# Master Informatics Eng.

2022/23

*A.J.Proença*

## The move from multicore to manycore

*(most slides & images are borrowed)*

# From multicore to manycore: key issues

## Tips to fit many cores into a single chip

- lower the compute capability of each core, but not too much
- use a scalable interconnection network fabric on-chip (NoC)
  - to minimize shared-cache/memory access latencies
  - to provide enough data communication bandwidth
  - to minimize traffic bottlenecks
- group cores in clusters to improve the quality of the NoC
- reduced cache size/levels (*beware: strong impact on performance*)
- smaller fabrication process (KNL: 14nm; Ice Lake: 10nm; Apple M1: 5nm)
- mix general-purpose PUs with application-oriented modules: GPUs for vector computing, TPU for tensor computing, …
- move to MCM/chiplets (*simpler chips have better wafer production yield*)

# Interconnect Fundamentals

- **Networks-on-chip**: an adapted follow-up of interconnection systems to link servers in supercomputers
- Key parameters that define a **NoC**:
  - **topology**: defines how the nodes and links are connected, namely all possible paths a message can take through the network
  - **routing algorithm**: selects the specific path a message will take from source to destination
  - **flow control protocol**: determines how a message actually traverses the assigned route
  - **router micro architecture**: implements the routing and flow control protocols and critically shapes its circuits

# *Manycore chips/packages: an overview*

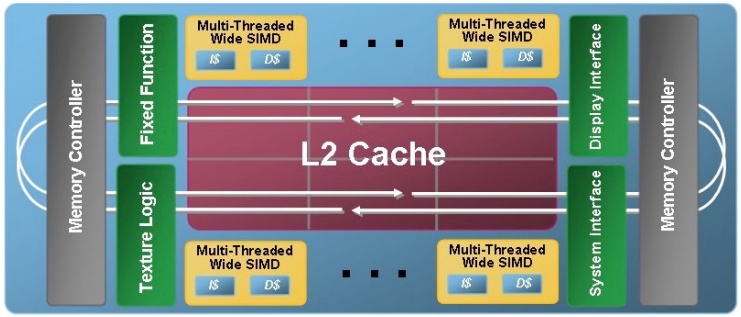
Key server chips/packages that addresses those issues:

- **Intel: from Intel MIC to the Xeon Scalable family**
- AMD: the Epyc Zen family
- ARM: key ARMv8 & v9 server-level competitors
  - Marvell ThunderX family
  - Fujitsu A64FX Arm chip
  - Neoverse hyperscale reference design for
    - Ampere Altra Arm
    - Amazon Graviton
  - Alibaba Yitian 710
  - Huawei HiSilicon Kunpeng 920
- Sunway: the SX260x0 family
- Cerebras: a Wafer Scale Engine
- Apple *(not server…)*: the SoC approach *(no chiplets!)*

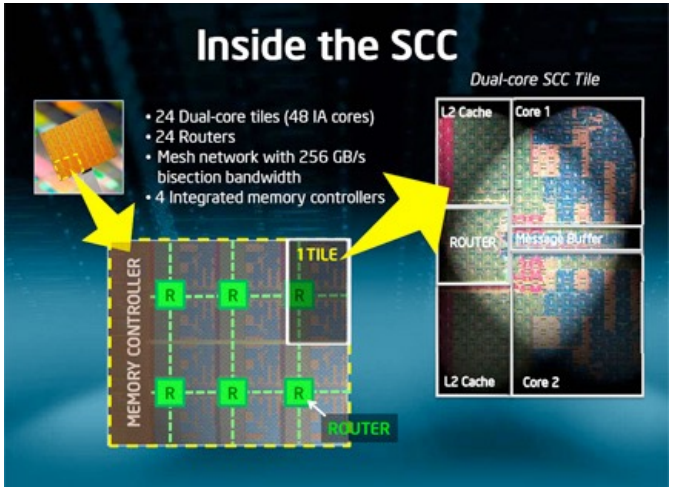
# Intel MIC: Many Integrated Core

## Intel evolution, from:

- Larrabee (80-core GPU) & SCC

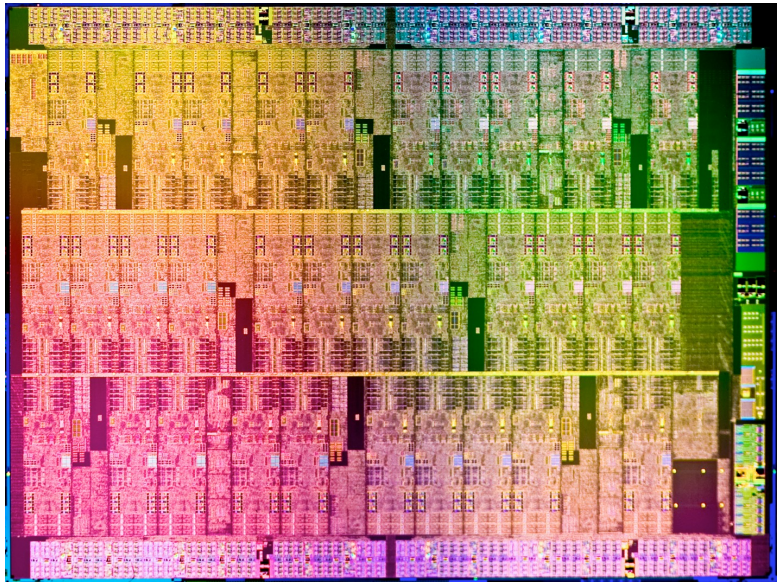



**S**ingle-chip
**C**loud
**C**omputer,
24x
dual-core tiles

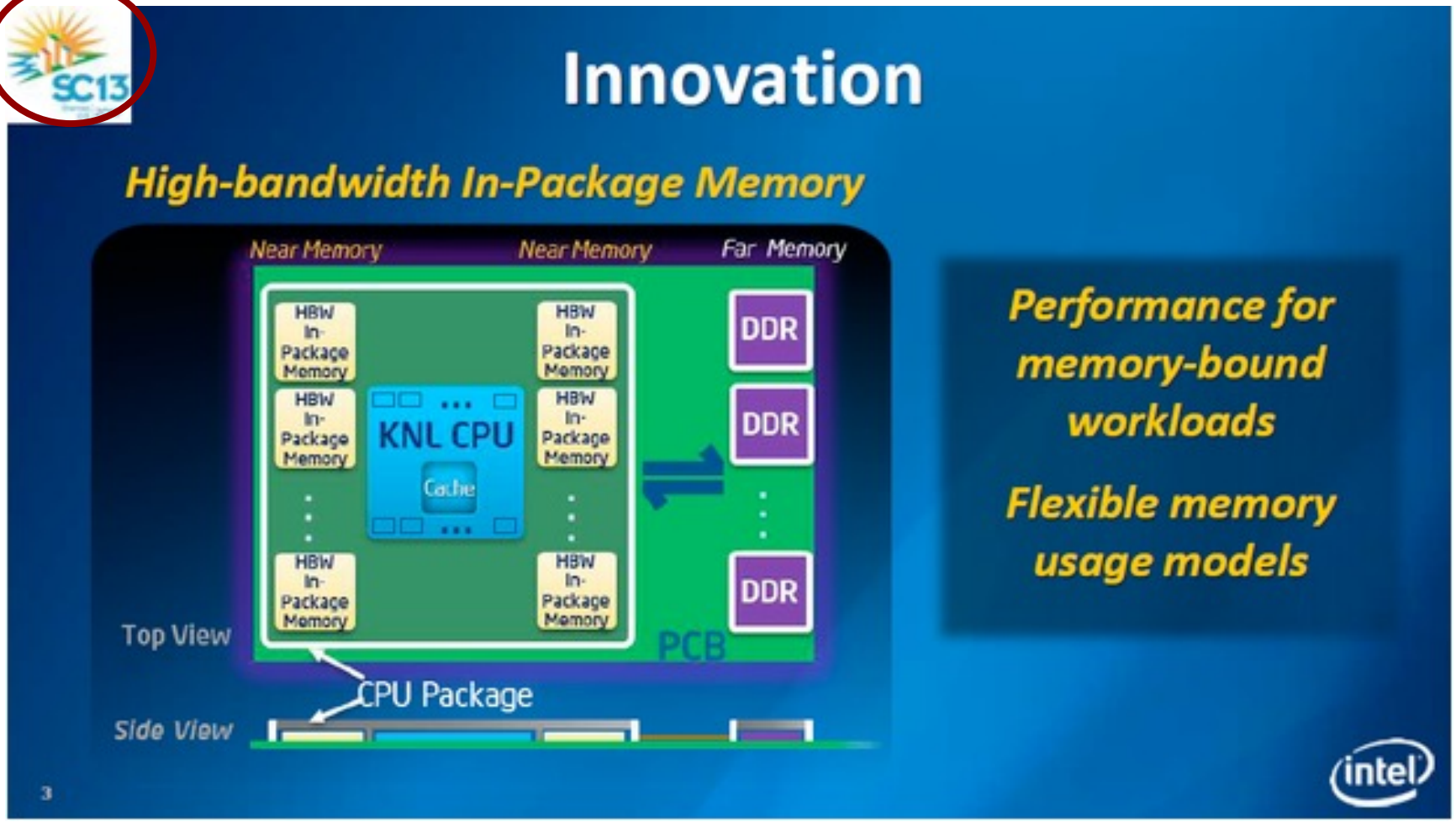



## to MIC:

- Knights Ferry (pre-production, Stampede)
- Knights Corner →
  Xeon Phi co-processor up to 61 Pentium cores
- Knights Landing (~~& Knights Mill...~~)
  Xeon Phi full processor up to 36x dual-core Atom tiles



# The Knights Landing architecture




Launched in June 2016
Discontinued in July 2018

# INTRODUCTION TO THE INTEL® XEON PHI™ PROCESSOR (CODENAME "KNIGHTS LANDING")

Dr. Harald Servat - HPC Software Engineer
Data Center Group – Innovation Performing and Architecture Group

Summer School in Advanced Scientific Computing 2016
~~February~~ June 21st, 2016 – Braga, Portugal

# INTEL® XEON PHI™ PROCESSOR FAMILY ARCHITECTURE OVERVIEW

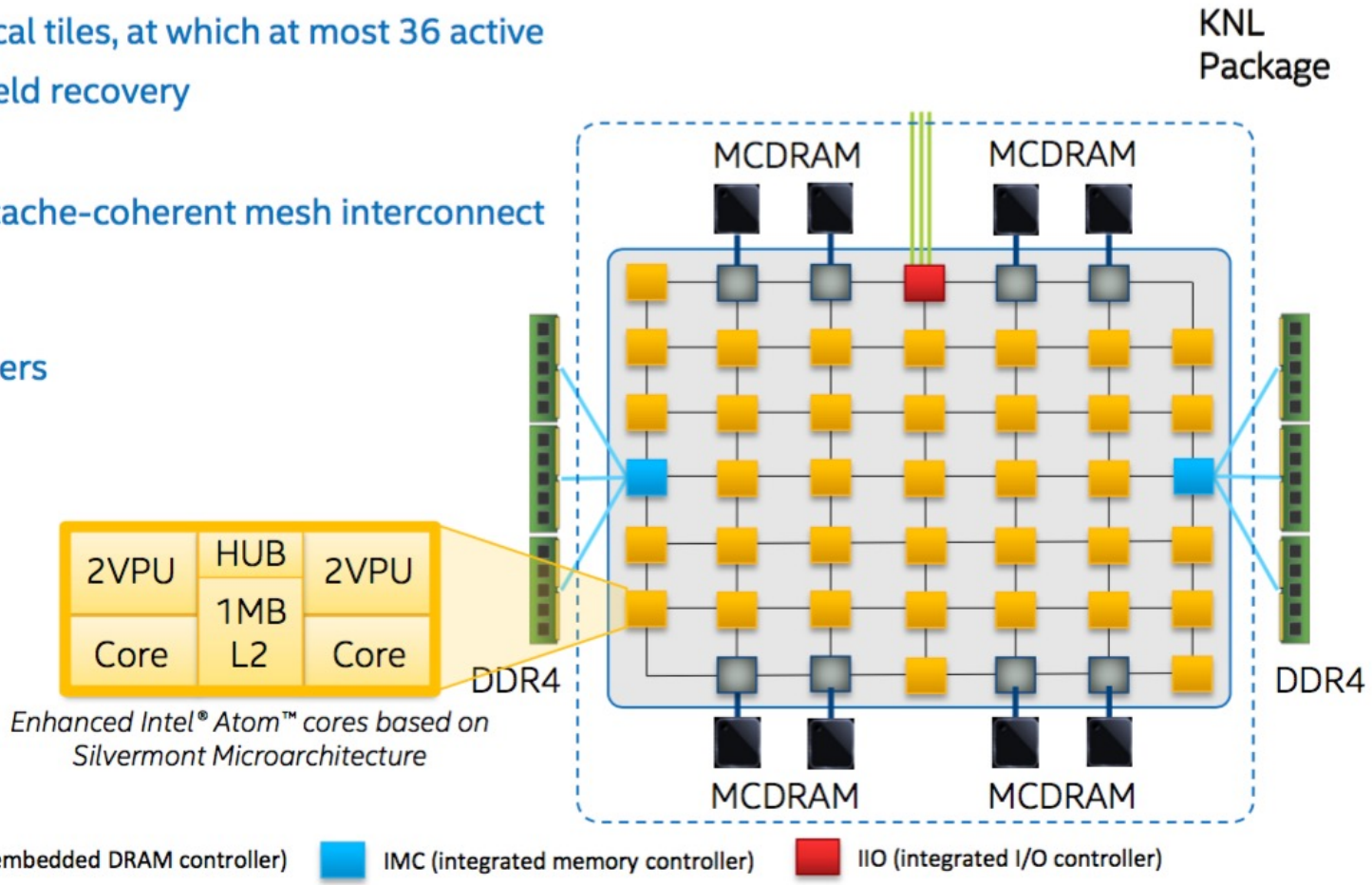
## Codenamed "Knights Landing" or KNL

Comprises 38 physical tiles, at which at most 36 active

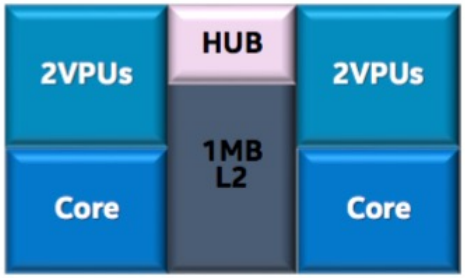
- Remaining for yield recovery

Introduces new 2D cache-coherent mesh interconnect (Untile)

- Tiles
- Memory controllers
- I/O controllers
- Other agents



# KNL PROCESSOR TILE




**Tile**

- 2 cores, each with 2 vector processing units (VPU)
- 1 MB L2-cache shared between the cores

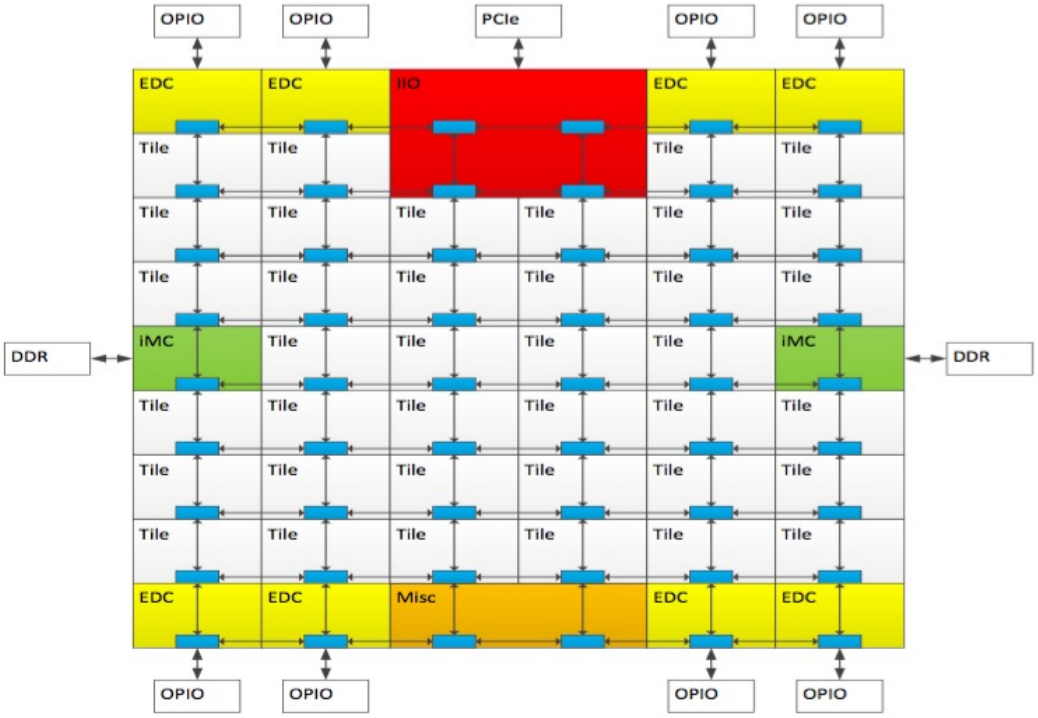
**Core**

- Binary compatible with Xeon
- Enhanced Silvermont (Atom)-based for HPC w/ 4 threads
- Out-of-order core
- 2-wide decode, 6-wide execute (2 int, 2 fp, 2 mem), 2-wide retire

**2 VPU**

- 512-bit SIMD (AVX512) 32SP/16DP per unit
- Legacy X87, SSE, AVX and AVX2 support

# KNL MESH INTERCONNECT




## Mesh of Rings

- Every row and column is a ring
- YX routing: Go in Y → Turn → Go in X
  - 1 cycle to go in Y, 2 cycles to go in X
- Messages arbitrate at injection and on turn

## Mesh at fixed frequency of 1.7 GHz

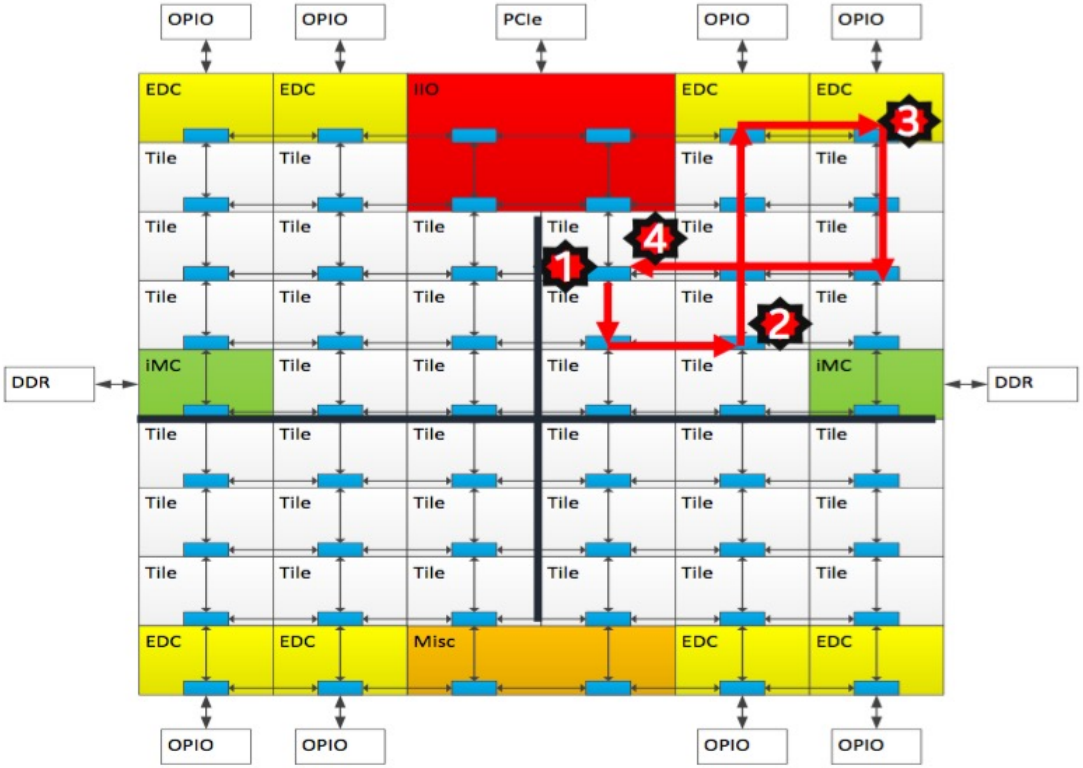
## Distributed Directory Coherence protocol

## KNL supports Three Cluster Modes

1) All-to-all
2) Quadrant
3) Sub-NUMA Clustering

## Selection done at boot time.

# CLUSTER MODE: SUB-NUMA CLUSTERING (SNC4)




**Each Quadrant (Cluster) exposed as a separate NUMA domain to OS**

**Analogous to 4-socket Xeon**

**SW Visible**

Typical Read L2 miss

1. L2 miss encountered
2. Send request to the distributed directory
3. Miss in the directory. Forward to memory
4. Memory sends the data to the requestor

