Intel Skylake

Ryan Estep
Vishakh Suresh Babu

A brief introduction

- Intel Skylake microarchitecture
  - 2015
  - designed for 14nm process
  - preceded by Broadwell
- Intel development process
  - Tick-Tock
  - Process-Architecture-Optimization
- Tick was new process, adapting old architectures
- Tock was designing new microarchitecture

Intel development roadmap			
Cycle	Process	Introduction	Microarchitecture
Tock	32 nm	2010	Sandy Bridge
Tick	22 nm	2011	Ivy Bridge
Tock	22 nm	2013	Haswell
Tick/ Process	14 nm	2014	Broadwell
Architecture	14 nm	2015	Skylake (Client)
Optimization	14 nm+	2016	Kaby Lake
Optimization	14 nm++	2017	Coffee Lake, Skylake (Server)
Optimization	14 nm++	2018	Amber Lake, Whiskey Lake
Optimization	14 nm++	2019	Cascade Lake
Optimization	14 nm++	2020	Cooper Lake, Comet Lake
Optimization	14 nm++	2021	Rocket Lake

Source : [5]

Die shot

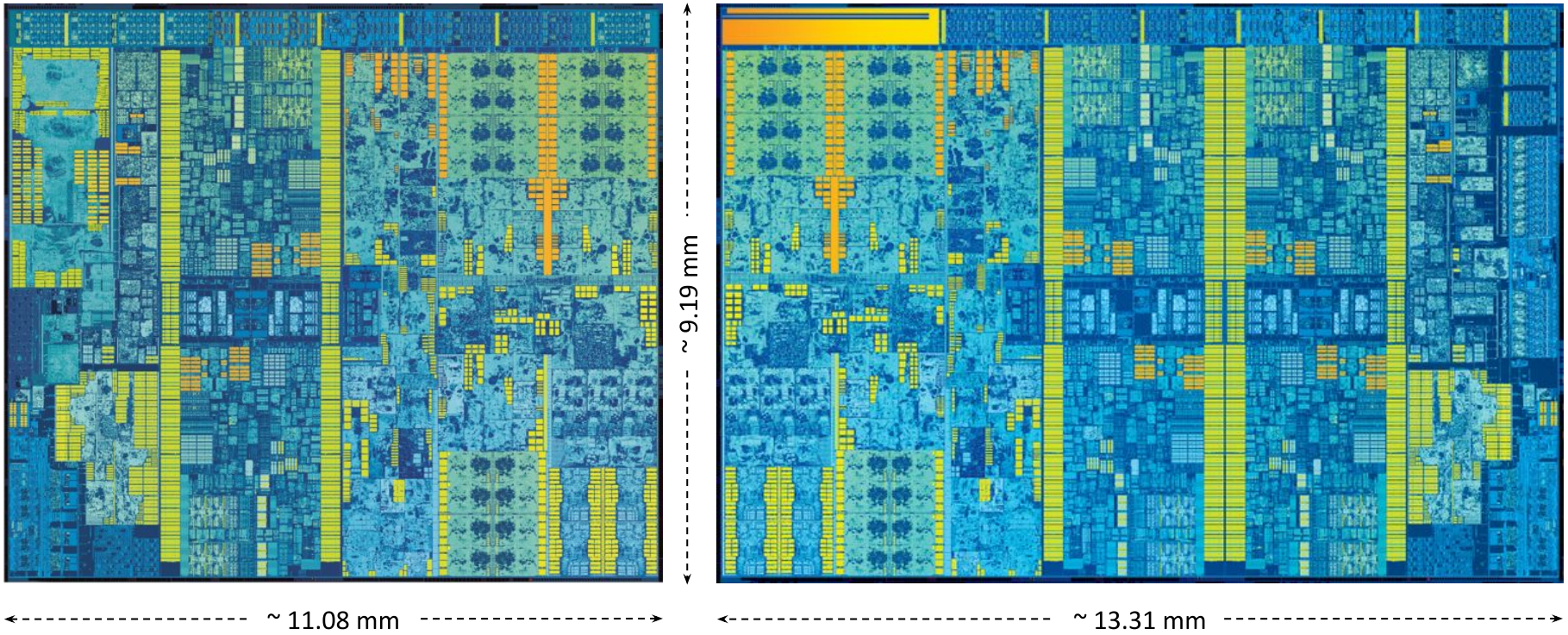


Image source : [5]

Die shot (cont'd)

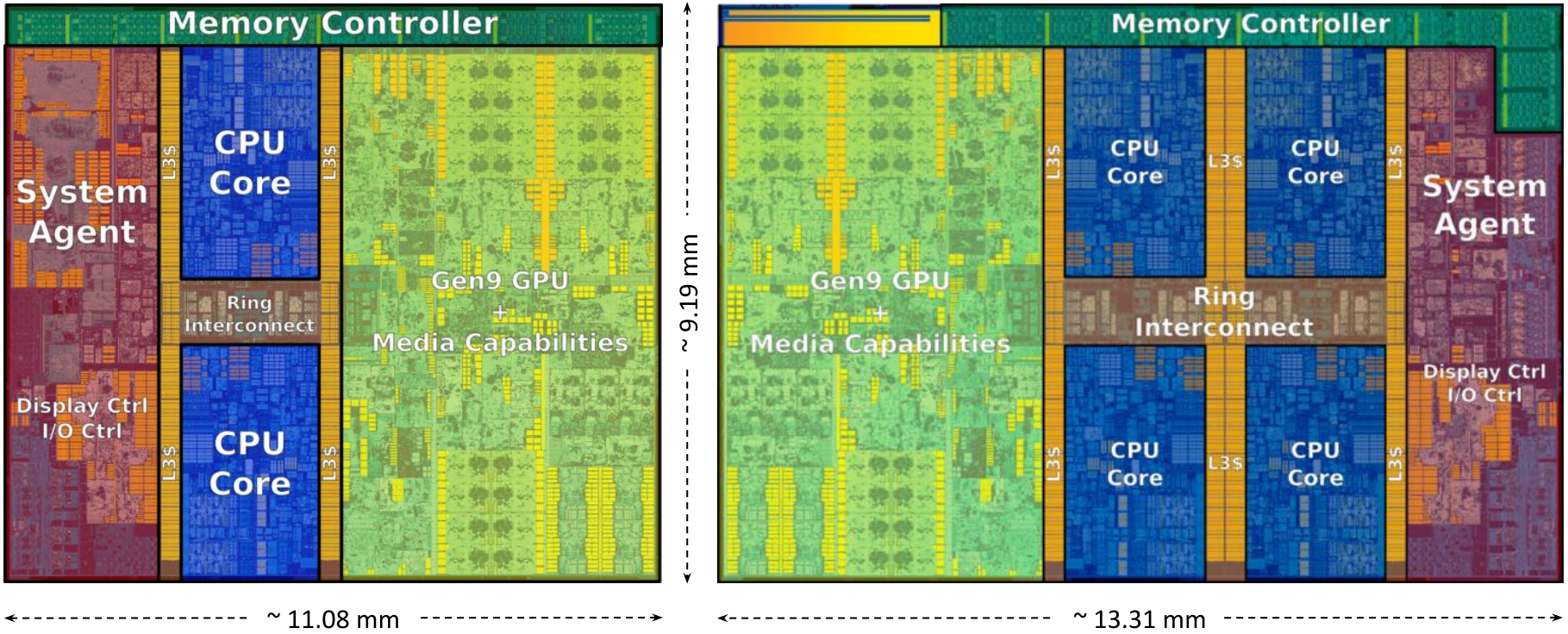


Image source : [5]

