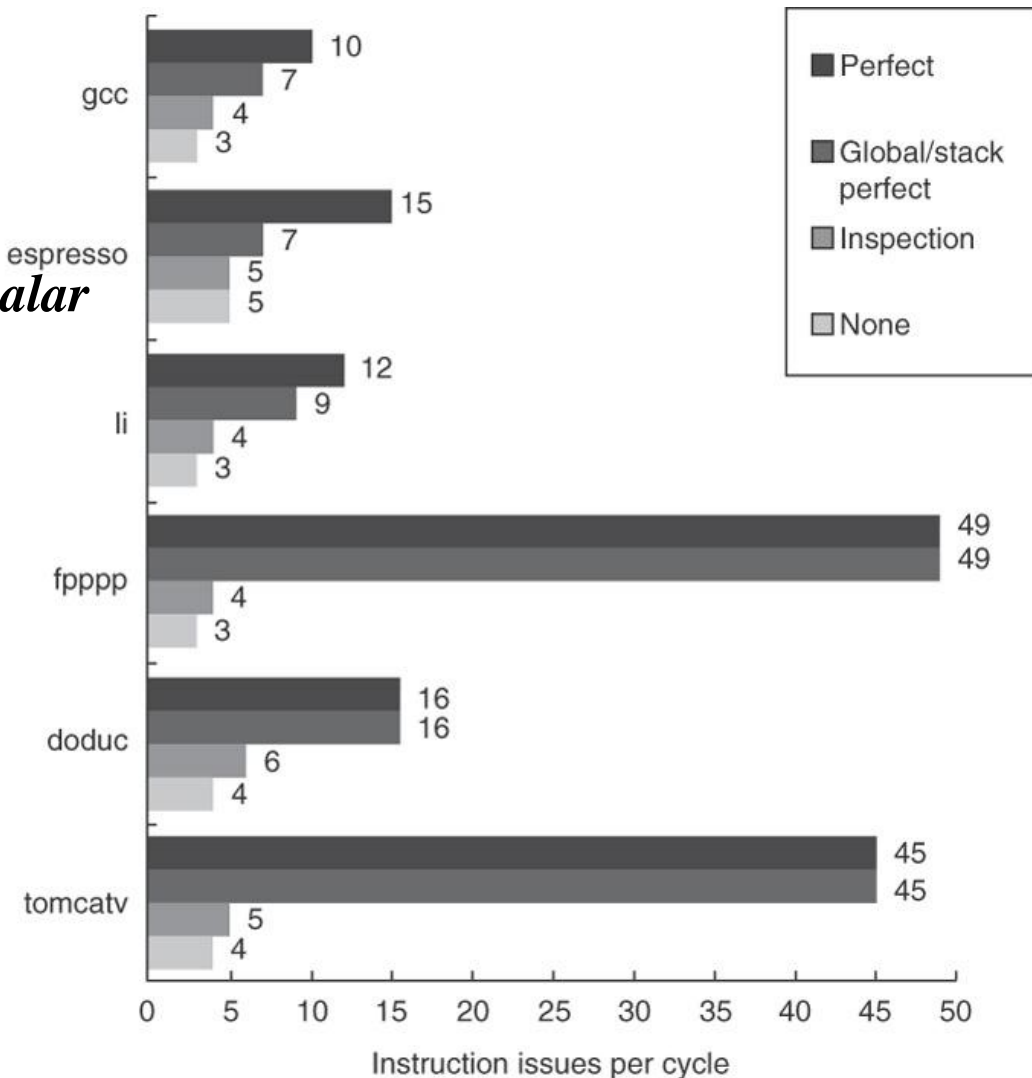
ILP Saturation II – Limits of ILP



Even with a very aggressive superscalar microarchitecture

- ✓ 2K window
- ✓ Max. 64 instruction issues per cycle
- ✓ 8K entry tournament predictors
- ✓ 2K jump and return predictors
- ✓ 256 integer and 256 FP registers

Available ILP is only 3 ~ 6!

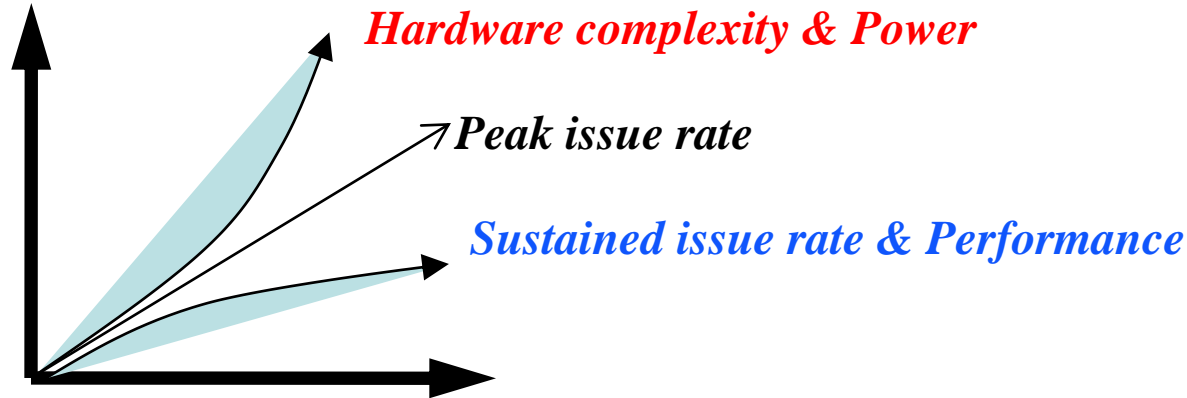


© 2007 Elsevier, Inc. All rights reserved.