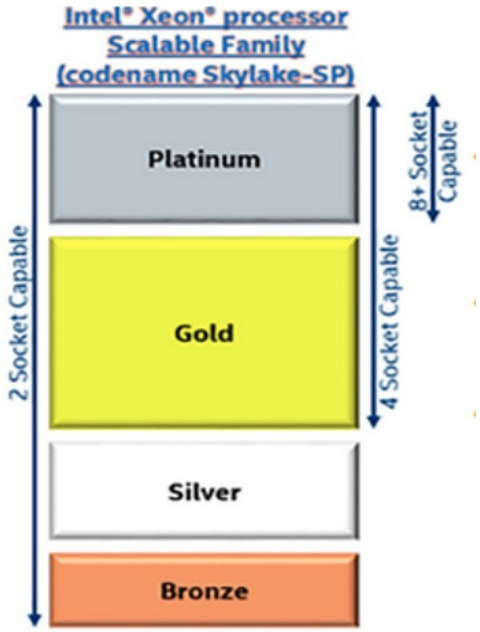
# *Intel Xeon Scalable Processor*
*(formerly code-named Skylake-SP)*

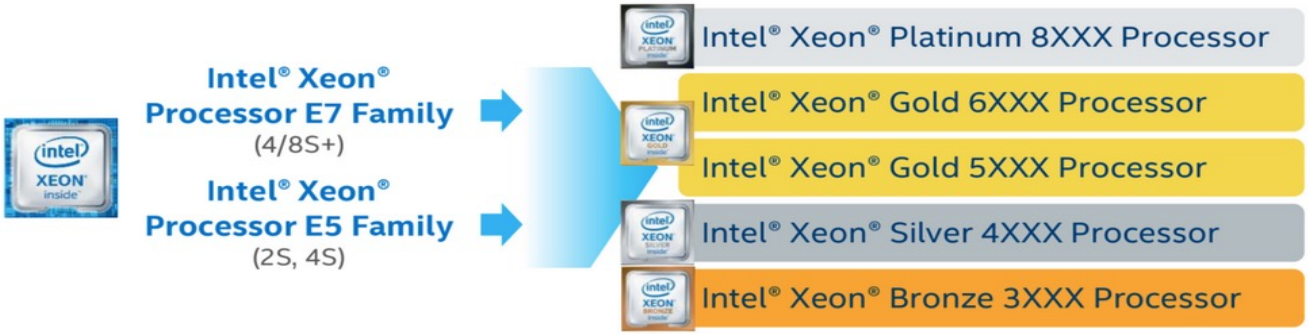
## First Generation Intel® Xeon® Scalable Processor

Introduced in July 2017

- Skylake-SP core microarchitecture with data center specific enhancements
- Intel® AVX-512 with 32 DP flops per cycle per core
- Data center optimized cache hierarchy – 1MB L2 per core, non-inclusive L3
- New Intel® Mesh architecture
- Enhanced 6 channel memory subsystem
- 48 lanes of PCIe Gen3 with integrated DMA, NTB, and VMD devices
- New Intel® Ultra Path Interconnect (Intel® UPI)

| Features | Intel® Xeon® Scalable Processor |
|---|---|
| Cores and Threads Per CPU | Up to 28 cores and 56 threads |
| Last-level Cache (LLC) | Up to 38.5 MB (non-inclusive) |
| QPI/UPI Speed (GT/s) | Up to 3x UPI @ 10.4 GT/s |
| PCIe* Lanes/ Controllers | Up to 48 / 12 / PCIe 3.0 (2.5, 5, 8 GT/s) |
| Memory Population | Up to 6 channels of up to 2 RDIMMs, LRDIMMs, or 3DS LRDIMMs |
| Max Memory Speed | Up to 2666 MHz |



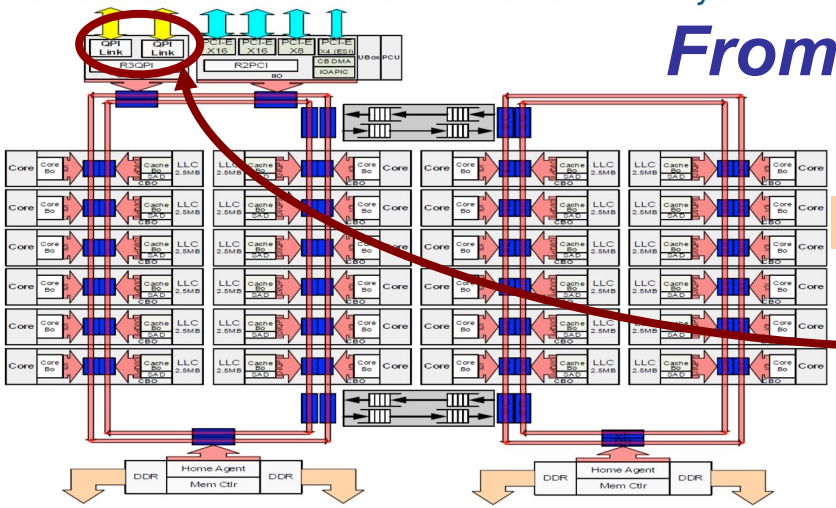






https://www.servethehome.com/resetting-scalable-expectations-intel-cascade-lake-sp-at-hot-chips-30/

Intel® Xeon® Processor E5 v4 Product Family HCC

# From Broadwell to Skylake (server): the move from ring to mesh



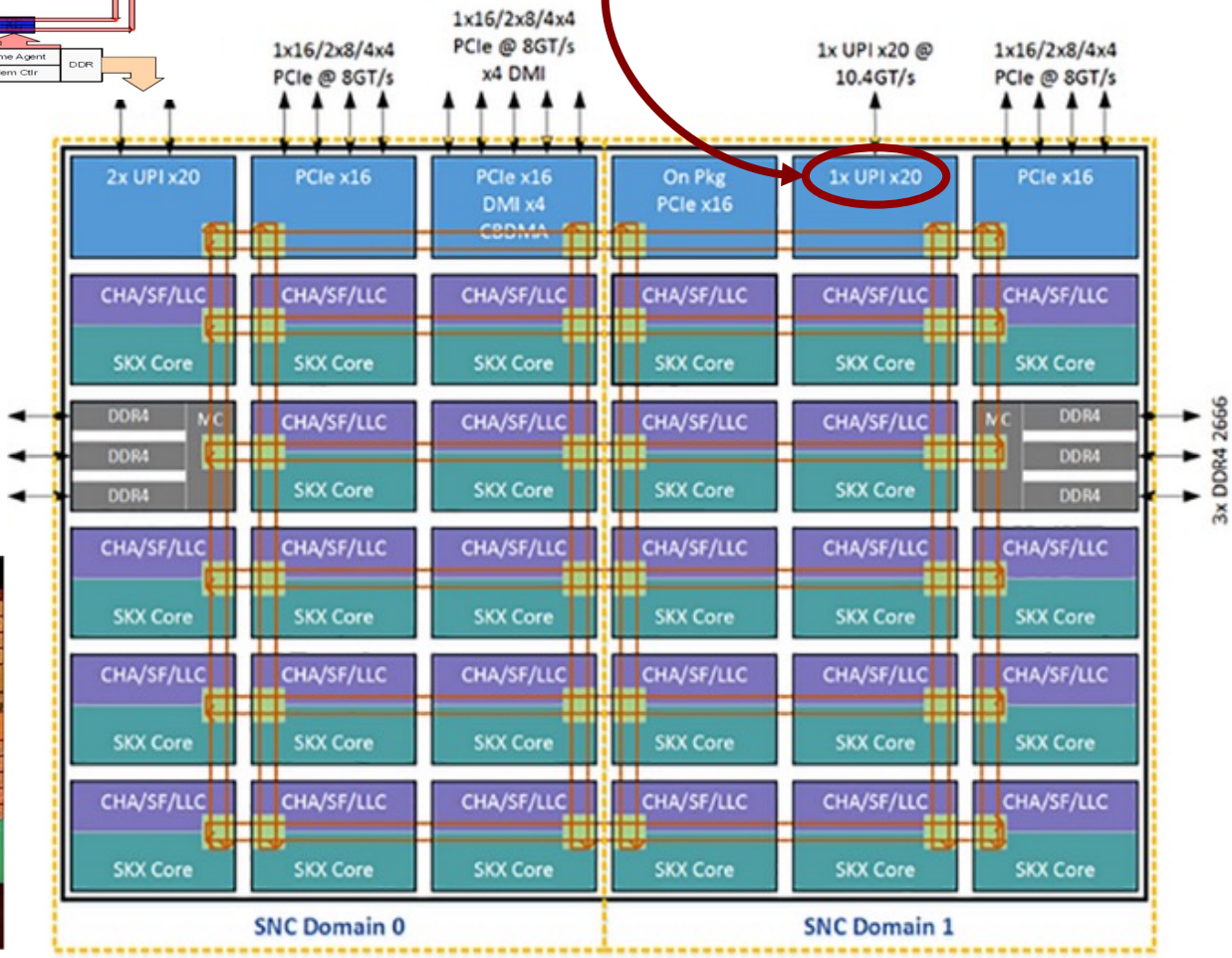

**UPI** required for dual-socket
*(**U**ltra **P**ath **I**nterconnect)*

Broadwell

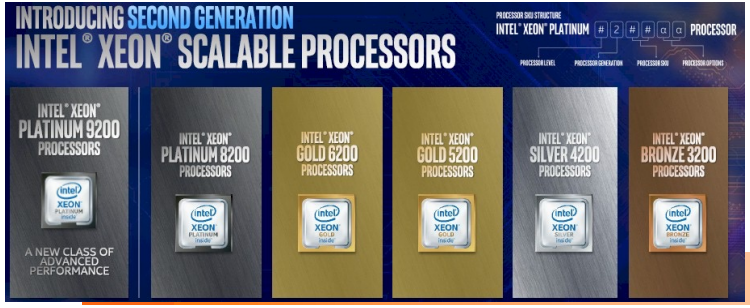
ring interconnection does not scale for large #cores

Skylake (server)

*(mesh follows KNL)*


https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview



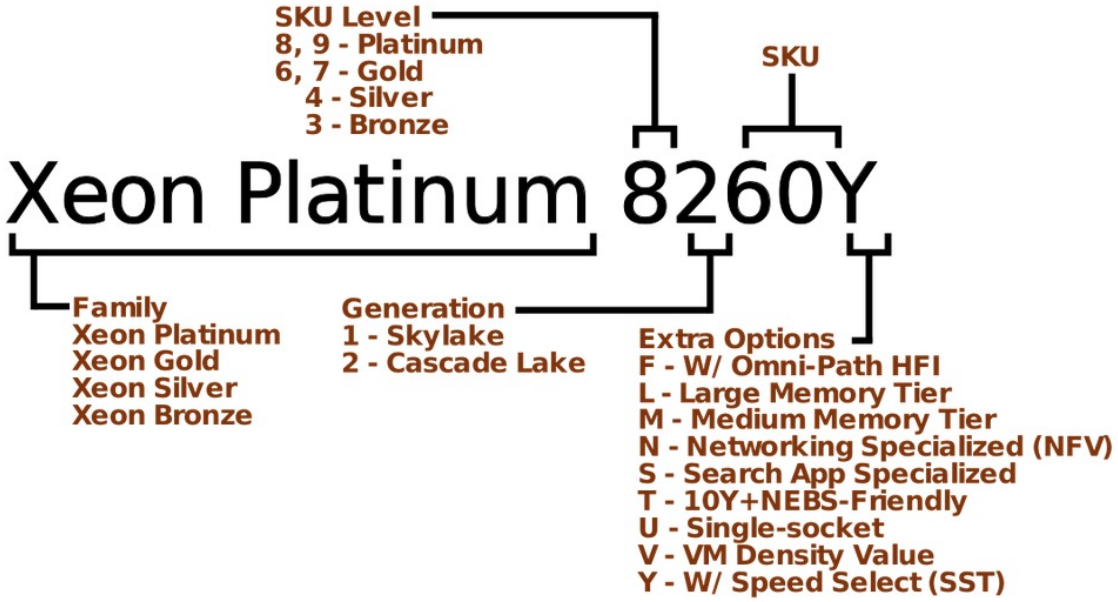

# 2nd Gen Intel Xeon Scalable Family: Cascade Lake *(launched April 2019)*

**1st Gen (Skylake)**

- Xeon Platinum 8100
- Xeon Gold 6100
- Xeon Gold 5100
- Xeon Silver 4100
- Xeon Bronze 3100

**2nd Gen (Cascade Lake)**

- Xeon Platinum 9200 –> 2x 8200 dies, 28+28 cores
- Xeon Platinum 8200 –> max 28 cores
- Xeon Gold 6200
- Xeon Gold 5200
- Xeon Silver 4200
- Xeon Bronze 3200




- **still 14 nm process!**


# Cascade Lake : partial view of a core architecture

Register Alias Table (RAT)
2x4 µOP
µOP µOP µOP µOP µOP µOP
Branch Order Buffer (BOB) (48-entry)
Load
FP
Int Vect
Int
Store
Common Data Buses (CDBs)
Move Elimination
Rename / Allocate / Retirement
ReOrder Buffer (224 entries)
Ones Idioms
Zeroing Idioms
µOP µOP µOP µOP µOP µOP µOP µOP
Integer Physical Register File
(180 Registers)
Scheduler
Unified Reservation Station (RS)
(97 entries)
Vector Physical Register File
(168 Registers)
Port 0 Port 1 Port 5 Port 6 Port 2 Port 3 Port 4 Port 7
µOP µOP µOP µOP µOP µOP µOP µOP
INT ALU
INT DIV
INT ALU
INT MUL
INT ALU
LEA
INT ALU
Branch
AGU
Load Data
AGU
Load Data
Store Data
AGU
INT Vect ALU
INT Vect MUL
FP FMA
INT Vect ALU
INT Vect MUL
FP FMA
INT Vect ALU
INT Vect MUL
FP FMA
INT Vect ALU
INT Vect MUL
FP FMA
512b fused
512b (zmm only)
(optional)
512bit/cycle
AES
Vect String
FP DIV
Branch
Bit Scan
EUs
Execution Engine
64B/cycle
Unified STLB
L2 Cache
1 MiB 16-Way
64B/cycle
To L3
Store Buffer & Forwarding
(56 entries)
64B/cycle
Load Buffer
(72 entries)
64B/cycle
64B/cycle
L1 Data Cache
32KiB 8-Way
Data TLB
64B/cycle
Line Fill Buffers (LFB)
(10 entries)
Memory Subsystem

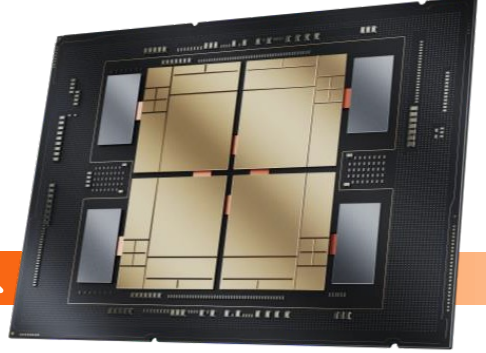

# 3rd Gen Intel Xeon Scalable Family: Ice Lake *(launched April 2021)*

*https://www.anandtech.com/show/16594/intel-3rd-gen-xeon-scalable-review*

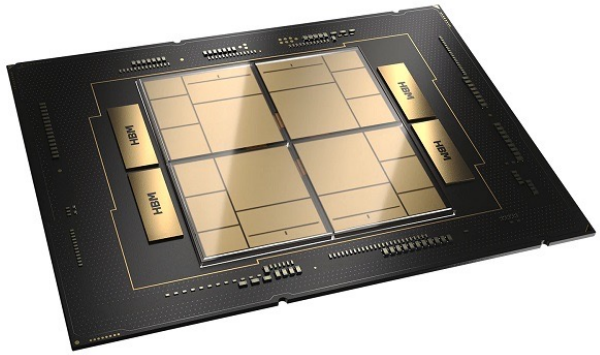
## Key notes:

- **up to 40 cores and 60 MiB L3** (*platinium*)
- **no bronze chips**
- **connects up 2 devices**
- **yet with 10 nm process**

| Intel Xeon Comparison: 3rd Gen vs 2nd Gen Peak vs Peak | | |
|---|---|---|
| Xeon Platinum 8380 | AnandTech | Xeon Platinum 8280 |
| 40 / 80 | **Cores / Threads** | 28 / 56 |
| 2900 / 3400 / 3000 | **Base / ST / MT Freq** | 2700 / 4000 / 3300 |
| 50 MB + 60 MB | **L2 + L3 Cache** | 28 MB + 38.5 MB |
| 270 W | **TDP** | 205 W |
| PCIe 4.0 x64 | **PCIe** | PCIe 3.0 x48 |
| 8 x DDR4-3200 | **DRAM Support** | 6 x DDR4-2933 |
| 4 TB | **DRAM Capacity** | 1 TB |



# 4th Gen Intel Xeon Max Family: Saphire Rapids (January 2023)




# 4th Gen Intel Xeon Max Family: Saphire Rapids (January 2023)

## Next-Generation Intel Xeon Scalable Processors

Unique Capabilities Optimized for HPC and AI Acceleration


**Breakthrough Technology**

DDR5 — Increased Memory BW

PCIE 5 — High Throughput

CXL 1.1 — Next-gen IO

**Built-In AI Acceleration**

Intel® Advanced Matrix Extensions (AMX)

Increased Deep Learning Inference and Training Performance

**Agility and Scalability**

Hardware Enhanced Security

Intel® Speed Select Technology

Broad Software Optimization

NEW

**High Bandwidth Memory**

Significant performance increase for bandwidth-bound workloads

Up to 64 GB HBM2e

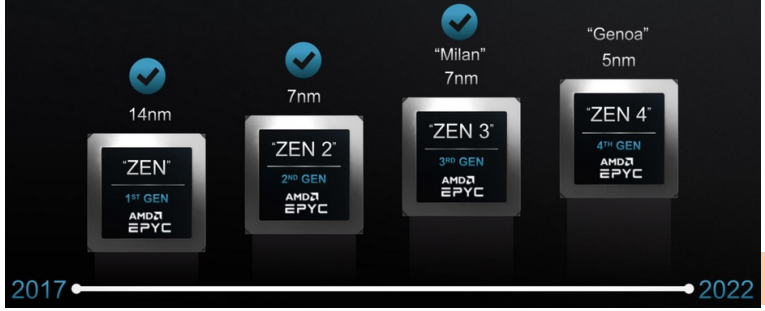
intel.


