2a - Dynamic Instruction Level Parallelism (ILP)

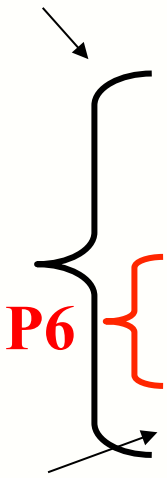
- The **P6** microarchitecture (Pentium III)
- The **NetBurst** microarchitecture (Pentium 4)
- The **Intel Core Microarchitecture** (Core e Core 2)
- The **Nehalem/Westmere** microarchitecture (Core i7-i5-i3)
- The **Sandy / Ivy bridge** microarchitecture (Core i7-i5-i3)

in part 2b – Dynamic ILP

- The **AMD K10** microarchitecture (Phenom, Opteron, Athlon)
- Next Intel and AMD microarchitectures
- INTEL vs AMD
- GPU as GP CPUs
- SPARC T3 Rainbow Fall

The Intel P6 microarchitecture

The Intel P6 microarchitecture belongs to the familty of INTEL processors that begins with 4004 and ends with Pentium III (Tanenbaum, Fig. 1.11):



Chip	Date	MHz	Transistors	Memory	Notes
4004	4/1971	0.108	2300	640	First microprocessor on a chip
8008	4/1972	0.108	3500	16 KB	First 8-bit microprocessor
8080	4/1974	2	6000	64 KB	First general-purpose CPU on a chip
8086	6/1978	5–10	29,000	1 MB	First 16-bit CPU on a chip
8088	6/1979	5–8	29,000	1 MB	Used in IBM PC
80286	2/1982	8–12	134,000	16 MB	Memory protection present
80386	10/1985	16–33	275,000	4 GB	First 32-bit CPU
80486	4/1989	25–100	1.2M	4 GB	Built-in 8-KB cache memory
Pentium	3/1993	60–233	3.1M	4 GB	Two pipelines; later models had MMX
Pentium Pro	3/1995	150–200	5.5M	4 GB	Two levels of cache built in
Pentium II	5/1997	233–450	7.5M	4 GB	Pentium Pro plus MMX instructions
Pentium III	2/1999	650–1400	9.5M	4 GB	SSE Instructions for 3D graphics
Pentium 4	11/2000	1300–3800	42M	4 GB	Hyperthreading; more SSE instructions

N.B: some recent versions of Pentium 4 used the Intel 64 ISA

The Intel P6 microarchitecture

- A basic feature of this processors, and a major reason for their success, is backward compatibility: Core (2) Duo processors still execute programs written and compiled for 8086.
- This characteristic has required a lot of efforts and design solutions that made processors architecture very complex and cumbersome:
  - *"Difficult to explain and impossible to love"*
    Hennessy & Patterson
  - *"The x86 isn't all that complex, it just doesn't make a lot of sense"*
    Mike Johnson, AMD

The Intel P6 microarchitecture

- The microarchitecture code name P6 was first introduced with the Pentium Pro, and it is a major departure from that of the previous processor, Pentium: no longer two pipelines, but a superscalar structure capable of issuing concurrently multiple instructions
  - pentium Pro → Pentium II: adds MMX instructions (MultiMedia Extension), already available in Pentium MMX.
  - Pentium II → Pentium III: adds SSE instructions (Streaming SIMD Extension – SIMD = Single Instruction Stream – Multiple Data Stream), for 3D applications

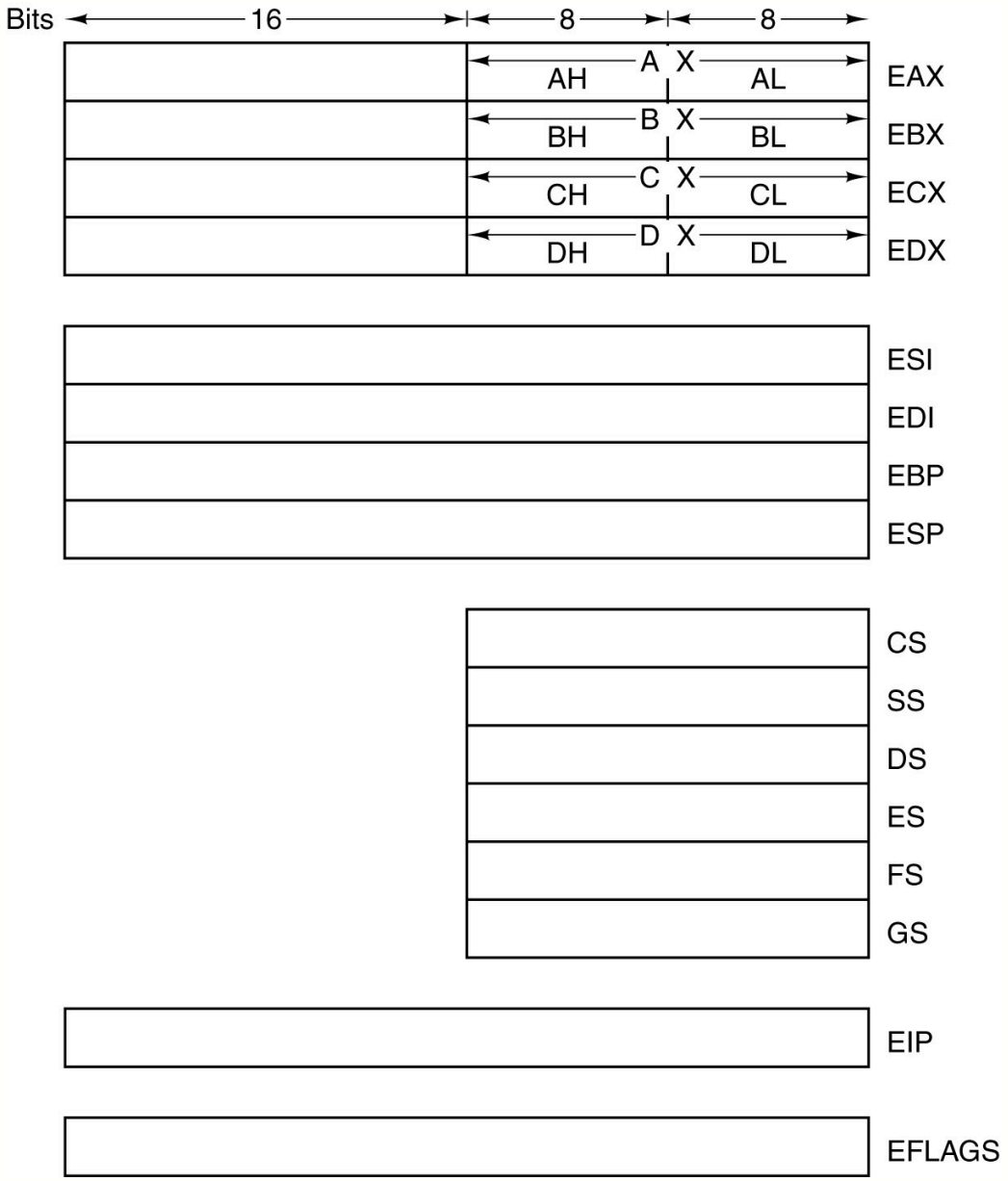
The Intel P6 microarchitecture

- Besides, the various models of the P6 architecture differentiate mainly on clock speed and on cache system (Hennessy-Patterson, Fig. 3.47):

