COMP 122

DR JEFF SOFTWARE
INDIE APP DEVELOPER
© Jeff Drobman
2016-2022

Rev. 3-9-22

Computer Org & ASSEMBLY Programming

ARM SoC

Qualcomm, Samsung, Google, Tesla

Dr Jeff Drobman

website ⇨ drjeffsoftware.com/classroom.html

email ⇨ jeffrey.drobman@csun.edu

Index

- ❖Phone CPU's
  - ❑ Apple Phones (separate slide set)
  - ❑ Qualcomm (Snapdragon) → slide 9
  - ❑ MediaTek → slide 19
  - ❑ Samsung (Exynos) → slide 23
  - ❑ Google → slide 40
- ❖ Others
  - ❑ Tesla → slide 54
  - ❑ Amazon → slide 85
  - ❑ SMIC China → slide 87

Section

Phones

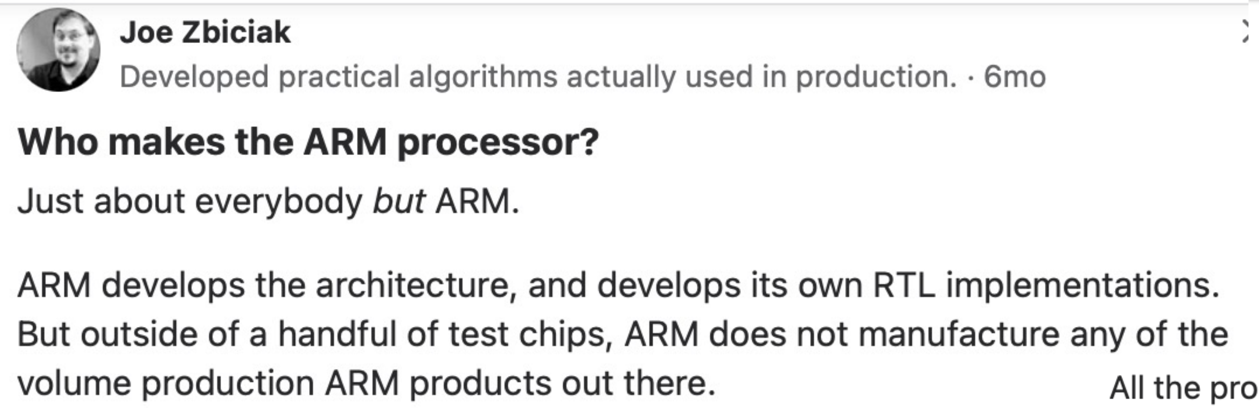
CSUN CALIFORNIA STATE UNIVERSITY NORTHRIDGE

COMP122

Quora

ARM Chips (SoC)

DR JEFF SOFTWARE INDIE APP DEVELOPER



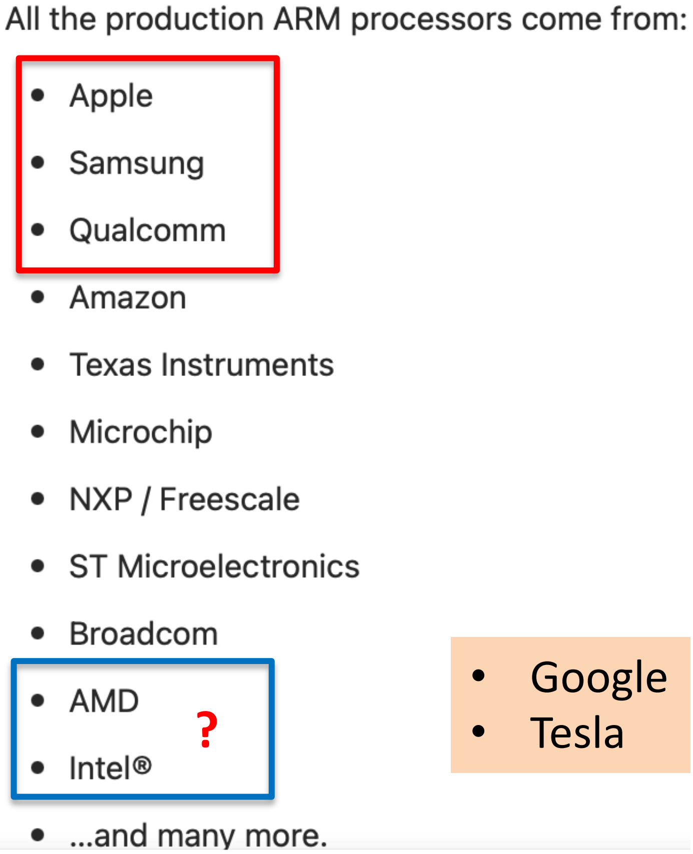
Joe Zbiciak

Developed practical algorithms actually used in production. · 6mo

Who makes the ARM processor?

Just about everybody *but* ARM.

ARM develops the architecture, and develops its own RTL implementations. But outside of a handful of test chips, ARM does not manufacture any of the volume production ARM products out there.

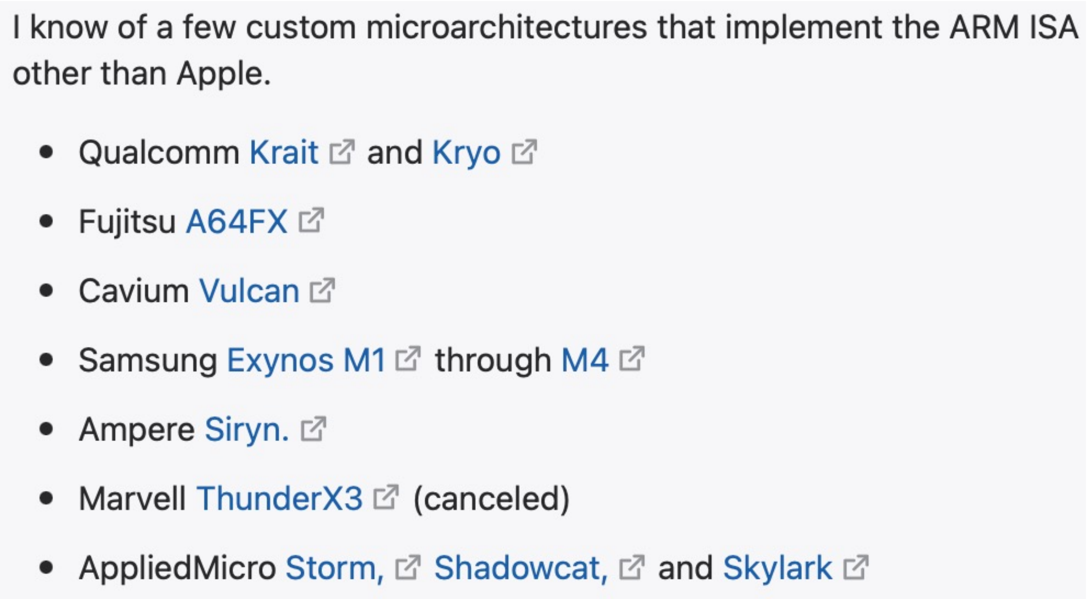


All the production ARM processors come from:

- Apple
- Samsung
- Qualcomm
- Amazon
- Texas Instruments
- Microchip
- NXP / Freescale
- ST Microelectronics
- Broadcom
- AMD
- Intel®
- ...and many more.

?

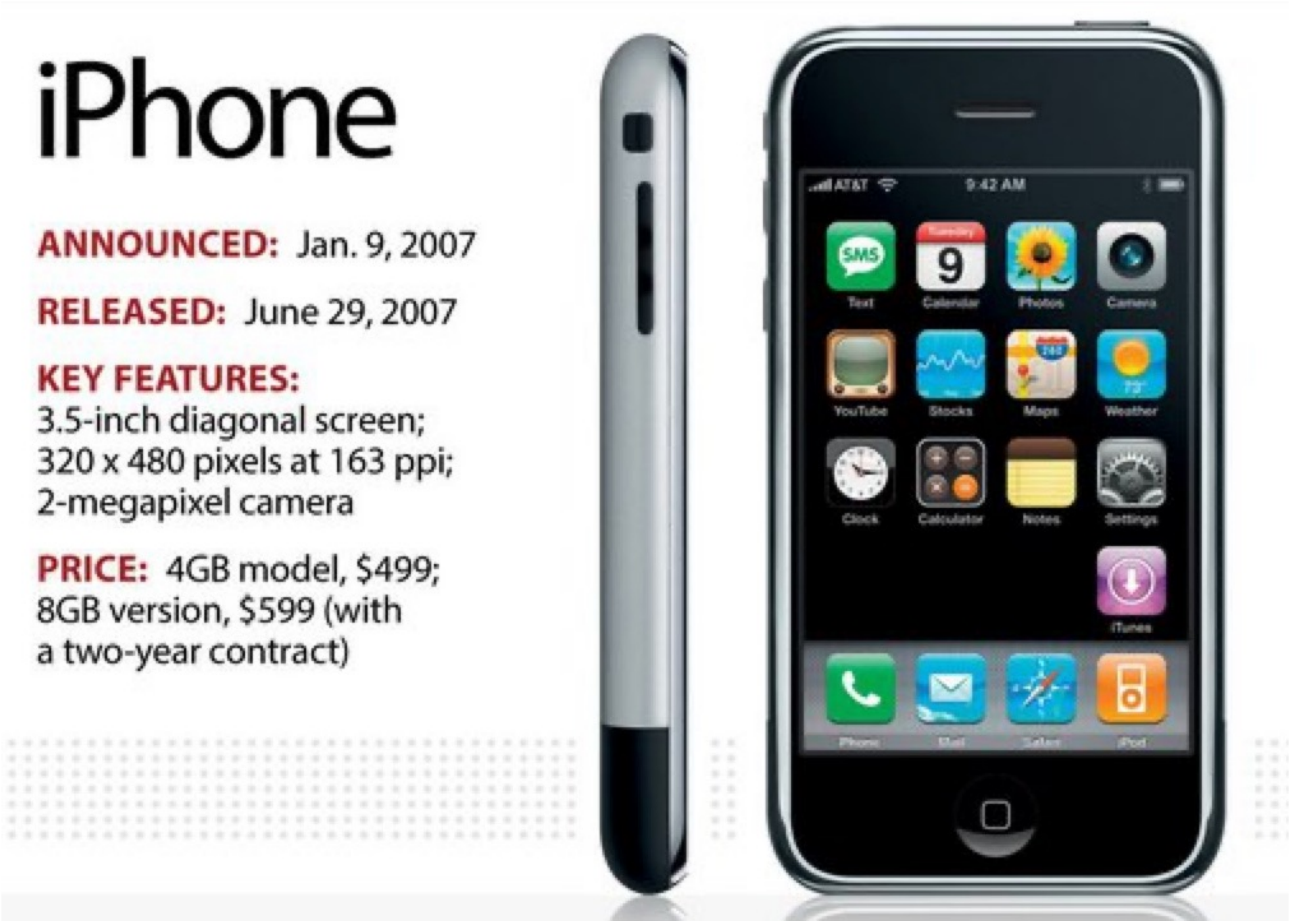
- Google
- Tesla



I know of a few custom microarchitectures that implement the ARM ISA other than Apple.

- Qualcomm Krait and Kryo
- Fujitsu A64FX
- Cavium Vulcan
- Samsung Exynos M1 through M4
- Ampere Siryn.
- Marvell ThunderX3 (canceled)
- AppliedMicro Storm, Shadowcat, and Skylark

1st iPhone



iPhone

ANNOUNCED: Jan. 9, 2007

RELEASED: June 29, 2007

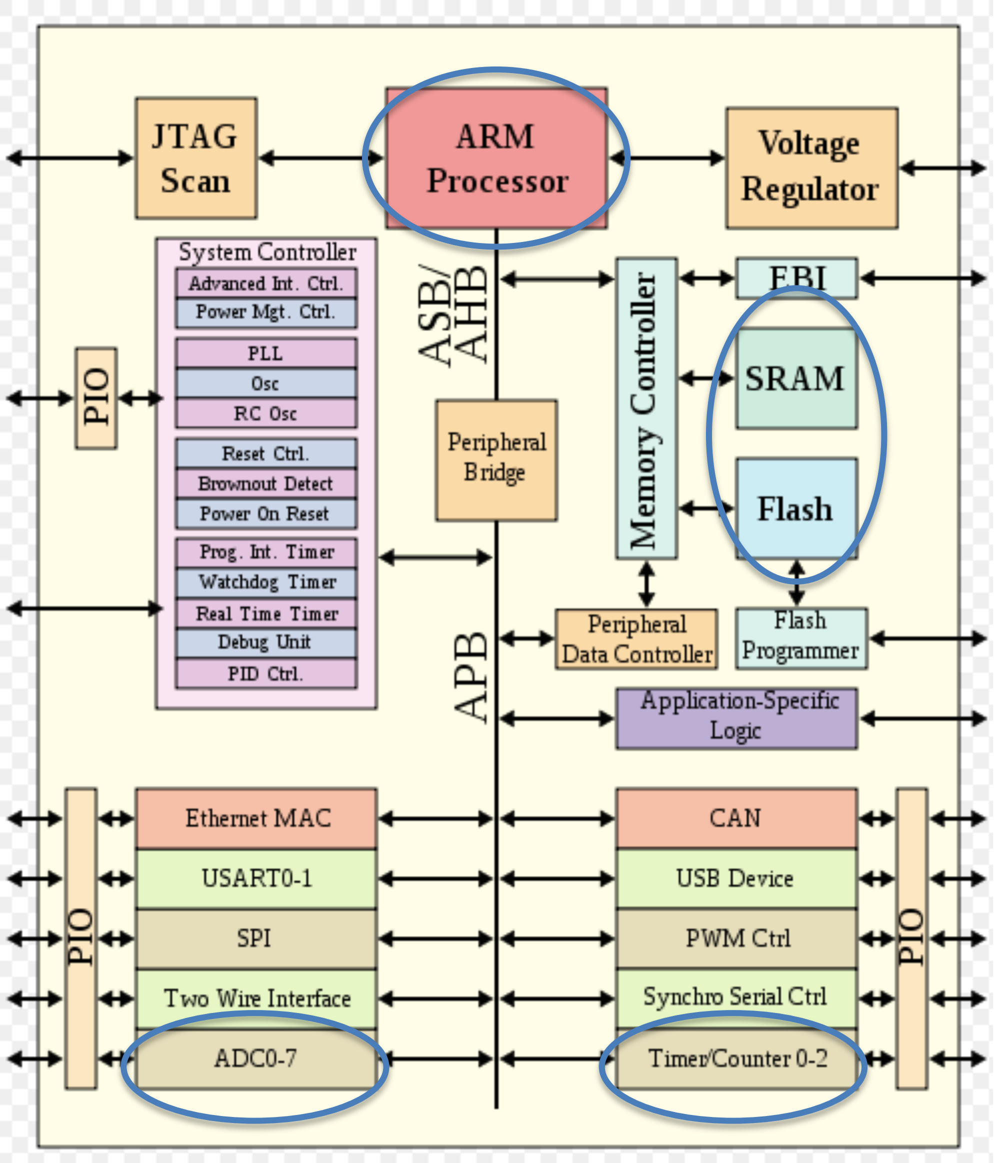
KEY FEATURES:
3.5-inch diagonal screen;
320 x 480 pixels at 163 ppi;
2-megapixel camera

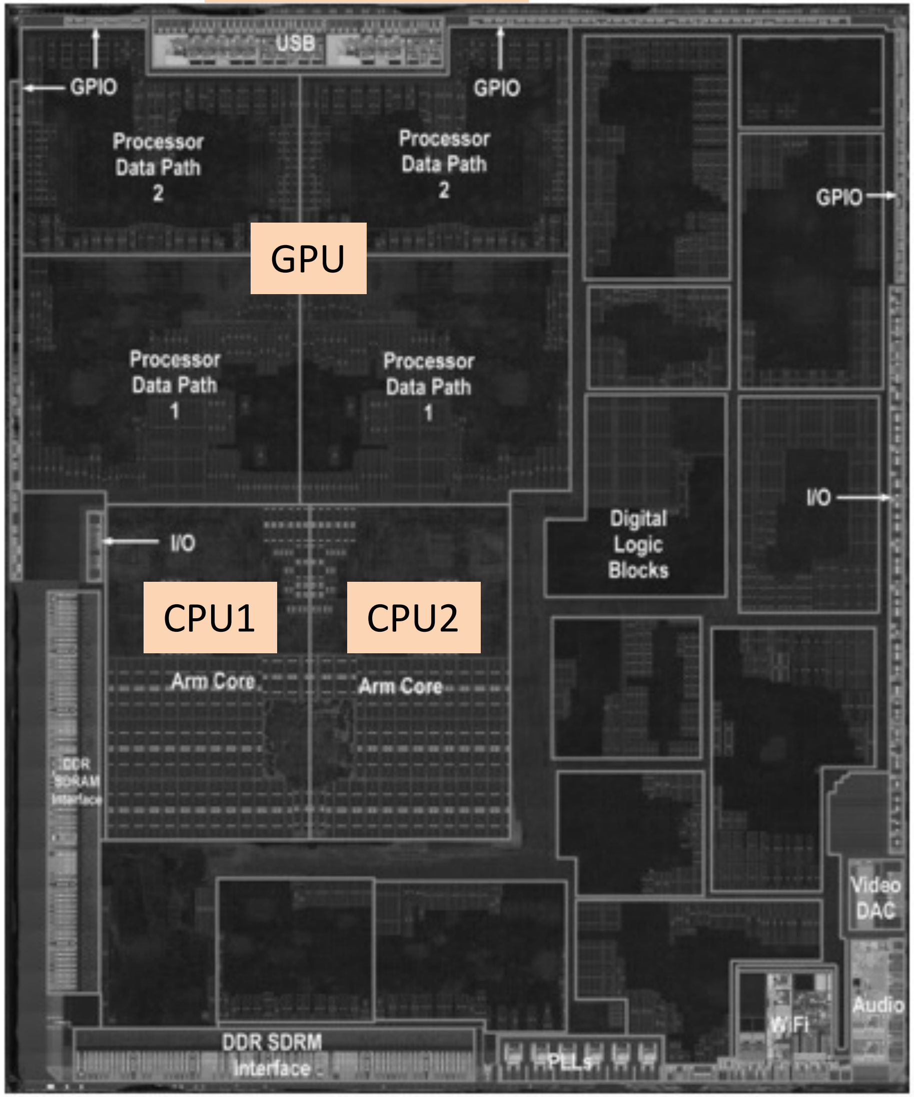
PRICE: 4GB model, $499;
8GB version, $599 (with
a two-year contract)

ARM SoC/A5

Hennessy & Patterson

Apple ARM **A5** 12.1x10.2mm





Apple ARM SoC

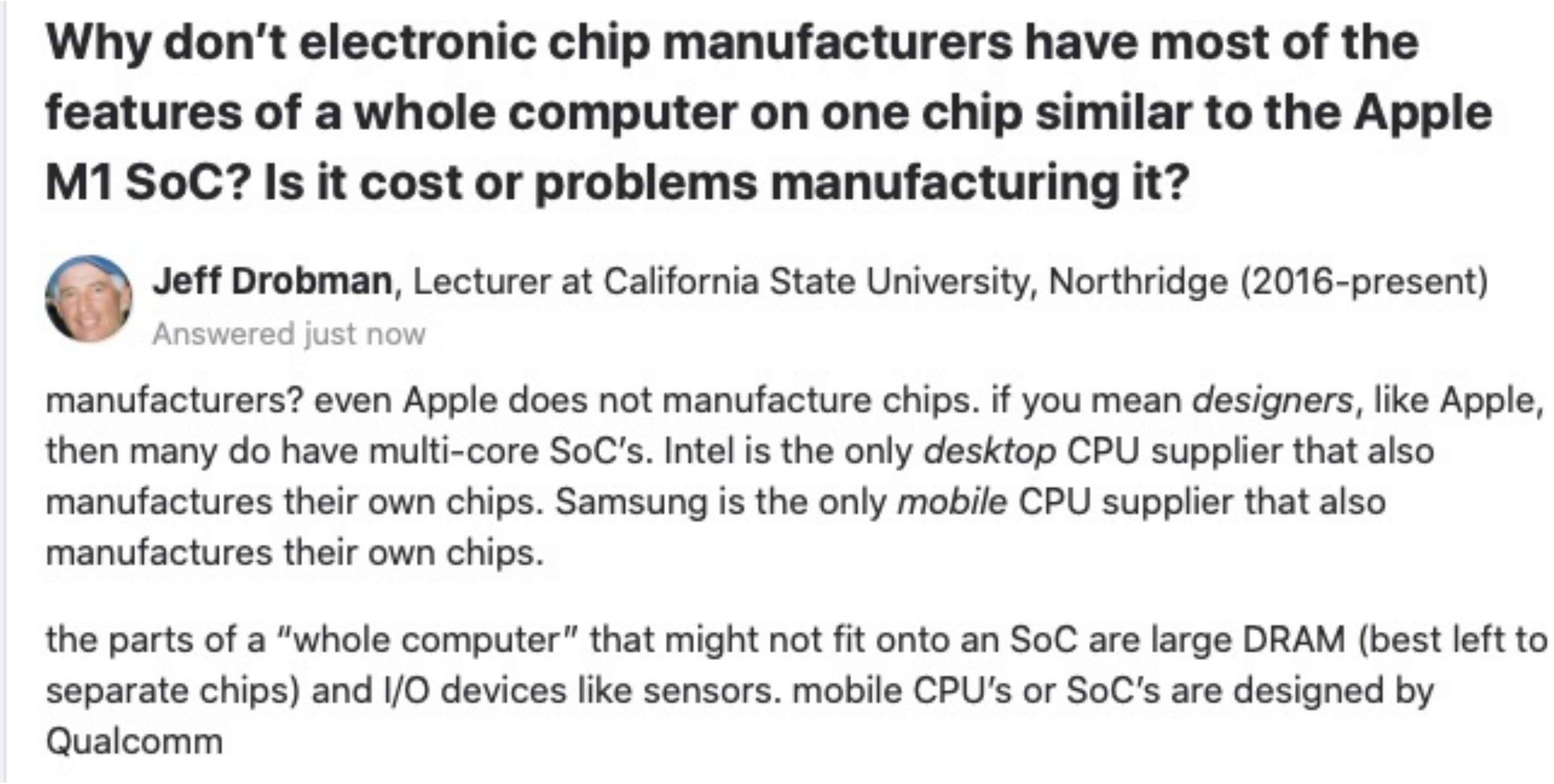
Apple ARM SoC

the first ARM designs were for LOW POWER for portable devices. to achieve low power, the CPU was designed as simple RISC ISA and low clock frequency. Apple iPhones have used ARM from day 1, since they too initially didn't need high compute performance. over time, ARM models have evolved into more powerful models, including a 64-bit ISA -- necessary for today's computers. so now the time has come to start switching to ARM, mainly due to ARM being a licensable ISA and core that can be designed into anyone's SoC like Apple and many others do. I also note that the ARM ISA has evolved from a simple RISC to a more complex, CISC-like one.

Apple has just announced they will replace x86 CPU's on their **Macs** with their own ARM-based **A13** (or next generation A14/15). makes sense for them to use their own chips now that they are powerful enough. this will also give Apple the same **AI** performance capabilities across all their hardware devices (e.g., Siri) -- way more AI power than any x86 chips.

The reason -- historically: Apple has upgraded their Macs for the same reason any company does: to be competitive they have to use a top performance CPU. so Apple switched from the 6502 (Apple II) to the M68000 in the 1st Mac, then upgraded to the Mot PPC. but then Mot stopped making PPC's, so Apple had nowhere else to go but x86 (Intel or AMD) – for Macs. note they have been making their own custom ARM-based chips (A series) for their phones.
Apple has long made their own chips with ARM CPU's since 2007 in their iPhones, designing an ever more powerful SoC (A4-A13). These new **A13** SoC's are now much better than the Intel x86 chips in overall performance -- including machine learning (ML) via on-chip GPU cores plus neural engine -- and with superior power management.

SoC's



Why don't electronic chip manufacturers have most of the features of a whole computer on one chip similar to the Apple M1 SoC? Is it cost or problems manufacturing it?

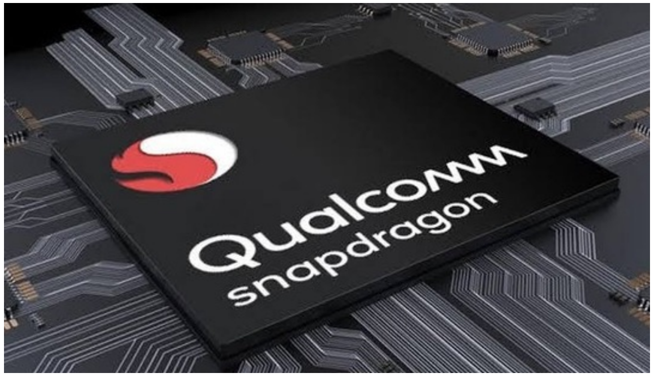
Jeff Drobman, Lecturer at California State University, Northridge (2016-present)
Answered just now

manufacturers? even Apple does not manufacture chips. if you mean *designers*, like Apple, then many do have multi-core SoC's. Intel is the only *desktop* CPU supplier that also manufactures their own chips. Samsung is the only *mobile* CPU supplier that also manufactures their own chips.

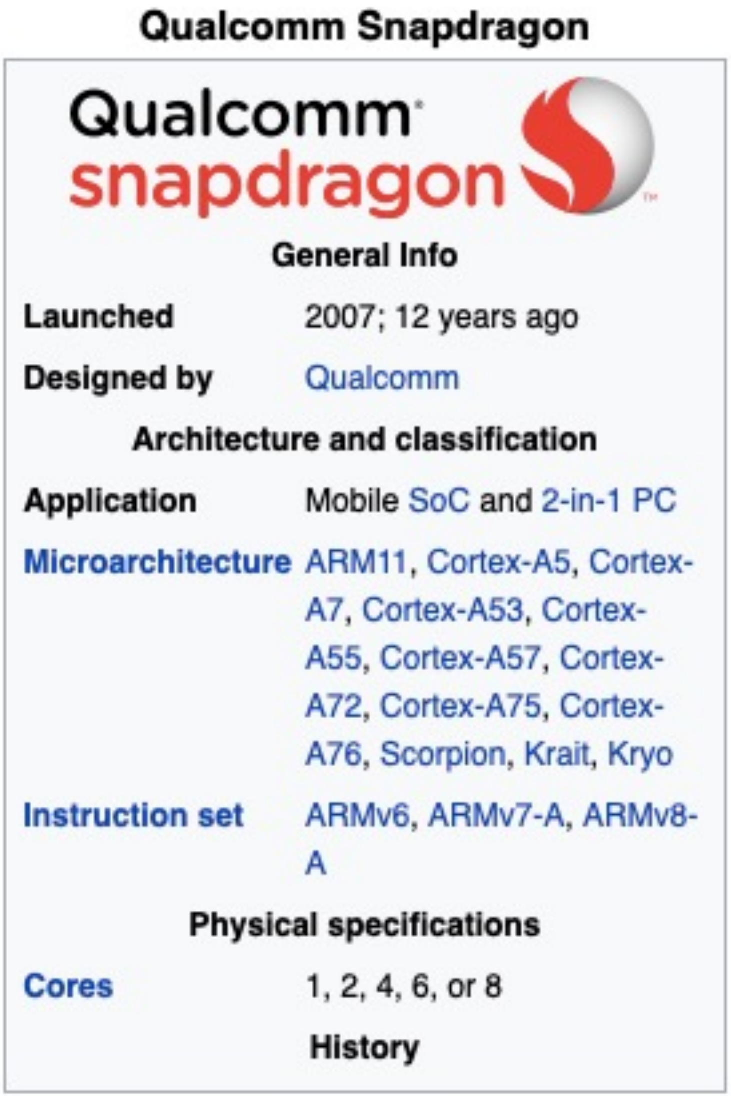
the parts of a "whole computer" that might not fit onto an SoC are large DRAM (best left to separate chips) and I/O devices like sensors. mobile CPU's or SoC's are designed by Qualcomm

Section

Qualcomm *Snapdragon*



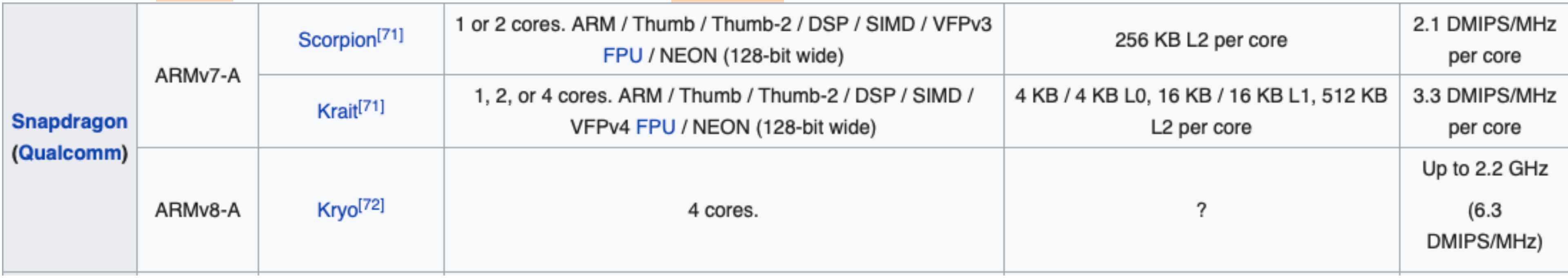
Qualcomm Snapdragon



Qualcomm Snapdragon

General Info	
Launched	2007; 12 years ago
Designed by	Qualcomm
Architecture and classification	
Application	Mobile SoC and 2-in-1 PC
Microarchitecture	ARM11, Cortex-A5, Cortex-A7, Cortex-A53, Cortex-A55, Cortex-A57, Cortex-A72, Cortex-A75, Cortex-A76, Scorpion, Krait, Kryo
Instruction set	ARMv6, ARMv7-A, ARMv8-A
Physical specifications	
Cores	1, 2, 4, 6, or 8
History	

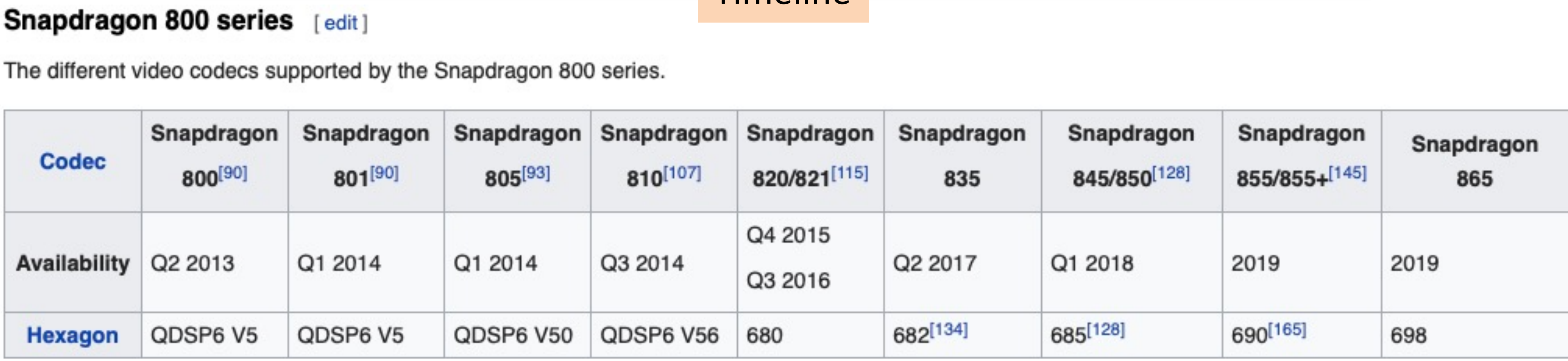
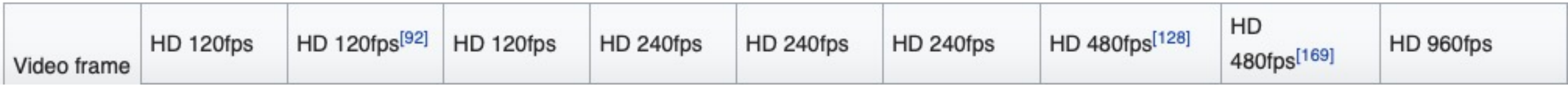
Snapdragon ARM SoC



	ISA		Cores	Cache	Speed
Snapdragon (Qualcomm)	ARMv7-A	Scorpion[71]	1 or 2 cores. ARM / Thumb / Thumb-2 / DSP / SIMD / VFPv3 FPU / NEON (128-bit wide)	256 KB L2 per core	2.1 DMIPS/MHz per core
		Krait[71]	1, 2, or 4 cores. ARM / Thumb / Thumb-2 / DSP / SIMD / VFPv4 FPU / NEON (128-bit wide)	4 KB / 4 KB L0, 16 KB / 16 KB L1, 512 KB L2 per core	3.3 DMIPS/MHz per core
	ARMv8-A	Kryo[72]	4 cores.	?	Up to 2.2 GHz (6.3 DMIPS/MHz)

Qualcomm Snapdragon

Timeline

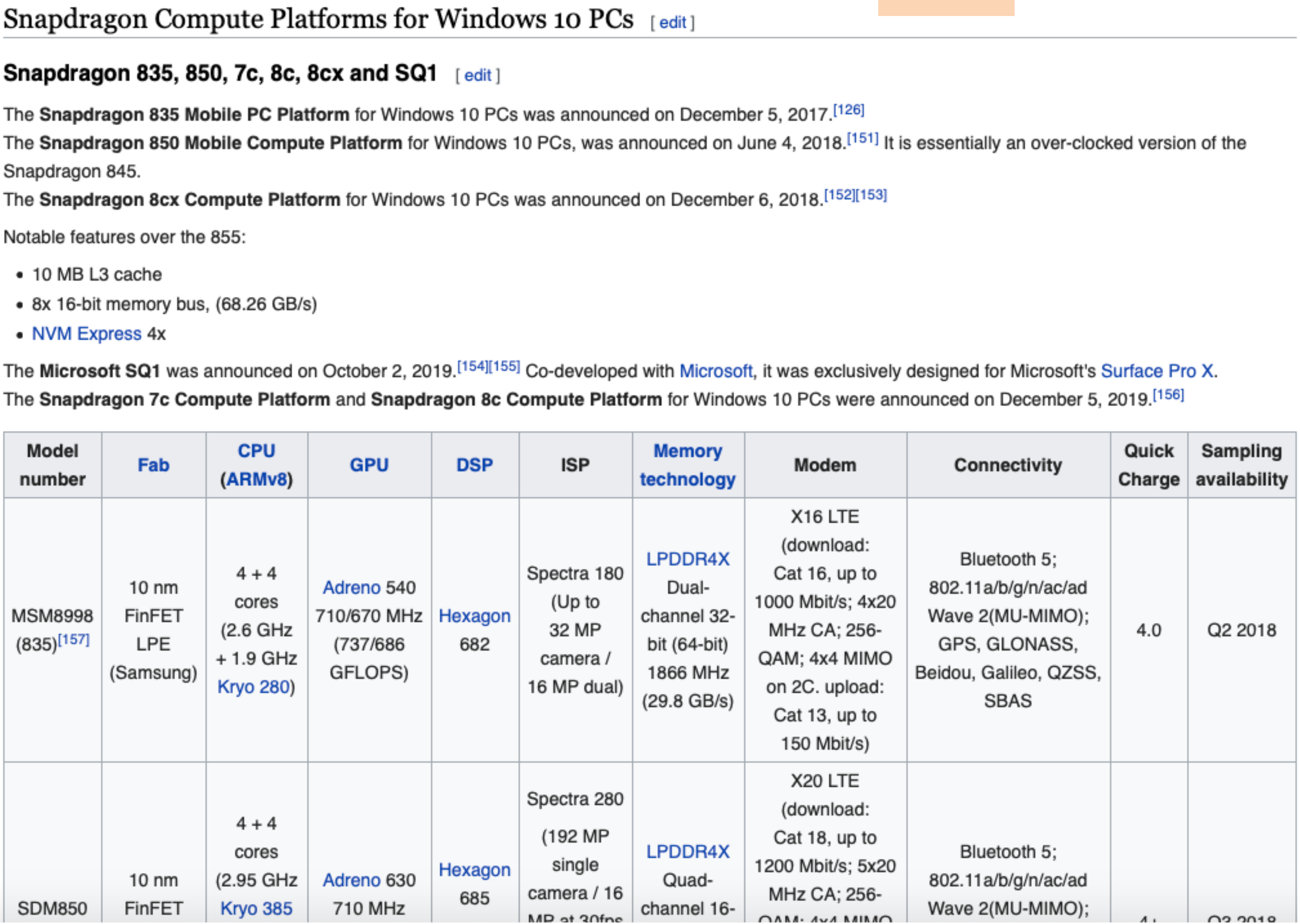
Snapdragon 800 series [edit]

The different video codecs supported by the Snapdragon 800 series.

