Intel Skylake

- App Req vs Tech Constraints
- Skylake System Overview
- Skylake Processor
- Skylake Memory
- Skylake Network
- Skylake System Manager
Application Requirements:
Low-Power, High-Performance, Scalable Design

Converged core: Single microarchitecture that scales from tablet to server

Performance
- Legacy Code Performance Improvements
- New Technologies to Extract Greater Parallelism

Modularity
- Increased power/performance range
- Greater number of supported products
- Support for SoC designs

Power Innovations
- Active Power Reduction
- Idle Power Reduction
- Focused on Full Platform, not just CPU

Goal: Achieve new levels of power reduction without compromising performance
Technology Constraints: Power Consumption
Technology Constraints: New Devices

- **2003**: 90 nm
  - Invented SiGe Strained Silicon

- **2005**: 65 nm
  - 2nd Gen. SiGe Strained Silicon

- **2007**: 45 nm
  - Invented Gate-Last High-k Metal Gate

- **2009**: 32 nm
  - 2nd Gen. Gate-Last High-k Metal Gate

- **2011**: 22 nm
  - First to Implement Tri-Gate

---

**Strained Silicon**

**High-k Metal Gate**

**Tri-Gate**
Traditional Planar Transistors

Silicon Substrate

Oxide

Gate

Source

Drain

High-k Dielectric

2D planar transistors form a conducting channel in the silicon region under the gate electrode.

3D tri-gate transistors form conducting channels on three sides of a vertical fin structure.

Tri-Gate Transistors
Tri-Gate Transistors

32 nm Planar Transistors

22 nm Tri-Gate Transistors
Intel Skylake

- App Req vs Tech Constraints
- **Skylake System Overview**
- Skylake Processor
- Skylake Memory
- Skylake Network
- Skylake System Manager
Skylake System

Intel® Core™ i7-6700K Processor
Intel® HD Graphics 530

Intel® Z170 Chipset

DMI 3.0
8 Gb/s each x1

DDDR4/DDR3L Up to 2133/1600 MHz

DDDR4/DDR3L Up to 2133/1600 MHz

Intel® High Definition Audio

6x SATA ports, eSATA; Port Disable

Intel® Rapid Storage Technology for PCI Express Storage

Intel® Smart Sound Technology

SPI

Up to 6 Gb/s

Intel® Integrated 10/100/1000 MAC

USB 3.0 Ports
14x USB 2.0 Ports
XHCI; USB Port Disable

Three Independent Displays Support

Up to 20x PCI Express* 3.0

PCI Express* 3.0 Graphics

1x16 lanes

2x8 lanes

1x8 and 2x4 lanes

PCle x1

SMBus

Intel® Ethernet Connection

Intel® ME 11 Firmware and BIOS Support

Intel® Extreme Tuning Utility Support

Intel® Device Protection Technology with Boot Guard
• ~1.7B transistors in ~122 mm² in a 22nm tri-gate process
• Four out-of-order cores each with two SMT threads running at 4.0-4.2 GHz
• Three-level cache hierarchy with last-level on-chip cache capacity of 8MB
• Max thermal design power of 91W
• 2 DDR4 DRAM memory controllers, 34.1 GB/s max memory bandwidth
• Integrated 3D graphics processor running at 350 MHz to 1.15 GHz
• Pipelined bus on-chip network connecting cores, last-level cache banks, and GPU
Intel Skylake: Block Diagram
### Design-Time Modularity to Meet Scalability Application Requirement

<table>
<thead>
<tr>
<th>2 in 1 Detachables, Tablets and Compute Stick</th>
<th>Thin Light Notebooks, Portable AIO, Minis and Conference Room</th>
<th>Ultimate Mobile Performance, Mobile Workstations</th>
<th>Desktop Performance to Value, AIO and Minis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Y-SERIES</strong></td>
<td><strong>U-SERIES</strong></td>
<td><strong>H-SERIES</strong></td>
<td><strong>S-SERIES</strong></td>
</tr>
<tr>
<td>5 Dies 4 Packages</td>
<td>5 Dies 4 Packages</td>
<td>5 Dies 4 Packages</td>
<td>5 Dies 4 Packages</td>
</tr>
<tr>
<td><strong>Dies</strong></td>
<td>2+2</td>
<td>2+2</td>
<td>2+2</td>
</tr>
<tr>
<td><strong>Package (mm)</strong></td>
<td>BGA 1515</td>
<td>BGA 1356</td>
<td>BGA 1440</td>
</tr>
<tr>
<td></td>
<td>20 x 16.5</td>
<td>42 x 24</td>
<td>42 x 28</td>
</tr>
<tr>
<td><strong>TDP (W)</strong></td>
<td>4.5</td>
<td>15</td>
<td>15, 28</td>
</tr>
<tr>
<td><strong>Chipset</strong></td>
<td>Integrated 6th Gen Intel® Core™ Platform I/O</td>
<td>Intel® 100 Series chipset (23mm x 23mm)</td>
<td></td>
</tr>
</tbody>
</table>
Scalable Design in Haswell Microarchitecture