Processor	year	clock (MHz)	L1 (I – D)	L2
Pentium Pro	1995	100-200	8K + 8K	256k – 1M
Pentium II	1998	233-450	16k + 16k	256k – 512k
Pentium II Xeon	1999	400-450	16k + 16k	512k – 2M
Celeron	1999	500-900	16k + 16k	128k
Pentium III	1999	450-1100	16k + 16k	256k – 512k
Pentium III Xeon	2000	700-900	16k + 16k	1M – 2M

The Intel P6 microarchitecture

- "Visible" registers in P6 architecture: the first 8 are used also as general-purpose registers.
- Registers CS—GS are remains of 8088, and are often unused.
- Furthermore, there are 8 floating point registers.
- (Tanenbaum, Fig. 5.3)



The Intel P6 microarchitecture

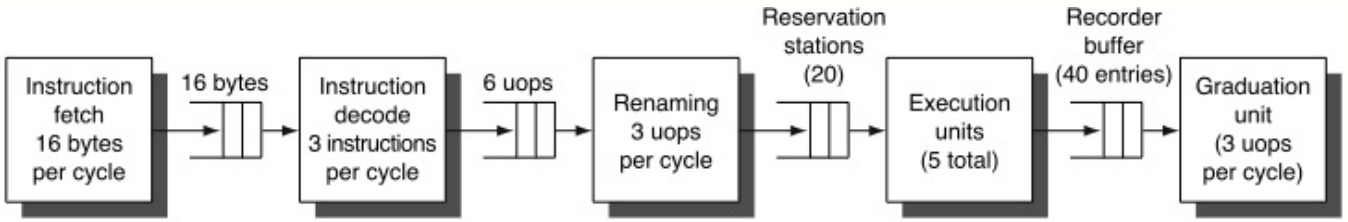
- P6 actually specifies a dynamically scheduled processor that starts embedding RISC design concepts.
- IA-32 instructions, typically CISC (variable length format from 1 to 17 bytes (!) with at least 10 different addressing modes) are translated into 72-bit RISC microinstructions (the so-called uops) executed in pipelining mode.
- Most IA-32 instructions can be translated with 1 to 4 uops.

The Intel P6 microarchitecture

- IA-32 instructions IA-32 requiring more than 4 uops (e.g. string manipulation) are not translated directly, rather they are executed through conventional microcode stored in a 8K x 72bit ROM holding sequences of uops.
- At each clock cycle, up to 3 IA-32 instructions are fetched, decoded and translated into sequences of uops.
- A maximum of 6 uops per clock cycle can be generated in the translation phase.

The microarchitecture Intel P6

- uops are executed in a dynamically scheduled pipeline using reservation stations with speculation through a ROB.
- Up to three uops per clock cycle can:
  - be forwarded to reservation stations
  - carry out commit (called “graduation” - Hennessy-Patterson, Fig. 3.49)



The Intel P6 pipeline

- The P6 pipeline consists of 14 stages:

a) 8 stages to fetch in-order IA-32 instructions, decode and issue. Specifically, 1 cycle is required to estabilish the IA-32 instruction length, and 2 cycles to generate the corresponding uops

- a 512-entry branch target buffer for deciding on branch speculation
- register renaming is also carried out in these stages by dispatching to the 20 available reservation stations and to one of the 40 entries in the ROB

The Intel P6 pipeline

b) 3 stages are used for out-of-order execution, in five functional units:

- ALU, for integer operations
- FP unit for multiplication and division
- branch unit
- LOAD unit
- STORE unit

c) 3 stages are required for commit.

Intel P6: some data

- With a fairly large number of benchmarks, the following average figures have been obtained:
  - 1 uop *committed* per clock cycle (beware, this just an average, in 23% of case, 3 uops commit per clock cycle).
  - 1,37 uops to execute a IA-32 instruction
  - 1,15 clock cycles to complete a IA-32 instruction, in integer programs (so 0,87 IA-32 instructions completed per clock cycle)
  - 2 clock cycles to complete a IA-32 instruction, in FP programs (0,5 IA-32 FP instructions per clock cycle)
- The last two figures show that P6 was designed with a focus on “integer” applications

The microarchitecture NetBurst: Pentium 4

- NetBurst is the name of Pentium 4 microarchitecture.
- NetBurst is a departure from P6, especially in some key features of the pipeline to have a very high-speed clock and a larger number of committed instructions per clock cycle.
- The resulting processor is extremely quick, but power consumption is very high, with the dissipation problems associated with it.

The NetBurst microarchitecture

1. NetBurst has a pipeline with at least 20 stages, but in the last versions of Pentium 4 (e.g, Prescott) the stages grew to 31, 2 of them just to move the result from one point to another in the chip. This increase in the number of stages is necessary to decrease the clock cycle duration. In Pentium 4 too, up to 3 IA-32 instructions can be committed per clock cycle.
2. Speculation applies a technique different from the ROB, that allows to have up to 128 instructions uop waiting for commit (a similar technique has been used also in CPU MIPS R10K and Alpha 21264)

The NetBurst microarchitecture

3. There are 7 functional units, instead of 5 in P6: an integer ALU and a floating point unit have been added.
4. ALUs have a double frequency clock, and each can execute two integer operations per clock cycle. An *aggressive* data cache decreases by 1 the number of clock cycles required to execute a LOAD, with respect to Pentium III.
5. A *trace cache* is used as L1 cache for instructions uops (up to 12k uops). P6 instead uses an ordinary cache for instructions, which stored IA-32 instructions (rather than uops) that could be possibly translated uselessly many times.