Codec	Snapdragon 800[90]	Snapdragon 801[90]	Snapdragon 805[93]	Snapdragon 810[107]	Snapdragon 820/821[115]	Snapdragon 835	Snapdragon 845/850[128]	Snapdragon 855/855+[145]	Snapdragon 865
Availability	Q2 2013	Q1 2014	Q1 2014	Q3 2014	Q4 2015 Q3 2016	Q2 2017	Q1 2018	2019	2019
Hexagon	QDSP6 V5	QDSP6 V5	QDSP6 V50	QDSP6 V56	680	682[134]	685[128]	690[165]	698
Video frame	HD 120fps	HD 120fps[92]	HD 120fps	HD 240fps	HD 240fps	HD 240fps	HD 480fps[128]	HD 480fps[169]	HD 960fps

Qualcomm Snapdragon

Win 10



Snapdragon Compute Platforms for Windows 10 PCs [edit]

Snapdragon 835, 850, 7c, 8c, 8cx and SQ1 [edit]

The **Snapdragon 835 Mobile PC Platform** for Windows 10 PCs was announced on December 5, 2017.[126]
The **Snapdragon 850 Mobile Compute Platform** for Windows 10 PCs, was announced on June 4, 2018.[151] It is essentially an over-clocked version of the Snapdragon 845.
The **Snapdragon 8cx Compute Platform** for Windows 10 PCs was announced on December 6, 2018.[152][153]

Notable features over the 855:

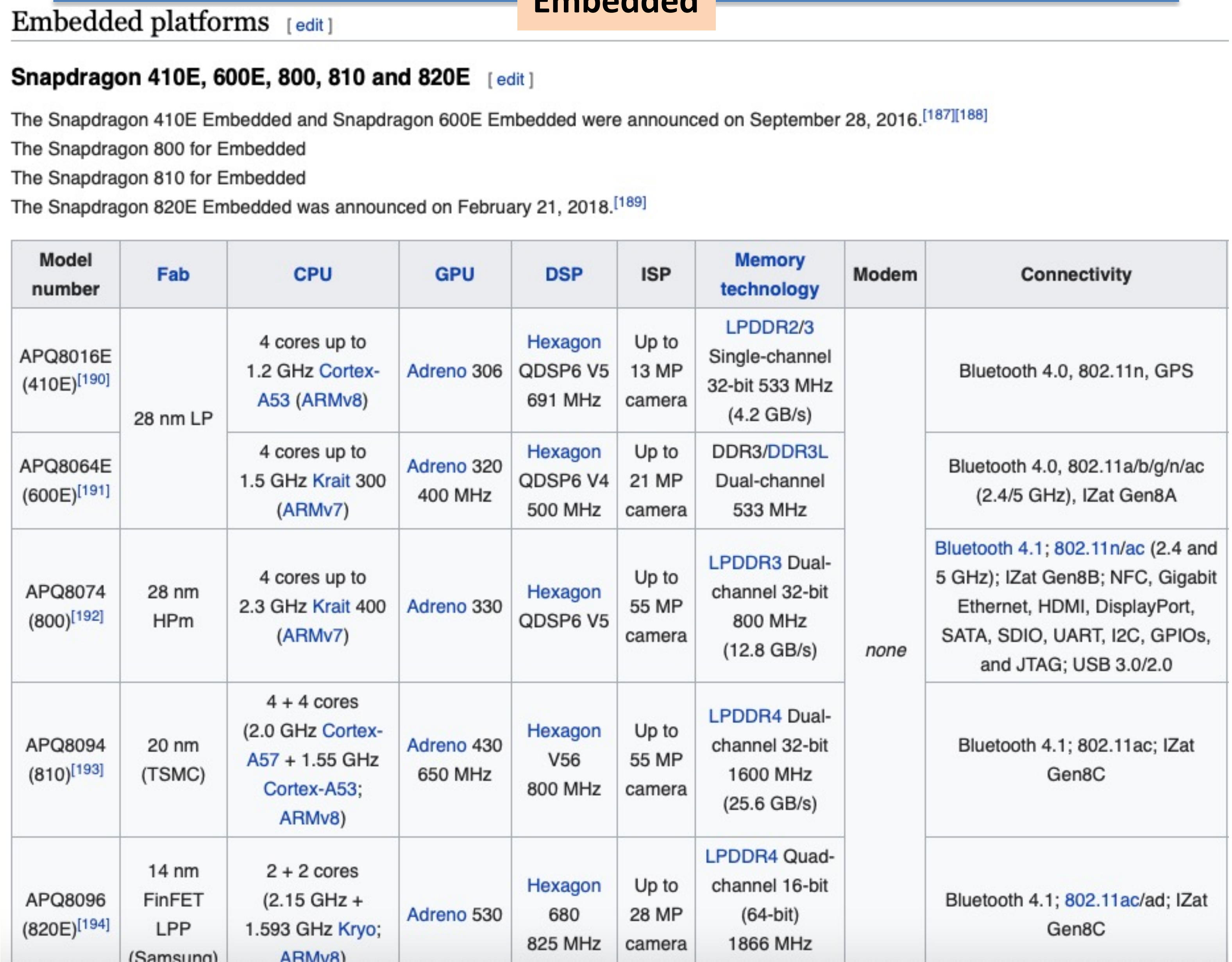
- 10 MB L3 cache
- 8x 16-bit memory bus, (68.26 GB/s)
- NVM Express 4x

The **Microsoft SQ1** was announced on October 2, 2019.[154][155] Co-developed with Microsoft, it was exclusively designed for Microsoft's Surface Pro X.
The **Snapdragon 7c Compute Platform** and **Snapdragon 8c Compute Platform** for Windows 10 PCs were announced on December 5, 2019.[156]

Model number	Fab	CPU (ARMv8)	GPU	DSP	ISP	Memory technology	Modem	Connectivity	Quick Charge	Sampling availability
MSM8998 (835)[157]	10 nm FinFET LPE (Samsung)	4 + 4 cores (2.6 GHz + 1.9 GHz Kryo 280)	Adreno 540 710/670 MHz (737/686 GFLOPS)	Hexagon 682	Spectra 180 (Up to 32 MP camera / 16 MP dual)	LPDDR4X Dual-channel 32-bit (64-bit) 1866 MHz (29.8 GB/s)	X16 LTE (download: Cat 16, up to 1000 Mbit/s; 4x20 MHz CA; 256-QAM; 4x4 MIMO on 2C. upload: Cat 13, up to 150 Mbit/s)	Bluetooth 5; 802.11a/b/g/n/ac/ad Wave 2(MU-MIMO); GPS, GLONASS, Beidou, Galileo, QZSS, SBAS	4.0	Q2 2018
SDM850	10 nm FinFET	4 + 4 cores (2.95 GHz Kryo 385	Adreno 630 710 MHz	Hexagon 685	Spectra 280 (192 MP single camera / 16 MP at 30fps	LPDDR4X Quad-channel 16-	X20 LTE (download: Cat 18, up to 1200 Mbit/s; 5x20 MHz CA; 256-QAM; 4x4 MIMO	Bluetooth 5; 802.11a/b/g/n/ac/ad Wave 2(MU-MIMO);	4+	Q3 2018

Qualcomm Snapdragon

Embedded



Embedded platforms [edit]

Snapdragon 410E, 600E, 800, 810 and 820E [edit]

The Snapdragon 410E Embedded and Snapdragon 600E Embedded were announced on September 28, 2016.[187][188]

The Snapdragon 800 for Embedded

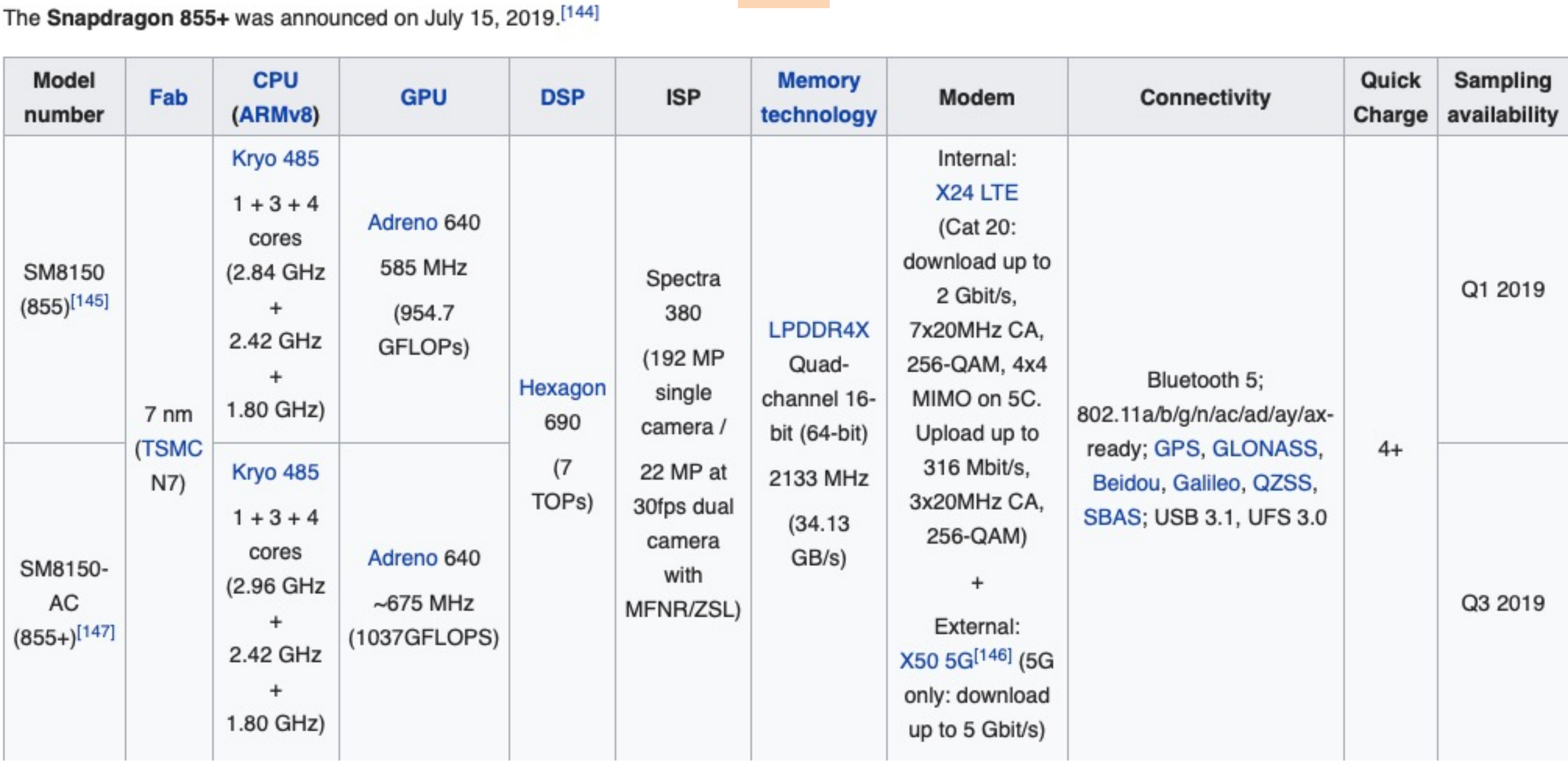
The Snapdragon 810 for Embedded

The Snapdragon 820E Embedded was announced on February 21, 2018.[189]

Model number	Fab	CPU	GPU	DSP	ISP	Memory technology	Modem	Connectivity
APQ8016E (410E)[190]	28 nm LP	4 cores up to 1.2 GHz Cortex-A53 (ARMv8)	Adreno 306	Hexagon QDSP6 V5 691 MHz	Up to 13 MP camera	LPDDR2/3 Single-channel 32-bit 533 MHz (4.2 GB/s)	*none*	Bluetooth 4.0, 802.11n, GPS
APQ8064E (600E)[191]		4 cores up to 1.5 GHz Krait 300 (ARMv7)	Adreno 320 400 MHz	Hexagon QDSP6 V4 500 MHz	Up to 21 MP camera	DDR3/DDR3L Dual-channel 533 MHz		Bluetooth 4.0, 802.11a/b/g/n/ac (2.4/5 GHz), IZat Gen8A
APQ8074 (800)[192]	28 nm HPm	4 cores up to 2.3 GHz Krait 400 (ARMv7)	Adreno 330	Hexagon QDSP6 V5	Up to 55 MP camera	LPDDR3 Dual-channel 32-bit 800 MHz (12.8 GB/s)		Bluetooth 4.1; 802.11n/ac (2.4 and 5 GHz); IZat Gen8B; NFC, Gigabit Ethernet, HDMI, DisplayPort, SATA, SDIO, UART, I2C, GPIOs, and JTAG; USB 3.0/2.0
APQ8094 (810)[193]	20 nm (TSMC)	4 + 4 cores (2.0 GHz Cortex-A57 + 1.55 GHz Cortex-A53; ARMv8)	Adreno 430 650 MHz	Hexagon V56 800 MHz	Up to 55 MP camera	LPDDR4 Dual-channel 32-bit 1600 MHz (25.6 GB/s)		Bluetooth 4.1; 802.11ac; IZat Gen8C
APQ8096 (820E)[194]	14 nm FinFET LPP (Samsung)	2 + 2 cores (2.15 GHz + 1.593 GHz Kryo; ARMv8)	Adreno 530	Hexagon 680 825 MHz	Up to 28 MP camera	LPDDR4 Quad-channel 16-bit (64-bit) 1866 MHz		Bluetooth 4.1; 802.11ac/ad; IZat Gen8C

Qualcomm Snapdragon

855

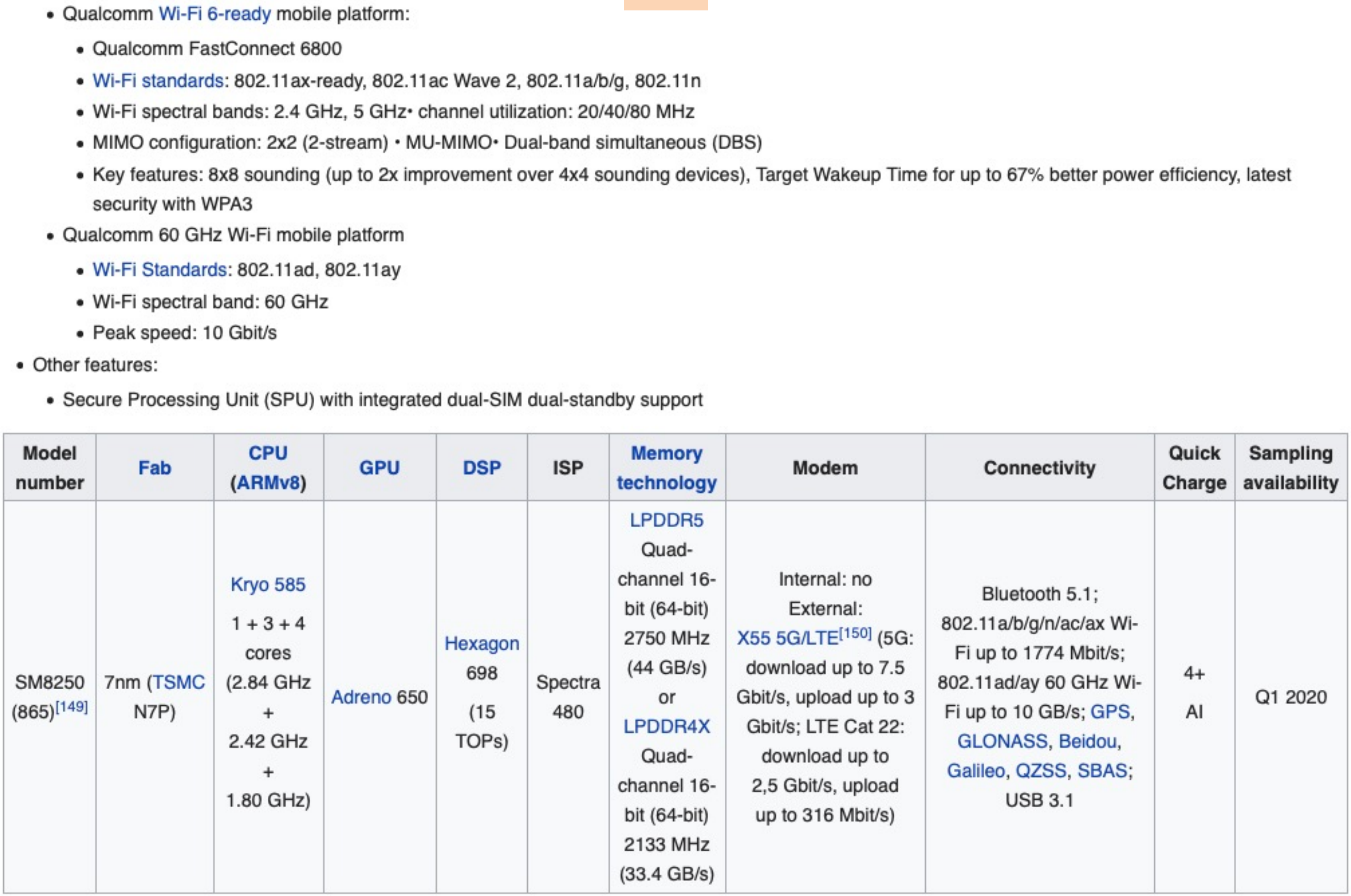


The **Snapdragon 855+** was announced on July 15, 2019.[144]

Model number	Fab	CPU (ARMv8)	GPU	DSP	ISP	Memory technology	Modem	Connectivity	Quick Charge	Sampling availability
SM8150 (855)[145]	7 nm (TSMC N7)	Kryo 485 1 + 3 + 4 cores (2.84 GHz + 2.42 GHz + 1.80 GHz)	Adreno 640 585 MHz (954.7 GFLOPs)	Hexagon 690 (7 TOPs)	Spectra 380 (192 MP single camera / 22 MP at 30fps dual camera with MFNR/ZSL)	LPDDR4X Quad-channel 16-bit (64-bit) 2133 MHz (34.13 GB/s)	Internal: X24 LTE (Cat 20: download up to 2 Gbit/s, 7x20MHz CA, 256-QAM, 4x4 MIMO on 5C. Upload up to 316 Mbit/s, 3x20MHz CA, 256-QAM) + External: X50 5G[146] (5G only: download up to 5 Gbit/s)	Bluetooth 5; 802.11a/b/g/n/ac/ad/ay/ax-ready; GPS, GLONASS, Beidou, Galileo, QZSS, SBAS; USB 3.1, UFS 3.0	4+	Q1 2019
SM8150-AC (855+)[147]		Kryo 485 1 + 3 + 4 cores (2.96 GHz + 2.42 GHz + 1.80 GHz)	Adreno 640 ~675 MHz (1037GFLOPS)							Q3 2019

Qualcomm Snapdragon

865

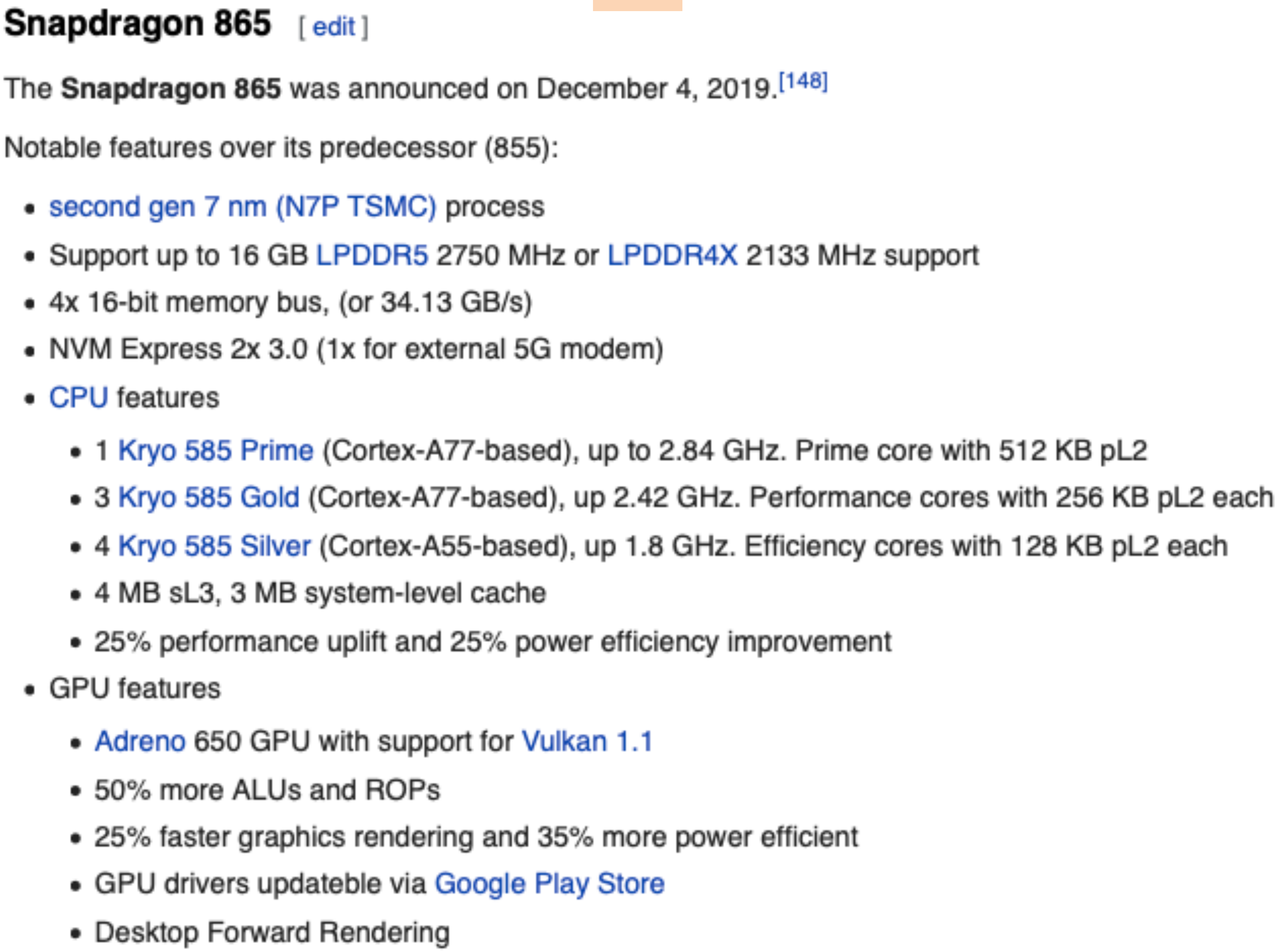


- Qualcomm Wi-Fi 6-ready mobile platform:
  - Qualcomm FastConnect 6800
  - Wi-Fi standards: 802.11ax-ready, 802.11ac Wave 2, 802.11a/b/g, 802.11n
  - Wi-Fi spectral bands: 2.4 GHz, 5 GHz• channel utilization: 20/40/80 MHz
  - MIMO configuration: 2x2 (2-stream) • MU-MIMO• Dual-band simultaneous (DBS)
  - Key features: 8x8 sounding (up to 2x improvement over 4x4 sounding devices), Target Wakeup Time for up to 67% better power efficiency, latest security with WPA3
- Qualcomm 60 GHz Wi-Fi mobile platform
  - Wi-Fi Standards: 802.11ad, 802.11ay
  - Wi-Fi spectral band: 60 GHz
  - Peak speed: 10 Gbit/s
- Other features:
  - Secure Processing Unit (SPU) with integrated dual-SIM dual-standby support

Model number	Fab	CPU (ARMv8)	GPU	DSP	ISP	Memory technology	Modem	Connectivity	Quick Charge	Sampling availability
SM8250 (865)[149]	7nm (TSMC N7P)	Kryo 585 1 + 3 + 4 cores (2.84 GHz + 2.42 GHz + 1.80 GHz)	Adreno 650	Hexagon 698 (15 TOPs)	Spectra 480	LPDDR5 Quad-channel 16-bit (64-bit) 2750 MHz (44 GB/s) or LPDDR4X Quad-channel 16-bit (64-bit) 2133 MHz (33.4 GB/s)	Internal: no External: X55 5G/LTE[150] (5G: download up to 7.5 Gbit/s, upload up to 3 Gbit/s; LTE Cat 22: download up to 2,5 Gbit/s, upload up to 316 Mbit/s)	Bluetooth 5.1; 802.11a/b/g/n/ac/ax Wi-Fi up to 1774 Mbit/s; 802.11ad/ay 60 GHz Wi-Fi up to 10 GB/s; GPS, GLONASS, Beidou, Galileo, QZSS, SBAS; USB 3.1	4+ AI	Q1 2020

Qualcomm Snapdragon

865



Snapdragon 865 [edit]

The **Snapdragon 865** was announced on December 4, 2019.[148]

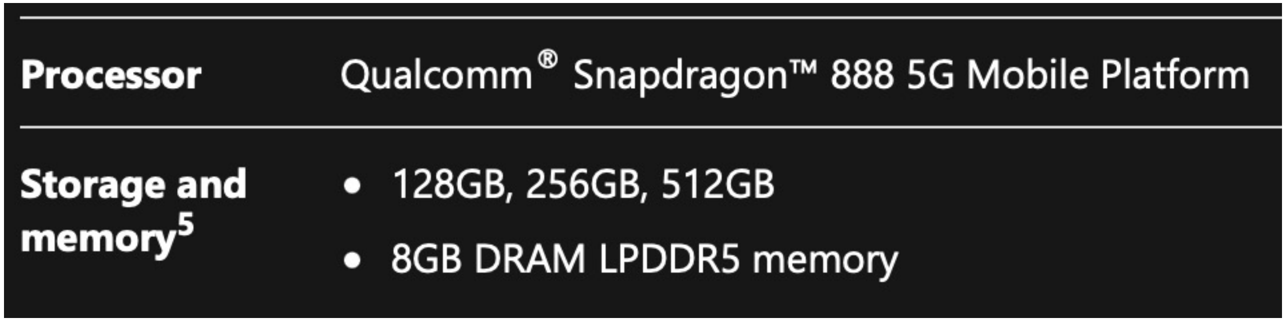
Notable features over its predecessor (855):

- second gen 7 nm (N7P TSMC) process
- Support up to 16 GB LPDDR5 2750 MHz or LPDDR4X 2133 MHz support
- 4x 16-bit memory bus, (or 34.13 GB/s)
- NVM Express 2x 3.0 (1x for external 5G modem)
- CPU features
  - 1 Kryo 585 Prime (Cortex-A77-based), up to 2.84 GHz. Prime core with 512 KB pL2
  - 3 Kryo 585 Gold (Cortex-A77-based), up 2.42 GHz. Performance cores with 256 KB pL2 each
  - 4 Kryo 585 Silver (Cortex-A55-based), up 1.8 GHz. Efficiency cores with 128 KB pL2 each
  - 4 MB sL3, 3 MB system-level cache
  - 25% performance uplift and 25% power efficiency improvement
- GPU features
  - Adreno 650 GPU with support for Vulkan 1.1
  - 50% more ALUs and ROPs
  - 25% faster graphics rendering and 35% more power efficient
  - GPU drivers updateble via Google Play Store
  - Desktop Forward Rendering

Qualcomm Snapdragon

Surface Duo 2

Two screens, limitless possibilities



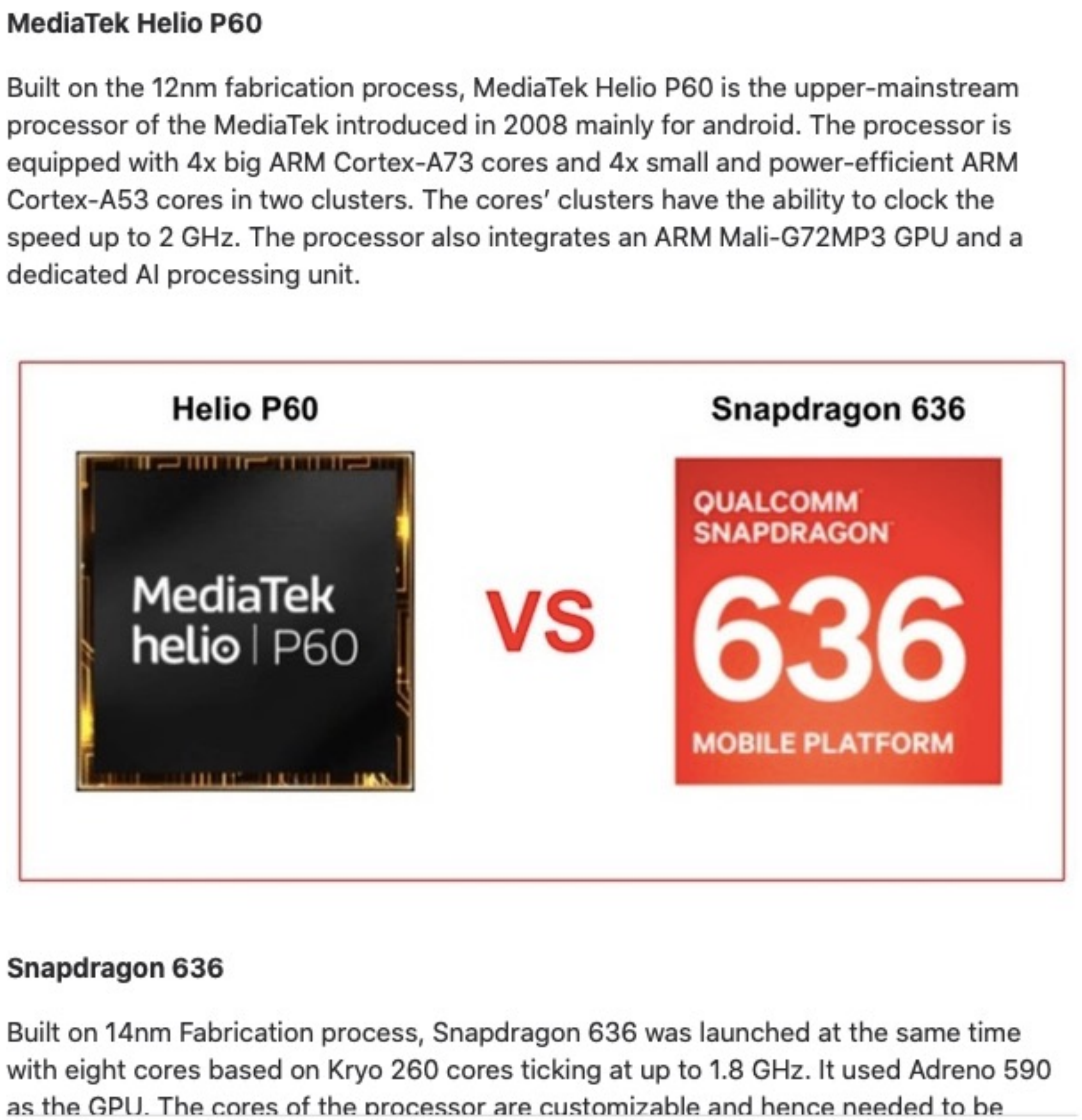
Processor	Qualcomm® Snapdragon™ 888 5G Mobile Platform
Storage and memory[5]	- 128GB, 256GB, 512GB - 8GB DRAM LPDDR5 memory

Section

MediaTek

MediaTek vs. Snapdragon

MediaTek vs. Snapdragon



MediaTek Helio P60

Built on the 12nm fabrication process, MediaTek Helio P60 is the upper-mainstream processor of the MediaTek introduced in 2008 mainly for android. The processor is equipped with 4x big ARM Cortex-A73 cores and 4x small and power-efficient ARM Cortex-A53 cores in two clusters. The cores' clusters have the ability to clock the speed up to 2 GHz. The processor also integrates an ARM Mali-G72MP3 GPU and a dedicated AI processing unit.

