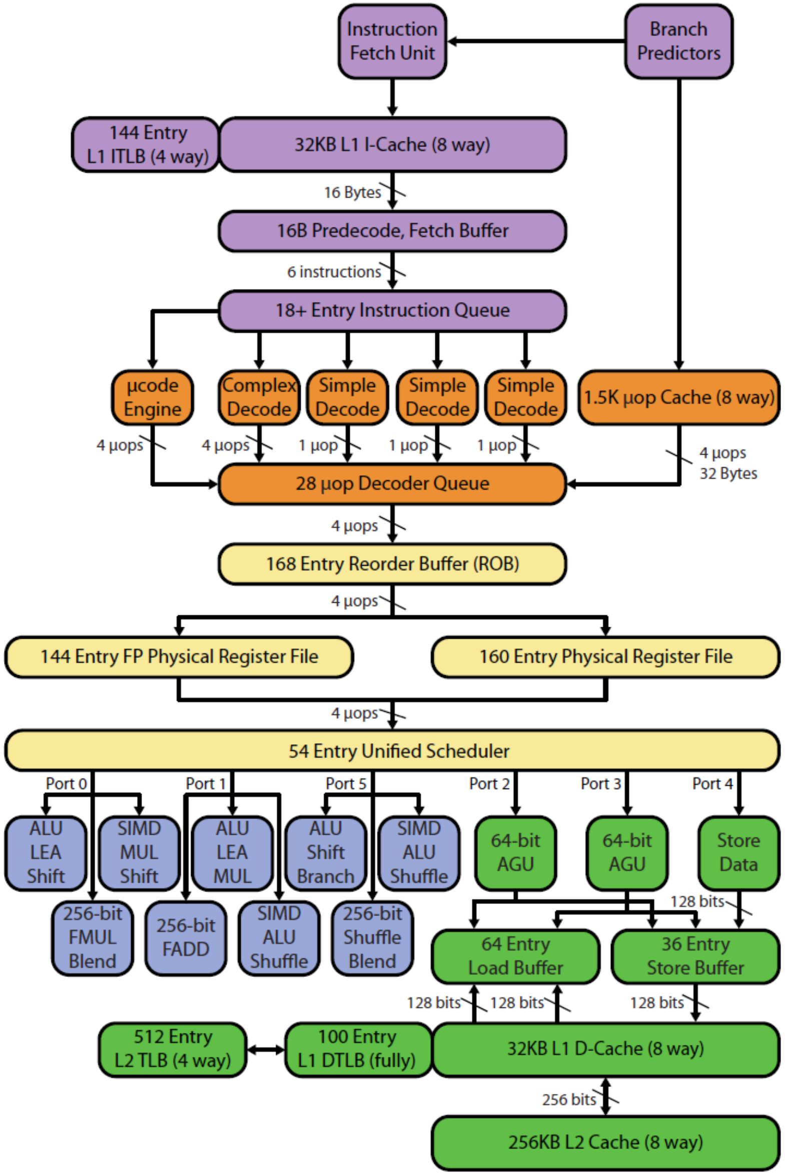
CPU microarchitectures (2000-2018)

CPU Microarchitecture

- Branch prediction
  - conditions, indirect branches, call/return pairs
  - speculative execution
- Instruction decoder
  - loop cache (simple loops up to ~20 instructions)
  - conversion to micro-ops (1:1, 1:N, 2:1)
  - stack-pointer simulator
- Renamer
  - 16 architectural integer registers mapped to ~160 physical
    - similarly for FP/vector registers
- Out-of-order execution
  - ~40 micro-ops in simultaneous execution (RS) from a window of ~300 (ROB)
  - retirement: memory/register stores in-order in background
  - store forwarding: loads retrieve values from waiting stores
  - speculative loads: no waiting for waiting stores (to unknown addresses)

Intel vs. AMD architectures (realworldtech.com)

Sandy Bridge



Instruction Fetch Unit

Branch Predictors

144 Entry L1 ITLB (4 way)

32KB L1 I-Cache (8 way)