**4x** 14-core tiles
4x 2 mem ch

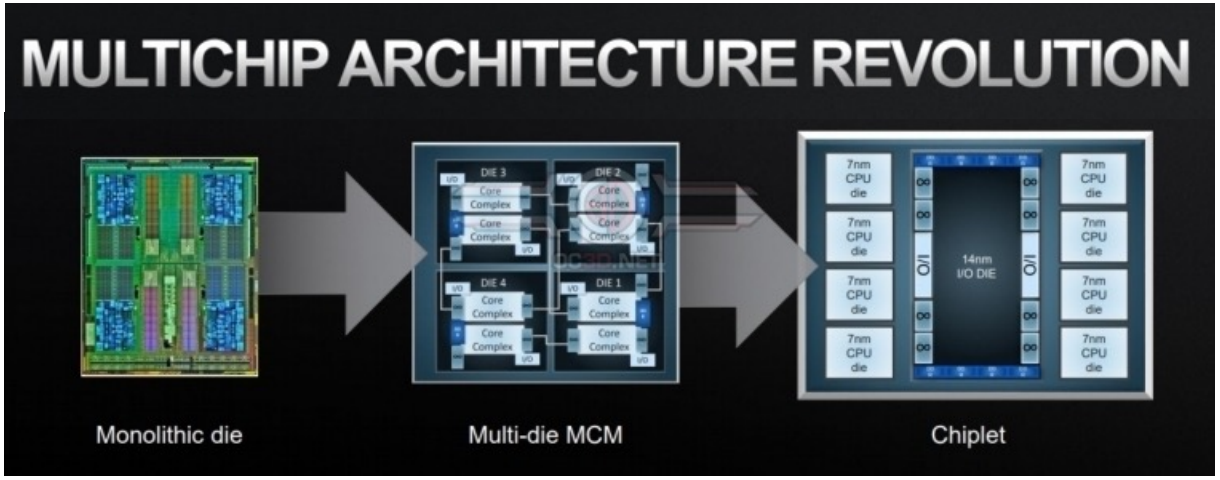
# *Manycore chips/packages: an overview*

Key server chips/packages that addresses those issues:

- Intel: from Intel MIC to the Xeon Scalable family
- **AMD: the Epyc Zen family**
- ARM: key ARMv8 & v9 server-level competitors
  - Marvell ThunderX family
  - Fujitsu A64FX Arm chip
  - Neoverse hyperscale reference design for
    - Ampere Altra Arm
    - Amazon Graviton
  - Alibaba Yitian 710
  - Huawei HiSilicon Kunpeng 920
- Sunway: the SX260x0 family
- Cerebras: a Wafer Scale Engine
- Apple *(not server…)*: the SoC approach *(no chiplets!)*

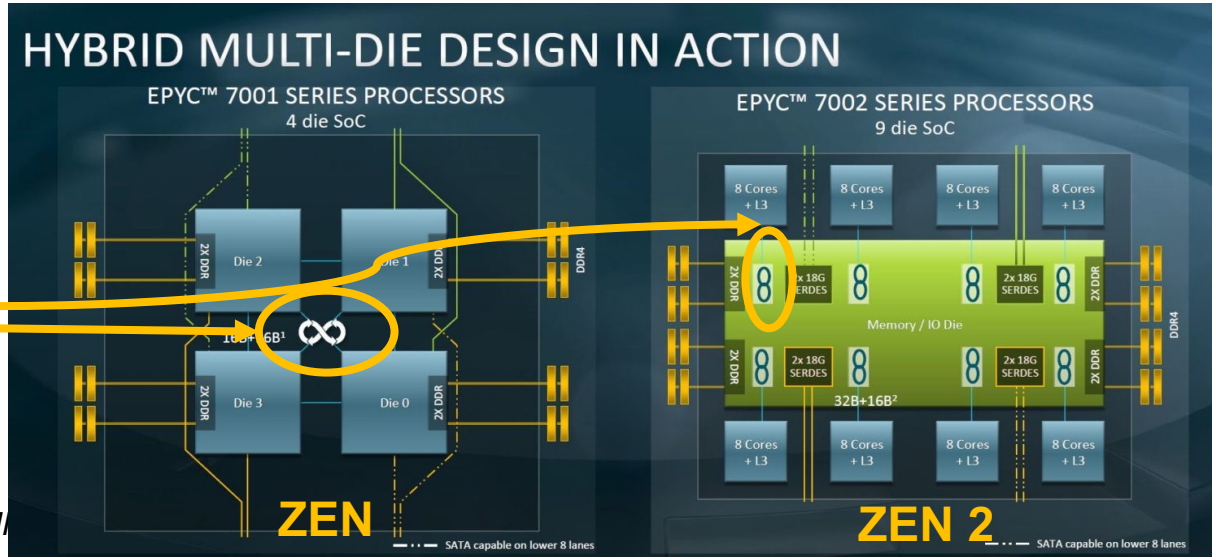



# *Key Intel Xeon competitor: AMD Epyc (Zen, Zen 2, 3, 4)*



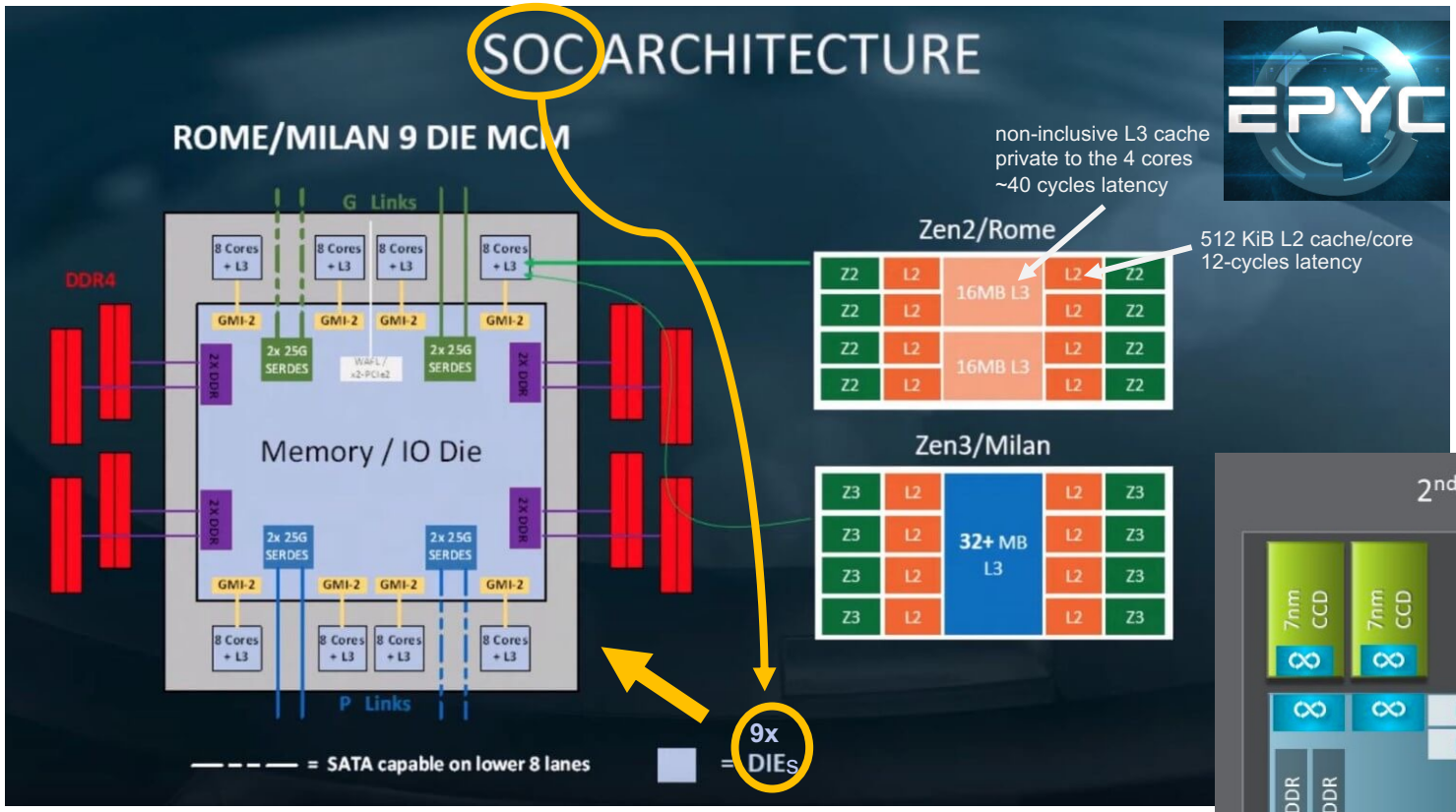


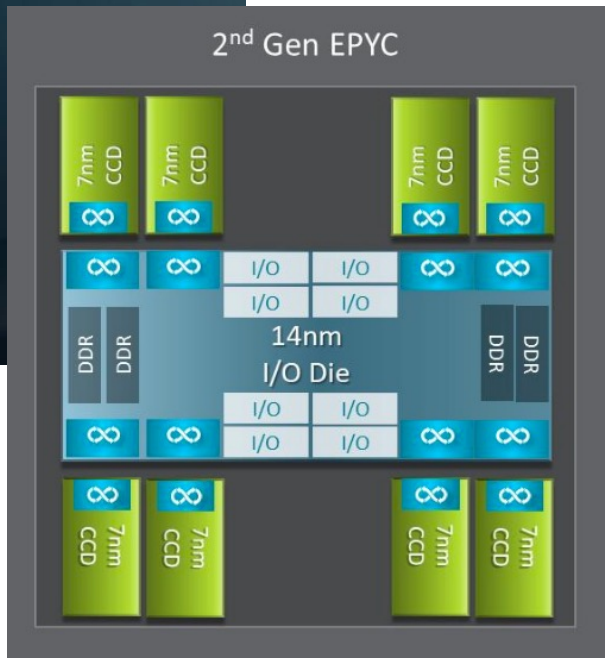
**Infinity**:
AMD interconnection fabric,
a superset of HyperTransport




*AJProença, Parallel Computing, MEI, U*


# AMD Epyc: from Zen 2 (Rome) to Zen 3 (Milan)



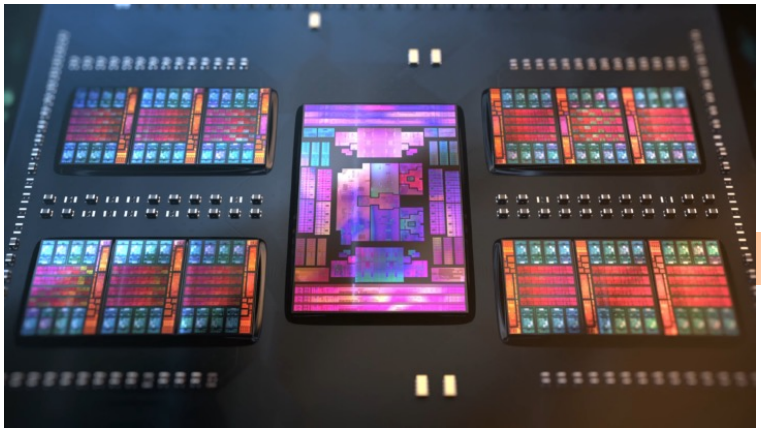




8x 8-core CCD dies at 7 nm
1x I/O die at 14 nm

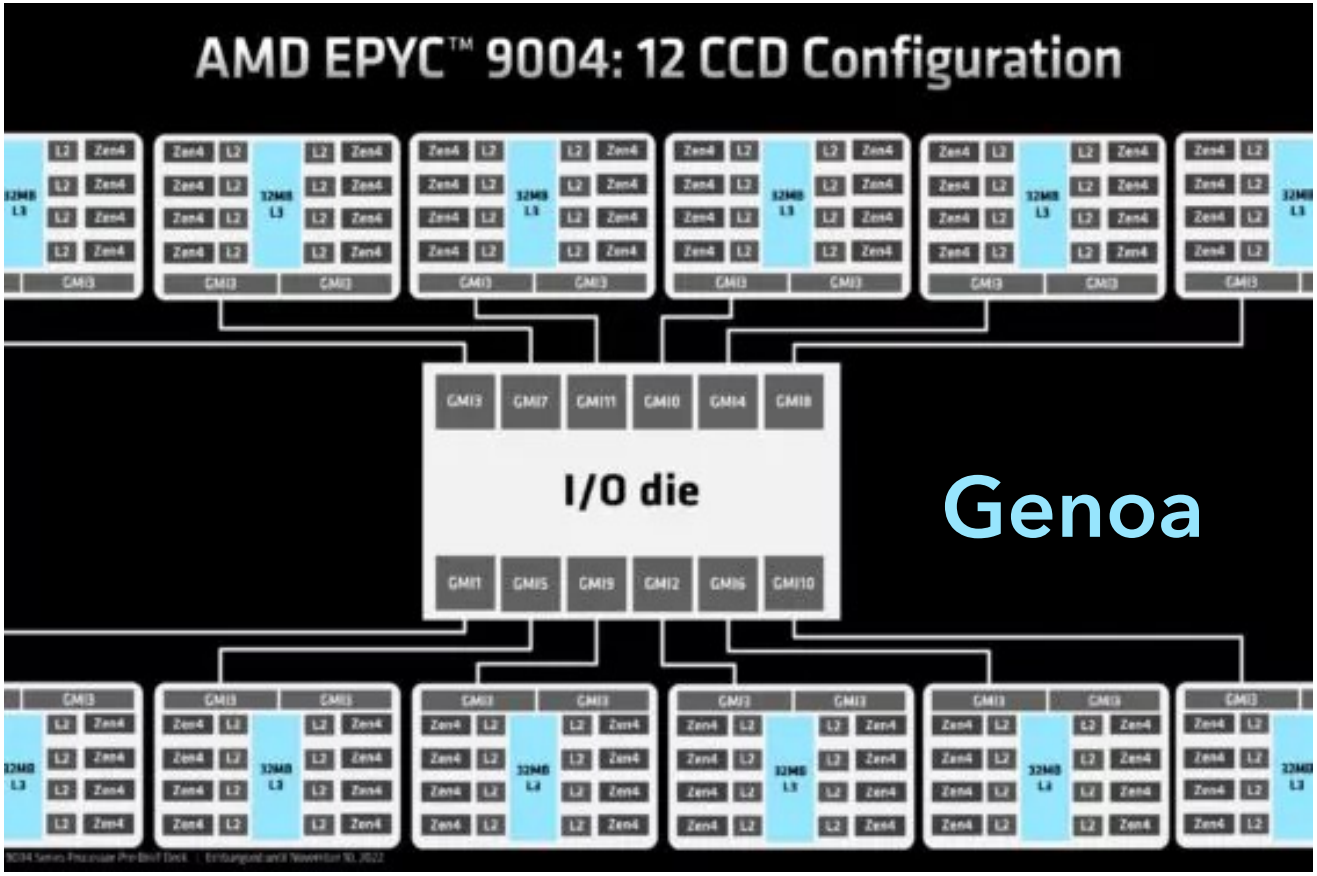


https://wccftech.com/amd-epyc-rome-7nm-server-cpu-official-launch-64-core-128-thread-128-pcie-gen4/

# AMD 4th Gen Epyc: Zen 4 (Genoa & Bergamo)



**Genoa**: up to 5nm 96 cores /192 threads
BFLOAT16, VNNU, AVX-512
1 MiB/core L2 cache
384 MiB L3 cache
12 DDR5 mem chan
128 PCIe Gen 5.0
*(launched Nov'22)*




**Bergamo**:
- up to 128 Zen 4c cores
- for cloud-native computing

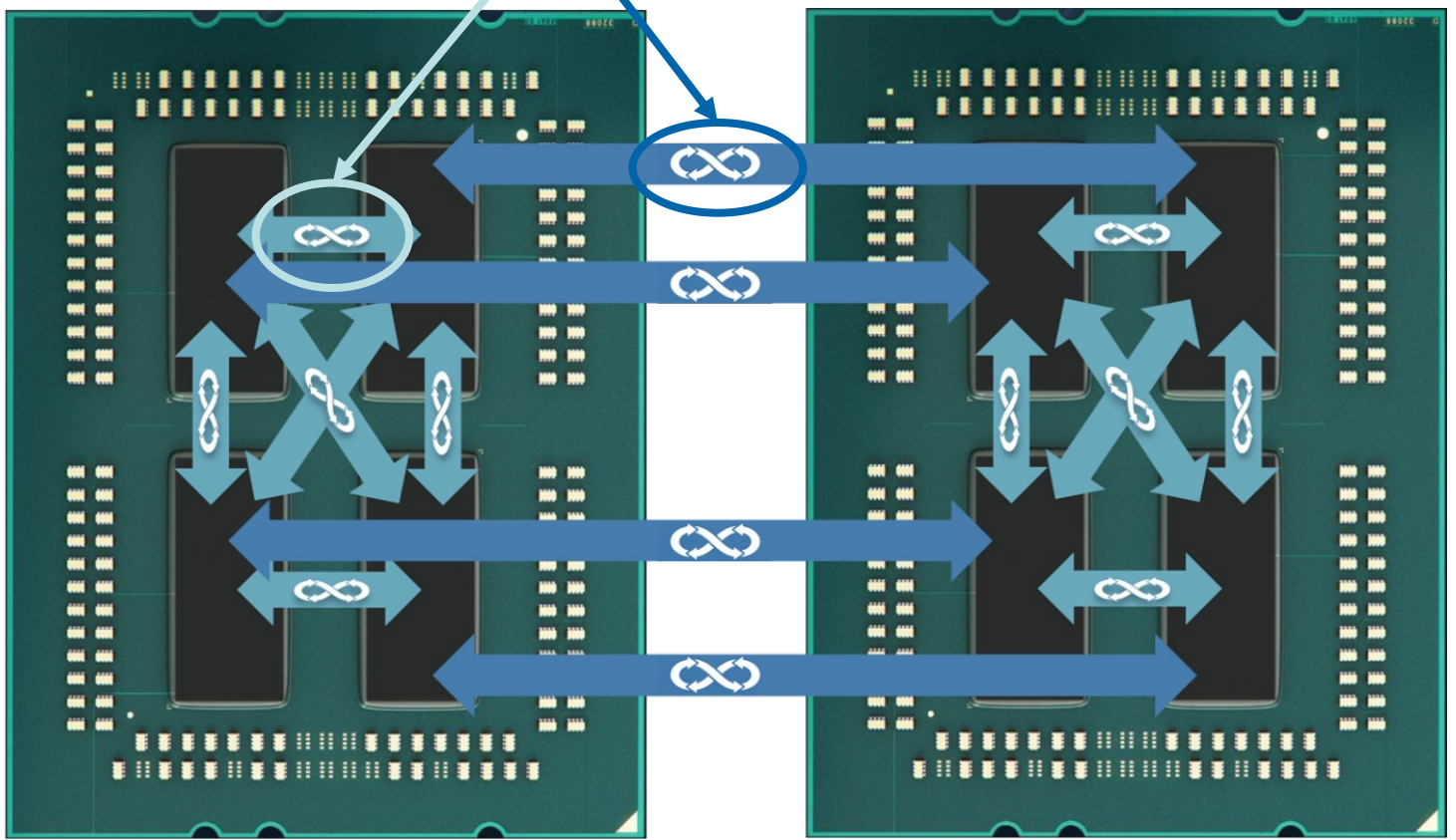
*(expected 1H 2023)*

https://wccftech.com/amd-4th-gen-epyc-genoa-zen-4-cpus-official-launch-96-cores-192-threads-worlds-fastest-server-chips/




## **Infinity Fabric** (IF)

- AMD system interconnect architecture, a 256-wide bi-directional crossbar:
  - **Infinity Fabric On-Package** (IFOP): die-to-die communication in same package
  - **Infinity Fabric InterSocket** (IFIS): for package-to-package communications



# *Manycore chips/packages: an overview*

Key server chips/packages that addresses those issues:

- Intel: from Intel MIC to the Xeon Scalable family
- AMD: the Epyc Zen family
- **ARM: key ARMv8 & v9 server-level competitors**
  - **Marvell ThunderX family**
  - **Fujitsu A64FX Arm chip**
  - **Neoverse hyperscale reference design for**
    - **Ampere Altra Arm**
    - **Amazon Graviton**
  - **Alibaba Yitian 710**
  - **Huawei HiSilicon Kunpeng 920**
- Sunway: the SX260x0 family
- Cerebras: a Wafer Scale Engine
- Apple *(not server)*: the SoC approach *(no chiplets!)*

Suppport for at least dual-socket

# ARM brand: a bit of history...

*Current owner of Arm Holdings: NVidia*

## ARM architecture

From Wikipedia, the free encyclopedia

**ARM**, previously **Advanced RISC Machine**, originally **Acorn RISC Machine**, is a family of reduced instruction set computing (RISC) architectures for computer processors, configured for various environments. Arm Holdings develops the architecture and licenses it to other companies, who design their own products that implement one of those architectures—including systems-on-chips (SoC) and systems-on-modules (SoM) that incorporate memory, interfaces, radios, etc. It also designs cores that implement this instruction set and licenses these designs to a number of companies that incorporate those core designs into their own products.