The NetBurst microarchitecture

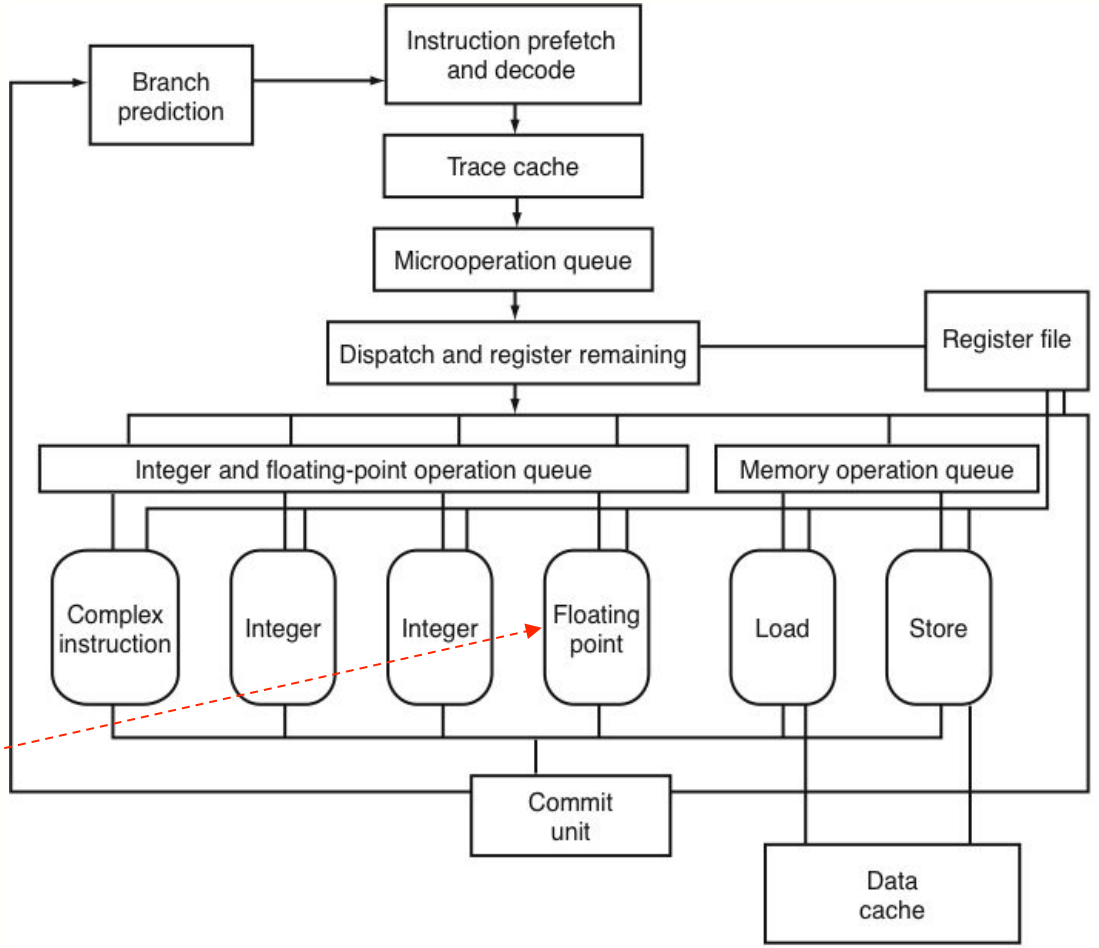
6. The branch target buffer is 8-times larger than in P6 (up to 4096 branches), with a more advanced prediction algorithm (a decrease in wrong predictions by 33% with respect to Pentium III). If the branch is not in the BTB, a static branch prediction is applied.
7. An enriched set of FP instructions allowing to execute two FP operations per instruction.
8. The L2 cache has grown from 256K up to 2 MB. The L1 data cache is 8-way *set-associative* (the *trace cache* and *set-associative* caches will be covered in the chapter on the cache)
9. Hyper-threading: a new architectural feature

The NetBurst microarchitecture

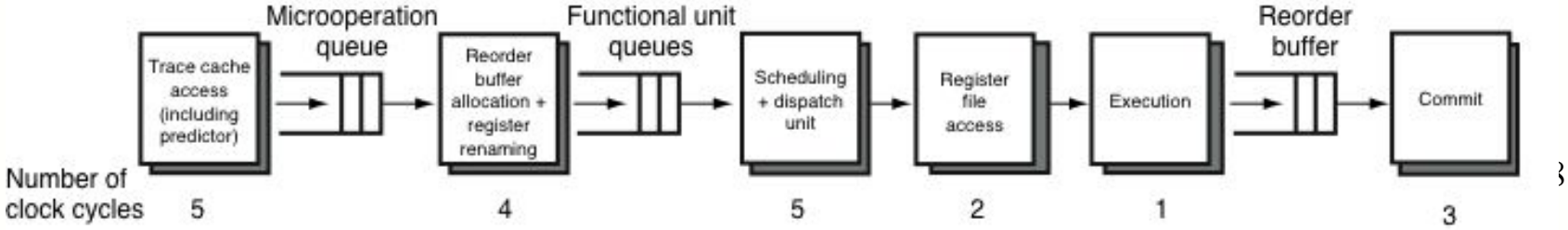
- This improvements make the Pentium 4 handle up to 126 instructions uops concurrently, including 42 LOAD and 24 STORE!
- The reservation stations offer 128 rename registers. 8 are used to represent the integer registers“visible in the ISA”, and are mapped dynamically onto the internal registerss.
- Compared to Pentium III, Pentium 4 has an increases of performances less than linear in the clock speed, in integer applications, and more than linear, in FP ones.

The NetBurst microarchitecture

Pentium 4 data path. The number of pipeline phases has increased with successive versions of Pentium 4 (Patterson-Hennessy fig. 6.50 e 6.51)