16 Bytes

16B Predecode, Fetch Buffer

6 instructions

18+ Entry Instruction Queue

µcode Engine

Complex Decode

Simple Decode

Simple Decode

Simple Decode

1.5K µop Cache (8 way)

4 µops

4 µops

1 µop

1 µop

1 µop

4 µops

32 Bytes

28 µop Decoder Queue

4 µops

168 Entry Reorder Buffer (ROB)

4 µops

144 Entry FP Physical Register File

160 Entry Physical Register File

4 µops

54 Entry Unified Scheduler

Port 0

Port 1

Port 5

Port 2

Port 3

Port 4

ALU LEA Shift

SIMD MUL Shift

ALU LEA MUL

ALU Shift Branch

SIMD ALU Shuffle

64-bit AGU

64-bit AGU

Store Data

256-bit FMUL Blend

256-bit FADD

SIMD ALU Shuffle

256-bit Shuffle Blend

128 bits

64 Entry Load Buffer

36 Entry Store Buffer

128 bits

128 bits

128 bits

512 Entry L2 TLB (4 way)

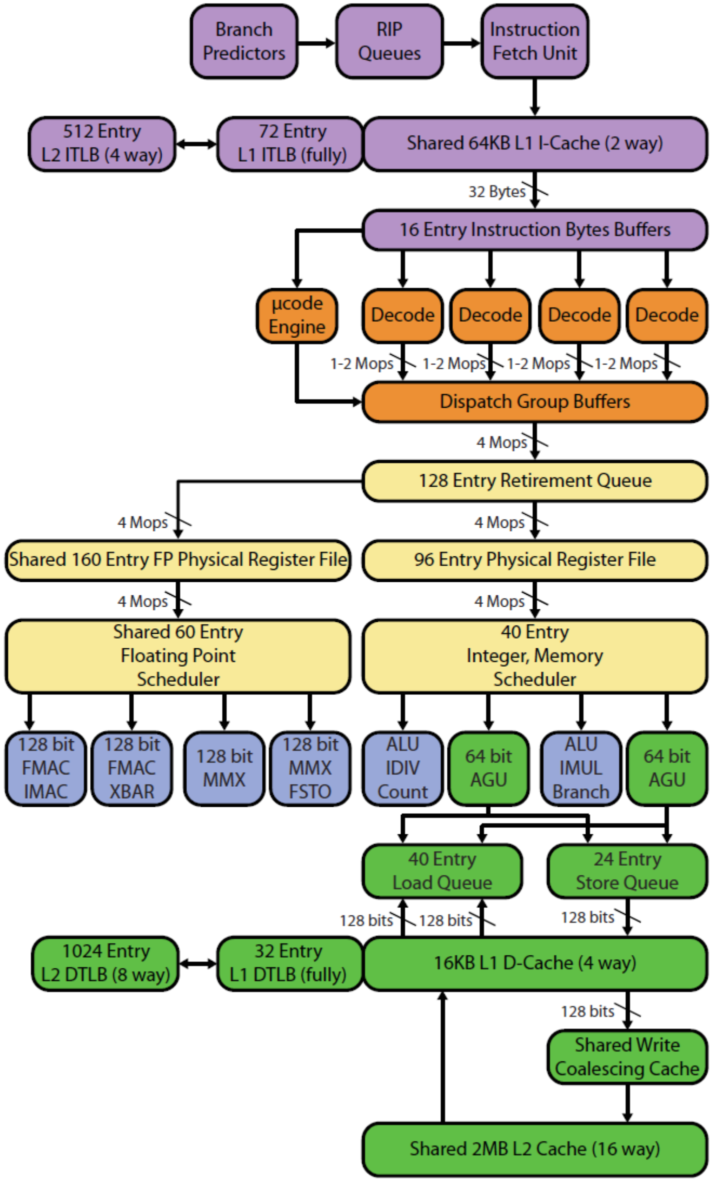
100 Entry L1 DTLB (fully)

32KB L1 D-Cache (8 way)

256 bits

256KB L2 Cache (8 way)

Bulldozer



Branch Predictors

RIP Queues

Instruction Fetch Unit

512 Entry L2 ITLB (4 way)

72 Entry L1 ITLB (fully)

Shared 64KB L1 I-Cache (2 way)