Processors that have a RISC architecture typically require fewer transistors than those with a complex instruction set computing (CISC) architecture (such as the x86 processors found in most personal computers), which

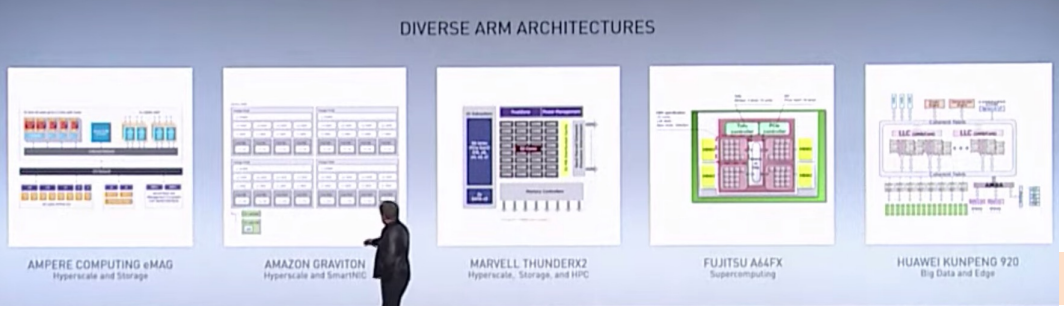
**ARM architectures**

The ARM logo

| | |
|---|---|
| **Designer** | Arm Holdings |
| **Bits** | 32-bit, 64-bit |
| **Introduced** | 1985; 34 years ago |
| **Design** | RISC |
| **Type** | Register-Register |
| **Branching** | Condition code, compare and branch |
| **Open** | Proprietary |

# HPCs with ARMv8: server-level competitors

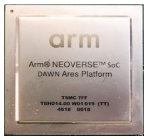
*(back in 2019…)*





1. Marvell ThunderX product family **(dead?...)**


2. Fujitsu A64FX Arm chip **(in MACC…)**




3. Neoverse N1 hyperscale reference design **(…)**

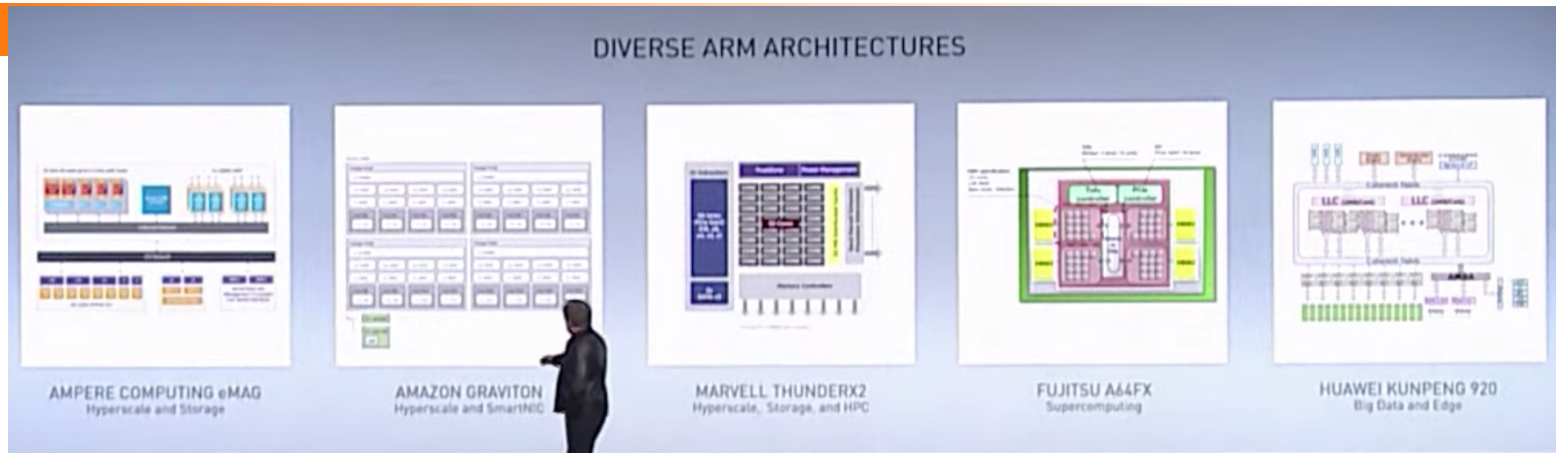

4. Ampere Altra Arm Processor **(…)**

5. Amazon Graviton **(…)**


6. Huawei HiSilicon Kunpeng 920 **(dead?...)**

https://www.eweek.com/pc-hardware/eweek-s-top-vendors-server-processor-makers

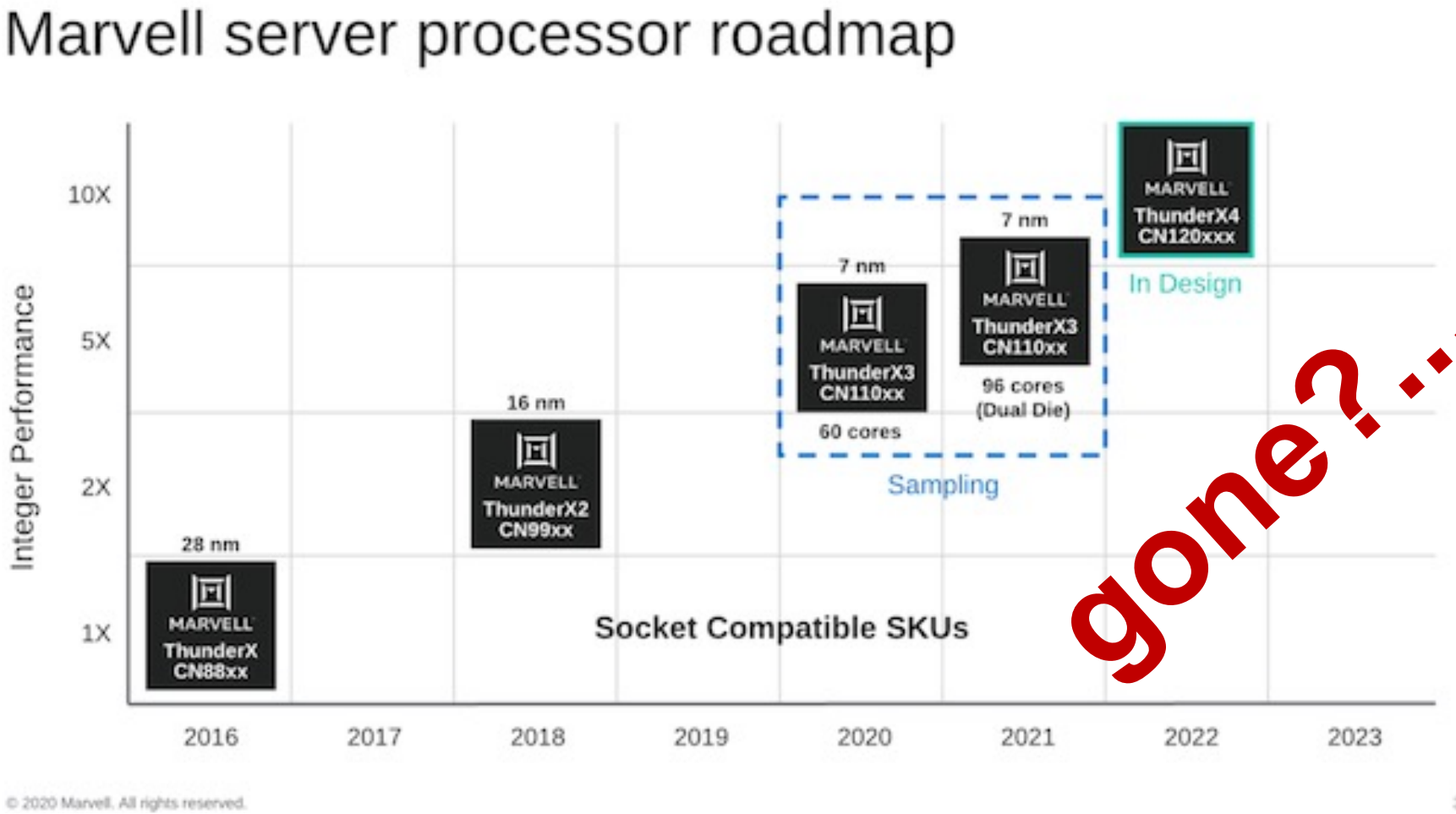
# HPCs with ARMv8: server-level competitors
*(back in 2019…)*





1. **Marvell ThunderX product family (dead?...)**
2. Fujitsu A64FX Arm chip **(in MACC…)**
3. Neoverse N1 hyperscale reference design **(…)**
4. Ampere Altra Arm Processor **(…)**
5. Amazon Graviton **(…)**
6. Huawei HiSilicon Kunpeng 920 **(dead?...)**

# Marvell Server Processor Roadmap

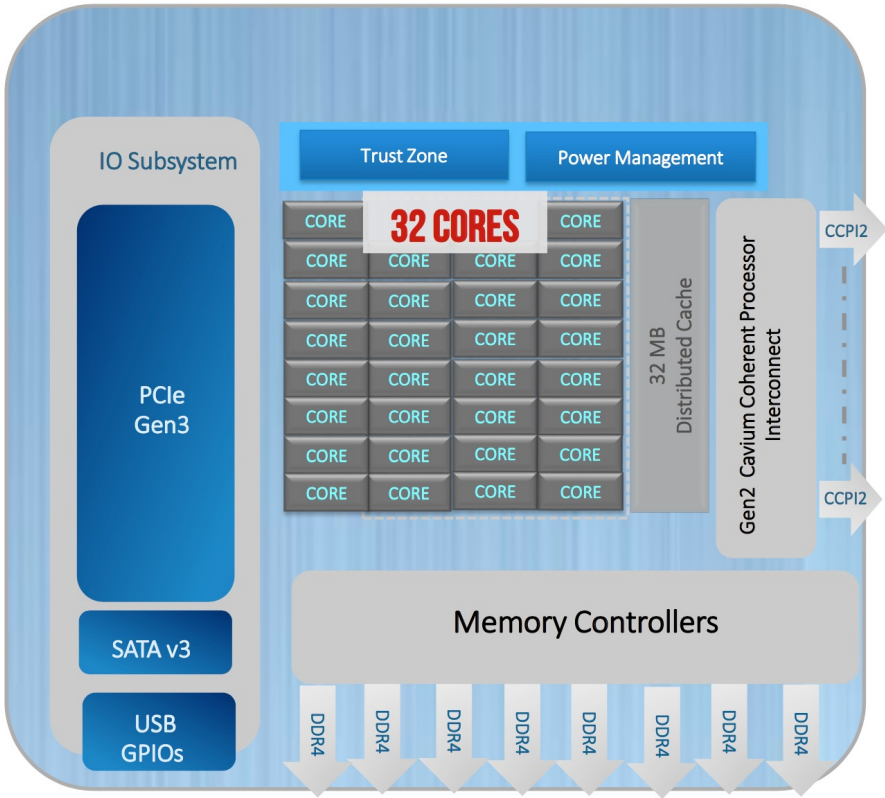



gone?...

https://www.anandtech.com/show/15995/hot-chips-2020-marvell-details-thunderx3


# The Marvell/Cavium ThunderX2

## THUNDERX2® Family Key Features



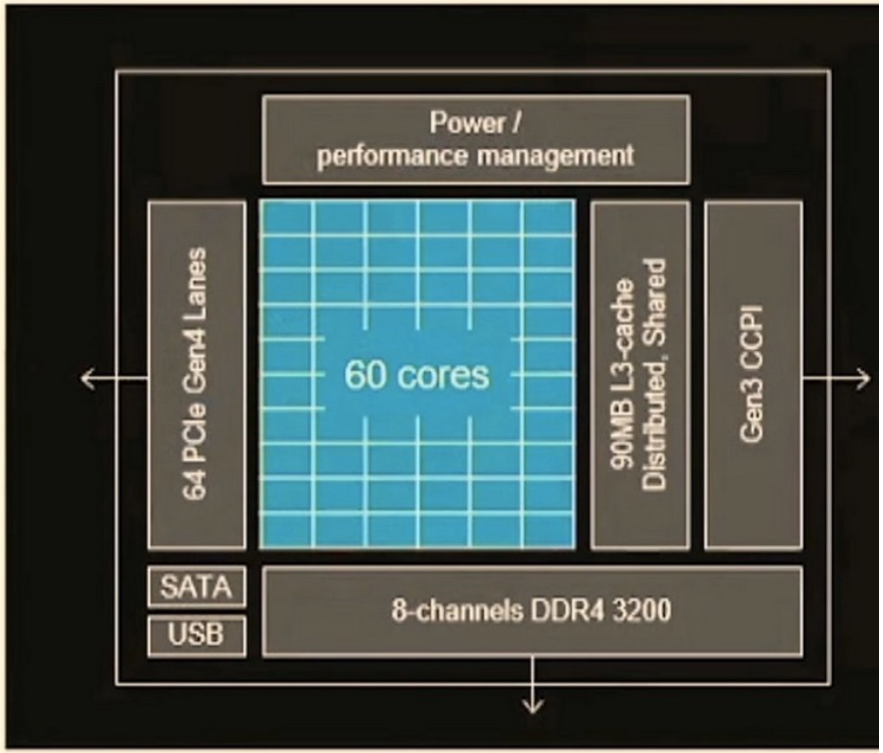

- Up to 32 custom Armv8.1 cores, up to 2.5GHz
- Full OoO, 1, 2, 4 threads per core
- 1S and 2S Configuration
- Up to 8 DDR4-2667 Memory Controllers, 1 & 2 DPC
- Up to 56 lanes of PCIe, 14 PCIe controllers
- Full SoC: Integrated SATAv3 USB3 and GPIOs
- Server class RAS & Virtualization
- Extensive Power Management
- LGA and BGA for most flexibility
- 40+ SKUs
- Volume SKU List Price: $1795 (180W) to $800 (75W)

https://www.hpcwire.com/2018/05/07/cavium-announces-ga-for-thunderx2-touts-oem-support/


# Next generation: ThunderX3

## ThunderX3™ overview

- Single die: Up to 60 cores
- Dual die: Up to 96 cores
- Arm v8.3 with select v8.4/v8.5 features
- 30% single thread gain at equal frequency over ThunderX2
- Up to four threads per core
- High bandwidth switched ring interconnect
- Up to 8 DDR4-3200 channels
- Single die: 2X-3X perf over ThunderX2 at equal power
  - Further gains from dual die
- Up to 64 PCIe Gen4, 16 PCIe controllers
- Fine grain power monitoring/management
- TSMC 7nm





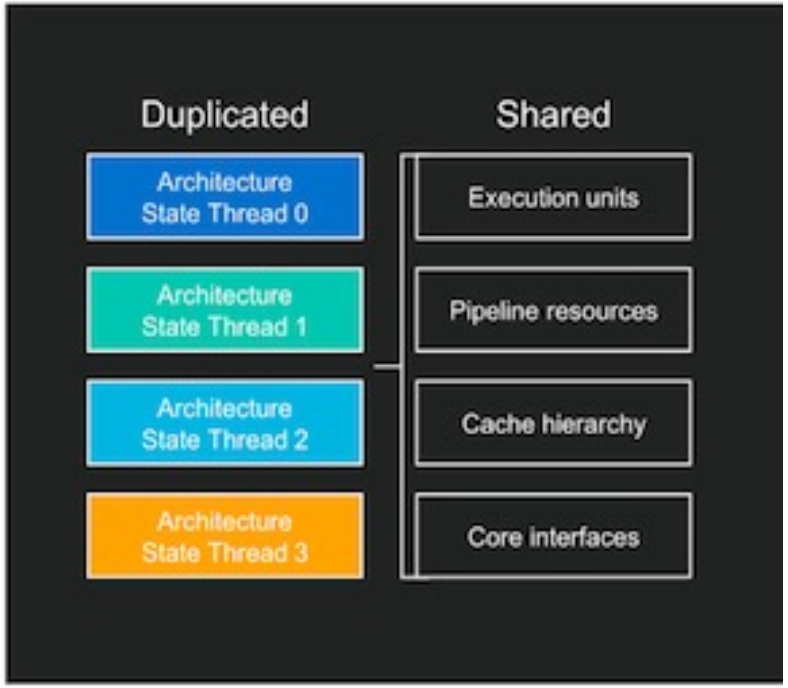

4

https://www.anandtech.com/show/15995/hot-chips-2020-marvell-details-thunderx3


# 4-way SMT in ThunderX3

## Multithread execution

- Four hardware threads per core
- Each thread includes full copy of Arm architecture state
- Threads share core pipeline resources
- To OS each thread appears as a regular Arm CPU
  - So four CPUs per core
- Area impact of 4-way SMT relative to no SMT: ~5%
- ThunderX3 has 60 cores / 240 threads per die

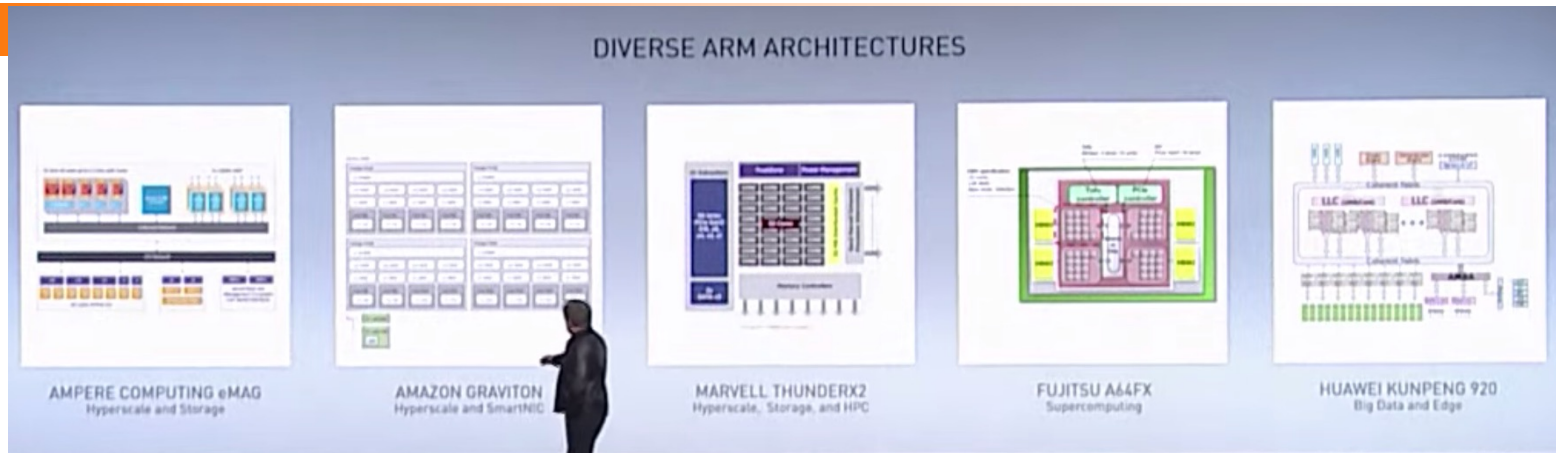





12

https://www.anandtech.com/show/15995/hot-chips-2020-marvell-details-thunderx3

# *HPCs with ARMv8: server-level competitors*
*(back in 2019…)*



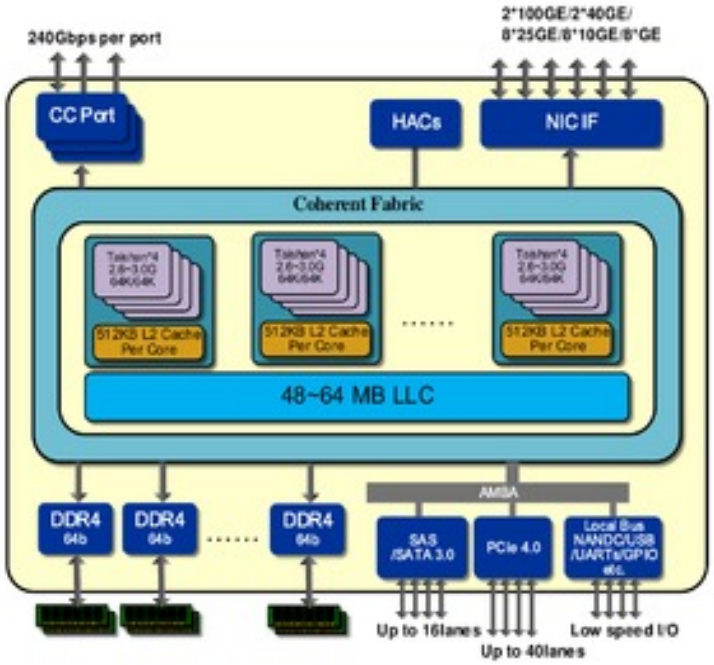

1. Marvell ThunderX product family (dead?...)
2. Fujitsu A64FX Arm chip **(in MACC…)**
3. Neoverse N1 hyperscale reference design **(…)**
4. Ampere Altra Arm Processor **(…)**
5. Amazon Graviton **(…)**
6. **Huawei HiSilicon Kunpeng 920 (dead?...)**

# The Huawei Kunpeng 920
*(previously known as HiSilicon Hi1620)*

*(launched in 2019)*

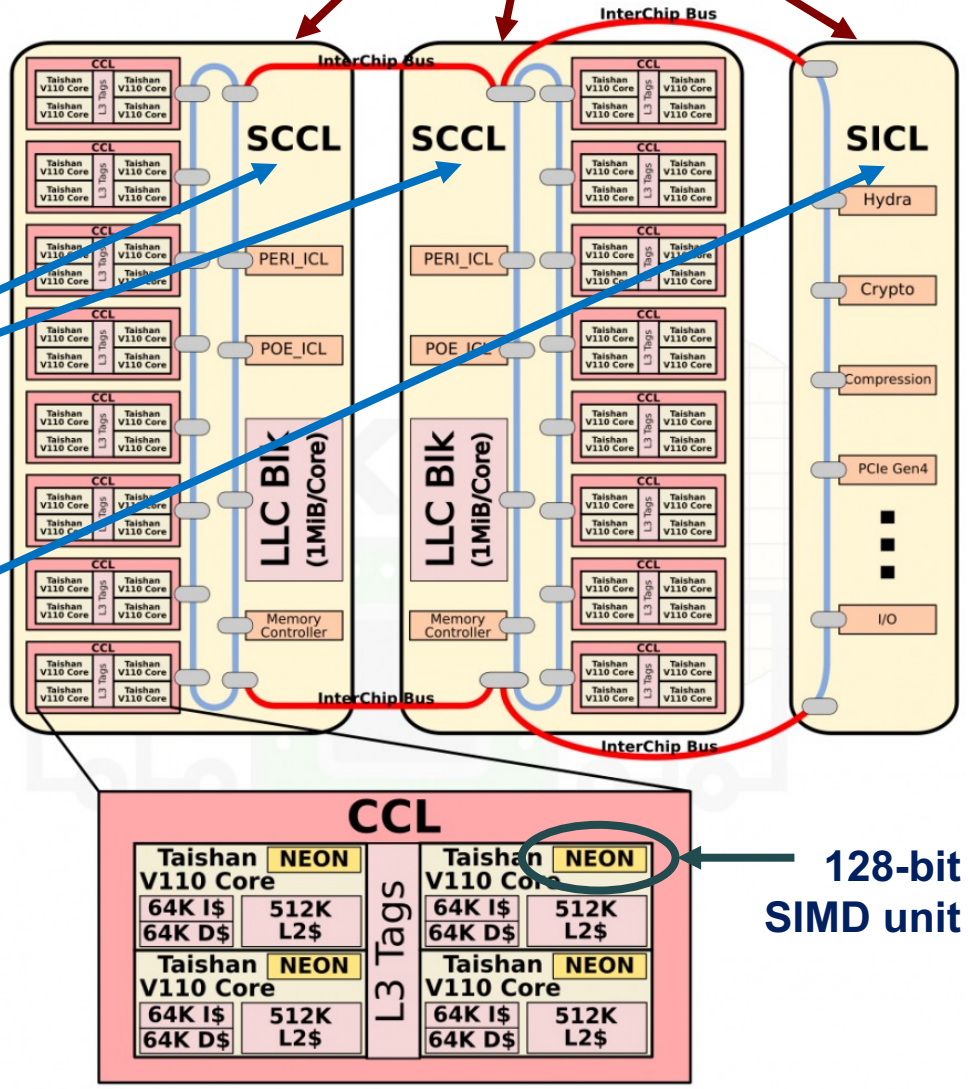
Huawei Kunpeng 920: based on TaiShan V110 core, a semi-custom ARM Cortex-A72