ILP Saturation III – Power Inefficiency



- ❑ Increasing issue rate is not energy efficient



- ❑ Increasing clock rate is also not energy efficient
 - Increasing clock rate will increase transistor switching frequency
 - Faster clock needs deeper pipeline, but the pipelining overhead grows faster
- ❑ Existing processors already reach the power limit
 - 1.6GHz Itanium 2 consumes 130W of power!
 - **Temperature problem** –Pentium power density passes that of a hot plate ('98) and would pass a nuclear reactor in 2005, and a rocket nozzle in 2010.
- ❑ Higher IPC and higher clock speed have been pushed to their limit!

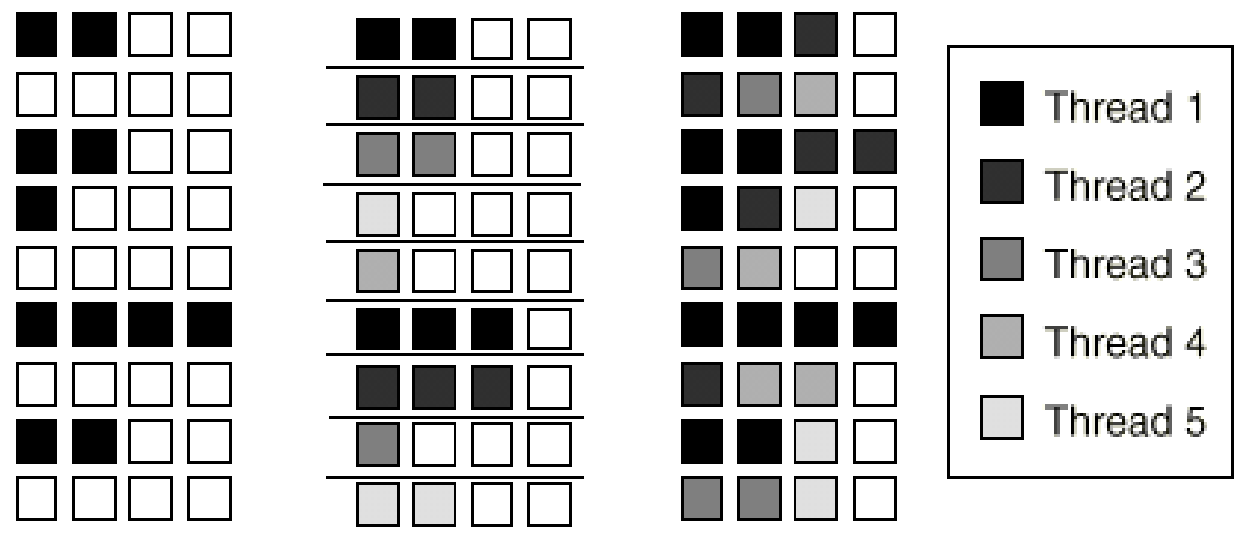
TLP Era I - Multithreading



□ *Multithreading*

- Interleave multiple independent threads into the pipeline every cycle
 - Each thread has its own PC, RF, branch prediction structures but shares instruction pipelines and backend execution units
- Increase resource utilization & throughput for multiple-issue processors
 - Improve total system throughput (IPC) at the expense of compromised single program performance

