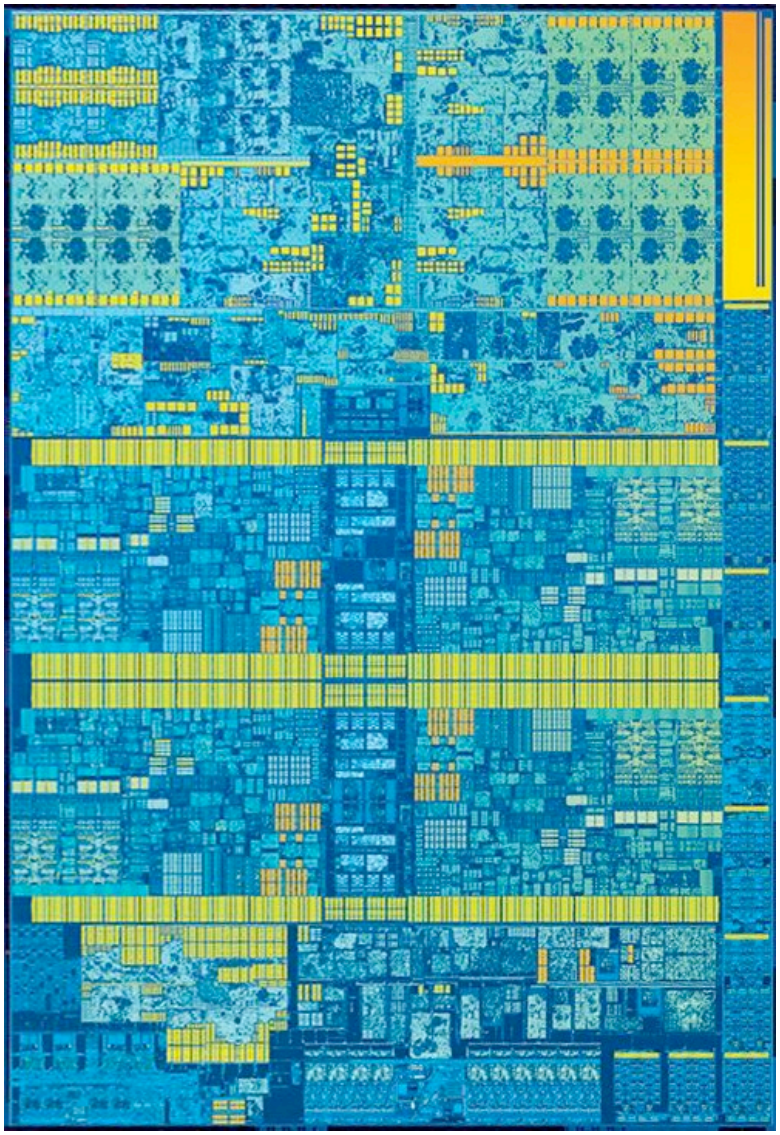
ECE 4750 Computer Architecture
Intel Skylake

Christopher Batten
School of Electrical and Computer Engineering
Cornell University

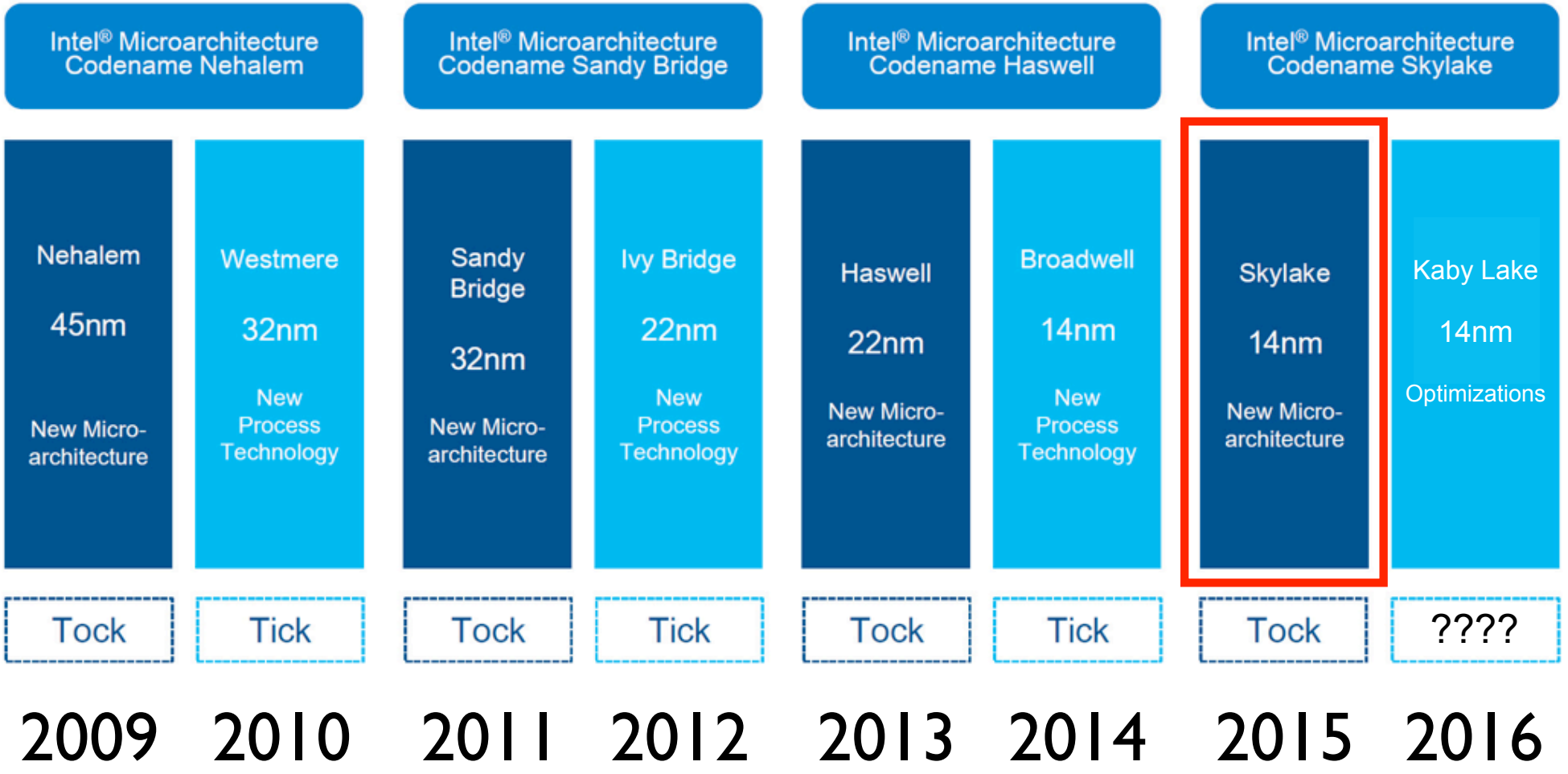
http://www.csl.cornell.edu/courses/ece4750



Intel Skylake

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

Intel Tick/Tock Product Releases



Application Requirements: Low-Power, High-Performance, Scalable Design

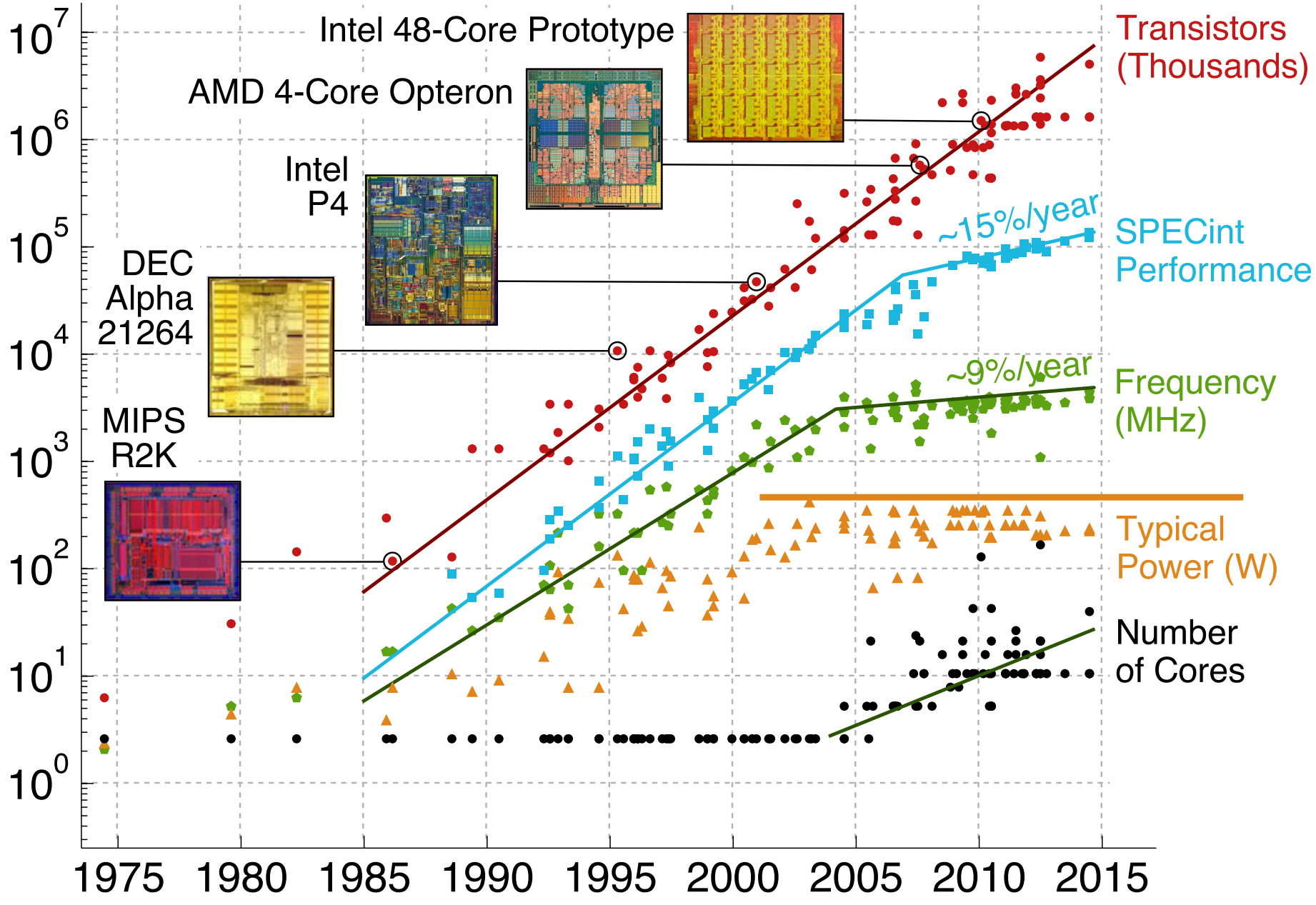
Client Implementations | Server Implementations

Converged core: Single microarchitecture that scales from tablet to server

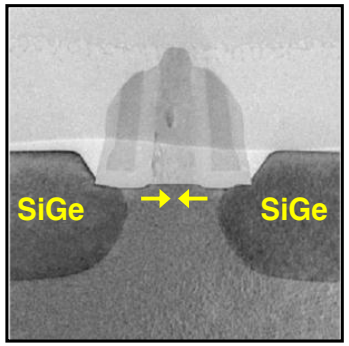
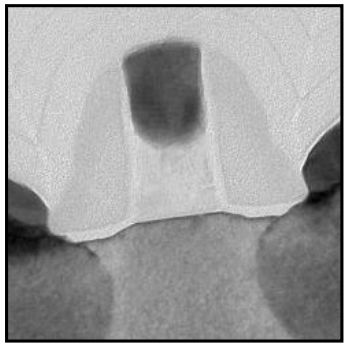
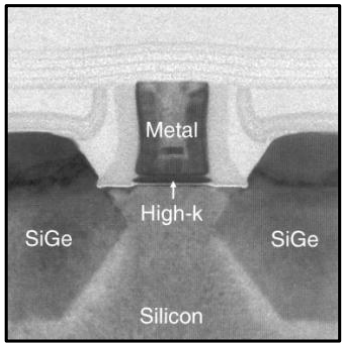
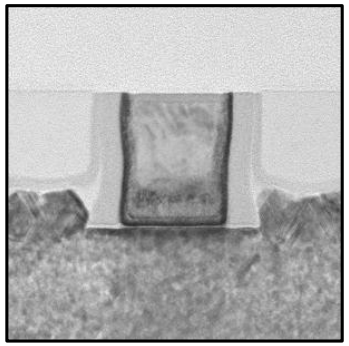
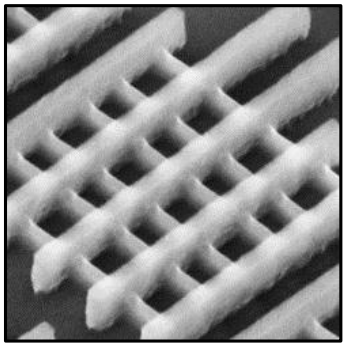
Performance	Modularity	Power Innovations
• Legacy Code Performance Improvements • New Technologies to Extract Greater Parallelism	• Increased power/performance range • Greater number of supported products • Support for SoC designs	• Active Power Reduction • Idle Power Reduction • Focused on Full Platform, not just CPU

Goal: Achieve new levels of power reduction without compromising performance

Technology Constraints: Power Consumption



Technology Constraints: New Devices

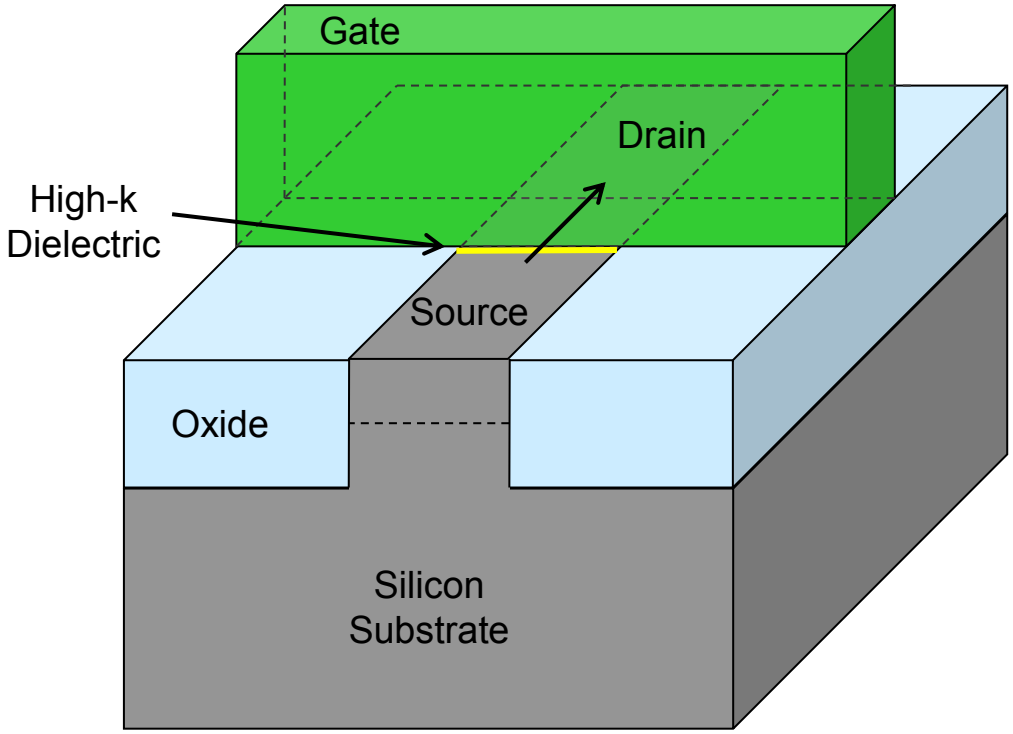
2003	2005	2007	2009	2011
90 nm	65 nm	45 nm	32 nm	22 nm
				
Invented SiGe Strained Silicon	2nd Gen. SiGe Strained Silicon	Invented Gate-Last High-k Metal Gate	2nd Gen. Gate-Last High-k Metal Gate	First to Implement Tri-Gate

