A Brief History of Intel CPU Microarchitectures

Xiao-Feng Li
xiaofeng.li@gmail.com

2013-02-10
Notes

- The materials are only for my personal use.
  - Not representing Intel opinions
  - Not a complete list of Intel microprocessors
  - Not specifications of Intel microprocessors
Intel Pre-Processor Devices

• Intel founded in 1968
• Intel 3101, 1969
  – Intel first product
  – World first solid state memory device
  – 16 x 4-bit SRAM
• Intel 1103, 1970
  – World first DRAM product, 1K-bit PMOS
  – Used in HP 98000 series computers
  – By 1972, world bestselling memory chip, defeating magnetic memory
Moore’s Law

  - “The complexity for minimum component costs has increased at a rate of roughly a factor of two per year.”
  - Moore refined it to “every two years” in 1975
  - Also quoted as “every 18 months” by David House, (referring to performance)
  - Most popular formulation: #transistors/IC

- Carver Mead coined it as Moore's law around 1970
  - “Tall & Thin engineers”

- Ultimate limit of Moore’s Law
  - No one knows
  - How to use the capability? Resource limit?
# Intel MCS Family

<table>
<thead>
<tr>
<th>MCS Family</th>
<th>Intel CPU</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCS-4</td>
<td>4004</td>
<td></td>
</tr>
<tr>
<td>MCS-40</td>
<td>4040</td>
<td>MCS-40 sometimes refers also to the MCS-4 family</td>
</tr>
<tr>
<td>MCS-8</td>
<td>8008</td>
<td></td>
</tr>
<tr>
<td>MCS-80</td>
<td>8080</td>
<td>MCS-80 sometimes refers also to the MCS-8 family</td>
</tr>
<tr>
<td>MCS-85</td>
<td>8085</td>
<td>Sometimes refers to the MCS-80 and MCS-8, sometimes as the MCS-80/85 family</td>
</tr>
<tr>
<td>MCS-86</td>
<td>8086, 8088, 80186, 80188, 80286, 80386, 80486, Pentiums</td>
<td></td>
</tr>
</tbody>
</table>
Intel 4004, 1971

- World first “general purpose” micro-processor
- Lead designers
  - Ted Hoff, Federico Faggin, Stan Mazor, Masatoshi Shim
- Data
  - Word width: 4-bit
  - 2300 transistors
  - Clock: 108KHz/500/740
  - 46 instructions
  - Registers: 16 x 4-bit
  - Stack: 12 x 4-bit
  - Address space
    - 1Kb of program, 4Kb of data

2013/02/10
Brief history of Intel CPU uArch - xiaofeng.li@gmail.com
Intel 8008, 1972

- **World first 8-bit microprocessor**
- **Designers**
  - Ted Hoff, Stan Mazor, Hal Feeney, Federico Faggin
- **Data**
  - Word width: 8-bit
  - Clock: 800KHz
  - 3500 transistors
  - 48 instructions
  - Registers: 6 x 8-bit
  - Stack: 17 x 7-bit
  - Address space: 16KB
Intel 8080, 1974

- Lead designers
  - Federico Faggin (then to zilog), Masatoshi Shima, Stan Mazor
  - "The 8080 really created the microprocessor market"

- Used in MITS Altair 8800, 1975
  - “Microcomputer”
  - Also Intel Intellec-8

- Data
  - Word width: 8-bit
  - 4500 transistors
  - Clock: 2M-3MHz
  - Address space: 64KB
  - Registers: 6 x 8-bit
  - IO ports, Stack pointer

A follow up: 8085
Intel 16-bit Microprocessors

• Intel 8086, 1978 - first x86 family microprocessor
  – Source compatibility with 80xx lines – business win
  – 16-bit: all registers, internal and external buses
  – 29,000 transistors, 5MHz initially
  – 20-bit address bus - 4MB address space
    • 16-bit register - segmentation programming
• IBM PC selected 8088, 1981
• Intel 80286, 1982
  – 134,000 transistors, 6M-8MHz initially (0.21 IPC)
    • 10MHz → 1.5MIPS
  – Used by IBM PC/AT, 1984
  – Designed for multi-tasking with MMU “protection mode”

