Master Informatics Eng.

2020/21

A.J.Proença
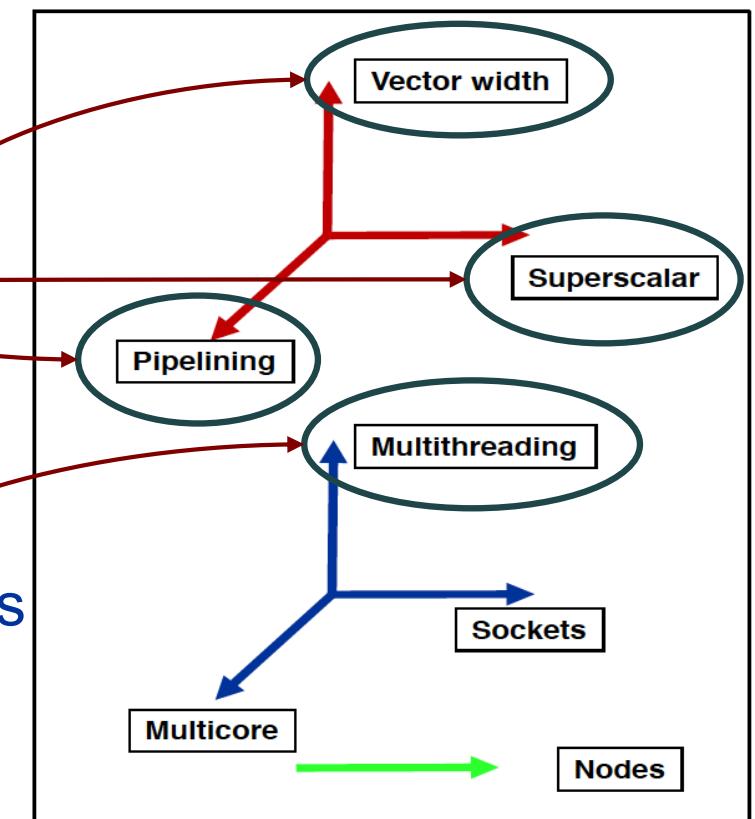
From ILP to Multithreading *(online)*

(most slides are borrowed)

Key issues for parallelism in a single-core
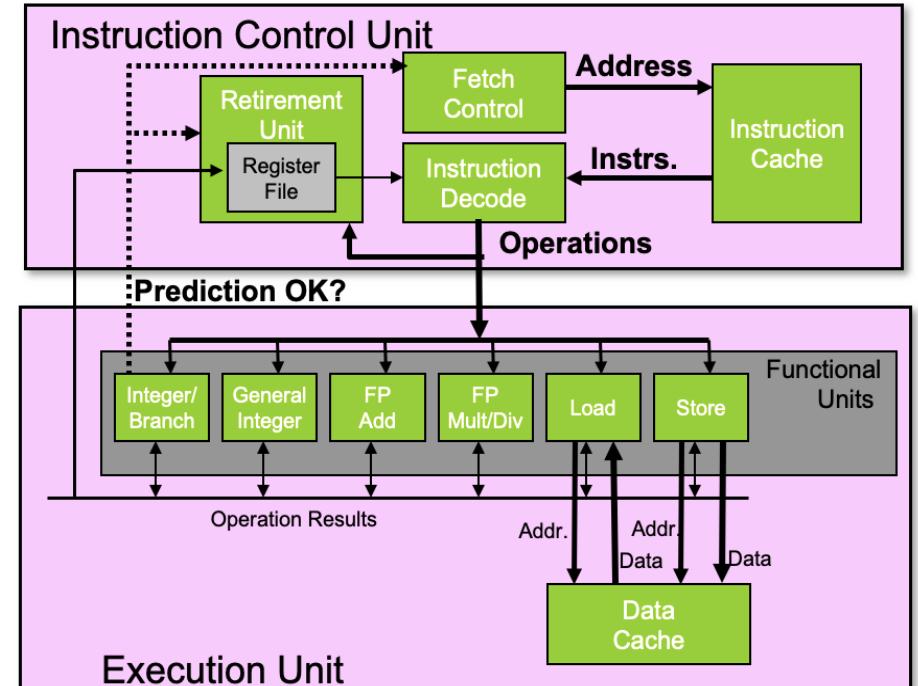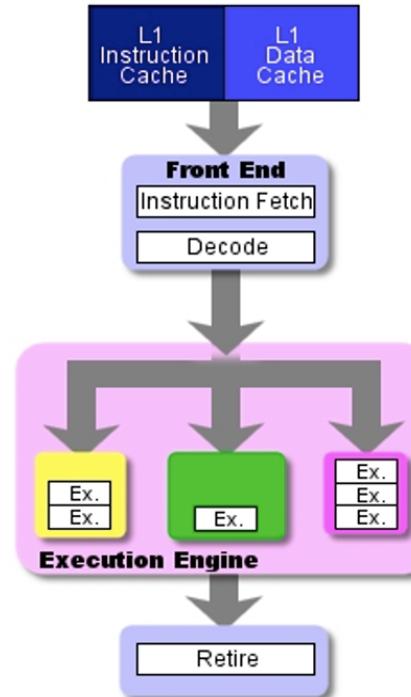
- **Currently under discussion:**

  - pipelining:
    reviewed in the `combine` example

  - superscalar:
    idem, but some more now

  - data parallelism:
    vector computers &
    vector extensions to scalar processors

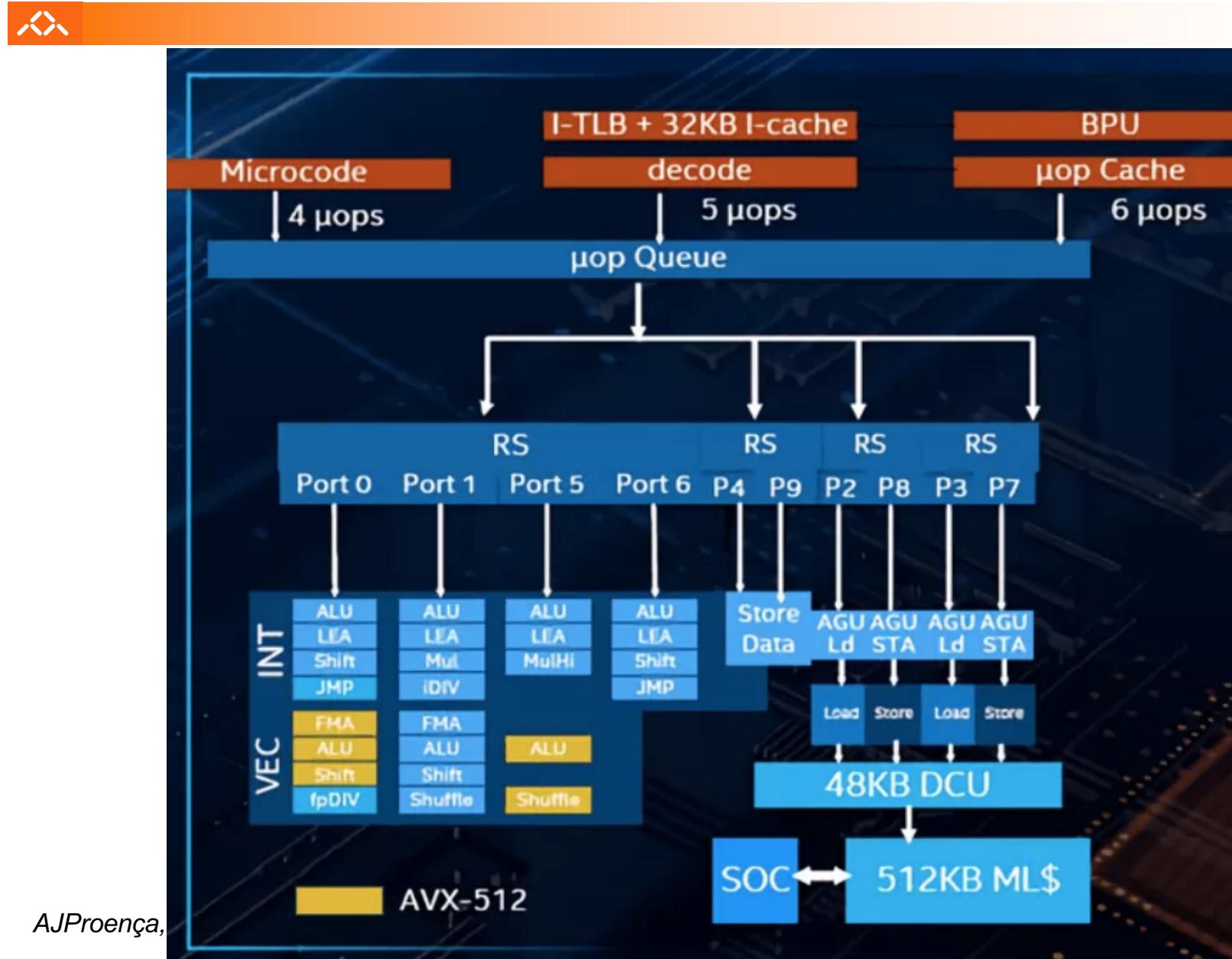 - multithreading:
    alternative approaches