Time (processor cycles) ↓



(a) Superscalar (b) Fine-Grain Multithreading (c) SMT

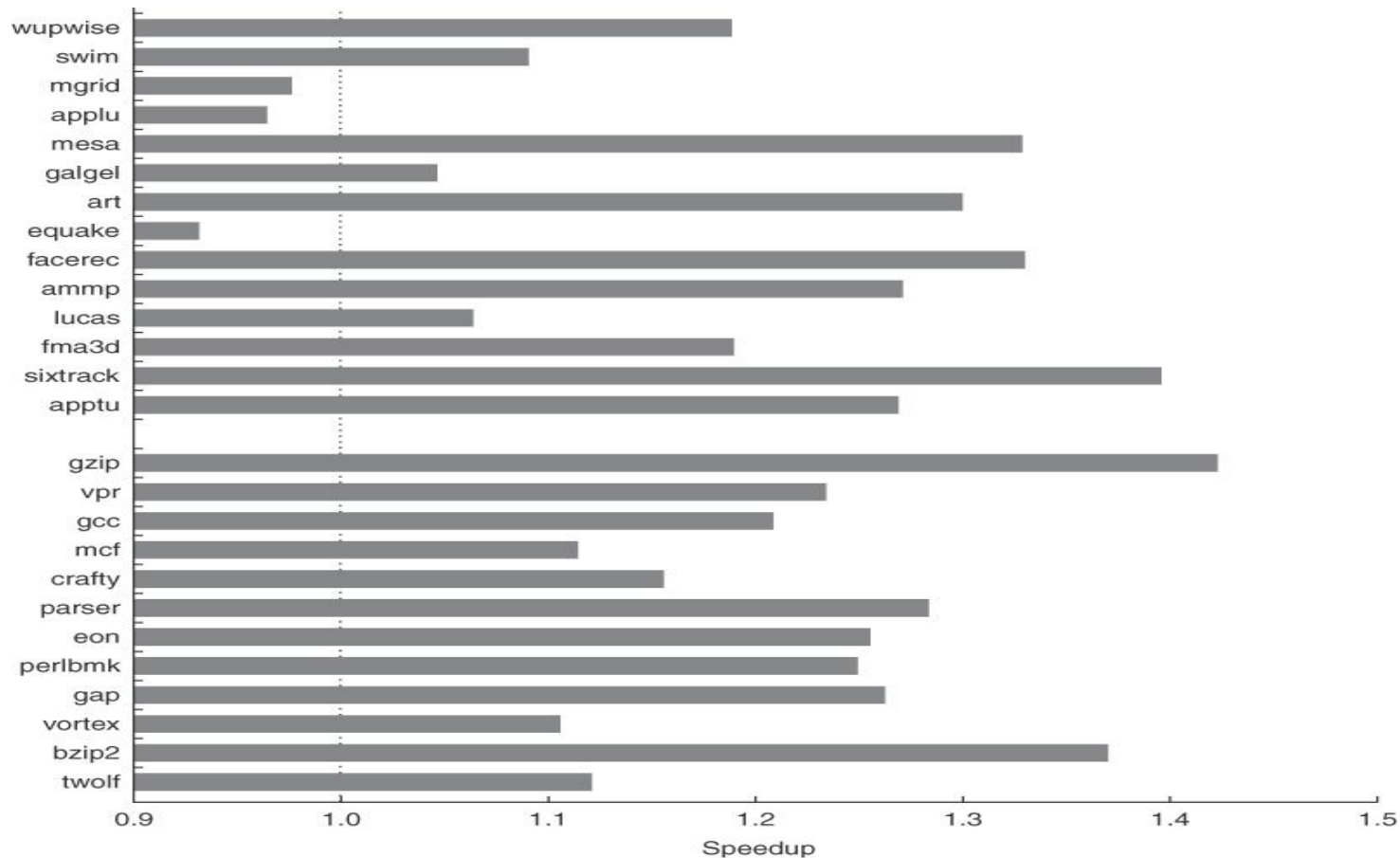
IEEE All rights reserved

TLP Era I - Multithreading



□ IBM 8-processor Power 5 with SMT (2 threads per core)

- Run two copies of an application in SMT mode versus single-thread mode
- 23% improvement in SPECintRate and 16% improvement in SPECfpRate



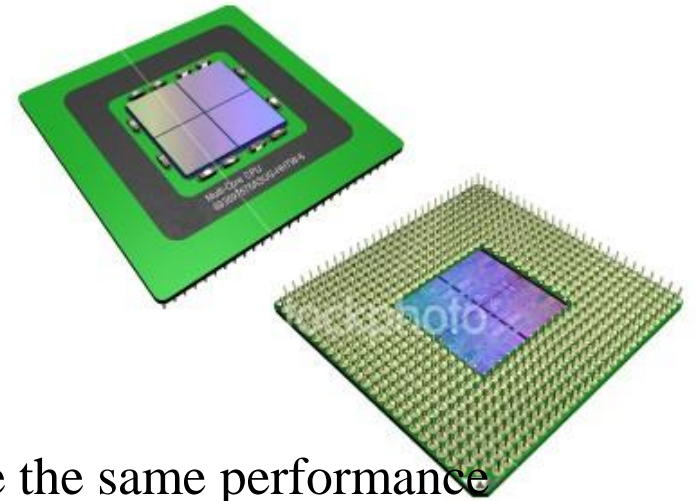
© 2007 Elsevier, Inc. All rights reserved.