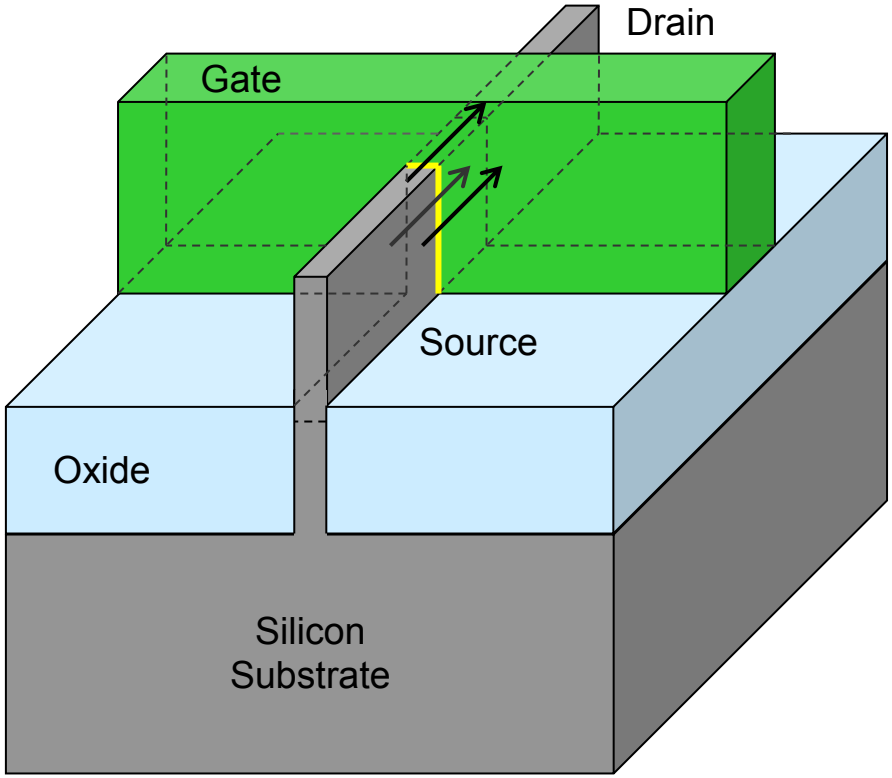
Strained Silicon

High-k Metal Gate

Tri-Gate

Tri-Gate Transistors



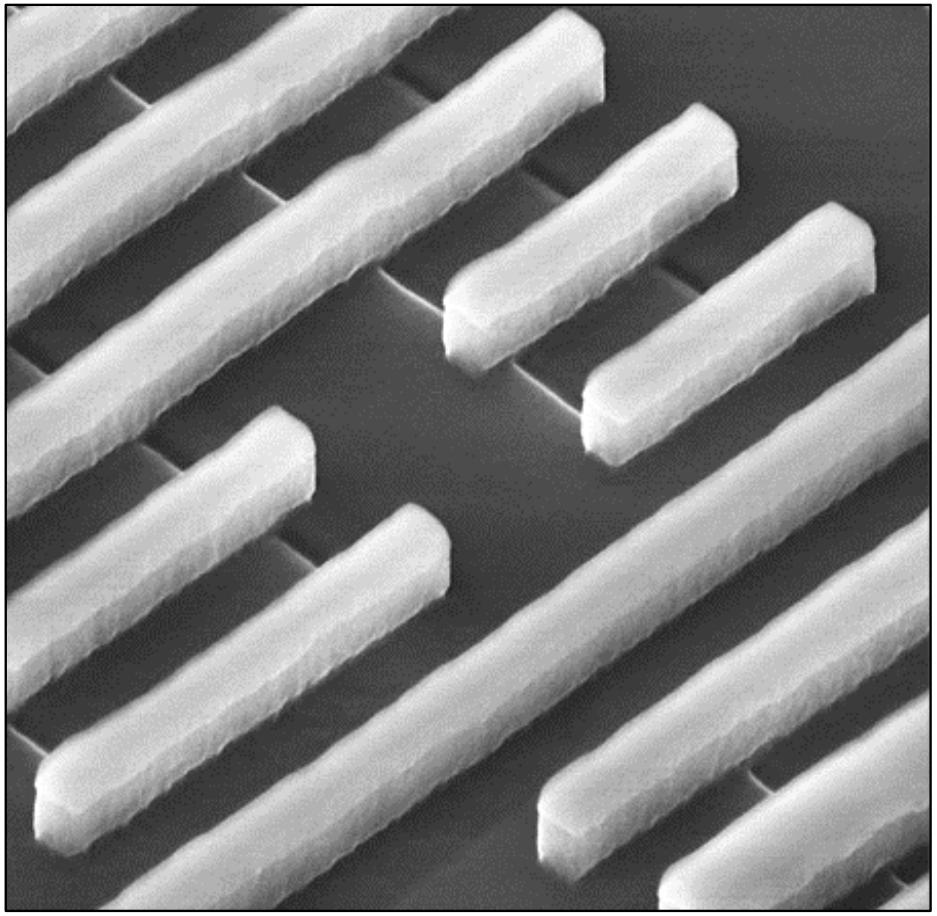


2D planar transistors form a conducting channel in silicon region under the gate electrode

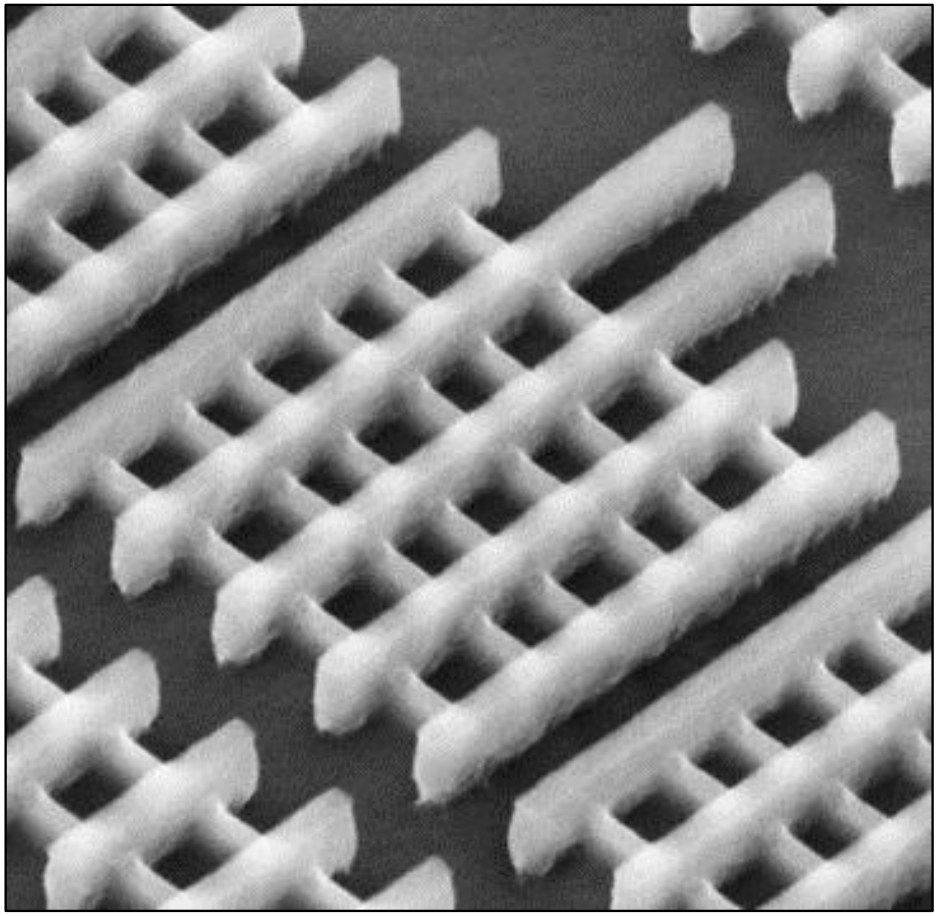
3D tri-gate transistors form conducting channels on three sides of a vertical fin structure

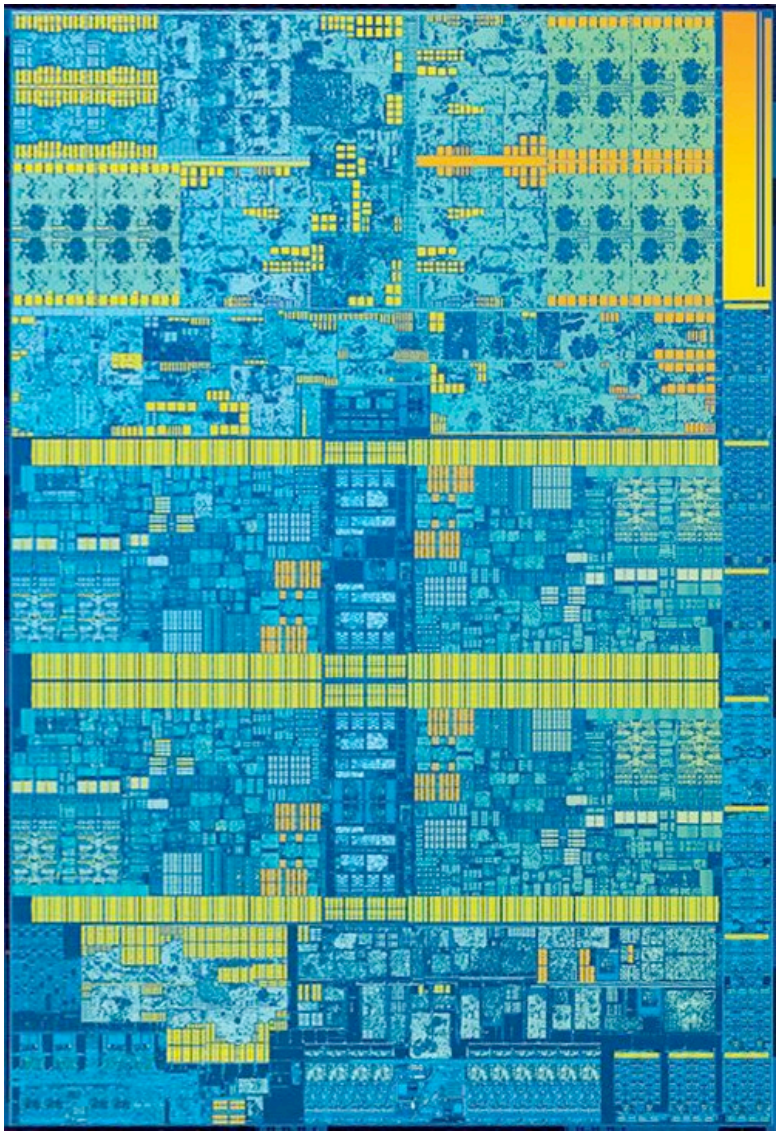
Tri-Gate Transistors

32 nm Planar Transistors



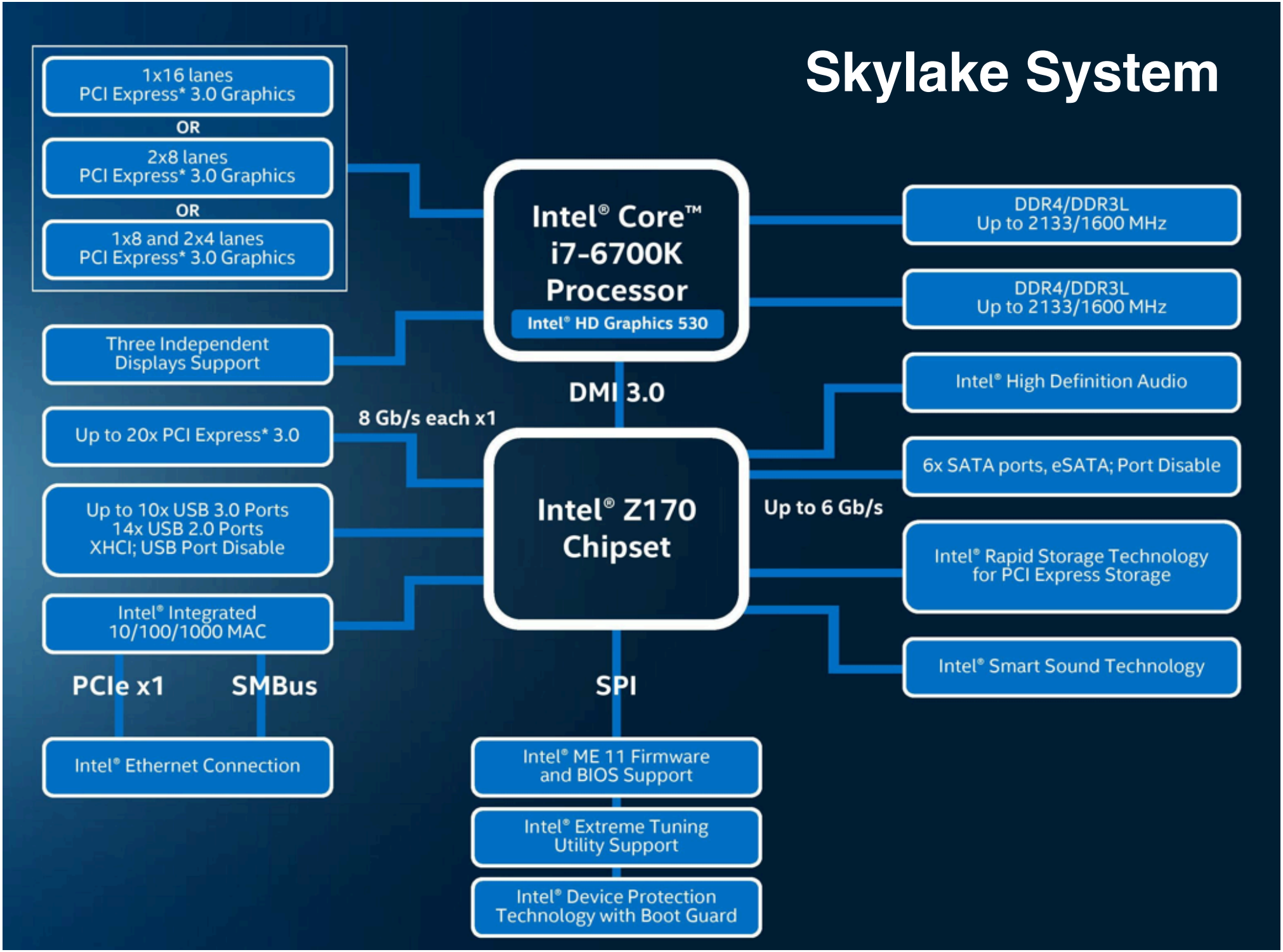
22 nm Tri-Gate Transistors



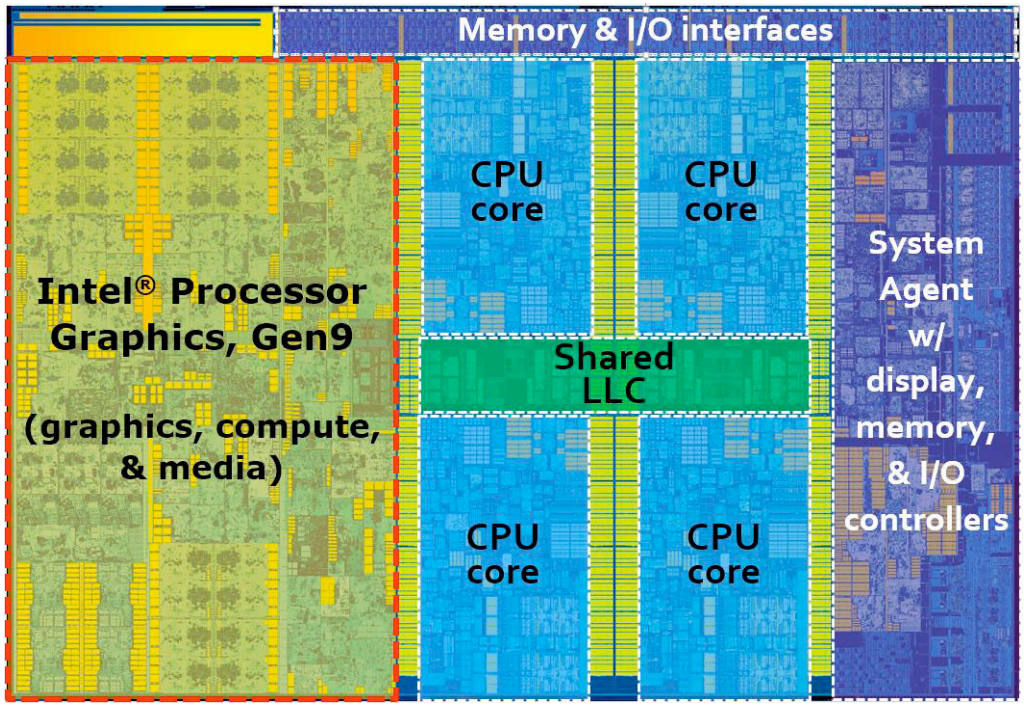


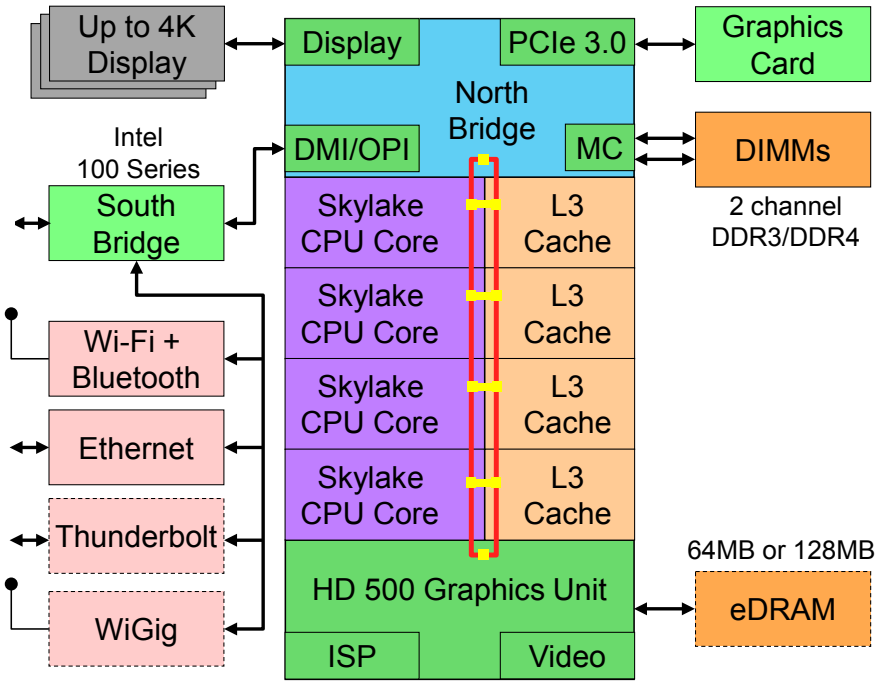
Intel Skylake

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


Skylake System
1x16 lanes
PCI Express* 3.0 Graphics
OR
2x8 lanes
PCI Express* 3.0 Graphics
OR
1x8 and 2x4 lanes
PCI Express* 3.0 Graphics
Intel® Core™
i7-6700K
Processor
Intel® HD Graphics 530
DDR4/DDR3L
Up to 2133/1600 MHz
DDR4/DDR3L
Up to 2133/1600 MHz
Three Independent
Displays Support
DMI 3.0
8 Gb/s each x1
Up to 20x PCI Express* 3.0
Intel® High Definition Audio
6x SATA ports, eSATA; Port Disable
Up to 6 Gb/s
Intel® Z170
Chipset
Up to 10x USB 3.0 Ports
14x USB 2.0 Ports
XHCI; USB Port Disable
Intel® Rapid Storage Technology
for PCI Express Storage
Intel® Integrated
10/100/1000 MAC
Intel® Smart Sound Technology
PCIe x1
SMBus
SPI
Intel® Ethernet Connection
Intel® ME 11 Firmware
and BIOS Support
Intel® Extreme Tuning
Utility Support
Intel® Device Protection
Technology with Boot Guard