Pipelining & superscalarity: a review

Topic addressed in the undergrad course through the `combine` example

- The analysed pipelines were only in the P6 **Execution Unit**, assuming that the **Instruction Control Unit** issues at each clock cycle all the required instructions for parallel execution

- The image suggests **(i)** a **3-way superscalar** engine and **(ii)** an execution engine with **6 functional units**

Intel Sunny Cove microarchitecture: 30 functional units

Comments to the slides on performance evaluation *(1)*

- **Assembly version for `combine4`**
  - data type: *integer* ; operation: *multiplication*

```
.L24:                           # Loop:
  imull (%eax,%edx,4),%ecx      # t *= data[i]
  incl   %edx                   # i++
  cmpl   %esi,%edx              # i:length
  jl     .L24                   # if < goto Loop
```

- **Translating 1st iteration into RISC-like instructions**

```
load   (%eax,%edx.0,4)  ➔  t.1
imull t.1, %ecx.0        ➔  %ecx.1
incl   %edx.0           ➔  %edx.1
cmpl   %esi, %edx.1     ➔  cc.1
jl     -taken cc.1
```

3+miss penalty?
+4
+1   Expected duration:
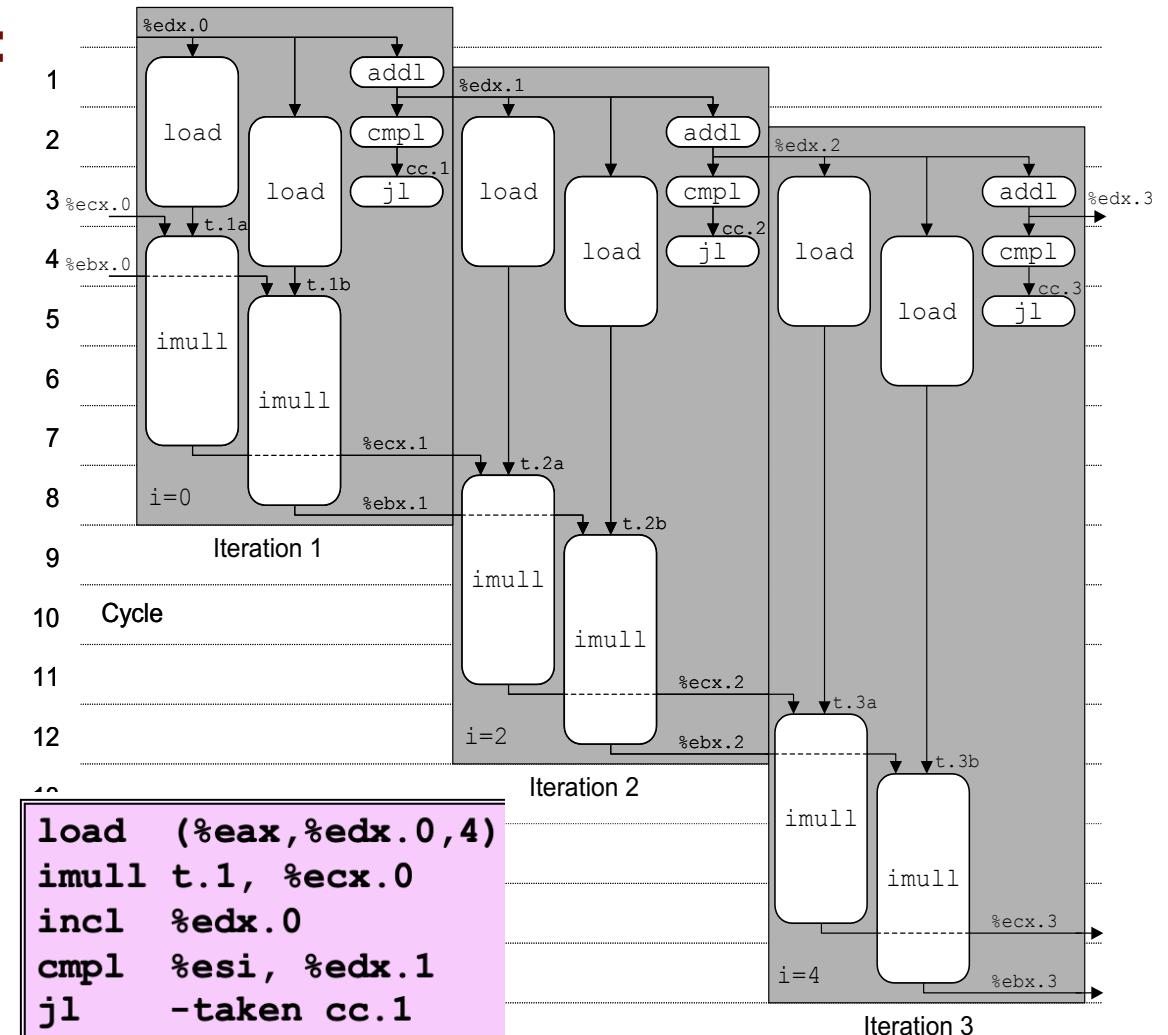+1   10+ clock cycles
+1   per vector element