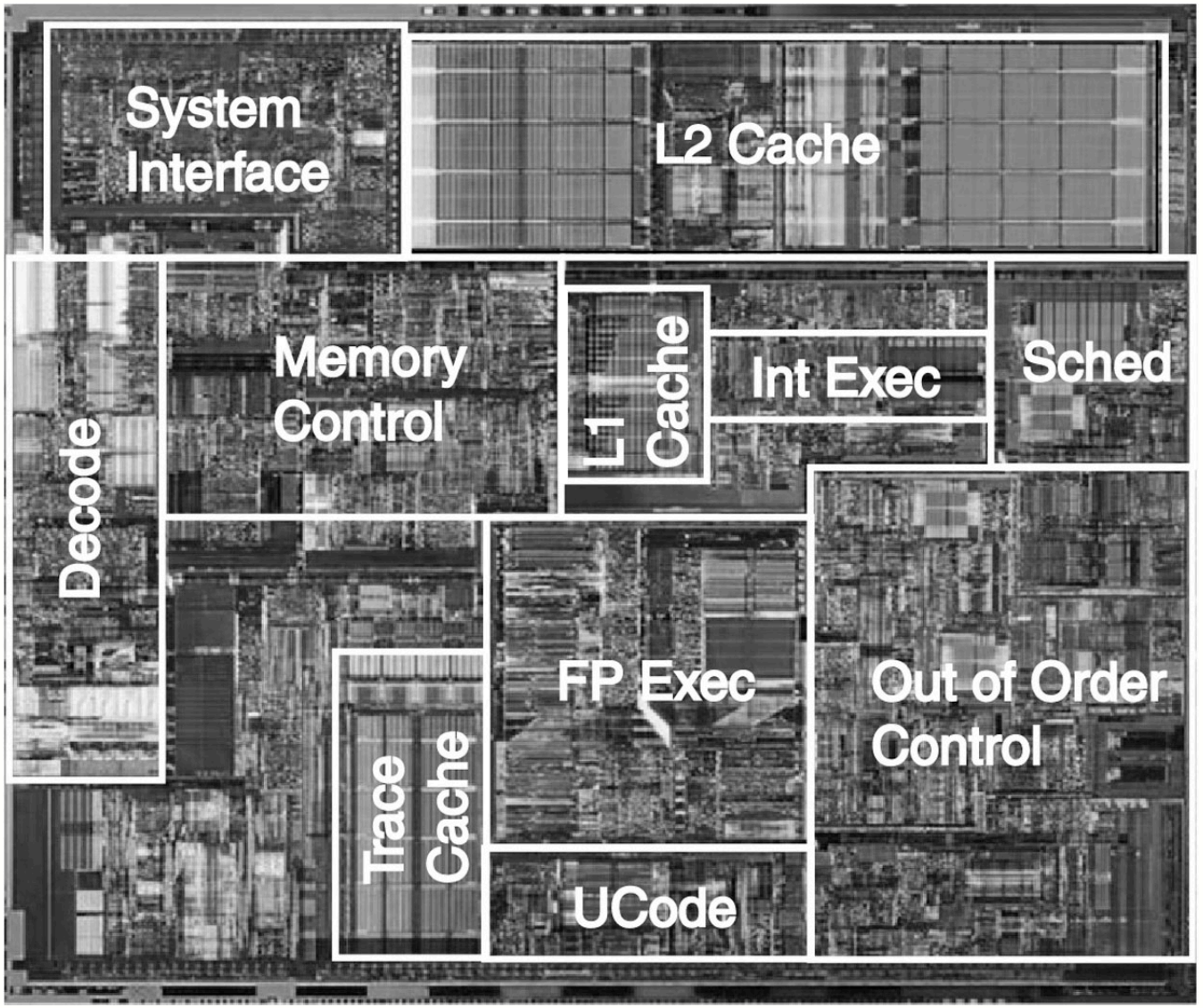


Actually 2 F.U.



Pentium 4

Pentium 4 (Tanenbaum, Fig. 1.12)



Intel Core Microarchitecture, namely Core 2 (Duo)

- NetBurst architecture has been controversial in its advantages: the high rise in clock speed has not produced a corresponding rise in performances, at least in integer applications.
- As an example, in the integer section of SPEC2000 benchmark (gcc + vortex) performance ratio of a 1 Ghz Pentium III over a 1.7 Ghz pentium 4 is 1.26: that is, performance has increased by 26%, with a clock quicker by 70%.
- Moreover, NetBurst processors have high power consumption: model Pentium 4 Prescott with a 3.5 Ghz clock can draw up to 130 watt.

Intel Core Microarchitecture

- Since 2005 Intel decided to return to an updated version of P6 microarchitecture, with a 12-14 stage pipeline.
- Meanwhile, it sets up a "rebranding" operation, to highlight the departure from NetBurst/Pentium 4.
- In january 2006 the first Core Duo (*Yonah*) enters the market: it is the first dual core processor fabricated with the 65 nanometers technology.
  - But it was not the first Intel dual core: actually, this was the Pentium D, set up with two Pentium 4 (Cedar Mill) on separate, side-to-side dies
- Core Duo is based on the P6 version already used in Pentium M, designed for the "mobile" market: its main feature is indeed low power consumption.

Intel Core Microarchitecture

- Core Duo has 2 Mbytes of shared L2 cache, a 12-stage pipeline with a clock frequency that ranges up to 2.5 Ghz.
- It has roughly 151 milion transistors, and does not yet support Intel 64-bit extended instruction set (already available in Pentium D and in some Prescott). There is a Core Solo version to deploy dies where one of the two cores is faulty.

Intel Core Microarchitecture

- At the end of July 2006 Intel releases Core 2 Duo, that marks the "definite" dismittal of the Pentium brand and the return to a single processor microarchitecture both for the desktop and for the mobile products.
- Actually, in Pentium era, Pentium 4 (NetBurst microarchitecture) was used in desktop systems, and Pentium M (P6 microarchitecture) in notebooks.
- The Core 2 Duo family initially consists of **Merom** (mobile) and **Conroe** (desktop) processors, both fabricated with 65 nm technology. The differences are in the speed of the bus, the socket, and a better power optimization in Merom.
- In 2007/2008 Intel introduces the 45 nm technology in processors **Penryn** (mobile) e **Wolfdale** (desktop), and a clock frequecy beyond 3 GHZ

Intel Core Microarchitecture

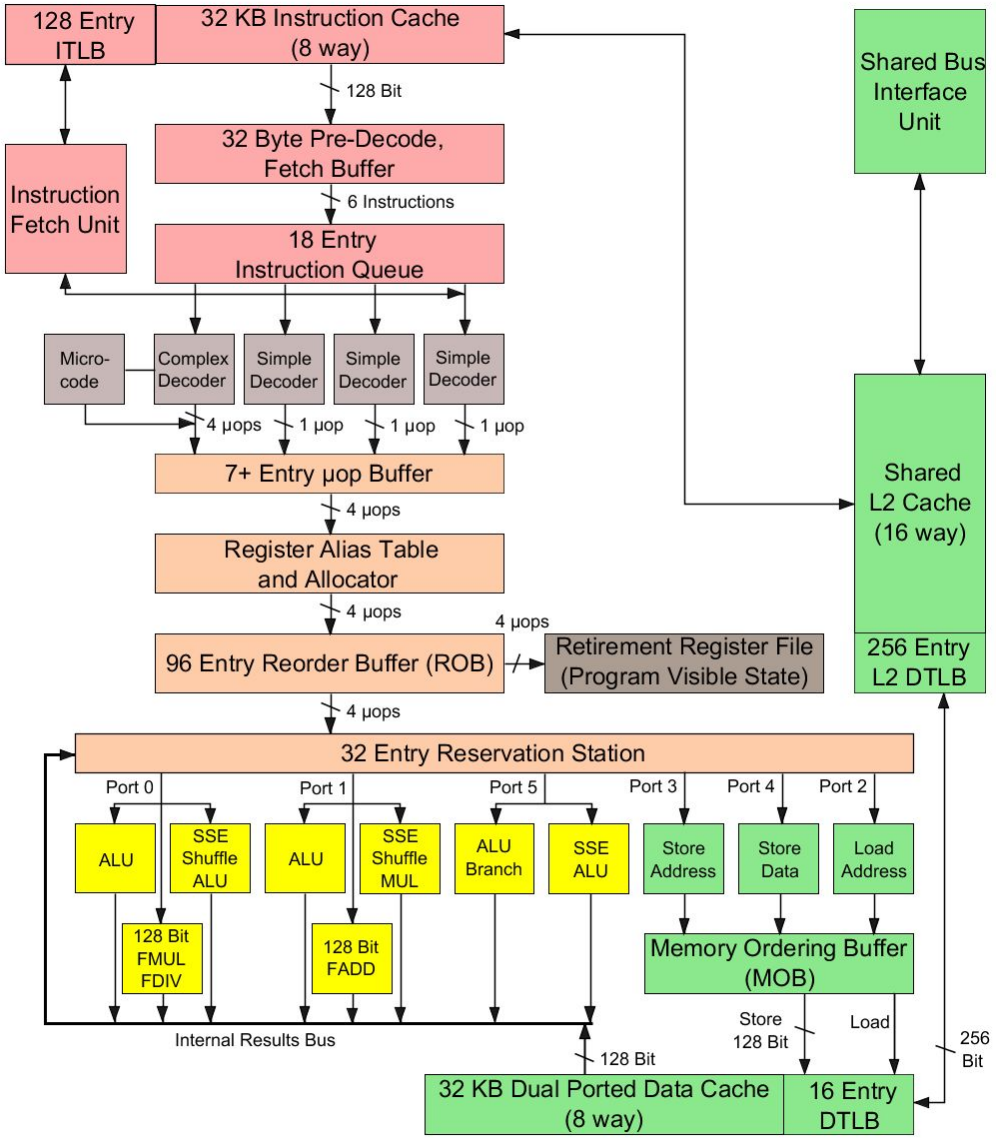
- With the rebranding operation, Core 2 Duo is renamed **Intel Core Microarchitecture**, and it leverages the new 64-bit technology known as **Intel 64** (as well as **EM64T** or **IA-32e**)
- Core 2 Duo has a 4 Mbyte, shared L2 cache, dynamically allocated to the two cores; each core has private 32 Kbyte L1 D-cache and I-cache (no trace cache). These values change with models, though
- The pipeline has 14 stages, as in previous P6 versions, but the new microarchitecture is capable of translating up to 4 IA-32e instructions per clock cycle (against 3 in Pentium III e 4).
- Hyper-threading is dismissed, (deemed a major reason for the excessive power consumption of Pentium 4), substituted for by the double core (however, it will be back in Nehalem microarchitecture)