32 Bytes

16 Entry Instruction Bytes Buffers

µcode Engine

Decode

Decode

Decode

Decode

1-2 Mops

1-2 Mops

1-2 Mops

1-2 Mops

Dispatch Group Buffers

4 Mops

128 Entry Retirement Queue

4 Mops

4 Mops

Shared 160 Entry FP Physical Register File

96 Entry Physical Register File

4 Mops

4 Mops

Shared 60 Entry Floating Point Scheduler

40 Entry Integer, Memory Scheduler

128 bit FMAC IMAC

128 bit FMAC XBAR

128 bit MMX

128 bit MMX FSTO

ALU IDIV Count

64 bit AGU

ALU IMUL Branch

64 bit AGU

40 Entry Load Queue

24 Entry Store Queue

128 bits

128 bits

128 bits

1024 Entry L2 DTLB (8 way)

32 Entry L1 DTLB (fully)

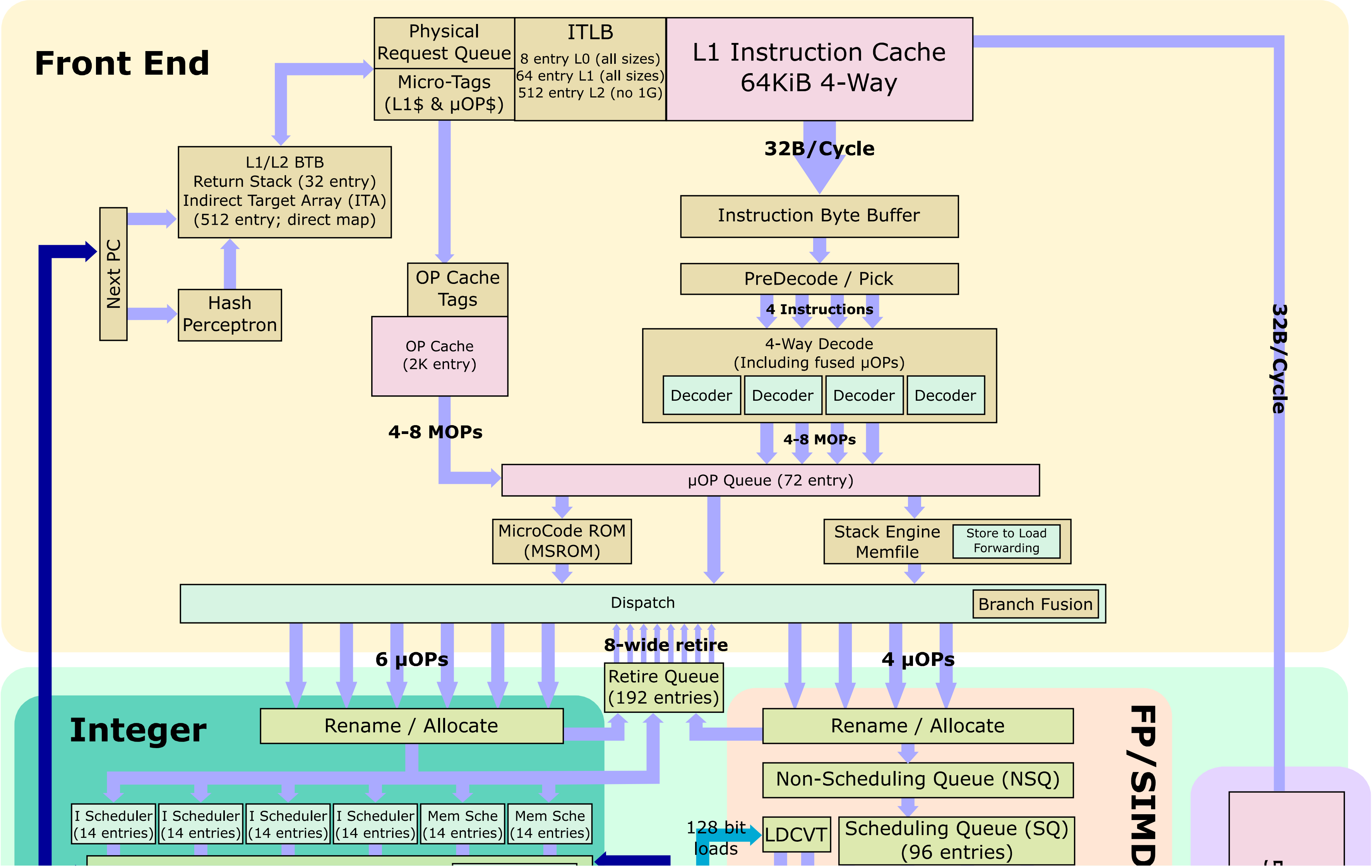
16KB L1 D-Cache (4 way)

