Intel Micro-Architecture (Ice Lake/Sunny Cove)

Goal: Fast overview of one of Intel's main processors

Highlights:

- Superscalar
- Speculative Execution
- Register Renaming
- 14-deep pipeline

A bit of x86 instruction set

X86 Milestones – Evolution of the instruction set

Some relevant steps (not all):

1974: 8080 8-bit, 2MHz, 6k transistors

1978: 8086 16-bit, 5-10MHz, 29k transistors

1980: 8087 floating point coprocessor

1982: 80286 16-bit, 6-12.5MHz, 134k transistors, 24-bit address space

1985: 80386 32-bit, 16-33MHz, 256k transistors, 256B code cache

1989: 80486 32-bit, 25MHz, 1.2M transistors, 8KB L1, 5-stage pipe

1992: Pentium 32-bit, 60-66MHz, 3.3M transistors, 16KB L1, L2, branch predict, superscalar (CPI=0.5).

1995: Pentium Pro, 32-bit, 200MHz, 5.5M transistors, CPI=1/3, 12-stage pipeline, out-of-order execute, predicated instructions, 4-bit branch history.

1996: Pentium MMX, 150-233MHz, 4.5M transistors, SIMD (single instruction multiple data) instructions.

2000: Pentium 4, 1.3-3.0GHz, 42M transistors, 20-deep pipeline, symmetric multithreading

2006: Core 2 Duo, 64-bit, 1.0-2.3GHz, 291M transistors, 14-stage pipeline, multi-core

2008: Nehalem/i7, 1.73-3.46 GHz, 2.6B transistors, quad/octo-core, SMT, shared L3

2013: Haswell/i7, 1.9-4.6GHz, 1.4B transistors, 2-8 cores, on-die GPU, 2 branch units

2019: Ice Lake/Sunny Cove …

Backwards compatible

X86 Operands

16x64-bit registers plus special-purpose registers (Flag, segments, etc).

2-operand instructions:

Source/Destination operand Type	Second source operand
Register	Register
Register	Immediate
Register	Memory
Memory	Register
Memory	Immediate

Multiple data memory addressing modes:

Mode	Description
Register Indirect	Mem[Reg[id]]
Base + 8/32/64-bit displace	Mem[Reg[id]+displace]
Base + scaled index	Mem[Reg[id]+Reg[id2]*2^{scale}], scale=0..3
Base + scaled + 8/32/64 displace	Mem[Reg[id]+Reg[id2]*2^{scale}+displace]

Instructions

Data movement: Move, push, pop

Arithmetic & logic: test, integer, decimal math, etc.

Control Flow: conditional & unconditional jumps, calls, returns

String instructions: string move, compare (legacy from 8080, not much used)

Streaming SIMD (MMX, SSE)

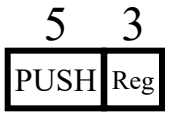
- Single instruction, multiple data (i.e. 4x8-bit adds simultaneously)
- Intended for multi-media

Instruction Encoding

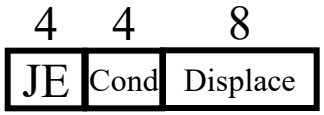
Range from 1-byte to 17-byte!

- Opcode says bitwidth of 8-bit/32-bit…
- May have extra byte to indicate addressing mode
- Extra byte for scaled index mode.

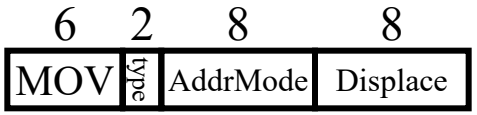
Stack Operation (PUSH)



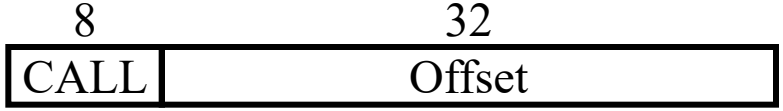
Conditional Branch (JE)



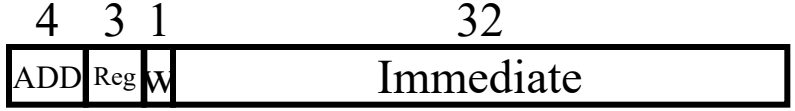
LDUR/STUR (MOV)