Intel Core Microarchitecture

- Core 2 Duo contains roughly 290 million transistors (Pentium 4 EE contained some 190) and have a maxim power consumption of 45-65 watt (on the average, roughly half that of Pentium 4).
- Low power consumption is obtained by dynamically “turning off” sections of the datapath not required in instruction execution (as an instance, floating point units are turned off during integer instruction execution). Furthermore, the processor can scale down clock frequency, during periods of low utilization.

Intel Core Microarchitecture

The Intel Core Microarchitecture (fonte: Wikipedia)



Intel Core 2 Architecture

Intel 64 Instruction Set Architecture

- The *Intel Core Microarchitecture* in Core 2 Duo (as well as a few recent versions of *NetBurst*, but not *P6* in Core Duo) adopts a 64 bit instruction set known as **Intel 64** (not to be confused with IA-64 instruction set, still an Intel product, to be discussed in another chapter).
- Intel 64 has also been termed **IA-32e** and **EM64T** (Extended Memory 64 bit Technology). Actually, it is the Intel version of the **x86-64** ISA introduced originally by AMD (later renamed **AMD64**), that extends and includes the IA-32 ISA.
- It was Intel to follow AMD path, not vice-versa ! (the differences between Intel 64 and AMD64 specs are minimal, and touch only compiler and operating system development).

Intel 64 Instruction Set Architecture

- Intel 64 describes a 64-bit architecture and a 64-bit instruction set, with 16 general registers (instead of 8 as in IA-32) and 16 SSE registers (instead of 8) for multimedial applications.

- The logical address space is extended to 2^{48} byte (it is 2^{32} in IA-32), and the physical one is 2^{40} byte (2^{36} in IA-32).

- Intel 64 has a *compatibility mode* to run 32 and 16-bit applications as if on a normal 32-bit processor. There is of course also a *native mode*, for the new 64-bit applications.

Processors and nanometers

- What is actually meant when one says that a processor is fabricated with a *xx* nm technology ?
- This unit has been defined by a group of experts with support from the association of the major semiconductor producers in the world.
- The expert group delivered (in 1998) the so-called **International Technology Roadmap for Semiconductors** that, on the basis of past developments, tries to forecast the technology evolution in the near future.
- "xx nanometers technology" implies the ability to produce semiconductors chips, notably DRAM memories, where two cells are separated by a distance of xx nanometers.

Processors and nanometers

- Semiconductor producers believe that 11 nm is a barrier that cannot be overcome, because of physical effects that would prevent correct operations of the devices.
- The dates quoted are just estimates, some producers already have (small scale) processes on the 22/16 nm
- (Source: wikipedia)

10 µm — 1971
3 µm — 1975
1.5 µm — 1982
1 µm — 1985
800 nm (0.80 µm) — 1989
600 nm (0.60 µm) — 1994
350 nm (0.35 µm) — 1995
250 nm (0.25 µm) — 1998
180 nm (0.18 µm) — 1999
130 nm (0.13 µm) — 2000
90 nm — 2002
65 nm — 2006
45 nm — 2008
32 nm — 2010
22 nm — 2011
16 nm — approx. 2013
11 nm — approx. 2016

Nehalem/Westmere microarchitecture

- In the second half of 2008, there appears the new **Nehalem microarchitecture**, that is to replace the Intel core microarchitecture

- Processors based on this microarchitecture are still fabricated with 45 nm technology, but at the begin of 2010 starts the production with 32 nm technology, the **Westmere microarchitecture** (initially labeled **Nehalem-C**). In what follows, the code name Nehalem will be used for both.

- Many processors families based on these microarchitectures are grouped with the denominations i7, i5 e i3. Single processors: *Bloomfield*, *Lynnfield*, *Clarkdale* for desktop, *Clarksfield* and *Arrandale* for mobile.

Nehalem/Westmere microarchitecture

- The Nehalem microarchitecture combines some of the features of *Intel Core microarchitecture* and *Netburst*
- Processors are produced in 2, 4 and 6 cores versions, with a clock frequency up to 3,3 GHz. The quad-core version features some 730 million transistors.
- **Hyperthreading** is again available, 2 threads per core (so, a quad-core has 8 logical CPUs)
- There is a second level **branch prediction**, more two-bit predictors are used in parallel
- Also, a two-level **Translation Lookaside Buffer** (TLB is actually a cache for the active page table)