Snapdragon 636

Built on 14nm Fabrication process, Snapdragon 636 was launched at the same time with eight cores based on Kryo 260 cores ticking at up to 1.8 GHz. It used Adreno 590 as the GPU. The cores of the processor are customizable and hence needed to be

MediaTek vs. Snapdragon

MediaTek helio | P22

POWERING THE NEW **PREMIUM**

UNPARALLELED PHOTOGRAPHY
AI Backend Features with NeuroPilot

INVENTIVE PICTURE AND VIDEO QUALITY
20:9 ratio with HD+ quality

POWER-PACKED PERFORMANCE
Octa Core CPU with ARM Cortex-A53 at 2.0 GHz

ASTOUNDING ARCHITECTURE
Latest TSMC 12nm FinFET production process

EFFICIENCY REDEFINED
TAS (2.0) for optimal signal quality at lowest power use

OUTSTANDING CONNECTIVITY
Dual 4G VoLTE Support

This Chip Changes Everything

Section

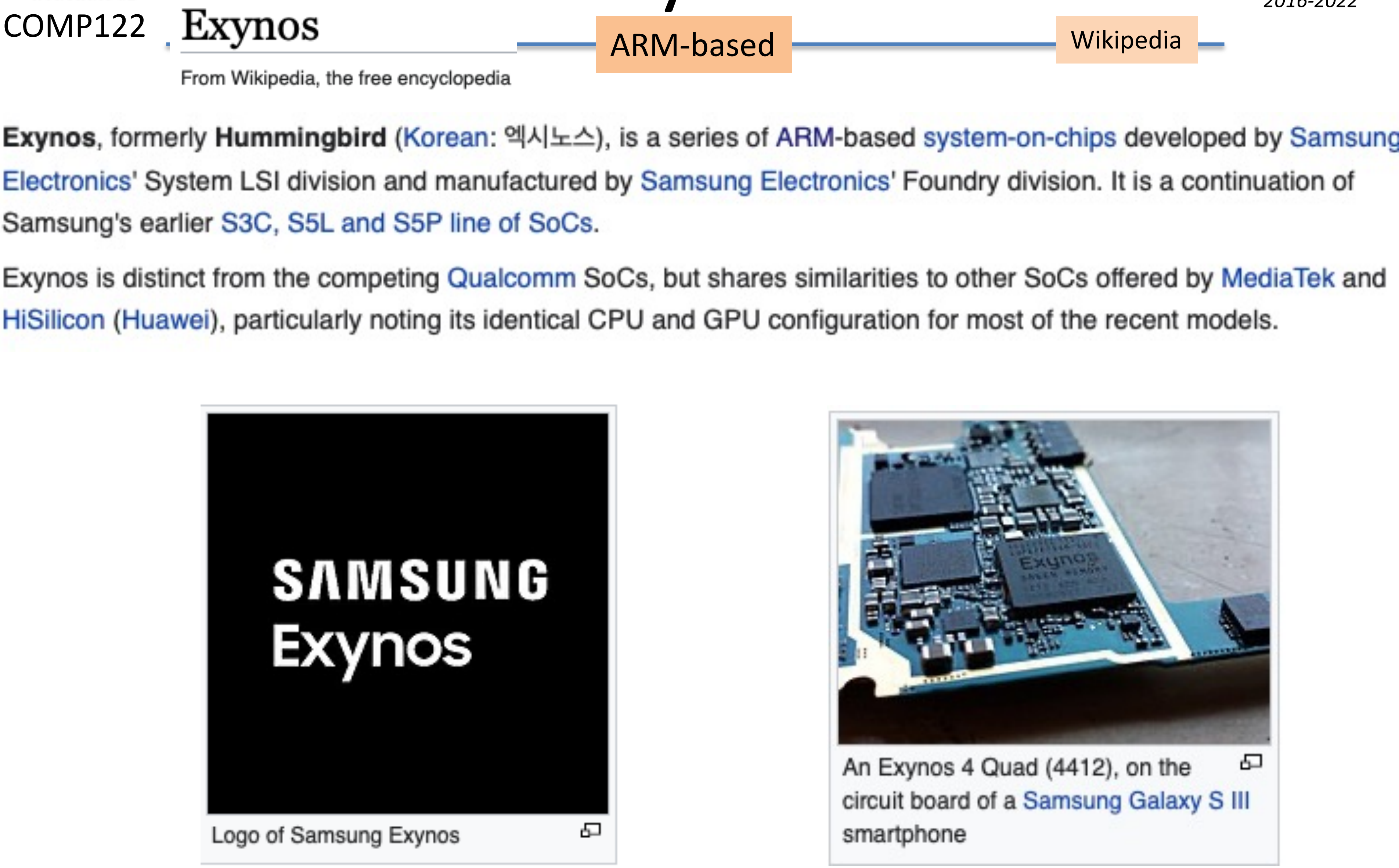
Samsung

Exynos

Exynos

ARM-based

Wikipedia



Exynos

From Wikipedia, the free encyclopedia

Exynos, formerly **Hummingbird** (Korean: 엑시노스), is a series of ARM-based system-on-chips developed by Samsung Electronics' System LSI division and manufactured by Samsung Electronics' Foundry division. It is a continuation of Samsung's earlier S3C, S5L and S5P line of SoCs.

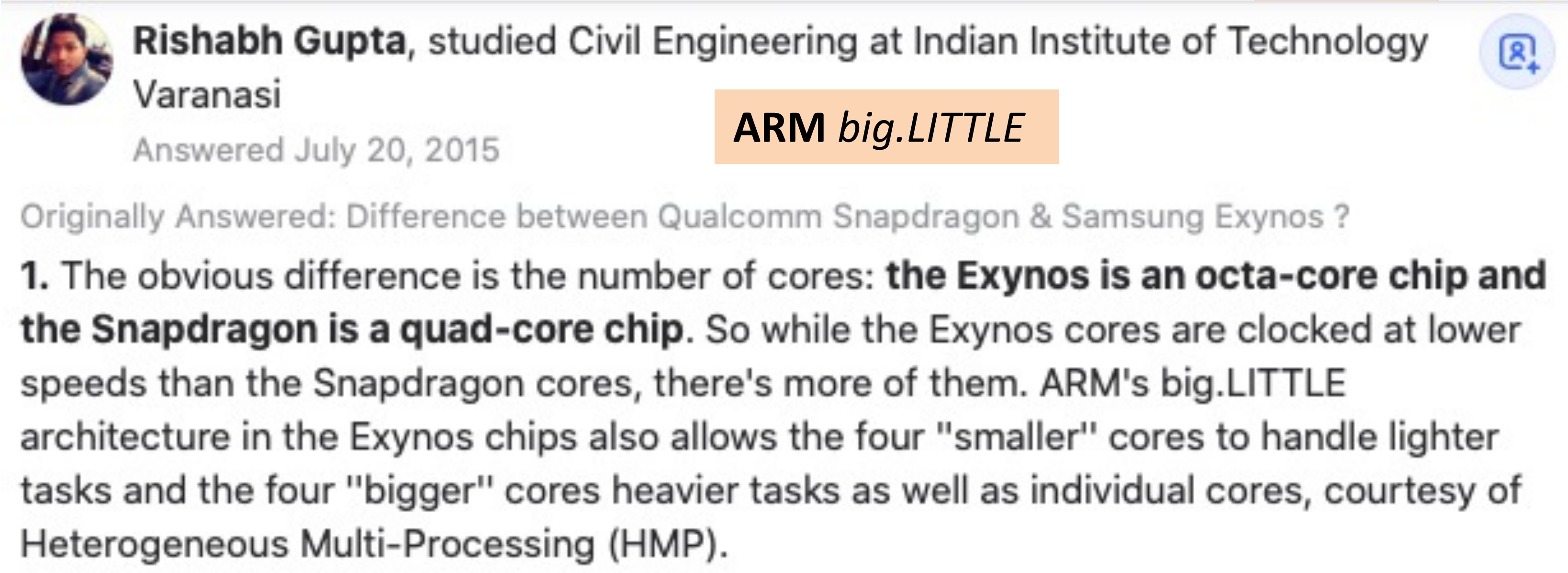
Exynos is distinct from the competing Qualcomm SoCs, but shares similarities to other SoCs offered by MediaTek and HiSilicon (Huawei), particularly noting its identical CPU and GPU configuration for most of the recent models.

Logo of Samsung Exynos

An Exynos 4 Quad (4412), on the circuit board of a Samsung Galaxy S III smartphone

Exynos

Wikipedia



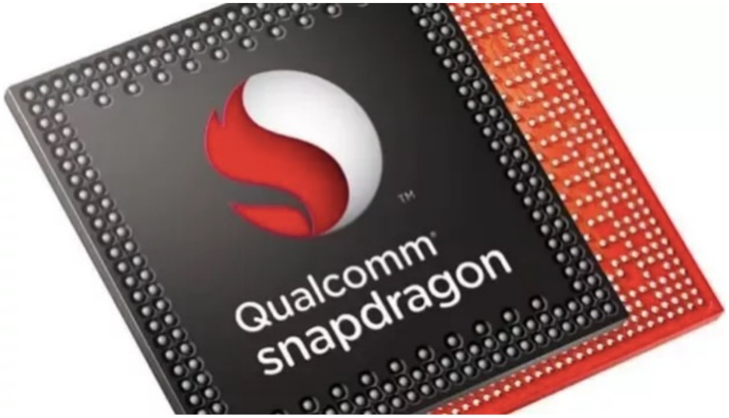
Rishabh Gupta, studied Civil Engineering at Indian Institute of Technology Varanasi

Answered July 20, 2015

ARM *big.LITTLE*

Originally Answered: Difference between Qualcomm Snapdragon & Samsung Exynos ?

1. The obvious difference is the number of cores: **the Exynos is an octa-core chip and the Snapdragon is a quad-core chip**. So while the Exynos cores are clocked at lower speeds than the Snapdragon cores, there's more of them. ARM's big.LITTLE architecture in the Exynos chips also allows the four "smaller" cores to handle lighter tasks and the four "bigger" cores heavier tasks as well as individual cores, courtesy of Heterogeneous Multi-Processing (HMP).

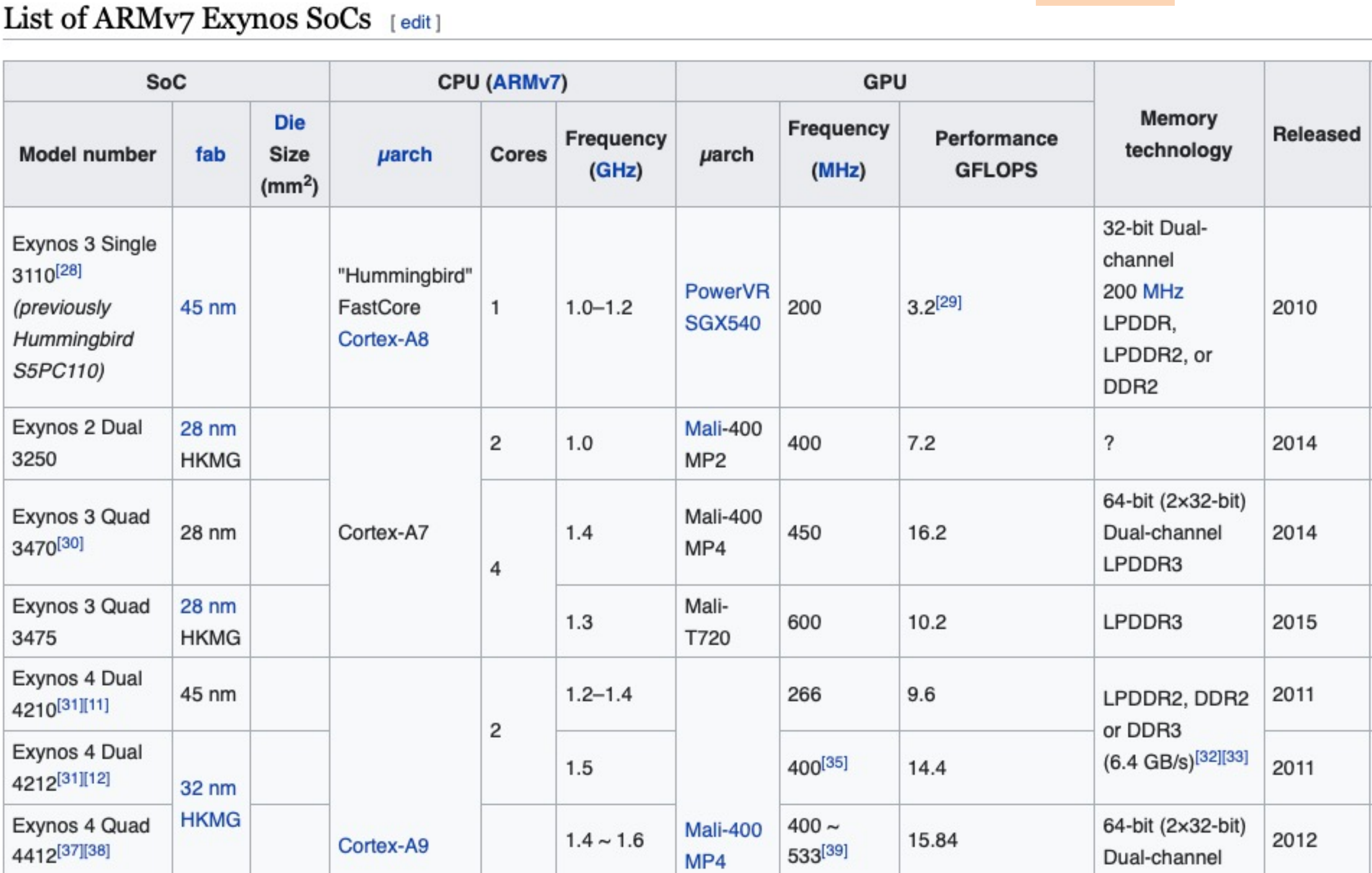


CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP122

Exynos

DR JEFF SOFTWARE
INDIE APP DEVELOPER
© Jeff Drobman
2016-2022

Wikipedia

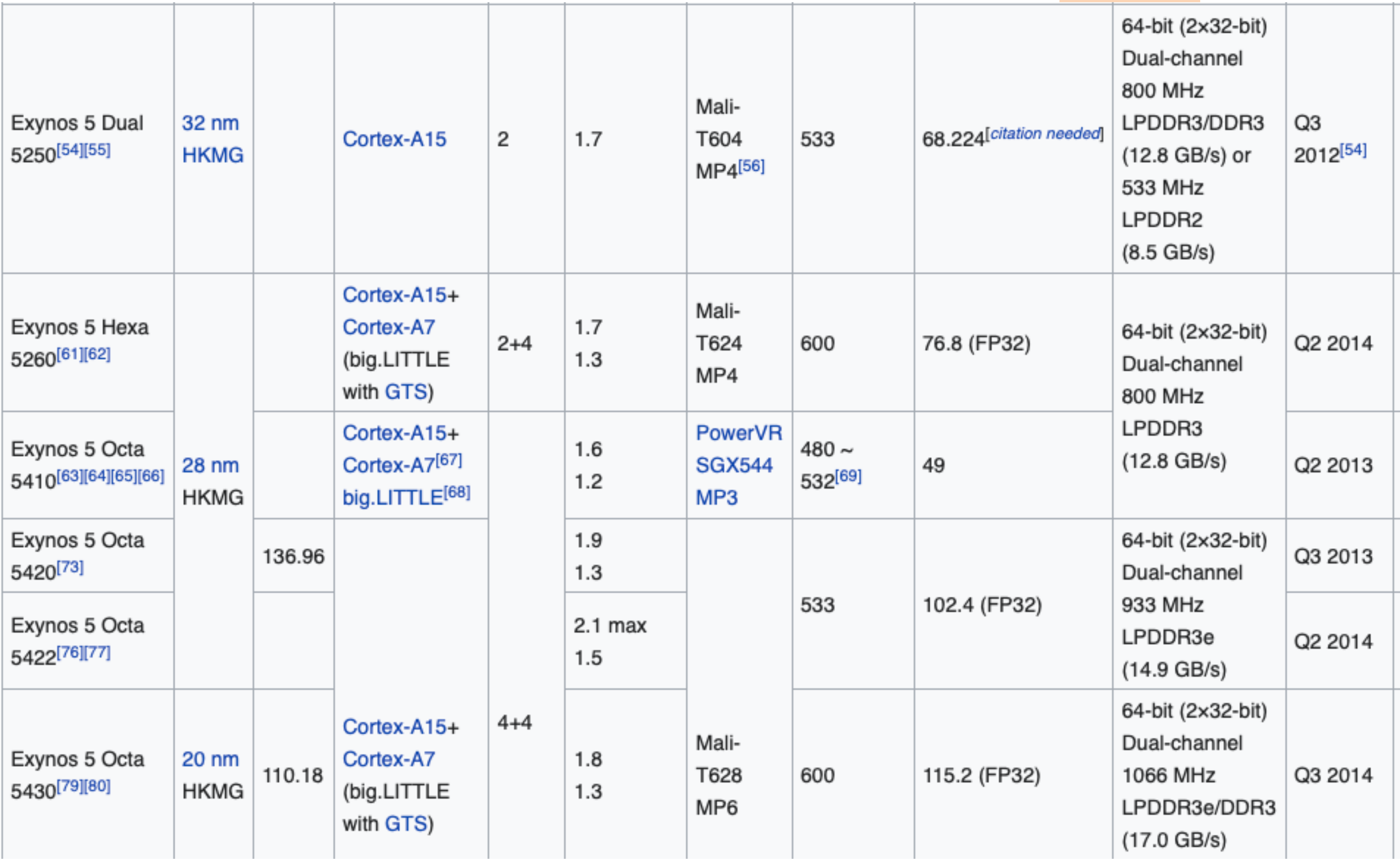


List of ARMv7 Exynos SoCs [edit]

SoC			CPU (ARMv7)			GPU			Memory technology	Released
Model number	fab	Die Size (mm²)	µarch	Cores	Frequency (GHz)	µarch	Frequency (MHz)	Performance GFLOPS		
Exynos 3 Single 3110[28] *(previously Hummingbird S5PC110)*	45 nm		"Hummingbird" FastCore Cortex-A8	1	1.0–1.2	PowerVR SGX540	200	3.2[29]	32-bit Dual-channel 200 MHz LPDDR, LPDDR2, or DDR2	2010
Exynos 2 Dual 3250	28 nm HKMG		Cortex-A7	2	1.0	Mali-400 MP2	400	7.2	?	2014
Exynos 3 Quad 3470[30]	28 nm			4	1.4	Mali-400 MP4	450	16.2	64-bit (2×32-bit) Dual-channel LPDDR3	2014
Exynos 3 Quad 3475	28 nm HKMG				1.3	Mali-T720	600	10.2	LPDDR3	2015
Exynos 4 Dual 4210[31][11]	45 nm		Cortex-A9	2	1.2–1.4	Mali-400 MP4	266	9.6	LPDDR2, DDR2 or DDR3 (6.4 GB/s)[32][33]	2011
Exynos 4 Dual 4212[31][12]	32 nm HKMG				1.5		400[35]	14.4		2011
Exynos 4 Quad 4412[37][38]					1.4 ~ 1.6		400 ~ 533[39]	15.84	64-bit (2×32-bit) Dual-channel	2012

Exynos

Wikipedia



Exynos 5 Dual 5250[54][55]	32 nm HKMG		Cortex-A15	2	1.7	Mali-T604 MP4[56]	533	68.224[citation needed]	64-bit (2x32-bit) Dual-channel 800 MHz LPDDR3/DDR3 (12.8 GB/s) or 533 MHz LPDDR2 (8.5 GB/s)	Q3 2012[54]
Exynos 5 Hexa 5260[61][62]	28 nm HKMG		Cortex-A15+ Cortex-A7 (big.LITTLE with GTS)	2+4	1.7 1.3	Mali-T624 MP4	600	76.8 (FP32)	64-bit (2x32-bit) Dual-channel 800 MHz LPDDR3 (12.8 GB/s)	Q2 2014
Exynos 5 Octa 5410[63][64][65][66]			Cortex-A15+ Cortex-A7[67] big.LITTLE[68]	4+4	1.6 1.2	PowerVR SGX544 MP3	480 ~ 532[69]	49		Q2 2013
Exynos 5 Octa 5420[73]		136.96	Cortex-A15+ Cortex-A7 (big.LITTLE with GTS)		1.9 1.3	Mali-T628 MP6	533	102.4 (FP32)	64-bit (2x32-bit) Dual-channel 933 MHz LPDDR3e (14.9 GB/s)	Q3 2013
Exynos 5 Octa 5422[76][77]					2.1 max 1.5					Q2 2014
Exynos 5 Octa 5430[79][80]	20 nm HKMG	110.18			1.8 1.3		600	115.2 (FP32)	64-bit (2x32-bit) Dual-channel 1066 MHz LPDDR3e/DDR3 (17.0 GB/s)	Q3 2014

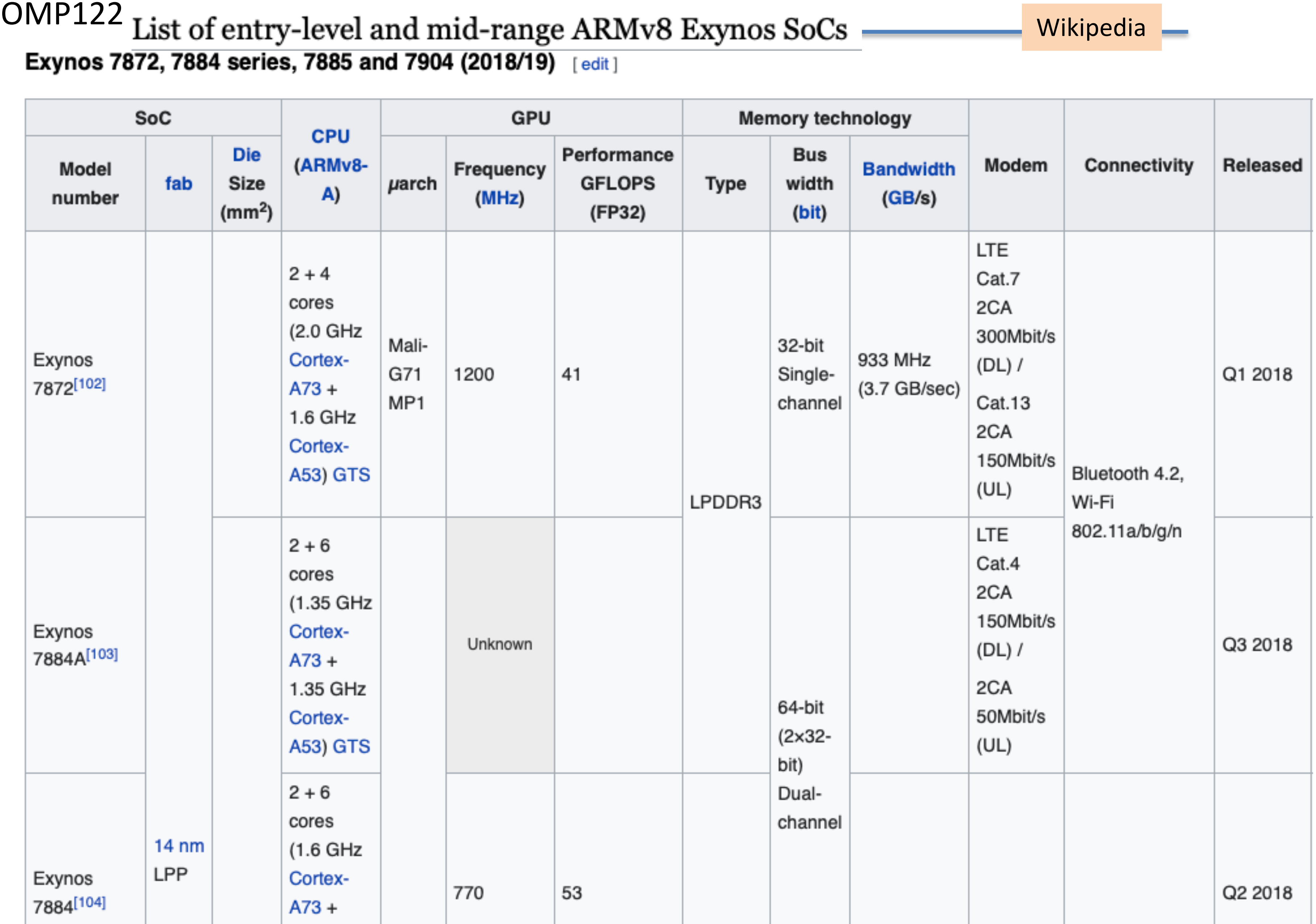
CSUN CALIFORNIA STATE UNIVERSITY NORTHRIDGE

COMP122

Exynos

DR JEFF SOFTWARE INDIE APP DEVELOPER

© Jeff Drobman 2016-2022



Wikipedia

List of entry-level and mid-range ARMv8 Exynos SoCs

Exynos 7872, 7884 series, 7885 and 7904 (2018/19) [edit]

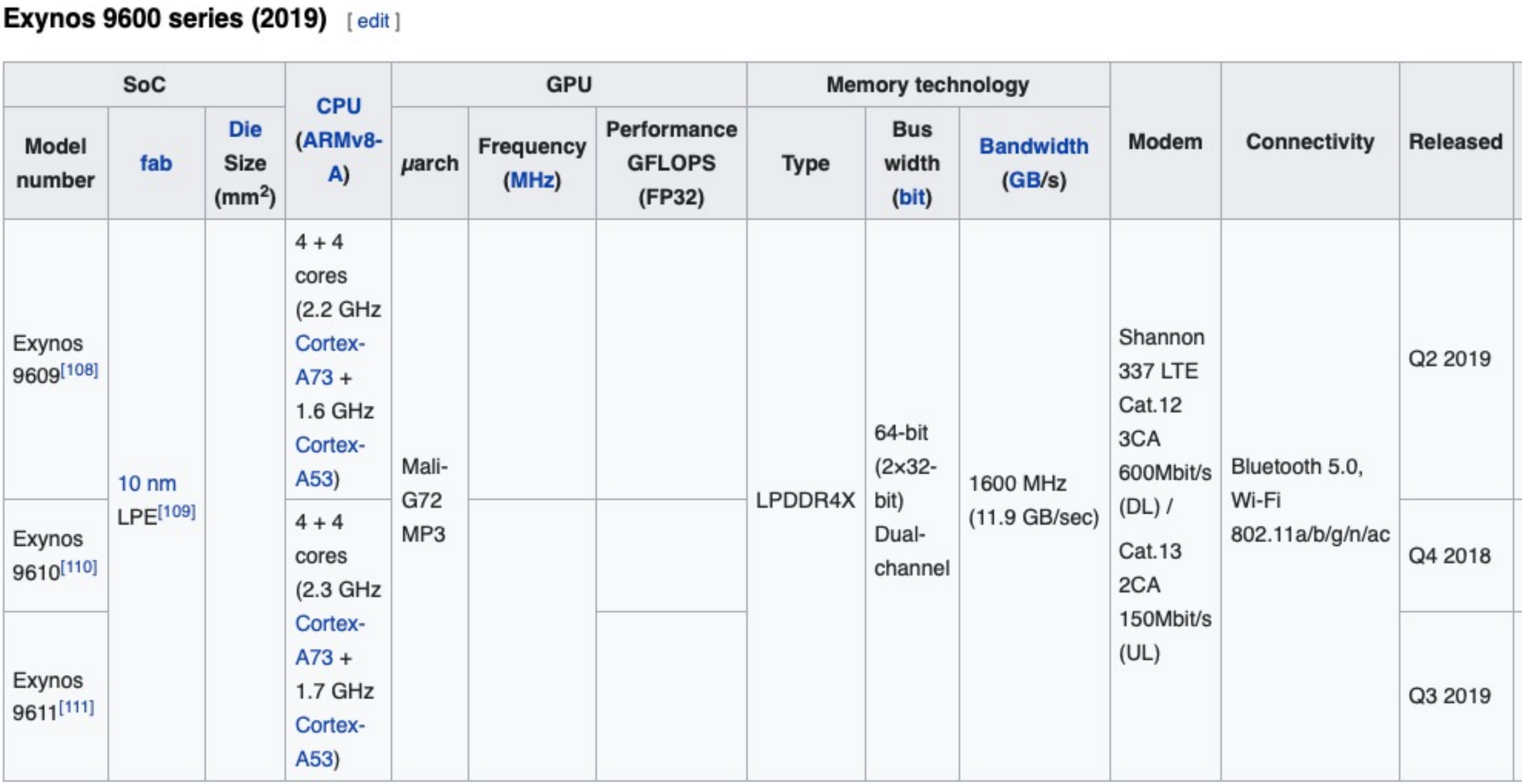
SoC			CPU (ARMv8-A)	GPU			Memory technology			Modem	Connectivity	Released
Model number	fab	Die Size (mm^2)		µarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)			
Exynos 7872[102]	14 nm LPP		2 + 4 cores (2.0 GHz Cortex-A73 + 1.6 GHz Cortex-A53) GTS	Mali-G71 MP1	1200	41	LPDDR3	32-bit Single-channel	933 MHz (3.7 GB/sec)	LTE Cat.7 2CA 300Mbit/s (DL) / Cat.13 2CA 150Mbit/s (UL)	Bluetooth 4.2, Wi-Fi 802.11a/b/g/n	Q1 2018
Exynos 7884A[103]			2 + 6 cores (1.35 GHz Cortex-A73 + 1.35 GHz Cortex-A53) GTS		Unknown			64-bit (2×32-bit) Dual-channel		LTE Cat.4 2CA 150Mbit/s (DL) / 2CA 50Mbit/s (UL)		Q3 2018
Exynos 7884[104]			2 + 6 cores (1.6 GHz Cortex-A73 +		770	53						Q2 2018

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP122

Exynos

DR JEFF SOFTWARE
INDIE APP DEVELOPER
© Jeff Drobman
2016-2022

Wikipedia



Exynos 9600 series (2019) [edit]

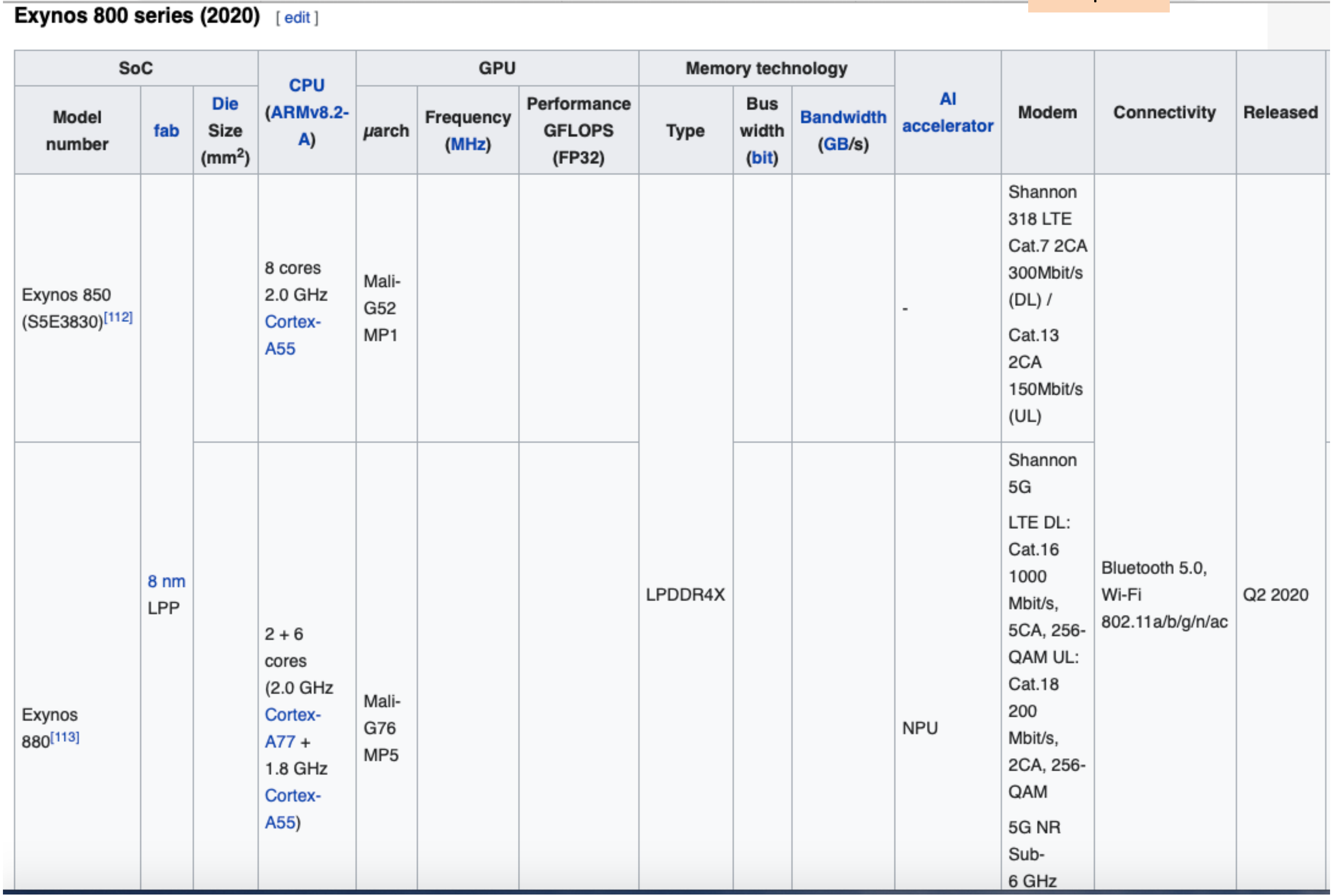
SoC			CPU (ARMv8-A)	GPU			Memory technology			Modem	Connectivity	Released
Model number	fab	Die Size (mm²)		µarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)			
Exynos 9609[108]	10 nm LPE[109]		4 + 4 cores (2.2 GHz Cortex-A73 + 1.6 GHz Cortex-A53)	Mali-G72 MP3			LPDDR4X	64-bit (2x32-bit) Dual-channel	1600 MHz (11.9 GB/sec)	Shannon 337 LTE Cat.12 3CA 600Mbit/s (DL) / Cat.13 2CA 150Mbit/s (UL)	Bluetooth 5.0, Wi-Fi 802.11a/b/g/n/ac	Q2 2019
Exynos 9610[110]			4 + 4 cores (2.3 GHz Cortex-A73 + 1.7 GHz Cortex-A53)									Q4 2018
Exynos 9611[111]												Q3 2019