Timings in clock cycles

Comments to the slides on performance evaluation (2)
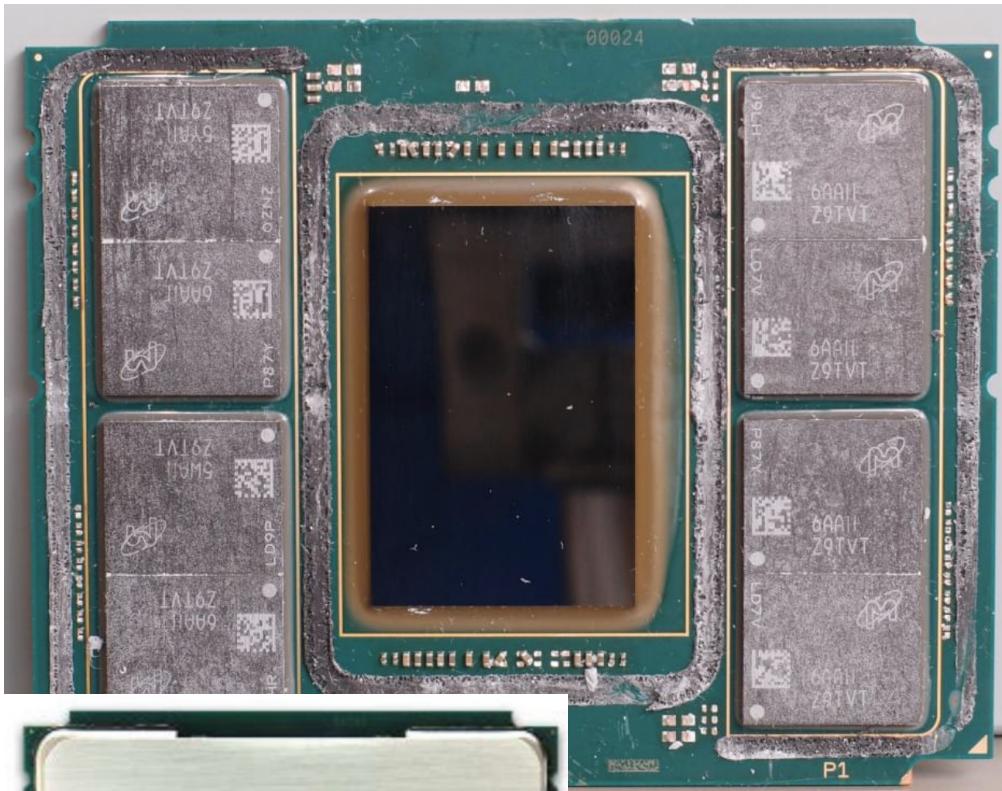
Features that lead to CPE=2:

- **in the hardware**
  - pipelined execution units with 1 clock-cycle/issue
  - mem hierarchy with cache
  - out-of-order execution
  - at least 5-way superscalar
  - more 1 arithm & 1 load units
  - speculative jump

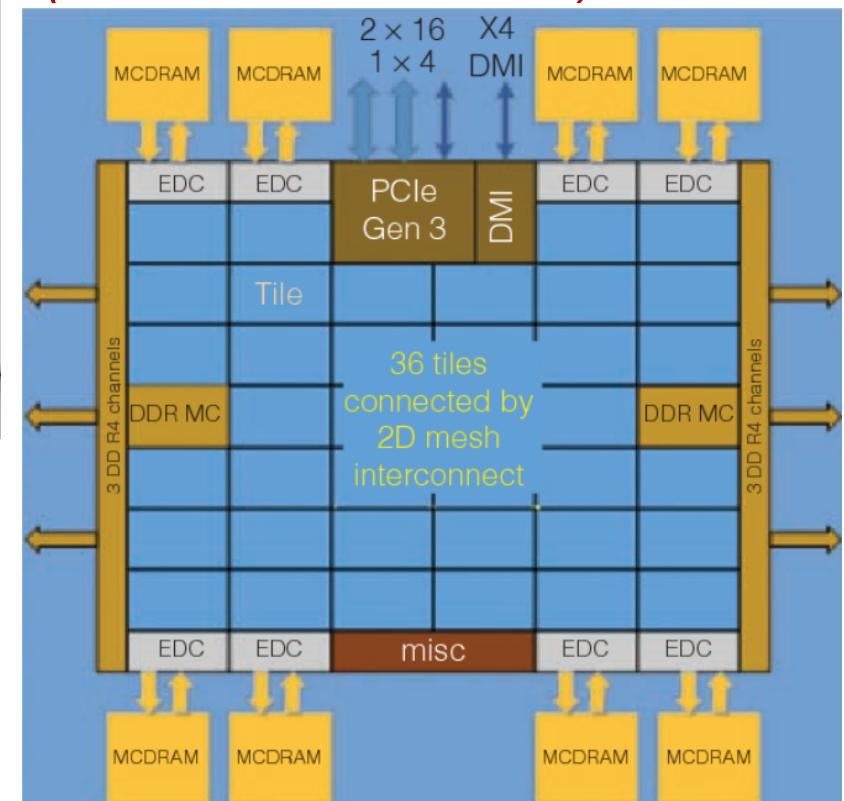- **at the code level**
  - loop unroll 2x
  - 2-way parallelism


```
load   (%eax,%edx.0,4)
imull  t.1, %ecx.0
incl   %edx.0
cmpl   %esi, %edx.1
jl     -taken cc.1
```

Previous questions: max number of physical cores?

Intel

Xeon Phi package:
up to 72 cores
(discontinued in 2018)
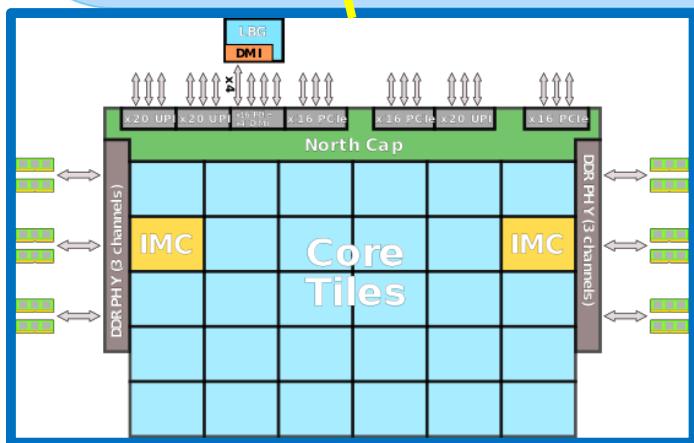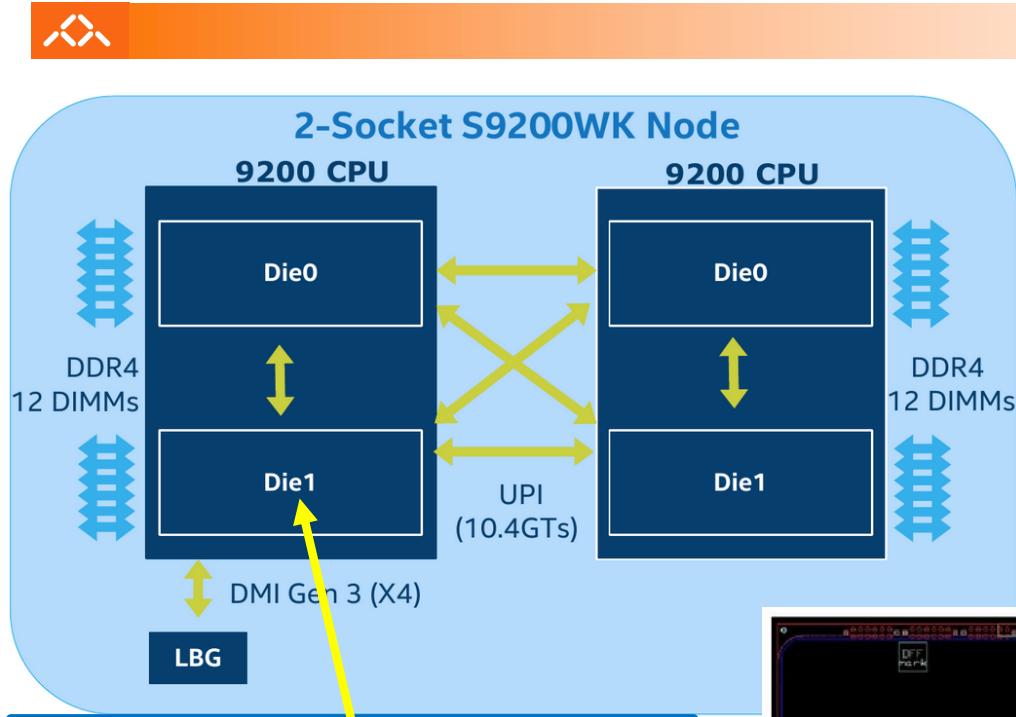
MiEI, UMinho, 2020/21
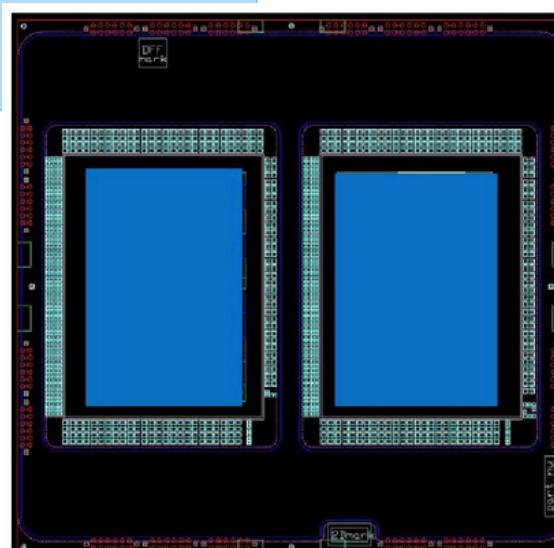
Previous questions: max number of physical cores?
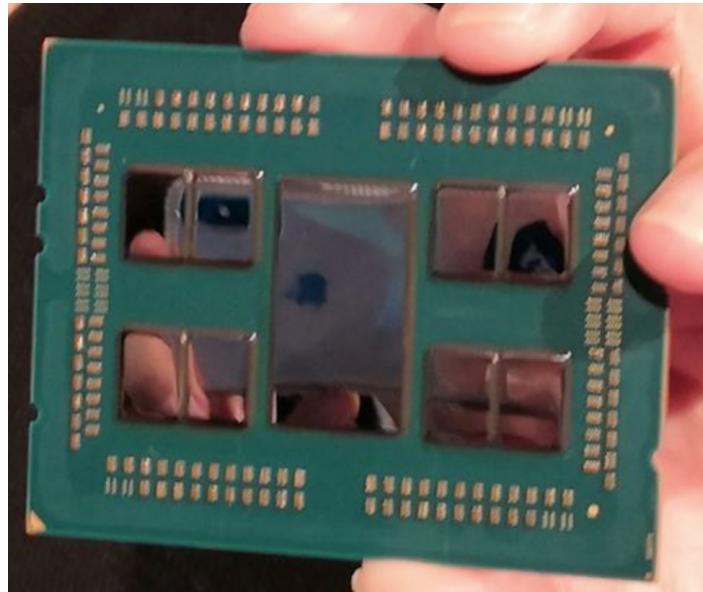
Intel