Intel Haswell i7-4770K
85W @ 3.5 GHz

Intel Haswell 3560Y
6W @ 880 MHz
Intel Skylake

- App Req vs Tech Constraints
- Skylake System Overview
- **Skylake Processor**
- Skylake Memory
- Skylake Network
- Skylake System Manager
Growth in Instruction Sets Over Time

Three key additions in Skylake

- AVX512: 512-bit SIMD Extensions
- SGX: Software Guarded Execution
- MPX: Memory Protection Extensions
AVX512: 512-bit SIMD Extensions
SGX: Software Guarded Extensions

CPU instructions used by applications to protect critical secrets from unauthorized access:

- Against software attacks originated at any privilege level
- Against many hardware based attacks

Applications are modified (split) into trusted and untrusted parts:

- Trusted part of application is protected via encryption by Intel hardware
- Intel® Software Guard Extensions (Intel® SGX) does not protect untrusted part of application OS support
- Intel plans to enable Intel SGX on Windows® 7 and 8.x platforms
- Intel is collaborating with Microsoft® on native support in future release of Windows operating system
MPX: Memory Protection Extensions

What happens if we use $A[128]$? 

Assume array element address is in $r1$

```
BNDCU $r00, 0(r1)  # check address vs. upper bound
BNDCU $r00, 0(r1)  # check address vs. upper bound
```

If the check fails cause an exception
IO Fetch

~5 cycles

IO Decode

~14 cycles

OOO Issue and Late Commit

Integer/FP Functional Units with OOO Writeback

Load/Store Execution

Branch Pred → Instruction Fetch Unit

L1 ITLB → 32KB L1 I$ (8-way)

>20B

16B Predecode, Fetch Buf

6 x86 Instructions

2x20 Instruction Queue

5 x86 Instructions

μcode store

Complex Decode

Simple Decode

Simple Decode

Simple Decode

Simple Decode

Simple Decode

1.5K μop Cache

2x64 μop Decode Queue

5 μops

6 μops

Retire Unit

224 Entry ROB

180 Integer Registers

168 FP Registers

48 Entry BR Order Buffer

72 Entry Load Buffer

56 Entry Store Buffer

2x4 μops

97 Entry Unified Scheduler

Port0 → ALU Branch Shift

Port1 → ALU LEA

Port5 → ALU LEA

Port6 → ALU Branch Shift

Port2 → Load Store Addr

Port3 → Load Store Addr

Port4 → Store Addr

Port7 → Store Data

L1 DTLB → 32KB L1 D$ (8-way)

32B

L2 DTLB

256KB L2 Cache (8-way)
IO Fetch and Decode

- Complex CISC instructions are broken into much simpler, almost RISC-like, micro-ops
- Predecoder handles variable length encoding (1-15B), finds x86 instruction boundaries and inserts into instruction queue
- Parallel decoders are used to transform x86 instructions into uops; can decode either five “simple” x86 instructions (decodes into 1 uop) per cycle or one “complex” x86 instruction (decodes into 1-4 fused uops)
- Very complex instructions fall back to a microcoded control unit
- uop cache acts as a kind of L0 instruction cache that holds decoded uops and enables much of the front-end to be shut down to save power
- uop decode queue can be used as a special loop cache
IO Fetch and Decode

- Skylake predictor has changed but little is known about it; more is known about Sandy Bridge (2gen old)
- Sandy Bridge predictor has a misprediction latency of \(~15\) cycles for branches in uop cache
- Sandy Bridge predictor uses a “two-level predictor with 32b global history buffer and a history pattern table of unknown size”
- Sandy Bridge uses a BTB for both L1 I$ and uop cache; “conditional jumps are less efficient if there are more than 3 branch instructions per 16 bytes of code”
- Sandy Bridge uses a return address stack predictor with 16 entries
• **Integer/FP Registers** are the physical registers used for register renaming

• **Load Buffer** and **Store Buffer** are the finished load/store buffers

• **Branch Order Buffer** is used to store snapshots of the rename tables to recover from mispredicted branches

• **Unified Scheduler** is a centralized issue queue

• Can rename and insert into the IQ up to six fused uops per cycle; can commit up to four uops per thread per cycle; since fused uop can encode two uops peak throughput is eight uops/cycle
Can “issue” (dispatch) up to eight instructions per cycle to eight “dispatch ports”, which is just several arithmetic units collected into a functional unit.