Pipeline

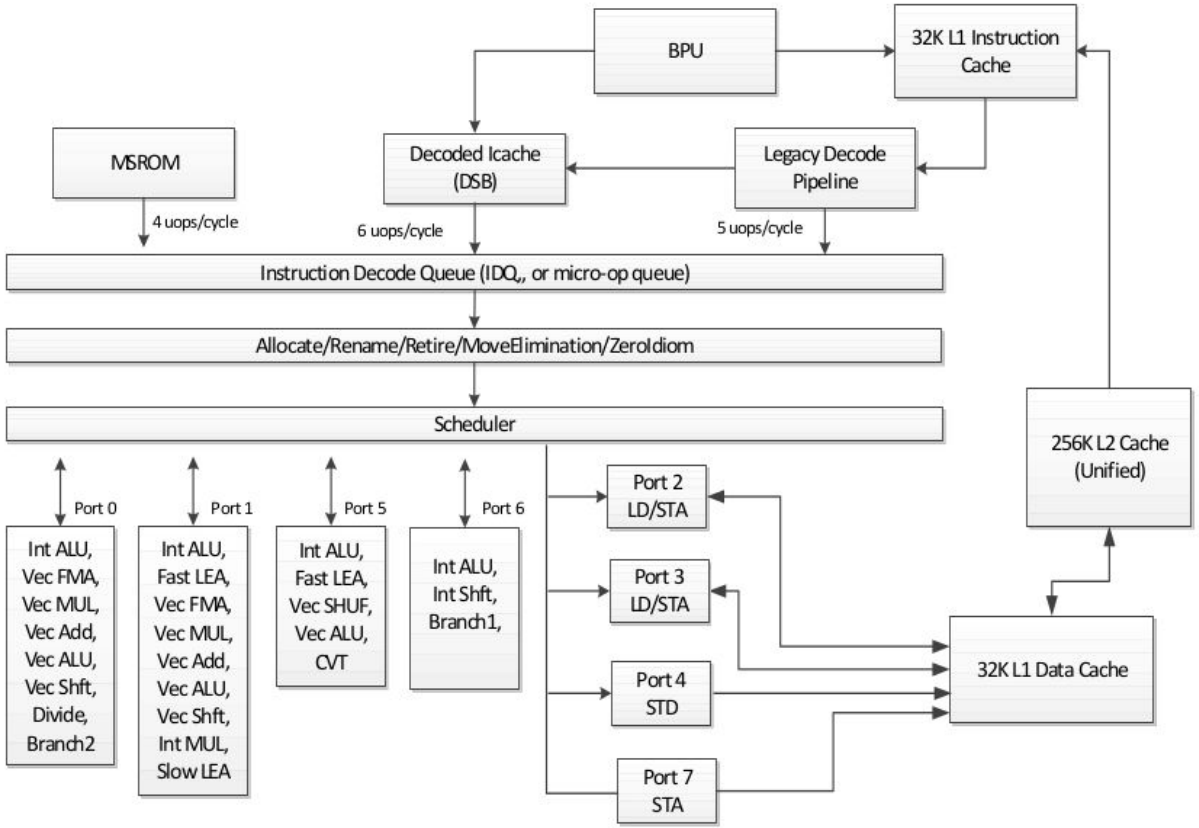


Image source : Figure 2-6, [1]

Pipeline (cont'd)

IO Fetch

IO Decode

OOO Issue and Late Commit

Integer/FP Functional Units with OOO Writeback

Load/Store Execution

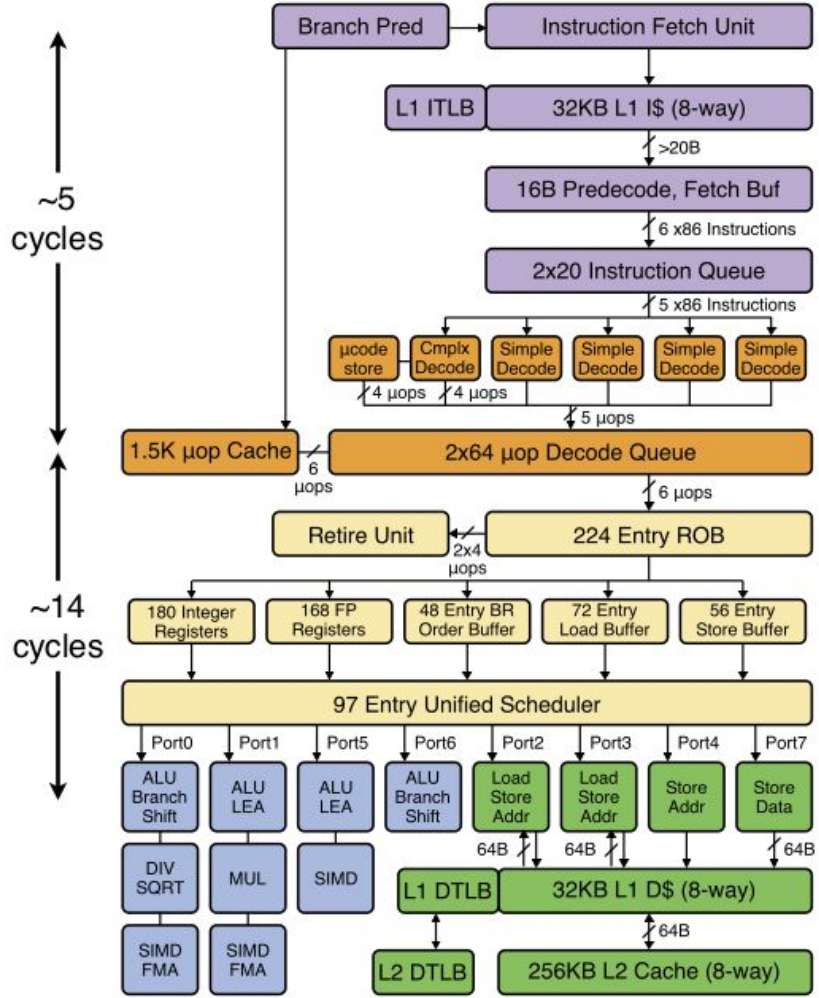


Image source : Page 21, [3]