Nehalem/Westmere microarchitecture

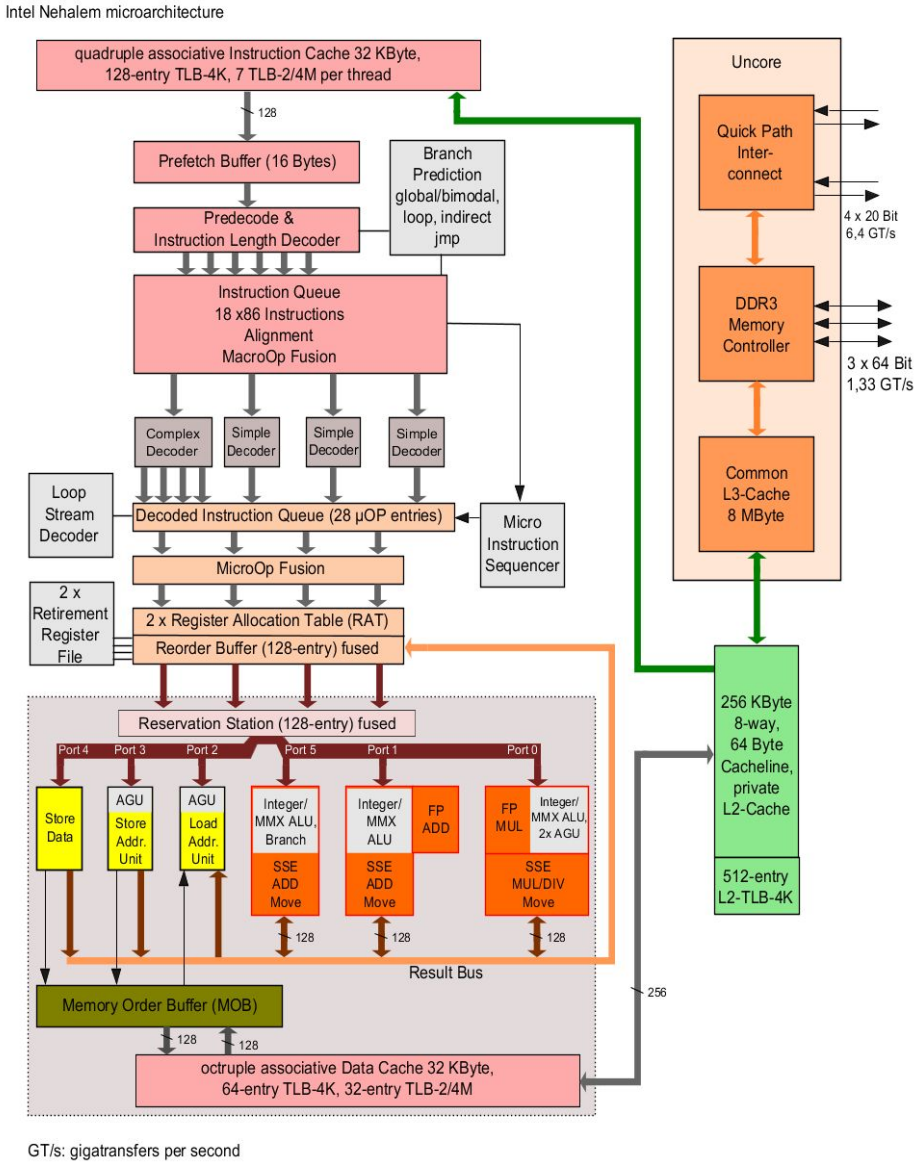
- The Nehalem microarchitecture tries to optimize memory accesses, and introduces a **third level cache**:
  - L1: 32 $KB_{istruzions}$ + 32 KB_{data} per core
  - L2: 256 KB per core
  - L3: 4 – 8 MB shared among cores
- Clear advantages over Core Microarchitecture

CPU	L1	L2	L3
Nehalem(2.66GHz)	4 clock cycles	11 clock cycles	39 clock cycles
Penryn (2.66GHz)	3 clock cycles	15 clock cycles	N/A

- Furthermore, with reference to memory access time, Nehalem completes a memory request in 2/3 of the time required by Penryn, other conditions being equal. (source: www.anandtech.com):

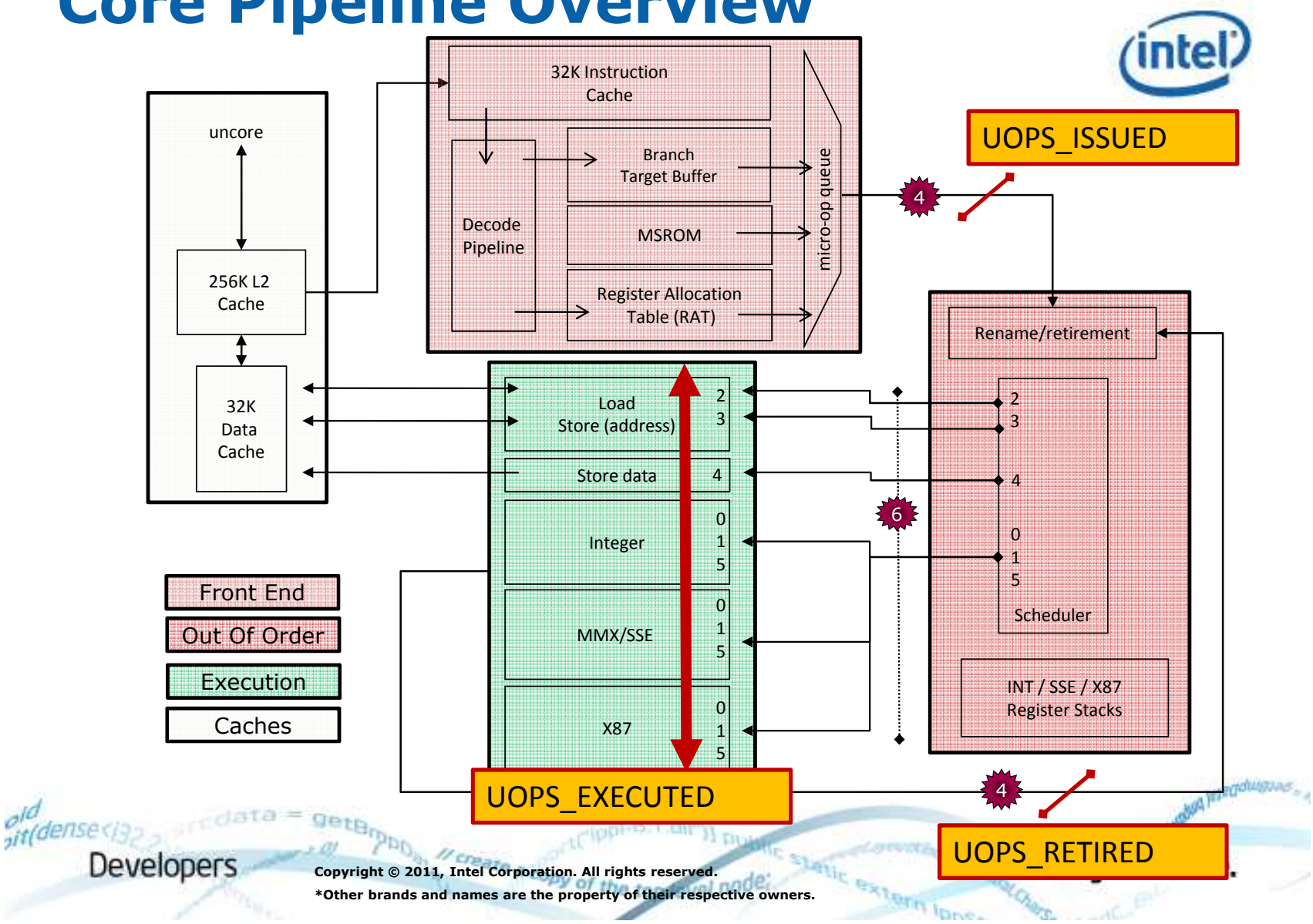
Nehalem/Westmere microarchitecture

- The microarchitecture Nehalem/Westmere (Source: wikipedia)
- Compared to Core M.A. on the average:
  - - 30% power consumption
  - +10% ÷ 25% performance
  - up to + 30% uops executed in parallel