“Every cycle, the 8 oldest, non-conflicting uops that are ready for execution are sent from the unified scheduler to the dispatch ports.”

“Execution units are arranged into stacks: integer, SIMD integer, and floating point. ... Each stack has different data types, different result forwarding networks, and potentially different registers.”

“Note that the divider on port 0 is not fully pipelined and is shared by all types of uops (integer, SIMD integer, and floating point)”
The new CPU increases the load/store bandwidth to 128 bytes per cycle. This bandwidth requires using the new AVX512 technology.

Haswell's microarchitecture also boosts instruction throughput by increasing the store buffer. The new CPU uses the same pipeline as its predecessors, including the same clock speed. Some minor tuning allows Skylake to achieve the peak performance of the CPU. For example, it increases the number of microarchitectural units, which are grouped into eight ports (not shown). The wider front end floods the CPU with instructions while the wider back end retires them faster. The wider back end also increases the number of instructions per cycle, a 50% improvement from when some instructions were retired at four per thread basis.

As in Haswell, Skylake implements a unified scheduler, a decoupling queue between front-end and back-end, and a unified allocation queue in Haswell on a per-thread basis. The unified scheduler in the front end reduces stalls. Skylake splits the unified scheduler into two execution resources, which are grouped into eight ports (not shown). Skylake also increases the number of microarchitectural units and two dispatch ports to the reorder buffer (ROB) size from 192 to 224 entries, as Table 1 shows.

Table 1: Size of Data Structures

<table>
<thead>
<tr>
<th></th>
<th>Nehalem</th>
<th>Sandy Bridge</th>
<th>Haswell</th>
<th>Skylake</th>
</tr>
</thead>
<tbody>
<tr>
<td>x86 Decoders</td>
<td>4 instr</td>
<td>4 instr</td>
<td>4 instr</td>
<td>5 instr</td>
</tr>
<tr>
<td>Max Instr/Cycle</td>
<td>4 ops</td>
<td>6 ops</td>
<td>8 ops</td>
<td>8 ops</td>
</tr>
<tr>
<td>Reorder Buffer</td>
<td>128 ops</td>
<td>168 ops</td>
<td>192 ops</td>
<td>224 ops</td>
</tr>
<tr>
<td>Load Buffer</td>
<td>48 loads</td>
<td>64 loads</td>
<td>72 loads</td>
<td>72 loads</td>
</tr>
<tr>
<td>Store Buffer</td>
<td>32 stores</td>
<td>36 stores</td>
<td>42 stores</td>
<td>56 stores</td>
</tr>
<tr>
<td>Scheduler</td>
<td>36 entries</td>
<td>54 entries</td>
<td>60 entries</td>
<td>97 entries</td>
</tr>
<tr>
<td>Integer Rename</td>
<td>In ROB</td>
<td>160 regs</td>
<td>168 regs</td>
<td>180 regs</td>
</tr>
<tr>
<td>FP Rename</td>
<td>In ROB</td>
<td>144 regs</td>
<td>168 regs</td>
<td>168 regs</td>
</tr>
<tr>
<td>Allocation Queue</td>
<td>28/thread</td>
<td>28/thread</td>
<td>56 total</td>
<td>64/thread</td>
</tr>
</tbody>
</table>

- **Reorder Buffer** is the number of entries in the reorder buffer (ROB)
- **Load Buffer** is the number of entries in the finished store buffer (FLB)
- **Store Buffer** is the number of entries in the finished store buffer (FSB)
- **Scheduler** is the number of entries in the centralized issue queue (IQ)
- **Integer/FP Rename** is the number of physical integer and floating point registers
- **Allocation Queue** is a decoupling queue between front-end and back-end
- Nehalem to Sandy Bridge transitioned from value- to pointer-based register renaming
Maco-Op Fusion

Combines a compare x86 instruction and a jump x86 instruction into a single micro-op for the entire pipeline.

```assembly
cmp eax, ecx
jl  loop
```