Xeon Platinum 9282 package:
56 cores
2-socket node: 112 cores

Package Top side view

Package bottom side view
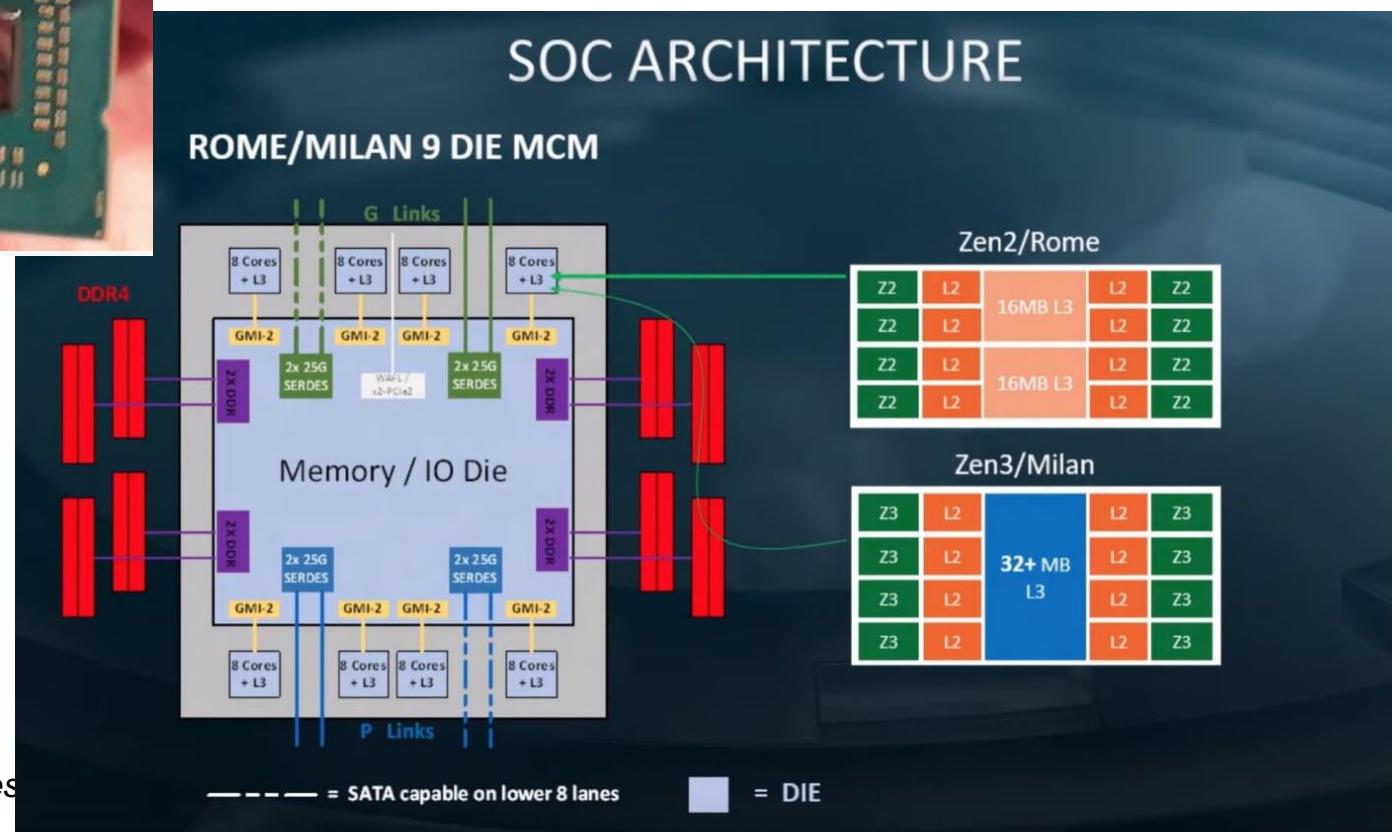
Previous questions: max number of physical cores?

AMD

Epyc Rome & Milan: **64 cores**

AJProença, Advanced Architectures

Previous questions: max number of physical cores?
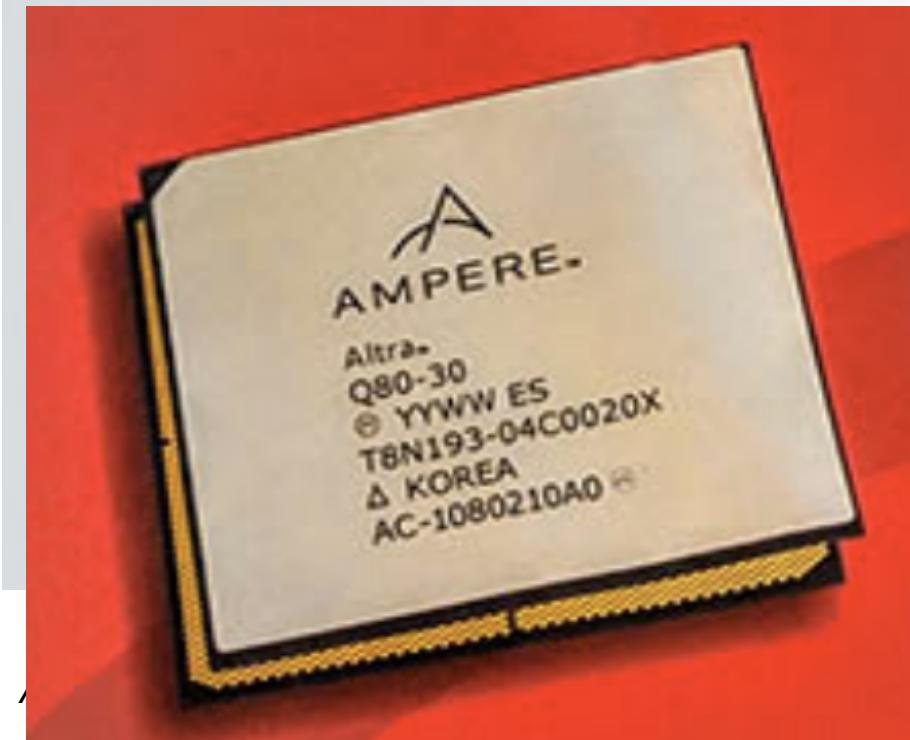
Ampere™ Altra™ processor complex