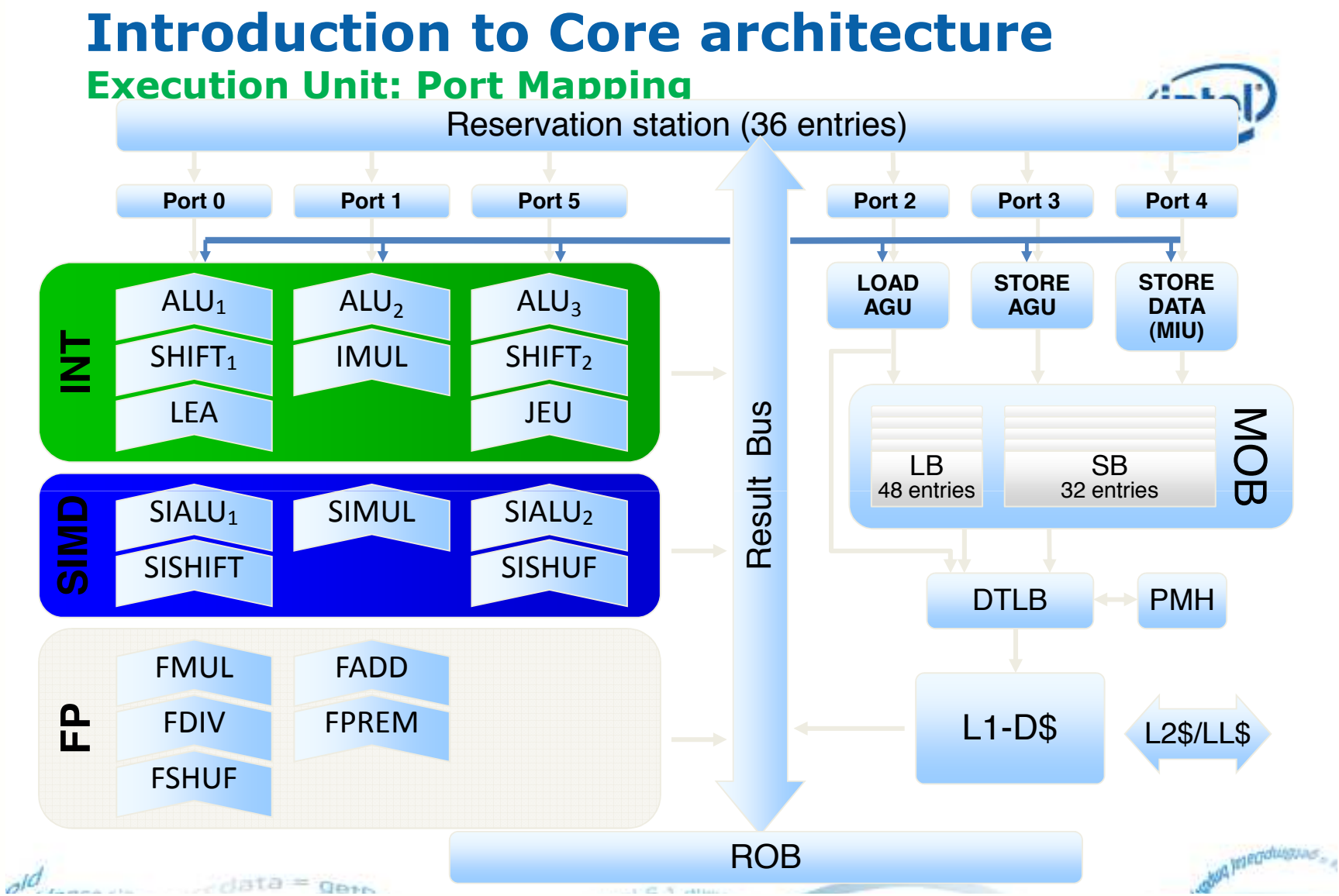


Nehalem/Westmere microarchitecture

Core Pipeline Overview

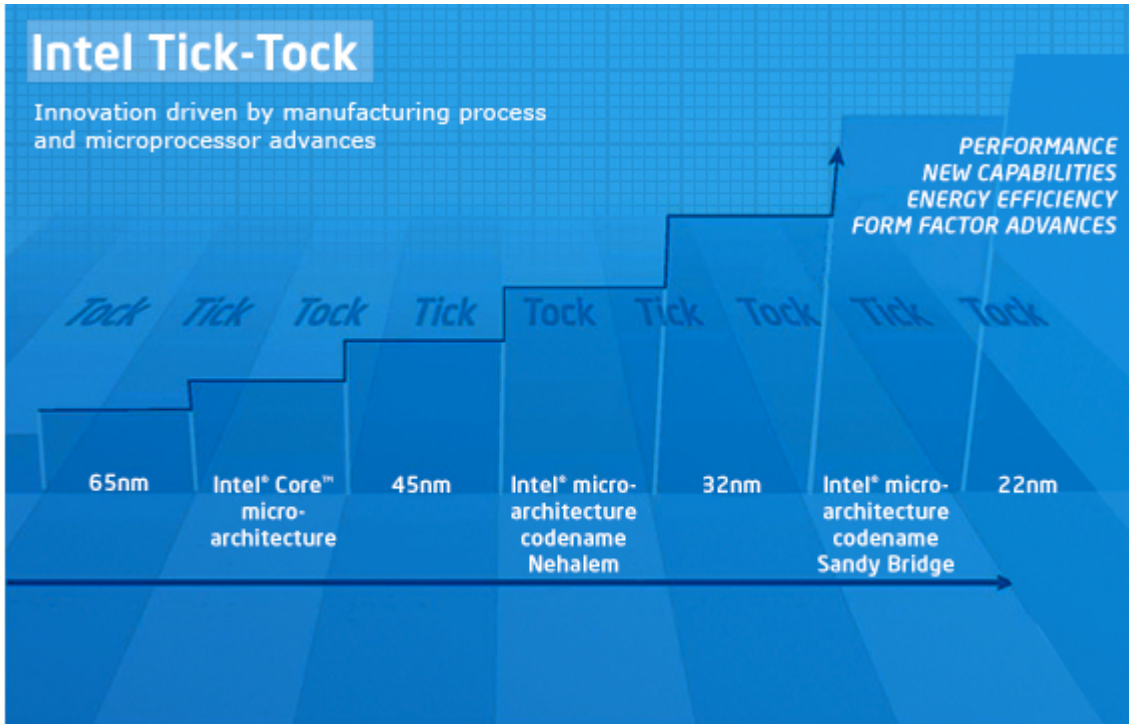


Nehalem/Westmere microarchitecture



Intel Next Microarchitectures

- Intel defines its production strategy "tick – tock":
- **Year 1, Tick**: a new fabrication process is developed (actually, fewer nanometers, more transistors) that increases the efficiency of the current microarchitecture
- **Year 2, Tock**: a new microarchitecture is developed, leveraging on the technology of the previous year