Front end overview

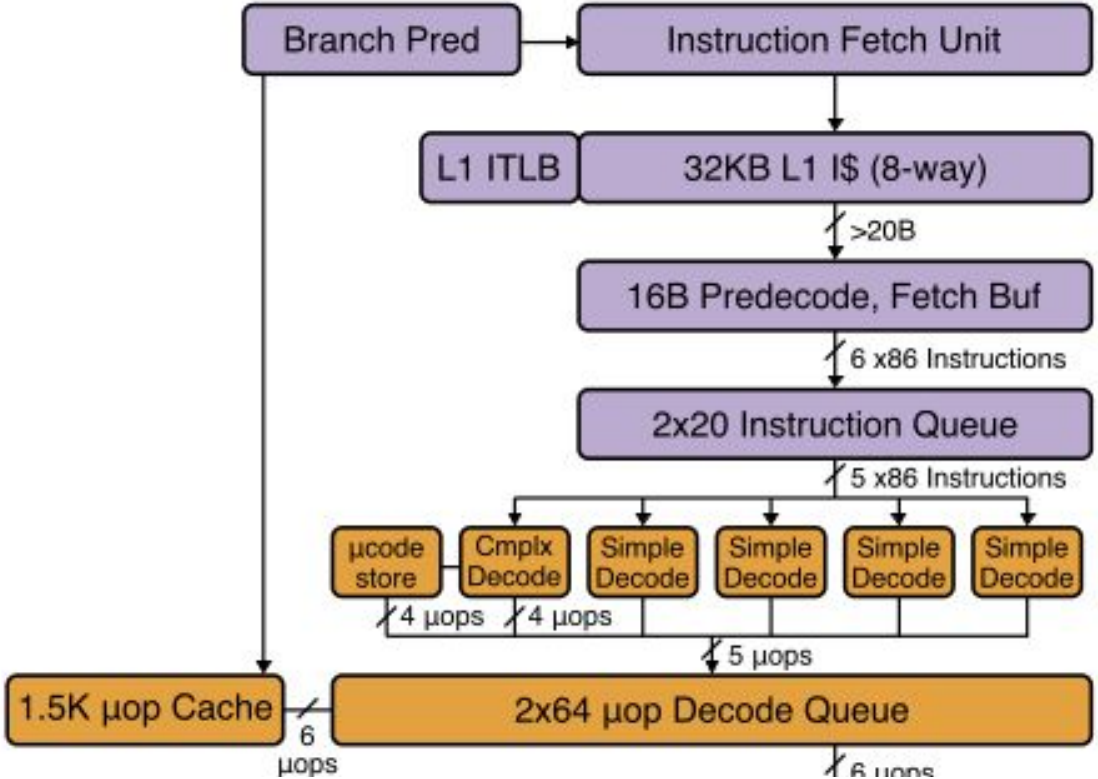
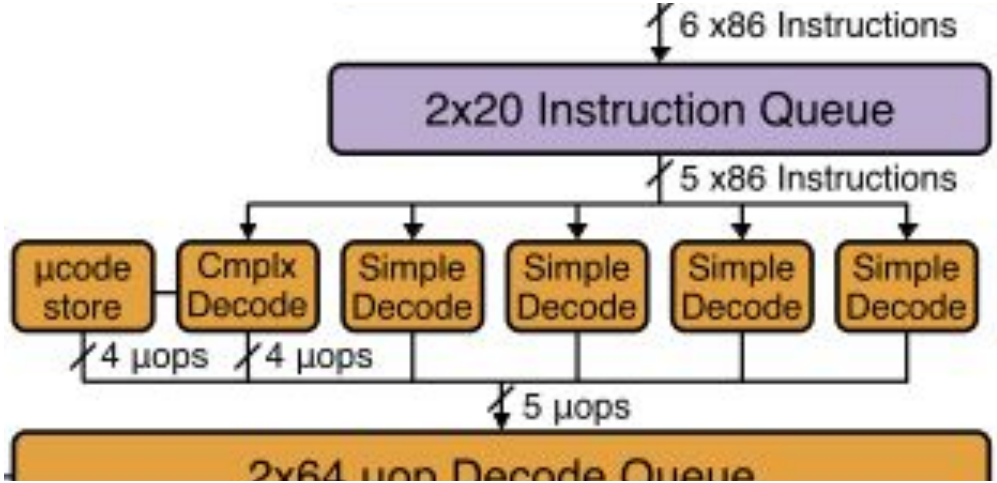


Image source : Page 21, [3]

MOP fusion & Decoding

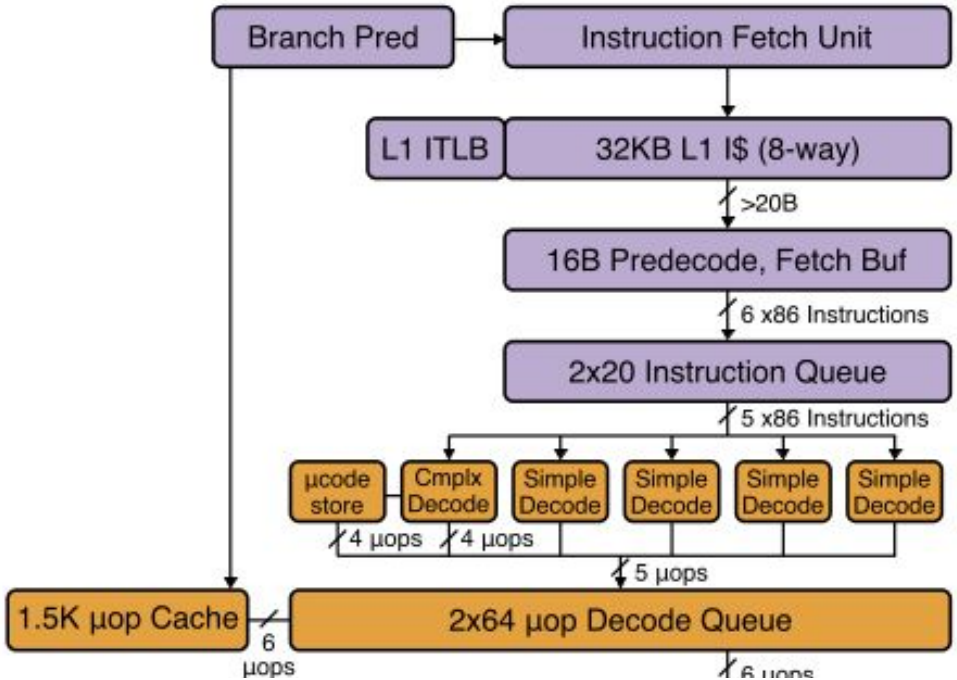
- Pre-decoding buffer
  - mark instruction boundaries
  - prefix decoding (e.g. branches)
- IQ has the ability to fuse MOPs into a single instruction
  - improved bandwidth
- Decode complex and variable MOPs into fixed size µ-ops



Source : Page 2-18, [1]; Page 22 [3] *Image source* : Page 21, [3]

µ-op cache

- µ-op cache (or Data Stream Buffer)-has cache lines of decoded µ-ops ready
- Bypasses the entire other path to IDQ (immensely preferred path)
- 1536 µ-ops → 32 sets, 8 lines/set, 6 µ-ops/line
- Competitively shared
- Hit rate > 80%
    - “Hot spots” ~100%



Source : Page 2-18, [1]; Page 150, [2] *Image source* : Page 21, [3]