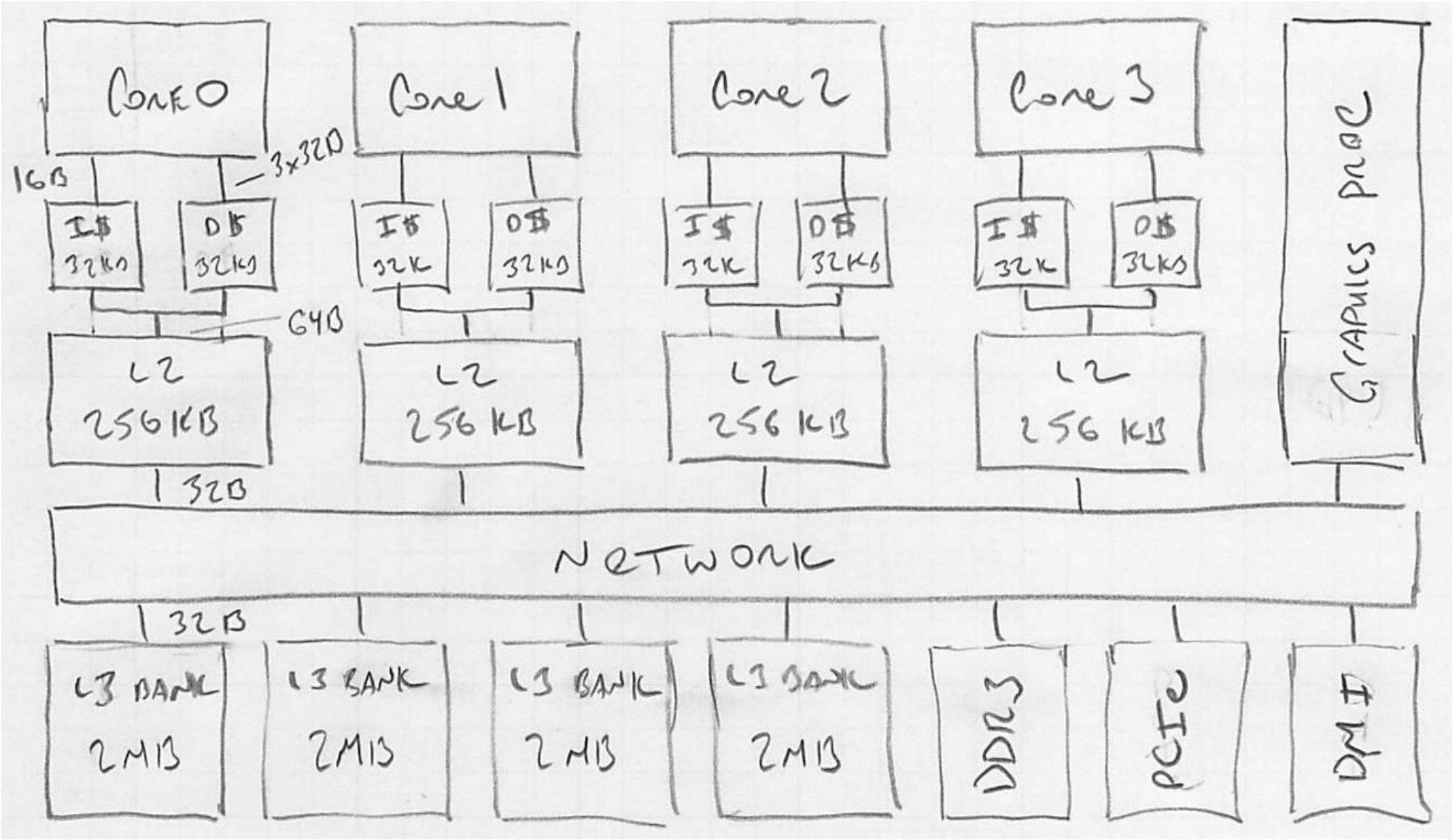
Intel Skylake i7-6700K, Fall 2015





- ~1.7B transistors in ~122 mm^2 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

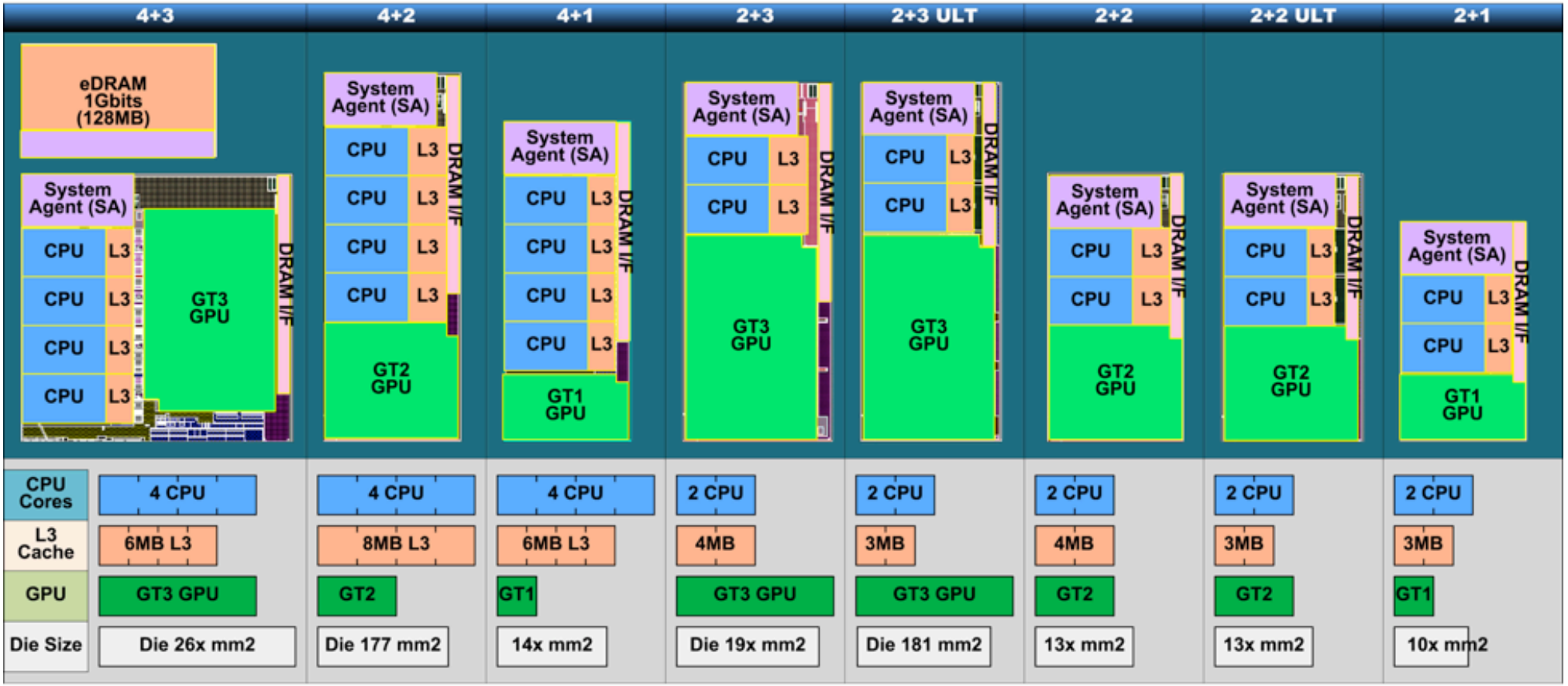
Intel Skylake: Block Diagram



Design-Time Modularity to Meet Scalability Application Requirement

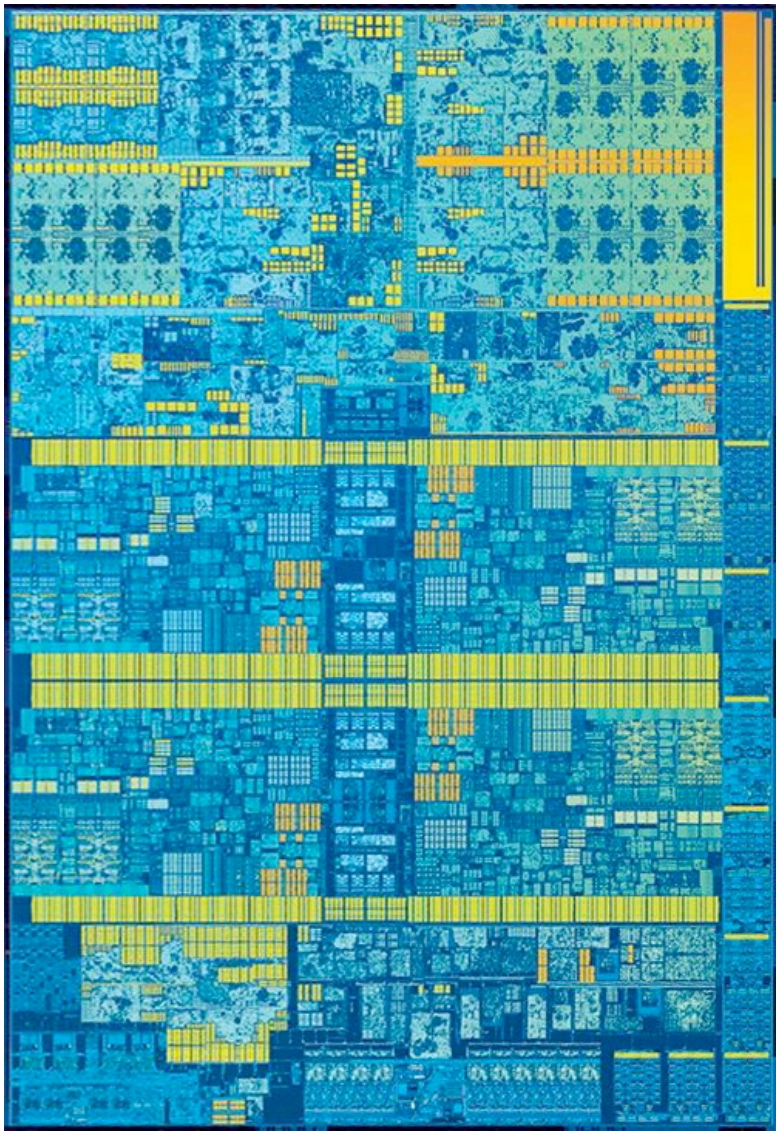
	2 in 1 Detachables, Tablets and Compute Stick	Thin Light Notebooks, Portable AIO, Minis and Conference Room		Ultimate Mobile Performance, Mobile Workstations		Desktop Performance to Value, AIO and Minis	
	Y-SERIES	**U-SERIES**		**H-SERIES**		**S-SERIES**	
5 Dies 4 Packages							
Dies	2+2	2+2	2+3e	4+2	4+4e	2+2	4+2
Package (mm)	BGA 1515	BGA 1356		BGA 1440		LGA 1151	
	20 x 16.5	42 x 24		42 x 28		37.5 x 37.5	
TDP (W)	4.5	15	15, 28	45		35, 65	35, 65, 91(“K”)
Chipset	Integrated 6th Gen Intel® Core™ Platform I/O			Intel® 100 Series chipset (23mm x 23mm)			