128 bits

Shared Write Coalescing Cache

Shared 2MB L2 Cache (16 way)

AMD Zen+ (2018) [wikichip.org]



Front End

Physical Request Queue

Micro-Tags (L1$ & μOP$)

ITLB
8 entry L0 (all sizes)
64 entry L1 (all sizes)
512 entry L2 (no 1G)

L1 Instruction Cache
64KiB 4-Way

32B/Cycle

L1/L2 BTB
Return Stack (32 entry)
Indirect Target Array (ITA)
(512 entry; direct map)

Next PC

Hash Perceptron

OP Cache Tags

OP Cache (2K entry)

Instruction Byte Buffer

PreDecode / Pick

4 Instructions

4-Way Decode
(Including fused μOPs)

Decoder | Decoder | Decoder | Decoder

4-8 MOPs

4-8 MOPs

μOP Queue (72 entry)

MicroCode ROM (MSROM)

Stack Engine Memfile

Store to Load Forwarding

32B/Cycle

Dispatch

Branch Fusion

6 μOPs

8-wide retire

Retire Queue (192 entries)

4 μOPs

Integer

Rename / Allocate

Rename / Allocate

Non-Scheduling Queue (NSQ)

FP/SIMD

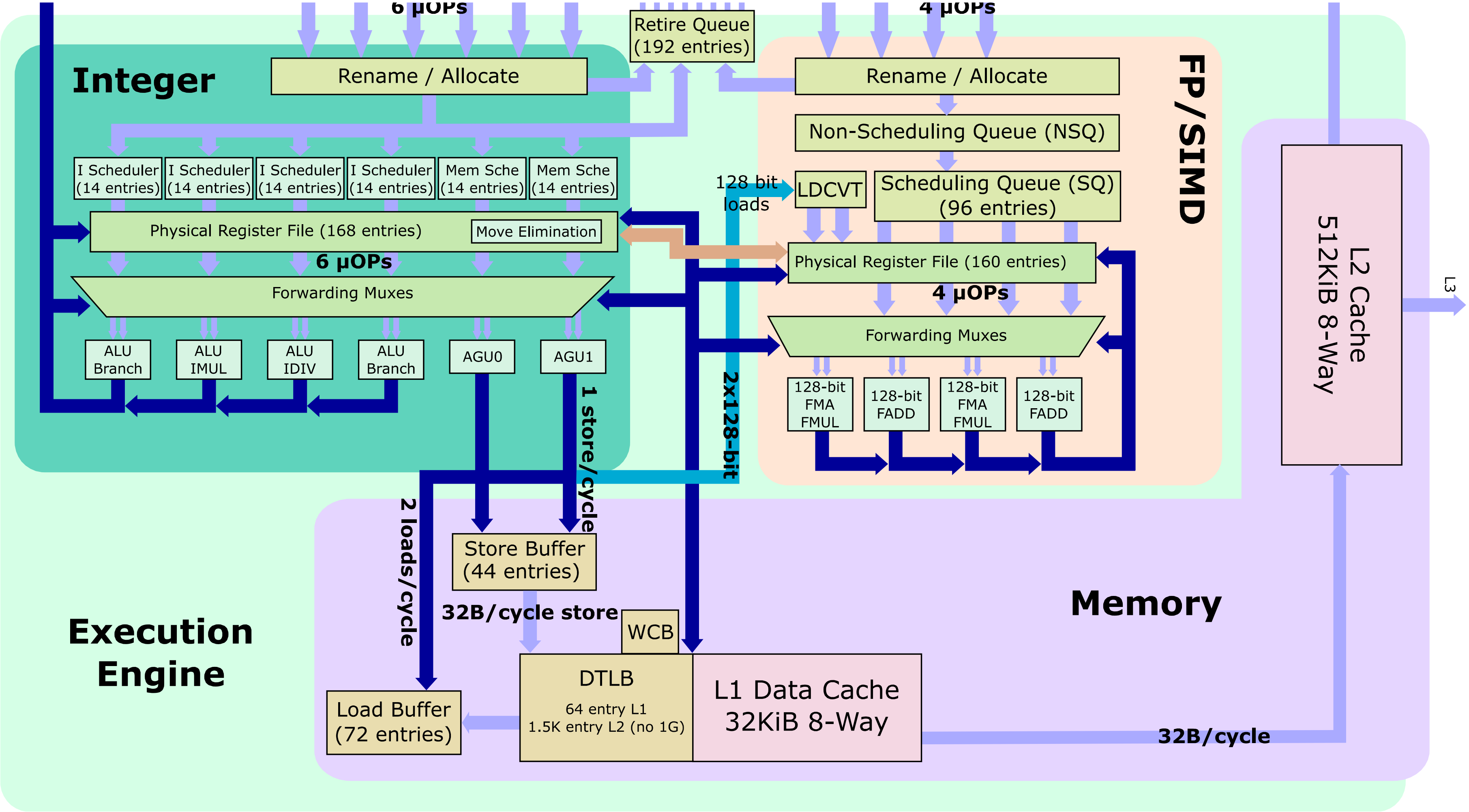
I Scheduler (14 entries) | I Scheduler (14 entries) | I Scheduler (14 entries) | I Scheduler (14 entries) | Mem Sche (14 entries) | Mem Sche (14 entries)