## Hi1620 Specifications Overview




| | |
|---|---|
| CPU core | Up to 64 ARMv 8.2 cores, 3.0 GHz, 48-bit physical address<br>4 issue OoO superscalar design<br>64 KB L1 I Cache and 64 KB L1 D cache 128-bit SIMD unit |
| L2 cache | 512 KB private per core, 24 MB total |
| L3 cache | 48 MB shared for all (1 MB/core), Partitioned |
| Memory | 8-channel DDR4-2400/2666/2933/3200<br>16 ranks/channel, 1DPC and 2DPC configurations<br>x4/x8 support<br>ECC, SDDC, DDDC |
| PCIe | 40 lanes of PCIe Gen4.0 16x |
| Integrated I/O | 8 lanes of ETH, Combo MACs, supporting 2 x 100GE, 2 x 40GE, 8 x 25GE/10GE,10 x GE, supporting SR-IOV<br>RoCEv2/RoCEv1<br>x4 USB<br>3.0 x8 SAS 3.0 x2 SATA 3.0 |
| Crypto engine | AES, DES/3DES, MD5, SHA1, SHA2, HMAC, CMAC<br>Up to 100 Gbit/s |
| Compression | GZIP, LZS, LZ4<br>Up to 40 Gbit/s (compress)/100 Gbit/s (decompression) |
| RAID | RAID5/6, DIF, XOR, PQ acceleration |
| CCIX | Cache coherency interface for accelerator, like Xilinx FPGA<br>World's 1st CCIX solution |
| Scale-up | Coherent SMP interface for 2P/4P<br>3*240Gbps bandwidth |
| Power | TDP ~150 W (48C 2.6 GHz) |

# The Huawei Kunpeng 920: a multi-chip 48-64 cores

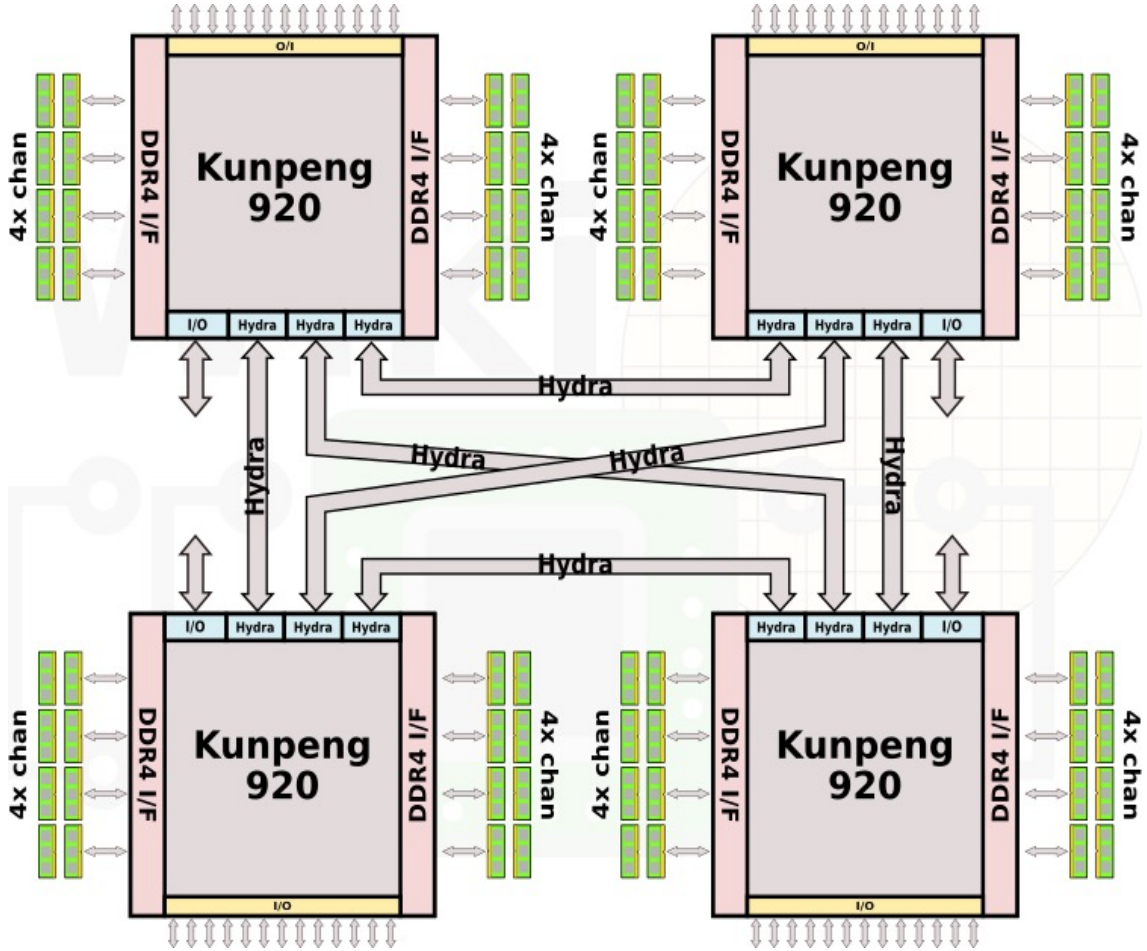



**SCCL: Super CPU Cluster**

**SICL: Super IO Cluster**

**CCL: CPU Clusters**

# The Huawei Kunpeng 920: multi-socket support




**Next-Gen Kunpeng 930:**

- higher-performance core
- SMT support
- Arm SVE (vector comp)
- expected in 2021

Kunpeng 930

gone?...

1. Marvell ThunderX product family **(dead?...)**
2. Fujitsu A64FX Arm chip **(in MACC…)**
3. Neoverse N1 hyperscale reference design **(…)**
4. Ampere Altra Arm Processor **(…)**
5. Amazon Graviton **(…)**
6. Huawei HiSilicon Kunpeng 920 **(…)**

*(2019)*

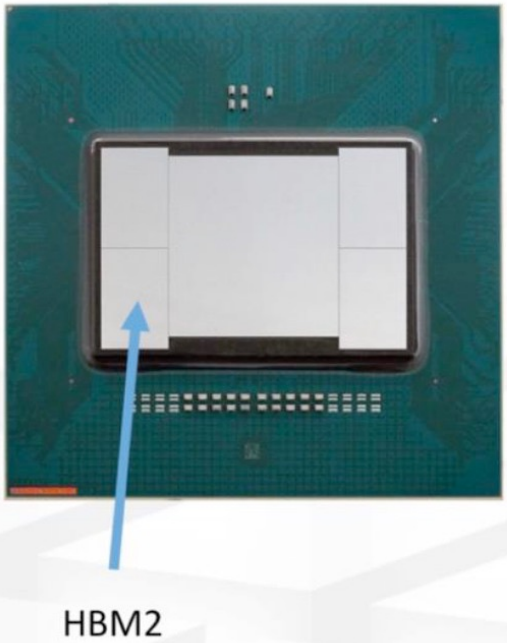
# *Current ARMv8 & v9 HPCs: server-level competitors*

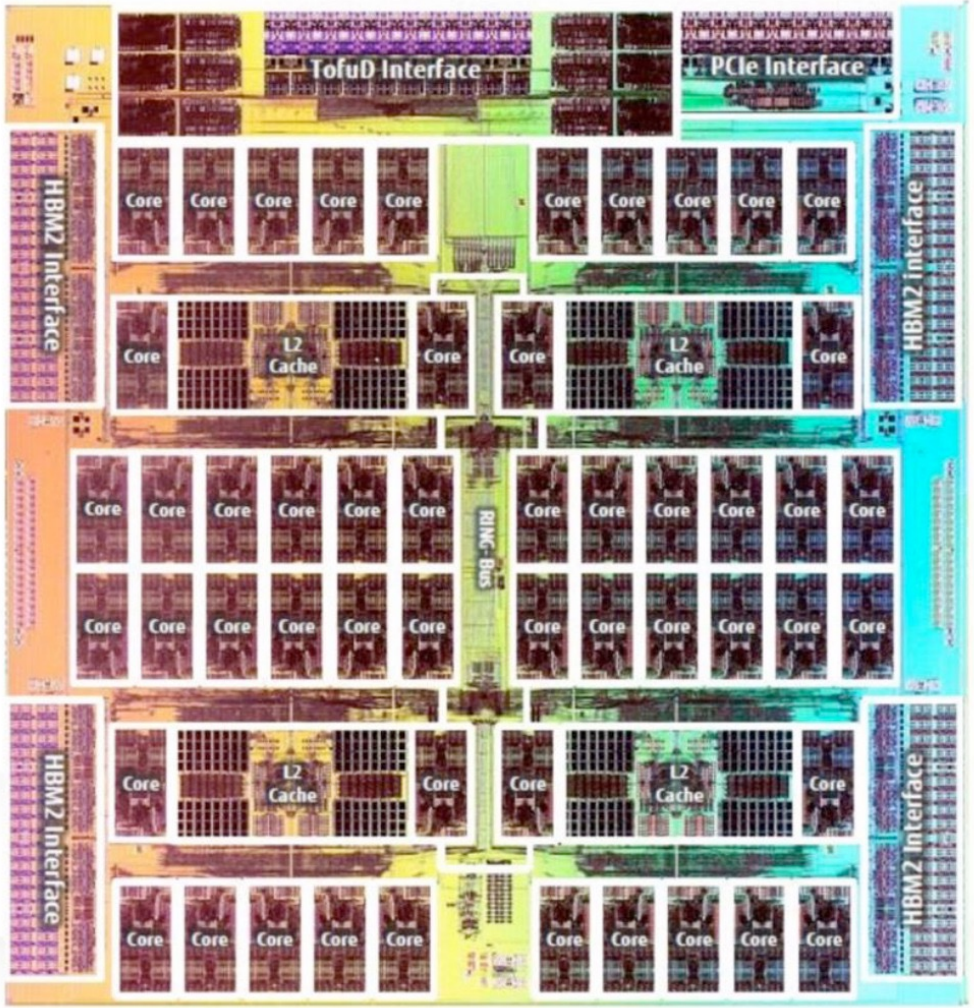
*(2022)*


# Fujitsu's A64FX ARM Chip

*(#1 in June'20 TOP500)*

- TSMC 7nm FinFET
- CoWoS technologies for HBM2



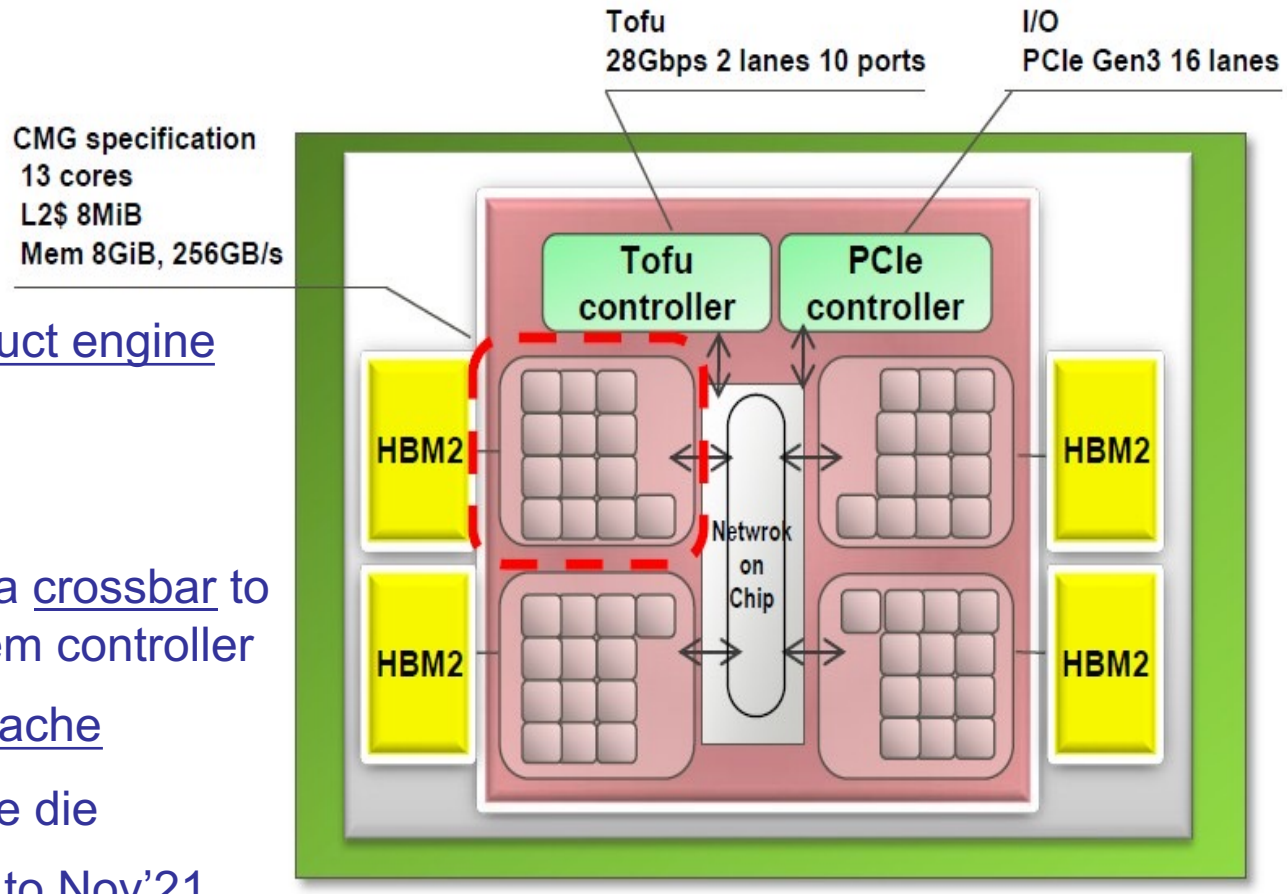



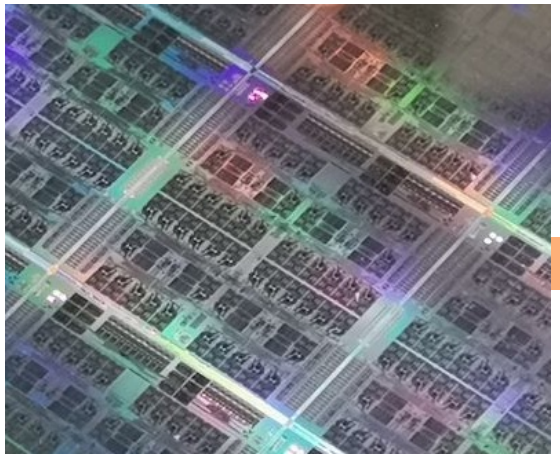

https://www.scientific-computing.com/sites/default/files/content/white-paper/pdfs/EXDCI-2-White%20paper%20Future%20HPC%20technologies.pdf /

# Fujitsu's A64FX Arm Chip: 48+4 cores

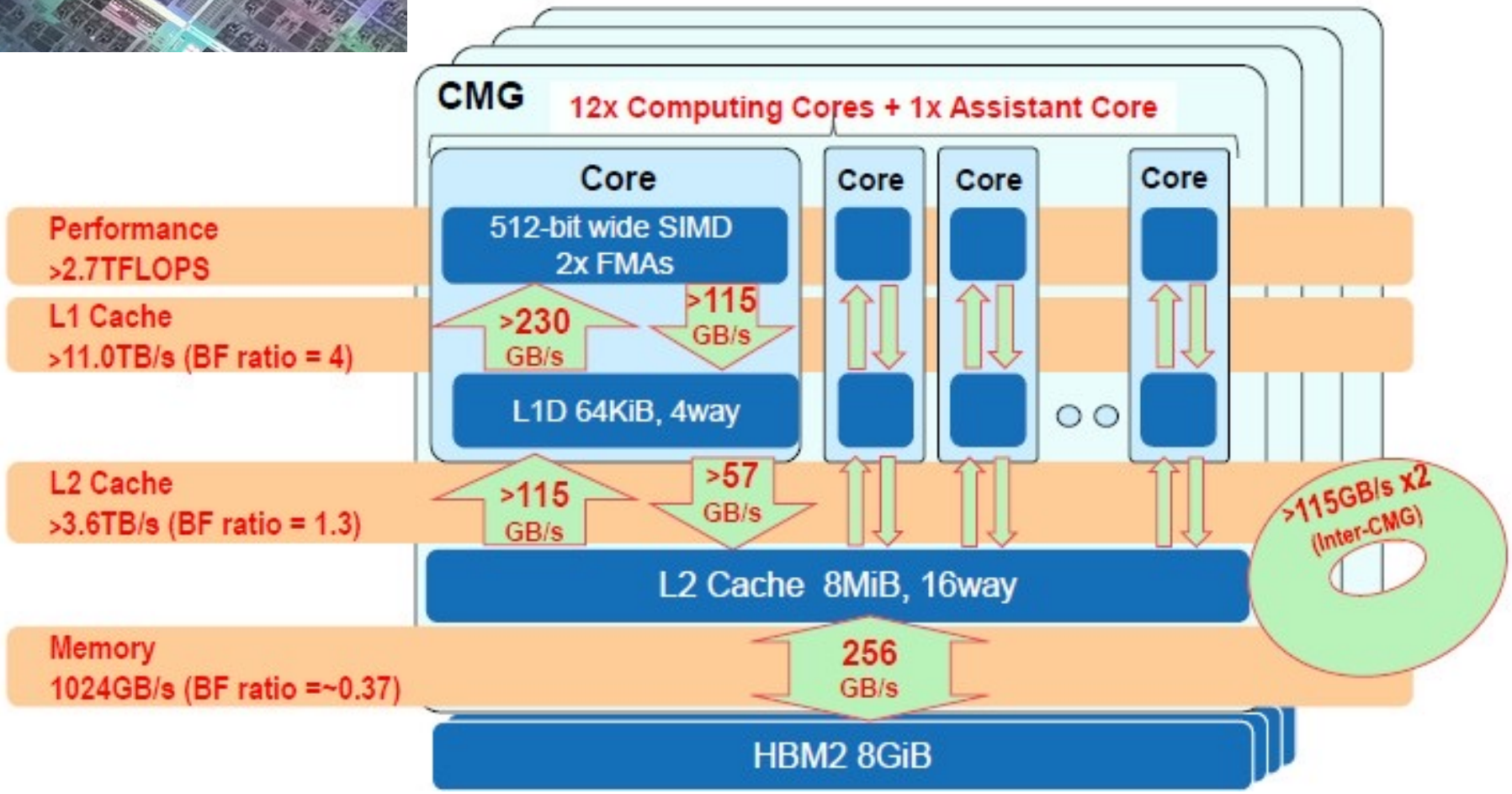
## A64FX Arm

- Feature size: 7 nm
- Armv8.2-A spec with 512-bit SVE extensions
- HP math and a dot-product engine
- 4 core-memory groups
- NoC: a double ring bus
- cores in CMG linked by a crossbar to L2 cache & to HBM2 mem controller
- 8 MiB L2 cache; no L3 cache
- a Tofu-D controller on the die
- #1 TOP500 from Jun'20 to Nov'21 uses A64FX package