Execution engine overview

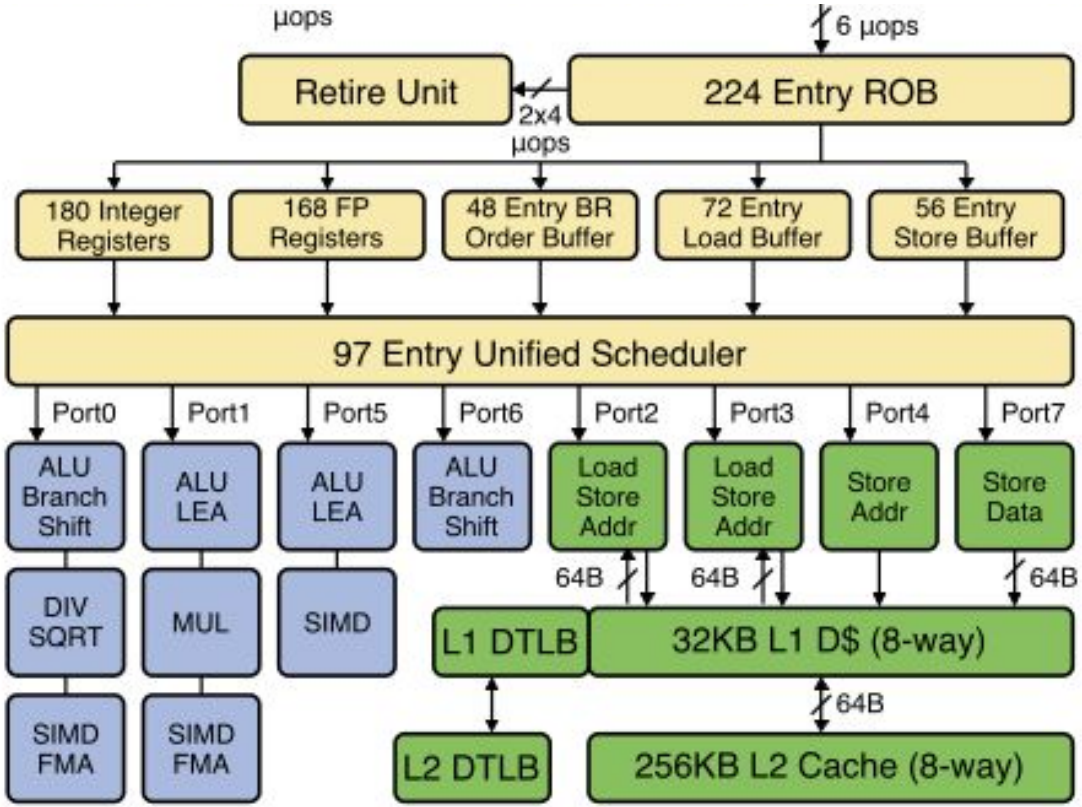
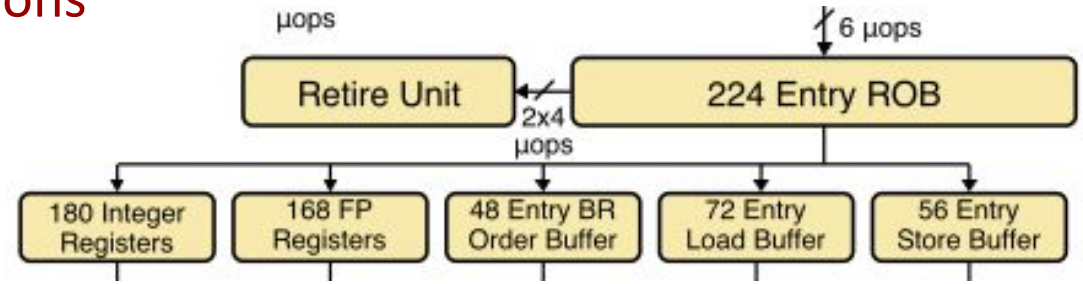


Image source : Page 21, [3]

Renaming & optimizations

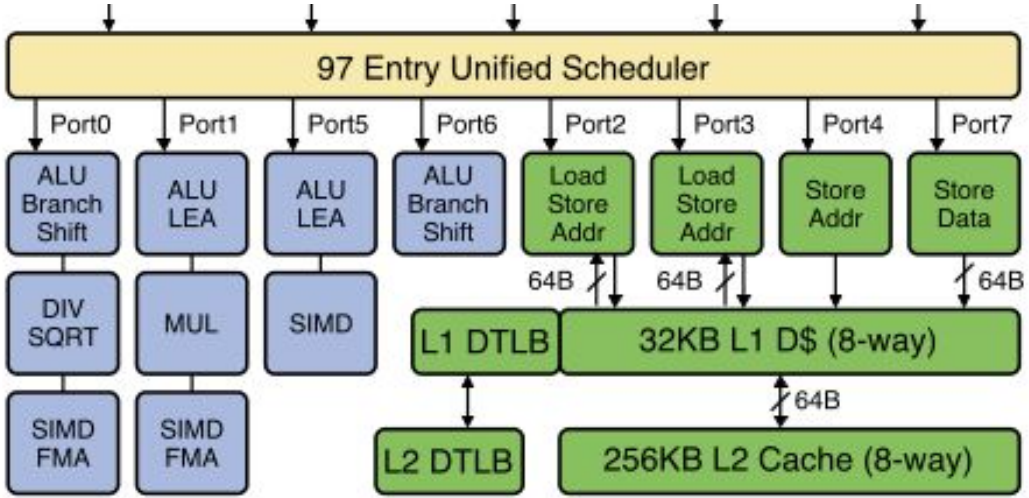


- Reorder Buffer for OoO Execution
  - in-order commit
  - increased size from predecessors
- Register Alias Table maps architectural registers to physical registers
- Speculative Execution
  - branch Order Buffer for mispeculation
- Renaming optimizations include Move Elimination, Zero or Ones Idiom

Source : Page 2-20, [1]; Page 24, 28 [3] *Image source* : Page 21, [3]

Scheduler & EUs

- Unified Reservation Station
- Scheduler for sorting µ-ops between ports and holding them until EU is ready
  - competitively shared and increased in size from predecessors
  - OoO oldest ready
- Ports are balanced between instructions for maximum performance



Source : Page 2-21, [1]; Page 24 [3] *Image source* : Page 21, [3]