128 bit loads

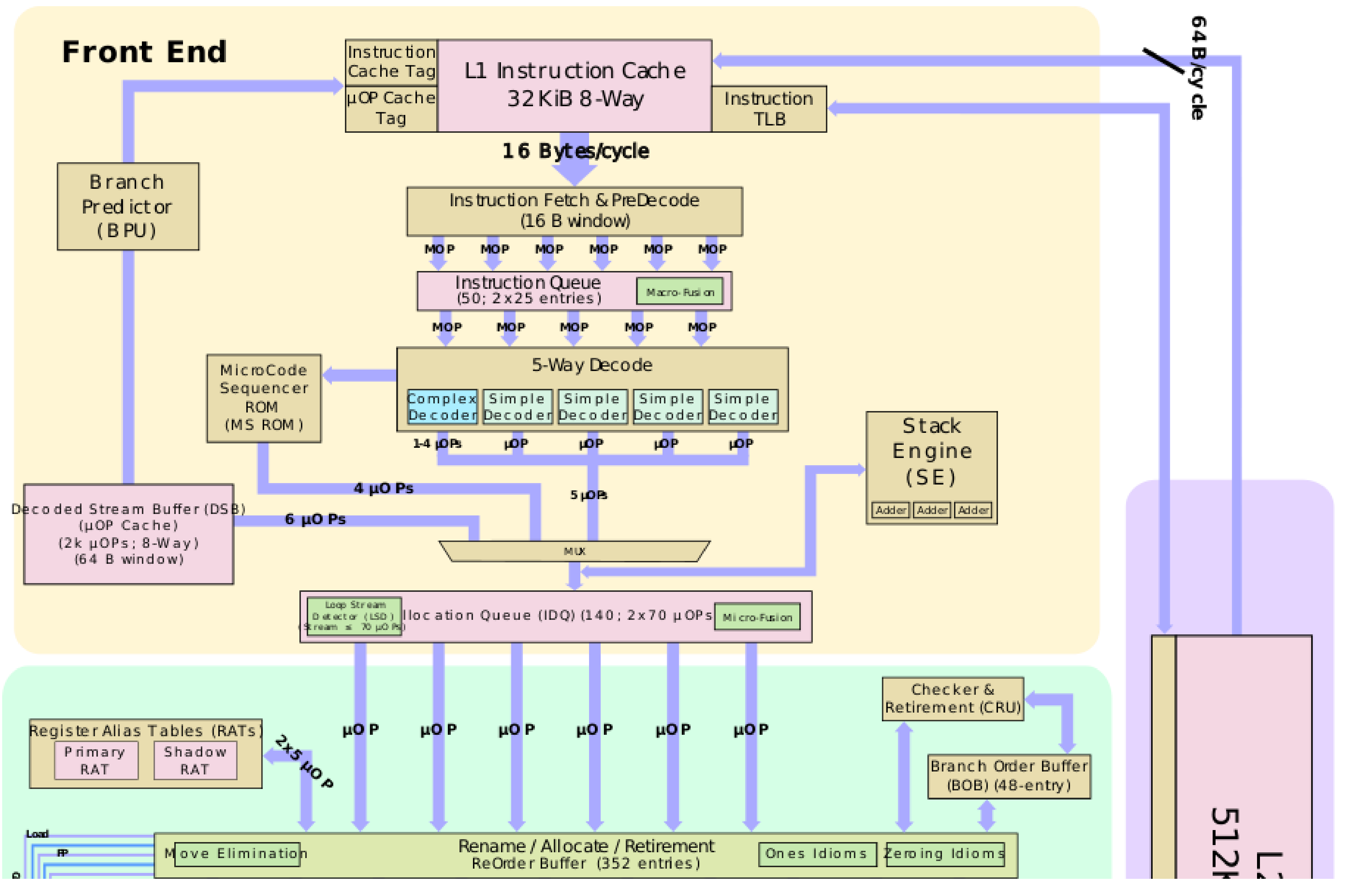
LDCVT

Scheduling Queue (SQ) (96 entries)