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP122

Exynos

DR JEFF SOFTWARE
INDIE APP DEVELOPER
© Jeff Drobman
2016-2022

Wikipedia



Exynos 800 series (2020) [edit]

SoC			CPU (ARMv8.2-A)	GPU			Memory technology			AI accelerator	Modem	Connectivity	Released
Model number	fab	Die Size (mm^2)		µarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)				
Exynos 850 (S5E3830)[112]	8 nm LPP		8 cores 2.0 GHz Cortex-A55	Mali-G52 MP1			LPDDR4X			-	Shannon 318 LTE Cat.7 2CA 300Mbit/s (DL) / Cat.13 2CA 150Mbit/s (UL)	Bluetooth 5.0, Wi-Fi 802.11a/b/g/n/ac	Q2 2020
Exynos 880[113]			2 + 6 cores (2.0 GHz Cortex-A77 + 1.8 GHz Cortex-A55)	Mali-G76 MP5						NPU	Shannon 5G LTE DL: Cat.16 1000 Mbit/s, 5CA, 256-QAM UL: Cat.18 200 Mbit/s, 2CA, 256-QAM 5G NR Sub-6 GHz		

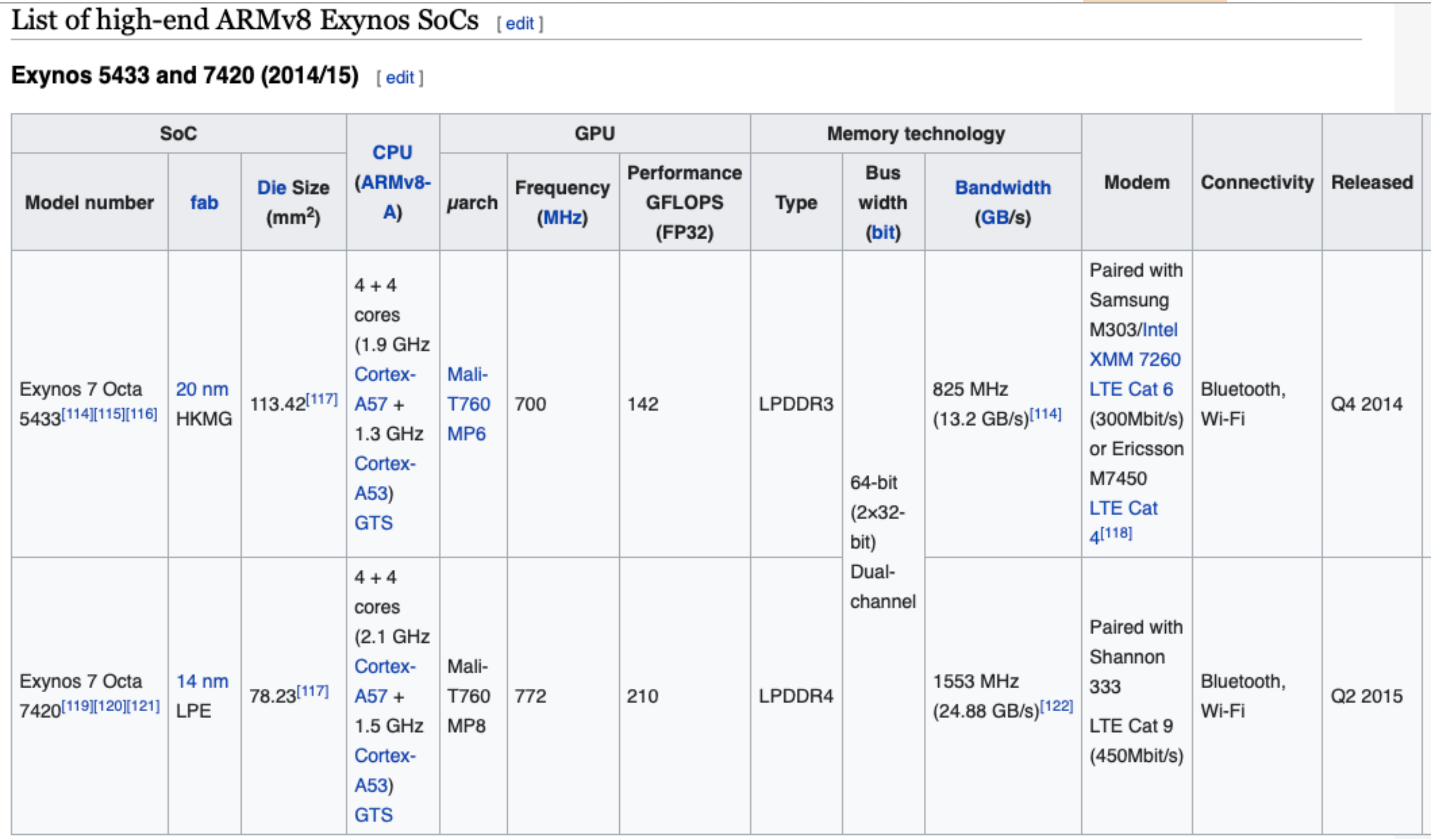
CSUN
CALIFORNIA
STATE UNIVERSITY
NORTHRIDGE

COMP122

Exynos

DR JEFF SOFTWARE
INDIE APP DEVELOPER
© Jeff Drobman
2016-2022

Wikipedia



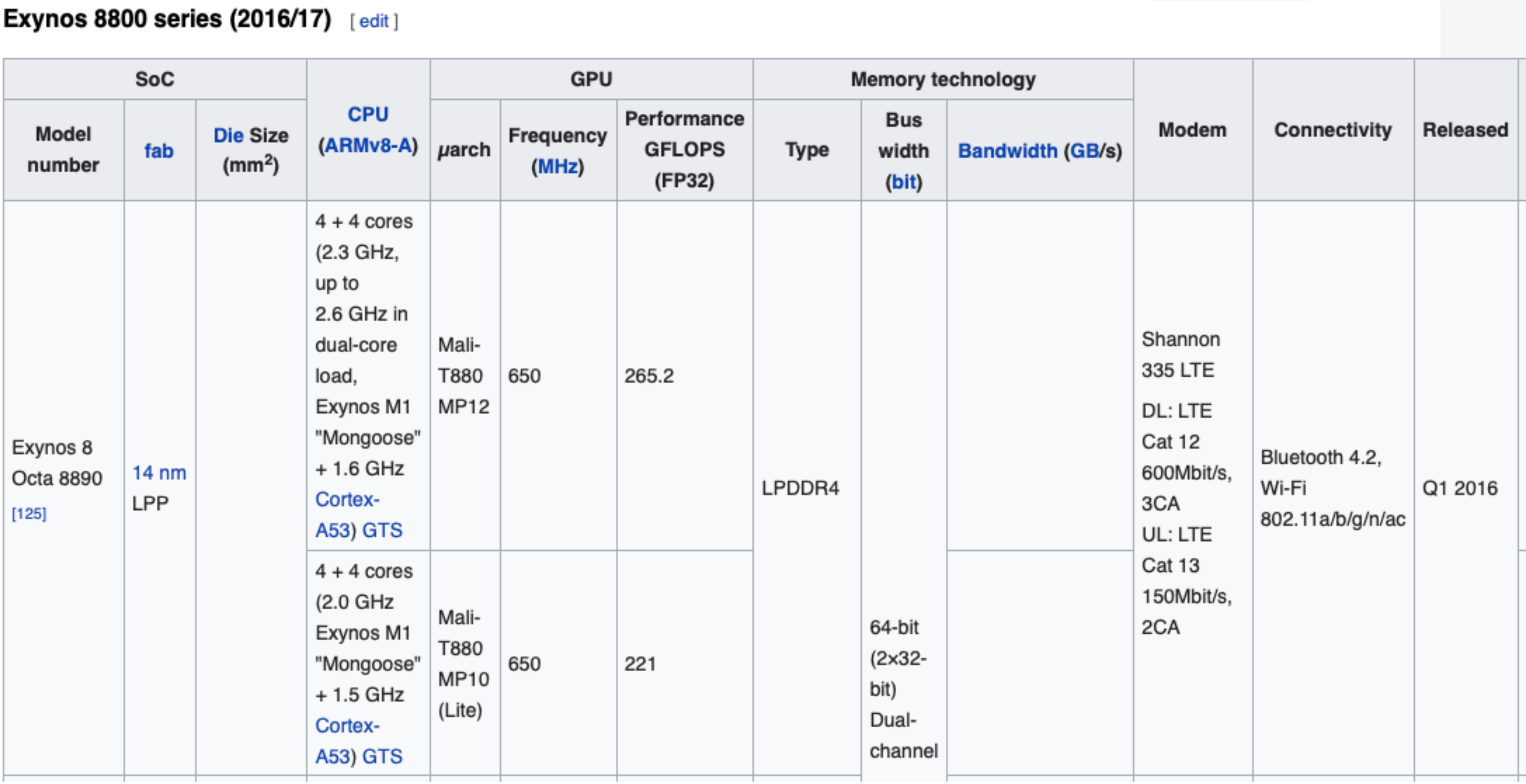
List of high-end ARMv8 Exynos SoCs [edit]

Exynos 5433 and 7420 (2014/15) [edit]

SoC			CPU (ARMv8-A)	GPU			Memory technology			Modem	Connectivity	Released
Model number	fab	Die Size (mm^2)		µarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)			
Exynos 7 Octa 5433[114][115][116]	20 nm HKMG	113.42[117]	4 + 4 cores (1.9 GHz Cortex-A57 + 1.3 GHz Cortex-A53) GTS	Mali-T760 MP6	700	142	LPDDR3	64-bit (2x32-bit) Dual-channel	825 MHz (13.2 GB/s)[114]	Paired with Samsung M303/Intel XMM 7260 LTE Cat 6 (300Mbit/s) or Ericsson M7450 LTE Cat 4[118]	Bluetooth, Wi-Fi	Q4 2014
Exynos 7 Octa 7420[119][120][121]	14 nm LPE	78.23[117]	4 + 4 cores (2.1 GHz Cortex-A57 + 1.5 GHz Cortex-A53) GTS	Mali-T760 MP8	772	210	LPDDR4		1553 MHz (24.88 GB/s)[122]	Paired with Shannon 333 LTE Cat 9 (450Mbit/s)	Bluetooth, Wi-Fi	Q2 2015

Exynos

Wikipedia



Exynos 8800 series (2016/17) [edit]

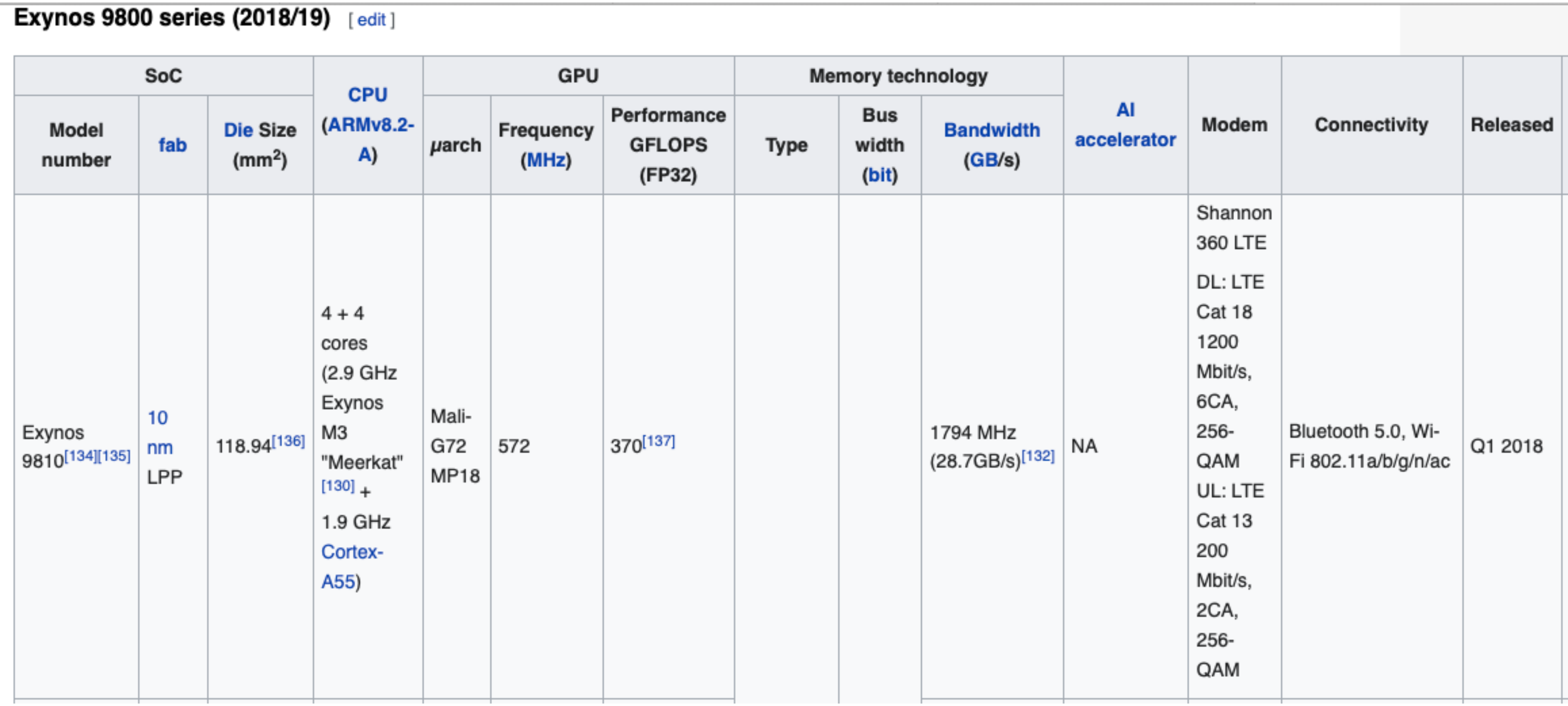
SoC			CPU (ARMv8-A)	GPU			Memory technology			Modem	Connectivity	Released
Model number	fab	Die Size (mm^2)		µarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)			
Exynos 8 Octa 8890 [125]	14 nm LPP		4 + 4 cores (2.3 GHz, up to 2.6 GHz in dual-core load, Exynos M1 "Mongoose" + 1.6 GHz Cortex-A53) GTS	Mali-T880 MP12	650	265.2	LPDDR4	64-bit (2x32-bit) Dual-channel		Shannon 335 LTE DL: LTE Cat 12 600Mbit/s, 3CA UL: LTE Cat 13 150Mbit/s, 2CA	Bluetooth 4.2, Wi-Fi 802.11a/b/g/n/ac	Q1 2016
			4 + 4 cores (2.0 GHz Exynos M1 "Mongoose" + 1.5 GHz Cortex-A53) GTS	Mali-T880 MP10 (Lite)	650	221						

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP122

Exynos

DR JEFF SOFTWARE
INDIE APP DEVELOPER
© Jeff Drobman
2016-2022

Wikipedia



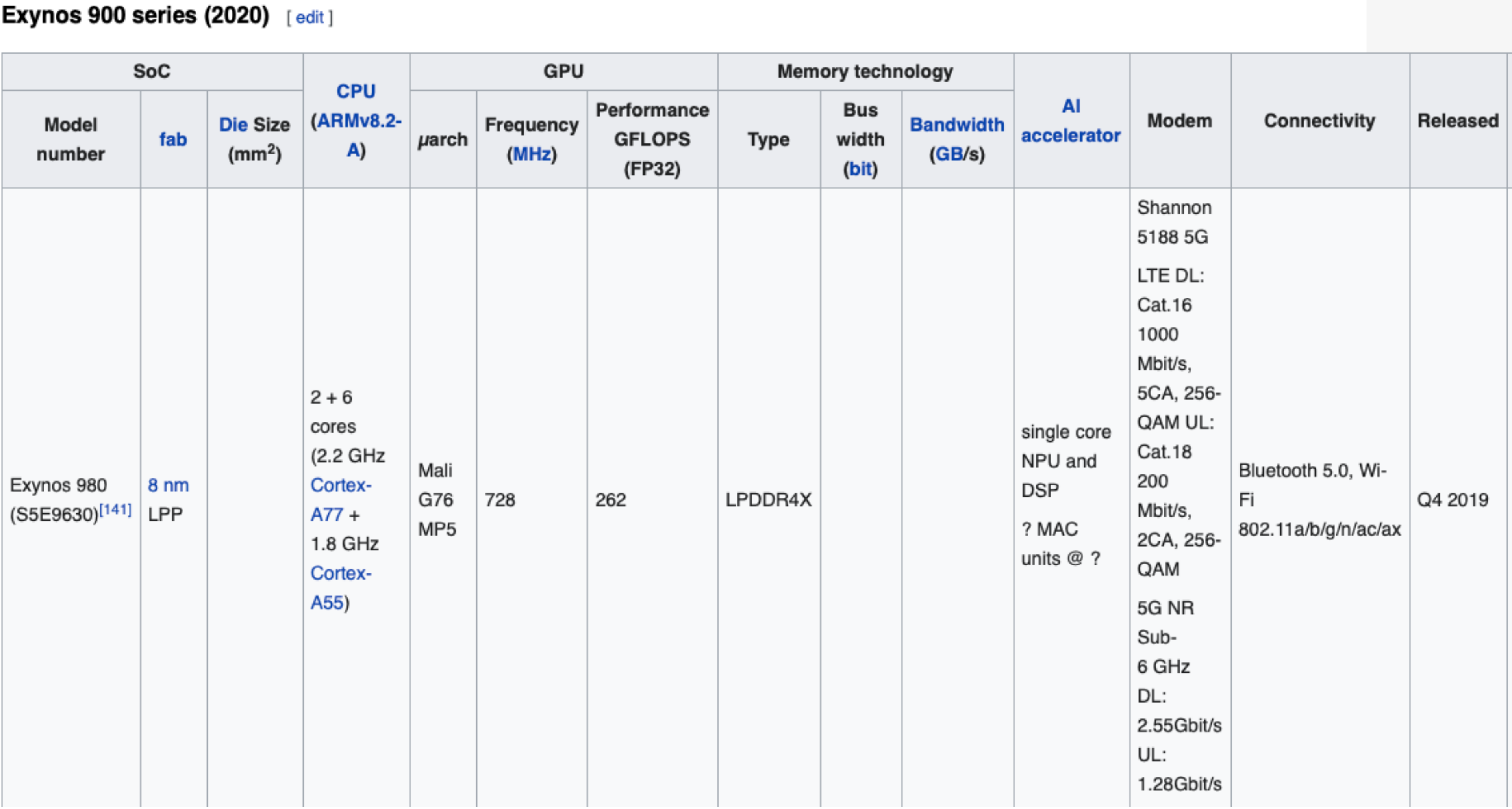
Exynos 9800 series (2018/19) [edit]

SoC			CPU (ARMv8.2-A)	GPU			Memory technology			AI accelerator	Modem	Connectivity	Released
Model number	fab	Die Size (mm^2)		μarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)				
Exynos 9810[134][135]	10 nm LPP	118.94[136]	4 + 4 cores (2.9 GHz Exynos M3 "Meerkat"[130] + 1.9 GHz Cortex-A55)	Mali-G72 MP18	572	370[137]			1794 MHz (28.7GB/s)[132]	NA	Shannon 360 LTE DL: LTE Cat 18 1200 Mbit/s, 6CA, 256-QAM UL: LTE Cat 13 200 Mbit/s, 2CA, 256-QAM	Bluetooth 5.0, Wi-Fi 802.11a/b/g/n/ac	Q1 2018

COMP122

Exynos

Wikipedia



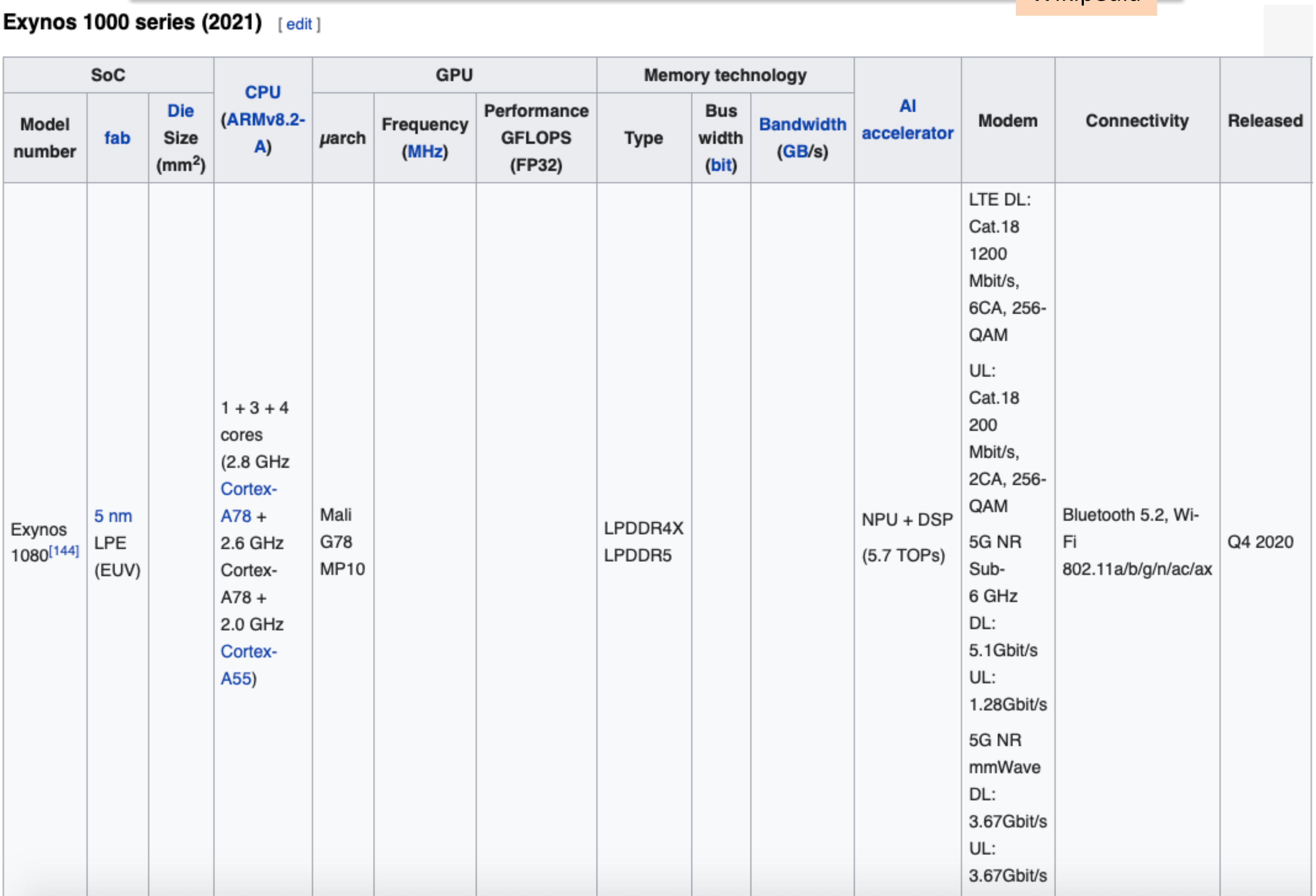
Exynos 900 series (2020)

[edit]

SoC			CPU (ARMv8.2-A)	GPU			Memory technology			AI accelerator	Modem	Connectivity	Released
Model number	fab	Die Size (mm²)		μarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)				
Exynos 980 (S5E9630)[141]	8 nm LPP		2 + 6 cores (2.2 GHz Cortex-A77 + 1.8 GHz Cortex-A55)	Mali G76 MP5	728	262	LPDDR4X			single core NPU and DSP ? MAC units @ ?	Shannon 5188 5G LTE DL: Cat.16 1000 Mbit/s, 5CA, 256-QAM UL: Cat.18 200 Mbit/s, 2CA, 256-QAM 5G NR Sub-6 GHz DL: 2.55Gbit/s UL: 1.28Gbit/s	Bluetooth 5.0, Wi-Fi 802.11a/b/g/n/ac/ax	Q4 2019

Exynos

Wikipedia

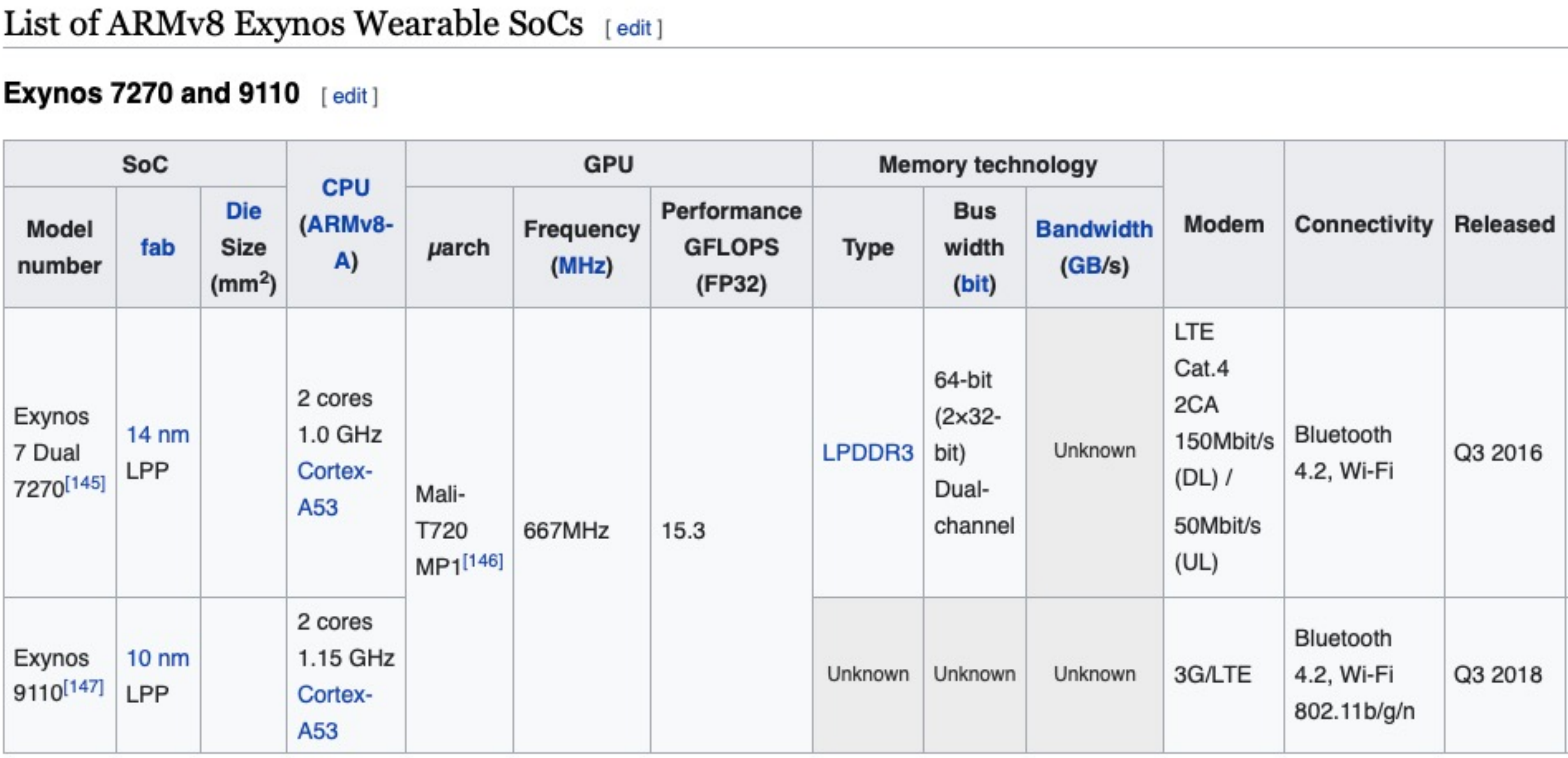


Exynos 1000 series (2021) [edit]

SoC			CPU (ARMv8.2-A)	GPU			Memory technology			AI accelerator	Modem	Connectivity	Released
Model number	fab	Die Size (mm^2)		µarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)				
Exynos 1080[144]	5 nm LPE (EUV)		1 + 3 + 4 cores (2.8 GHz Cortex-A78 + 2.6 GHz Cortex-A78 + 2.0 GHz Cortex-A55)	Mali G78 MP10			LPDDR4X LPDDR5			NPU + DSP (5.7 TOPs)	LTE DL: Cat.18 1200 Mbit/s, 6CA, 256-QAM UL: Cat.18 200 Mbit/s, 2CA, 256-QAM 5G NR Sub-6 GHz DL: 5.1Gbit/s UL: 1.28Gbit/s 5G NR mmWave DL: 3.67Gbit/s UL: 3.67Gbit/s	Bluetooth 5.2, Wi-Fi 802.11a/b/g/n/ac/ax	Q4 2020

Exynos

Wikipedia



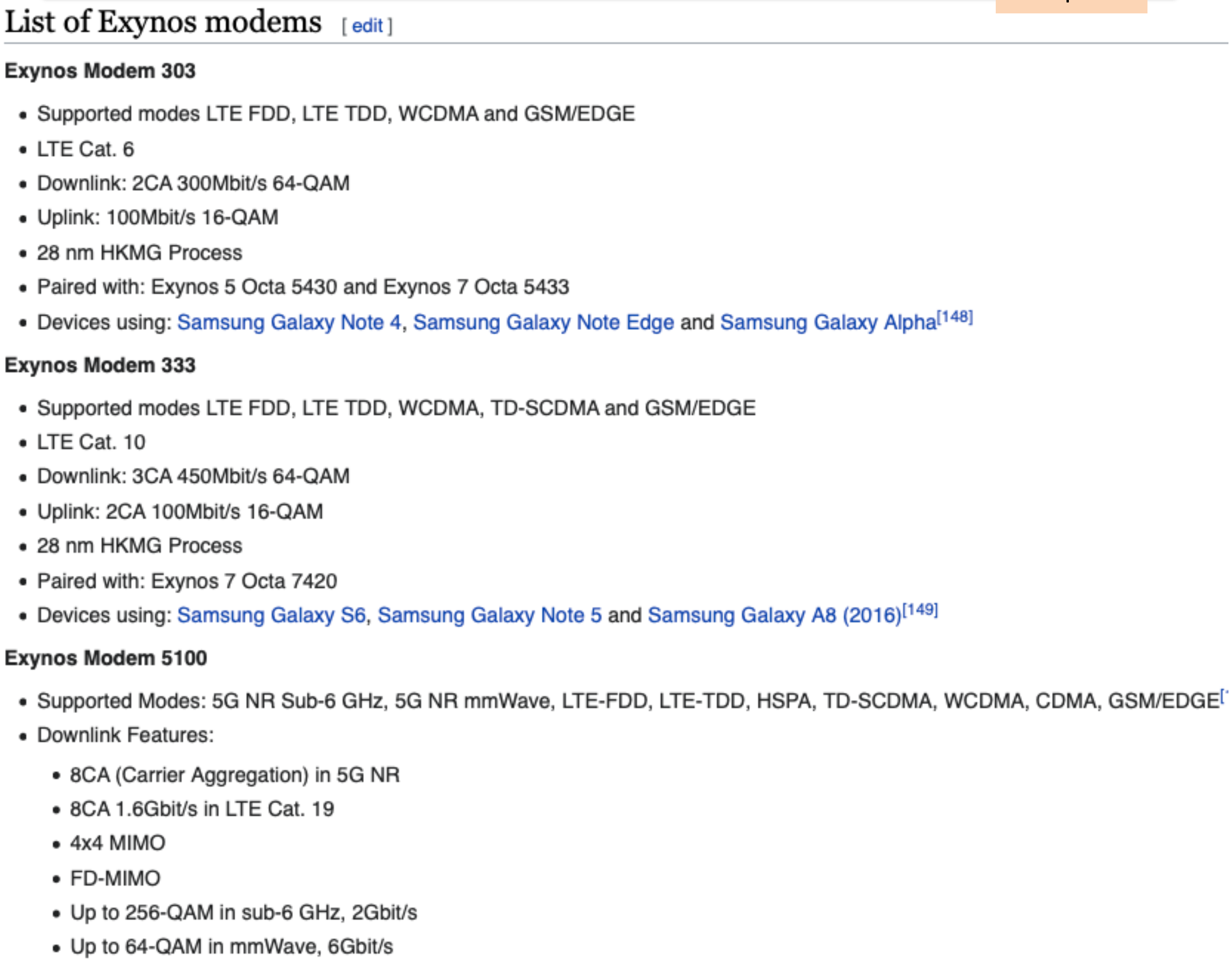
List of ARMv8 Exynos Wearable SoCs [edit]