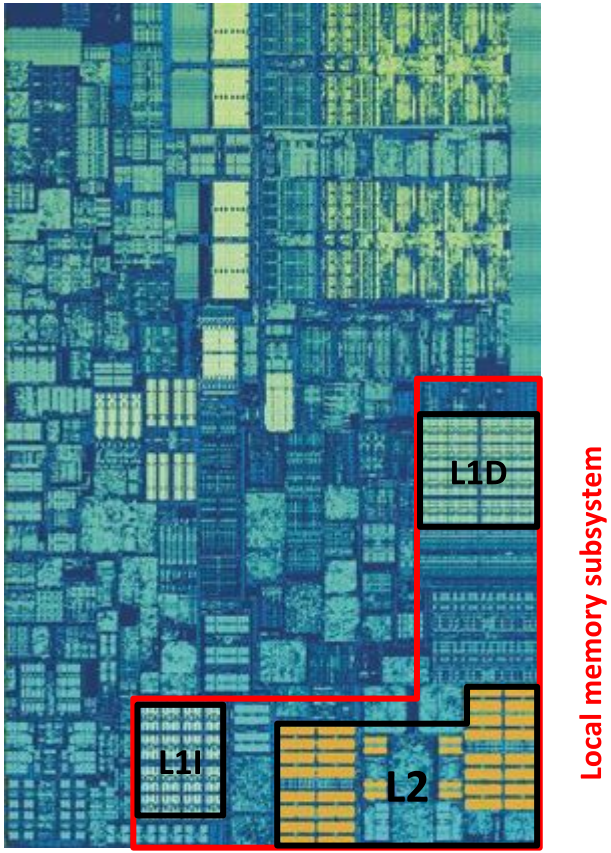
Memory subsystem overview

- Caches :
  - L0 μ-op cache
  - 3-level cache hierarchy
    - L1 cache
    - L2 cache
    - L3 cache/ LLC
  - eDRAM (on Skylake GPUs)
- TLB

Source : Page 2-19, [1]

Cache hierarchy

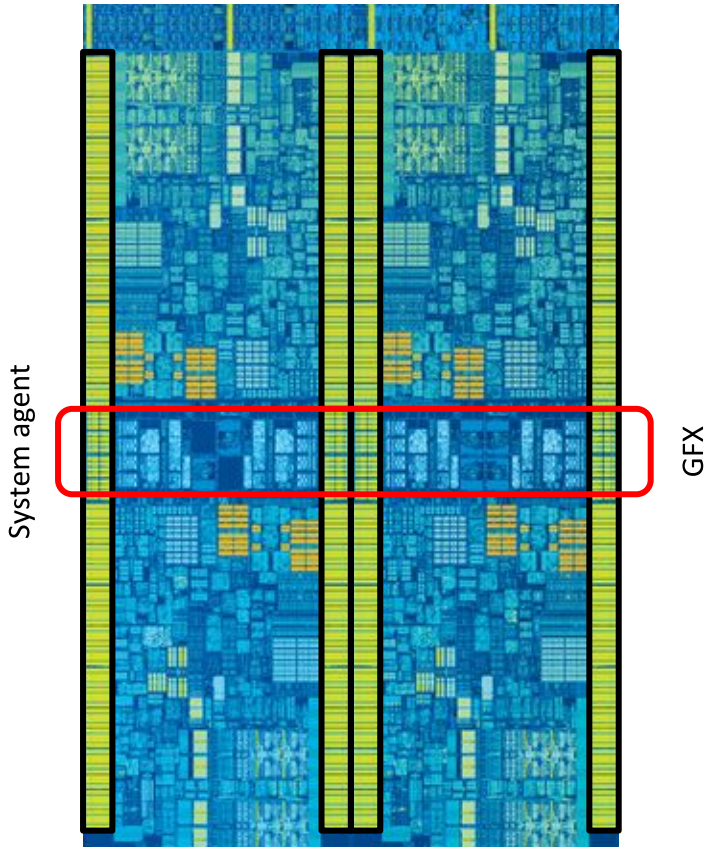
- *L1 cache* :
  - separate Instruction and Data caches
  - shared by 2 threads on the same core
  - L1D bandwidths :
    - load : 64 B/ cycle
    - store : 32 B/ cycle
- *L2 cache* :
  - unified
  - non-inclusive of L1
  - 64 B/ cycle bandwidth to L1



Source : Page 2-20, [1] *Image source* : [5]

Cache hierarchy (cont'd)

- *L3 cache/ LLC* :
  - inclusive of L2
  - shared among all cores
  - split into slices connected by 4 rings :
    - data, request, acknowledgement & snoop
    - to increase the bandwidth
    - uses an undocumented hash function, mapping cache lines almost evenly across slices
  - per core bandwidths (@ ring clock) :
    - read & write : 32 B/ cycle (two times that of Haswell



Source : Page 2-20, [1] *Image source* : [5]

Cache parameters

Level	Capacity	Associativity	Line size (bytes)	Fastest latency (cycles)	Update policy
L1I	32 KB	8	64	N/A	N/A
L1D	32 KB	8	64	4	writeback
L2	256 KB	4	64	12	writeback
L3	Up to 2 MB per core	Up to 16 ways	64	44	writeback

Data source : Page 2-20, [1]

Cache parameters (cont'd)

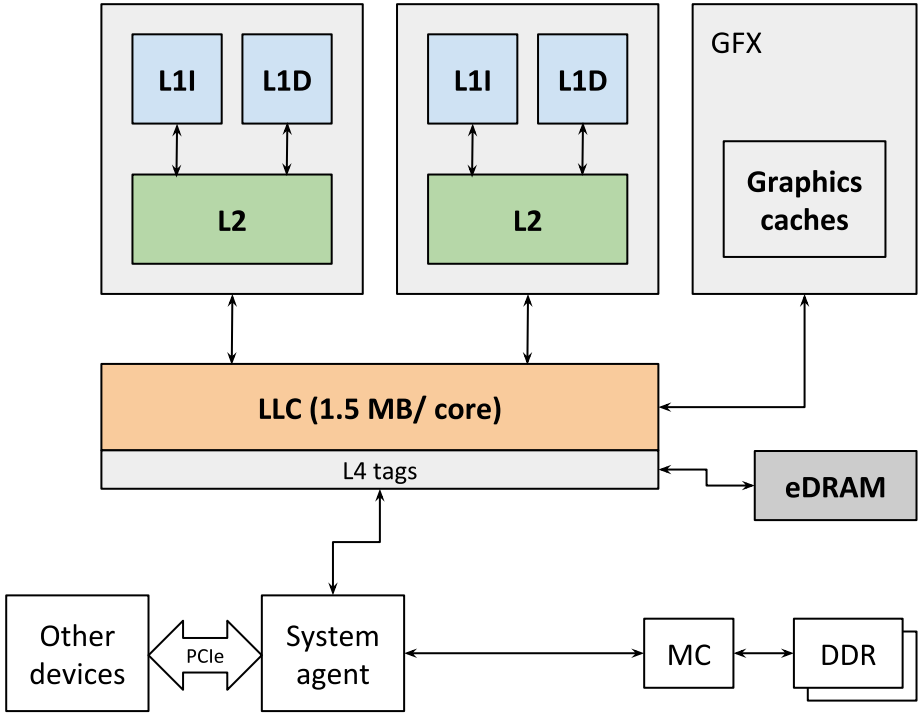
- L2 cache has be reduced from an 8-way (in Haswell) to 4-way set associative.
  - Theoretically, half the associativity ⇒ ↑ in miss rate.
  - Practically,
    - ↓ in power on a successful data access
    - saves area on the silicon die
  - ↑ in miss rate countered by
    - doubling bandwidth to L2 misses
    - improvement in cache and page miss handling
  - Net effect : A performance comparable to Haswell @ a reduced power consumption.