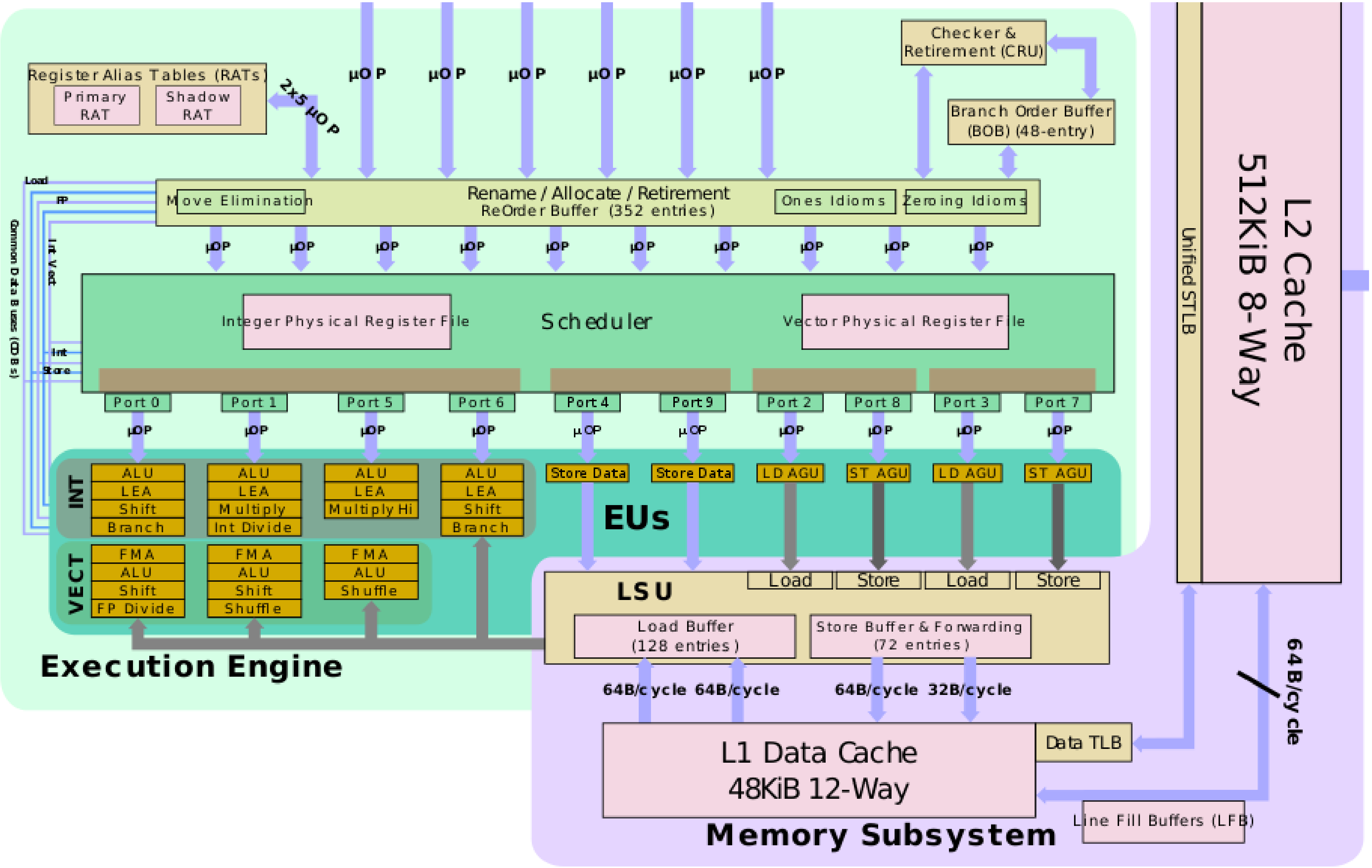
AMD Zen+ (2018) [wikichip.org]