TLP Era II - Multicore



□ Multicore

- Single-chip multiprocessing
- Easy to design and verify functionally
- Excellent performance/watt
 - $P_{\text{dyn}} = \alpha C_L * V_{\text{DD}}^2 * F$
 - Dual core at half clock speed can achieve the same performance (throughput) but with only $\frac{1}{4}$ of the power consumption !
 - ▼ Dual core consumes $2 * C * 0.5^2 V * 0.5 F = 0.25 CV^2 F$
- Packaging, cooling, reliability
 - Power also determines the *cost of packaging/cooling*.
 - Chip temperature must be limited to avoid *reliability* issue and *leakage power dissipation*.
- Improved throughput with minor degradation in single program performance
 - For multiprocessing workloads and multi-threaded applications



Today's Microprocessor



□ Intel Core 2 Quad Processor (code name “Yorkfield”)

➤ Technology

- 45nm process, 820M transistors, $2 \times 107 \text{ mm}^2$ dies
- 2.83 GHz, two 64-bit dual-core dies in one MCM package

➤ Core microarchitecture

- Next generation *multi-core* microarchitecture introduced in Q1 2006
 - ▼ Derived from P6 microarchitecture
- Optimized for multi-cores and lower power consumption
 - ▼ Lower clock speeds for lower power but higher performance
 - ▼ 1/2 power (up to 65W) but more performance compared to dual-core Pentium D
 - ▼ 14-stage 4-issue out-of-order (OOO) pipeline
- 64bit Intel architecture (x86-64)

➤ 2 unified 6MB L2 Caches

➤ 1333MHz system bus



Intel Corp. All rights reserved

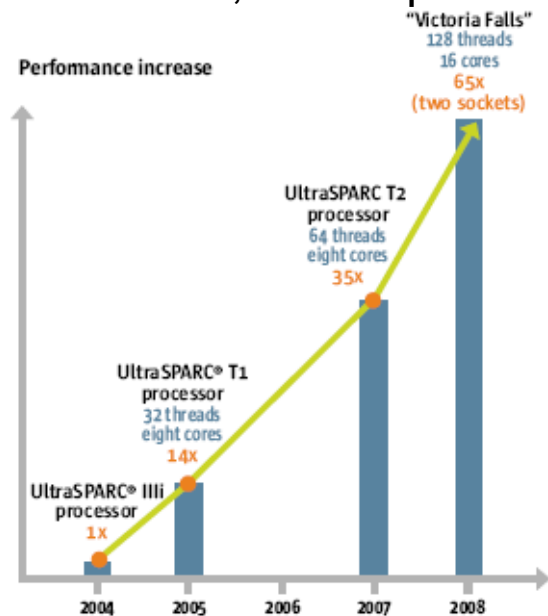
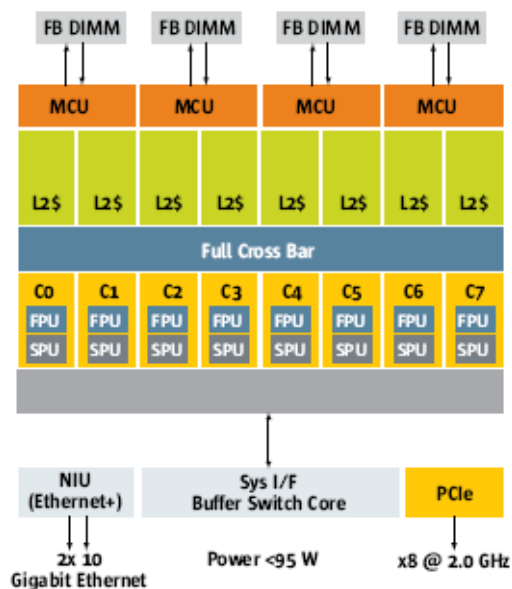
Today's Microprocessor

□ Sun UltraSPARC T2 processor (“Niagara II”)

- Multithreaded multicore technology
 - Eight 1.4 GHz cores, 8 threads per core → total 64 threads
 - 65nm process, 1831 pin BGA, 503M transistors, 84W power consumption
- Core microarchitecture
 - Two issue 8-stage instruction pipelines & pipelined FPU per core
- 4MB L2 – 8 banks, 64 FB DIMMs, 60+ GB/s memory bandwidth
- Security coprocessor per core and dual 10GB Ethernet, PCI Express