- Only works for specific versions of comparison and jump instructions
- There can be no other instructions between the compare and jump instructions
- Both instructions must be in a single 16-byte aligned block

Micro-Op Fusion

Combines two micro-ops together (load plus integer op, split stores) so that they only take a single ROB and IQ entry, but the fused micro-op is split such that two micro-ops are issued to two different execution units.

```assembly
mov [esi], eax ; 1 fused uop
add eax, [esi] ; 1 fused uop
add [esi], eax ; 2 uops + 1 fused
```

- Decoding becomes more efficient, because instructions that generate one fused uop can use the simple decoders
- Reduces pressure on register renaming and commit pipeline stages
- Capacity of IQ and ROB are increased since fused uop only uses one entry
**Move Elimination**

Moving a value from one register to another does not require any “real” work. Simply update the rename table so r3 now points to the same physical register as r1. Only perform move elimination if r1 is ready.

**Zero Idiom**

Zero’ing out a register is very common but requires very little “real” work.

```
xor eax, ecx
```

Allocate a fresh destination register in the rename stage, but then immediately clear the value in this destination register to zero.

Both techniques enable specific instructions to use no execution resources!
Multithreading & SIMD

- **SMT** enables two threads to share much of the OOO pipeline, although some data-structures are statically partitioned between the two threads.

- **Subword-SIMD** can process 512 bits of integer or floating-point data with a single instruction, where this data is carved into 64x8b, 32x16b, 16x32b, or 8x64b.
Intel Skylake

- App Req vs Tech Constraints
- Skylake System Overview
- Skylake Processor
- **Skylake Memory**
- Skylake Network
- Skylake System Manager
Memory System

- Skylake can sustain two loads and one store of **512b per cycle**
- Uses split stores with the store address generation uop being sent to the Store AGU execution unit and the store data being sent to a separate execution unit.
- L1 DTLB: “There are 64, 32, and 4 entries respectively for 4KB, 2MB, and 1GB pages, all the translation arrays are still 4-way associative.”
- “Misses in the L1 DTLB are serviced by the unified L2 TLB” which has 1024 entries and is 8-way associative.
“A dedicated store AGU is slightly less expensive than a more general AGU. Store uops only need to write the address (and eventually data) into the store buffer. In contrast, load uops must write into the load buffer and also probe the store buffer to check for any forwarding or conflicts.”

- L2 can sustain refilling a complete 64B cache line into the L2 per cycle
- L2 is private to the core and is “neither inclusive nor exclusive of the L1 data cache.”
- L2 is non-blocking and sustain up to 16 outstanding misses
<table>
<thead>
<tr>
<th>Metric</th>
<th>Nehalem</th>
<th>Sandy Bridge</th>
<th>Haswell</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Instruction Cache</td>
<td>32K, 4-way</td>
<td>32K, 8-way</td>
<td>32K, 8-way</td>
</tr>
<tr>
<td>L1 Data Cache</td>
<td>32K, 8-way</td>
<td>32K, 8-way</td>
<td>32K, 8-way</td>
</tr>
<tr>
<td>Fastest Load-to-use</td>
<td>4 cycles</td>
<td>4 cycles</td>
<td>4 cycles</td>
</tr>
<tr>
<td>Load bandwidth</td>
<td>16 Bytes/cycle</td>
<td>32 Bytes/cycle (banked)</td>
<td>64 Bytes/cycle</td>
</tr>
<tr>
<td>Store bandwidth</td>
<td>16 Bytes/cycle</td>
<td>16 Bytes/cycle</td>
<td>32 Bytes/cycle</td>
</tr>
<tr>
<td>L2 Unified Cache</td>
<td>256K, 8-way</td>
<td>256K, 8-way</td>
<td>256K, 8-way</td>
</tr>
<tr>
<td>Fastest load-to-use</td>
<td>10 cycles</td>
<td>11 cycles</td>
<td>11 cycles</td>
</tr>
<tr>
<td>Bandwidth to L1</td>
<td>32 Bytes/cycle</td>
<td>32 Bytes/cycle</td>
<td>64 Bytes/cycle</td>
</tr>
<tr>
<td>L1 Instruction TLB</td>
<td>4K: 128, 4-way 2M/4M: 7/thread</td>
<td>4K: 128, 4-way 2M/4M: 8/thread</td>
<td>4K: 128, 4-way 2M/4M: 8/thread</td>
</tr>
<tr>
<td>L1 Data TLB</td>
<td>4K: 64, 4-way 2M/4M: 32, 4-way 1G: fractured</td>
<td>4K: 64, 4-way 2M/4M: 32, 4-way 1G: 4, 4-way</td>
<td>4K: 64, 4-way 2M/4M: 32, 4-way 1G: 4, 4-way</td>
</tr>
<tr>
<td>L2 Unified TLB</td>
<td>4K: 512, 4-way</td>
<td>4K: 512, 4-way</td>
<td>4K+2M shared: 1024, 8-way</td>
</tr>
</tbody>
</table>