Intel Ice Lake (2019) [wikichip.org]



Intel Ice Lake (2019) [wikichip.org]



Register Alias Tables (RATs)

Primary RAT

Shadow RAT

2x5 µOP

µOP µOP µOP µOP µOP µOP

Checker & Retirement (CRU)

Branch Order Buffer (BOB) (48-entry)

Load

FP

Int Vect

Common Data Buses (CDBs)

Move Elimination

Rename / Allocate / Retirement

ReOrder Buffer (352 entries)

Ones Idioms

Zeroing Idioms

µOP µOP µOP µOP µOP µOP µOP µOP µOP µOP

Integer Physical Register File

Scheduler

Vector Physical Register File

Int

Store

Port 0, Port 1, Port 5, Port 6, Port 4, Port 9, Port 2, Port 8, Port 3, Port 7

µOP

INT

ALU, LEA, Shift, Branch

ALU, LEA, Multiply, Int Divide

ALU, LEA, Multiply Hi

ALU, LEA, Shift, Branch

VECT

FMA, ALU, Shift, FP Divide

FMA, ALU, Shift, Shuffle

FMA, ALU, Shuffle

Store Data

Store Data

LD AGU

ST AGU

LD AGU

ST AGU

EUs

Execution Engine

LSU

Load, Store, Load, Store

Load Buffer (128 entries)

Store Buffer & Forwarding (72 entries)

64B/cycle 64B/cycle 64B/cycle 32B/cycle

L1 Data Cache 48KiB 12-Way

Data TLB

Line Fill Buffers (LFB)

Memory Subsystem

Unified STLB

L2 Cache 512KiB 8-Way

64B/cycle

CPU Microarchitecture (Intel x86/64, 2015..2023)

- In a cycle, CPU can (in theory) simultaneously perform:
  - Fetch: 16 B (avg. 4 instructions) from L1 instruction cache
  - Decode: 5..6 instructions
  - ALU: 3..5 simple operations (add/mul)
  - Memory load: 2..3 reads (up to 256..512 bits total) from L1 data cache
  - Memory store: 1..2 writes (up to 256..512 bits total) to L1 data cache

- Latency
    - the time between consuming operands and producing results
  - integer add: 1, mul: 4
  - FP add: 2..4, FP mul/FMA: 4
  - FP div: 11/14, integer div: ~18, data dependent
  - integer load: 5, FP load: 6 (L1 cache)
  - store address: 3
  - store data: 1..2 (retirement, in-order)