# Block diagram of the A64FX chip



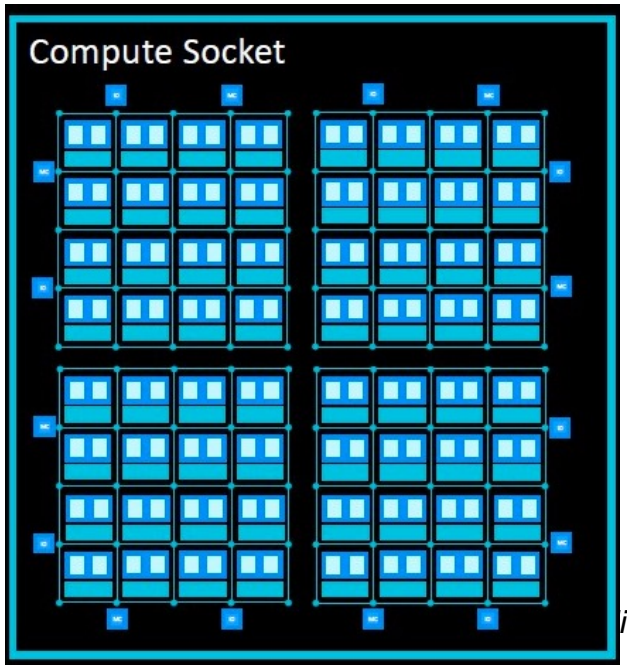

# Arm Neoverse N1

*(announced Feb'19)*

Hyperscale Datacenter

150W and beyond

64-128 core




arm Neoverse® N1 CPU

7 nm process
Based on Cortex A76

CoreSight™ multicore debug and trace

Core 1

Neoverse N1
Armv8-A (v8.2)
32/64-bit CPU

NEON™ AdvSIMD

Crypto

64KB I-Cache Parity

64KB D-Cache ECC

512kB/1MB Private L2 ECC

2

Direct-Connect to CMN-600 Mesh CHI

1x 256-bit AMBA® 5 CHI Direct-Connect

https://community.arm.com/developer/ip-products/processors/b/processors-ip-blog/posts/arm-neoverse-n1-platform-accelerating-the-transformation-to-a-scalable-cloud-to-edge-infrastructure


# Arm Neoverse N1

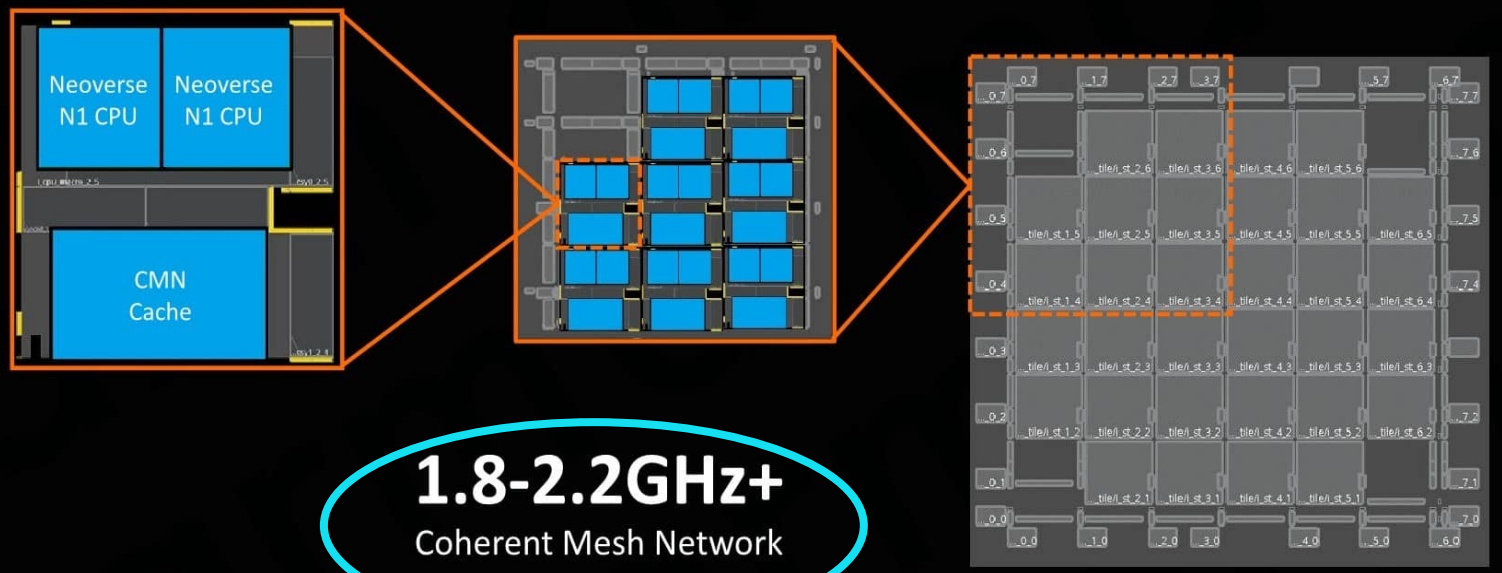
arm NEOVERSE

## Building hyperscale compute in 7nm

CPU Tile

Super Tile

Top-level Mesh



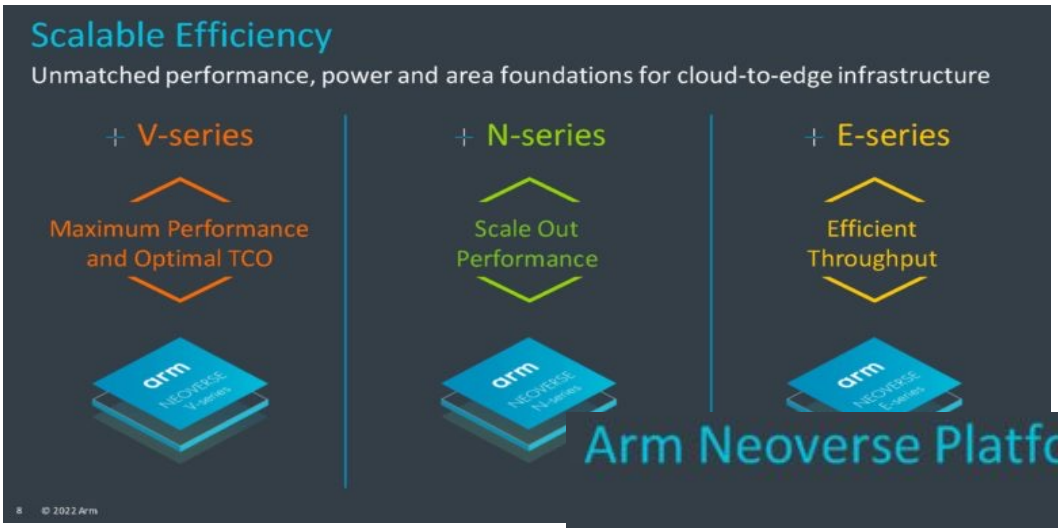

**1.8-2.2GHz+**

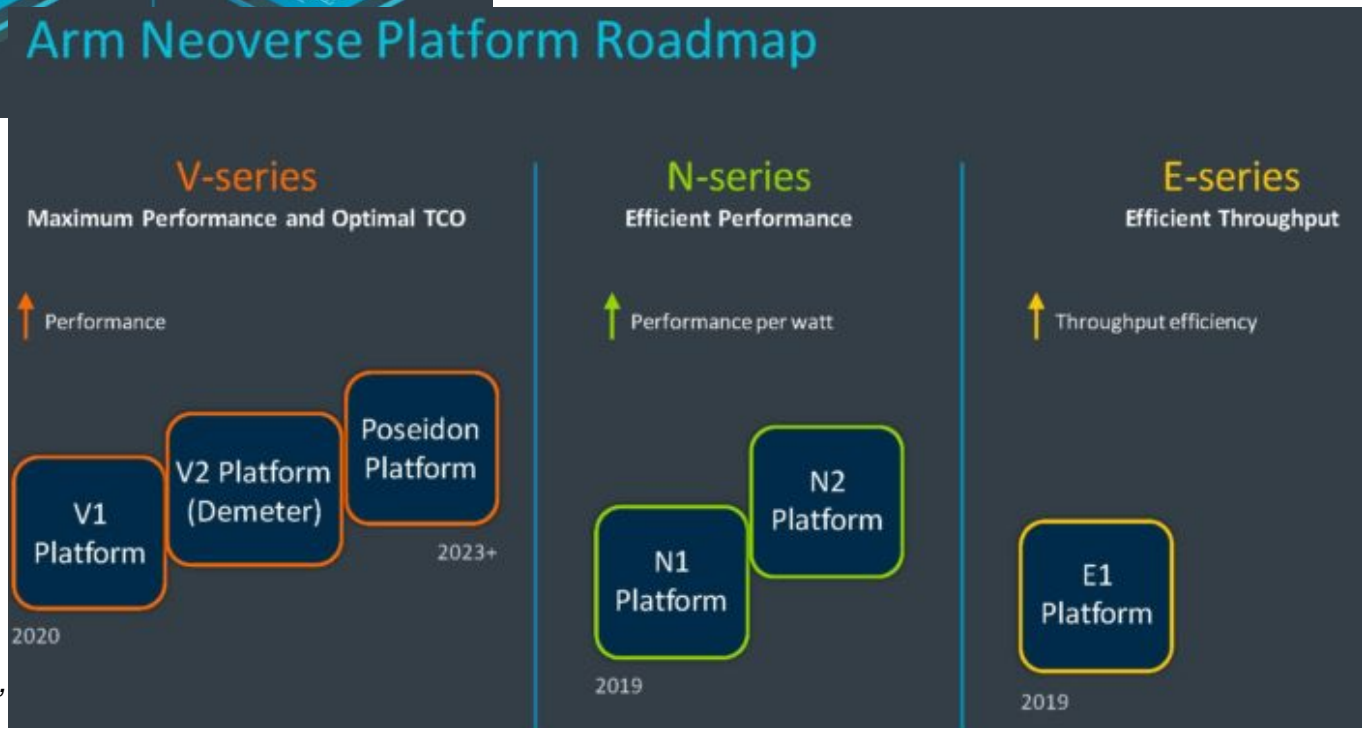
Coherent Mesh Network

CMN

# Arm Neoverse platform roadmap

https://www.servethehome.com/arm-neoverse-v2-cores-launched-for-nvidia-grace-and-cxl-2-0-pcie-gen5-cpus/

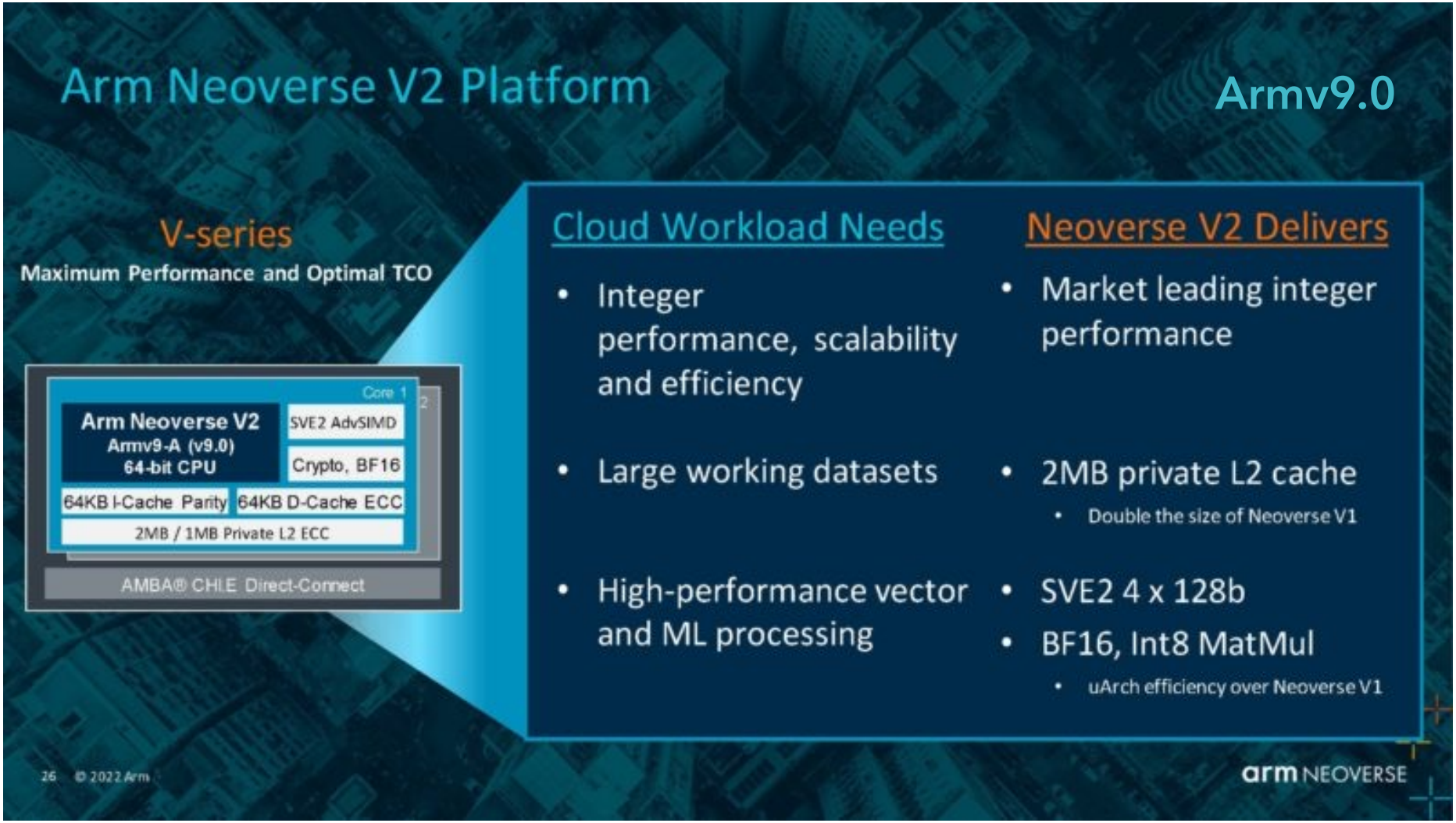



AJProença, Parallel Computing,

# Arm Neoverse V2 platform

https://www.servethehome.com/arm-neoverse-v2-cores-launched-for-nvidia-grace-and-cxl-2-0-pcie-gen5-cpus/




# Ampere Altra family: a SoC based on Neoverse N1

*(announced Mar'20)*

## Ampere™ Altra™ processor complex

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

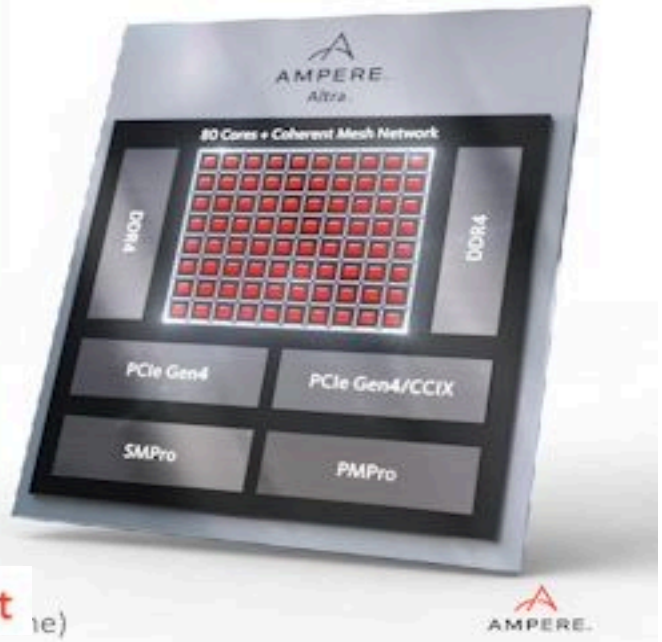
Arm v8.2+ features

Large Cache, all with ECC Protection
- 64 KB L1 I/D cache per core
- 1 MB L2 cache per core
- 32 MB system level cache

2x 128 SIMD Units

int8 and fp16 for ML inference performance

8 72-bit DDR4-3200 channels exceeding 200 GB/s per socket



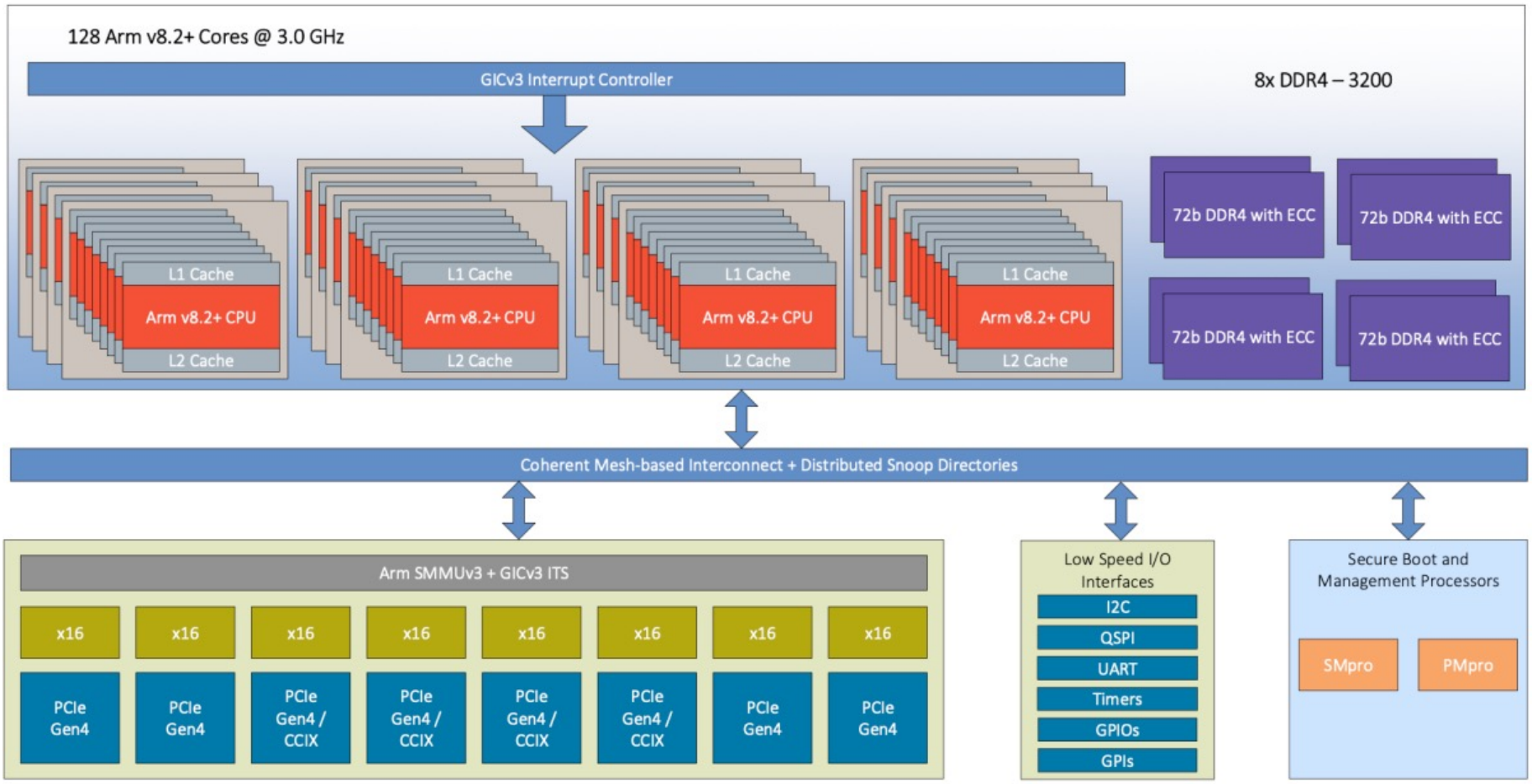

Ampere® Altra™ Max
7nm
- Up to 128 Cores

https://www.anandtech.com/show/15575/amperes-altra-80-core-n1-soc-for-hyperscalers-against-rome-and-xeon


# The Ampere Altra Max

*(launched in 2021)*

## Altra Max Block Diagram




https://www.tomshardware.com/news/ampere-altra-max-128-core-priced

# Amazon Web Services (AWS): Graviton2 with Arm Neoverse N1 cores

## AWS Graviton2

- 7 nm process
- up to 64 Arm Neoverse N1 cores
- 1 MiB L2 per core
- shared 32 MiB L3
- without SMT
- single-socket

## AWS Designing a 32-Core Arm Neoverse N1 CPU for Cloud Servers

by Anton Shilov on December 2, 2019 1:00 PM EST

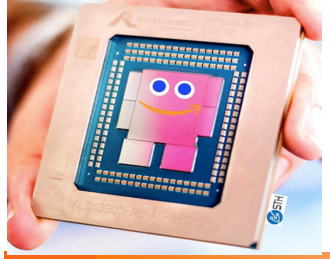
Posted in Servers | CPUs | Arm | Amazon | AWS | Neoverse N1