Scalable Design in Haswell Microarchitecture



Intel Haswell i7-4770K
85W @ 3.5 GHz

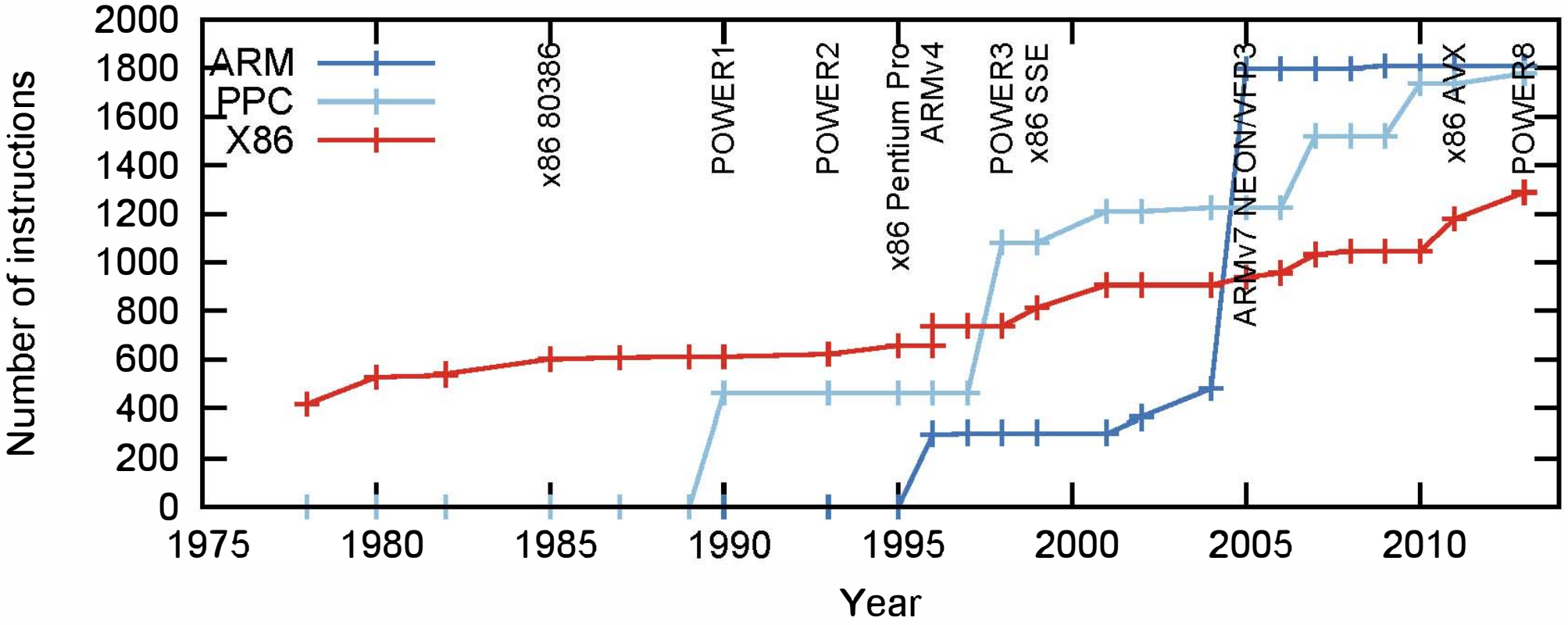
Intel Haswell 3560Y
6W @ 880 MHz



Intel Skylake

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

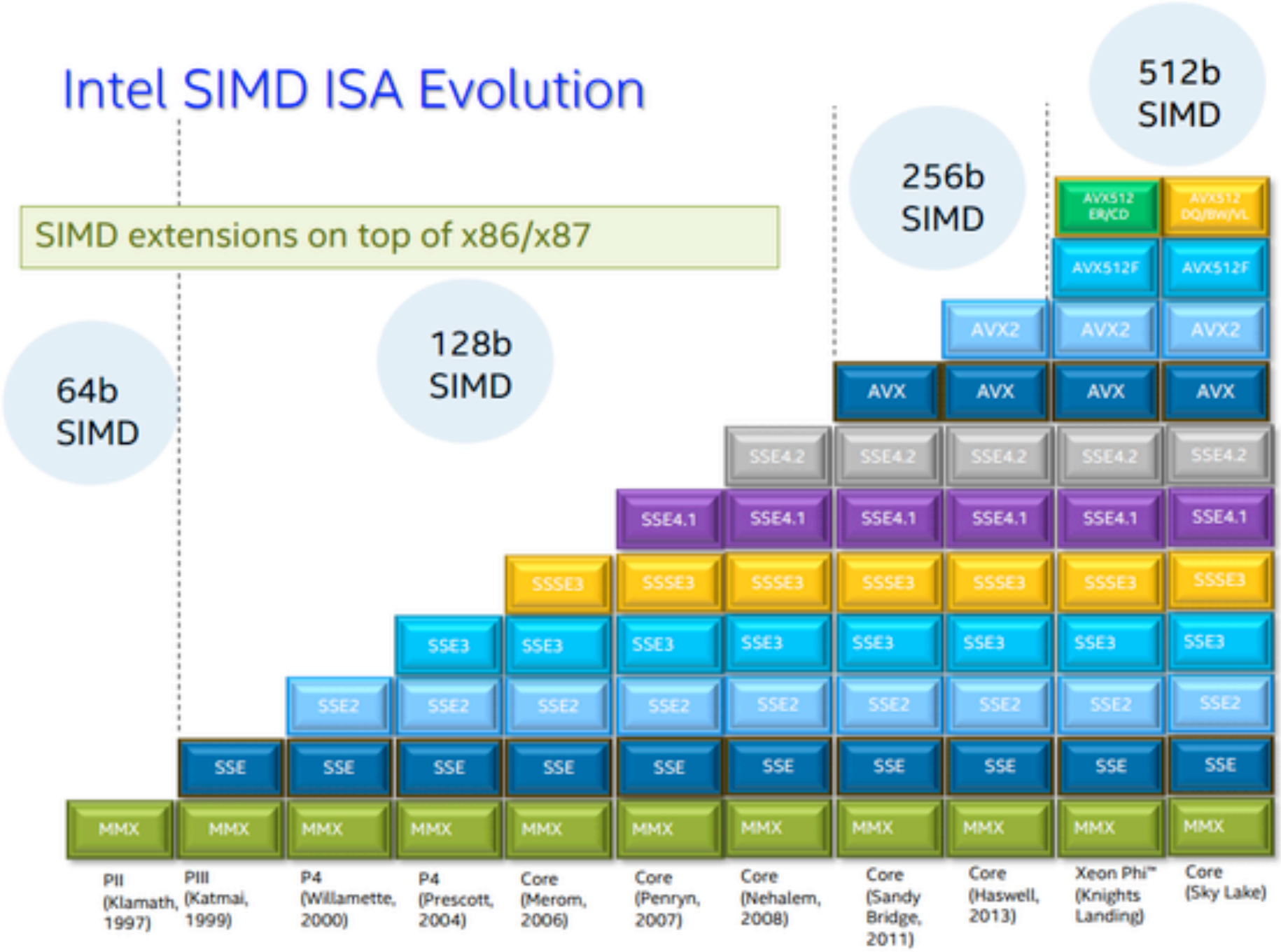
Growth in Instruction Sets Over Time



Three key additions in Skylake

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

AVX512: 512-bit SIMD Extensions



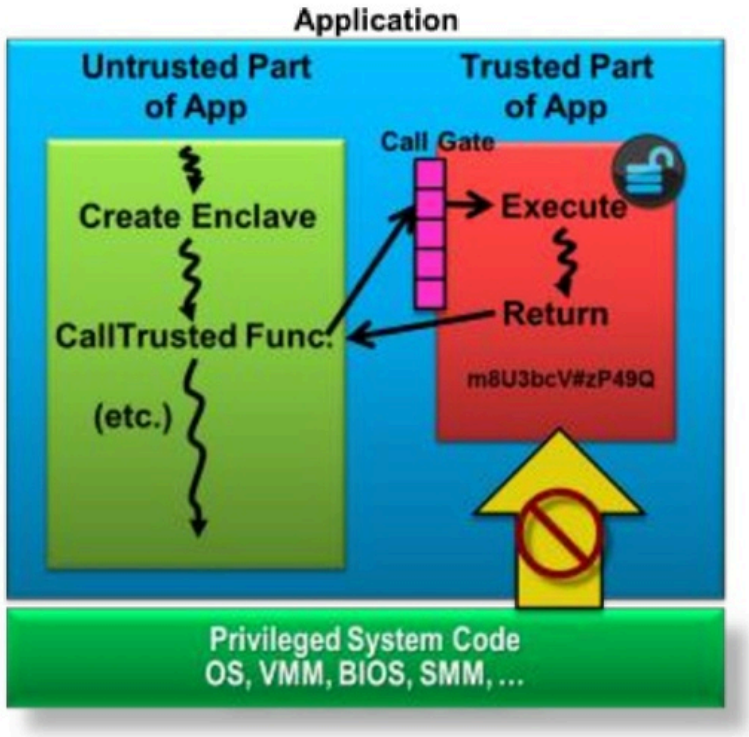
SGX: Software Guarded Extensions

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

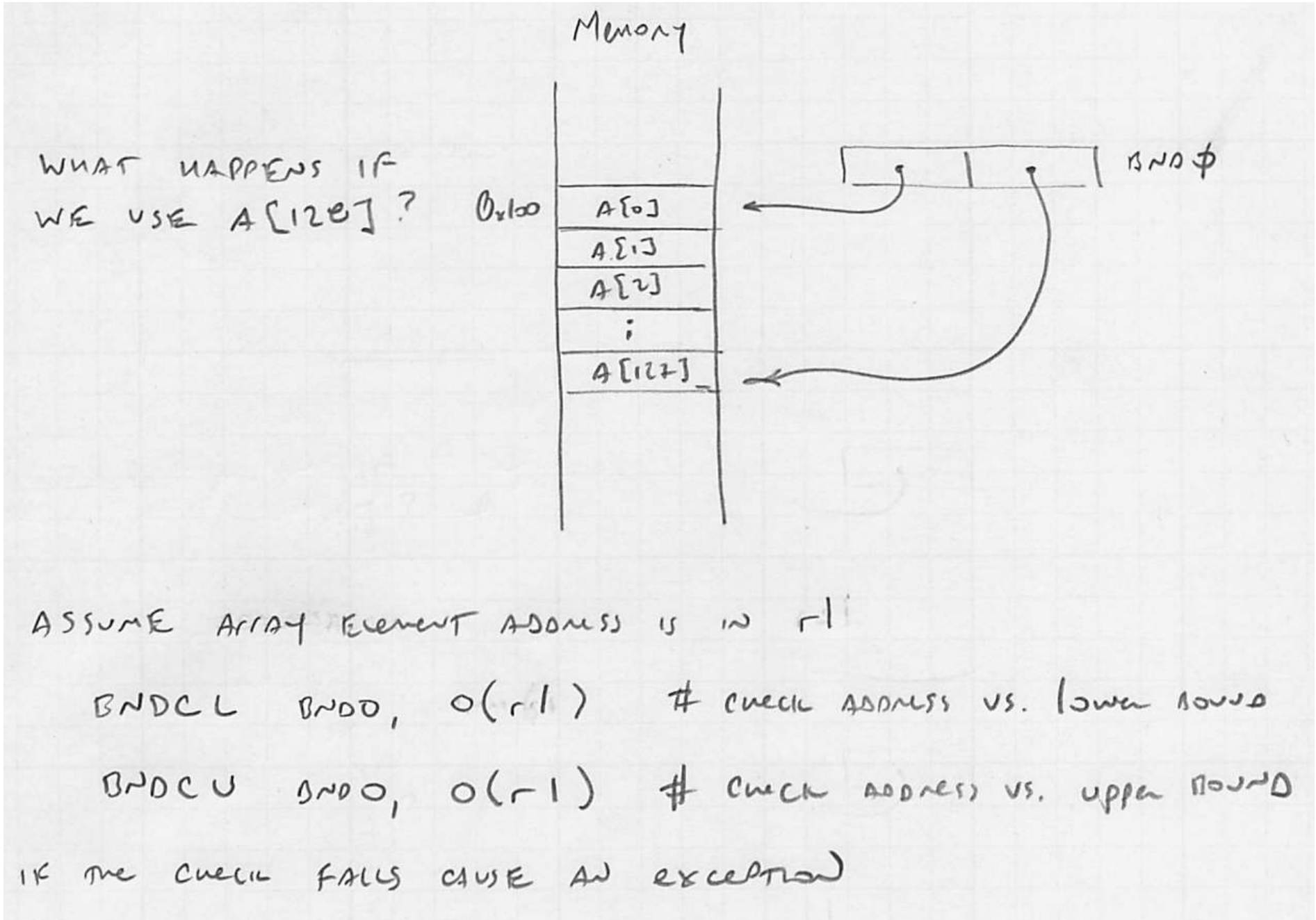
- Against software attacks originated at any privilege level
- Against many hardware based attacks

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

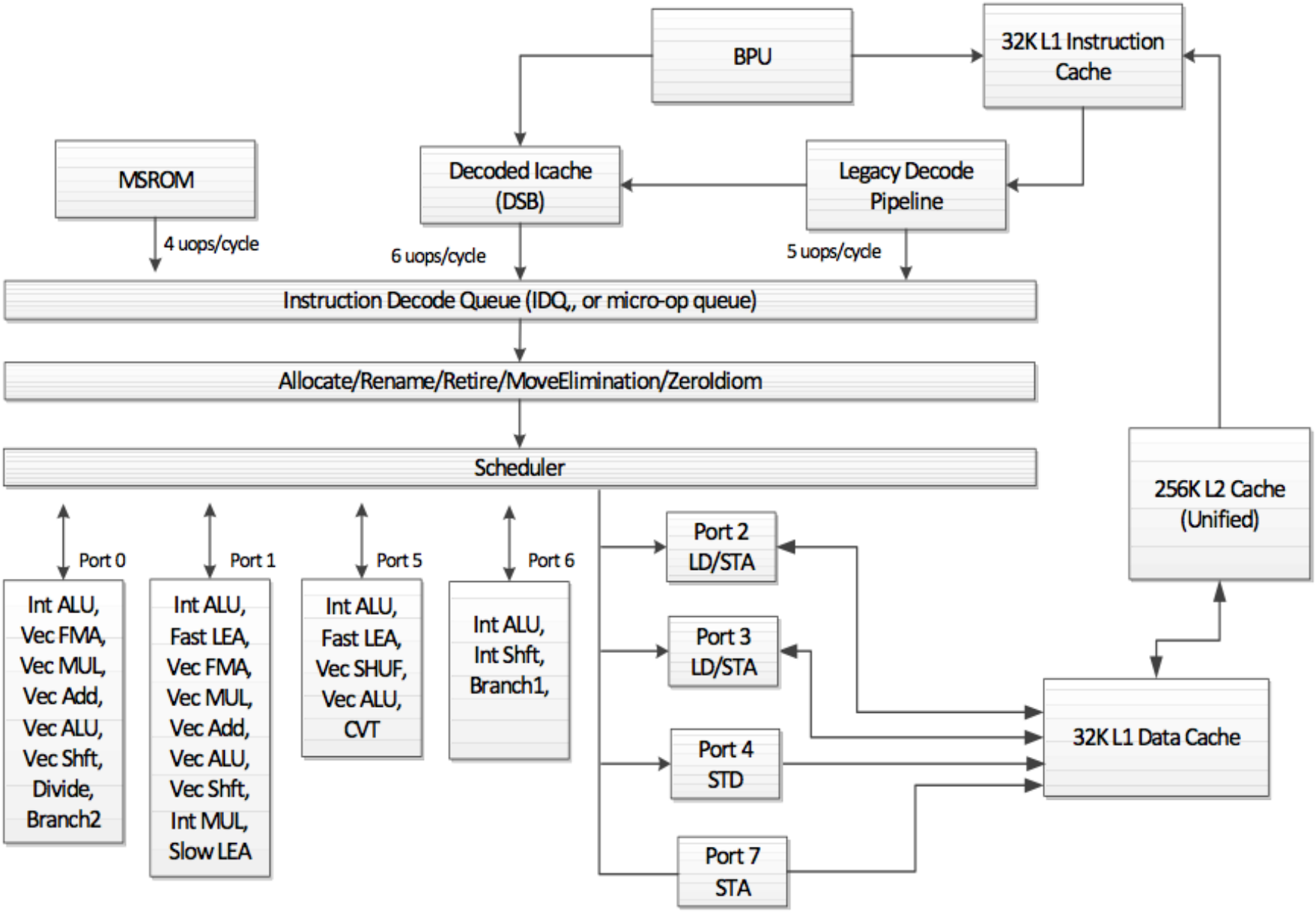


ASSUME ARRAY ELEMENT ADDRESS IS IN r1

```
BNDCL  BND0, 0(r1)   # CHECK ADDRESS VS. lower bound
BNDCU  BND0, 0(r1)   # CHECK ADDRESS VS. upper bound
```

IF THE CHECK FAILS CAUSE AN EXCEPTION

Processor Block Diagram



From Intel 64 and IA-32 Architectures Optimization Reference Manual

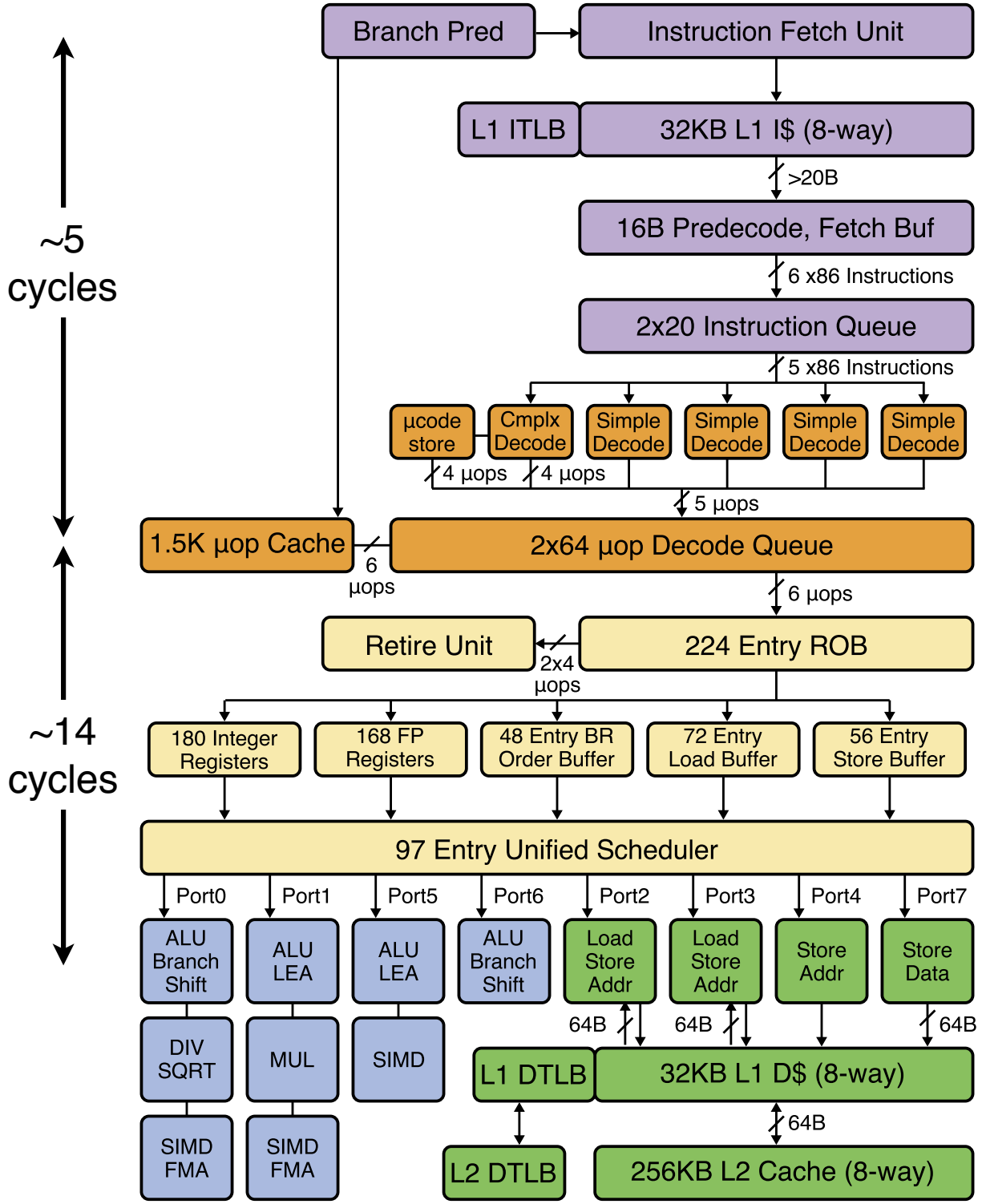
IO Fetch

IO Decode

OOO Issue and Late Commit

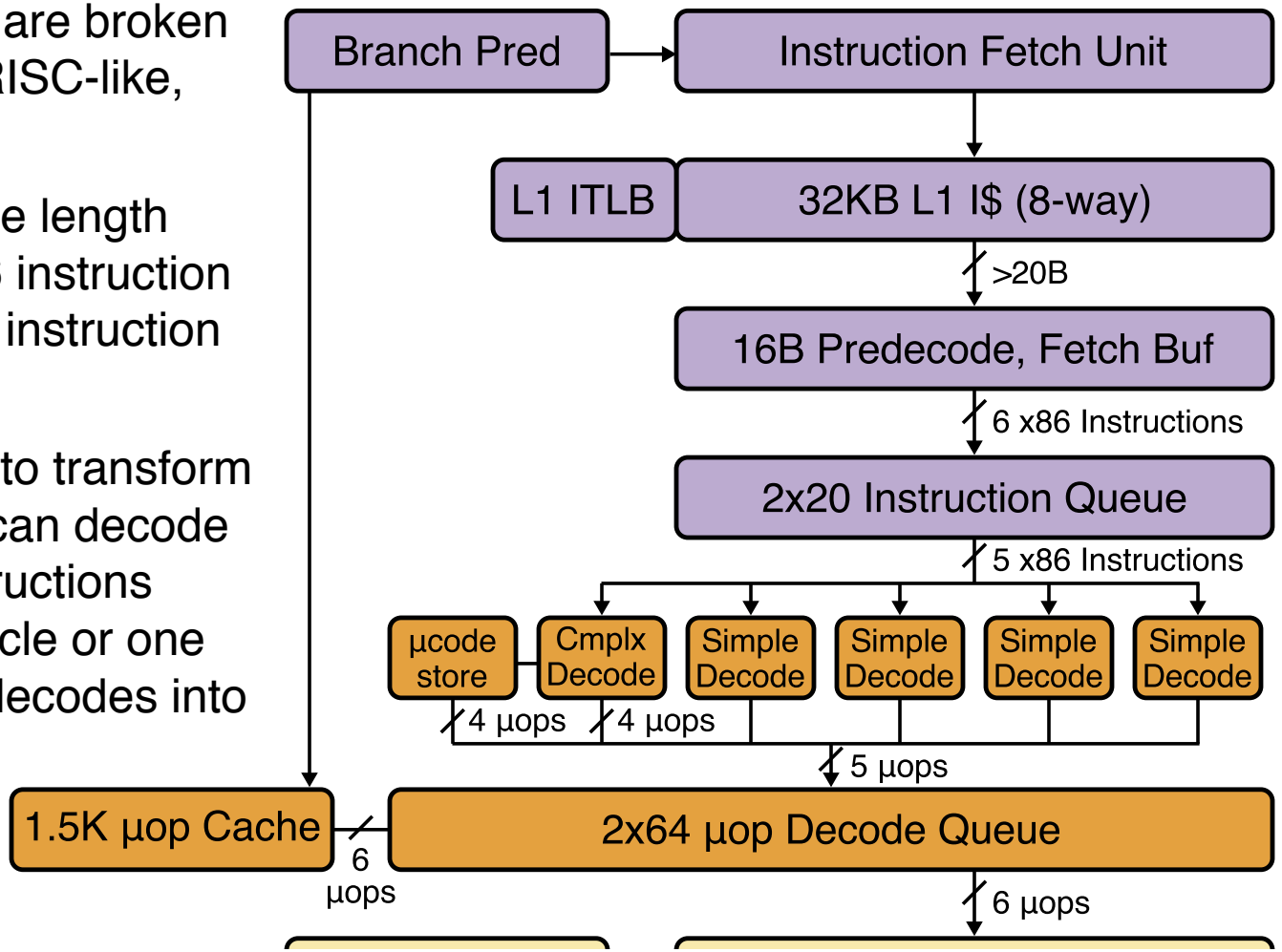
Integer/FP Functional Units with OOO Writeback