Exynos 7270 and 9110 [edit]

SoC			CPU (ARMv8-A)	GPU			Memory technology			Modem	Connectivity	Released
Model number	fab	Die Size (mm^2)		μarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)			
Exynos 7 Dual 7270[145]	14 nm LPP		2 cores 1.0 GHz Cortex-A53	Mali-T720 MP1[146]	667MHz	15.3	LPDDR3	64-bit (2×32-bit) Dual-channel	Unknown	LTE Cat.4 2CA 150Mbit/s (DL) / 50Mbit/s (UL)	Bluetooth 4.2, Wi-Fi	Q3 2016
Exynos 9110[147]	10 nm LPP		2 cores 1.15 GHz Cortex-A53				Unknown	Unknown	Unknown	3G/LTE	Bluetooth 4.2, Wi-Fi 802.11b/g/n	Q3 2018

Exynos

Wikipedia



List of Exynos modems [edit]

Exynos Modem 303

- Supported modes LTE FDD, LTE TDD, WCDMA and GSM/EDGE
- LTE Cat. 6
- Downlink: 2CA 300Mbit/s 64-QAM
- Uplink: 100Mbit/s 16-QAM
- 28 nm HKMG Process
- Paired with: Exynos 5 Octa 5430 and Exynos 7 Octa 5433
- Devices using: Samsung Galaxy Note 4, Samsung Galaxy Note Edge and Samsung Galaxy Alpha[148]

Exynos Modem 333

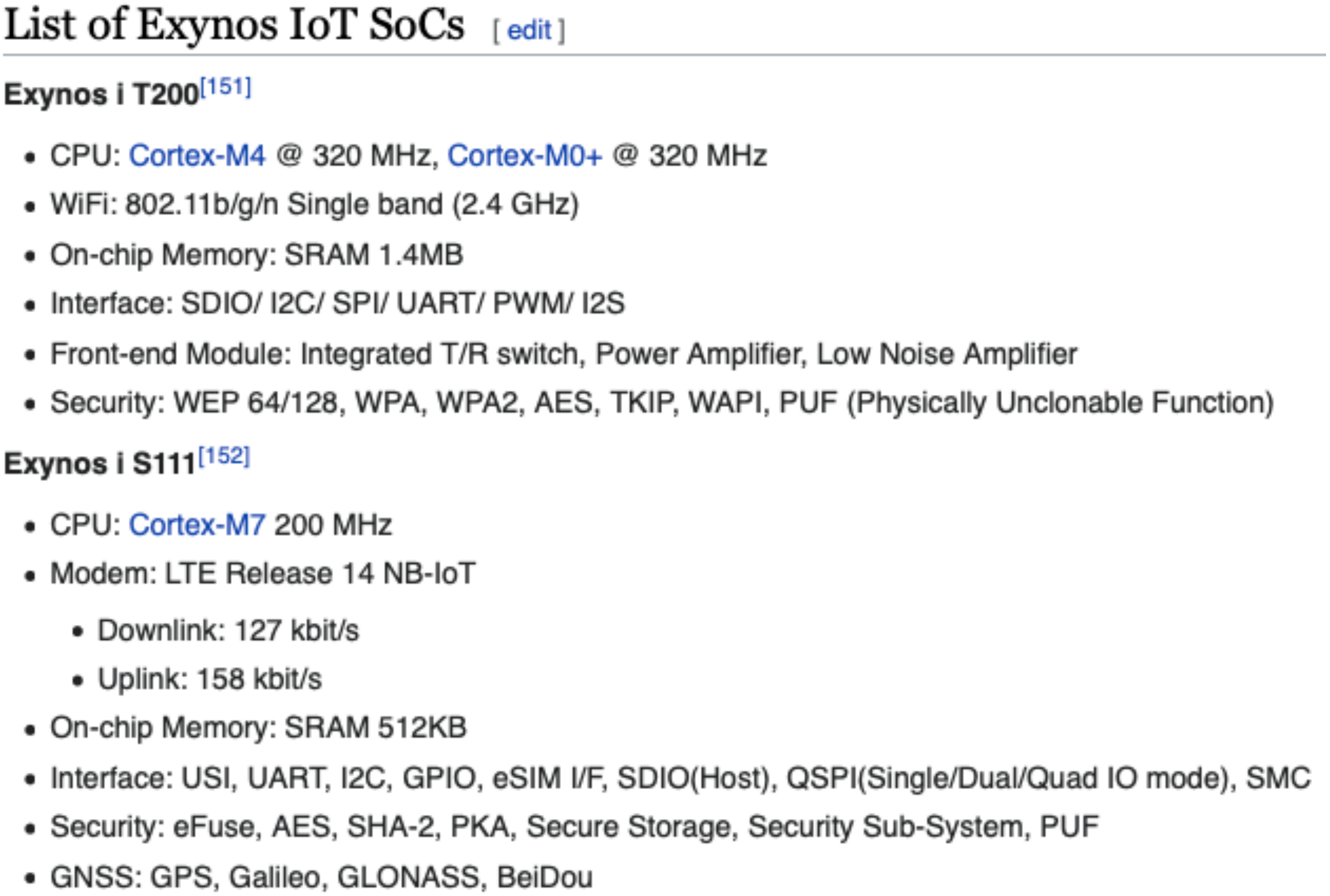
- Supported modes LTE FDD, LTE TDD, WCDMA, TD-SCDMA and GSM/EDGE
- LTE Cat. 10
- Downlink: 3CA 450Mbit/s 64-QAM
- Uplink: 2CA 100Mbit/s 16-QAM
- 28 nm HKMG Process
- Paired with: Exynos 7 Octa 7420
- Devices using: Samsung Galaxy S6, Samsung Galaxy Note 5 and Samsung Galaxy A8 (2016)[149]

Exynos Modem 5100

- Supported Modes: 5G NR Sub-6 GHz, 5G NR mmWave, LTE-FDD, LTE-TDD, HSPA, TD-SCDMA, WCDMA, CDMA, GSM/EDGE[
- Downlink Features:
  - 8CA (Carrier Aggregation) in 5G NR
  - 8CA 1.6Gbit/s in LTE Cat. 19
  - 4x4 MIMO
  - FD-MIMO
  - Up to 256-QAM in sub-6 GHz, 2Gbit/s
  - Up to 64-QAM in mmWave, 6Gbit/s

Exynos

Wikipedia



List of Exynos IoT SoCs [edit]

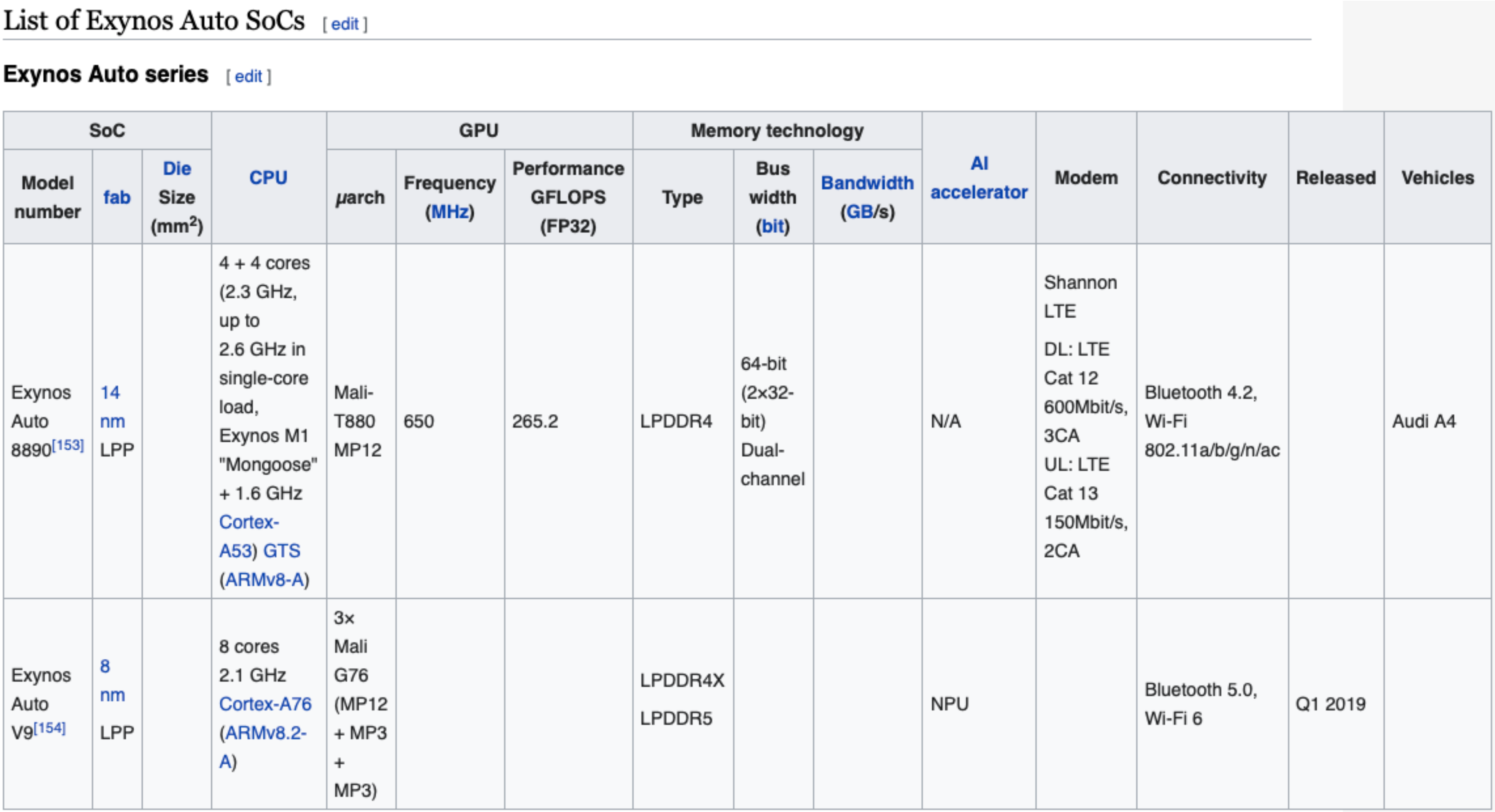
Exynos i T200[151]

- CPU: Cortex-M4 @ 320 MHz, Cortex-M0+ @ 320 MHz
- WiFi: 802.11b/g/n Single band (2.4 GHz)
- On-chip Memory: SRAM 1.4MB
- Interface: SDIO/ I2C/ SPI/ UART/ PWM/ I2S
- Front-end Module: Integrated T/R switch, Power Amplifier, Low Noise Amplifier
- Security: WEP 64/128, WPA, WPA2, AES, TKIP, WAPI, PUF (Physically Unclonable Function)

Exynos i S111[152]

- CPU: Cortex-M7 200 MHz
- Modem: LTE Release 14 NB-IoT
  - Downlink: 127 kbit/s
  - Uplink: 158 kbit/s
- On-chip Memory: SRAM 512KB
- Interface: USI, UART, I2C, GPIO, eSIM I/F, SDIO(Host), QSPI(Single/Dual/Quad IO mode), SMC
- Security: eFuse, AES, SHA-2, PKA, Secure Storage, Security Sub-System, PUF
- GNSS: GPS, Galileo, GLONASS, BeiDou

Exynos



List of Exynos Auto SoCs [edit]

Exynos Auto series [edit]

SoC			CPU	GPU			Memory technology			AI accelerator	Modem	Connectivity	Released	Vehicles
Model number	fab	Die Size (mm^2)		µarch	Frequency (MHz)	Performance GFLOPS (FP32)	Type	Bus width (bit)	Bandwidth (GB/s)					
Exynos Auto 8890[153]	14 nm LPP		4 + 4 cores (2.3 GHz, up to 2.6 GHz in single-core load, Exynos M1 "Mongoose" + 1.6 GHz Cortex-A53) GTS (ARMv8-A)	Mali-T880 MP12	650	265.2	LPDDR4	64-bit (2x32-bit) Dual-channel		N/A	Shannon LTE DL: LTE Cat 12 600Mbit/s, 3CA UL: LTE Cat 13 150Mbit/s, 2CA	Bluetooth 4.2, Wi-Fi 802.11a/b/g/n/ac		Audi A4
Exynos Auto V9[154]	8 nm LPP		8 cores 2.1 GHz Cortex-A76 (ARMv8.2-A)	3x Mali G76 (MP12 + MP3 + MP3)			LPDDR4X LPDDR5			NPU		Bluetooth 5.0, Wi-Fi 6	Q1 2019	

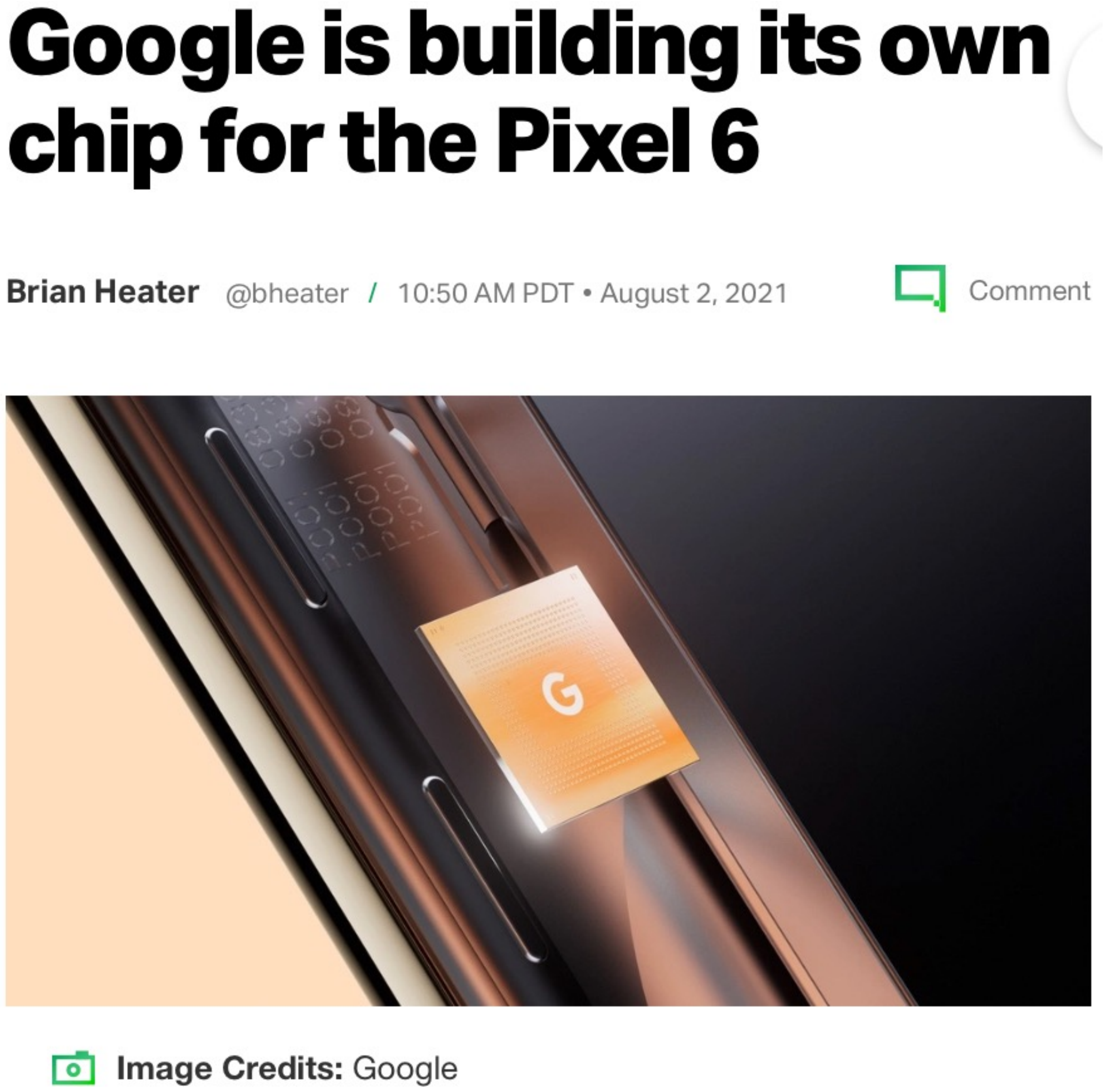
Chip Fab

Google

- *Pixel 6 Tensor*
- *TPU*

Google

Join Extra Crunch



Google is building its own chip for the Pixel 6

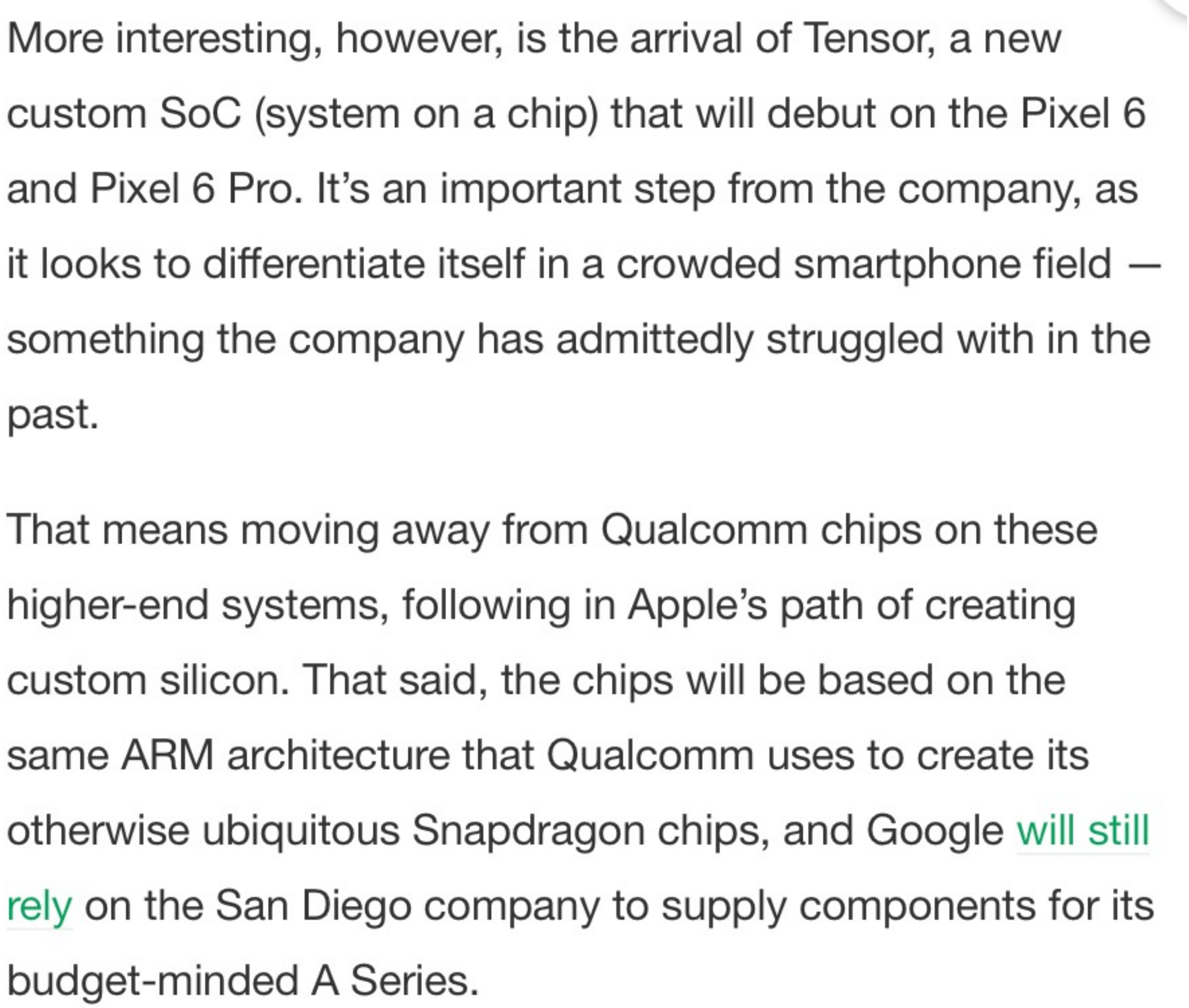
Brian Heater @bheater / 10:50 AM PDT • August 2, 2021

Comment

Image Credits: Google

Join Extra Crunch

Google



More interesting, however, is the arrival of Tensor, a new custom SoC (system on a chip) that will debut on the Pixel 6 and Pixel 6 Pro. It's an important step from the company, as it looks to differentiate itself in a crowded smartphone field — something the company has admittedly struggled with in the past.

That means moving away from Qualcomm chips on these higher-end systems, following in Apple's path of creating custom silicon. That said, the chips will be based on the same ARM architecture that Qualcomm uses to create its otherwise ubiquitous Snapdragon chips, and Google will still rely on the San Diego company to supply components for its budget-minded A Series.

Google Chips

One of the key challenges of computer design is how to pack chips and wiring in the most ergonomic fashion, maintaining power, speed and energy efficiency.

The recipe includes thousands of components that must communicate with one another flawlessly, all on a piece of real estate the size of a fingernail.

The process is known as **chip floor planning**, similar to what interior decorators do when laying out plans to dress up a room.

With digital circuitry, however, instead of using a one-floor plan, designers must consider integrated layouts within **multiple floors**.

As one tech publication referred to it recently, chip floor planning is **3-D Tetris**.

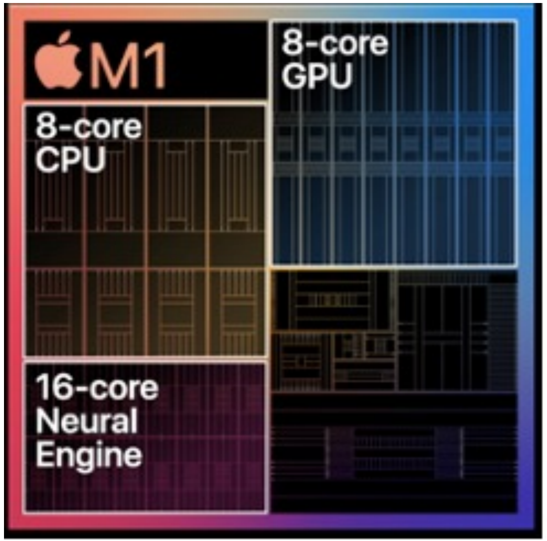
The process is time-consuming. And with continual improvement in chip components, laboriously calculated final designs become outdated fast. Chips are generally designed to last between two and five years, but there is constant pressure to shorten the time between upgrades.

Google Chips

AI/ML

Google researchers have just taken a giant leap in floor planning design. In a recent announcement, senior Google research engineers Anna Goldie and Azalia Mirhoseini said they have designed an algorithm that "learns" how to achieve optimum circuitry placement. It can do so in a fraction of the time currently required for such designing, analyzing potentially millions of possibilities instead of thousands, which is currently the norm. In doing so, it can provide chips that take advantage of the latest developments faster, cheaper and smaller.

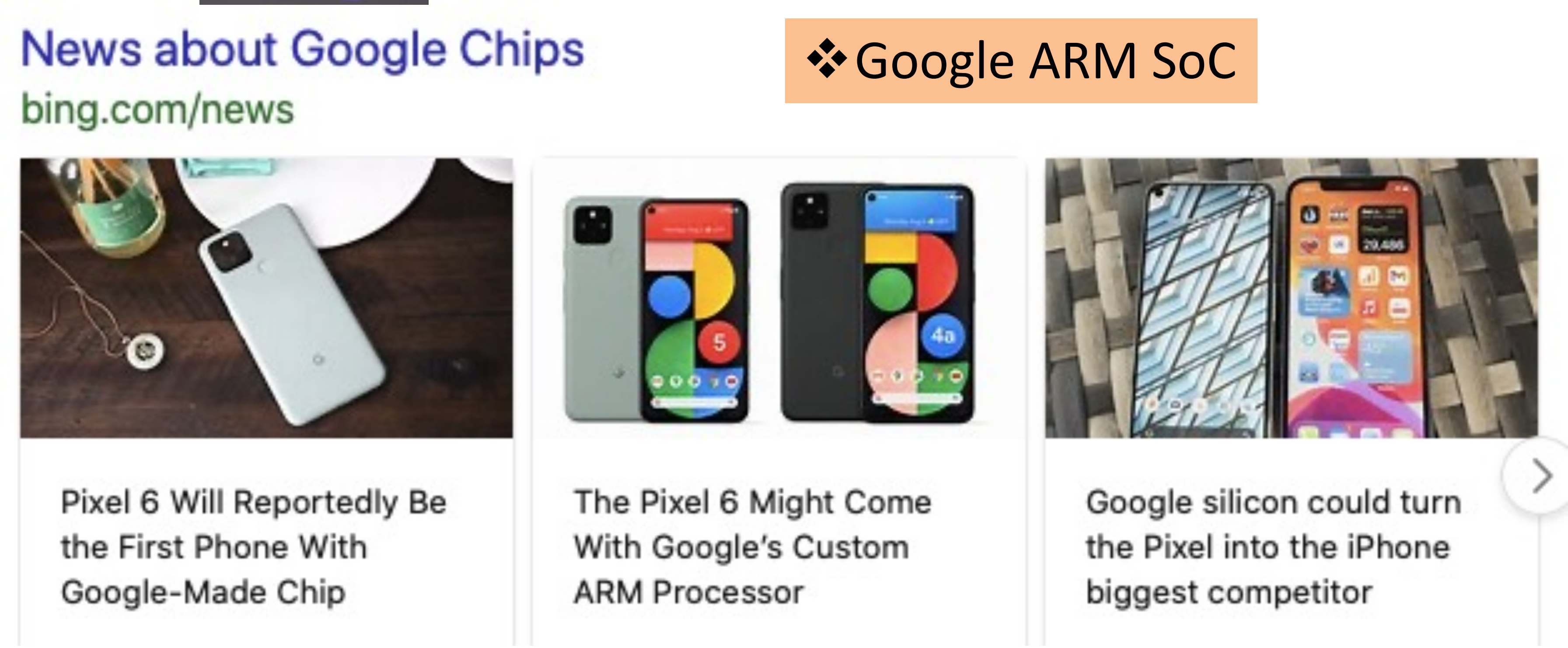
Goldie and Mirhoseini applied the concept of reinforcement learning to the new algorithm. The system generates "rewards" and "punishments" for each proposed design until the algorithm better recognizes the best approaches.



- ❖ Floor Planning

Google Chips

❖Google ARM SoC



News about Google Chips

bing.com/news

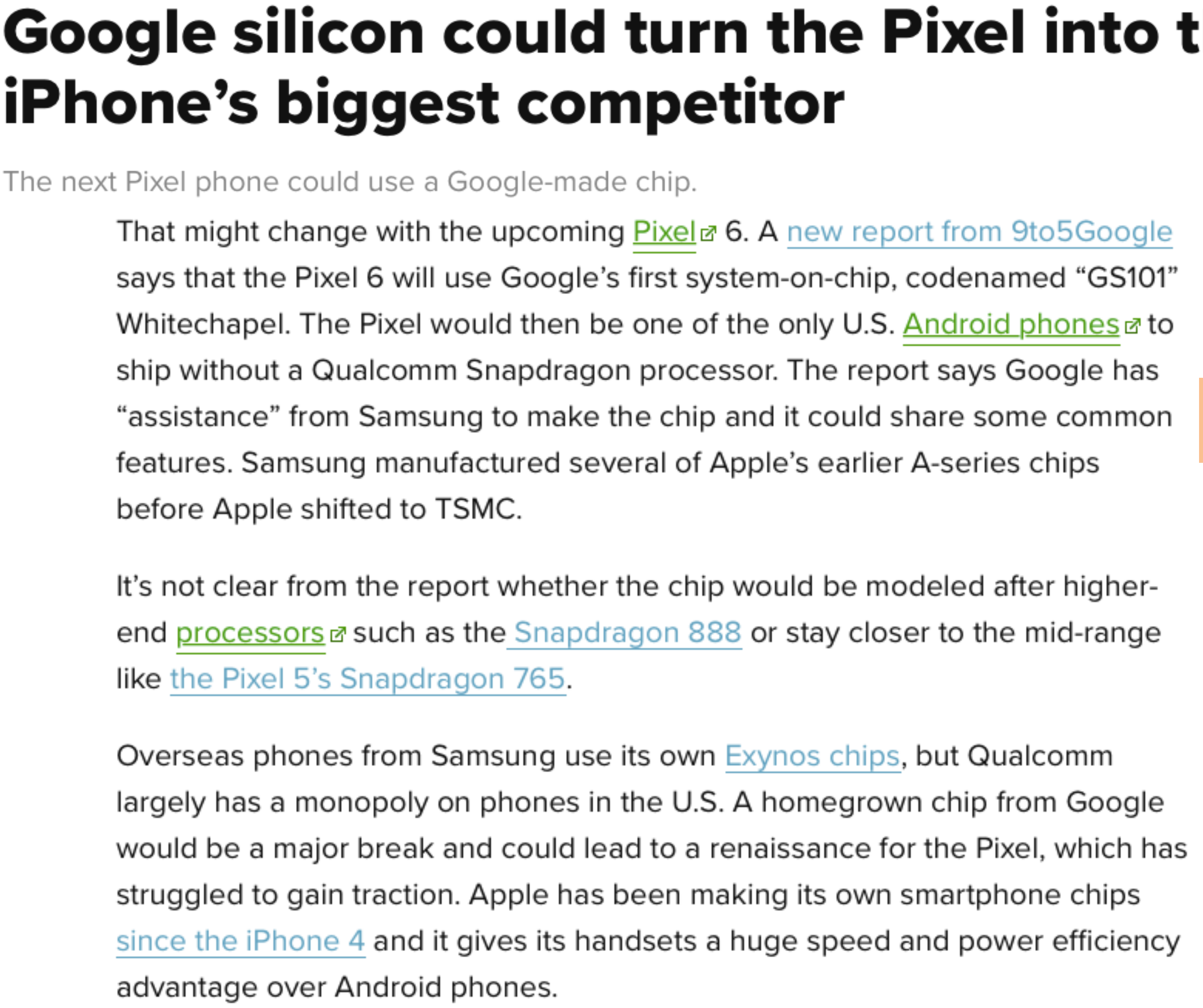
Pixel 6 Will Reportedly Be the First Phone With Google-Made Chip

The Pixel 6 Might Come With Google's Custom ARM Processor

Google silicon could turn the Pixel into the iPhone biggest competitor

Google Chips

Macworld



Google silicon could turn the Pixel into the iPhone's biggest competitor

The next Pixel phone could use a Google-made chip.

That might change with the upcoming Pixel 6. A new report from 9to5Google says that the Pixel 6 will use Google's first system-on-chip, codenamed "GS101" Whitechapel. The Pixel would then be one of the only U.S. Android phones to ship without a Qualcomm Snapdragon processor. The report says Google has "assistance" from Samsung to make the chip and it could share some common features. Samsung manufactured several of Apple's earlier A-series chips before Apple shifted to TSMC.

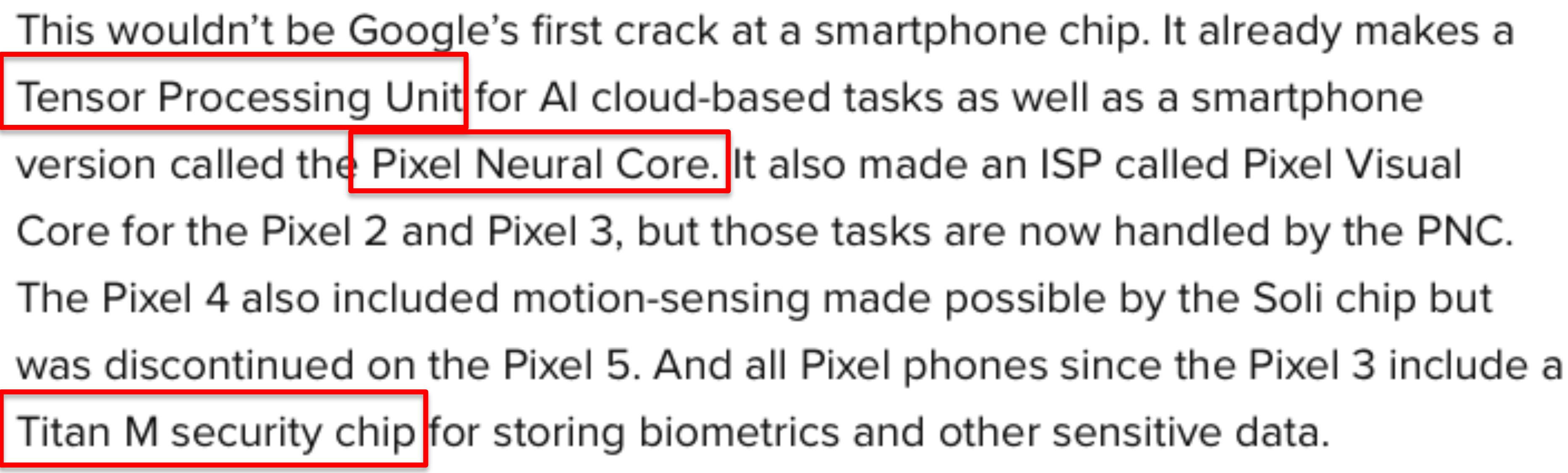
It's not clear from the report whether the chip would be modeled after higher-end processors such as the Snapdragon 888 or stay closer to the mid-range like the Pixel 5's Snapdragon 765.

Overseas phones from Samsung use its own Exynos chips, but Qualcomm largely has a monopoly on phones in the U.S. A homegrown chip from Google would be a major break and could lead to a renaissance for the Pixel, which has struggled to gain traction. Apple has been making its own smartphone chips since the iPhone 4 and it gives its handsets a huge speed and power efficiency advantage over Android phones.