Source : Page 2-20, [1]

eDRAM based cache

Haswell & Broadwell :

- eDRAM access through L4 tags in LLC.
- eDRAM acts like a victim cache for LLC.
- eDRAM fetches from processor :
  - earlier tag checking
  - faster
- Other devices require eDRAM data :
  - go through LLC & do the L4 tag conversion
  - slower

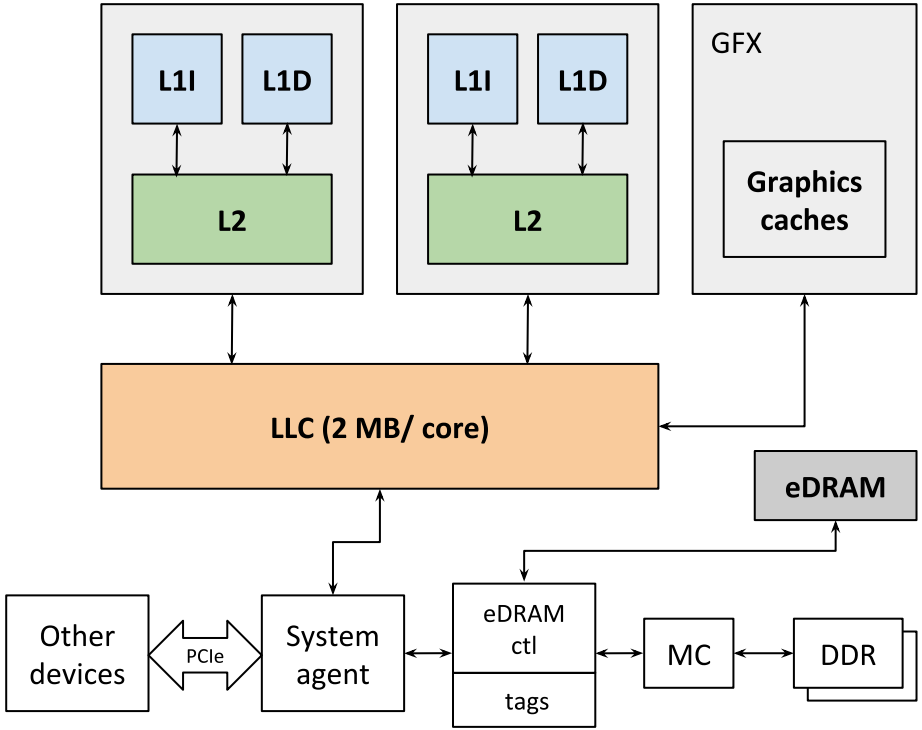


Source : [4] *Image reference* : Page 17, [7]

eDRAM based cache (cont'd)

Skylake :

- eDRAM behaves as a buffer!
- Other devices requiring eDRAM data do not need to navigate through the on-chip LLC.
- Graphics workloads need to circle around the system agent.
- All memory accesses through MC get looked up in eDRAM.
  - hit : use value from eDRAM.
  - miss : value stored on the eDRAM.
- Available in 2 sizes : 64 GB & 128 GB
  (48 EU) (72 EU)



Source : [4] *Image reference* : Page 18, [7]

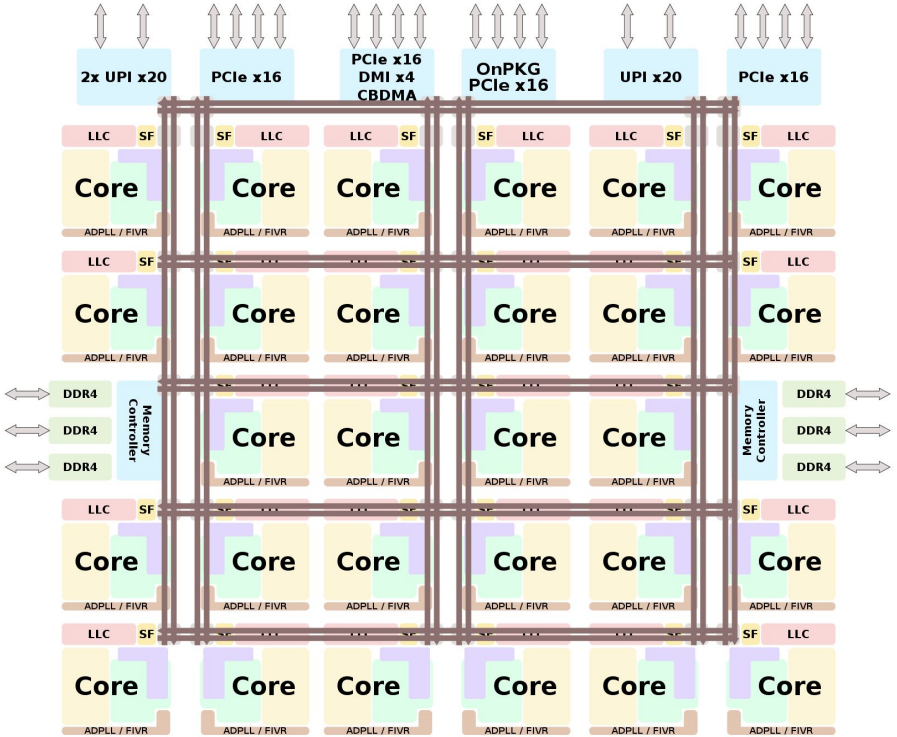
TLB parameters

Level	Page size	Entries	Associativity	Partition
ITLB	4 KB	128	8	dynamic
	2 MB/ 4 MB	8 per thread		fixed
DTLB	4 KB	64	4	fixed
	2 MB/ 4 MB	32	4	fixed
	1 GB	4	4	fixed
STLB	4 KB and 2 MB/ 4 MB	1536	12	fixed
	1 GB	16	4	fixed

Data source : Page 2-20, [1]

Parallelism summary

- Client Dual-core or quad-core
- Dual-thread
  - competitively shared
- Skylake (Server)
  - doubled bandwidth after front-end
  - mesh Interconnect
  - up to 28-cores (56 threads)
  - AVX-512



Source : Page 17, 29 [3]; [6] *Image source* : [6]

More special features

- Configurable core
  - Client (14nm)
  - Server (14nm+) ⟶ higher drive current, lower power
- Focus on graphics
  - wanted to improve performance and power consumption during video…
  - new IPU/ISP in mobile units
- Security technology
  - protection from attacks
  - SGX, MPX --now deprecated

- Speed Shift power management
- Turbo Boost Technology
  - turbo mode : cores run faster than the rated frequency
  - algorithmic overclocking

Source : Page 18, 19 [3]; [5]; [6]; [12]

Power management

- Previously, OS responsible for DVFS based on the current workload.
  - eg: CPU utilisation peaked ⇒ ↑ f to cope up with it
  - limitation : granularity of OS response time - 10s of milliseconds
- “Speed Shift” - new power management.
  - quickly alternate core frequencies in response to power loads
  - a new unit called Package Control Unit (PCU)
    - full-fledged microcontroller
    - collects and tracks many SoC statistics
  - speed shift kicks in ~1 ms

    OS bases P-state control can be as slow as 30 ms

Source : [5]

Skylake vs Kaby Lake

- *Turbo boost* :
  - Skylake : 3.1 GHz
  - Kaby Lake : 3.5 GHz
- *Encoding & decoding video codecs* (10-bit 4K HEVC video codecs as well as 4K VP9) :
  - Skylake : software support
  - Kaby Lake : hardware support

Playing	Battery-life improvement in Kaby Lake	Power consumption (W)	
		Skylake	Kaby Lake
10-bit 4K HEVC video	2.6 x	10.2	0.5
4K video on YouTube	1.7 x	5.8	0.8

Data source : [8]

References I

1. “Intel 64 and IA-32 Architectures Optimization Reference Manual.” *Intel Corporation*, https://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-optimization-manual.pdf.