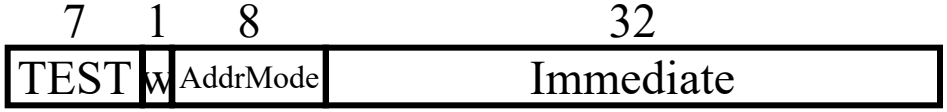
Branch w/link (CALL)



ADDI (ADD)



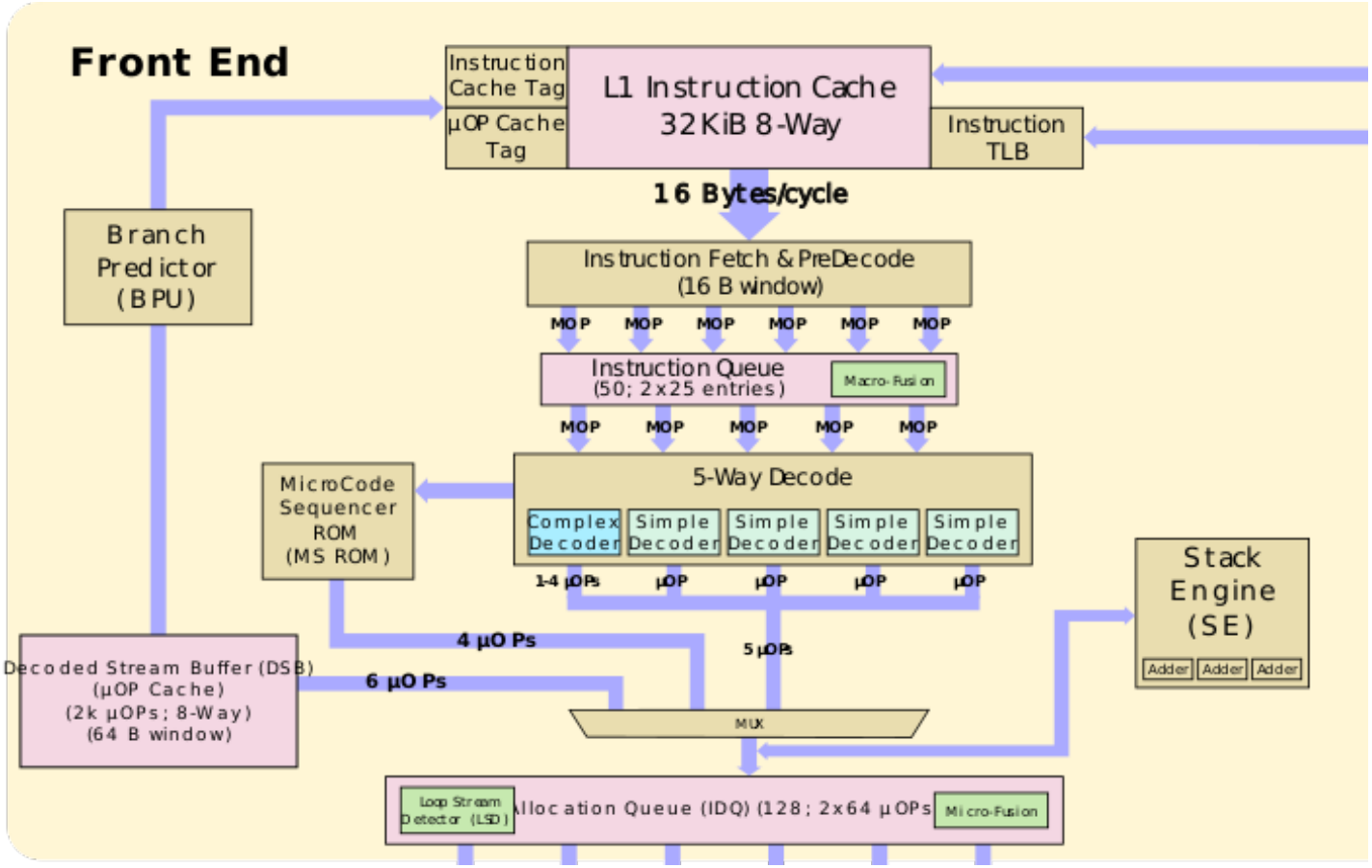
Set flag with AND (TEST)