- ❖ Google ARM SoC
- ❖ Snapdragon
- ❖ Samsung Exynos

Google Chips

Macworld



This wouldn't be Google's first crack at a smartphone chip. It already makes a Tensor Processing Unit for AI cloud-based tasks as well as a smartphone version called the Pixel Neural Core. It also made an ISP called Pixel Visual Core for the Pixel 2 and Pixel 3, but those tasks are now handled by the PNC. The Pixel 4 also included motion-sensing made possible by the Soli chip but was discontinued on the Pixel 5. And all Pixel phones since the Pixel 3 include a Titan M security chip for storing biometrics and other sensitive data.

- **TPU**
- PNC
- Titan M

Google Chips

The first Google chip will probably pack an **octa-core ARM CPU** with two Cortex-**A78**, two Cortex-**A76** and four Cortex-**A55** cores alongside **ARM Mali GPU** based on Samsung's 5nm manufacturing process.

Google *Whitechapel* is going to be an upper mid-range chip compared with Qualcomm's Snapdragon 700 series SoCs. Custom silicon will benefit Google over driver updates. It's not the first time the company is trying to come up with it's own chip. Google earlier collaborated with Intel in 2017 to develop *Pixel Visual Core* for the Pixel 2.

Google Chips

❖ 5G Modem

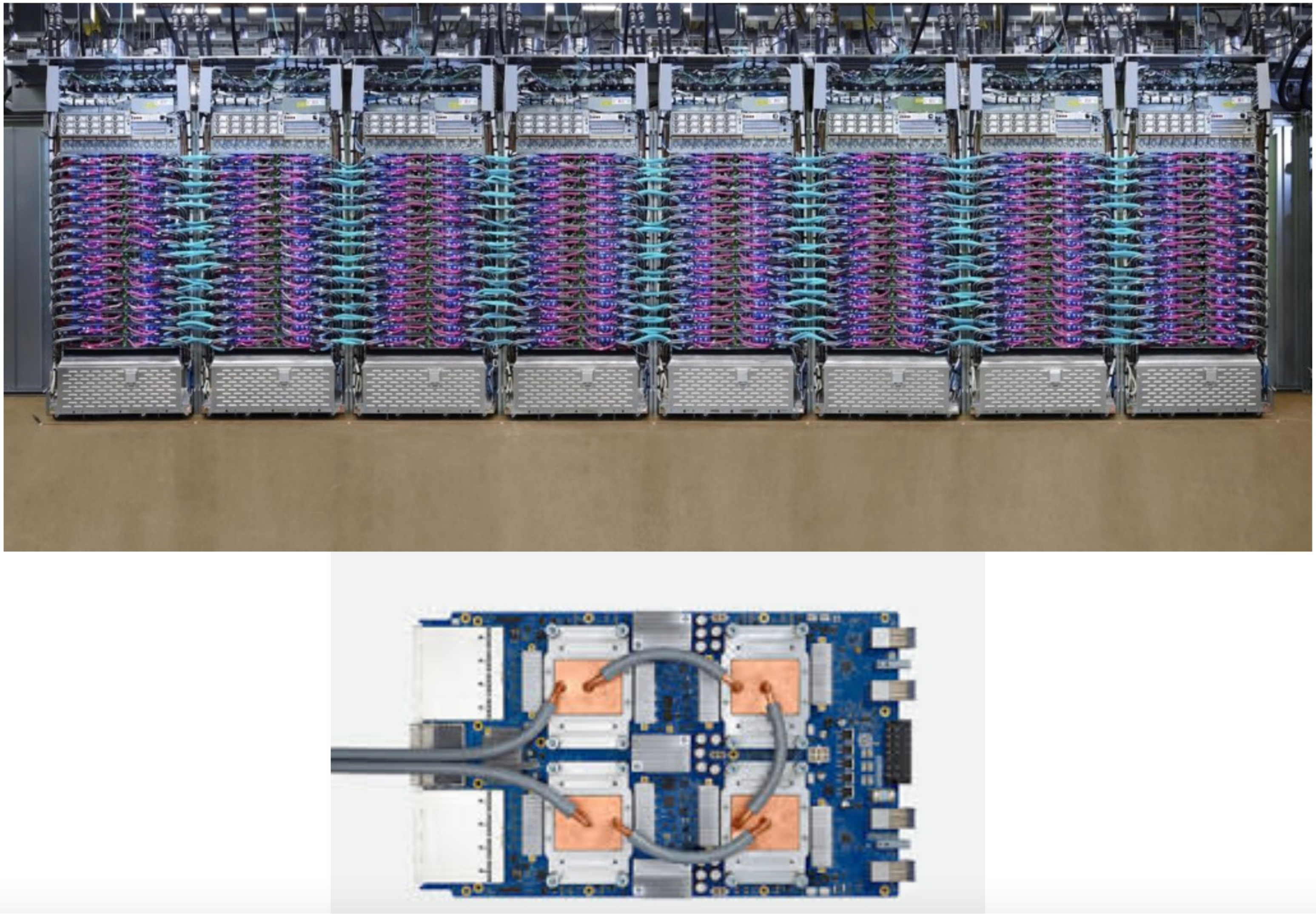
Google's probably going to have to use someone else's **5G modem**, and considering that Google is very closely tied to Verizon, it will likely still be a Qualcomm modem, possibly an x60 or x55. However, if Google wants to be forward-thinking, it should be an x60 or x65.

TPU

P&H

Figure 6.12.2: A TPUv3 supercomputer consisting of up to 1024 chips (left) (COD Figure 6.24).

It is about 6 ft tall and 40 ft long. A TPUv3 board (right) has four chips and uses liquid cooling.



TPU

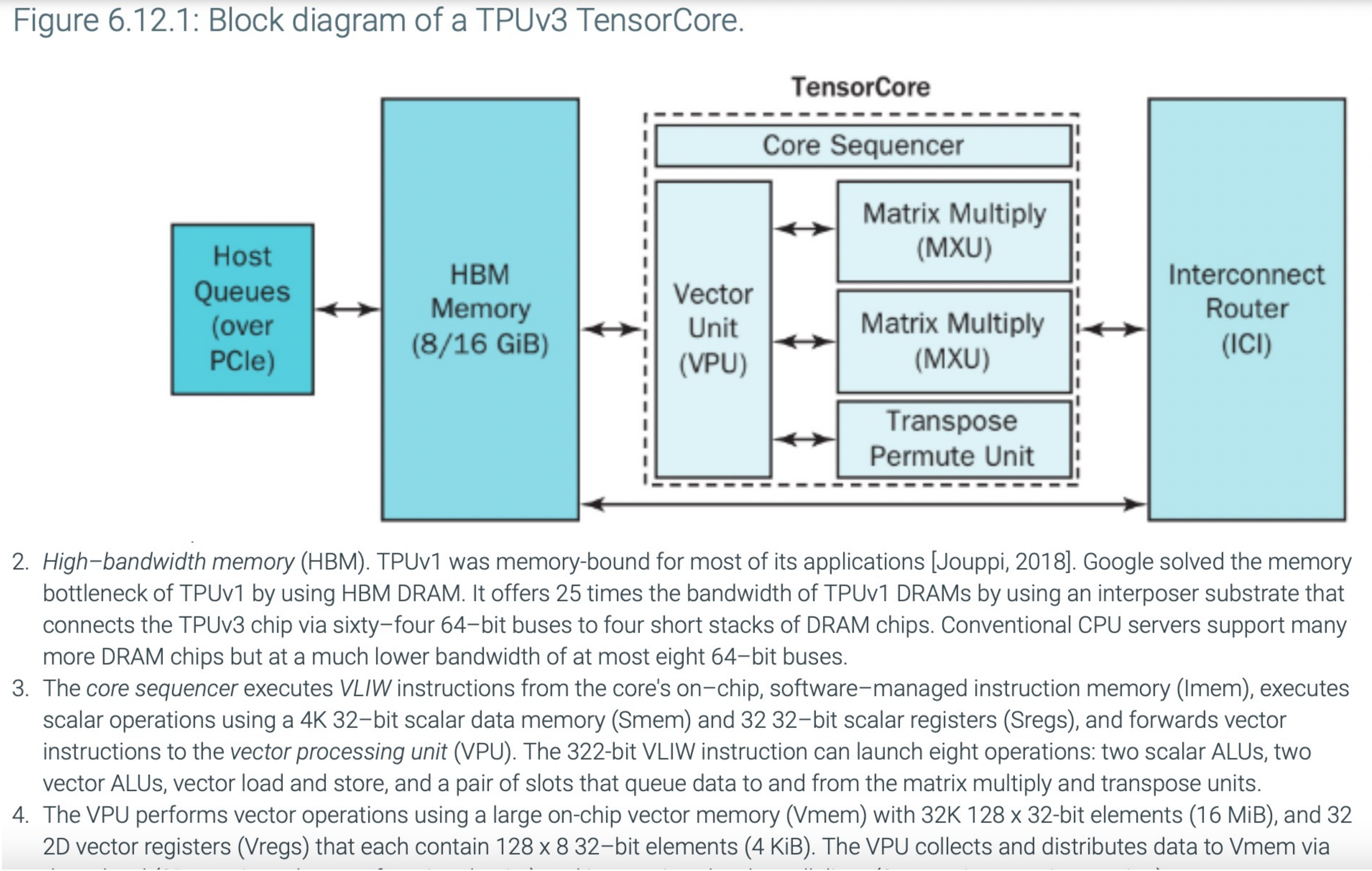
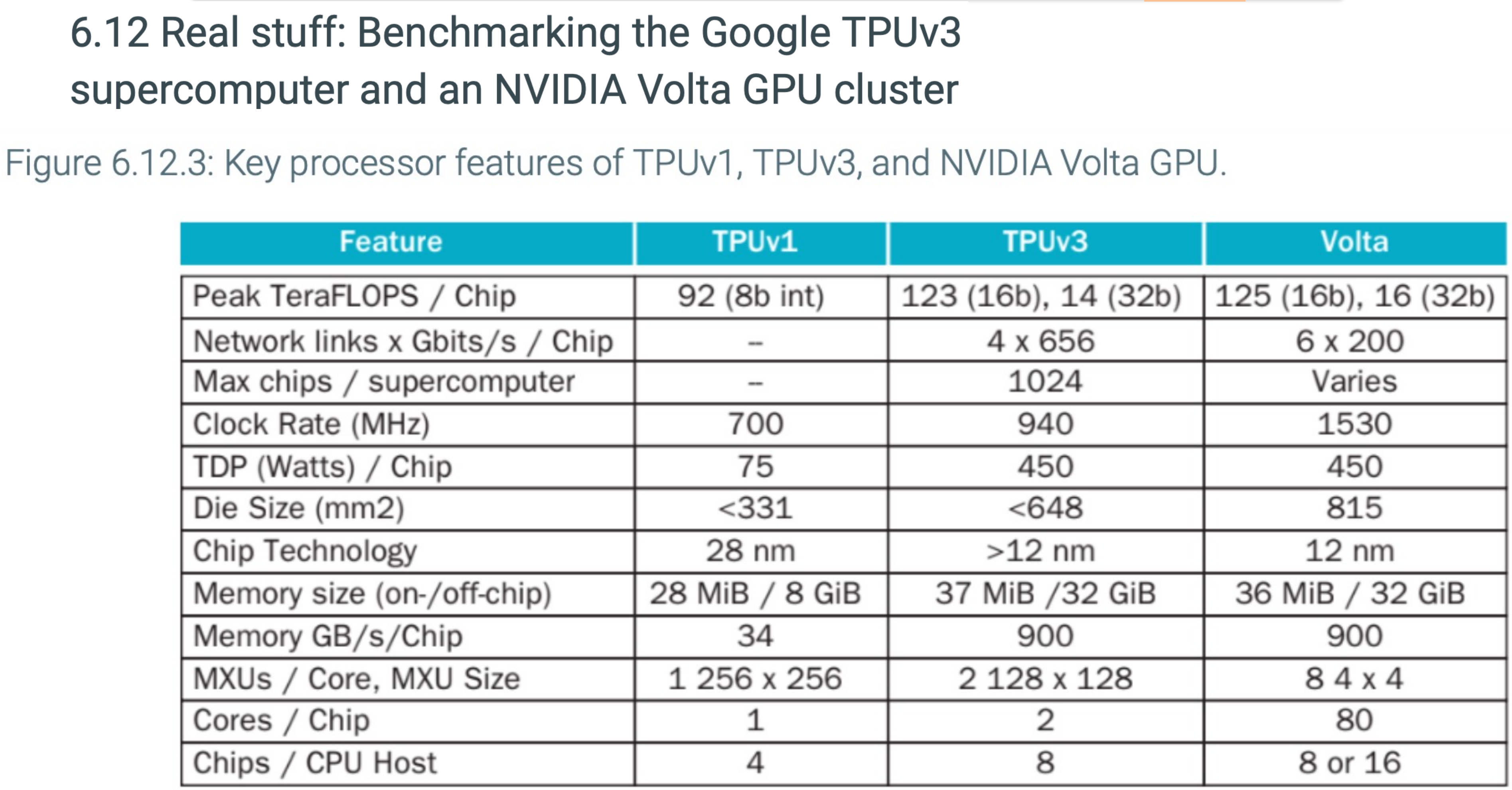


Figure 6.12.1: Block diagram of a TPUv3 TensorCore.

2. *High–bandwidth memory* (HBM). TPUv1 was memory-bound for most of its applications [Jouppi, 2018]. Google solved the memory bottleneck of TPUv1 by using HBM DRAM. It offers 25 times the bandwidth of TPUv1 DRAMs by using an interposer substrate that connects the TPUv3 chip via sixty–four 64–bit buses to four short stacks of DRAM chips. Conventional CPU servers support many more DRAM chips but at a much lower bandwidth of at most eight 64–bit buses.
3. The *core sequencer* executes *VLIW* instructions from the core's on–chip, software–managed instruction memory (Imem), executes scalar operations using a 4K 32–bit scalar data memory (Smem) and 32 32–bit scalar registers (Sregs), and forwards vector instructions to the *vector processing unit* (VPU). The 322-bit VLIW instruction can launch eight operations: two scalar ALUs, two vector ALUs, vector load and store, and a pair of slots that queue data to and from the matrix multiply and transpose units.
4. The VPU performs vector operations using a large on-chip vector memory (Vmem) with 32K 128 x 32-bit elements (16 MiB), and 32 2D vector registers (Vregs) that each contain 128 x 8 32–bit elements (4 KiB). The VPU collects and distributes data to Vmem via

TPU

P&H



6.12 Real stuff: Benchmarking the Google TPUv3 supercomputer and an NVIDIA Volta GPU cluster

Figure 6.12.3: Key processor features of TPUv1, TPUv3, and NVIDIA Volta GPU.

Feature	TPUv1	TPUv3	Volta
Peak TeraFLOPS / Chip	92 (8b int)	123 (16b), 14 (32b)	125 (16b), 16 (32b)
Network links x Gbits/s / Chip	--	4 x 656	6 x 200
Max chips / supercomputer	--	1024	Varies
Clock Rate (MHz)	700	940	1530
TDP (Watts) / Chip	75	450	450
Die Size (mm2)	<331	<648	815
Chip Technology	28 nm	>12 nm	12 nm
Memory size (on-/off-chip)	28 MiB / 8 GiB	37 MiB /32 GiB	36 MiB / 32 GiB
Memory GB/s/Chip	34	900	900
MXUs / Core, MXU Size	1 256 x 256	2 128 x 128	8 4 x 4
Cores / Chip	1	2	80
Chips / CPU Host	4	8	8 or 16

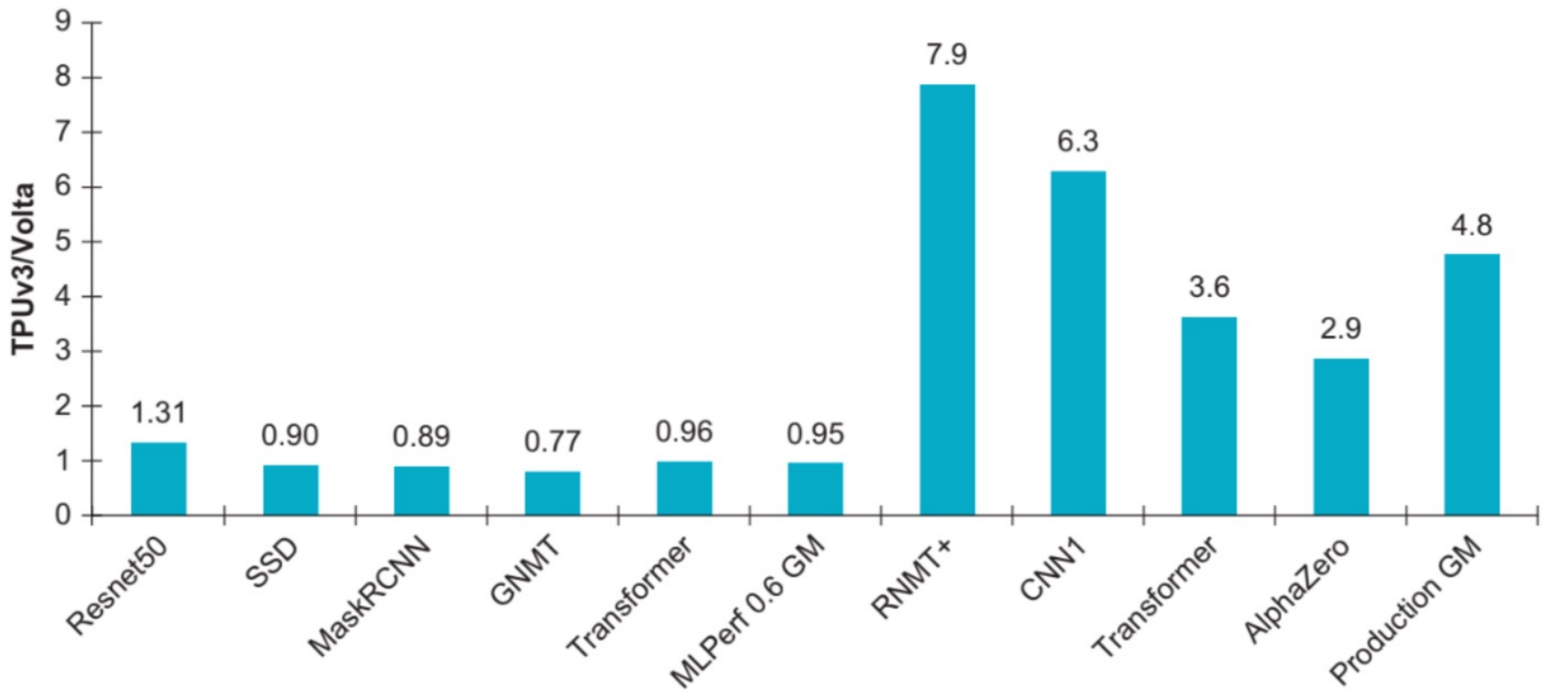
TPU

P&H

6.12 Real stuff: Benchmarking the Google TPUv3 supercomputer and an NVIDIA Volta GPU cluster

Figure 6.12.6

Performance per chip of TPUv3 relative to Volta for five MLPerf 0.6 benchmarks and four production applications.



Chip Fab

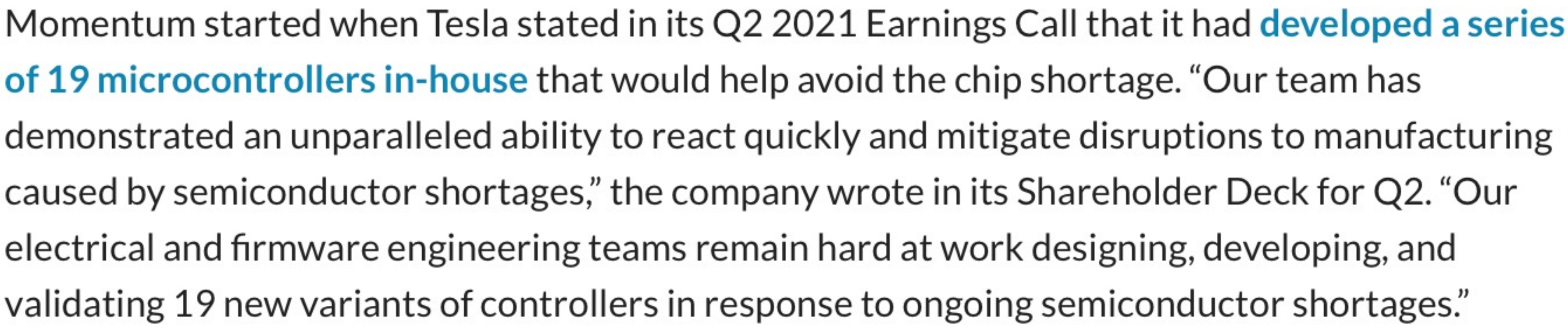
Tesla

- ❑ Supercomputer (Dojo)
- ❑ ARM SoC's

Tesla

TESLARATI 'How did Tesla find chips?'

19 new MCU's



Momentum started when Tesla stated in its Q2 2021 Earnings Call that it had **developed a series of 19 microcontrollers in-house** that would help avoid the chip shortage. "Our team has demonstrated an unparalleled ability to react quickly and mitigate disruptions to manufacturing caused by semiconductor shortages," the company wrote in its Shareholder Deck for Q2. "Our electrical and firmware engineering teams remain hard at work designing, developing, and validating 19 new variants of controllers in response to ongoing semiconductor shortages."

Tesla Chips

Quora

Did Tesla avert the semiconductor chip shortage with relative ease, while longstanding and prestigious automakers scaled back production due to the lack of supplies? Did Tesla keep itself stocked on the chips that keep its vehicles moving?

Jeff Drobman, Lecturer at California State University, Northridge (2016-present)

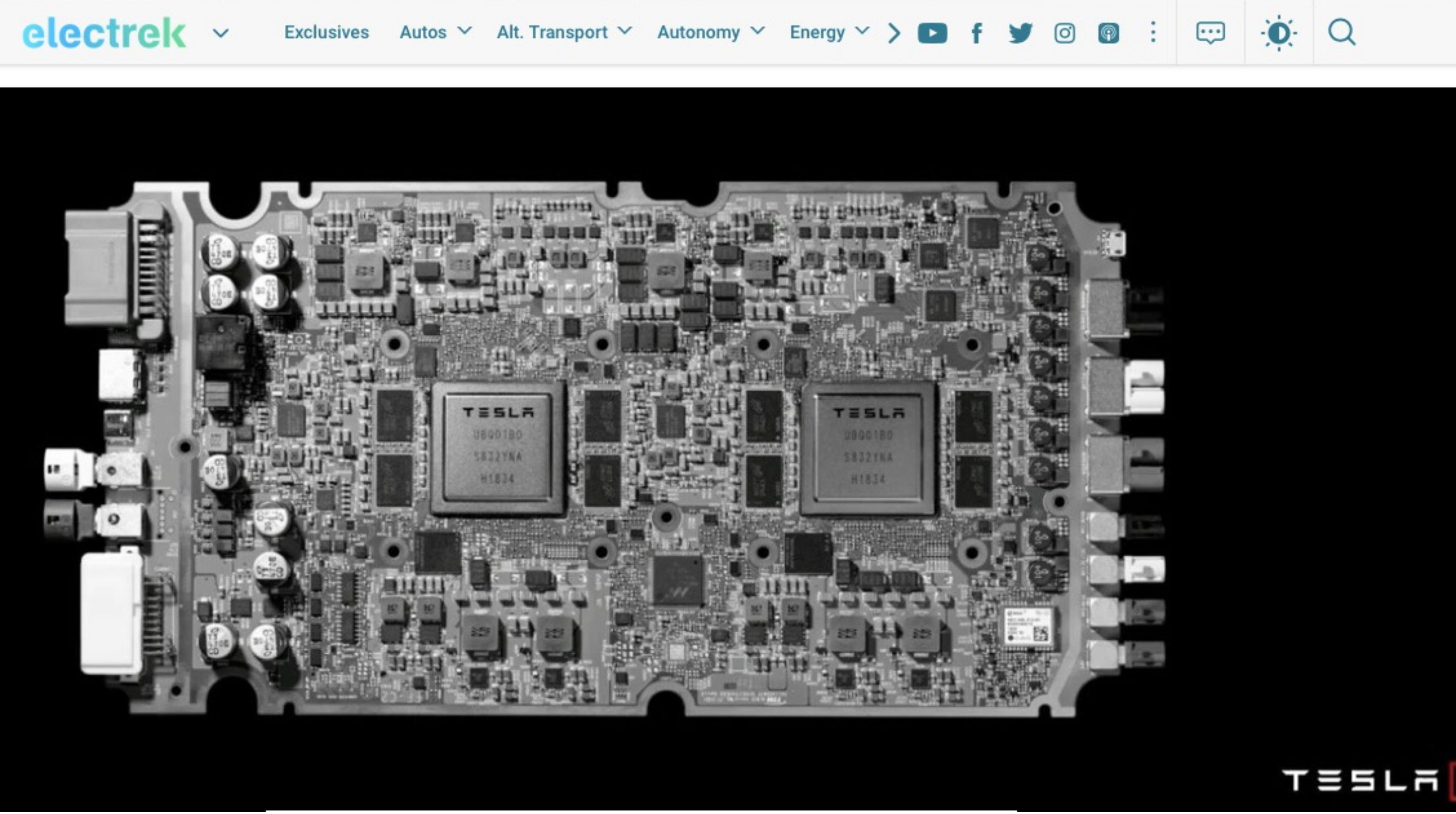
Answered just now

Yes. Tesla was prudent to well manage its chip inventory by purchasing months if not years out any critical chips. also, Tesla was wise to design their own microcontrollers (19) to better control its supply by choosing several chip manufacturers to supply them. also low-tech microcontrollers can be mass produced and cheaply — and they can be upgraded continuously via firmware downloads (software) over WiFi.

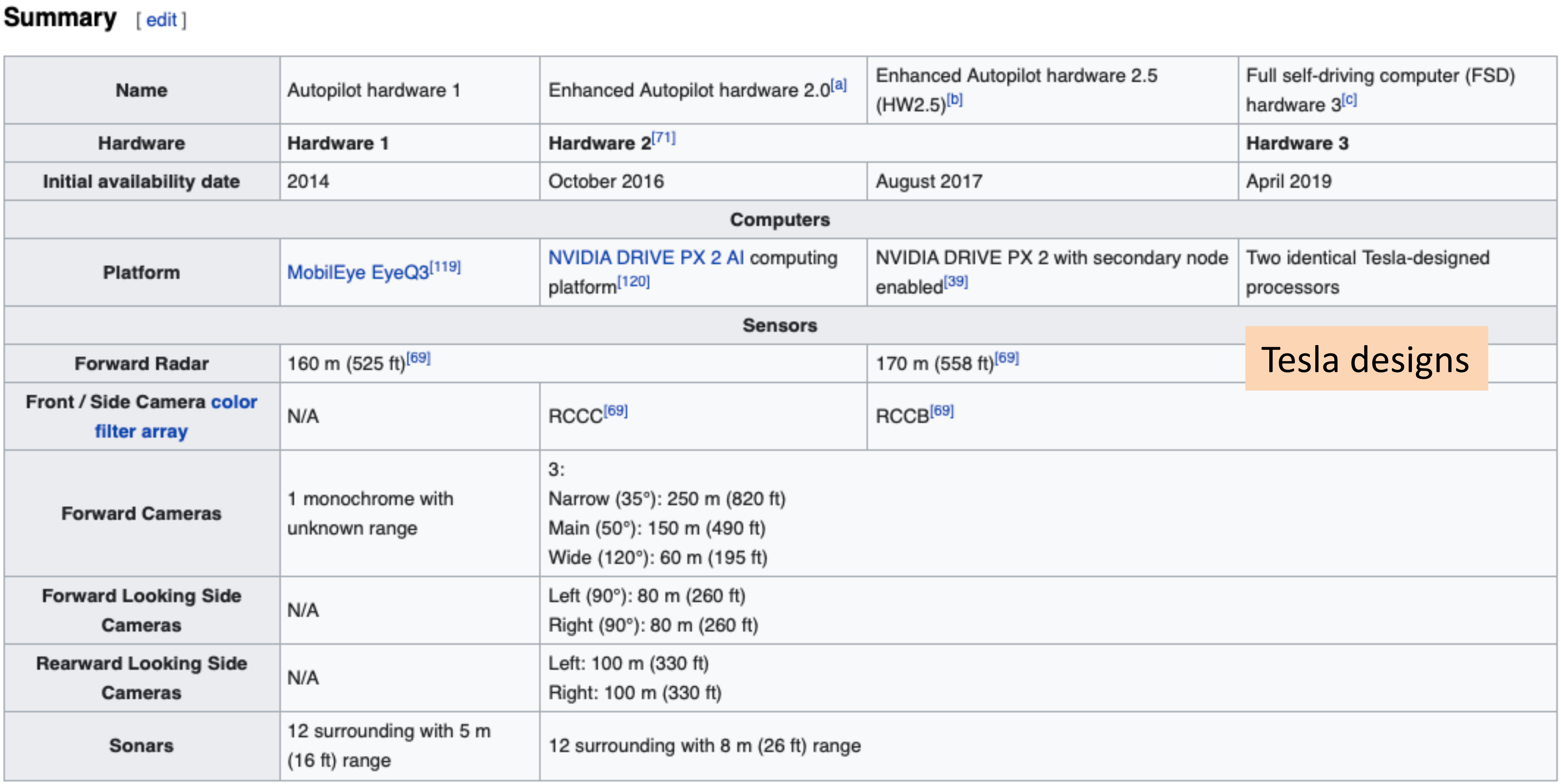
Tesla AI (Dojo SC)

TESLA AI DAY EXPECTED FOCUS:
AUTOPILOT IMPROVEMENT
ELON MUSK
TESLA CEO
TESLA EXPE
AI IMPROVI
TESLA EXPECTED TO HIGHLIGHT
AI IMPROVING MANUFACTURING

Tesla SoC


electrek
Exclusives
Autos
Alt. Transport
Autonomy
Energy
TESLA
TESLA
TESLA

Tesla Computers



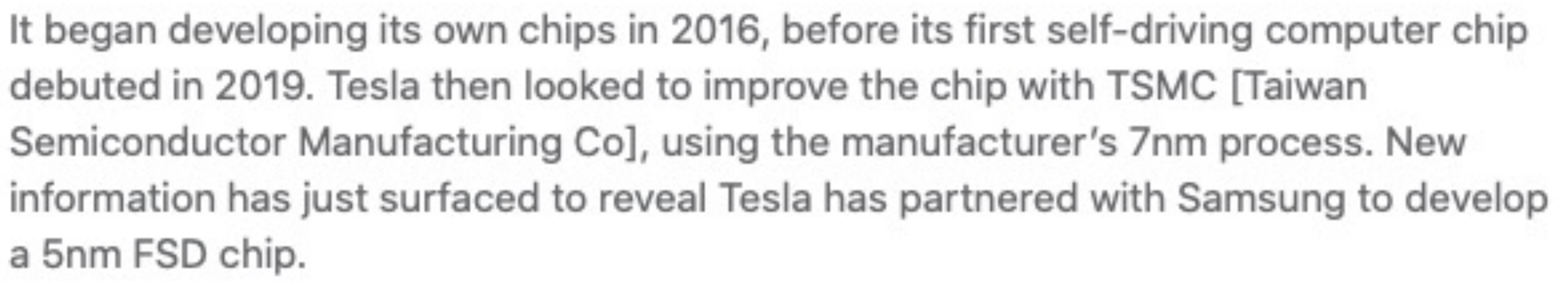
Summary [edit]

Name	Autopilot hardware 1	Enhanced Autopilot hardware 2.0[a]	Enhanced Autopilot hardware 2.5 (HW2.5)[b]	Full self-driving computer (FSD) hardware 3[c]
Hardware	**Hardware 1**	**Hardware 2**[71]		**Hardware 3**
Initial availability date	2014	October 2016	August 2017	April 2019
		Computers		
Platform	MobilEye EyeQ3[119]	NVIDIA DRIVE PX 2 AI computing platform[120]	NVIDIA DRIVE PX 2 with secondary node enabled[39]	Two identical Tesla-designed processors
		Sensors		
Forward Radar	160 m (525 ft)[69]		170 m (558 ft)[69]	
Front / Side Camera color filter array	N/A	RCCC[69]	RCCB[69]	
Forward Cameras	1 monochrome with unknown range	3: Narrow (35°): 250 m (820 ft) Main (50°): 150 m (490 ft) Wide (120°): 60 m (195 ft)		
Forward Looking Side Cameras	N/A	Left (90°): 80 m (260 ft) Right (90°): 80 m (260 ft)		
Rearward Looking Side Cameras	N/A	Left: 100 m (330 ft) Right: 100 m (330 ft)		
Sonars	12 surrounding with 5 m (16 ft) range	12 surrounding with 8 m (26 ft) range		

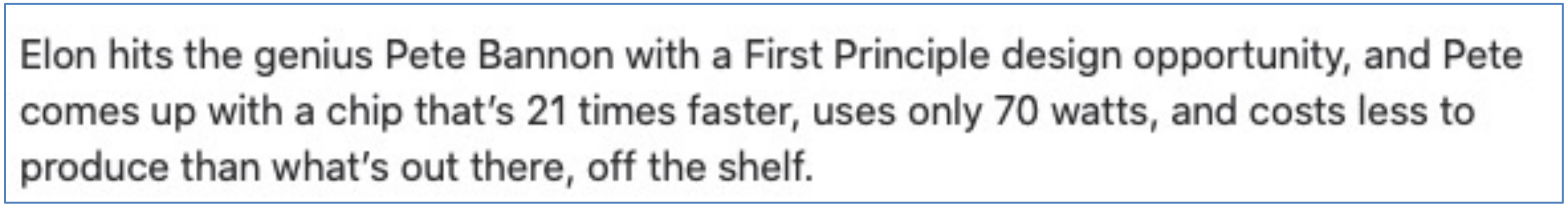
Tesla designs

Tesla Computers

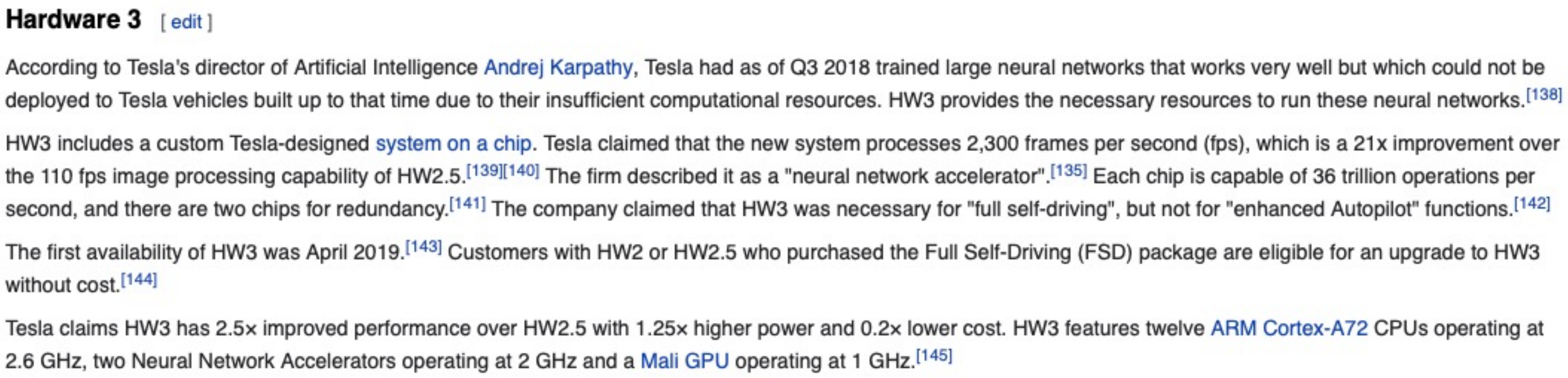
Tesla designs



It began developing its own chips in 2016, before its first self-driving computer chip debuted in 2019. Tesla then looked to improve the chip with TSMC [Taiwan Semiconductor Manufacturing Co], using the manufacturer's 7nm process. New information has just surfaced to reveal Tesla has partnered with Samsung to develop a 5nm FSD chip.