Then Microsoft and IBM split
Intel iAPX432, 1981

• Intel i432, Intel first 32-bit microprocessor design
  – “Intel Advanced Processor architecture”
  – Started in 1975 as the 8800, follow-on to the existing 8008 and 8080 CPUs
  – Intended purely 32-bit, to be Intel backbone in the 1980s, to support Ada, LISP, advanced computations
    • Micro-mainframe
  – HW supports to all the good terms
    • OO programming and capability-based addressing, Edsger Dijkstra’s on-the-fly parallel GC, multi-tasking and IPC, Multiprocessing, Fault tolerance, I/O
  – Problems: two-chip impl., lack of cache, bit-aligned var-len instructions, Ada compiler
  – Failed: ¼ performance of 286 as of 1982

2013/02/10
Brief history of Intel CPU uArch - xiaofeng.li@gmail.com
Intel x87 Family

• Intel 8087, 1980
  – First floating-point coprocessor for 8086 lines
  – Performance: +20% ~ 5x; 50,000 FLOPS
  – Floating registers form 8-level stack: st0~st7
  – 8-bit/16-bit
  – IEEE 754
• Intel 80287 – 16-bit
• Intel 80387, 80487 – 32-bit
• Starting from Intel 80486DX, Pentium and later has on-chip floating point unit
  – “DX” was used for on-chip FP capability
Intel 80386, 1985

- Intel first X86 32-bit flat memory model – 4GB space
  - 80386 instruction set, programming model, and binary encodings are the common denominator for all → IA-32, i386, x86
  - Paging to support VM, hardware debugging, first use of pipeline
  - Not necessarily a big performance improvement over 286
  - 275,000 transistors
  - 12MHz initially, later 33MHz → 11.4MIPS

- Compaq: first PC using 386, legitimize PC “clone” industry
- Andy Grove decided to single-source producing 386
  - Later changed in 1991 by AMD AM386
- Chief architect: John H. Crawford
Intel i960, 1985

• **Intel 80960, Intel first RISC microprocessor**
  – Best-selling embedded microcontroller at the time
  – After BiiN project, which was for high-end high-reliability processor jointly with Siemens
    • In response to i432 failure, avoid i432 problems
    • But, “Billions Invested In Nothing”
  – Lead: Glenford Myers
    • Intended to replace 80286/i386, and for UNIX systems (e.g., NeXT)
    • Removed all the “advanced” features of BiiN
    • Used Berkeley RISC (vs. Stanford), flat memory model, superscalar
  – Dropped after acquiring StrongARM in late 90’s
    • Price/perf/power no longer competitive
    • Team went to design another i386 processor – P6
Intel 80486, 1989

- Improvements
  - Atomic instructions
  - On-die 8KB SRAM cache
  - Tightly coupled pipelining: 1 IPC
    - 50MHz → 40MIPS on average and 50MIPS at peak
  - Integrated FPU (no longer need x87)
  - First chip exceeds 1M transistors

- Gaming is critical
  - 486 ended DOS games (Later, 3D ended 486)

- More manufacturers, AMD Am5x86, Cyrix Cx5x86, etc.

- Competitor
  - Motorola 68040 in Macintosh Quadra
Intel i860, 1989

• Entirely new RISC microprocessor
  – VLIW and high-performance FP operations
    • 32-bit ALU core, and 64-bit FPU (adder, multiplier, GPU)
    • Register sets: 32 x 32-bit integer, 16 x 64-bit FP
    • GPU uses FP registers as 8 x 128-bit, with SIMD (Influenced MMX)
    • 64/128-bit buses, fetch 2 x 32-bit instructions
  
• Dropped in mid-90’s
  – Compiler support was mission impossible
  – Context switch took 62 - 2000 cycles \(\rightarrow\) Unacceptable for GPCPU
  – Incompatible with X86, confusing the market with Intel 486 CISC

• Used in some parallel computers, graphic workstations
  – Windows NT (N-Ten) originally developed for i860 N10
  – NeXT, SGI, etc. used it as gfx accelerator
Intel Pentium, 1993

• Pentium means “5”, because court disallowed number-based trademark
  – Later “Pentium” was used in many Intel processors, no longer an micro-architecture branding – vs. “Celeron”