Oracle Inc. All rights reserved

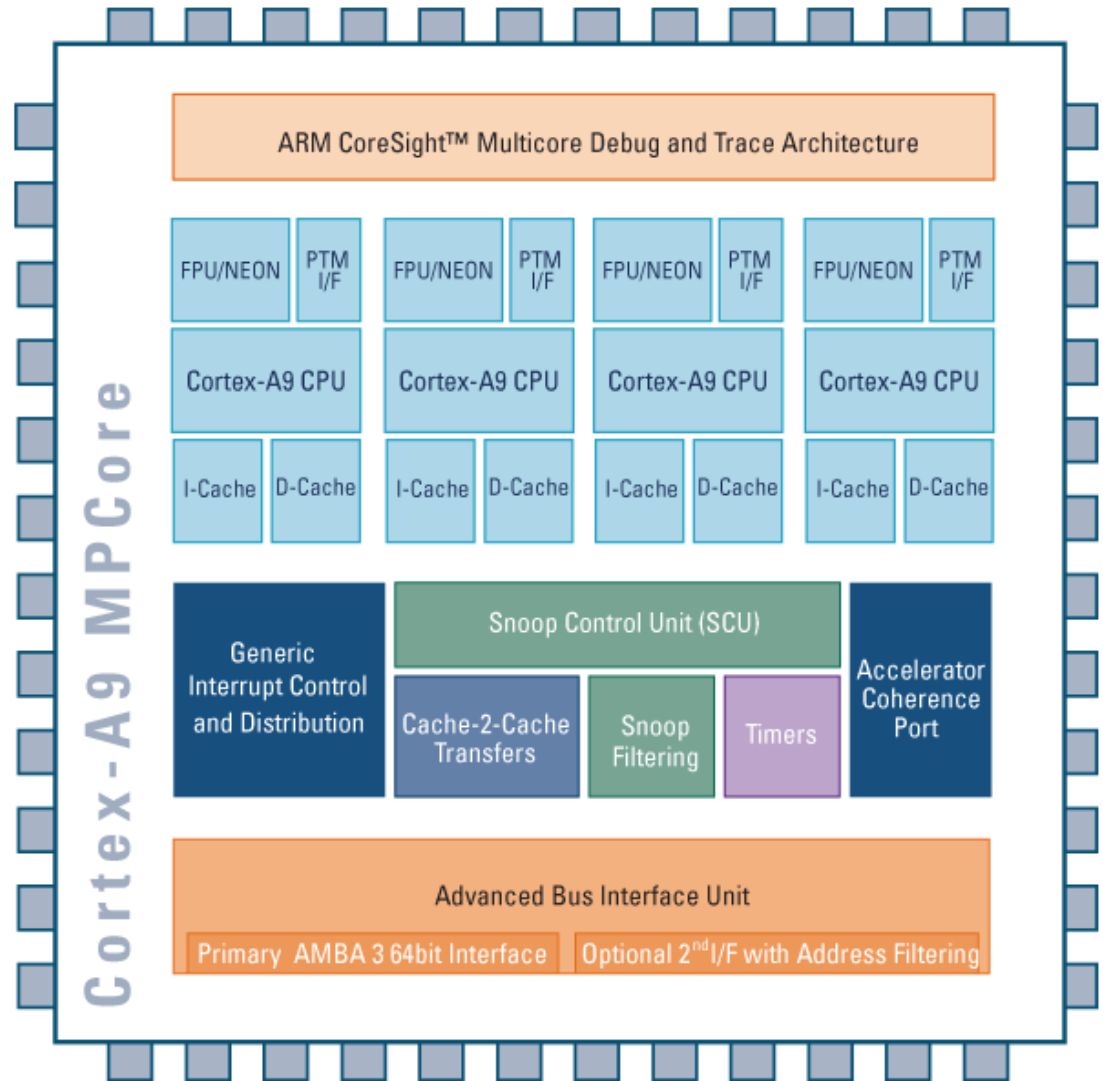


Today's Microprocessor

□ Cortex A-9 MPCore

- ARMv7 ISA
- Support complex OS and multiuser applications
- 2-issue superscalar 8-stage OOO pipeline
- FPU supports both SP and DP operations
- NEON SIMD media processing engine
- MPCore technology that can support 1 ~ 4 cores