Load/Store Execution



IO Fetch and Decode

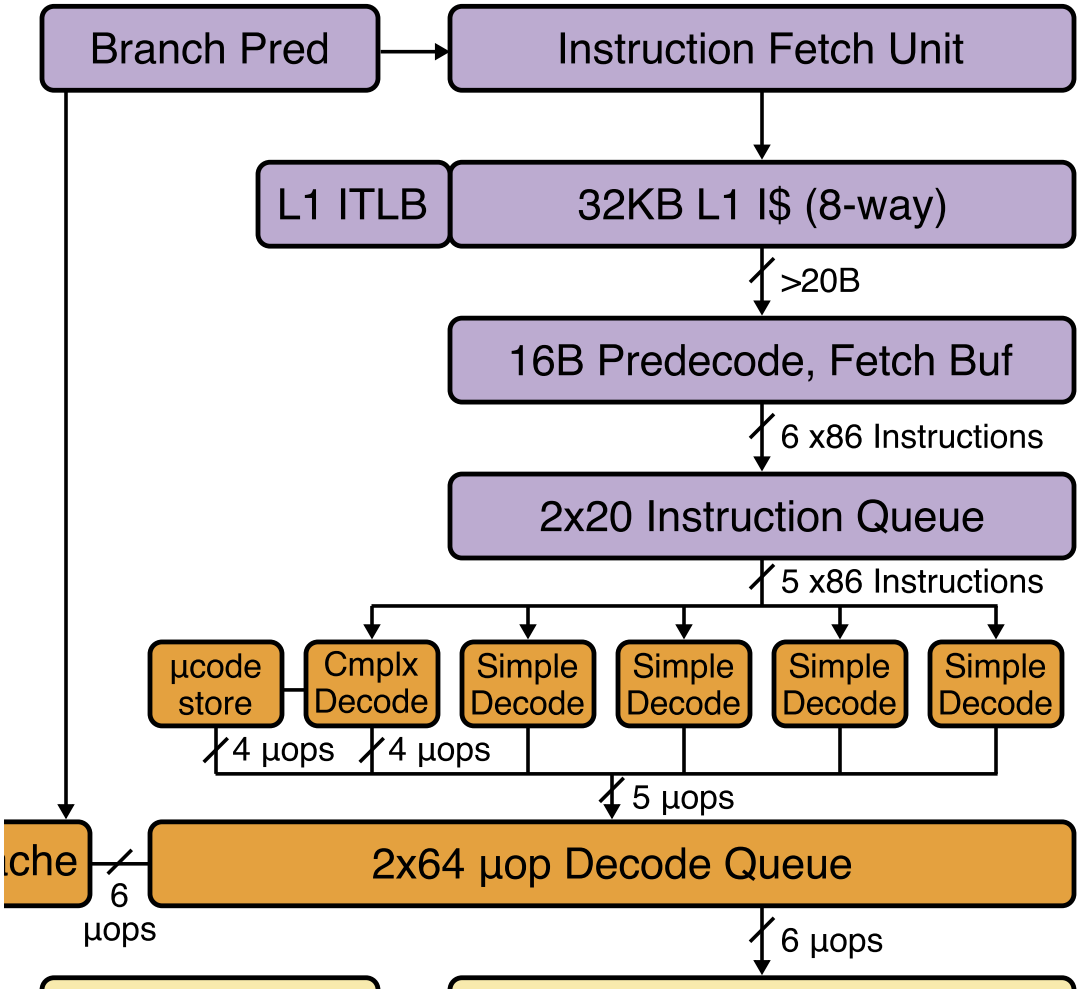
- Complex CISC instructions are broken into much simpler, almost RISC-like, **micro-ops**
- Predecoder handles variable length encoding (1-15B), finds x86 instruction boundaries and inserts into instruction queue
- Parallel decoders are used to transform x86 instructions into uops; can decode either five “simple” x86 instructions (decodes into 1 uop) per cycle or one “complex” x86 instruction (decodes into 1-4 fused uops)



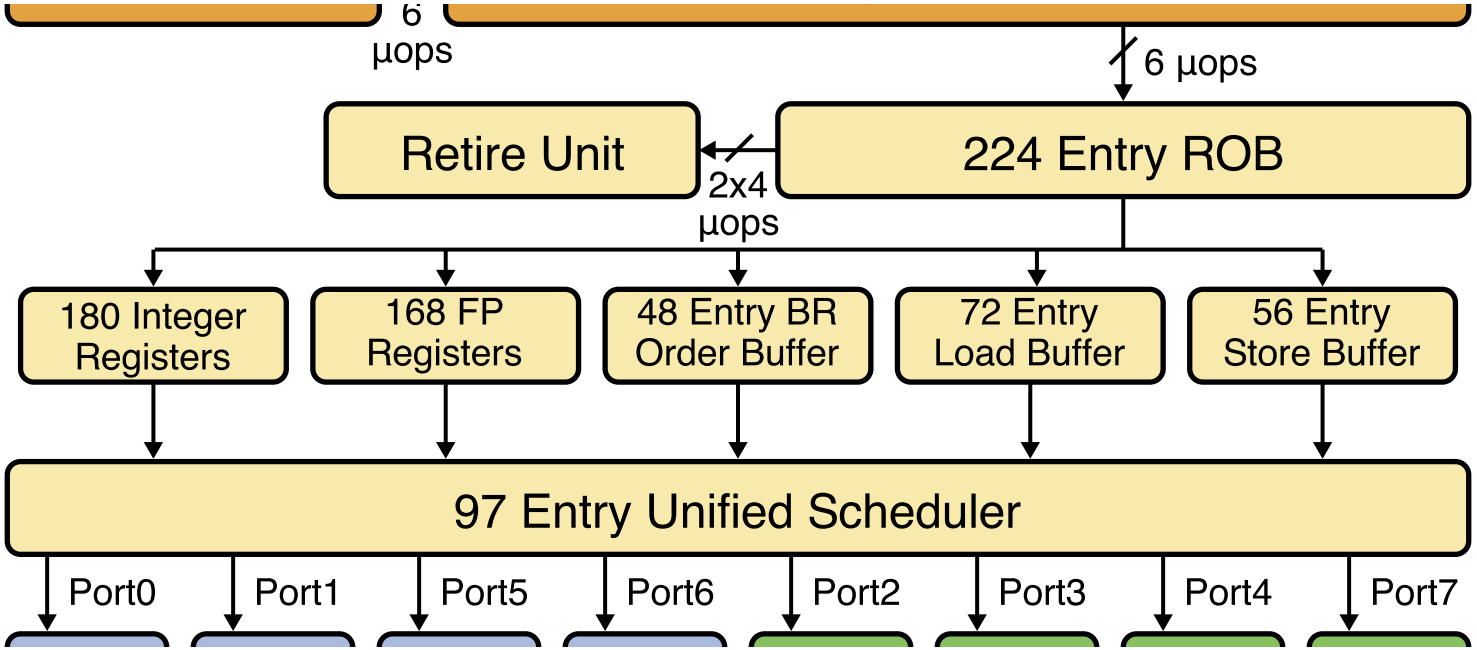
- Very complex instructions fall back to a microcoded control unit
- uop cache acts as a kind of L0 instruction cache that holds decoded uops and enables much of the front-end to be shut down to save power
- uop decode queue can be used as a special loop cache

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



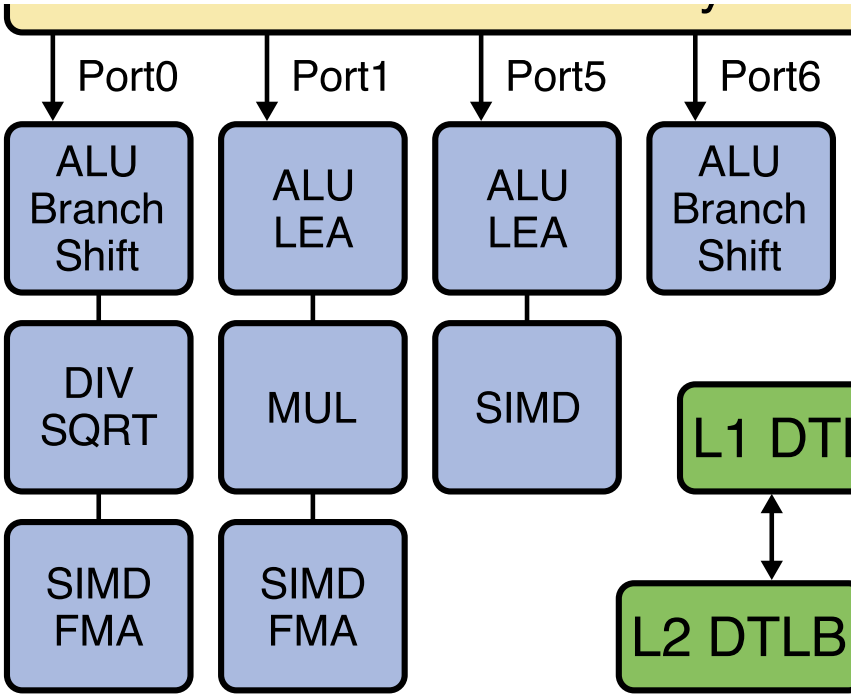
OOO Issue and Late Commit



- **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

Functional Units with OOO Writeback

- 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)”

Size of Data Structures

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

- **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 entires 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

```
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

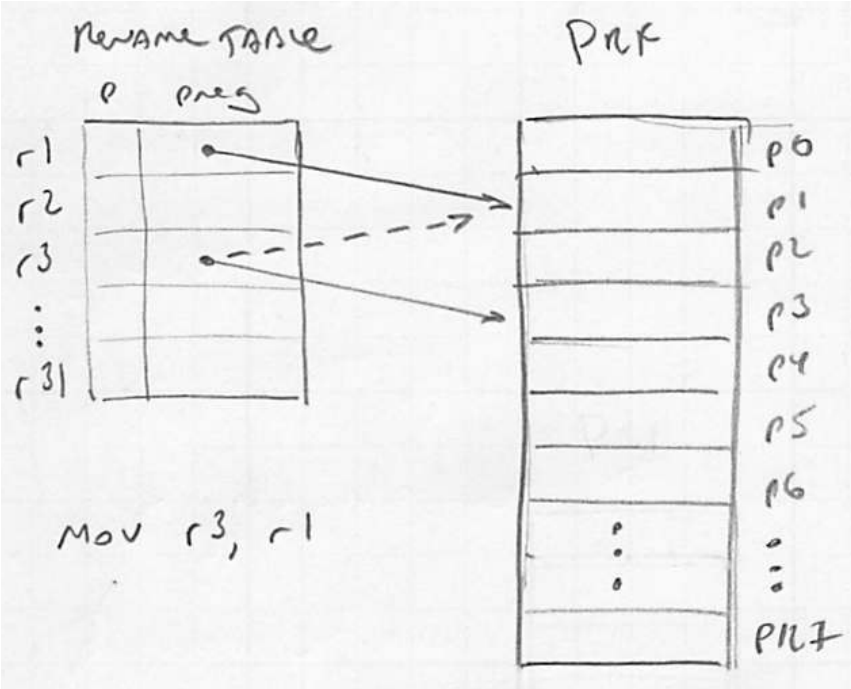
Combines two micro-ops together (load plus integer op, split stores) so that they only take a single ROB and IQ entry, but the fused micro-op is split such that two micro-ops are issued to two different execution units