Instruction Decoding

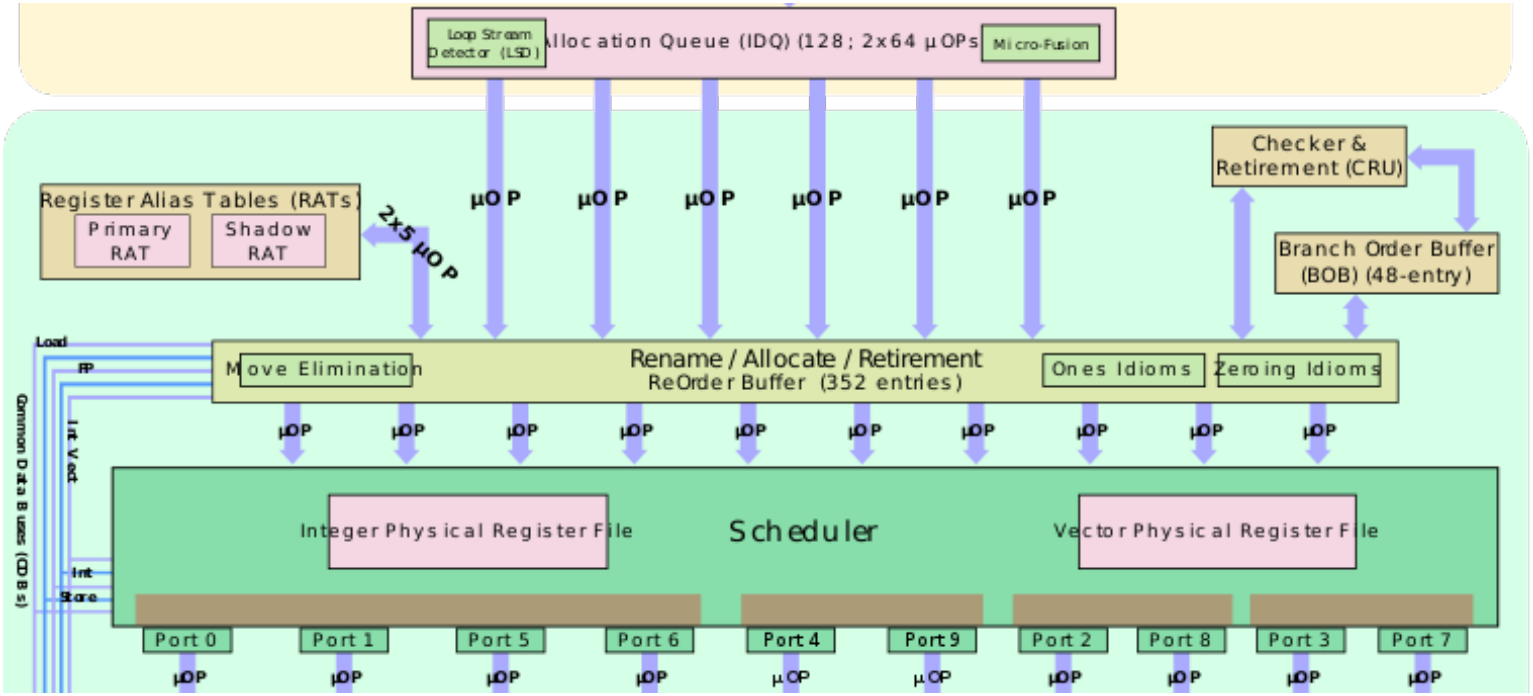
X86 instructions are essentially pseudo-instructions, converted to multiple RISC-like micro-ops