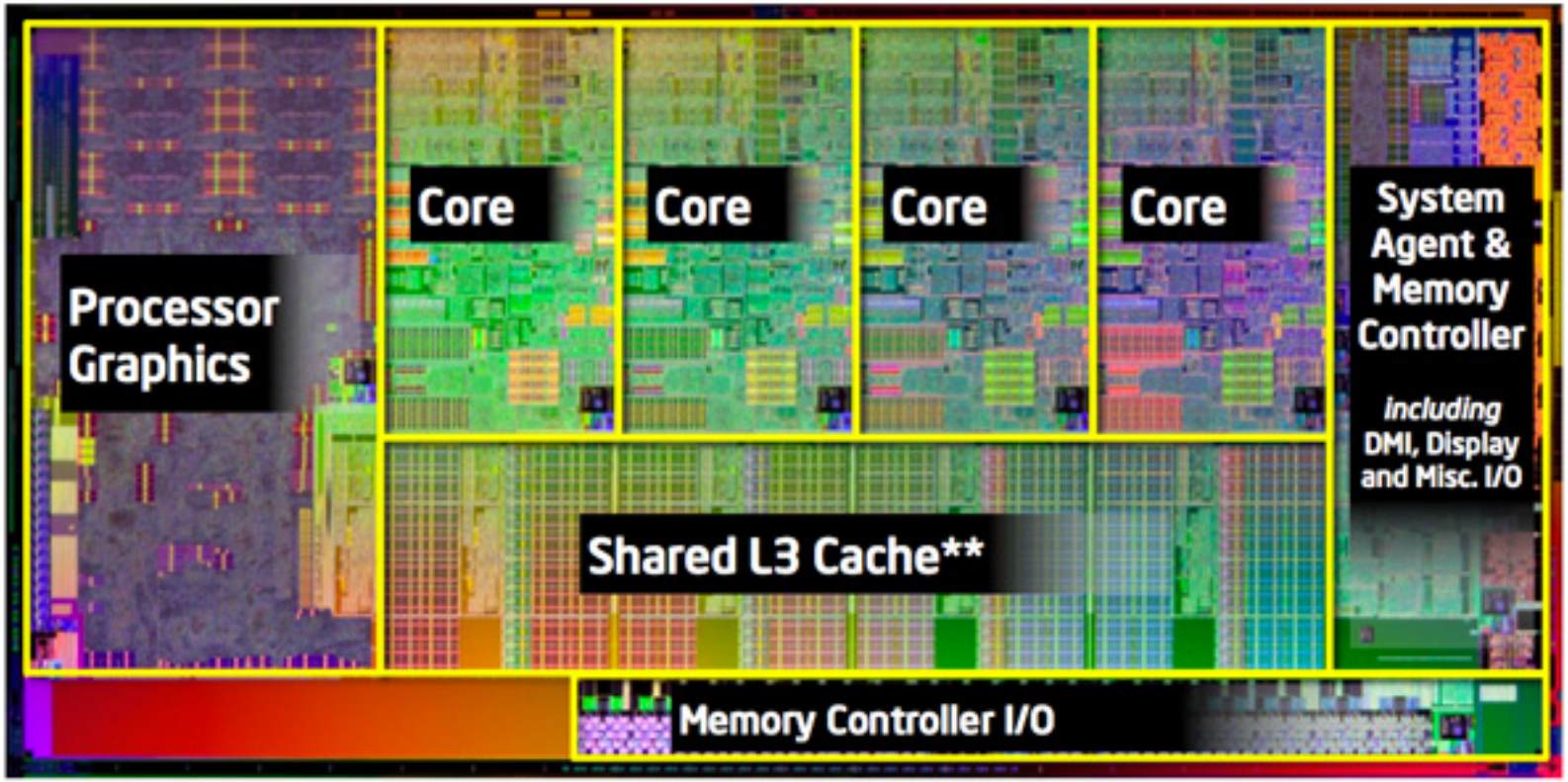


Intel Next Microarchitectures

- **Sandy Bridge** is the 32 nm microarchitecture, first processors on the market beginning 2011.
- Main characteritiscs of Sandy Bridge:
- L2 cache: 256/512 Kbyte, 8 clock cycles access time.
- L3 cache: up to 8 Mbyte, 25 clock cycles access time.
- Vector registers enlarged from 128 to 256 bits
- Clock frequency up to a 3,4 GHz (3,8 GHz with overclocking)
- GPU integrated on the same chip of processor. In Nehalem/ Westmere, the GPU is on a separate chip, CPU and GPU shared the same package (more on this later).

Intel's Next Microarchitectures

- The Sandy-bridge 4-core chip, with the GPU on the same die (source: AnandTech)



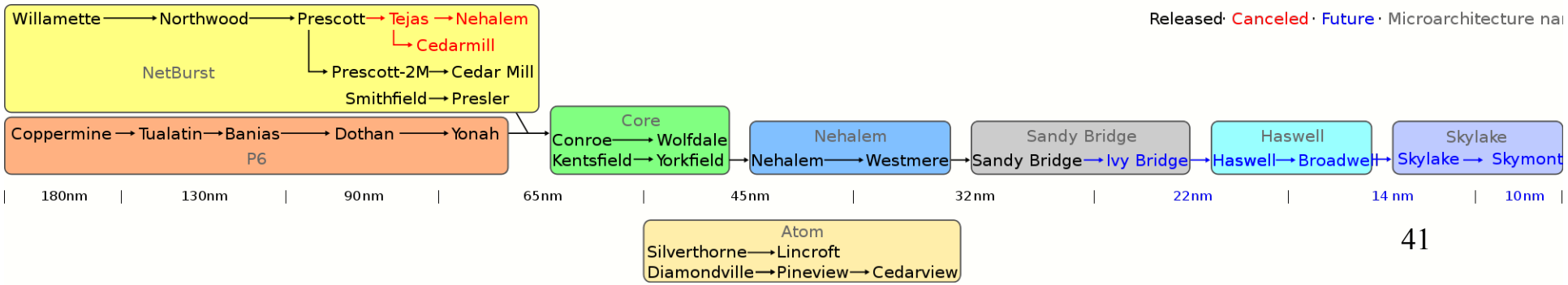
Intel Next Microarchitectures

Processor	Core Clock	Cores / Threads	L3 Cache	Price
Intel Core i7 2600K	3.4GHz	4 / 8	8MB	$317
Intel Core i7 2600	3.4GHz	4 / 8	8MB	$294
Intel Core i5 2500K	3.3GHz	4 / 4	6MB	$216
Intel Core i5 2500	3.3GHz	4 / 4	6MB	$205
Intel Core i3 2120	3.3GHz	2 / 4	3MB	$138
Intel Core i3 2100	2.93GHz	2 / 4	3MB	$117
Intel Pentium G850	2.9GHz	2 / 2	3MB	$86
Intel Pentium G620	2.6GHz	2 / 2	3MB	$64

- Some of the Sandy Bridge processors. The K versione K allows for an higher overclocking. Power consumption is in the range 65 - 95 watt. Low end processors are re-branded Pentium (source: AnandTech, august 2011)

Intel Next Microarchitectures

- According to Intel, Sandy Bridge yields 17% average increase in performance over a processor with Nehalem architecture, at the same clock frequency.
- In 2012 there should be on the market Ivy Bridge microarchitecture processors, equal to Sandy Bridge but with a 22 nm technology, and an icrease in performance of 20% over equivalent Sandy Bridge processors.
- The CPU Intel roadmap from Netburst onwards (wikipedia)



Intel Next Microarchitectures

- Codes for Intel processors families (varying in core number, clock speed and cache dimension):
  - core i7: -9xx; -8xx; -7xx; -6xx
  - core i5: -7xx; -6xx; -5xx; -4xx
  - core i3: -5xx; -3xx

Nehalem/Westmere architecture

  - core i7: -3xxx; -2xxx;
  - core i5: -2xxx
  - core i3: -2xxx

Sandy bridge architecture