• P5 micro-architecture
  – First X86 superscalar micro-architecture
    • Dual integer pipelines, separate D/I caches, 64-bit external data-bus
  – 60M-300MHz (75 MHz → 126.5 MIPS)
    • 60/66MHz 0.8um in 5v called “coffee warmer”
  – Competitors
    • X86: AMD K5/K6, Cyrix 6x86, etc.  Risc: M68060, PPC601, SPARC, MIPS, Alpha

• Pentium Overdrive package
  – Started to use a cooler
Intel MMX, 1996

• SIMD instruction set, introduced with P5
  – “Matrix Math Extensions”, mainly for graphics
  – 8 x 64-bit integer registers MM0 ~ MM7, alias of FPU ST0 ~ ST7
  – But Integer-only was not enough soon due to gfx cards
  – AMD 3DNow! in K6-2, 1998
    • Introduced single-precision FP
  – Intel introduced SSE, 1999
    • Started with Pentium-III
    • New XMM register set
    • 70 new instructions

• MMX in Xscale
  – iwMMXt : "Intel Wireless MMX Technology"
Intel Pentium Pro, 1995

• P6 (or i686), completely new apart from Pentium (P5)
  • #transistors: Pentium 3.1M, Pentium MMX 4.5M, Pentium Pro 5.5M
    – Out-of-order execution
      • Speculative execution, RISC-like micro-ops
      • Three pipelines, 2 integer, 1 fp
    – Innovative on-package level-2 cache
      • Manufacturing did allow on-die L2 cache
      • Same CPU clock rate, non-blocking, SMP advantage
      • Dies had to be bonded early → Low yield rate and high price
    – 36-bit address bus (PAE). 16-bit performance was low
    – Performance better than best RISC with SPECint95, but only about half with SPECfp95
Intel P6 Processors (cont.)

• **Pentium II, 1997, 7.5M transistors**
  – Slot replaced Socket with a daughterboard
    • Solved the issues of off-package L2 cache in PPro with half CPU clock
  – Implemented MMX, improved 16-bit performance
  – Celeron and Xeon, 1998
    • Celeron: no on-die L2-cache, 66MT/s FSB
      – To win low-end and to justify Xeon
    • Pentium II Xeon: L2-cache, 100MT/s, SMP

• **Pentium III, 1999**
  – Introduced SSE for FP and vector processing
  – On-die L2 cache with .18um Coppermine
  – PSN (Processor Serial Number) controversy
Intel SSE

- Intel Streaming SIMD Extensions, 1999 in PIII
  - MMX uses FP registers for SIMD data, and has only integer SIMD
  - SSE introduces separate XMM registers

**Intel SSE and Extensions**

- **SSE**
  - All on XMM, making MMX redundant
  - Pack/Unpack double-precise FP
  - Integer arithmetic

- **SSE2**
  - DSP-oriented support
  - Packed AddSub FP
  - Horizontally computation
  - Monitor/Mwait
  - Complex number support
  - Low overhead unaligned load
  - Multiply&add, Multiply&Round/Scale
  - Packed AddSub DWORDS
  - Packed align/sign/abs
  - Byte level shuffle
  - Advanced String Operations
  - Fast CRC
  - POPCNT

- **SSE3**
  - packed DWORD and QWORD arithmetic
  - Blending
  - Sums of absolute differences
  - Dot for AOS (Array of Structs) data
  - Packed Integer Min and Max
  - Floating Point Round
  - Register Insertion/Extraction
  - Packed Format Conversion
  - Packed Test and Set, Compare for Equal

- **SSSE3**
  - 256 bit

- **SSE4.1**
  - Up to 256-bit wide vector FP data
  - 3 and 4 operands support
  - Power efficient

- **SSE4.2**

- **AVX**
  - 256 bit

**Timeline**

- 1999: SSE
- 2000: SSE2
- 2001: SSE3
- 2005: SSSE3
- 2007: SSE4.1
- 2011: SSE4.2
- 2013: AVX

**Contact**

xiaofeng.li@gmail.com
Intel Xscale

• Intel acquired StrongARM from DEC, 1997
  – To replace the RISC processors i860 and i960
  – StrongARM implemented ARMv4 ISA
• Successor, Xscale implemented ARMv5
  – Seven-stage integer and an eight-stage memory superpipelined microarchitecture, 32KB data cache and 32KB instruction cache
• Xscale processor family
  – Application Processors (with the prefix PXA)
  – I/O Processors (with the prefix IOP)
  – Network Processors (with the prefix IXP)
  – Control Plane Processors (with the prefix IXC).
  – Consumer Electronics Processors (with the prefix CE)
• Intel sold Xscale PXA business to Marvell, 2006
Intel Itanium, 2001