ARM Ltd. All rights reserved

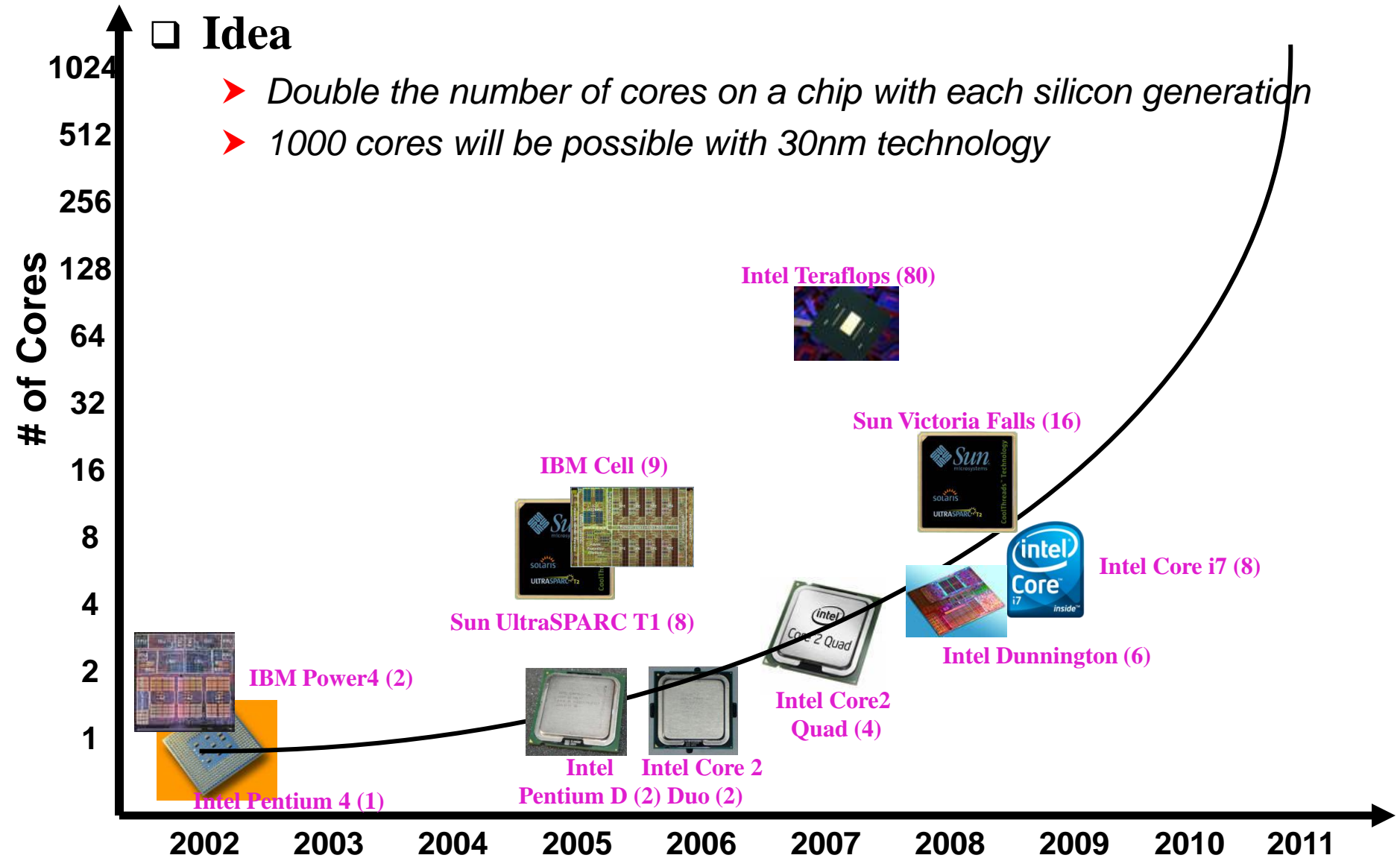


Future CPU Microarchitecture - MANYCORE



□ Idea

- Double the number of cores on a chip with each silicon generation
- 1000 cores will be possible with 30nm technology



Future CPU Microarchitecture - MANYCORE



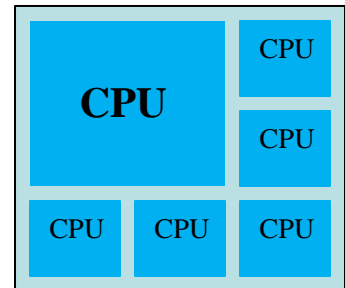
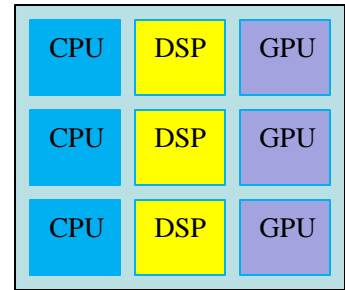
□ Architecture

➤ Core architecture

- Should be the *most efficient in MIPS/watt and MIPS/silicon*.
- Modestly pipelined (8~14 stages) in-order pipeline

➤ System architecture

- Heterogeneous vs. homogeneous MP
 - ▼ Heterogeneous in terms of functionality
 - ▼ Heterogeneous in terms of performance
 - ◆ Amdahl's Law
- Shared vs. distributed memory MP
 - ▼ Shared memory multicore
 - ◆ Most of existing multicores
 - ◆ Preserve the programming paradigm via binary compatibility and cache coherence
 - ▼ Distributed memory multicores
 - ◆ More scalable hardware and suitable for manycore architectures



Future CPU Microarchitecture I - MANYCORE

□ Issues

- On-chip interconnects
 - Buses and crossbar will not be scalable to 1000 cores!
 - Packet-switched point-to-point interconnects
 - ▼ Ring (IBM Cell), 2D/3D mesh/torus (RAW) networks
 - ▼ Can provide scalable bandwidth. But, how about latency?
- Cache coherence
 - Bus-based snooping protocols cannot be used!
 - Directory-based protocols for up to 100 cores
 - More simplified and flexible coherence protocols will be needed to leverage the improved bandwidth and low latency.
 - ▼ Caches can be adapted between private and shared configurations.
 - ▼ More direct control over the memory hierarchy. Or, software-managed caches
- Off-chip pin bandwidth
 - Manycores will unleash a much higher numbers of MIPS in a single chip.
 - More demand on IO pin bandwidth
 - ▼ Need to achieve 100 GB/s ~ 1TB/s memory bandwidth
 - More demand on DRAM out of total system silicon

Future CPU Microarchitecture I - MANYCORE

□ Projection

- Pin IO bandwidth cannot sustain the memory demands of manycores
- Multicores may work from 2 to 8 processors on a chip
- Diminishing returns as 16 or 32 processors are realized!
 - Just as returns fell with ILP beyond 4~6 issue now available
- But for applications with high TLP, manycore will be a good design choice
 - Network processors, Intel's RMS (Recognition, Mining, Synthesis)

Future CPU Architecture II – Multiple SoC



❑ Idea – System on Chip!