ARM

Ampere Altra: 80 cores

80 64-bit Arm CPU cores @ 3.0 GHz Turbo
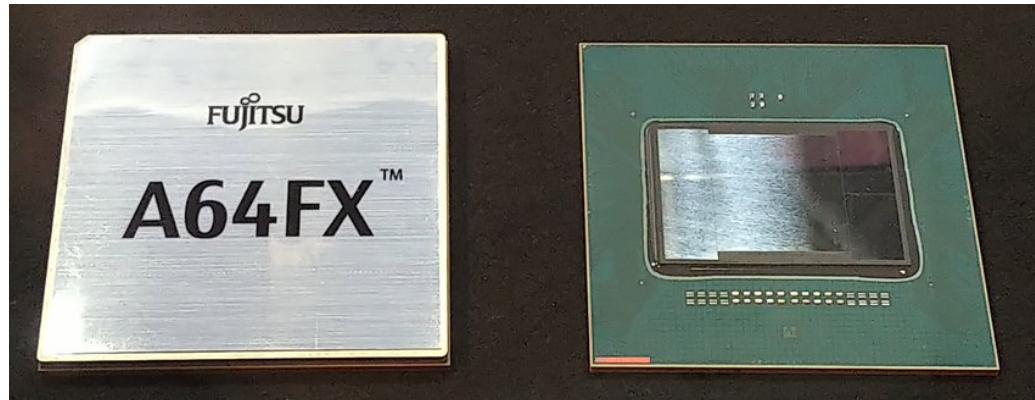- 4-Wide superscalar aggressive out-of-order execution
- Single threaded cores for performance and security isolation

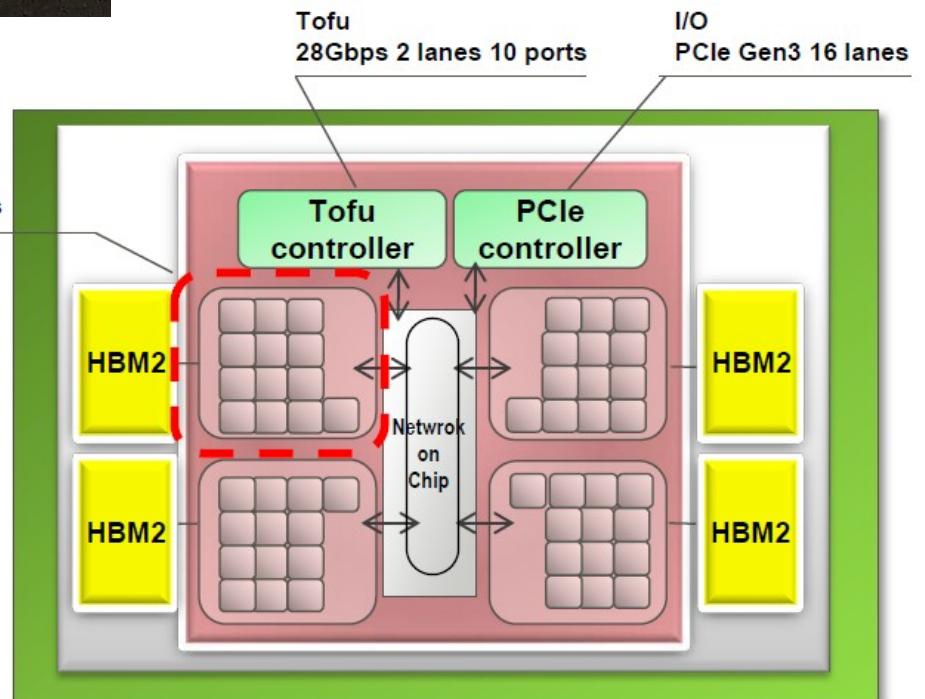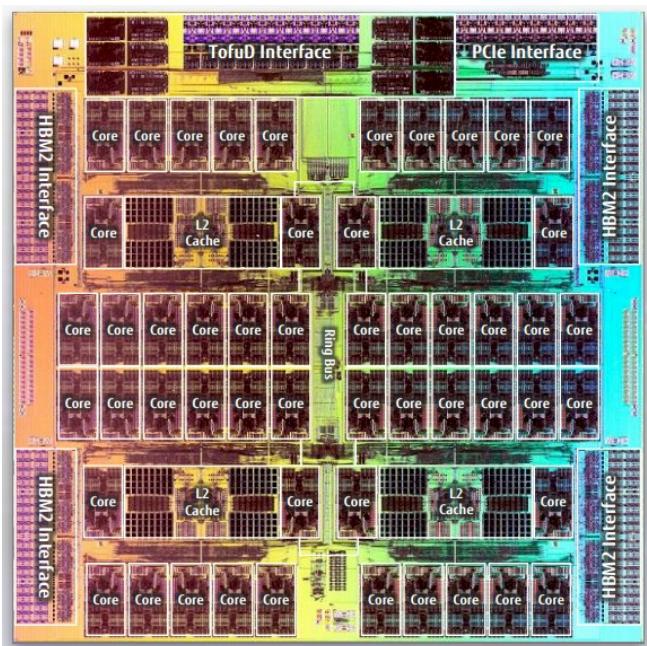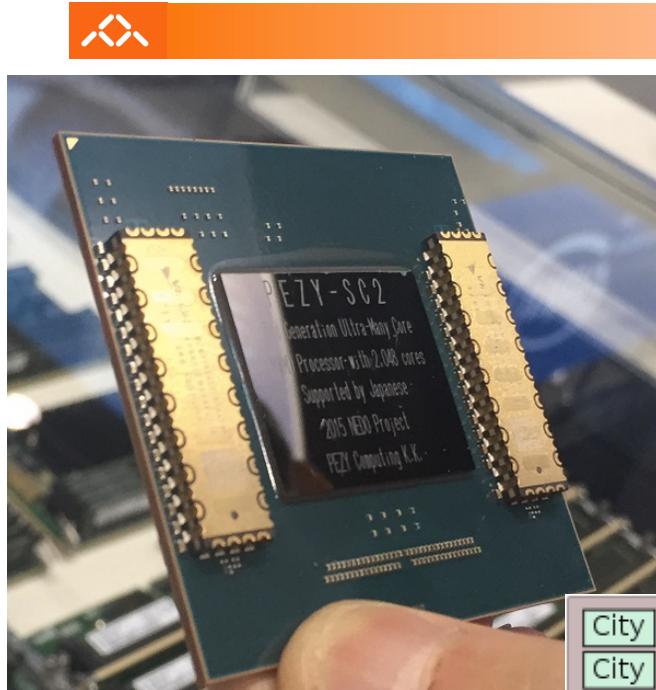# *Previous questions: max number of physical cores?*

ARM

Fujitsu A64FX Arm Chip:

48+4 cores

(in #1 TOP500, June 2020)

CMG specification
13 cores
L2$ 8MiB
Mem 8GiB, 256GB/s

Tofu
28Gbps 2 lanes 10 ports

I/O
PCIe Gen3 16 lanes

Previous questions: max number of physical cores?