**All caches use 64-byte lines**
Last-Level Cache

- **LLC shared** among all Cores, Graphics and Media
  - Graphics driver controls **which streams** are cached/coherent
  - **Any agent** can access all data in the LLC, independent of who allocated the line, after **memory range checks**
- Controlled LLC **way allocation** mechanism to prevent thrashing between Core/graphics
- Multiple coherency domains
  - **IA Domain** *(Fully coherent via cross-snoops)*
  - **Graphic domain** *(Graphics virtual caches, flushed to IA domain by graphics engine)*
  - **Non-Coherent domain** *(Display data, flushed to memory by graphics engine)*

**Much higher Graphics performance, DRAM power savings, more DRAM BW available for Cores**
In-Package Embedded DRAM L4 Cache
Intel Skylake

- App Req vs Tech Constraints
- Skylake System Overview
- Skylake Processor
- Skylake Memory
- **Skylake Network**
- Skylake System Manager
Pipelined Bus Interconnect

- **Ring-based** interconnect between Cores, Graphics, Last Level Cache (LLC) and System Agent domain
- Composed of **4 rings**
  - 32 Byte *Data* ring, *Request* ring, *Acknowledge* ring and *Snoop* ring
  - Fully pipelined at **core frequency/voltage**: bandwidth, latency and power scale with cores
- Massive ring **wire routing** runs over the LLC with no area impact
- Access on ring always picks the **shortest path** – minimize latency
- **Distributed arbitration**, sophisticated ring protocol to handle coherency, ordering, and core interface
- **Scalable to servers** with large number of processors

*High Bandwidth, Low Latency, Modular*
More Than Just Another Brickland in Intel's Wall

By David Kanter (February 24, 2014)

Although high-end server products are released infrequently, the impact of a new platform is great. Intel's unveiling of the Brickland platform is a perfect example. Like its predecessor, Brickland supports glueless four-socket and eight-socket servers, and it can be extended to larger systems using proprietary node controllers from third parties.

The cornerstone of Brickland is Ivy Bridge-EX, which will be billed as the Xeon E7 v2 family. Compared with the previous generation, Ivy Bridge-EX takes advantage of a new processor core and northbridge microarchitecture, new 22nm technology, and a new platform to deliver 1.7x to 2.3x higher performance on a variety of industry-standard benchmarks and customer applications. It also integrates PCIe 3.0 on the die and only uses up to 15% more power per socket.

Intel's mainstream server processors for one- and two-socket systems follow a fairly rapid cadence, but high-end and scalable server processors move at a slower march. For example, the company has introduced two new mainstream platforms and four different server processors since 2009. In contrast, the existing Boxboro-EX platform for high-end servers made its debut in 2010 along with the Xeon 7500 series (code-named Nehalem-EX) but only saw one upgrade: the Xeon E7 (code-named Westmere-EX) in 2Q11.

The Brickland platform brings comprehensive changes to high-end servers, as Figure 1 shows. The memory is still DDR3, but the bandwidth increases by 25% and, more importantly, the capacity triples. Intel changed the cache-coherence protocol from source snooping to a more complex and scalable home-snooping design to improve performance in four- and eight-socket systems. True to Moore's Law, Brickland shifts I/O from discrete chipsets to integrated PCIe links, substantially reducing system power, cost, and complexity. The processor and platform power
Two bidirectional rings with intermediate switches

Each ring has 32B (256b) channels

Total L3 capacity is 45MB

Ring and L3 operate at core frequency (2.5GHz)

Entire system is completely cache coherent

Supports transactional memory

$7,175!
One Die For Both Haswell-EP and -EX

Figure 2. Die micrograph of Haswell-EX/EP. Both the EX and EP versions use the same 662mm² chip. The high-core-count variants of Haswell-EP employ the same silicon as Haswell-EX, but with different I/O configurations. The latter chip enables more QPI links and fewer PCIe lanes, and it uses SMI2 memory interfaces. (Source: Intel)
Intel Skylake