Elon hits the genius Pete Bannon with a First Principle design opportunity, and Pete comes up with a chip that's 21 times faster, uses only 70 watts, and costs less to produce than what's out there, off the shelf.



Hardware 3 [edit]

According to Tesla's director of Artificial Intelligence Andrej Karpathy, Tesla had as of Q3 2018 trained large neural networks that works very well but which could not be deployed to Tesla vehicles built up to that time due to their insufficient computational resources. HW3 provides the necessary resources to run these neural networks.[138]

HW3 includes a custom Tesla-designed system on a chip. Tesla claimed that the new system processes 2,300 frames per second (fps), which is a 21x improvement over the 110 fps image processing capability of HW2.5.[139][140] The firm described it as a "neural network accelerator".[135] Each chip is capable of 36 trillion operations per second, and there are two chips for redundancy.[141] The company claimed that HW3 was necessary for "full self-driving", but not for "enhanced Autopilot" functions.[142]

The first availability of HW3 was April 2019.[143] Customers with HW2 or HW2.5 who purchased the Full Self-Driving (FSD) package are eligible for an upgrade to HW3 without cost.[144]

Tesla claims HW3 has 2.5× improved performance over HW2.5 with 1.25× higher power and 0.2× lower cost. HW3 features twelve ARM Cortex-A72 CPUs operating at 2.6 GHz, two Neural Network Accelerators operating at 2 GHz and a Mali GPU operating at 1 GHz.[145]

Tesla at TSMC

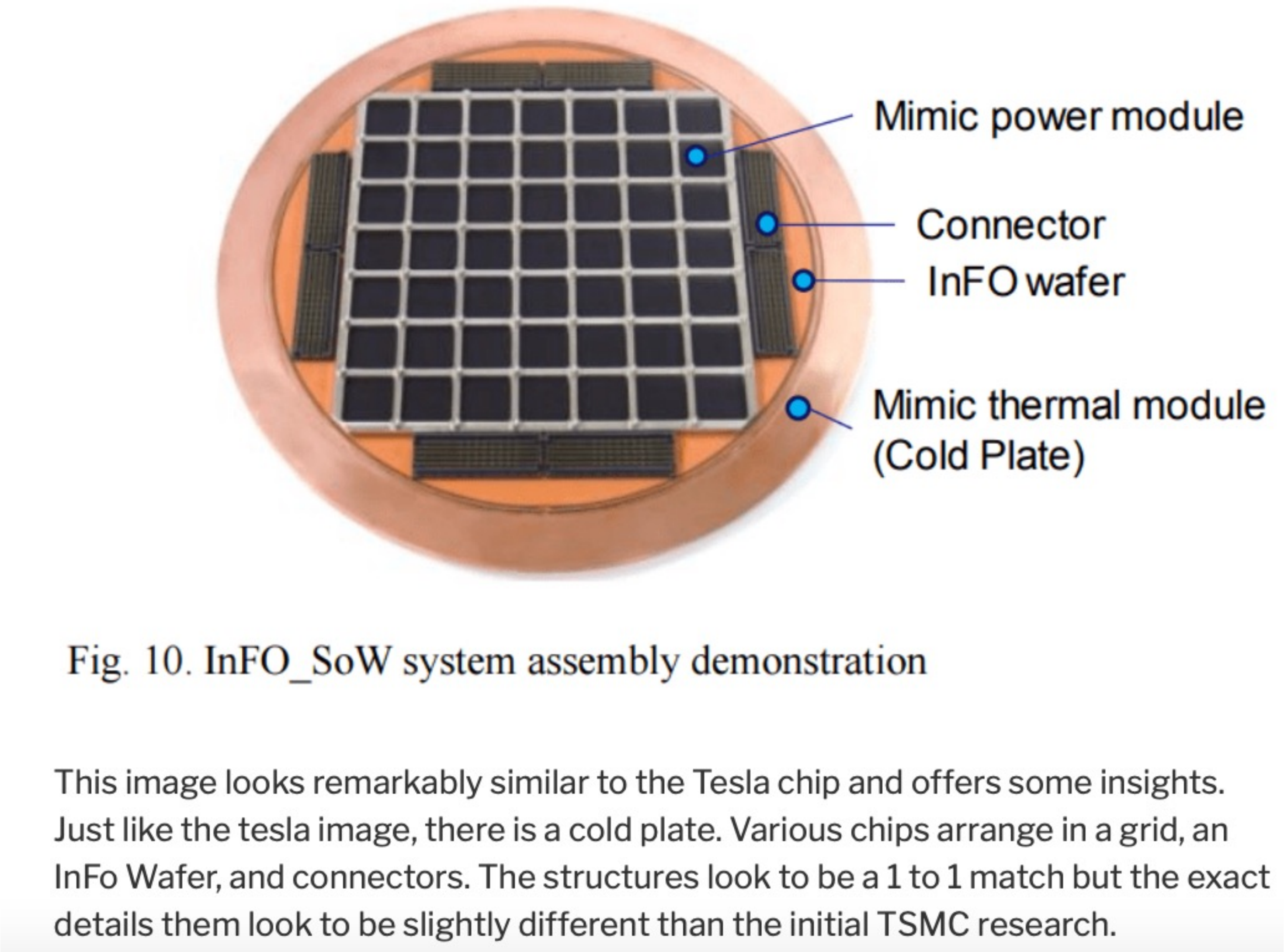


Fig. 10. InFO_SoW system assembly demonstration

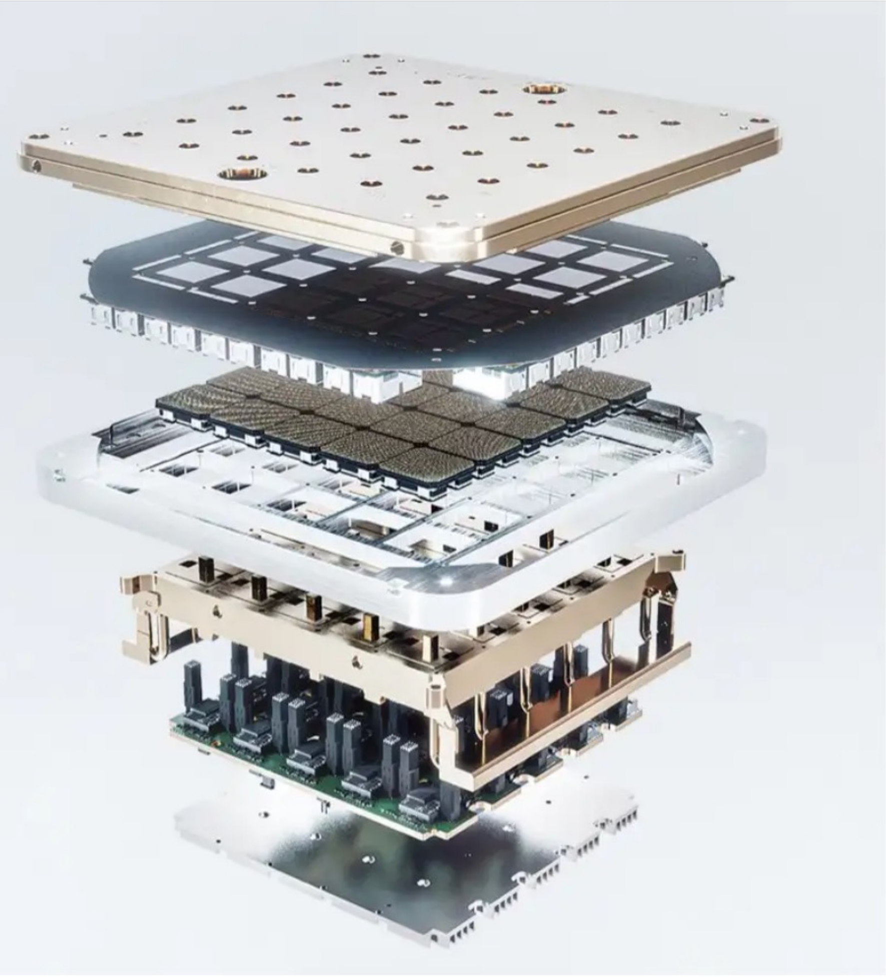
This image looks remarkably similar to the Tesla chip and offers some insights. Just like the tesla image, there is a cold plate. Various chips arrange in a grid, an InFo Wafer, and connectors. The structures look to be a 1 to 1 match but the exact details them look to be slightly different than the initial TSMC research.

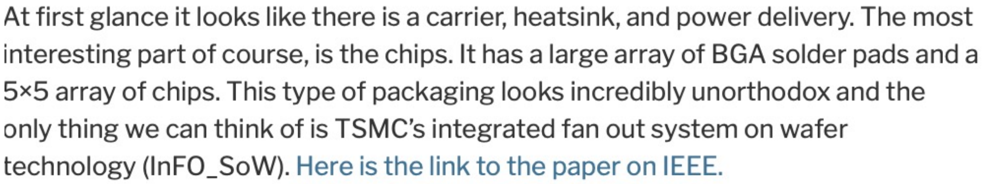
Tesla at TSMC

TSMC's Chiplets Integration

TSMC

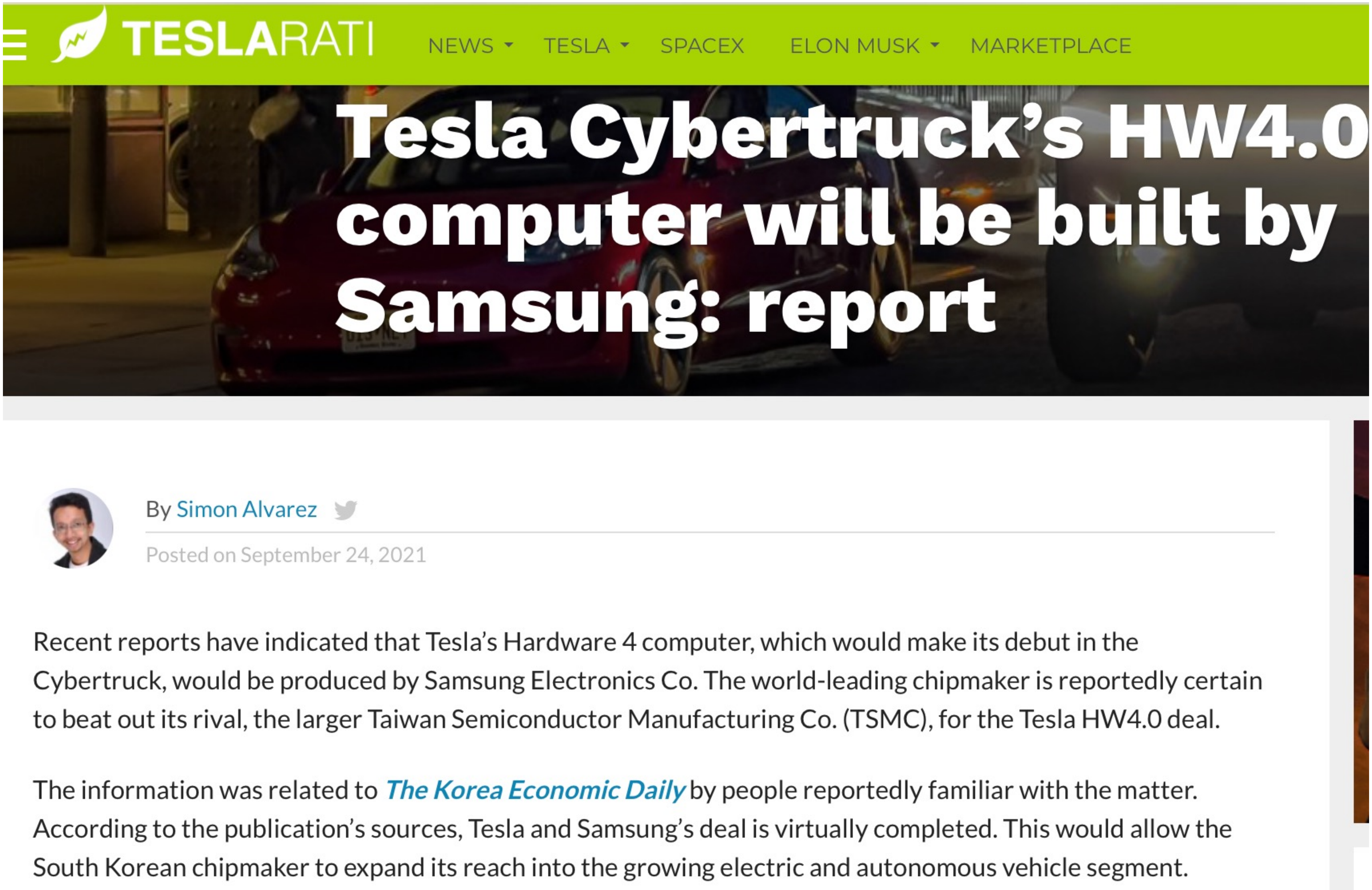
Tesla AI Day Supercomputer Chip Teaser | Is This The First Deployment Of TSMC InFO_SoW?
by Dylan Patel on 08-08-2021 at 10:00 am
Categories: Events, Samsung Foundry, TSMC





At first glance it looks like there is a carrier, heatsink, and power delivery. The most interesting part of course, is the chips. It has a large array of BGA solder pads and a 5×5 array of chips. This type of packaging looks incredibly unorthodox and the only thing we can think of is TSMC's integrated fan out system on wafer technology (InFO_SoW). Here is the link to the paper on IEEE.

Tesla HW 4.0



Tesla Cybertruck's HW4.0 computer will be built by Samsung: report

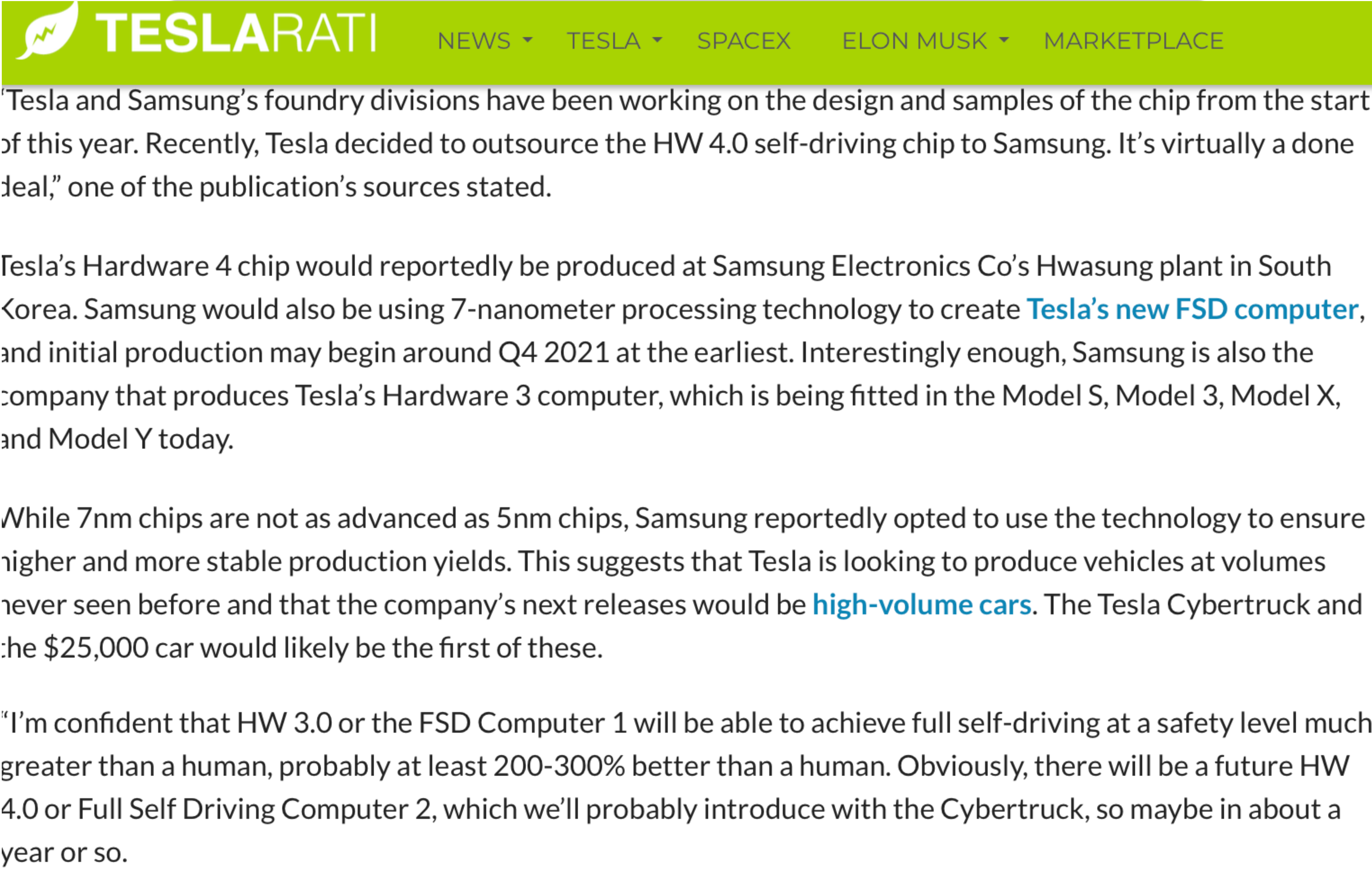
By Simon Alvarez

Posted on September 24, 2021

Recent reports have indicated that Tesla's Hardware 4 computer, which would make its debut in the Cybertruck, would be produced by Samsung Electronics Co. The world-leading chipmaker is reportedly certain to beat out its rival, the larger Taiwan Semiconductor Manufacturing Co. (TSMC), for the Tesla HW4.0 deal.

The information was related to ***The Korea Economic Daily*** by people reportedly familiar with the matter. According to the publication's sources, Tesla and Samsung's deal is virtually completed. This would allow the South Korean chipmaker to expand its reach into the growing electric and autonomous vehicle segment.

Tesla HW 4.0



"Tesla and Samsung's foundry divisions have been working on the design and samples of the chip from the start of this year. Recently, Tesla decided to outsource the HW 4.0 self-driving chip to Samsung. It's virtually a done deal," one of the publication's sources stated.

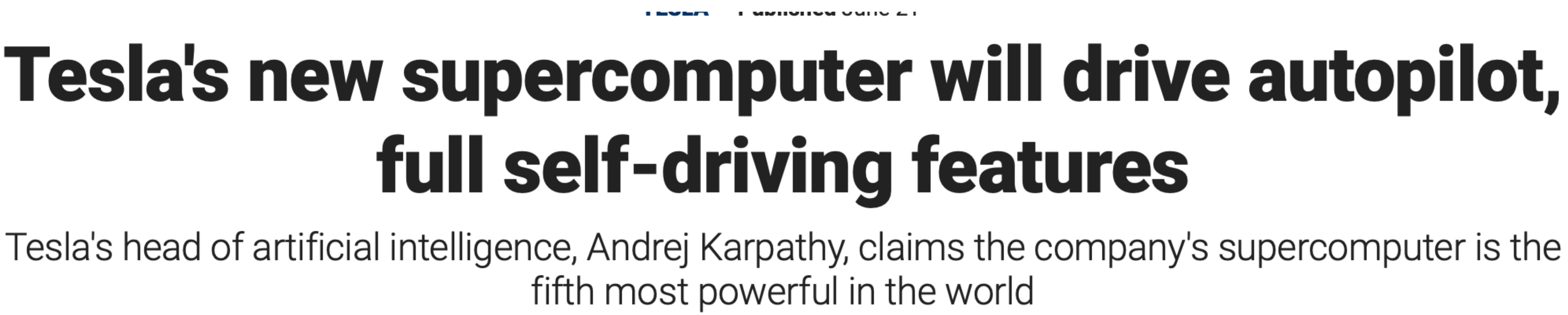
Tesla's Hardware 4 chip would reportedly be produced at Samsung Electronics Co's Hwasung plant in South Korea. Samsung would also be using 7-nanometer processing technology to create Tesla's new FSD computer, and initial production may begin around Q4 2021 at the earliest. Interestingly enough, Samsung is also the company that produces Tesla's Hardware 3 computer, which is being fitted in the Model S, Model 3, Model X, and Model Y today.

While 7nm chips are not as advanced as 5nm chips, Samsung reportedly opted to use the technology to ensure higher and more stable production yields. This suggests that Tesla is looking to produce vehicles at volumes never seen before and that the company's next releases would be high-volume cars. The Tesla Cybertruck and the $25,000 car would likely be the first of these.

"I'm confident that HW 3.0 or the FSD Computer 1 will be able to achieve full self-driving at a safety level much greater than a human, probably at least 200-300% better than a human. Obviously, there will be a future HW 4.0 or Full Self Driving Computer 2, which we'll probably introduce with the Cybertruck, so maybe in about a year or so.

Tesla Supercomputer

FOX BUSINESS



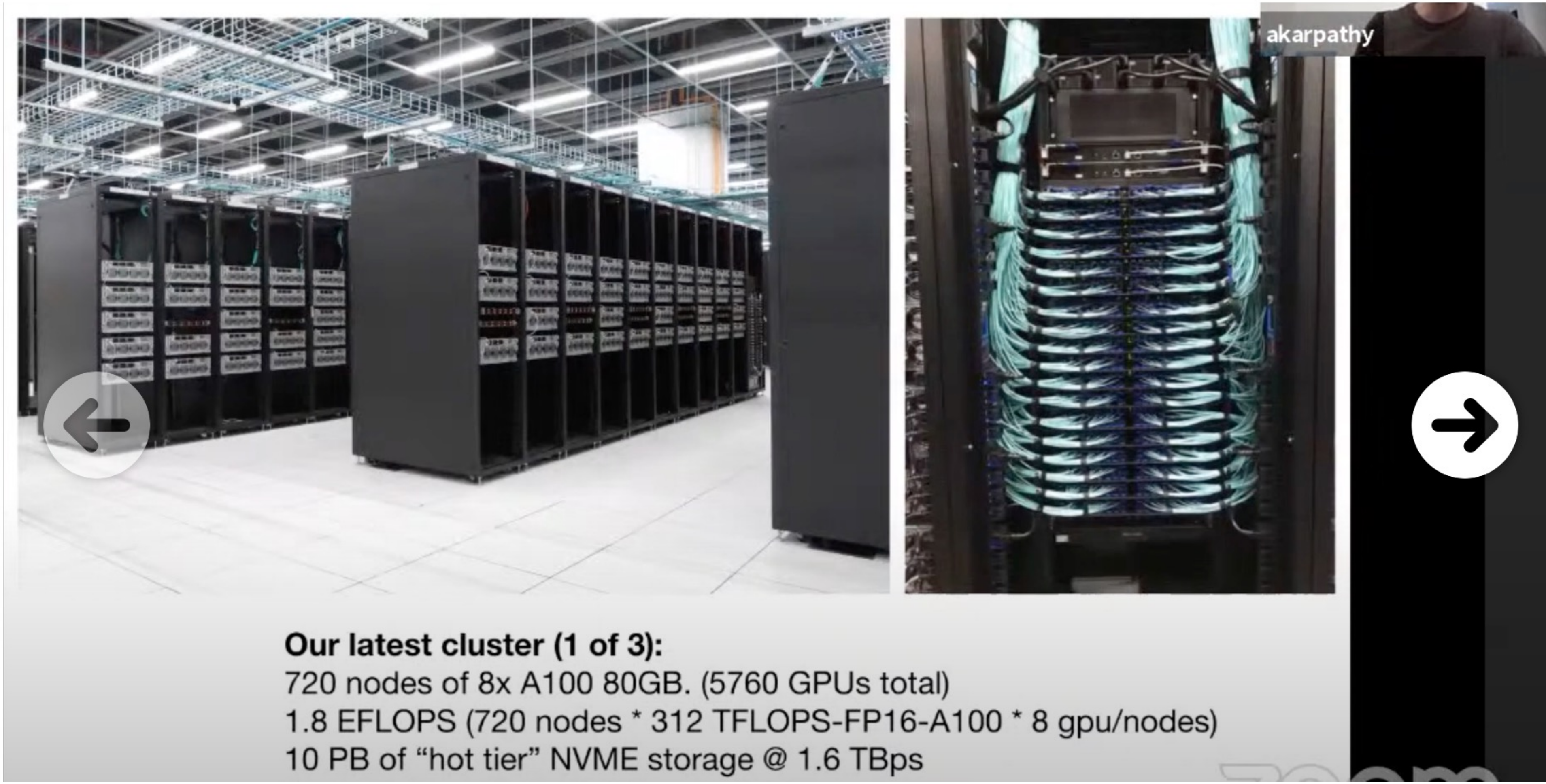
TESLA · Published June 21

Tesla's new supercomputer will drive autopilot, full self-driving features

Tesla's head of artificial intelligence, Andrej Karpathy, claims the company's supercomputer is the fifth most powerful in the world

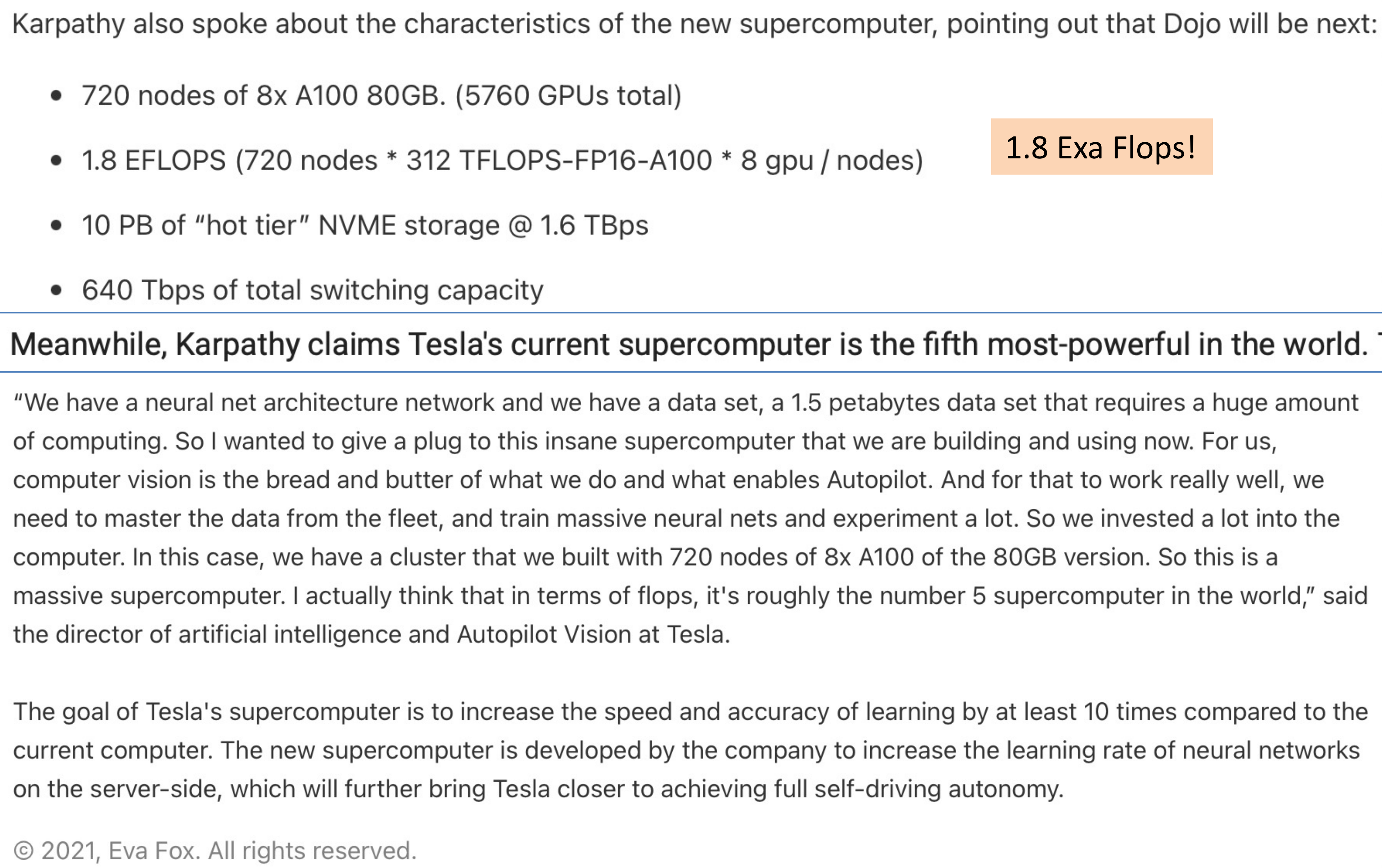
Tesla Supercomputer

Tesla's supercomputer is currently being used and further developed by the company to train neural networks, which are computer systems used to process vast amounts of data.



Our latest cluster (1 of 3):
720 nodes of 8x A100 80GB. (5760 GPUs total)
1.8 EFLOPS (720 nodes * 312 TFLOPS-FP16-A100 * 8 gpu/nodes)
10 PB of “hot tier” NVME storage @ 1.6 TBps

Tesla Supercomputer



Karpathy also spoke about the characteristics of the new supercomputer, pointing out that Dojo will be next:

- 720 nodes of 8x A100 80GB. (5760 GPUs total)
- 1.8 EFLOPS (720 nodes * 312 TFLOPS-FP16-A100 * 8 gpu / nodes)
- 10 PB of "hot tier" NVME storage @ 1.6 TBps
- 640 Tbps of total switching capacity

1.8 Exa Flops!

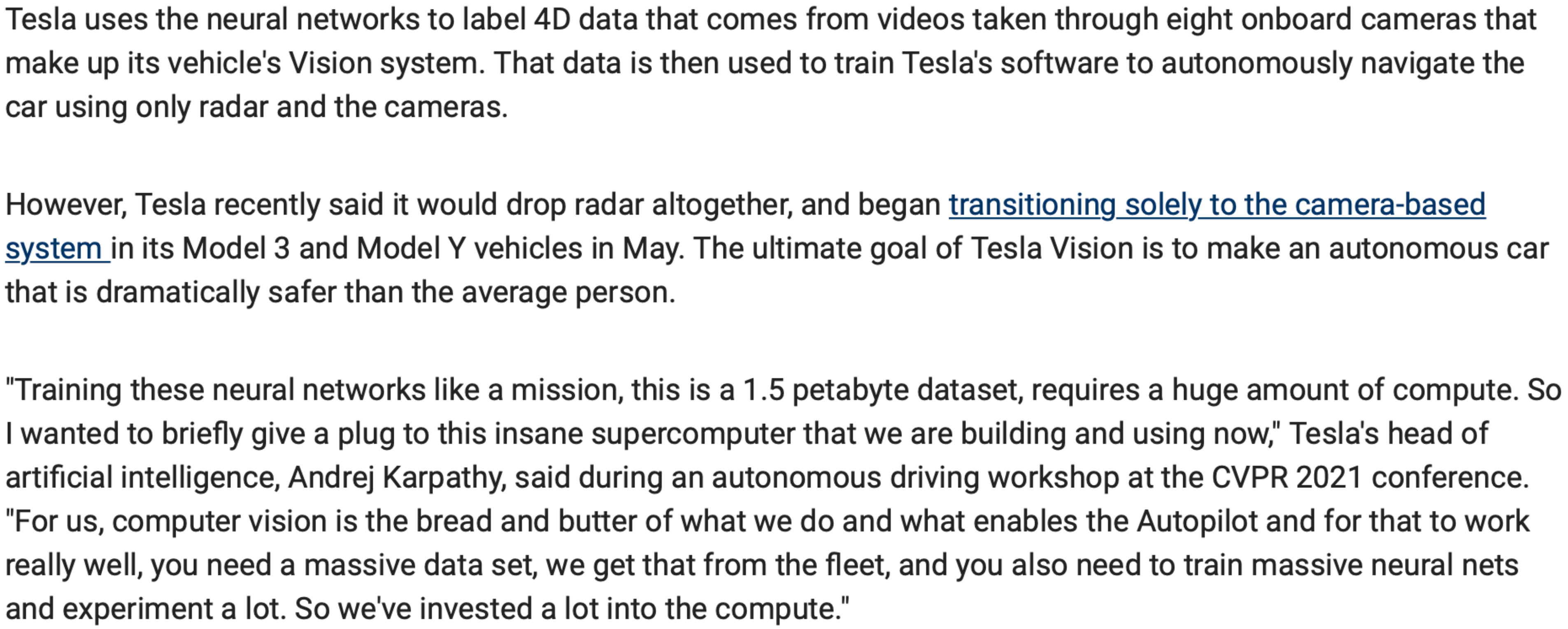
Meanwhile, Karpathy claims Tesla's current supercomputer is the fifth most-powerful in the world.