```
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 an 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
```

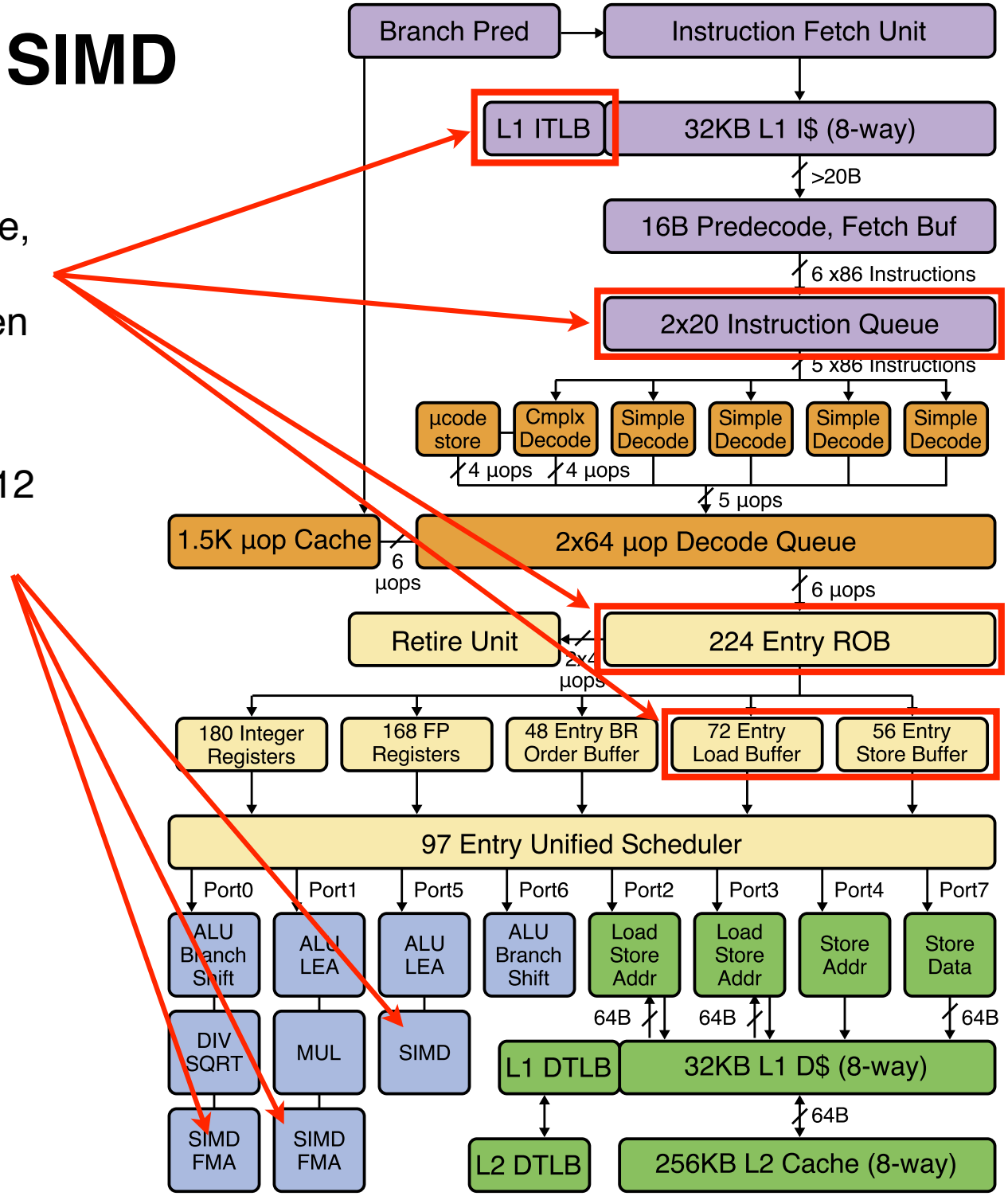
Allocate a fresh destination register in the rename stage, but then immediately clear the value in this destination register to zero.

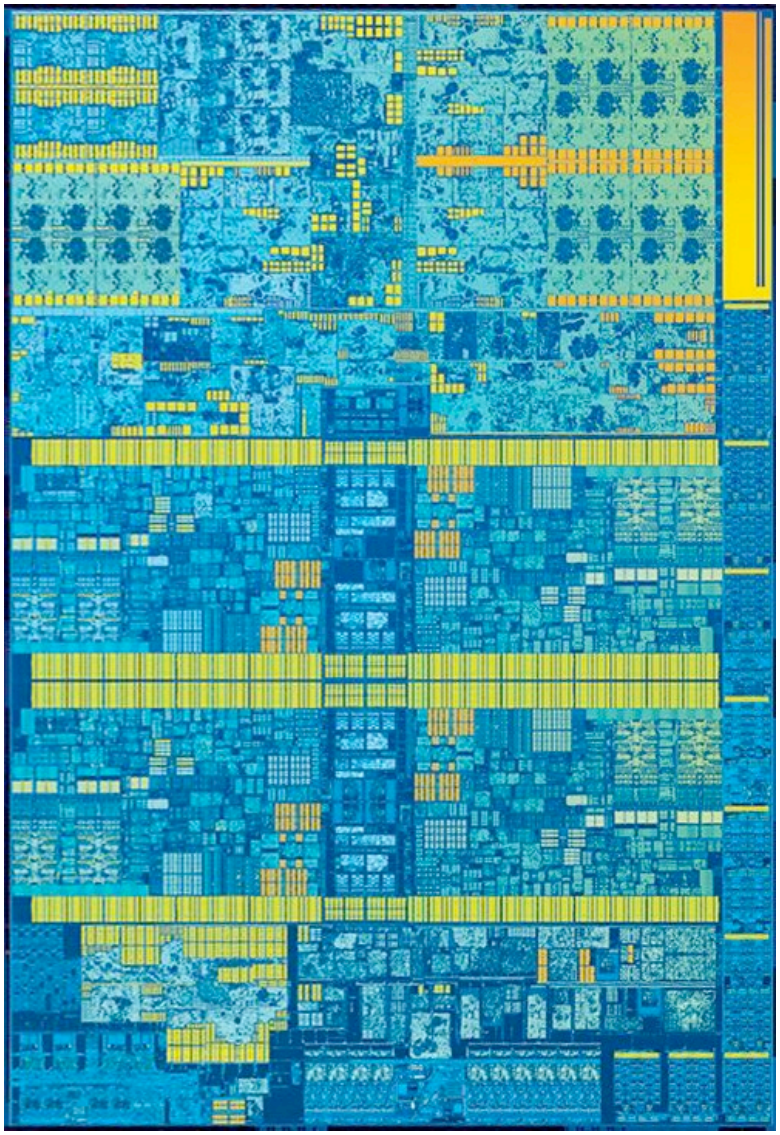
Both techniques enable specific instructions to use no execution resources!

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



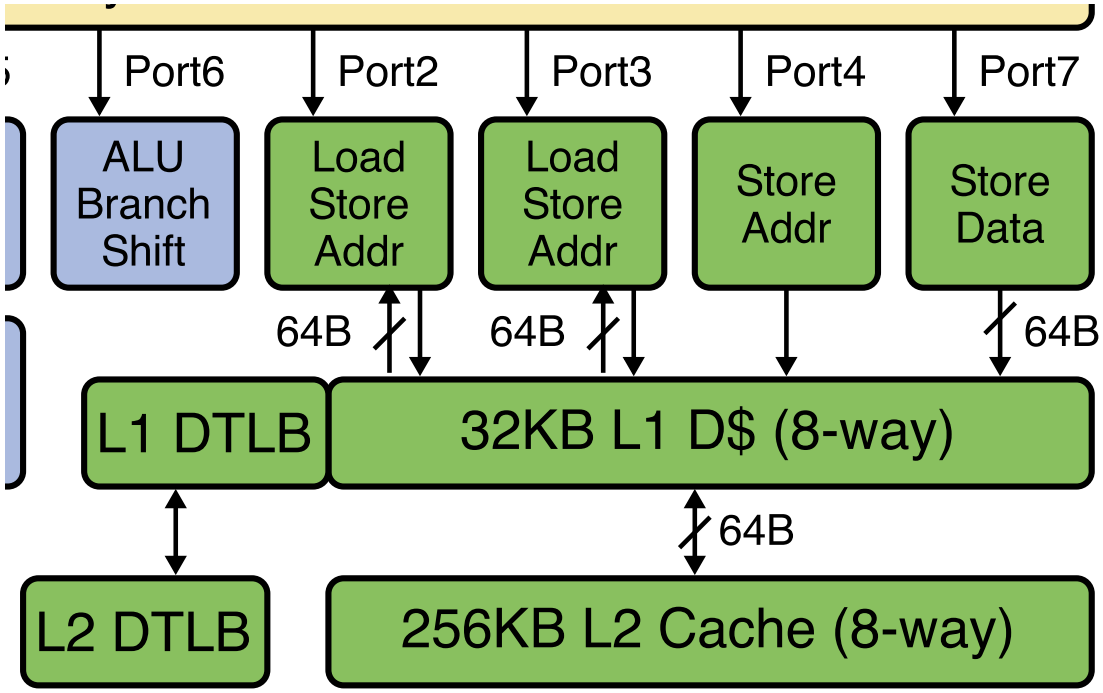


Intel Skylake

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

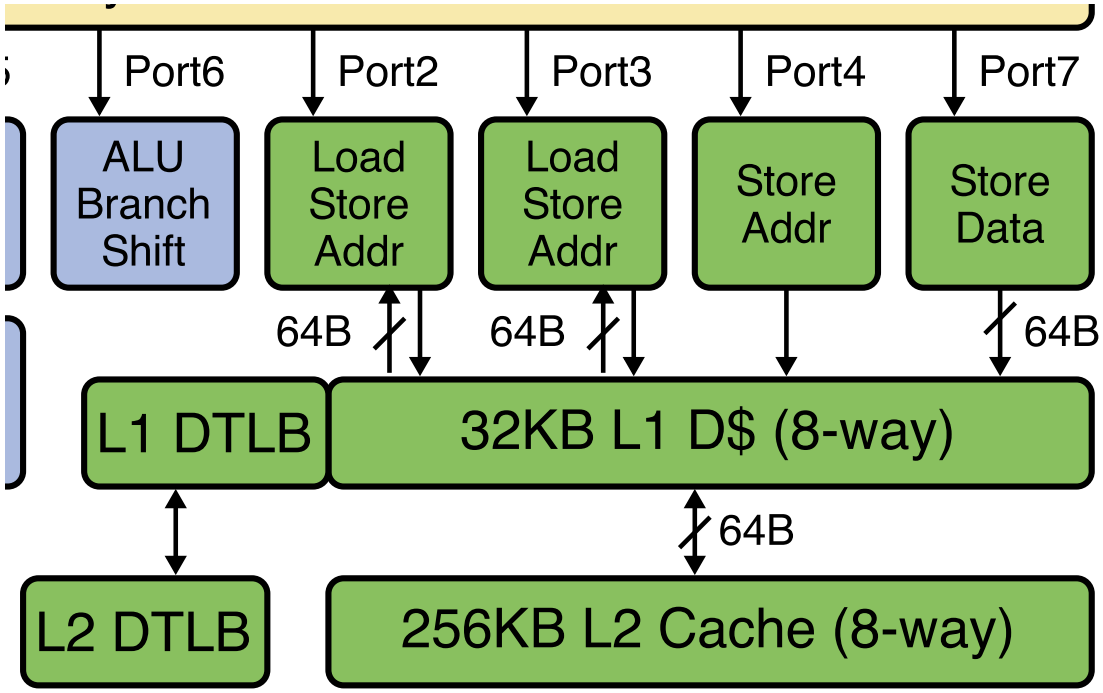
Memory System

- 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



Memory System

- “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



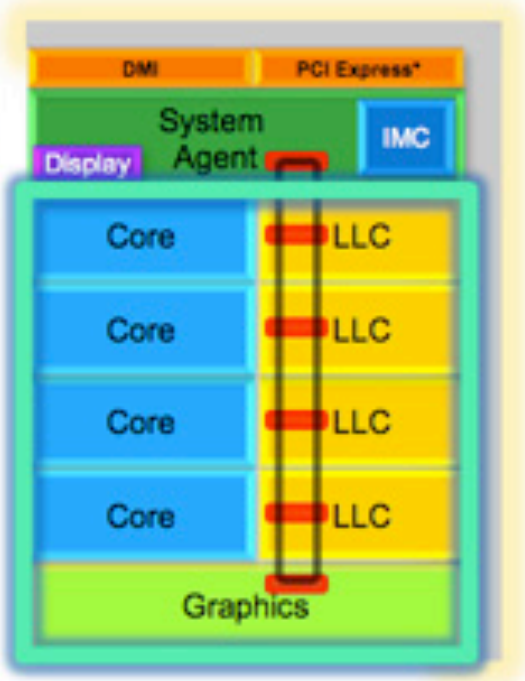
Memory System Parameters

Metric	Nehalem	Sandy Bridge	Haswell
L1 Instruction Cache	32K, 4-way	32K, 8-way	32K, 8-way
L1 Data Cache	32K, 8-way	32K, 8-way	32K, 8-way
Fastest Load-to-use	4 cycles	4 cycles	4 cycles
Load bandwidth	16 Bytes/cycle	32 Bytes/cycle (banked)	**64 Bytes/cycle**
Store bandwidth	16 Bytes/cycle	16 Bytes/cycle	**32 Bytes/cycle**
L2 Unified Cache	256K, 8-way	256K, 8-way	256K, 8-way
Fastest load-to-use	10 cycles	11 cycles	11 cycles
Bandwidth to L1	32 Bytes/cycle	32 Bytes/cycle	**64 Bytes/cycle**
L1 Instruction TLB	4K: 128, 4-way 2M/4M: 7/thread	4K: 128, 4-way 2M/4M: 8/thread	4K: 128, 4-way 2M/4M: 8/thread
L1 Data TLB	4K: 64, 4-way 2M/4M: 32, 4-way 1G: fractured	4K: 64, 4-way 2M/4M: 32, 4-way 1G: 4, 4-way	4K: 64, 4-way 2M/4M: 32, 4-way 1G: 4, 4-way
L2 Unified TLB	4K: 512, 4-way	4K: 512, 4-way	**4K+2M shared: 1024, 8-way**

All caches use 64-byte lines

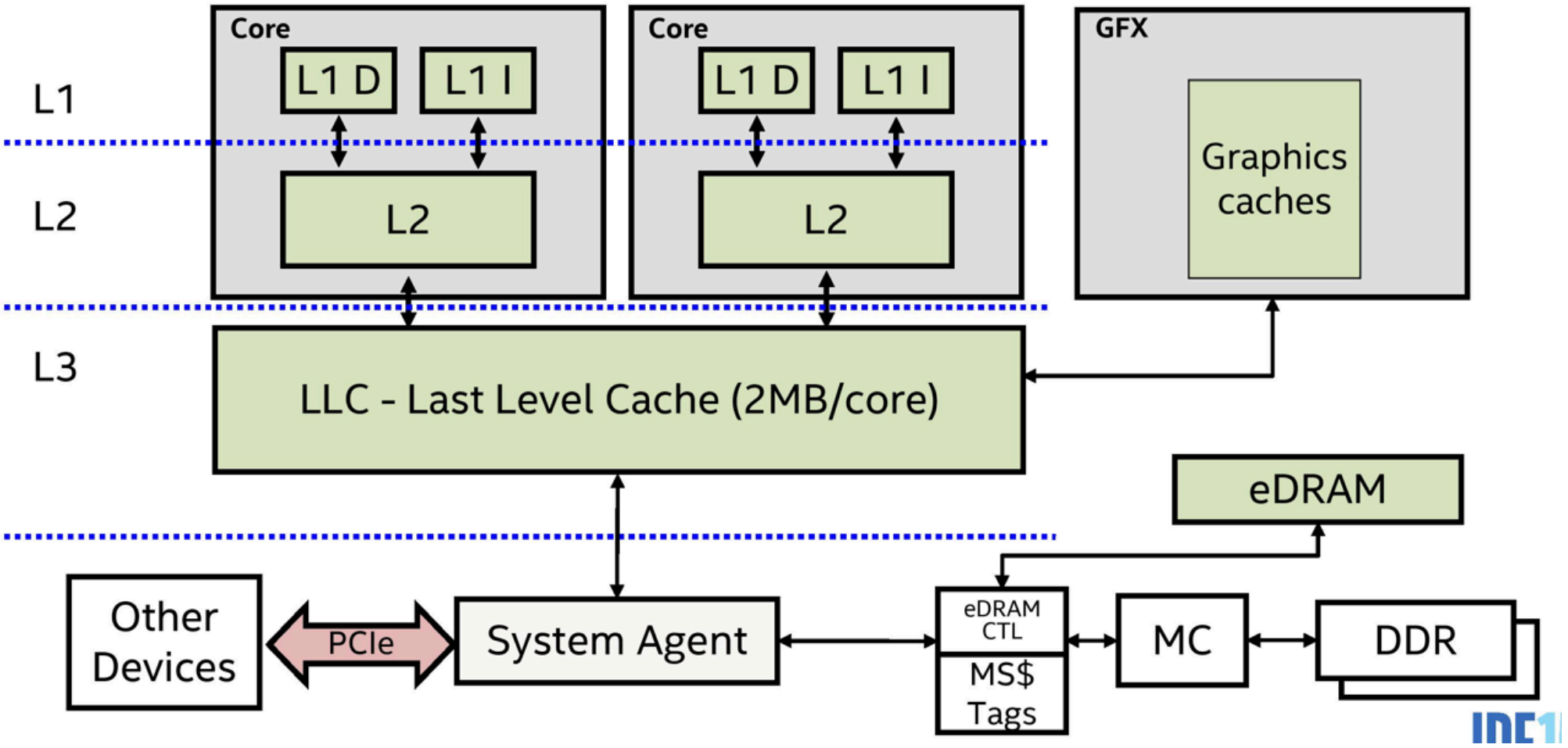
Last-Level Cache