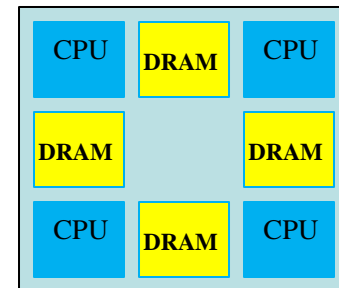
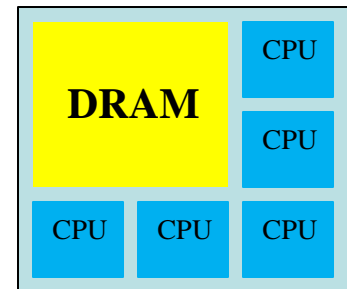
- Integrate main memory on chip
- Much higher memory bandwidth and reduced memory access latencies

❑ Memory hierarchy issue

- For memory expansion, off-chip DRAMs may need to be provided
 - This implies *multiple levels of DRAM in the memory hierarchy*
 - On-chip DRAMs can be used as a cache for the off-chip DRAM
- On-chip memory is divided into SRAMs and DRAMs
 - Should we use SRAMs for caches?

❑ Multiple systems on chip

- *Single monolithic DRAM shared by multiple cores*
- *Distributed DRAM blocks across multiple cores*



Intel Terascale processor



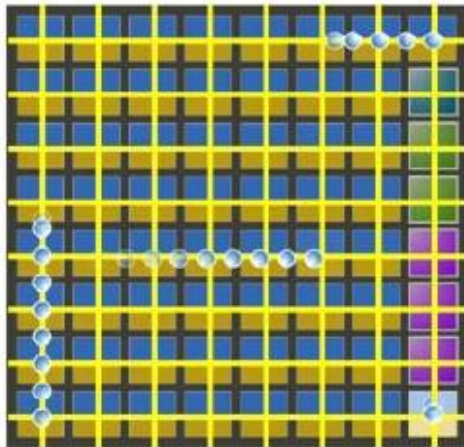
□ Features

- 80 3.13 GHz processor cores, 1.01 TFLOPS at 1.0V, 62W, 100M transistors
- 3D stacked memory
- Mesh interconnects – provides 80GB/s bandwidth

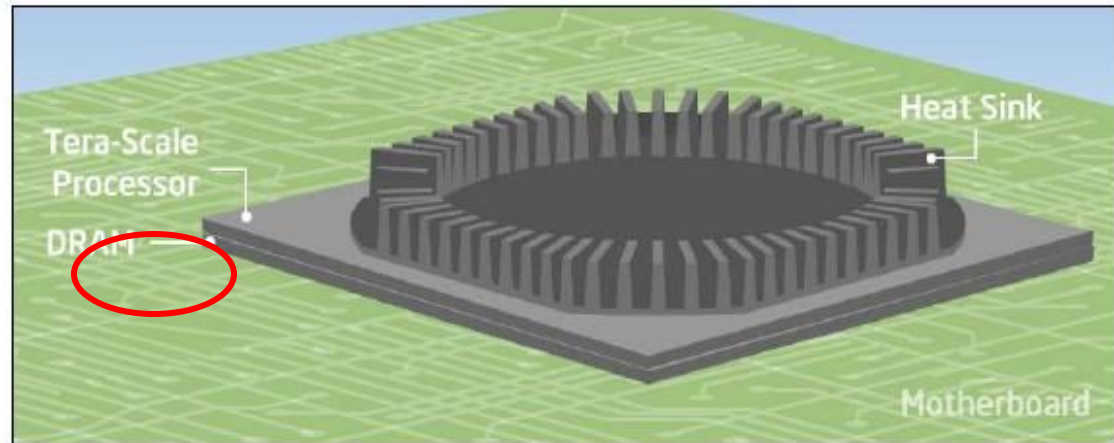
□ Challenges

- On-die power dissipation
- Off-chip memory bandwidth
- Cache hierarchy design and coherence