Japan

PEZY-SC2: **2048 cores**
   + 8x MIPS cores *(2017)*
PEZY-SC3: 8192 cores
   (due in 2019, but…)
PEZY-SC4: 16384 cores
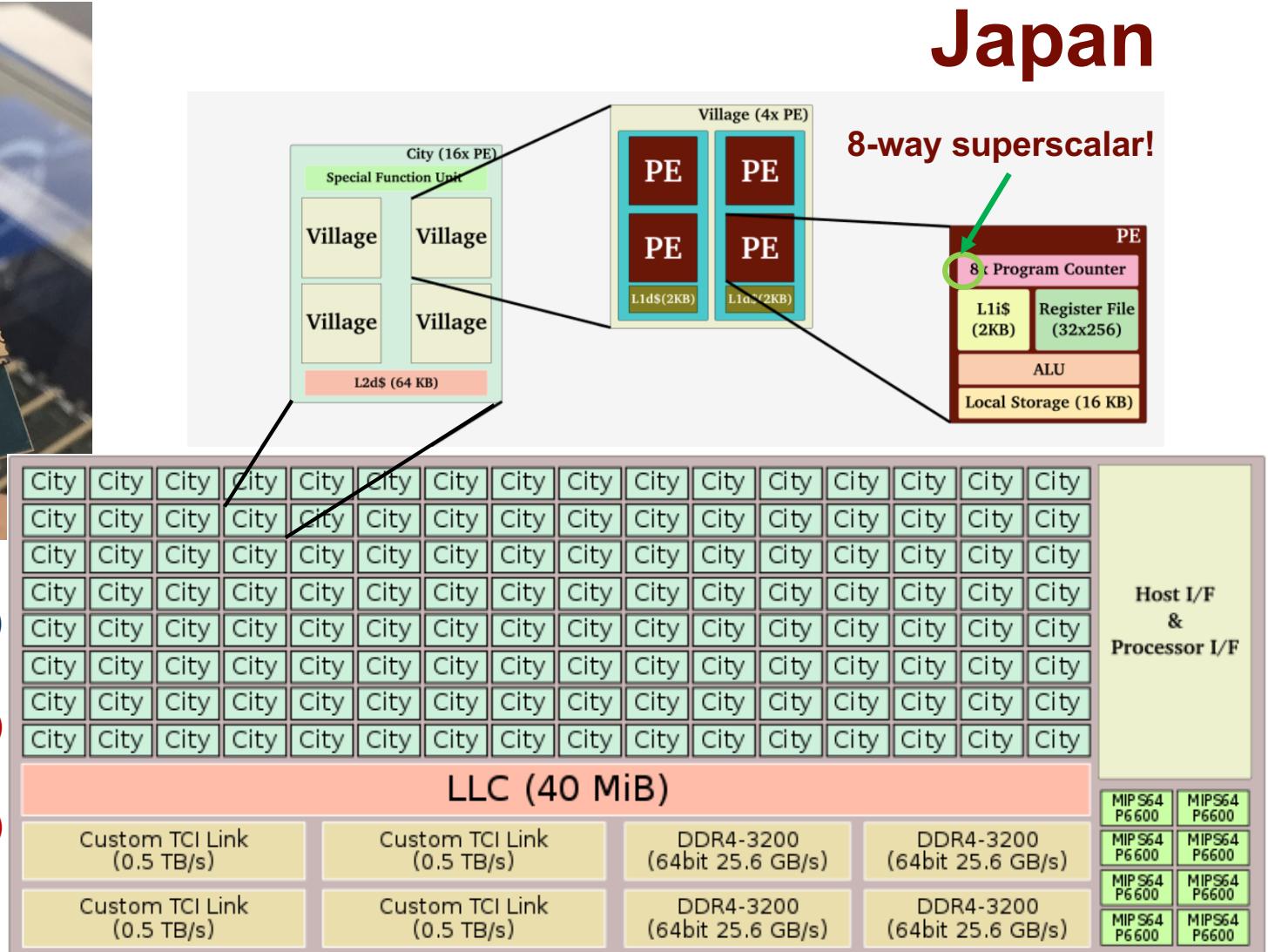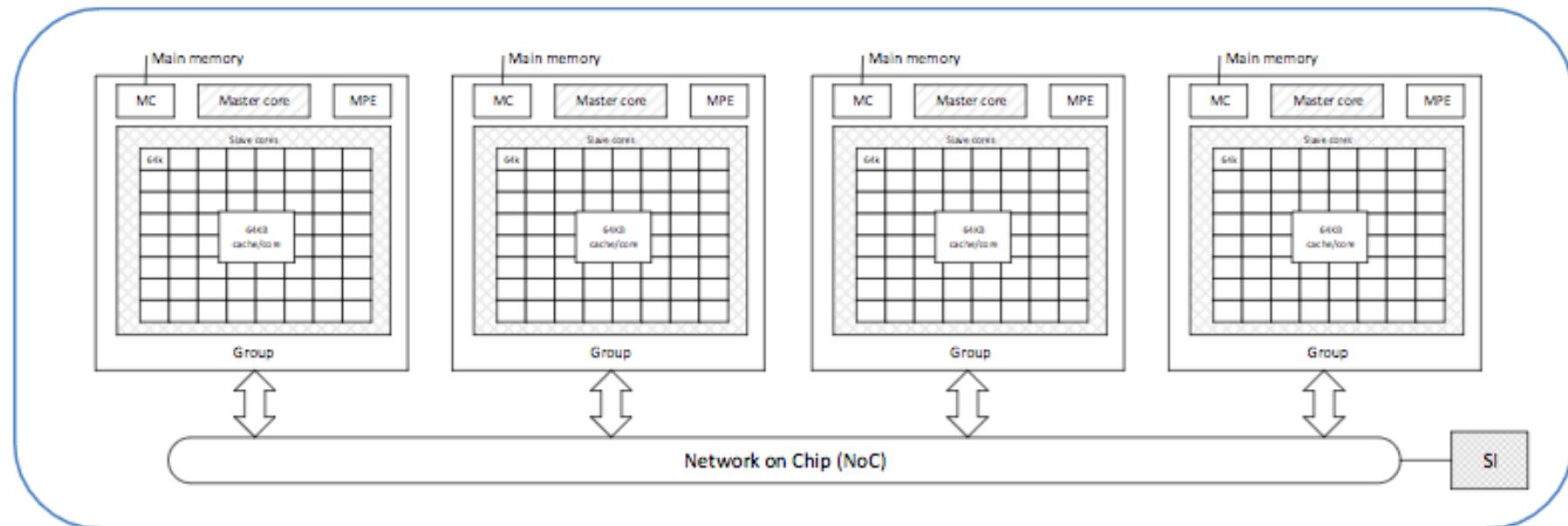   (due in 2020, but…)

Previous questions: max number of physical cores?

China

Sunway SW 26010:
256+4 cores
(in #1 TOP500, June 2016)

Previous questions: max number of physical cores?