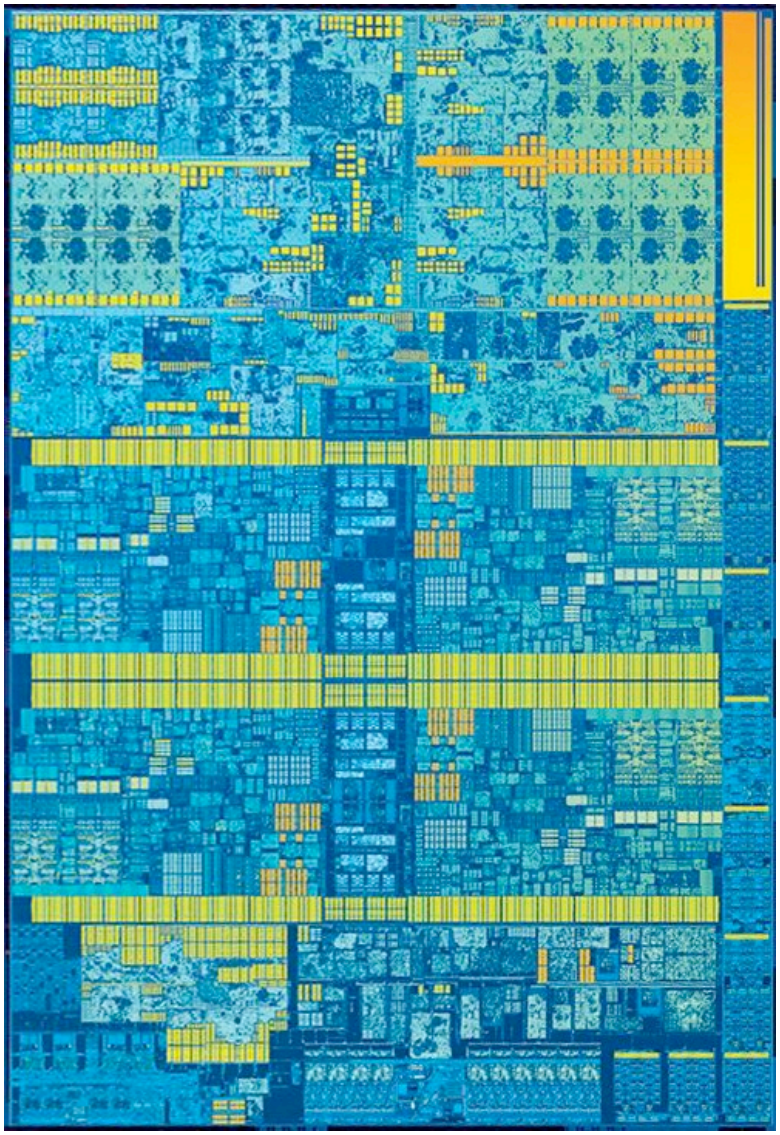
- **LLC shared** among all Cores, Graphics and Media
  - Graphics driver controls **which streams** are cached/coherent
  - **Any agent** can access all data in the LLC, independent of who allocated the line, after **memory range checks**
- Controlled LLC **way allocation** mechanism to prevent thrashing 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



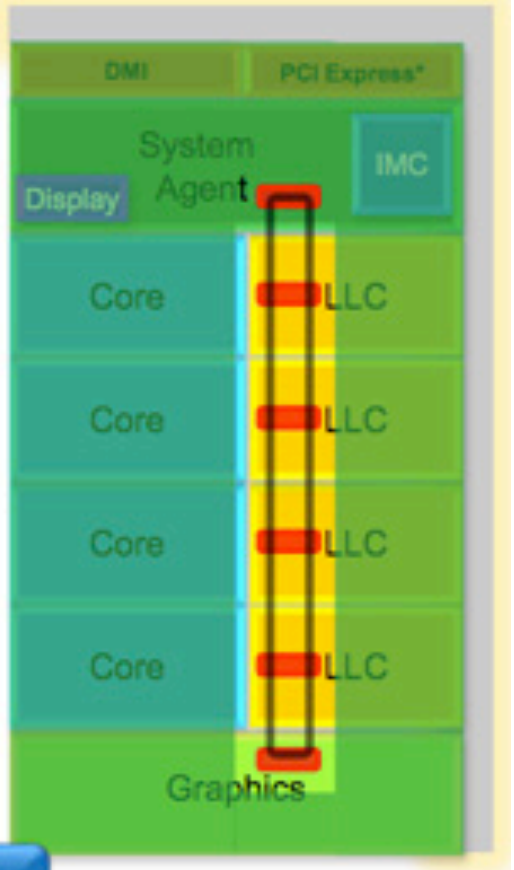


Intel Skylake

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

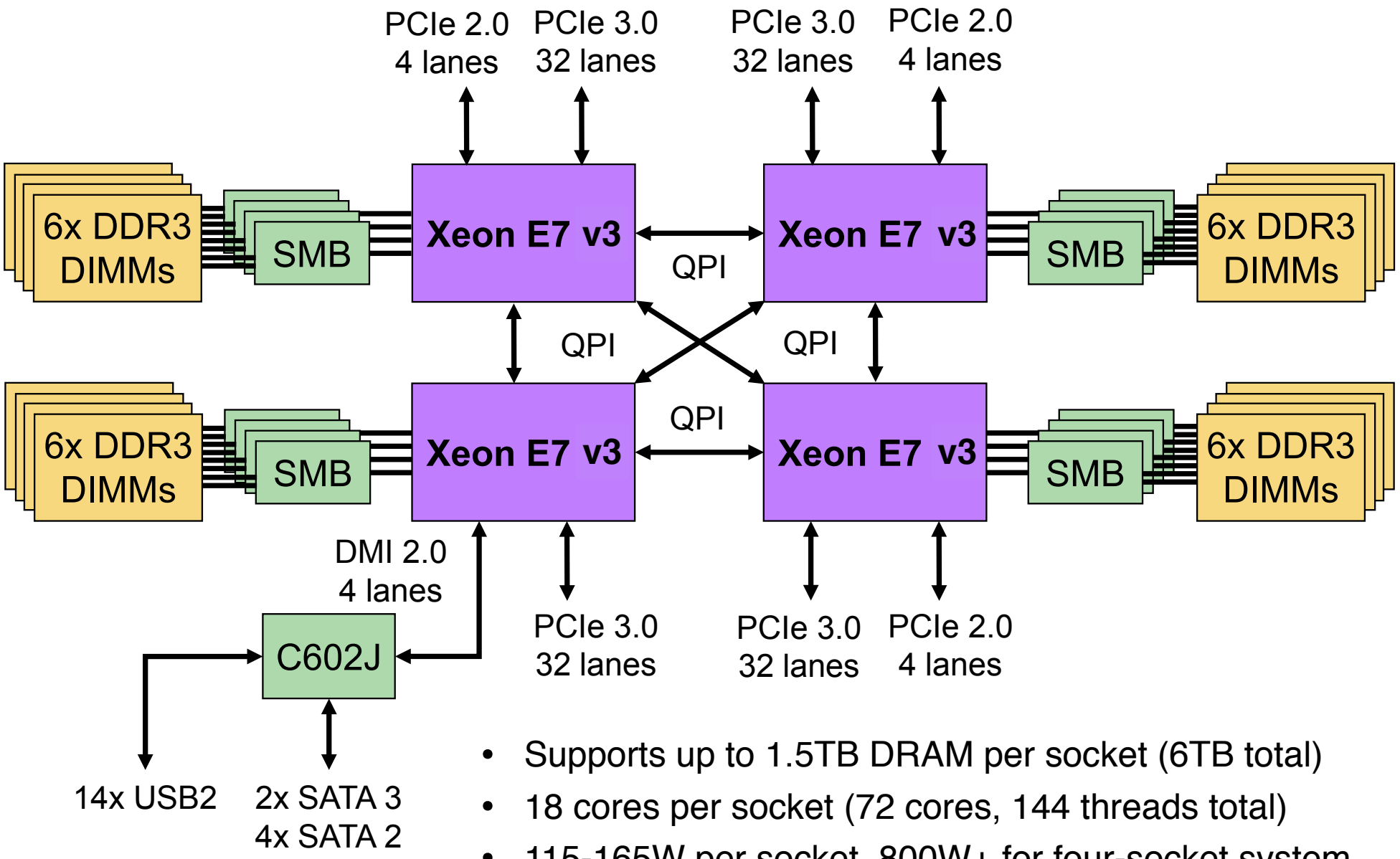
Pipelined Bus Interconnect

- **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



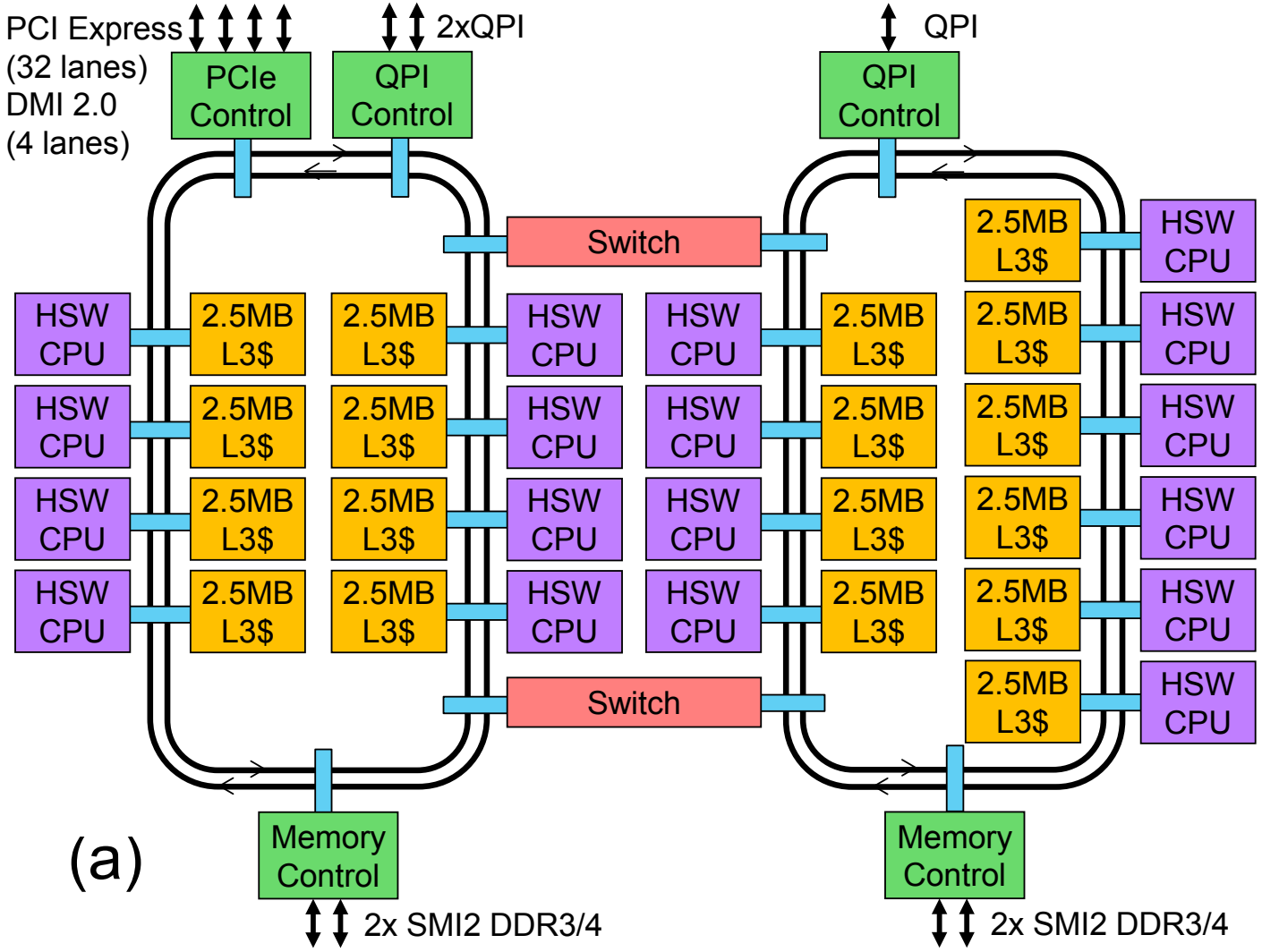
High Bandwidth, Low Latency, Modular

Intel Brickland Platform



- Supports up to 1.5TB DRAM per socket (6TB total)
- 18 cores per socket (72 cores, 144 threads total)
- 115-165W per socket, 800W+ for four-socket system
- Fully connected inter-socket network

Intel Haswell-EX Processor



- 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

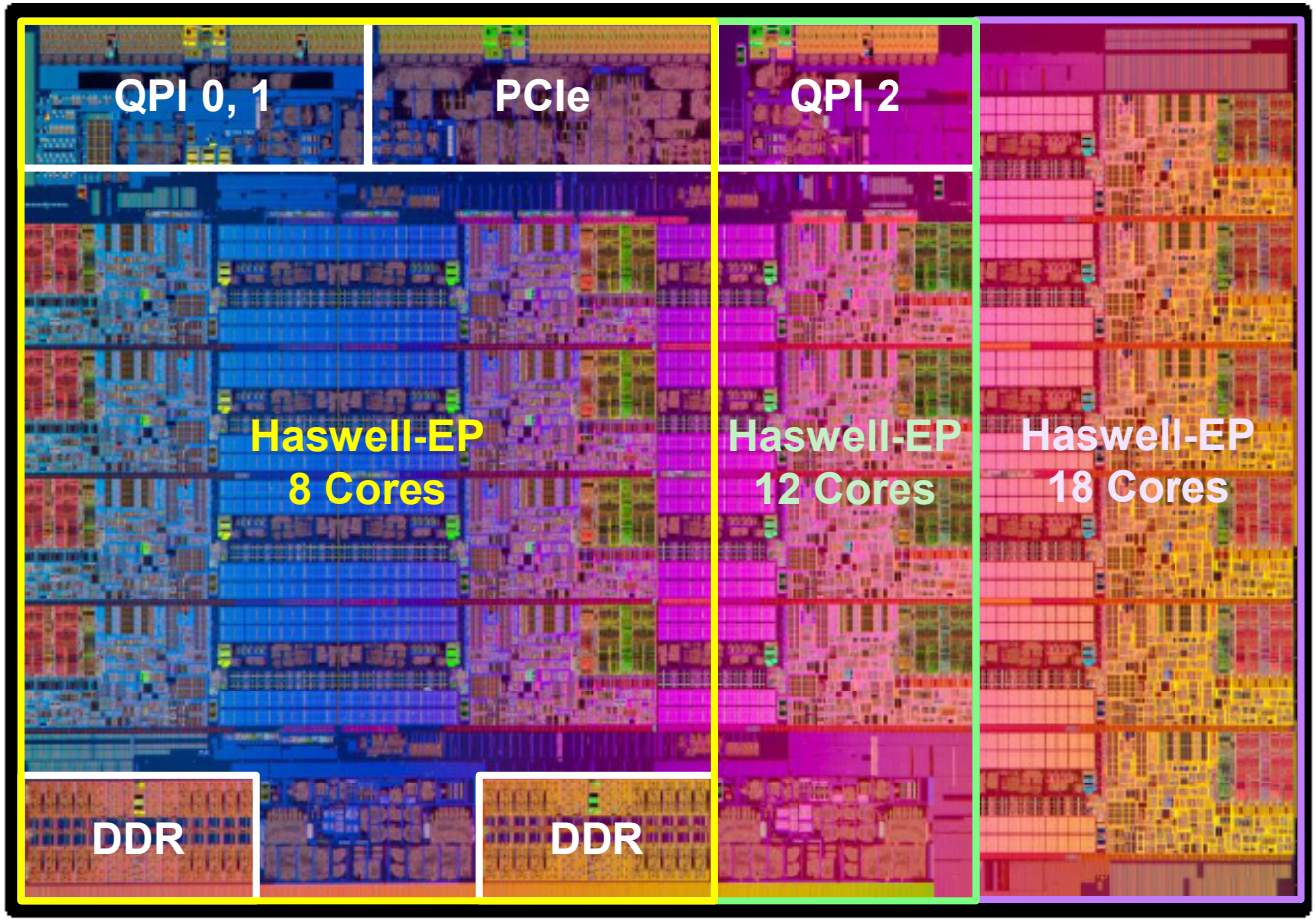
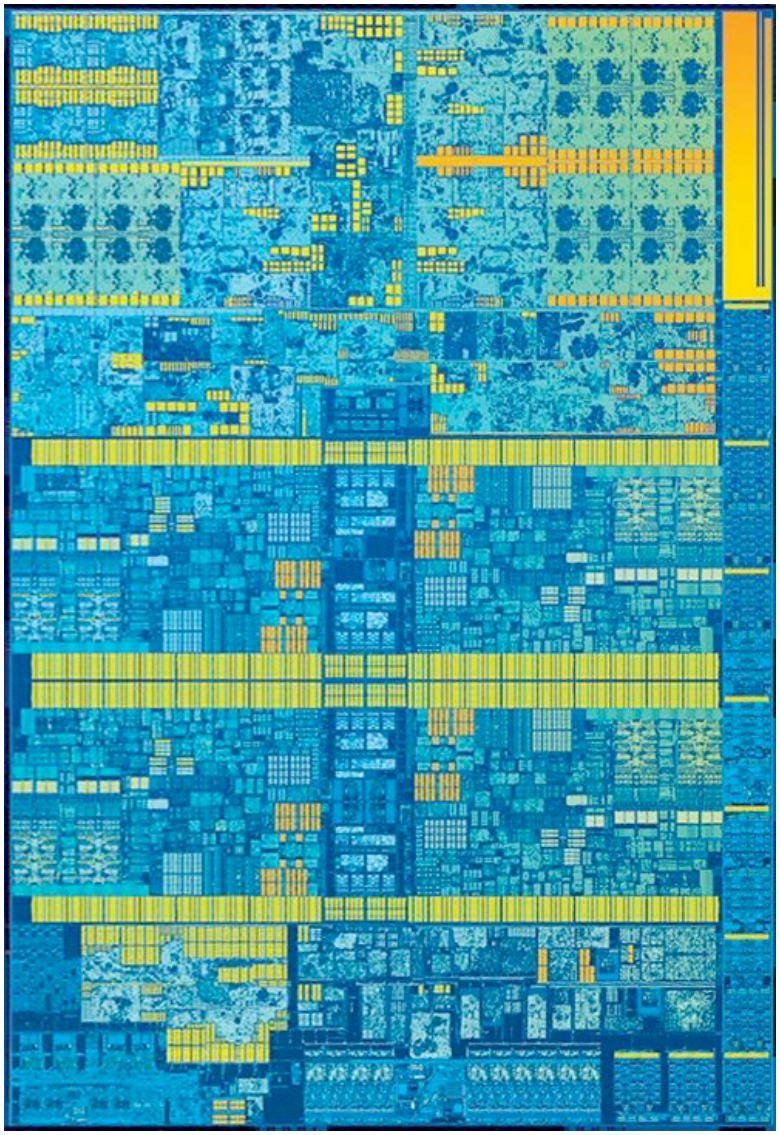


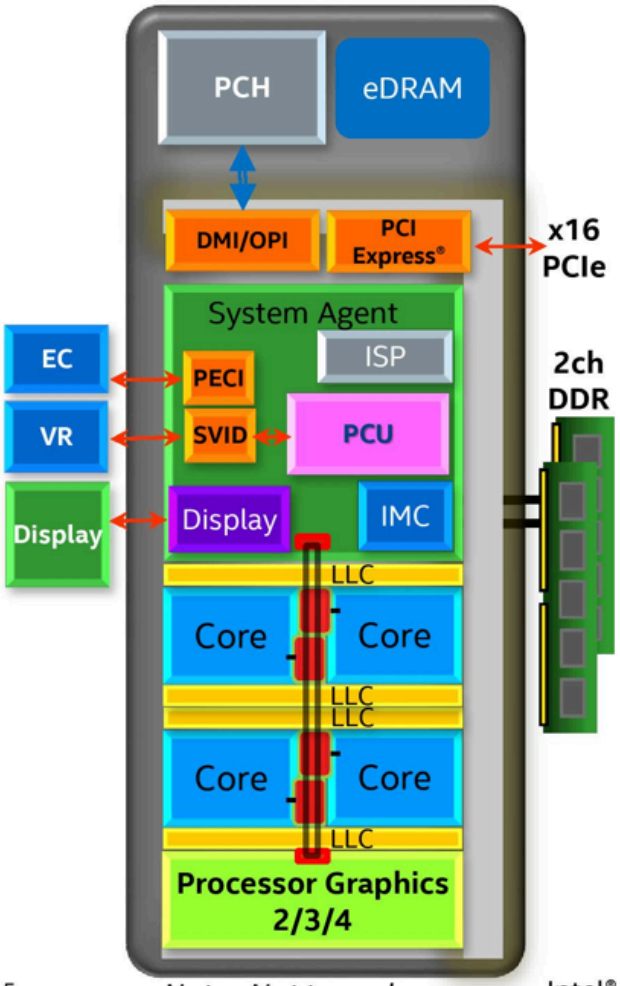
Figure 2. Die micrograph of Haswell-EX/EP. Both the EX and EP versions use the same 662mm2 chip. The high-core-count variants of Haswell-EP employ the same silicon as Haswell-EX, but with different I/O configurations. The latter chip enables more QPI links and fewer PCIe lanes, and it uses SMI2 memory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



Intel Skylake

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

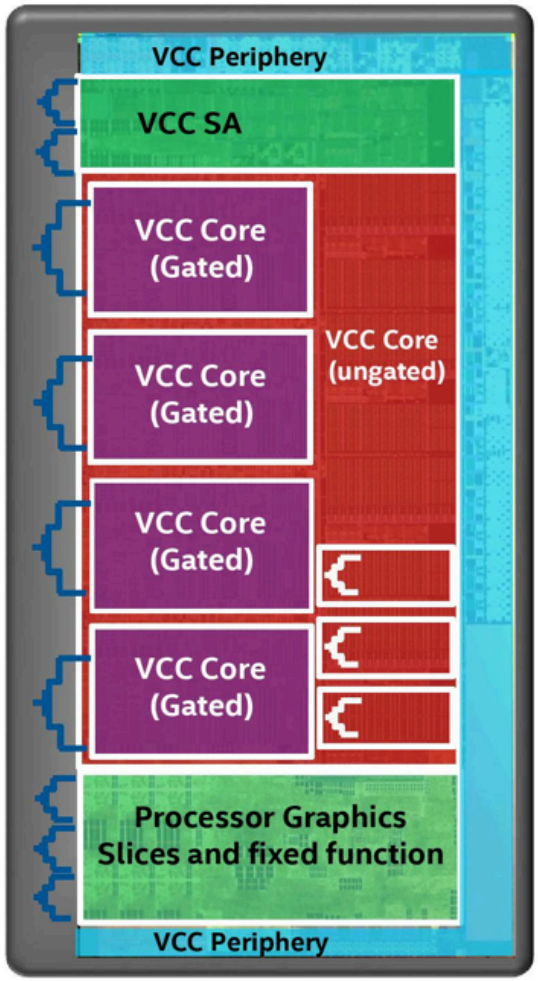
Skylake Overview – Power Management View



Note: Not to scale

- 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.