- App Req vs Tech Constraints
- Skylake System Overview
- Skylake Processor
- Skylake Memory
- Skylake Network
- Skylake System Manager
Skylake Overview – Power Management View

- Skylake is a SoC consisting of:
  - 2-4 CPU cores, Graphics, media, Ring interconnect, cache
  - Integrated System Agent (SA)
  - On package PCH and eDRAM

- Improved performance with aggressive power savings

- Package Control Unit (PCU):
  - Power management logic and controller firmware
  - Continues tracking of internal statistics
  - Collects internal and external power telemetry: iMon, Psys
  - Interface to higher power management hierarchies: OS, BIOS, EC, graphics driver, DPTF, etc.
Skylake Power Management ID Card

- Up to four independent variable Power domains:
  - CPU cores & ring, PG slice, PG logic and SA
- Other fixed SoC and PCH voltage rails
- High granularity power gating
  - Partial and full core gating, Sub slice Graphics gating, System agent, cache, ring and package power off
- Shared frequency for all Intel® Architecture cores
- Independent frequencies for ring, PG slice & logic
- SA GV for improved performance and battery life

*Note: Not to scale*
Trading Off Energy vs Performance

- Total Energy = Compute + system energy
- Compute energy $\sim f^2$
- $P_e = P_{\text{Most Efficient}}$
- SoC and system energy $\sim 1/f$

Energy vs Performance graph with curves indicating energy and performance trade-offs.
Intel® Turbo Boost Technology 2.0 delivers intelligent and energy efficient performance on demand.

---

**Efficient.**
- Adapts by varying turbo frequency to conserve energy depending upon the type of instructions.

**Dynamic.**
- Boosts power level to achieve performance gains for high intensity "dynamic" workloads.

**Intelligent.**
- Power averaging algorithm manages power and thermal headroom to optimize performance.
Legacy Energy-performance Control (P-state)

- **DVFS – Intel SpeedStep® Technology**
  - $P \sim V^2 \cdot f \cdot C_{dyn} + \text{leakage}(V) \sim f^3$
  - Performance comes at a cost of energy

- **Operating System performs P-state control**
  - P1-Pn frequency table enumerated via ACPI tables
  - Explicit P-state selection

- **Typically demand based algorithm**
  - Policies (AC/DC/Balanced, etc.)
  - Non regular workloads are hard to manage
  - Lower than Pn is used for critical conditions only
Intel® Speed Shift Technology - Hardware P-state

- Why change:
  - Highly dynamic power – Multi core, AVX, accelerators
  - Small form factors → large turbo range
  - Smarter power management enables better choices
    ▪ Finer grain and micro architectural observability
- How:
  - Expose entire frequency range
  - A new deal - OS and hardware share power/perf. control
    ▪ OS direct control when and where desired
    ▪ Autonomous control by PCU elsewhere
In one of its largest product introductions ever, Intel rolled out 48 new PC processors based on its new Skylake CPU design. Skylake uses the same 14nm process as Broadwell but offers several improvements. Perhaps the most significant is a move to hardware power management, a feature Intel calls SpeedShift. The sixth-generation CPU has instruction-set extensions dubbed SGX and MPX that work with Windows 10 to improve security, and AVX512 doubles the processing speed for vector operations. Owing to delays in the 14nm ramp, only the Y-Series Broadwell launched in 2014; mainstream notebook models appeared in January of this year, and a few high-end desktop models rolled out a few months ago. Now that the 14nm process is in full production, Skylake is launching with a broad set of notebook and desktop chips, including dual- and quad-core versions. This debut marks the first update in two years for mainstream desktop processors, which entirely skipped the Broadwell generation.

For systems that are moving up from the 22nm Haswell, Skylake provides significant gains in either clock speed or power. Compared with Broadwell, however, the gains are minimal; these products see little or no clock-speed increase. We estimate the enhancements to the CPU design yield about 5% more performance per clock on most PC applications. Skylake's advances are mainly outside the CPU: the processor includes a new image-processing engine (ISP) and adds DDR4 and PCI Express 3.0 interfaces, as Figure 1 shows. Revisions to the GPU and video engine boost performance per watt.

Most of the newly announced products are available immediately, but the Iris and Pentium models will ship in the fourth quarter, along with new vPro products. By early 2016, the architecture will extend all the way to the low-end Celeron brand as well as the high-end Iris Pro line. For the first time, Intel is developing different versions of the new CPU, omitting AVX512 for PCs but including it for servers. The latter version of Skylake will appear in Xeon E5 and E7 server processors in 2016.