Intel Corp. All rights reserved

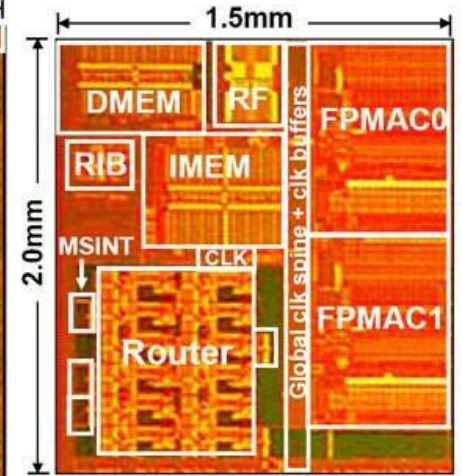
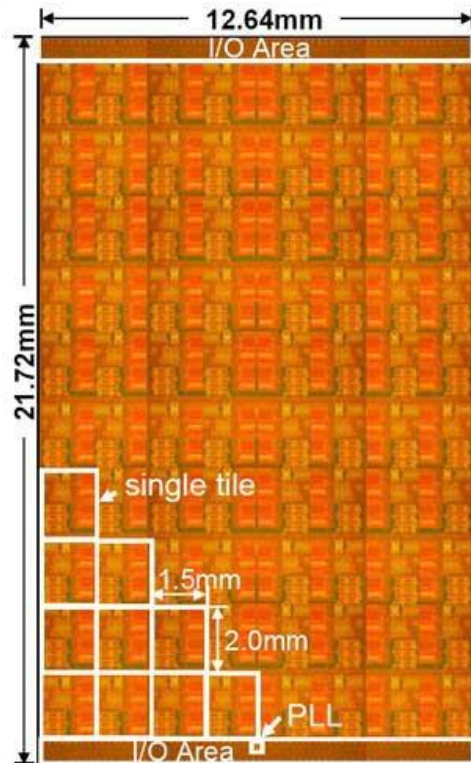
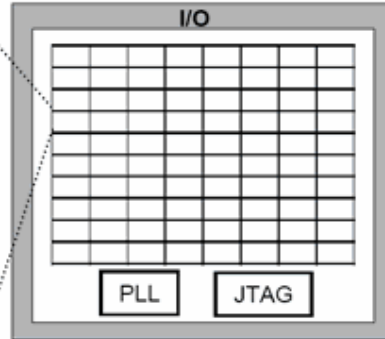
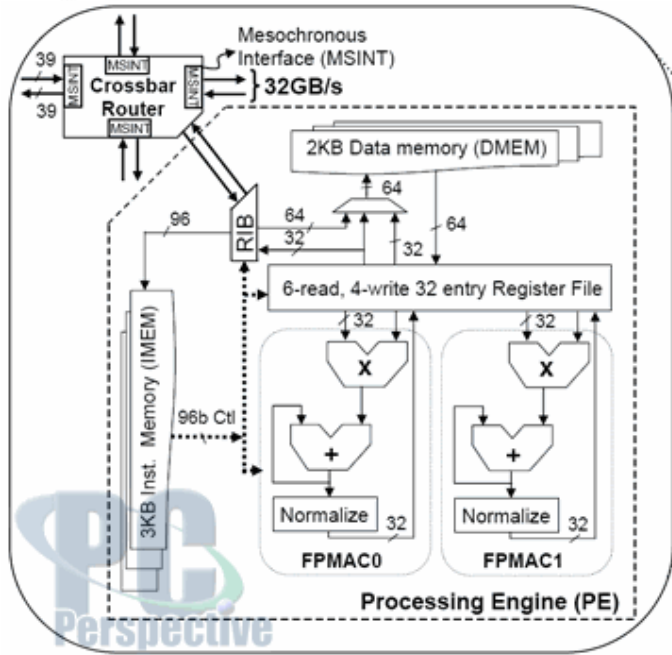


Example Mesh



The key technologies of this first Tera-scale Research Prototype are a mesh interconnect (left) and support for 3D stacked memory (above).

Intel Terascale processor



Technology	65nm CMOS Process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 Million
Die Area	275mm ²
Tile area	3mm ²
Package	1248 pin LGA, 14 layers, 343 signal pins

Intel Corp. All rights reserved

Trend - Change of Wisdoms



- ❑ **1. Power is free, but transistors are expensive.**
 - “Power wall”: Power is expensive, but transistors are “free”.
- ❑ **2. Regarding power, the only concern is dynamic power.**
 - For desktops/servers, static power due to leakage can be 40% of total power.
- ❑ **3. Can reveal more ILP via compilers/arch innovation.**
 - “ILP wall”: There are diminishing returns on finding more ILP.
- ❑ **4. Multiply is slow, but load and store is fast.**
 - “Memory wall”: Load and store is slow, but multiply is fast. 200 clocks to access DRAM, but FP multiplies may take only 4 clock cycles.
- ❑ **5. Uniprocessor performance doubles every 18 months.**
 - **Power Wall + Memory Wall + ILP Wall:** The doubling of uniprocessor performance may now take 5 years.
- ❑ **6. Don't bother parallelizing your application, as you can just wait and run it on a faster sequential computer.**
 - It will be a very long wait for a faster sequential computer.
- ❑ **7. Increasing clock frequency is the primary method of improving processor performance.**
 - Increasing parallelism is the primary method of improving processor performance.