• Originated from HP
  – EPIC: explicitly parallel instruction computing
  – 1994, worked with Intel on IA-64, to release product in 1998
  – All believed EPIC would supplant RISC and CISC
    • Compaq and SGI gave up Alpha and MIPS
    • Microsoft and SUN etc developed OSes for it
  – 1999, Intel named it Itanium

• Data
  – Speculation, prediction, predication, and renaming
  – 128 integer registers, 128 FP registers, 64 one-bit predicates, and eight branch registers
  – 128-bit instruction word has 3 insns, dual-issue, max 6 IPC
  – X86 support in HW initially and then purely in SW
Intel Pentium 4, 2000

• NetBurst microarchitecture (P68, successor to P6)
  – Pursue higher frequency, smaller IPC
    • Hyper Pipelined: 20-stage Willamette, 31-stage Prescott (vs. 10 in P6)
    • Rapid Execution Engine: Two ALUs in the core are double-pumped
    • Execution Trace Cache, SSE2, L3-cache (Extreme Edition)
    • Hyper-Threading Technology
  – Prescott: 90nm, SSE3, HT, Intel-64 (64-bit), 2004
    • But performance worse than Northwood with similar clock
    • Designed to be 10GHz, only achieved 3.8GHz
  – Pentium D: Dual-core Pentium4, 2005

• Abandoned in 2006:
  – High power consumption and heat intensity
  – Inability to increase clock speed, and inefficient pipeline
Intel 64

- Intel implementation of X86-64, the 64-bit extension of X86 ISA
  - AMD released spec in 2000, and first implementation in 2003, as a response to Itanium (was IA-64)
    - Intel adopted X86-64 due to AMD’s success over Itanium, released first X86-64 processor in 2004
    - Different names: AMD64 (official AMD name), Intel 64 (official Intel name), X86-64 or X64 (community names), etc.
  - Maintains 32-bit mode binary compatibility

- 64bit vs. 32bit
  - Bigger virtual space, wider operation, more registers
  - Not necessarily better performance, usually bigger code size

- X32: an ABI, not ISA, nor processor mode
  - 64-bit mode process with instructions encoding 32-bit address
Intel Pentium M, 2003

• From Pentium III, based on P6 uArch
  – FSB interface of Pentium 4, SSE2, much larger cache, improved decoding/issuing FE
    • L2 cache only switches on the portion being accessed
  – SpeedStep 3 tech, TDP: 5-27W
    • Dynamically variable clock frequency and core voltage
  – 1.6 GHz Pentium M performance > 2.4 GHz Pentium 4-M

• Next generation released as Intel Core brand, Jan 2006
  – Core Duo used in Macbook Pro, Core Solo in Mac Mini

• Core 2: Intel-64 Core uarch, July 2006
  – Larger cache, SSE4.1 in 45nm
  – Solo, Duo, Quad, Extreme

• No HT, no L3 cache, mostly
Intel Tick-Tock Model

- Introduced since 2007 to describe progress cadence
  - “Tick”: shrinking of process technology – same uArch
  - “Tock”: new microarchitecture – same process
  - Tick-Tock is expected alternating every year
    - Not really matched in reality though