CPU decodes X86 into micro-ops at runtime



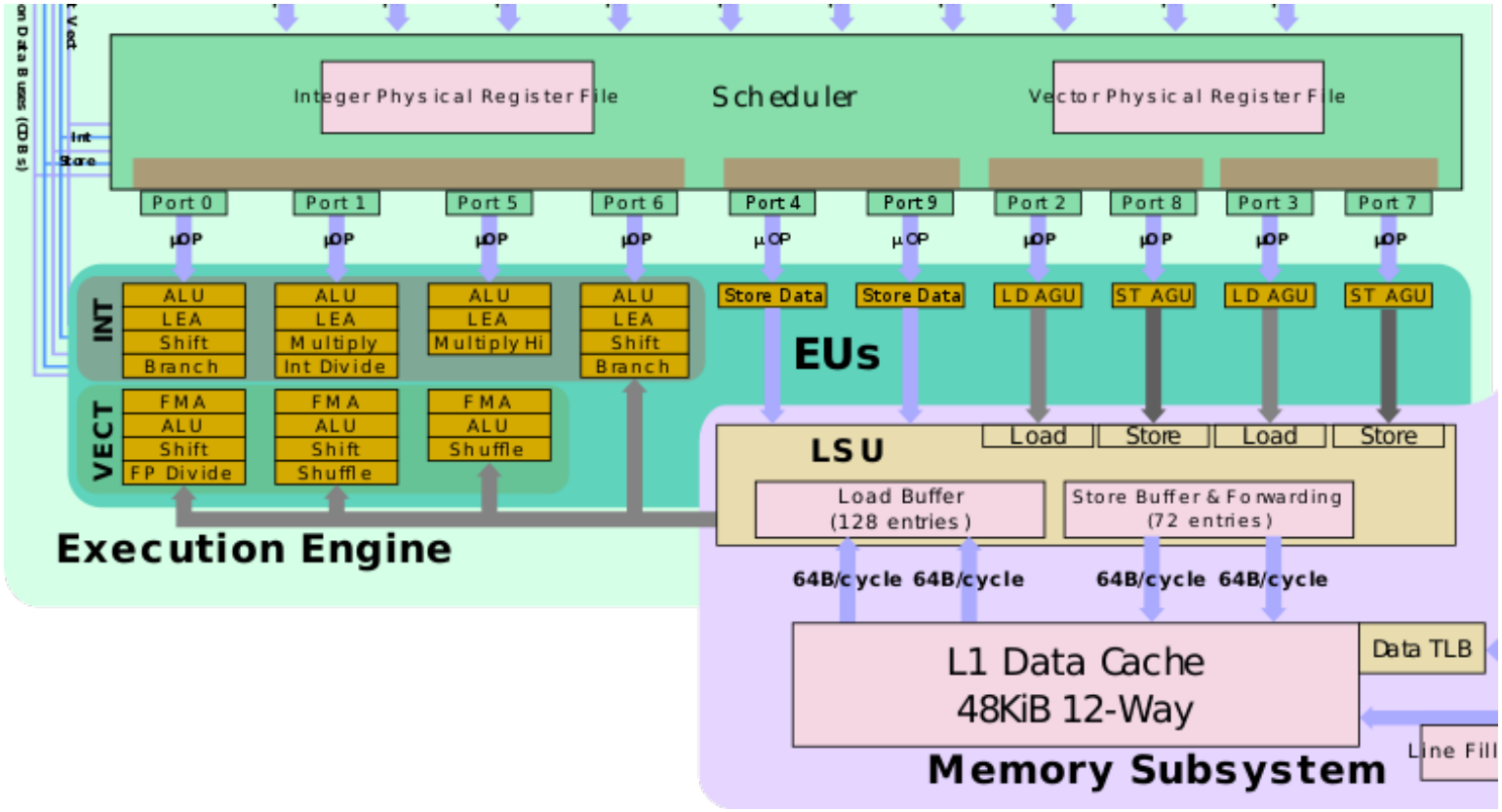
Picture from David Schor, wikichip.org

Instruction Scheduling



Picture from David Schor, wikichip.org

ALUs



Picture from David Schor, wikichip.org

Parallelism

CPU is ~8-way superscalar, ~14 pipeline stages (P4 had 20!)

- Superscalar picks from 352-Instruction window.
- Register renaming to 180 registers.

Each chip has 2-4 cores

Symmetric Multithreading (2-way per core)

On-chip GPU

Cache Organization

All 64 byte blocks, write-back.

(Per core) Split L1 Caches

- 32KB, 8-way Set Associative Instruction Cache
  - Can fetch 16B/cycle.
- 48KB, 12-way Set Associative Data Cache
  - 5-cycle latency. 128B/cycle loads & stores.