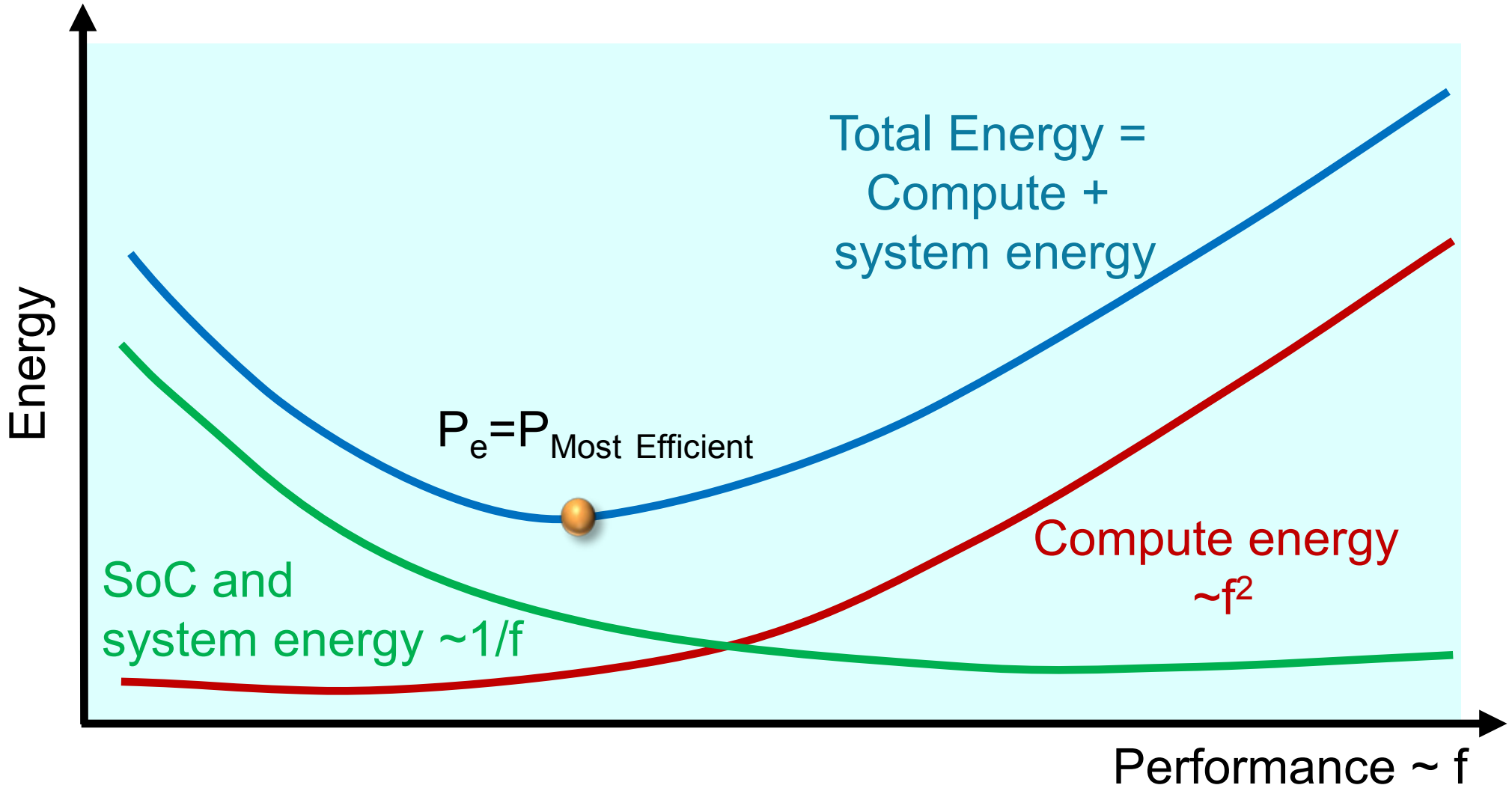
Skylake Power Management ID Card



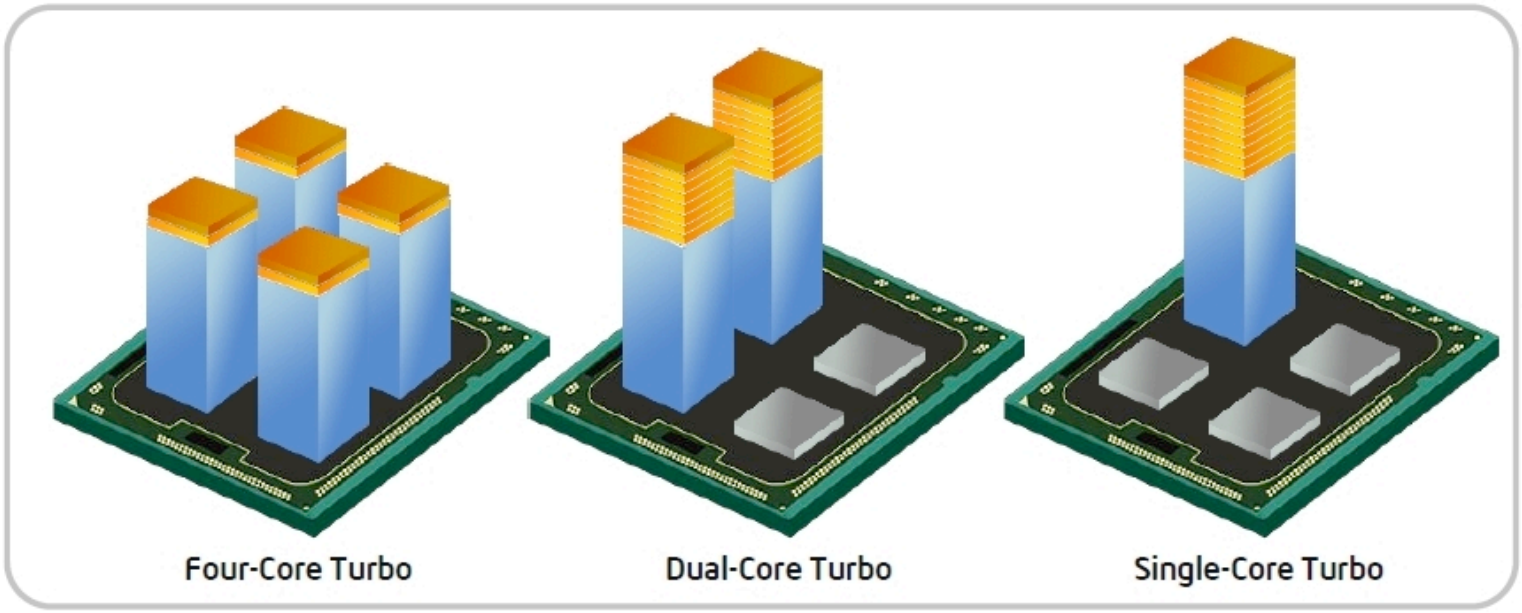
Note: Not to scale

- 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

Trading Off Energy vs Performance



Intel® Turbo Boost Technology 2.0



Efficient.

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

Dynamic.

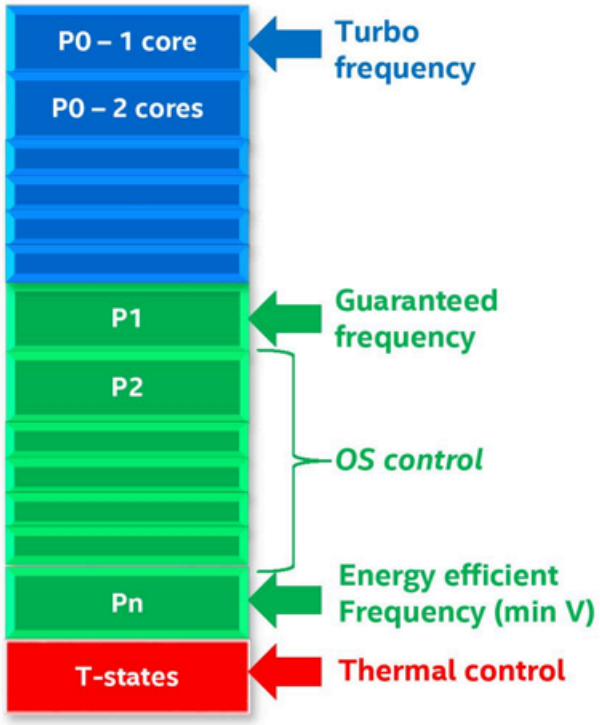
- Boosts power level to achieve performance gains for high intensity "dynamic" workloads

Intelligent.

- Power averaging algorithm manages power and thermal headroom to optimize performance

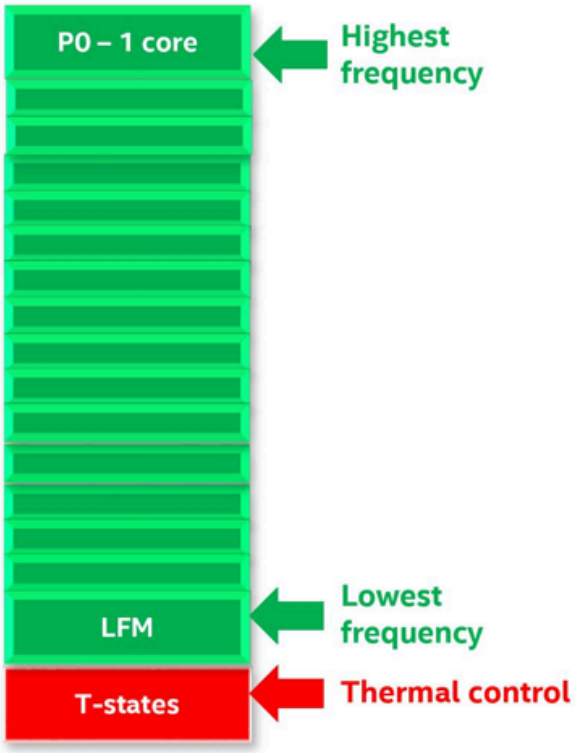
Intel® Turbo Boost Technology 2.0 delivers intelligent and energy efficient performance on demand

Legacy Energy-performance Control (P-state)



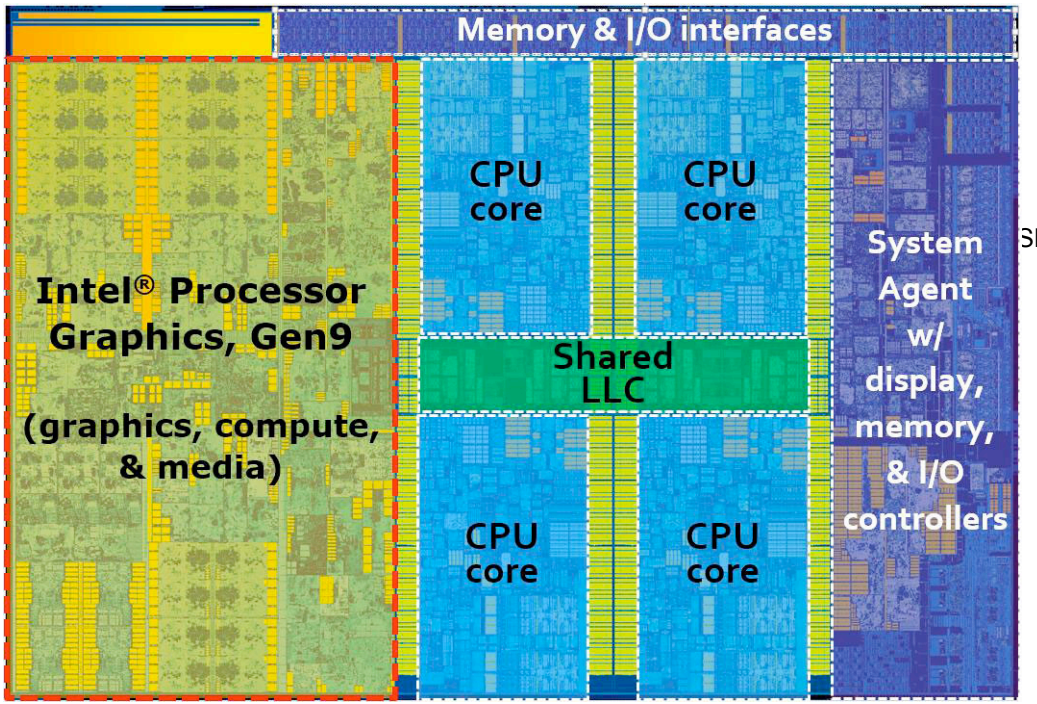
- DVFS – Intel SpeedStep® Technology
  - $P \sim V^2 \cdot f \cdot C_{dynn} + leakage(V) \sim f^3$
  - Performance comes at a cost of energy
- Operating System performs P-state control
  - P1-Pn frequency table enumerated via ACPI tables
  - Explicit P-state selection
- Typically demand based algorithm
  - Policies (AC/DC/Balanced, etc.)
  - Non regular workloads are hard to manage
  - Lower than Pn is used for critical conditions only

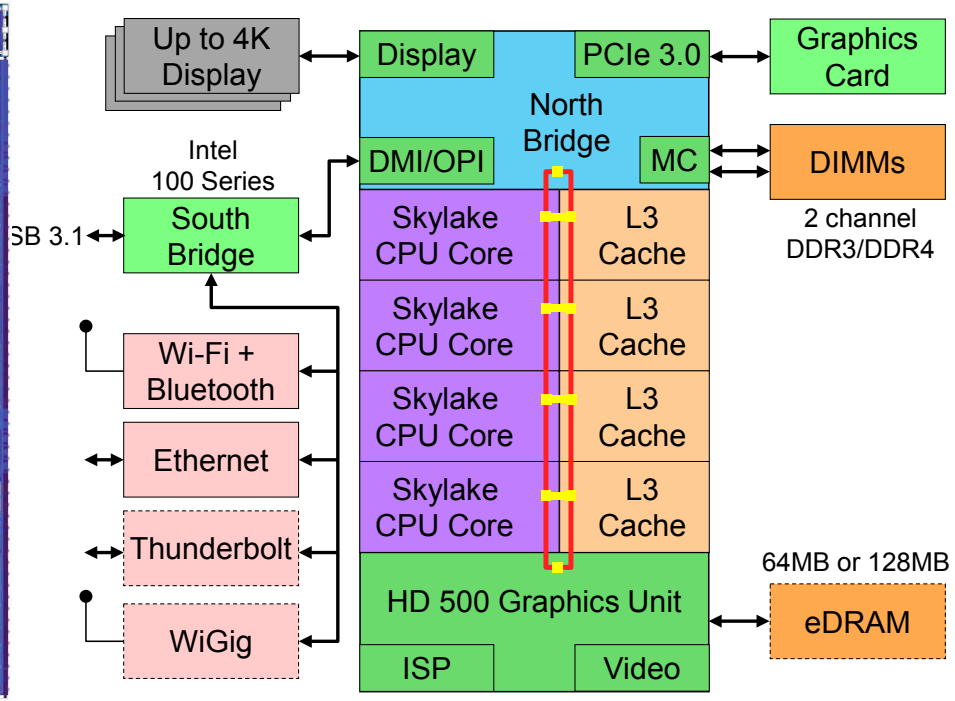
Intel® Speed Shift Technology - Hardware P-state



- 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

ECE 4750 Concepts





Processors

- Pipelining
- Superscalar Execution
- OOO Execution
- Register Renaming
- Memory Disambiguation
- Branch Prediction & Spec Exec
- SIMD Extensions
- Multithreading

Memories

- Multi-Level Caches
- Private/Shared Caches
- Consistency, Coherence
- Translation/Protection TLB

Networks

- On-Chip Ring, Inter-Socket All-to-All
- Routing, Buffering