2. Fog, A. “The microarchitecture of Intel, AMD and VIA CPUs.” *Technical University of Denmark*, https://www.agner.org/optimize/microarchitecture.pdf.

3. Batten, C. “ECE4750 Computer Architecture Intel Skylake.” *Cornell University*, https://www.csl.cornell.edu/courses/ece4750/2016f/handouts/ece4750-section-skylake.pdf.

4. The Intel Skylake Mobile and Desktop Launch, with Architecture Analysis.” *AnandTech*, https://www.anandtech.com/show/9582/intel-skylake-mobile-desktop-launch-architecture-analysis/.

5. “Skylake (Client) - Microarchitectures - Intel.” *WikiChip*, en.wikichip.org/wiki/intel/microarchitectures/skylake_(client).

6. “Skylake (Server) - Microarchitectures - Intel.” *WikiChip*, en.wikichip.org/wiki/intel/microarchitectures/skylake_(server).

7. “Skylake (Client) - Microarchitectures - Intel.” *WikiChip*, en.wikichip.org/w/images/8/8f/Technology_Insight_Intel%E2%80%99s_Next_Generation_Microarchitecture_Code_Name_Skylake.pdf.

8. “10 key things to know about Intel’s Kaby Lake CPUs.” *PCWorld*, https://www.pcworld.com/article/3111186/10-key-things-to-know-about-intels-kaby-lake-cpus.html.

References II

9. “Intel Core i7-6700HQ Processor Technical Specifications.” *Intel Corporation*, https://www.intel.com/content/www/us/en/products/processors/core/i7-processors/i7-6700hq.html.

10. “Intel Core i7-6700T Processor Technical Specifications.” *Intel Corporation*, https://www.intel.com/content/www/us/en/products/processors/core/core-vpro/i7-6700t.html.

11. “Intel Core i9-9960X X-series Processor Technical Specifications.” *Intel Corporation*, https://www.intel.com/content/www/us/en/products/processors/core/x-series/i9-9960x.html.

12. “Intel® Turbo Boost Technology 2.0” *Intel Corporation*, https://www.intel.in/content/www/in/en/architecture-and-technology/turbo-boost/turbo-boost-technology.html

Supplementary slides

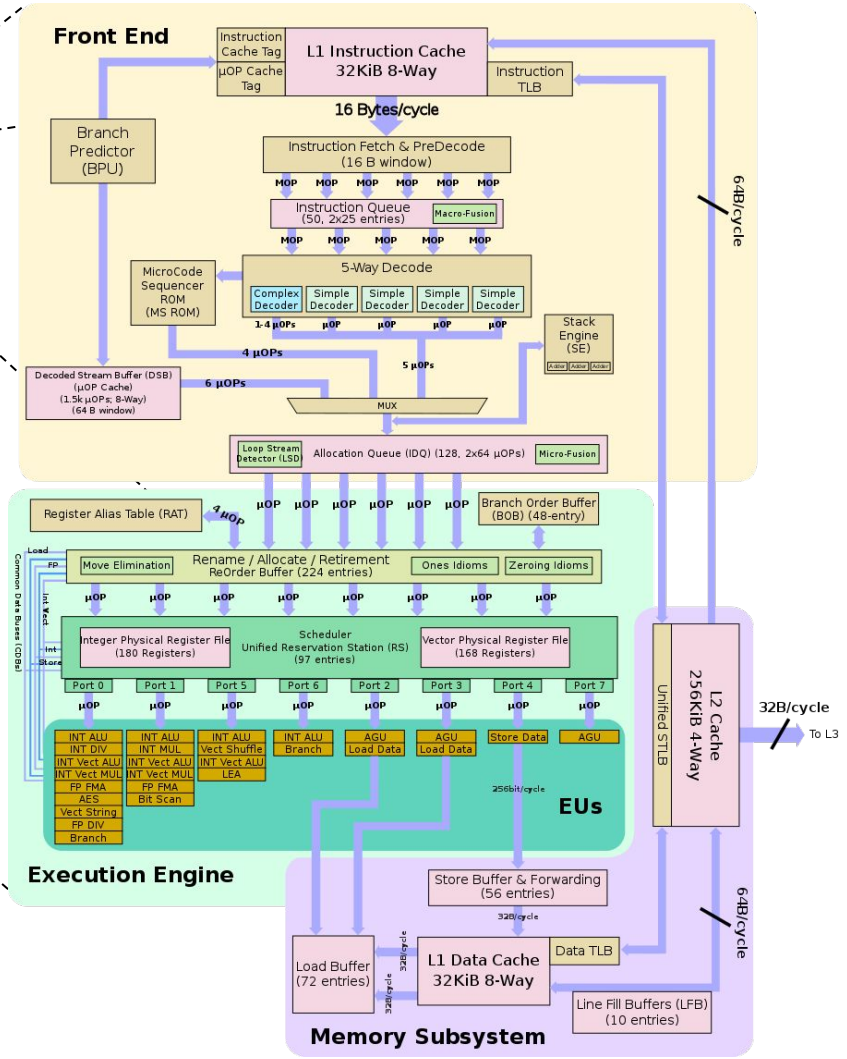
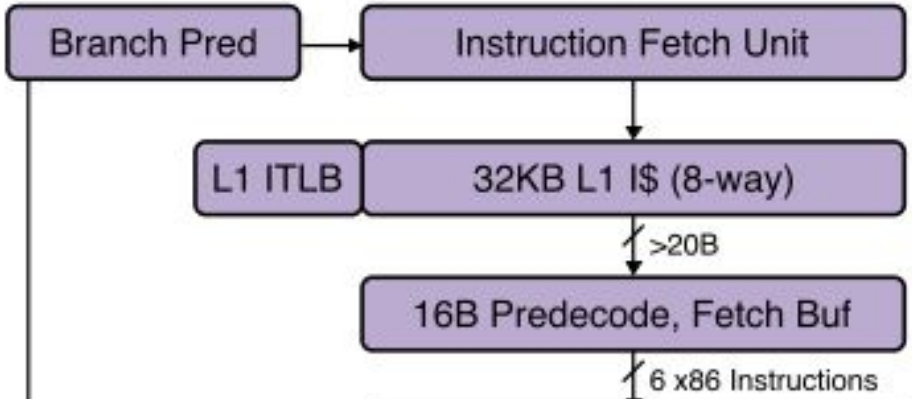