**Graviton1 Processor**

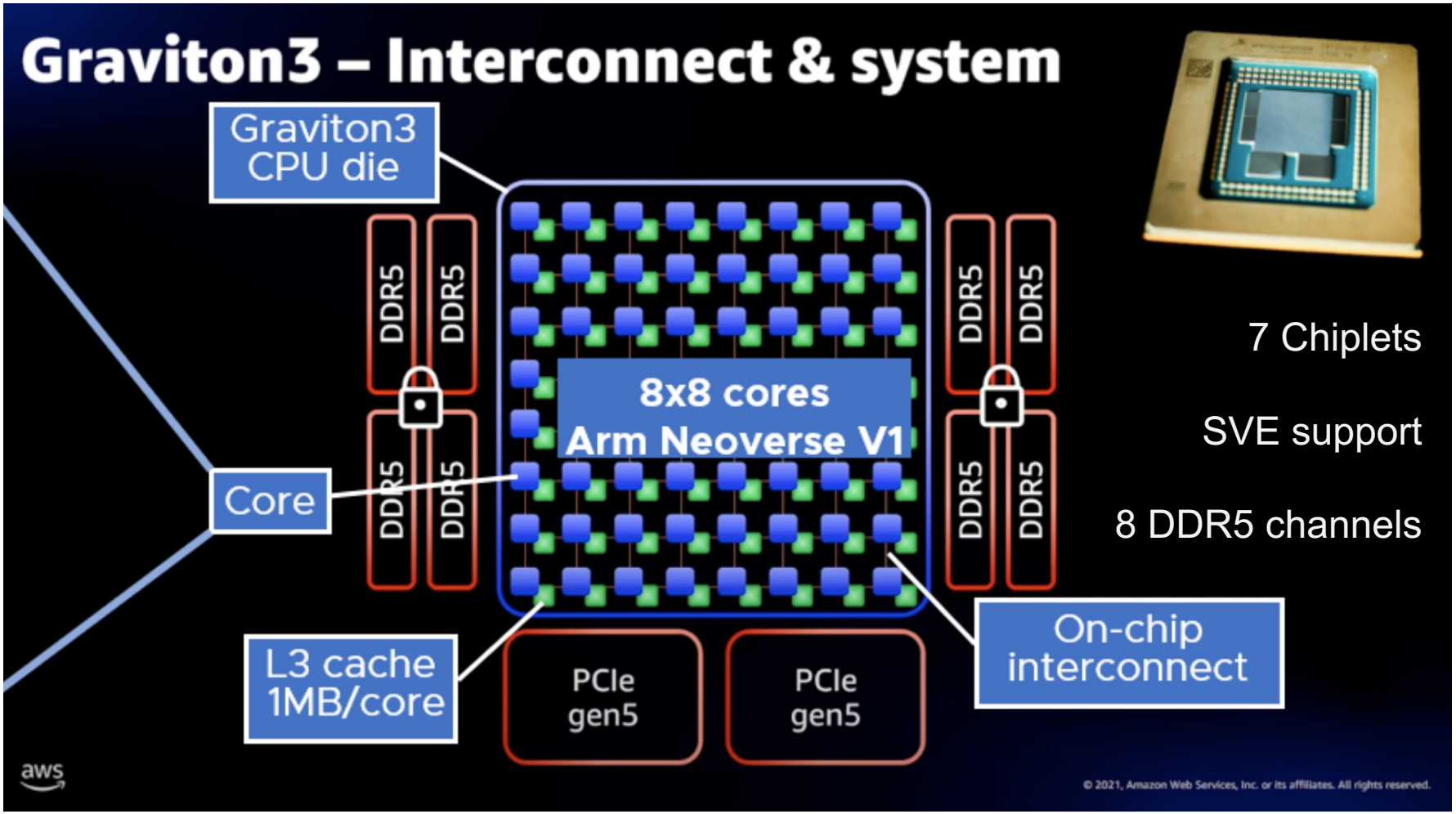
- First Arm-based processor in major cloud
- Built on 64-bit ARM Neoverse cores with AWS-designed 16 nm silicon
- Up to 16 vCPUs,10 Gbps enhanced networking, 3.5 Gbps EBS bandwidth

**Graviton2 Processor**

- Built with 64-bit Arm Neoverse cores with AWS-designed 7 nm silicon
- Up to 64 vCPUs, 25 Gbps enhanced networking, 18 Gbps EBS bandwidth
- **7x performance, 4x compute cores, 5x faster memory**



# AWS Graviton3 with Neoverse V1 cores




https://www.servethehome.com/amazon-aws-graviton3-hits-ga-with-3-sockets-per-motherboard-designs-tri-socket-arm/


# Alibaba Yitian 710

Alibaba Group | APSARA CONFERENCE

INVENT·EXPLORE·INSPIRE


## Yitian

**Born for Cloud**

Alibaba's first general-purpose chip: Yitian 710, compatible for different application scenarios

- ARMv9 Architecture
- 128 Arm cores
- DDR5 Integral
- PCIE5.0

YITIAN 710

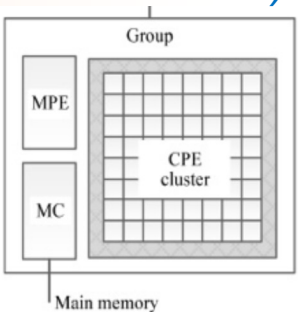
# *Manycore chips/packages: an overview*

Key server chips/packages that addresses those issues:

- Intel: from Intel MIC to the Xeon Scalable family
- AMD: the Epyc Zen family
- ARM: key ARMv8 & v9 server-level competitors
  - Marvell ThunderX family
  - Fujitsu A64FX Arm chip
  - Neoverse hyperscale reference design for
    - Ampere Altra Arm
    - Amazon Graviton
  - Alibaba Yitian 710
  - Huawei HiSilicon Kunpeng 920
- **Sunway: the SX260x0 family**
- Cerebras: a Wafer Scale Engine
- Apple *(not server…)*: the SoC approach *(no chiplets!)*

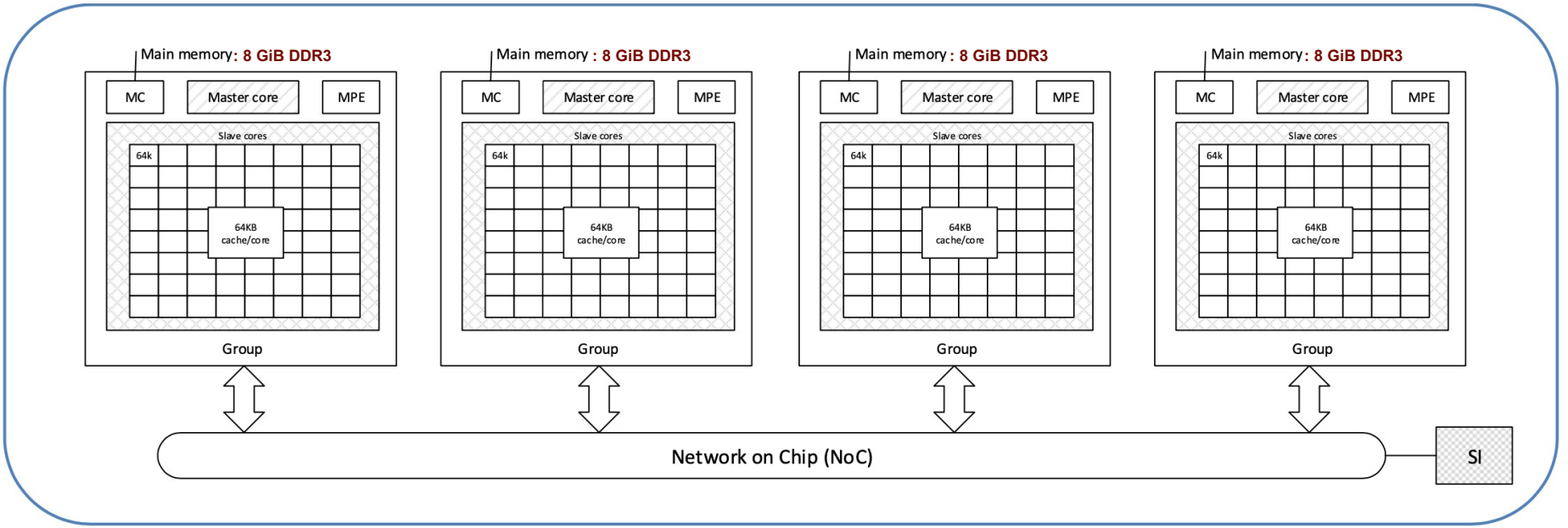

# The Shen Wei SW 26010 in SunWay TaihuLight

*(#1 in June'16 TOP500)*




## Shen Wei SW 26010 *(260 cores)*:

- 4 core groups (CG, as a SNC), connected via a NoC
- each CG has a Management Processing Element (MPE) a 8x8 mesh of 64 Computing Processing Elements (CPE Cluster) and a Memory Controller (MC)
- a CPE is a 64-bit RISC OoO *(out-of-order)* core w/ a 256-bit vector unit, no SMT, 16 KiB L1 instruction cache, and 64 KiB Scratch Pad Memory (no L2 cache)



# *Manycore chips/packages: an overview*

Key server chips/packages that addresses those issues:

- Intel: from Intel MIC to the Xeon Scalable family
- AMD: the Epyc Zen family
- ARM: key ARMv8 & v9 server-level competitors
  - Marvell ThunderX family
  - Fujitsu A64FX Arm chip
  - Neoverse hyperscale reference design for
    - Ampere Altra Arm
    - Amazon Graviton
  - Alibaba Yitian 710
  - Huawei HiSilicon Kunpeng 920
- Sunway: the SX260x0 family
- **Cerebras: a Wafer Scale Engine**
- Apple *(not server)*: the SoC approach *(no chiplets!)*

# Cerebras: a Wafer Scale Engine (WSE) (Aug'19)

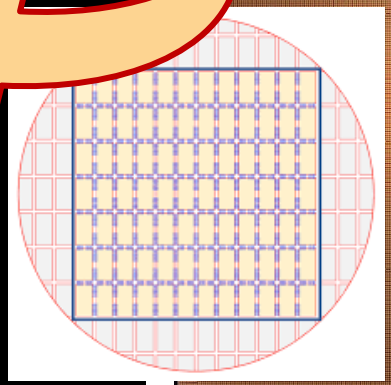
**Cerebras** Wafer Scale Engine (WSE):
the largest chip ever built

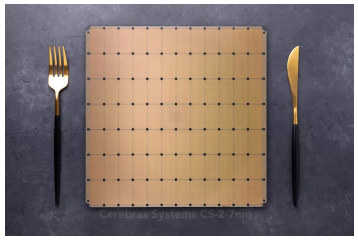
**46,225 mm² chip**
56x larger than the biggest GPU ever made

**400,000 core**
78x more cores

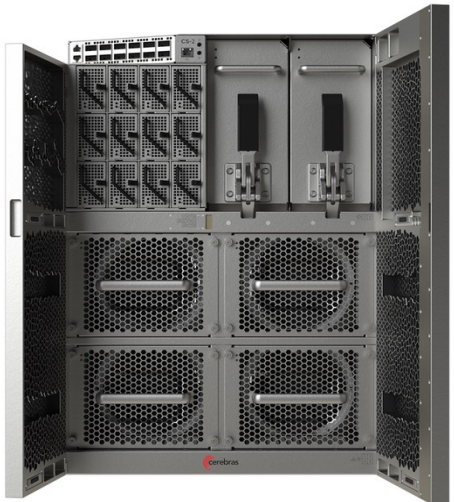
**18 GB on-chip SRAM**
3000x more on-chip memory

**100 Pb/s interconnect**
33,000x more bandwidth





# Cerebras:
# 2nd Gen Wafer Scale Engine (WSE2) (Q3 2021)

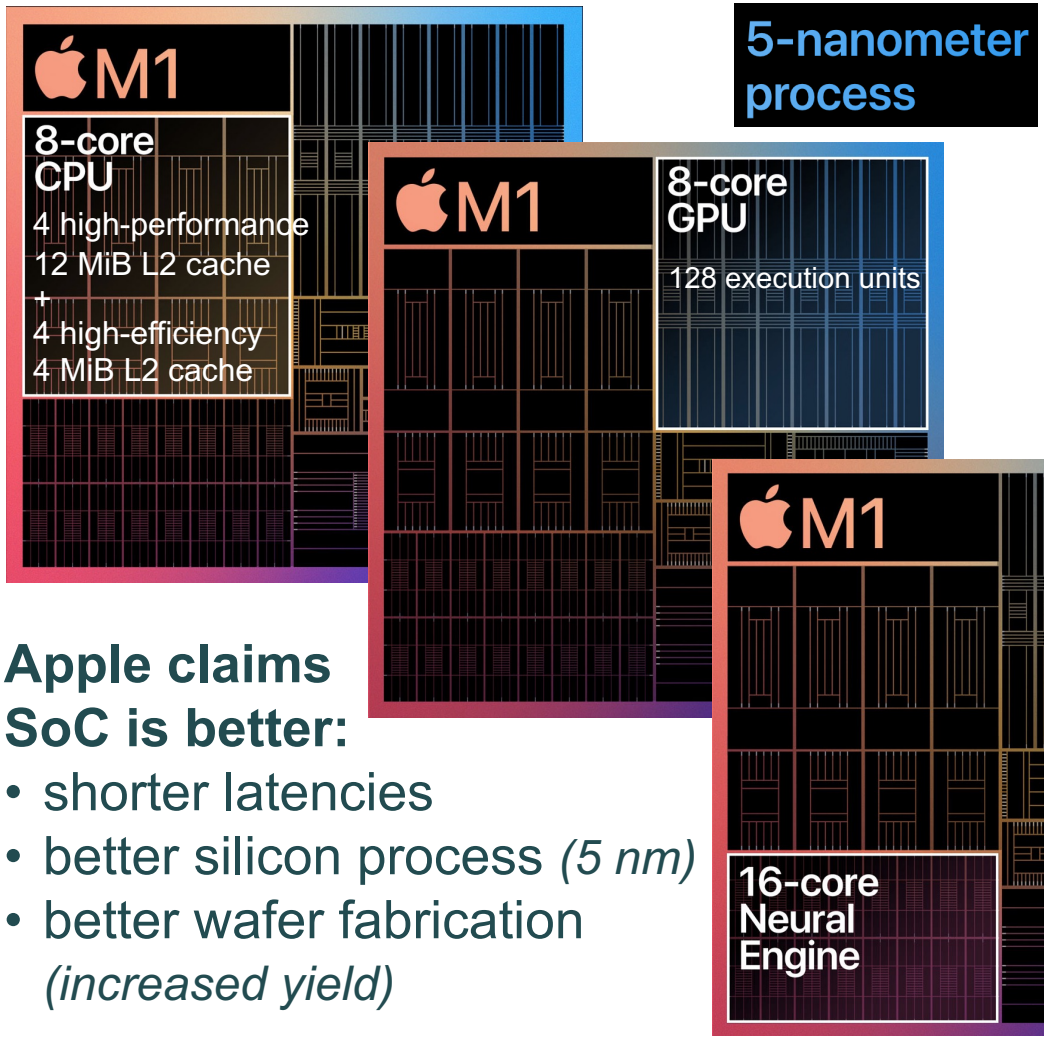
| | Cerebras Wafer Scale Engine 2 | Cerebras Wafer Scale Engine | Nvidia A100 |
|---|---|---|---|
| Process Node | TSMC 7nm | TSMC 16nm | TSMC 7nm N7 |
| AI Cores | 850,000 | 400,000 | 6,912 + 432 |
| Die Size | 46,255 mm2 | 46,255 mm2 | 826 mm2 |
| Transistors | 2.6 Trillion | 1.2 Trillion | 54 Billion |
| On-Chip SRAM Memory | 40 GB | 18 GB | 40 MB |
| Memory Bandwidth | 20 PB/s | 9 PB/s | 1,555 GB/s |
| Fabric Bandwidth | 220 Pb/s | 100 Pb/s | 600 GB/s |
| Power Consumption (System/Chip) | 20kW / 15kW | 20kW / 15kW | 250W (PCIe) / 400W (SXM) |

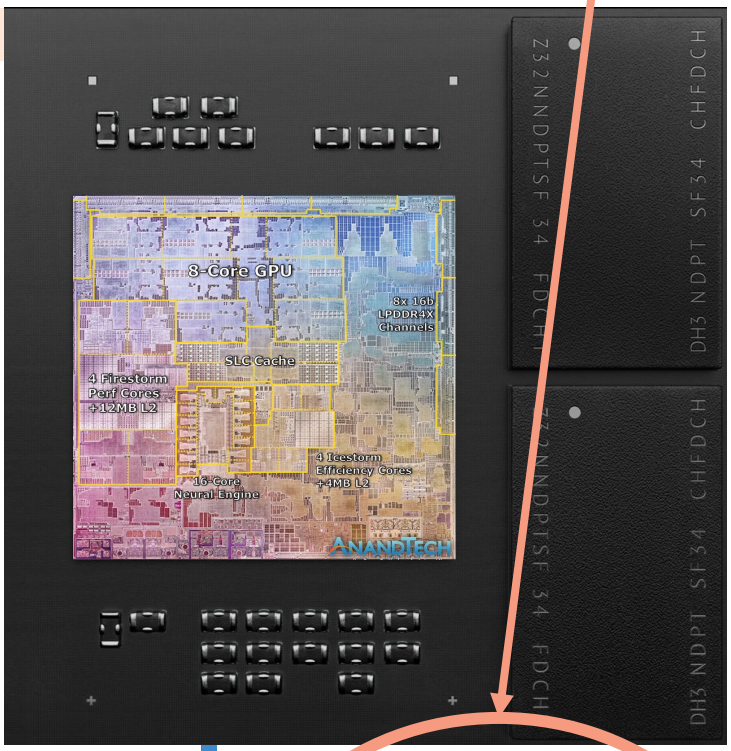
# *Manycore chips/packages: an overview*

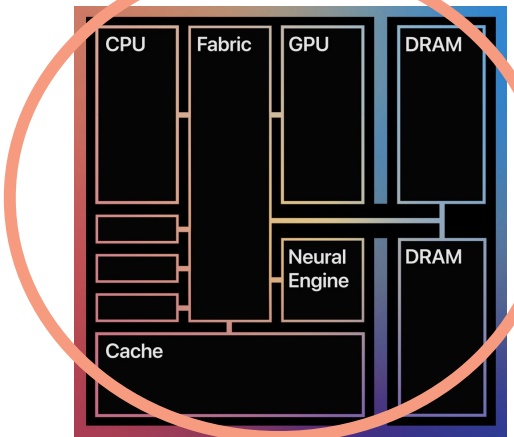
Key server chips/packages that addresses those issues:

- Intel: from Intel MIC to the Xeon Scalable family
- AMD: the Epyc Zen family
- ARM: key ARMv8 & v9 server-level competitors
  - Marvell ThunderX family
  - Fujitsu A64FX Arm chip
  - Neoverse hyperscale reference design for
    - Ampere Altra Arm
    - Amazon Graviton
  - Alibaba Yitian 710
  - Huawei HiSilicon Kunpeng 920
- Sunway: the SX260x0 family
- Cerebras: a Wafer Scale Engine
- **Apple** *(not server)***: the SoC approach** *(no chiplets!)*

# Apple M1 SoC
*(announced Nov'20)*

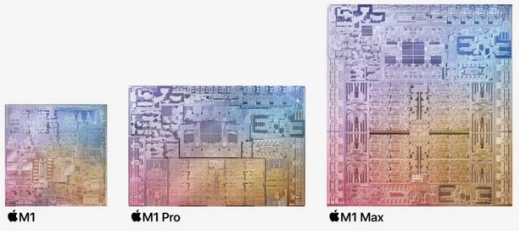



**Apple claims SoC is better:**

- shorter latencies
- better silicon process *(5 nm)*
- better wafer fabrication *(increased yield)*