<table>
<thead>
<tr>
<th>Architectural change</th>
<th>Codename</th>
<th>uArch</th>
<th>Process</th>
<th>Release date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tick</td>
<td>New Process</td>
<td></td>
<td>65 nm</td>
<td>Jan 5, 2006</td>
</tr>
<tr>
<td>Tock</td>
<td>New uArch</td>
<td>Conroe Core</td>
<td></td>
<td>July 27, 2006</td>
</tr>
<tr>
<td>Tock</td>
<td>New uArch</td>
<td>Nehalem Nehalem</td>
<td></td>
<td>Nov 17, 2008</td>
</tr>
<tr>
<td>Tick</td>
<td>New Process</td>
<td>Westmere</td>
<td>32 nm</td>
<td>Jan 4, 2010</td>
</tr>
<tr>
<td>Tock</td>
<td>New uArch</td>
<td>Sandy Bridge Sandy Bridge</td>
<td></td>
<td>Jan 9, 2011</td>
</tr>
<tr>
<td>Tick</td>
<td>New Process</td>
<td>Ivy Bridge</td>
<td>22 nm</td>
<td>2012</td>
</tr>
<tr>
<td>Tock</td>
<td>New uArch</td>
<td>Haswell Haswell</td>
<td></td>
<td>2013</td>
</tr>
<tr>
<td>Tick</td>
<td>New Process</td>
<td>Broadwell</td>
<td>14 nm</td>
<td>2014</td>
</tr>
<tr>
<td>Tock</td>
<td>New uArch</td>
<td>Skylake Skylake</td>
<td></td>
<td>2015</td>
</tr>
<tr>
<td>Tick</td>
<td>New Process</td>
<td>Cannonlake</td>
<td>10 nm</td>
<td>2017</td>
</tr>
<tr>
<td>Tock</td>
<td>New uArch</td>
<td>Ice Lake Ice Lake</td>
<td></td>
<td>2018</td>
</tr>
</tbody>
</table>
Intel Nehalem, 2008

• Successor of Core micro-architecture
  – Was planned as Netburst evolution, but then a completely different design of microarchitecture, 45nm

• Data
  – Multi-core, on-package GPU
  – Integrated memory controller, QPI replaced FSB
  – Integrated PCI-E and DMI replacing northbridge
  – HT, and shared L3 cache, 2nd-level branch predictor and TLB
  – SSE4.2, atomic overhead is reduced by 50%
  – Over Penryn, 20% gain performance/clock, 30% cut power/performance
  – Core i3, i5, i7, Celeron, Pentium, Xeon

• Tick: Westmere, 32nm
  – AES-NI, integrated graphics, VT 16-bit guest, 1GB page
Intel Atom Processors, 2008

- Based on Bonnell microarchitecture, 45nm
  - Dual-issue in order, 16-stage pipeline
  - On/off: SSEx, Intel-64, HT
  - TDP: \textit{n} watt
  - Only around 4\% of instructions produce multiple micro-ops
    - Significantly fewer than the P6 and NetBurst microarchitectures
    - Can contain both a load and a store with an ALU operation
    - Partial revival of old principle in P5 and 486 for perf/watt
  - For mobile and embedded devices
- Tick: Saltwell, 32nm, 2011
Accelerating to SoC
Intel Sandy Bridge, 2011

• New microarchitecture after Nehalem, 32nm
  – Shared L3 cache for cores, including GPU
  – Two load/store ops/cycle for memory channel
  – Ring bus interconnect between Cores, Graphics, Cache and System Agent Domain
  – AVX
  – Compared to Nehalem, 17% gain in performance/clock over Lynnfield, 2x graphics over Clarkdale

• Tick: Ivy Bridge, 22nm, 2012
  – 3D gates (tri-gate transistor)
# Pipeline Stages

<table>
<thead>
<tr>
<th>Microarchitecture</th>
<th>Pipeline stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>P5 (Pentium)</td>
<td>5</td>
</tr>
<tr>
<td>P6 (Pentium Pro)</td>
<td>14</td>
</tr>
<tr>
<td>P6 (Pentium 3)</td>
<td>10</td>
</tr>
<tr>
<td>NetBurst (Willamette)</td>
<td>20</td>
</tr>
<tr>
<td>NetBurst (Northwood)</td>
<td>20</td>
</tr>
<tr>
<td>NetBurst (Prescott)</td>
<td>31</td>
</tr>
<tr>
<td>NetBurst (Cedar Mill)</td>
<td>31</td>
</tr>
<tr>
<td>Core/NHM/SNB/HSW</td>
<td>14</td>
</tr>
<tr>
<td>Atom Bonnell</td>
<td>16</td>
</tr>
<tr>
<td>Silvermont/Airmont</td>
<td></td>
</tr>
</tbody>
</table>

---

2013/02/10

Brief history of Intel CPU uArch - xiaofeng.li@gmail.com
References