Cerebras Wafer Scale Engine (WSE):
the largest chip ever built)

Worldwide

46,225 mm^2 chip
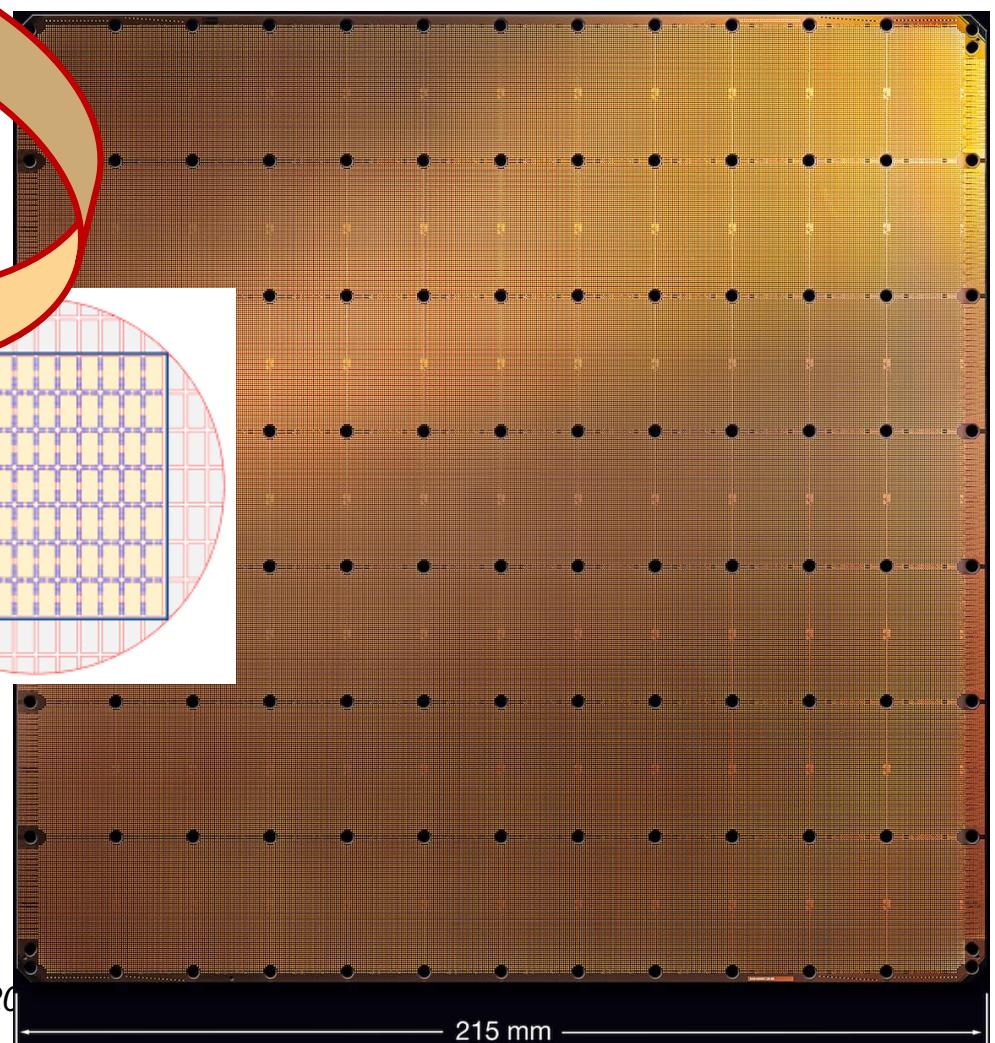56x larger than the biggest GPU ever made
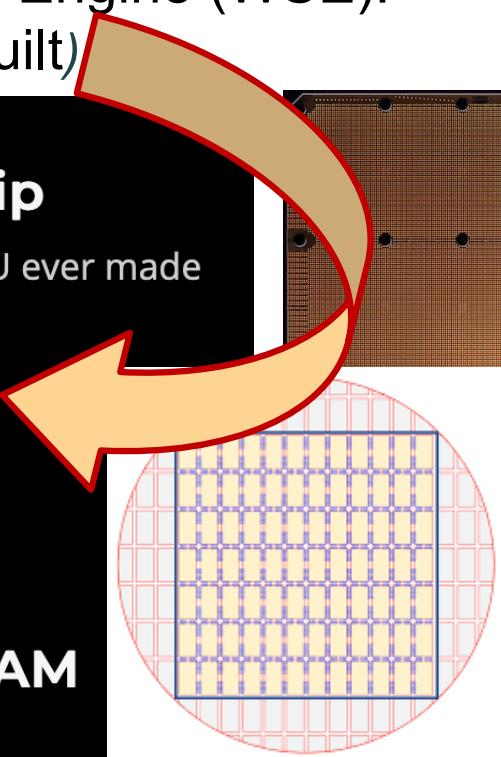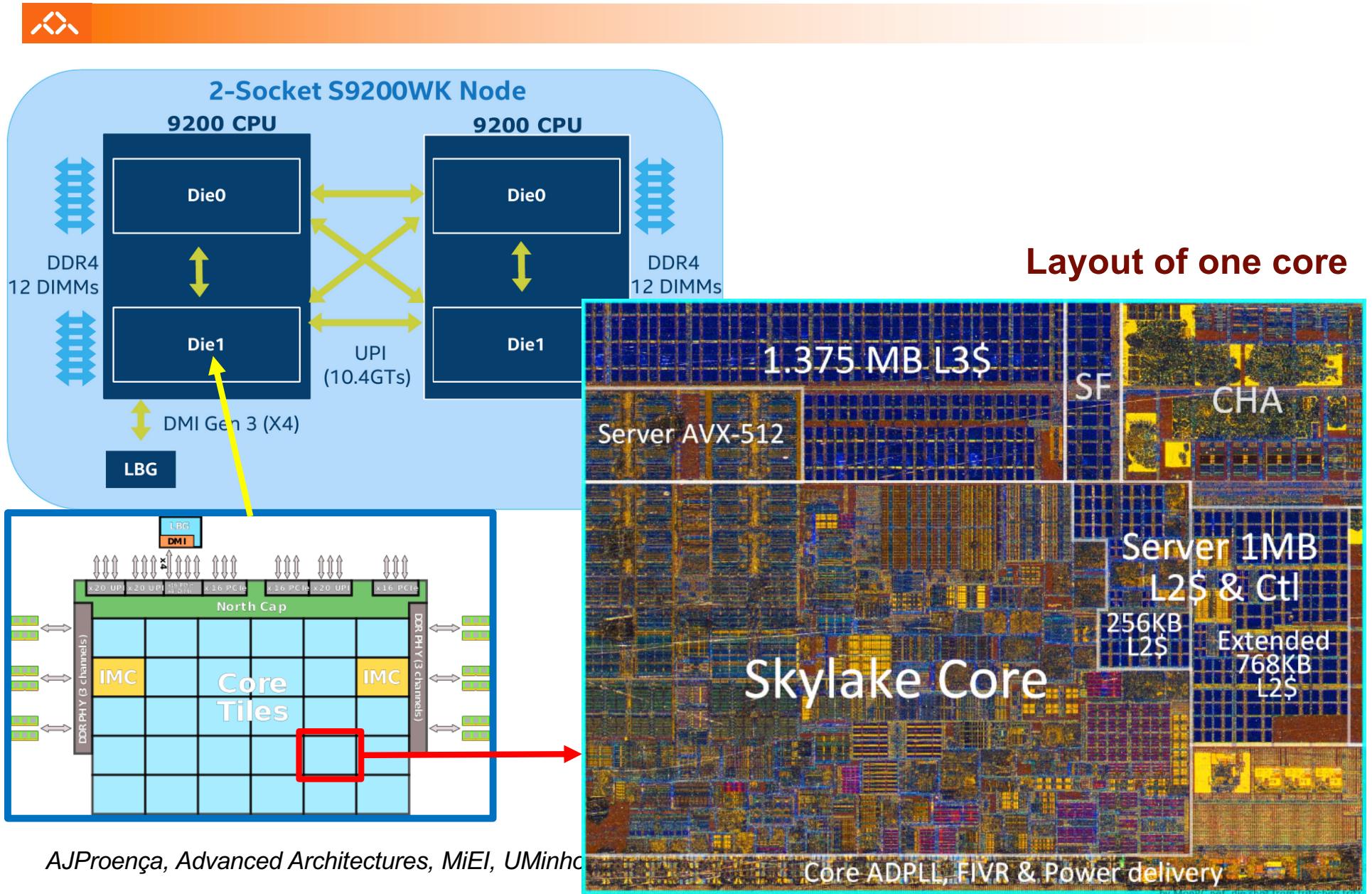
400,000 core
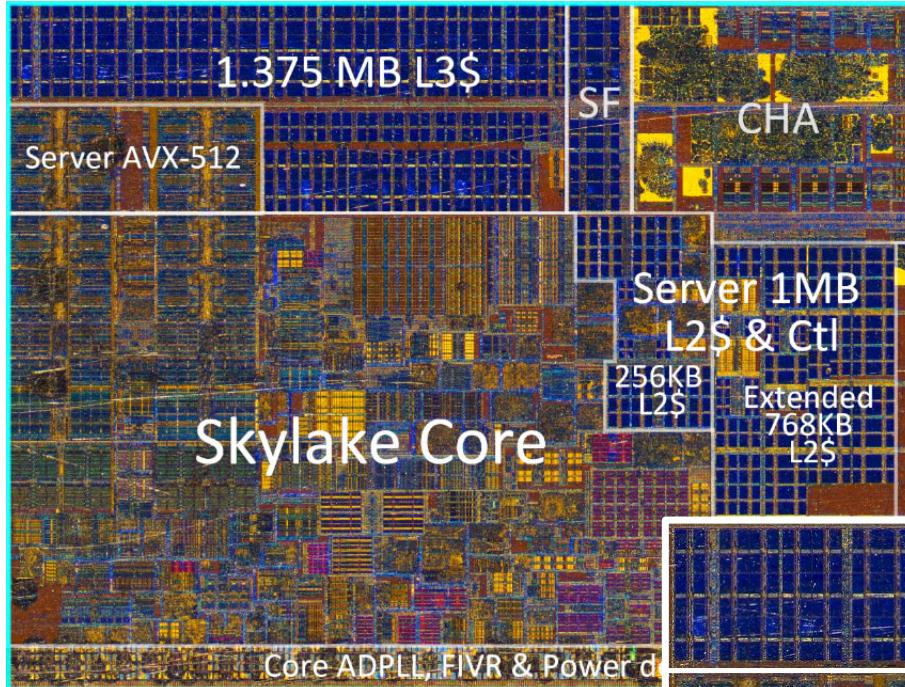78x more cores