(Per core) L2 Unified Cache

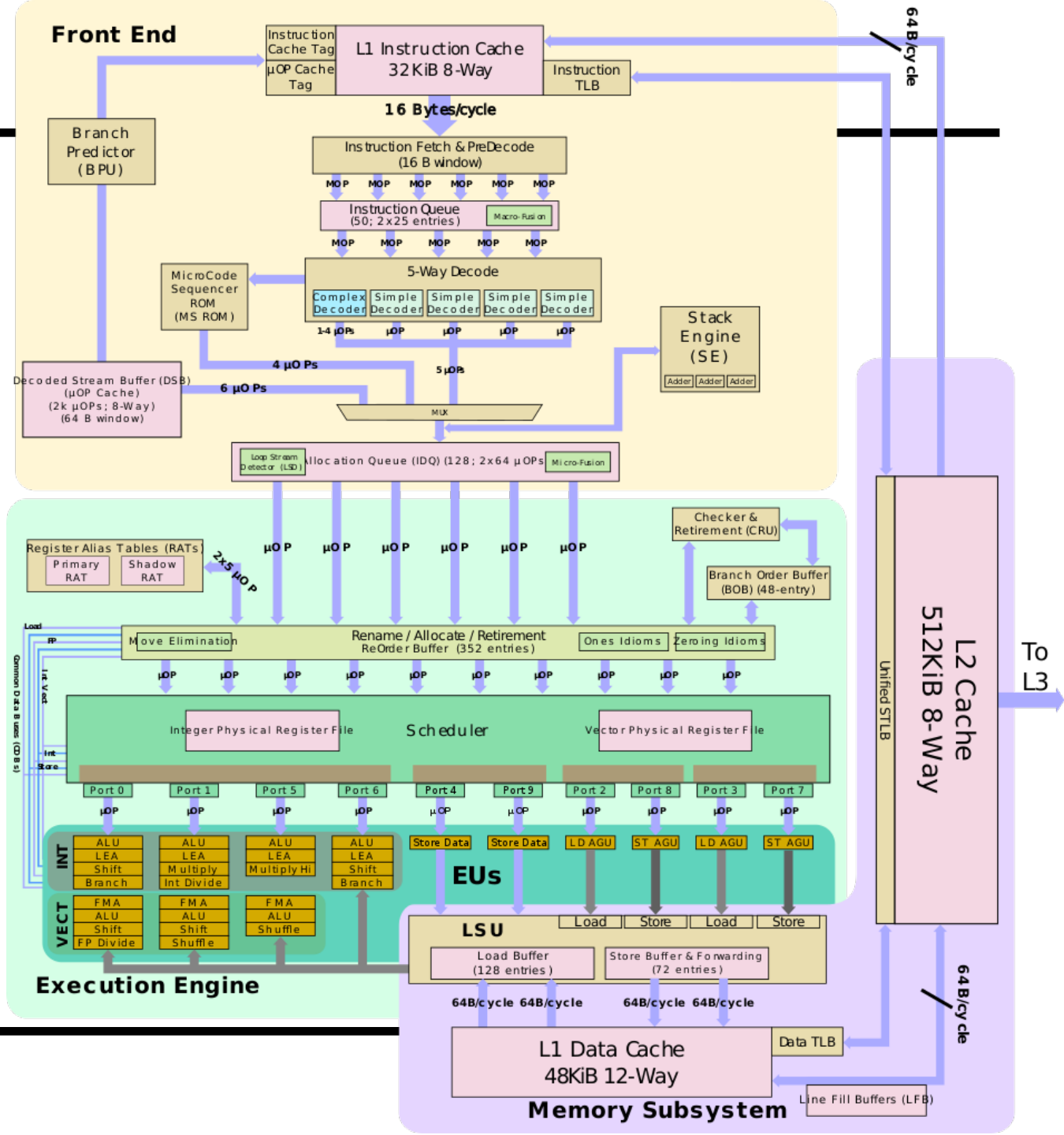
- 512KB, 8-way Set Associative
- non-inclusive
- 13-cycle latency