Image source : [5]

Fetch & Pre-decoding

- Fetching is dual-thread
  - Shared evenly
- 16B chunks of code
- Pre-decoding buffer
  - Mark instruction boundaries
  - Prefix decoding (e.g. branches)
- BPU-branch prediction
  - Further “vision” than predecessors

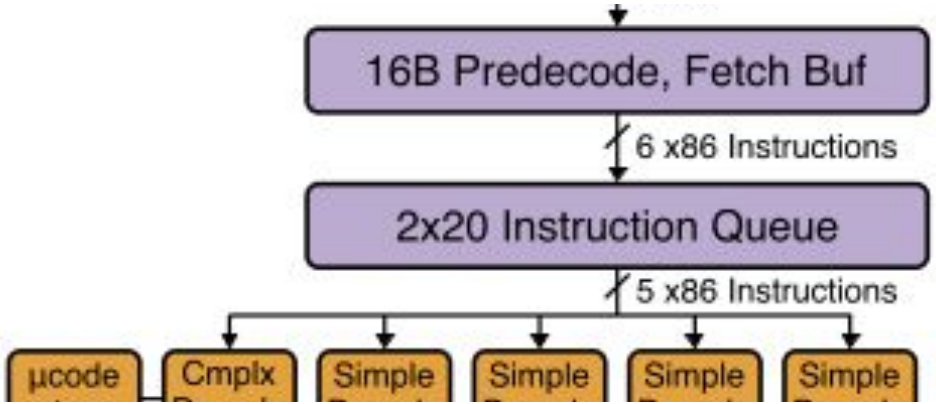


Source : Page 2-16, 2-17 [1]; Page 22 [3]

Image source : [3]

Instruction queue & MOP fusion

- 25 entries/thread
- Instruction queue holds macro-ops until the decoder is ready
- Has the ability to fuse MOPs into a single instruction
    - Improved bandwidth

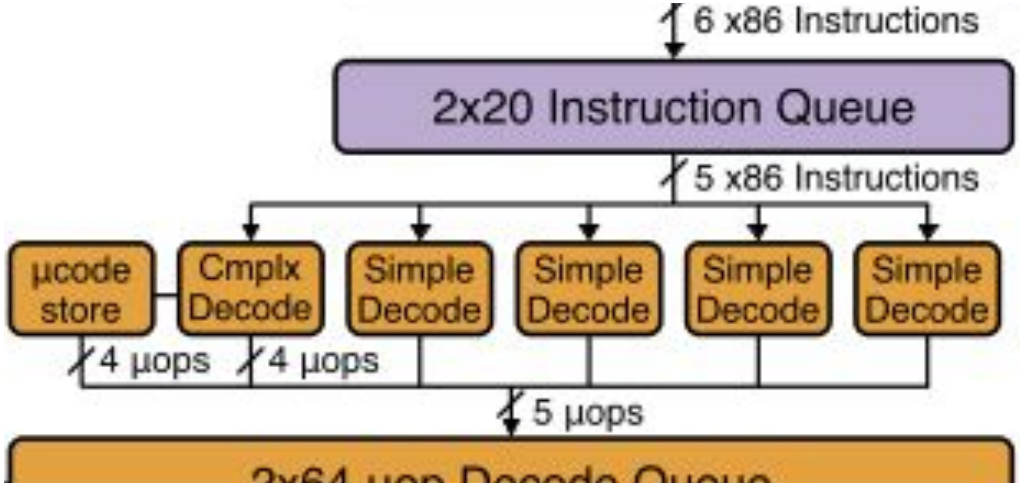


Source : Page 2-17, 2-18, [1]; Page 22 [3]

Image source : [3]

Decoding

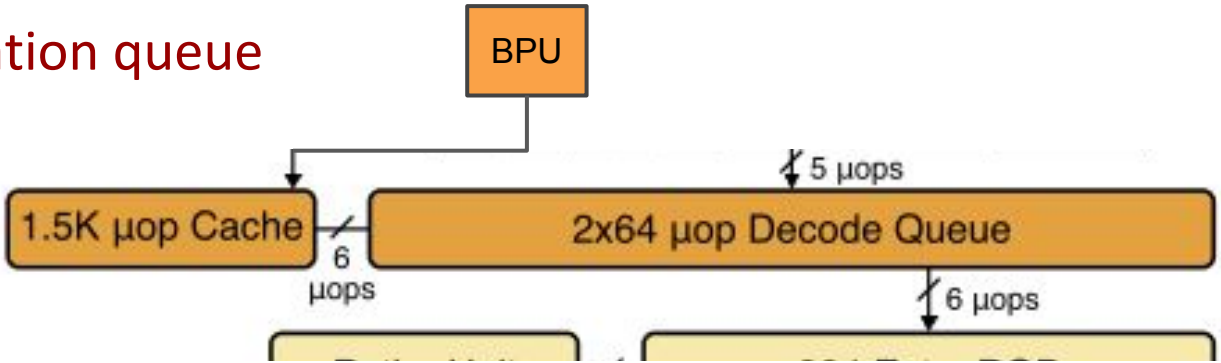
- 5-way decoder
  - 1 complex and 4 simple
- Decodes complex and variable MOPs into fixed size µ-ops
- Supports 5 µ-ops sent down the pipeline
- Complex decoder=1-4 µ-ops
- More than 4 µ-ops->microcode sequencer



Source : Page 2-18, [1]; Page 22 [3]

Image source : [3]

µ-op cache & Allocation queue



- Allocation queue (or Instruction Decode Queue)-interface between the in-order fetch/decode and OoO execution engine
  - Partitioned (non-competitive) 64 entries/thread
  - Loop stream detector detects loops and repeats µ-ops (server only)
- µ-op cache (or Data Stream Buffer)-has cache lines of decoded µ-ops ready
  - Bypasses the entire other path to IDQ (immensely preferred path)
  - 1536 µ-ops--32 sets, 8 lines/set, 6 µ-ops/line
- Competitively shared

Source : Page 2-18, Page 2-20 [1]; Page 150, [2] *Image source* : [3]

Intel Turbo Boost

- Some programs are memory-bound & some CPU-bound

  ⇒ need not always run the CPU at max frequency.

- Turbo Boost as an energy-η soln to this problem :
  - run at base clock speed for lighter workloads.
    - less power consumption
    - less heat dissipation
  - dynamically switch to a greater clock rate for heftier loads.
  - upto a max turbo boost frequency.
    - still within the safe power and temp limits.
  - “algorithmic overclocking”

Source : [12]

Intel Turbo Boost (cont'd)

Processor	Processor base frequency (GHz)	Max turbo boost frequency (GHz)	Comments
Intel Core i7-6700HQ	2.6	3.5	Mobile processor
Intel Core i7-6700T	2.8	3.6	Mainstream desktop processor
Intel Core i9-9960X X-series	3.1	4.4	High end processor

Data source : [9], [10], [11]