18 GB on-chip SRAM
3000x more on-chip memory

100 Pb/s interconnect
33,000x more bandwidth

215 mm

AJProença, Advanced Architectures, MiEI, UMinho, 2020

What is needed to increase the #cores in a chip?

2-Socket S9200WK Node

9200 CPU

9200 CPU

Die0 — Die0

DDR4 12 DIMMs

Die1 — Die1

UPI (10.4GTs)

DDR4 12 DIMMs

DMI Gen 3 (X4)

LBG

Layout of one core

1.375 MB L3$

SF

CHA

Server AVX-512

Server 1MB L2$ & Ctl

256KB L2$

Extended 768KB L2$

Skylake Core

Core ADPLL, FIVR & Power delivery
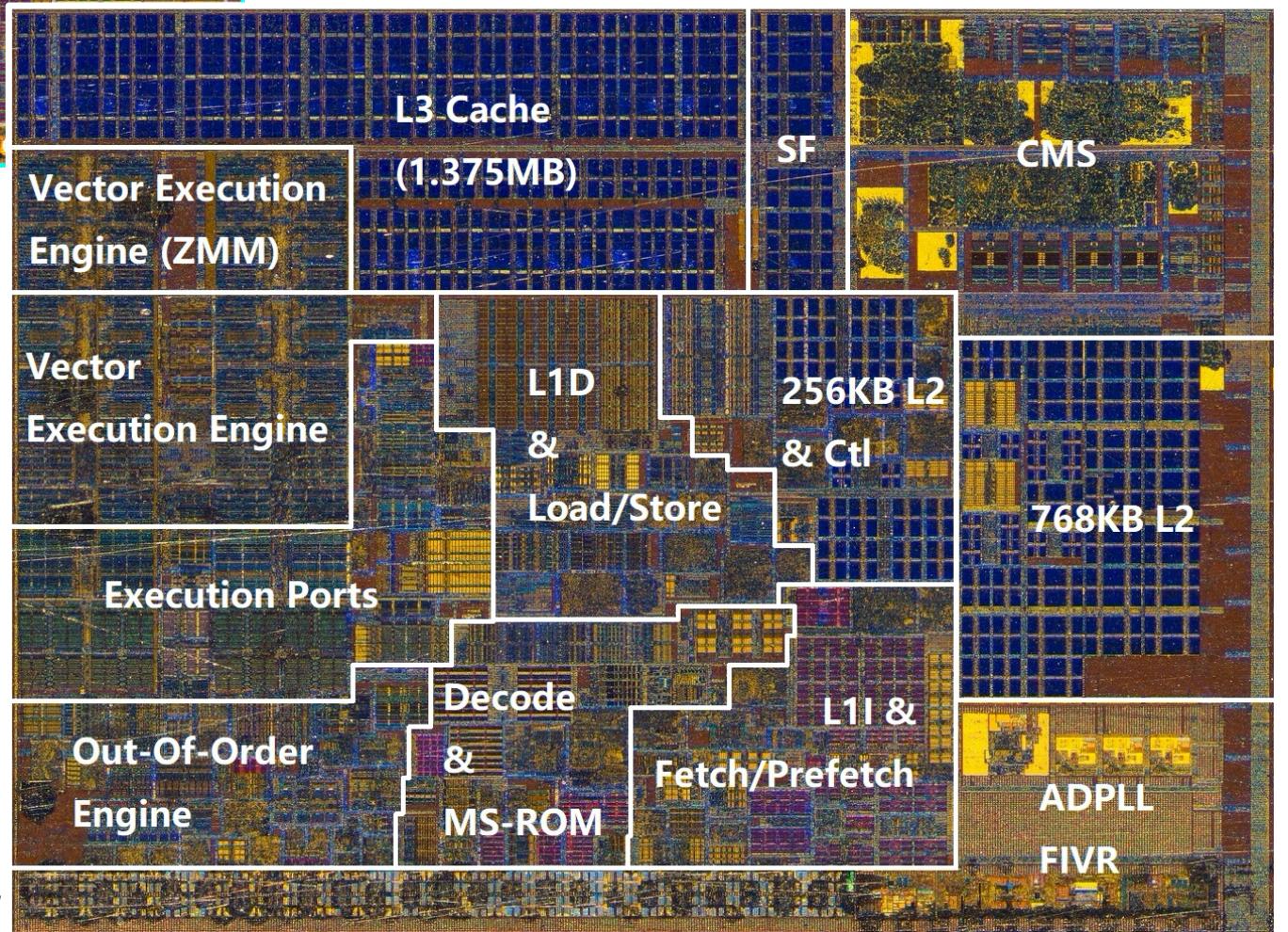
What is needed to increase the #cores in a chip?

Using the same microelectronics technology, **remove** parts from the core

Which parts?

- L3 cache
- AVX-512
- reduce L2 cache
- in-order exec
- less functional units
- …

AJProença, Advanced Architectures,

For each manufacturer identify the max hw support for SMT at each core (how many ways):

- Intel Xeon
- AMD Epyc
- Fujitsu Arm64FX
- IBM Power 9
- Sunway SW2610
- Apple A14