(Per-chip) L3 Unified Cache

- up to 8MB, 16-way Set Associative
- inclusive
- ~36-cycle latency

Pipeline

14-19 stages



Layout

Core: CPU

RING: interconnect

GT: GPU

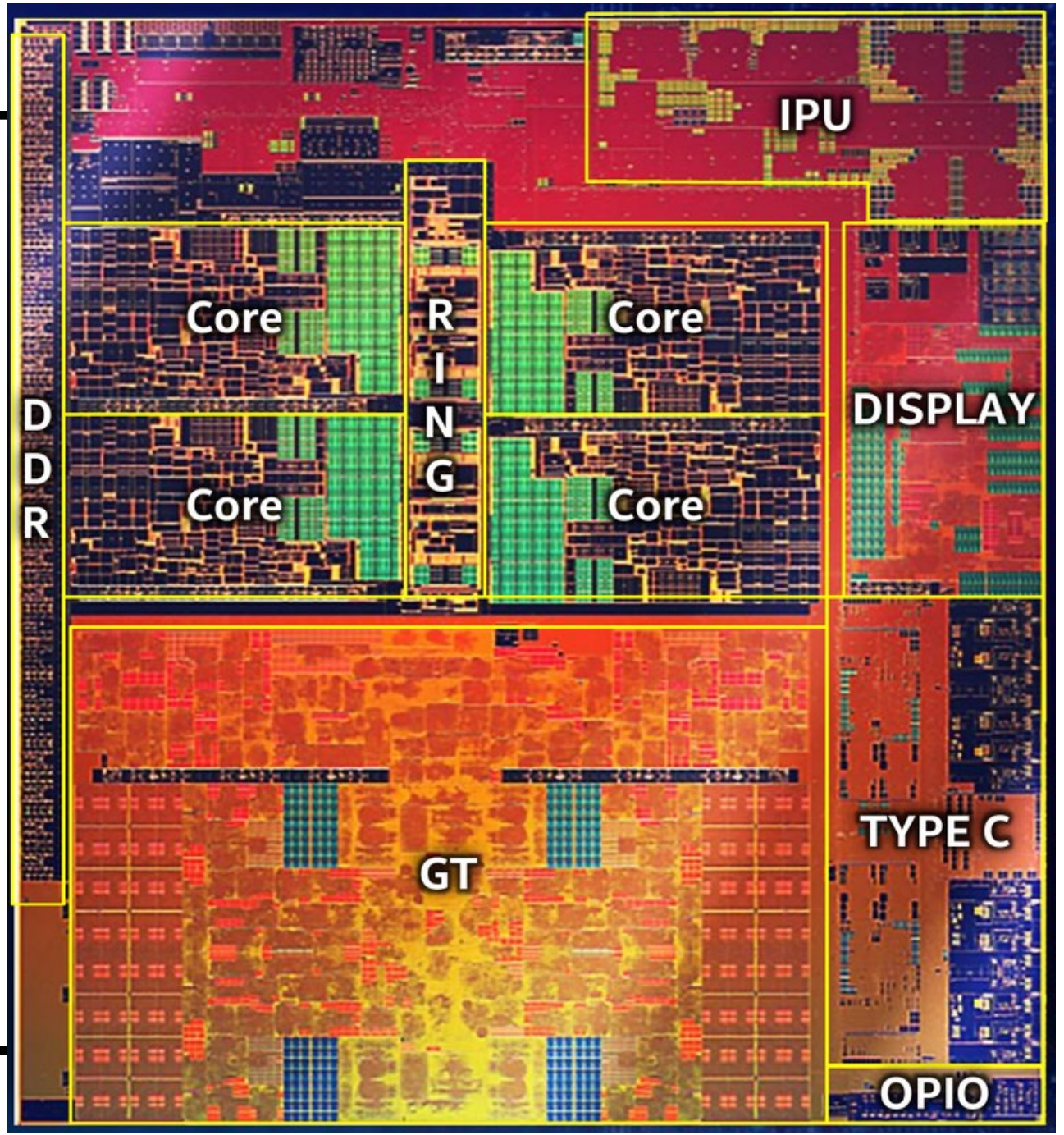
Type C: I/O port

Display: Display driver

IPU: Image Processing Unit

OPIO: on-package I/O

DDR: Memory interface



Intel Ice Lake Die