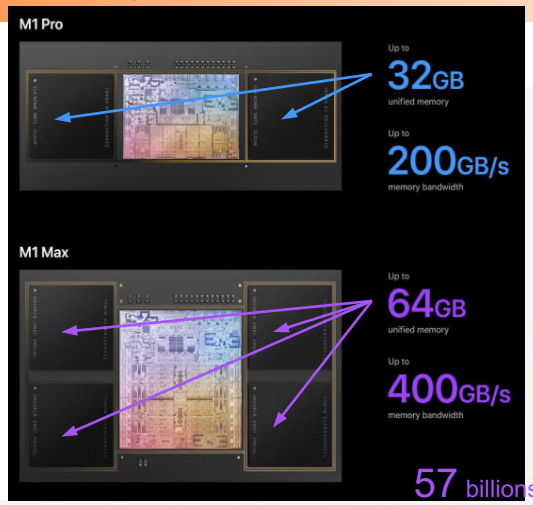

# *M1 successors: M1 Pro & M1 Max*

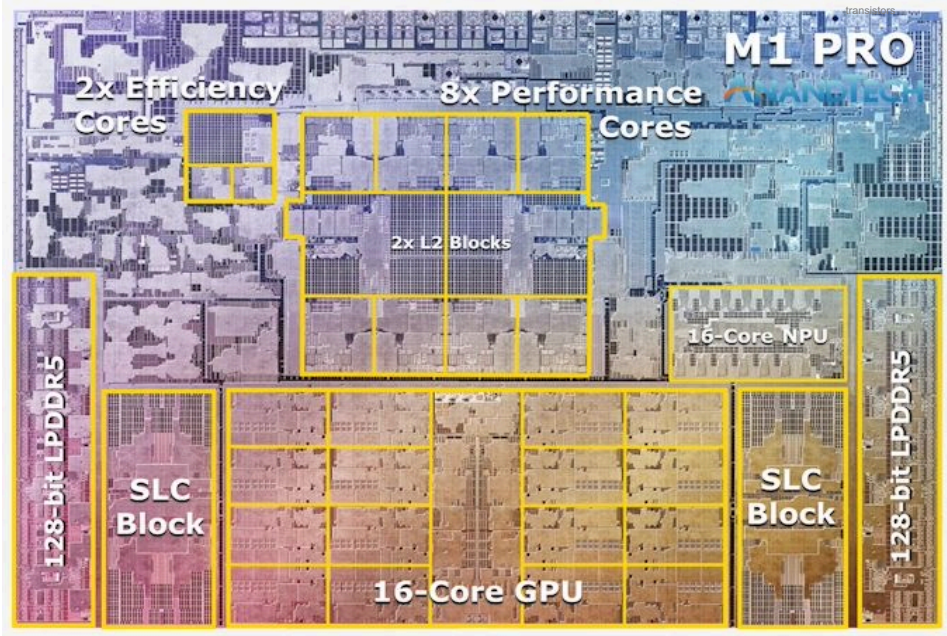
*(announced Oct'21)*

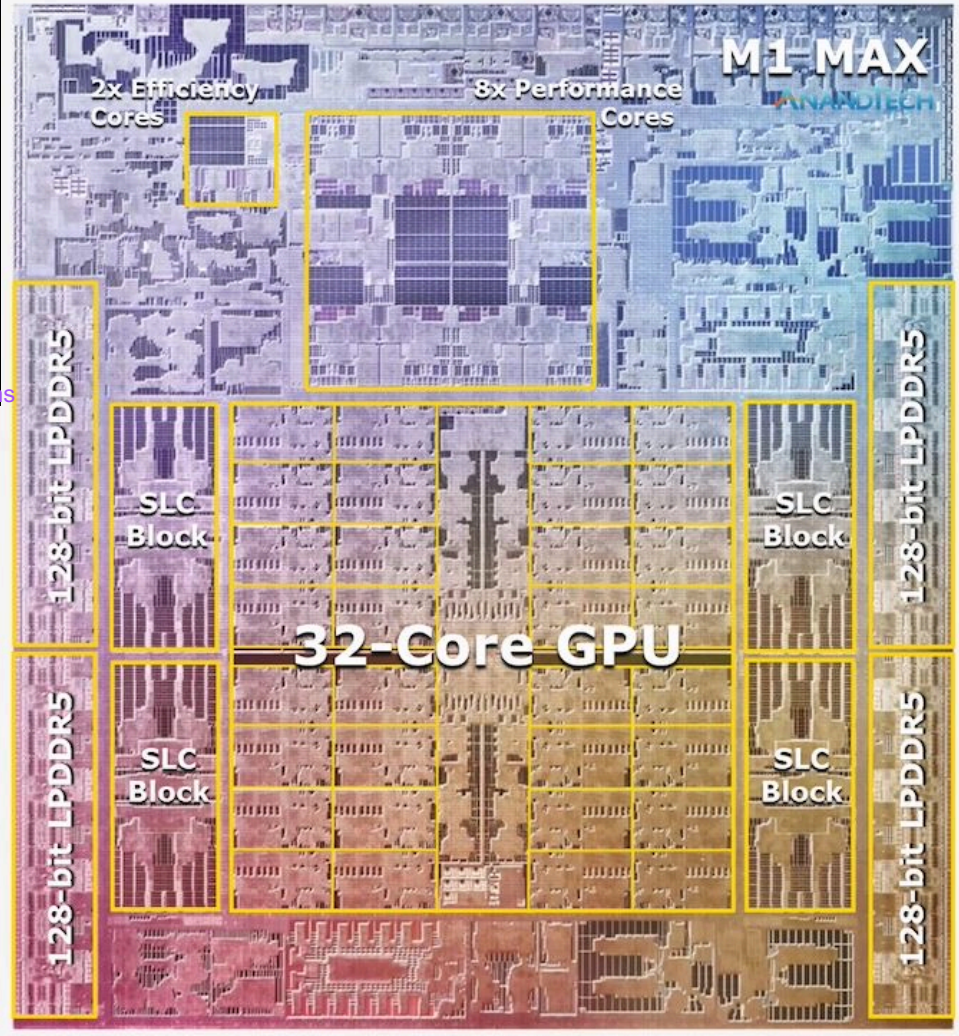
*https://www.anandtech.com/show/17019/apple-announced-m1-pro-m1-max-giant-new-socs-with-allout-performance*

## Mi Pro
256-bit LPDDR5-6400
⇒ 32B * 6.4GHz
≈ 200 GB/s

## Mi Max
512-bit LPDDR5-6400
⇒ 64B * 6.4GHz
≈ 400 GB/s



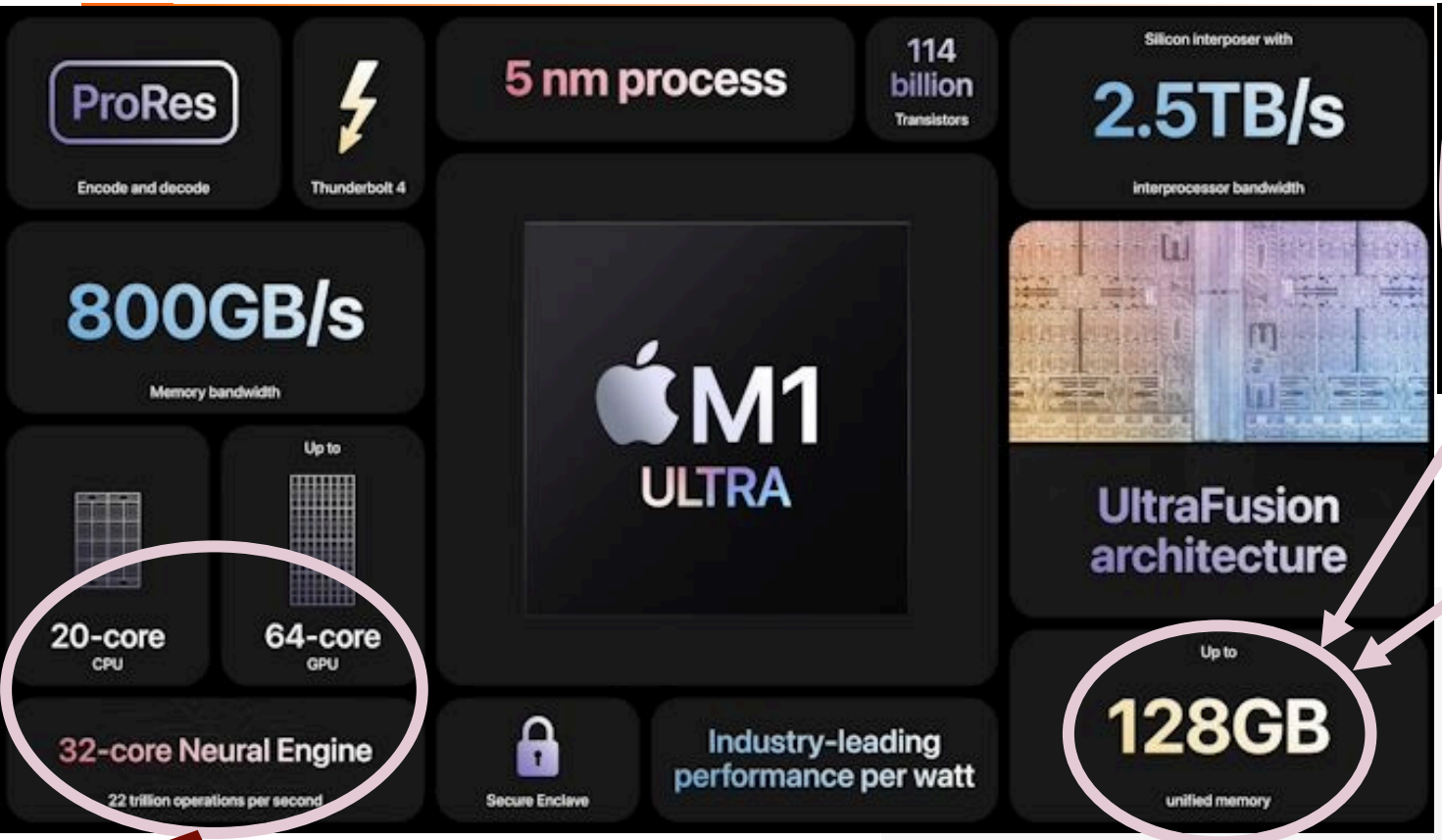



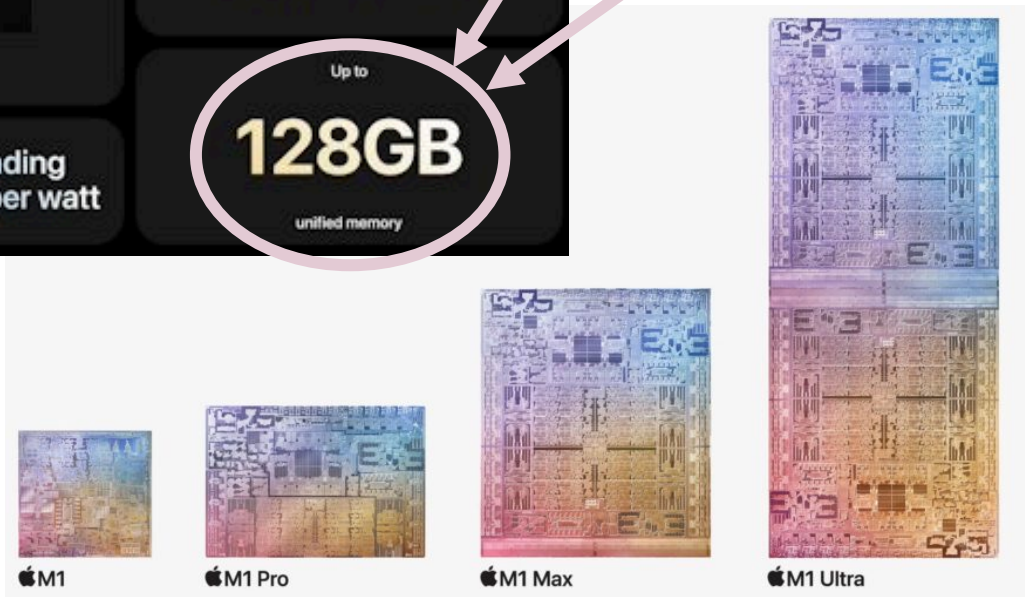
# *M1 successors: M1 Ultra*

*(announced Mar'22)*

*https://www.anandtech.com/show/17306/apple-announces-m1-ultra-combining-two-m1-maxes-for-even-more-performance*

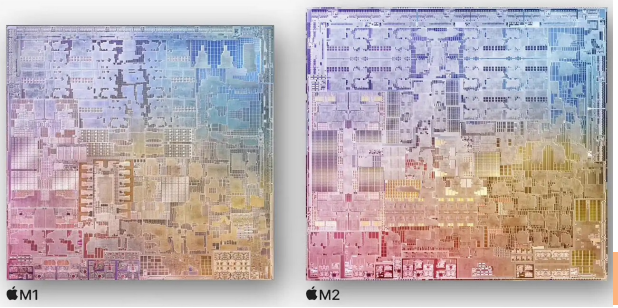



**M1 Ultra: 2x M1 Max**



# *M1 successors: a faster M2*

*(announced Jun'22)*




**Up to 16 GiB LPDDR4 in M1**

Up to 24GB LPDDR5 memory

High-performance media engine

40% Faster Neural Engine

Up to 15.8 trillion operations per second

16-core Neural Engine

Second-generation 5 nm technology

M2

8-core CPU

Up to 10-core GPU

**Same as M1**

**+2 than M1**

ProRes encode and decode

6K external display support

18% Faster CPU

35% Faster GPU

Over 20 billion transistors

Industry-leading performance per watt

50% More memory bandwidth

100GB/s Memory bandwidth

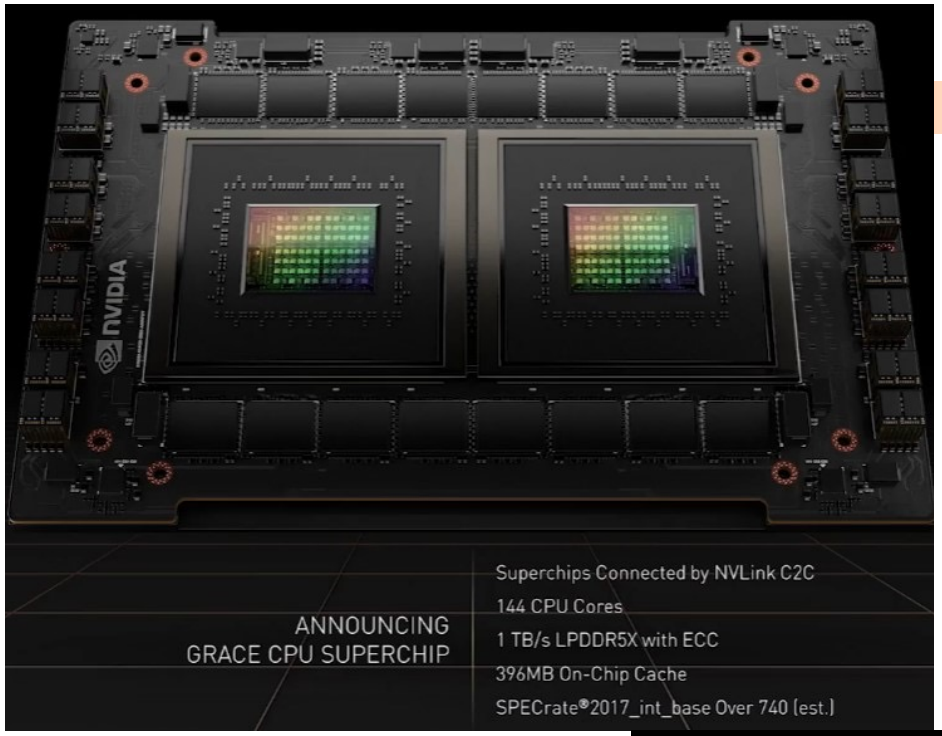
# And the M2 Extreme...

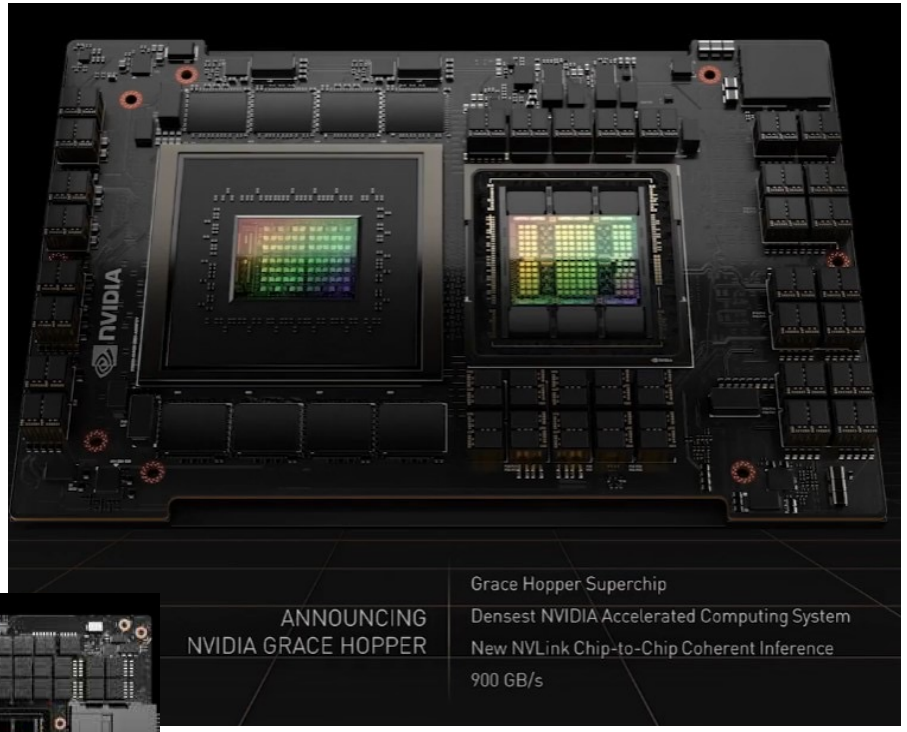
# *Manycore chips/packages: an overview*

Key server chips/packages that addresses those issues:

- Intel: from Intel MIC to the Xeon Scalable family
- AMD: the Epyc Zen family
- ARM: key ARMv8 & v9 server-level competitors
  - Marvell ThunderX family
  - Fujitsu A64FX Arm chip
  - Neoverse hyperscale reference design for
    - Ampere Altra Arm
    - Amazon Graviton
  - Alibaba Yitian 710
  - Huawei HiSilicon Kunpeng 920
- Sunway: the SX260x0 family
- Apple *(not server)*: the SoC approach *(no chiplets!)*
- Cerebras: a Wafer Scale Engine
- **NVidia** *(CPU+GPU)*: **the Grace Hopper Superchip**

# The NVidia Grace & Grace-Hopper Superchips




Grace superchip:

- 2x 72-core Arm v9.0
- 4 nm TSMC
- SVE2 suppport


https://www.servethehome.com/nvidia-gtc-2022-keynote-coverage-crazy-new-data-center-gear/


# The Turing Award

## Turing Award

From Wikipedia, the free encyclopedia

WIKIPEDIA
The Free Encyclopedia

The **ACM A. M. Turing Award** is an annual prize given by the Association for Computing Machinery (ACM) for contributions of lasting and major technical importance to computer science.[2] It is generally recognized as the highest distinction in computer science and is colloquially known as or often referred to as the "Nobel Prize of Computing".[3][4][5][6]

The award is named after Alan Turing, who was a British mathematician and reader in mathematics at the University of Manchester.

| | | |
|---|---|---|
| 2017 | John L. Hennessy |  |
| | David Patterson |  |

turing lecture

Nobel equivalent in Computer Science

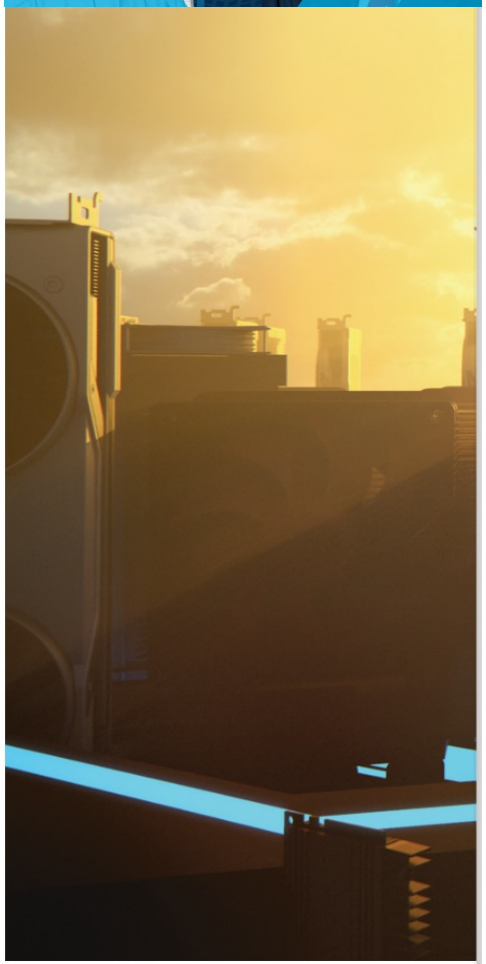

**And now?**

DOI:10.1145/3282307

**Innovations like domain-specific hardware, enhanced security, open instruction sets, and agile chip development will lead the way.**

BY JOHN L. HENNESSY AND DAVID A. PATTERSON

# A New Golden Age for Computer Architecture



WE BEGAN OUR Turing Lecture June 4, 2018[11] with a review of computer architecture since the 1960s. In addition to that review, here, we highlight current challenges and identify future opportunities, projecting another golden age for the field of computer architecture in the next decade, much like the 1980s when we did the research that led to our award, delivering gains in cost, energy, and security, as well as performance.

*"Those who cannot remember the past are condemned to repeat it."* —George Santayana, 1905

engineers, including ACM A.M. Turing Award laureate Fred Brooks, Jr., thought they could create a single ISA that would efficiently unify all four of these ISA bases.

They needed a technical solution