"We have a neural net architecture network and we have a data set, a 1.5 petabytes data set that requires a huge amount of computing. So I wanted to give a plug to this insane supercomputer that we are building and using now. For us, computer vision is the bread and butter of what we do and what enables Autopilot. And for that to work really well, we need to master the data from the fleet, and train massive neural nets and experiment a lot. So we invested a lot into the computer. In this case, we have a cluster that we built with 720 nodes of 8x A100 of the 80GB version. So this is a massive supercomputer. I actually think that in terms of flops, it's roughly the number 5 supercomputer in the world," said the director of artificial intelligence and Autopilot Vision at Tesla.

The goal of Tesla's supercomputer is to increase the speed and accuracy of learning by at least 10 times compared to the current computer. The new supercomputer is developed by the company to increase the learning rate of neural networks on the server-side, which will further bring Tesla closer to achieving full self-driving autonomy.

Tesla Supercomputer

FOX BUSINESS



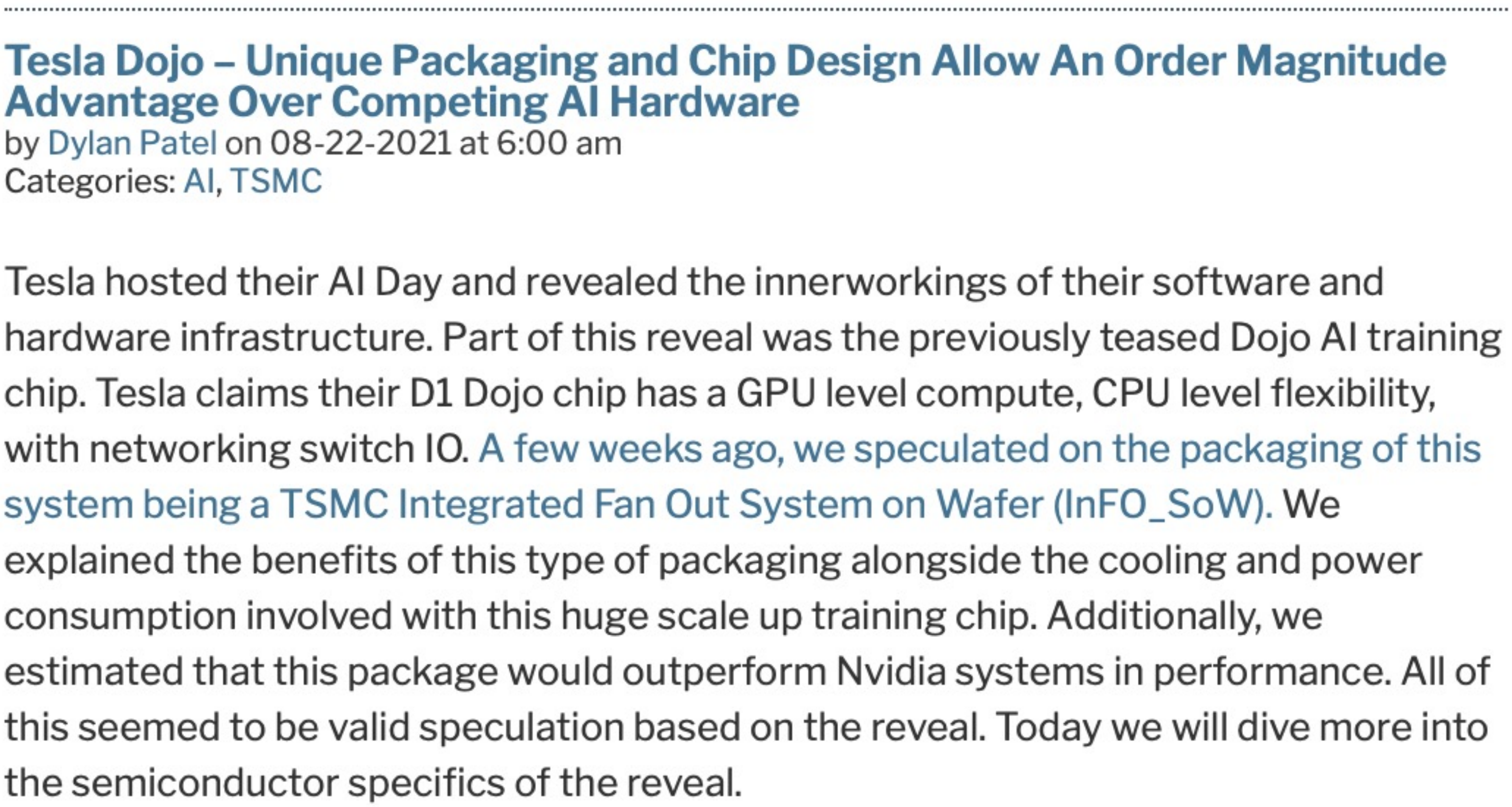
Tesla uses the neural networks to label 4D data that comes from videos taken through eight onboard cameras that make up its vehicle's Vision system. That data is then used to train Tesla's software to autonomously navigate the car using only radar and the cameras.

However, Tesla recently said it would drop radar altogether, and began transitioning solely to the camera-based system in its Model 3 and Model Y vehicles in May. The ultimate goal of Tesla Vision is to make an autonomous car that is dramatically safer than the average person.

"Training these neural networks like a mission, this is a 1.5 petabyte dataset, requires a huge amount of compute. So I wanted to briefly give a plug to this insane supercomputer that we are building and using now," Tesla's head of artificial intelligence, Andrej Karpathy, said during an autonomous driving workshop at the CVPR 2021 conference. "For us, computer vision is the bread and butter of what we do and what enables the Autopilot and for that to work really well, you need a massive data set, we get that from the fleet, and you also need to train massive neural nets and experiment a lot. So we've invested a lot into the compute."

Tesla Dojo

Aug 2021



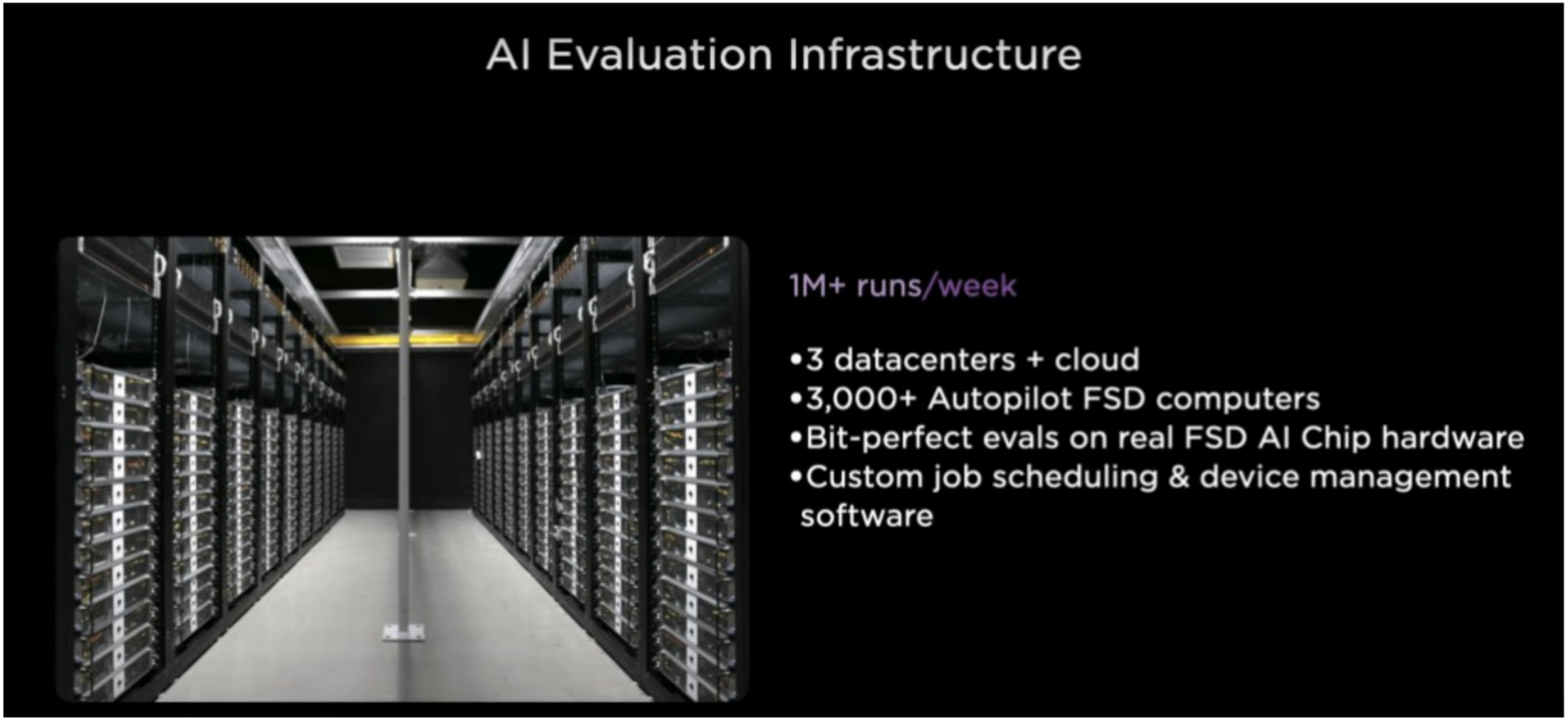
Tesla Dojo – Unique Packaging and Chip Design Allow An Order Magnitude Advantage Over Competing AI Hardware

by Dylan Patel on 08-22-2021 at 6:00 am
Categories: AI, TSMC

Tesla hosted their AI Day and revealed the innerworkings of their software and hardware infrastructure. Part of this reveal was the previously teased Dojo AI training chip. Tesla claims their D1 Dojo chip has a GPU level compute, CPU level flexibility, with networking switch IO. A few weeks ago, we speculated on the packaging of this system being a TSMC Integrated Fan Out System on Wafer (InFO_SoW). We explained the benefits of this type of packaging alongside the cooling and power consumption involved with this huge scale up training chip. Additionally, we estimated that this package would outperform Nvidia systems in performance. All of this seemed to be valid speculation based on the reveal. Today we will dive more into the semiconductor specifics of the reveal.

Tesla Dojo

Aug 2021



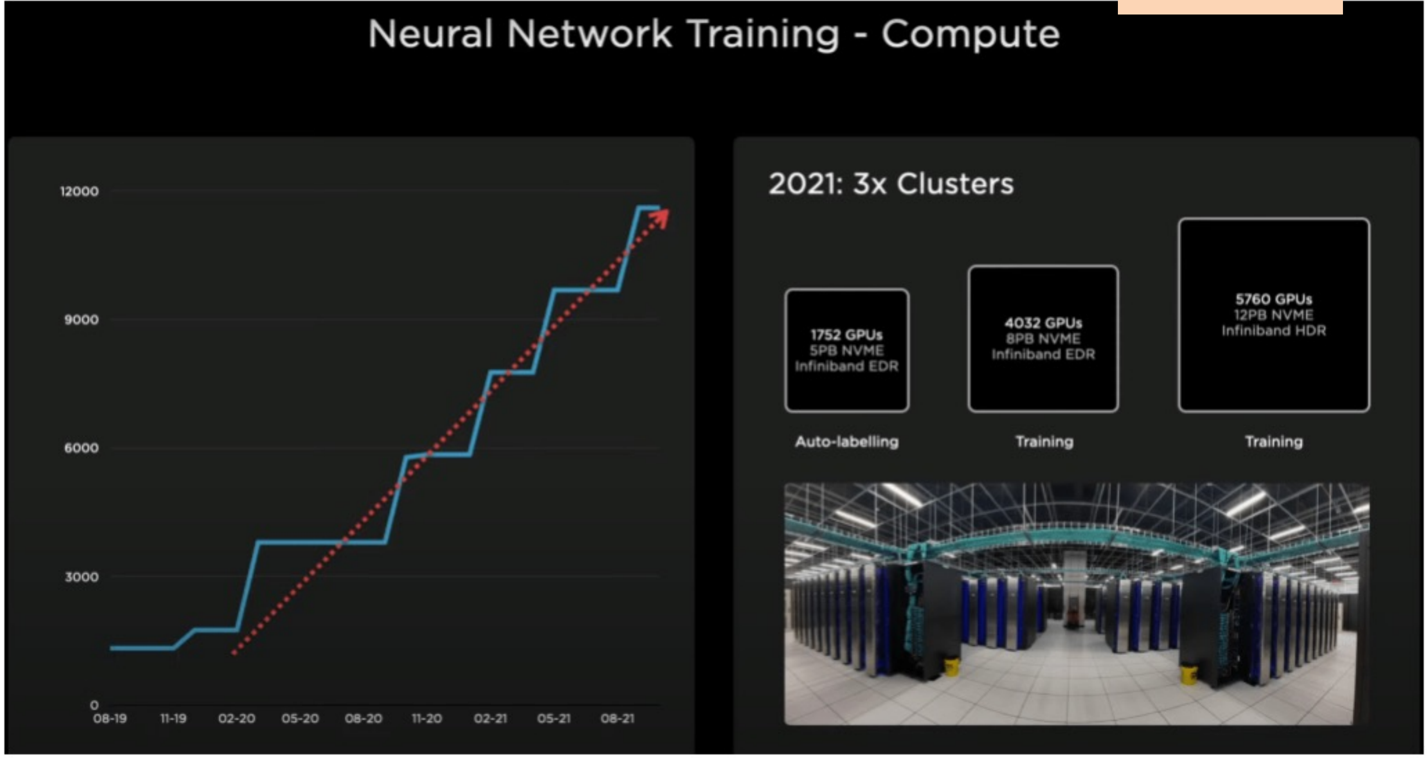
AI Evaluation Infrastructure

1M+ runs/week

- 3 datacenters + cloud
- 3,000+ Autopilot FSD computers
- Bit-perfect evals on real FSD AI Chip hardware
- Custom job scheduling & device management software

Tesla Dojo

Aug 2021



Neural Network Training - Compute

2021: 3x Clusters

1752 GPUs
5PB NVME
Infiniband EDR
Auto-labelling

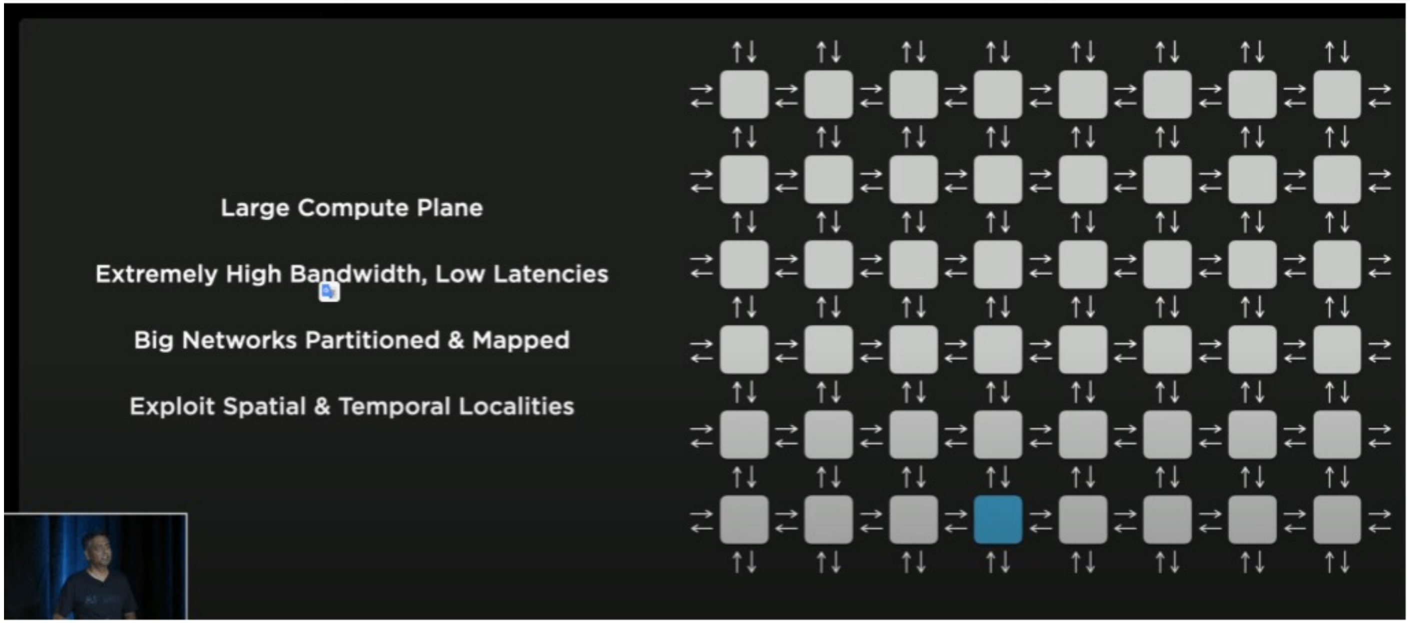
4032 GPUs
8PB NVME
Infiniband EDR
Training

5760 GPUs
12PB NVME
Infiniband HDR
Training

Tesla has been expanding the size of their GPU clusters for years. Their current training cluster would be the 5th largest supercomputer if Tesla stopped all real workloads, ran Linpack, and submitted it to the Top500 list. This scaling of performance is not enough for Tesla and their ambitions, so they set out on developing their own chip a few years ago, project Dojo. Tesla needed more performance to enable even larger and more complex neural networks in a power efficient and cost-effective manner.

Tesla Dojo

Aug 2021



Large Compute Plane

Extremely High Bandwidth, Low Latencies

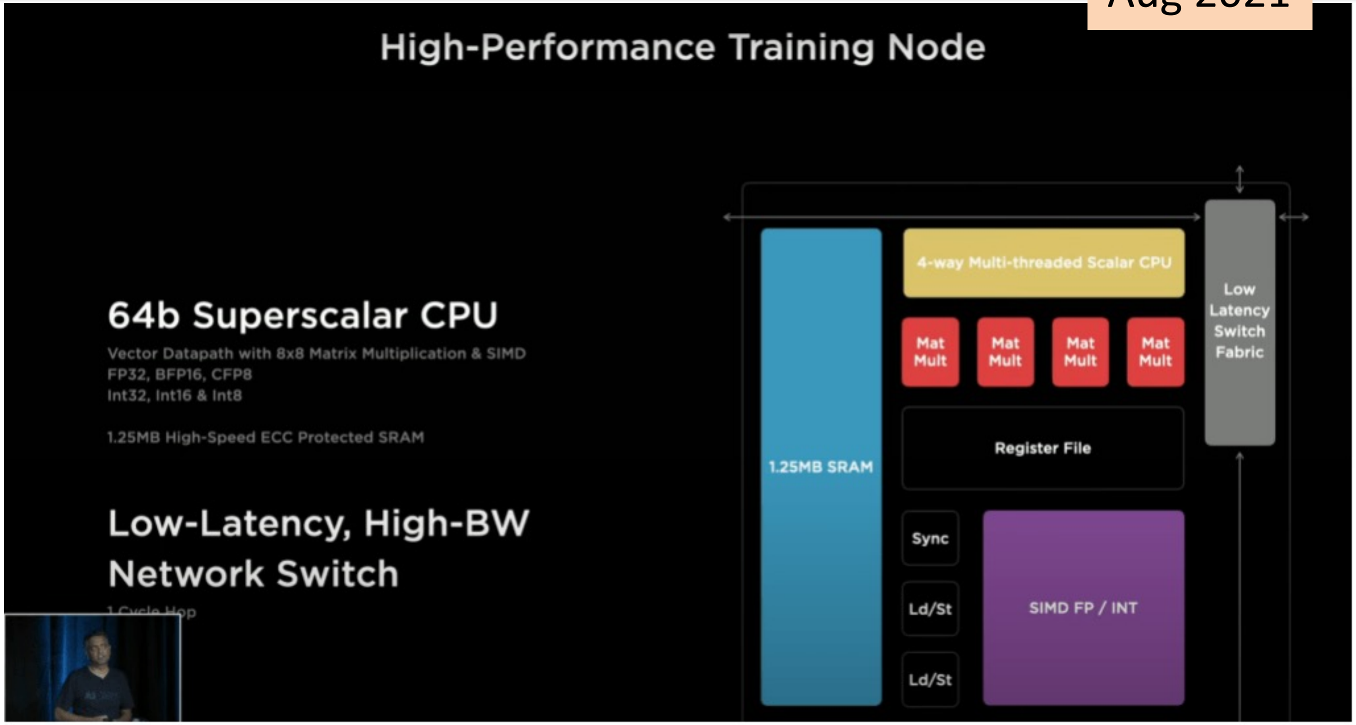
Big Networks Partitioned & Mapped

Exploit Spatial & Temporal Localities

Tesla's architectural solution was a distributed compute architecture. As we listened to their details, this architecture seems very similar to Cerberus. We analyzed the Cerebras Wafer Scale Engine and its architecture here. Every AI training architecture is a laid out in this manner, but the details of the compute elements, the network, and fabric vary widely. The biggest problem with these types of networks is scaling up bandwidth and retaining low latencies. In order to scale to larger networks, Tesla focused on these latter two especially. This influenced every part of their design from the chip fabrics to packaging.

Tesla Dojo

Aug 2021



High-Performance Training Node

64b Superscalar CPU

Vector Datapath with 8x8 Matrix Multiplication & SIMD
FP32, BFP16, CFP8
Int32, Int16 & Int8

1.25MB High-Speed ECC Protected SRAM

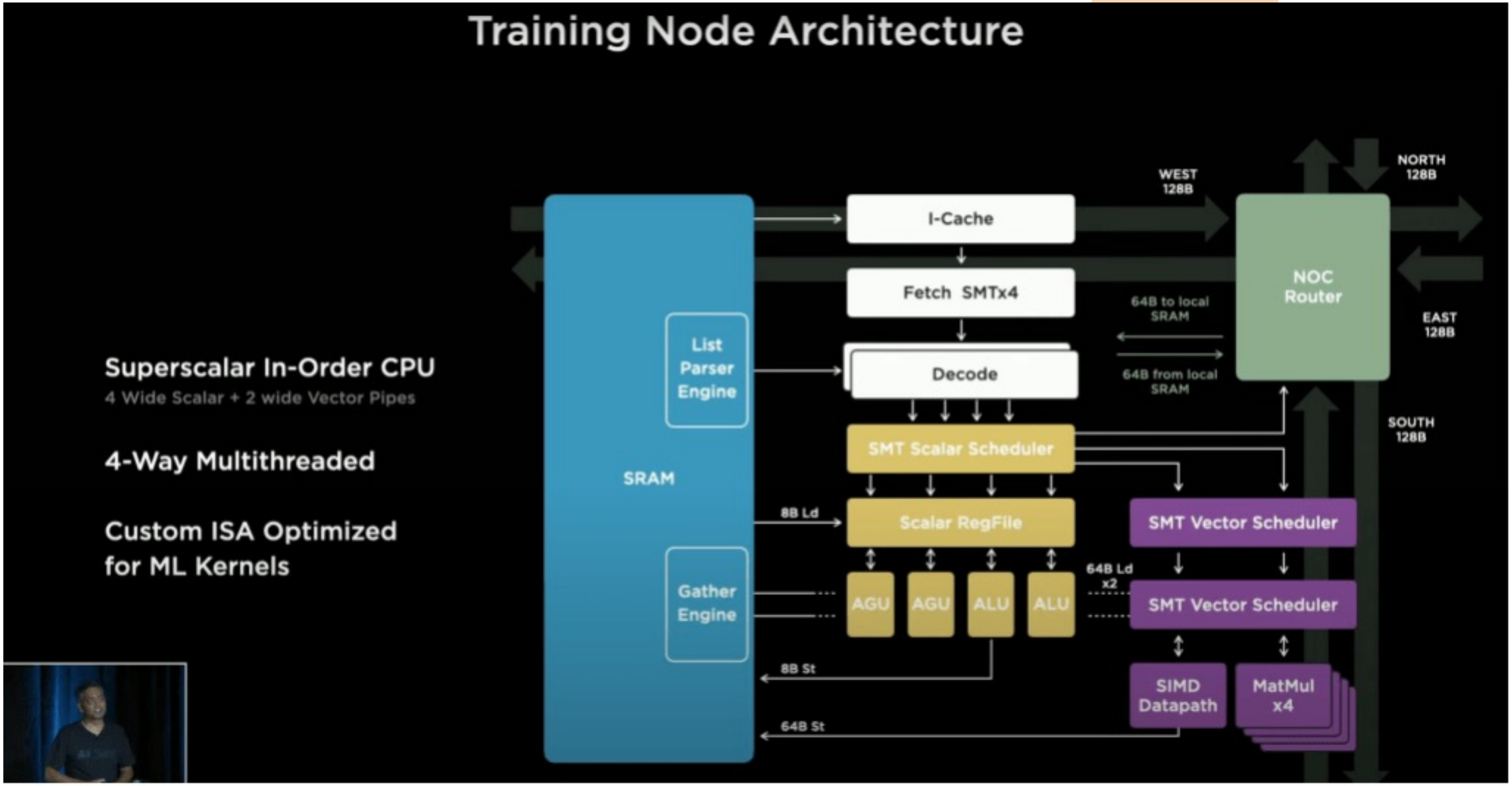
Low-Latency, High-BW Network Switch

1 Cycle Hop

The functional unit was designed to be traversable with 1 clock cycle, but large enough that synchronization overhead and software do not dominate the problem. As such they arrive at a design almost exactly like Cerebras. A mesh of individual units connected by a high-speed fabric which is routes communications between functional units in 1 clock. Each individual unit has a large 1.25MB SRAM scratchpad and multiple superscalar CPU cores with SIMD capabilities and matrix multiply units supporting all common data types. Additionally, they introduce a new data type called CFP8, configurable floating point 8. Each unit is capable of 1TFlop of BF16 or CFP8, 64GFlops FP32, and 512GB/s of bandwidth in each direction.

Tesla Dojo

Aug 2021



Training Node Architecture

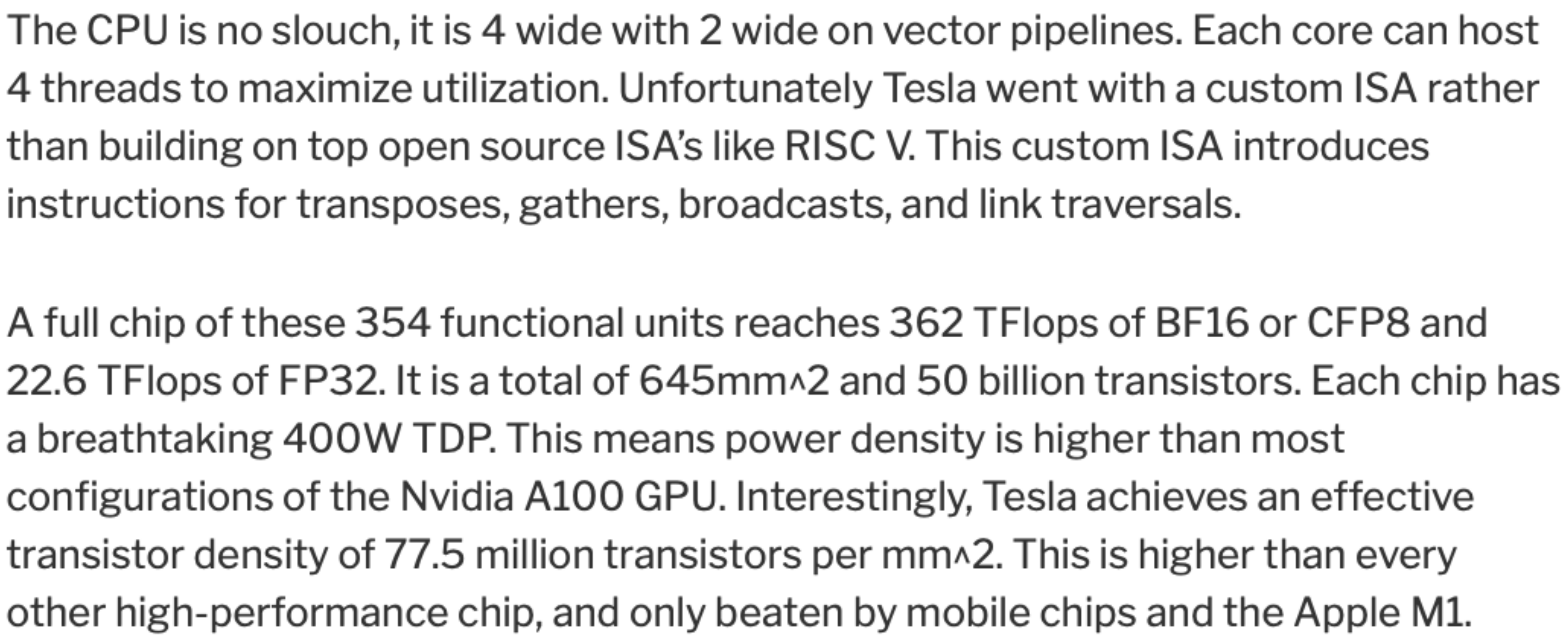
Superscalar In-Order CPU
4 Wide Scalar + 2 wide Vector Pipes

4-Way Multithreaded

Custom ISA Optimized for ML Kernels

Tesla Dojo

Aug 2021

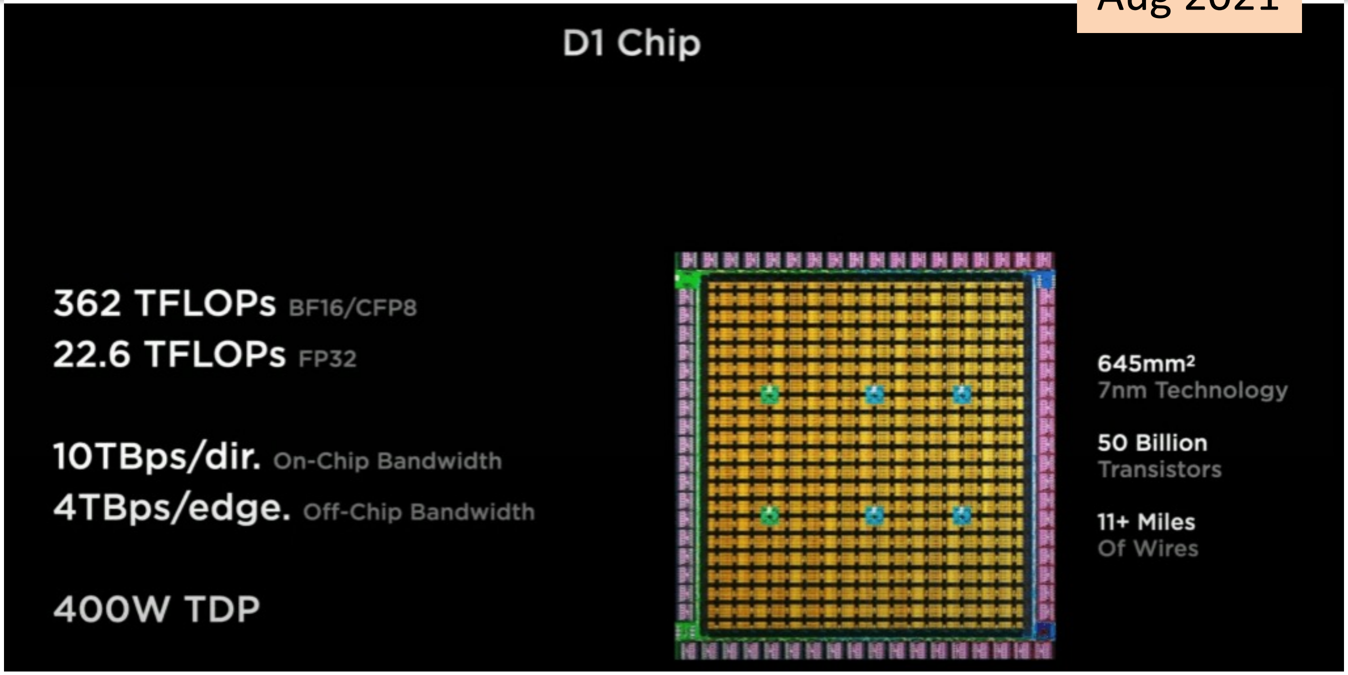


The CPU is no slouch, it is 4 wide with 2 wide on vector pipelines. Each core can host 4 threads to maximize utilization. Unfortunately Tesla went with a custom ISA rather than building on top open source ISA's like RISC V. This custom ISA introduces instructions for transposes, gathers, broadcasts, and link traversals.

A full chip of these 354 functional units reaches 362 TFlops of BF16 or CFP8 and 22.6 TFlops of FP32. It is a total of 645mm^2 and 50 billion transistors. Each chip has a breathtaking 400W TDP. This means power density is higher than most configurations of the Nvidia A100 GPU. Interestingly, Tesla achieves an effective transistor density of 77.5 million transistors per mm^2. This is higher than every other high-performance chip, and only beaten by mobile chips and the Apple M1.

Tesla Dojo

Aug 2021



D1 Chip

362 TFLOPs BF16/CFP8
22.6 TFLOPs FP32

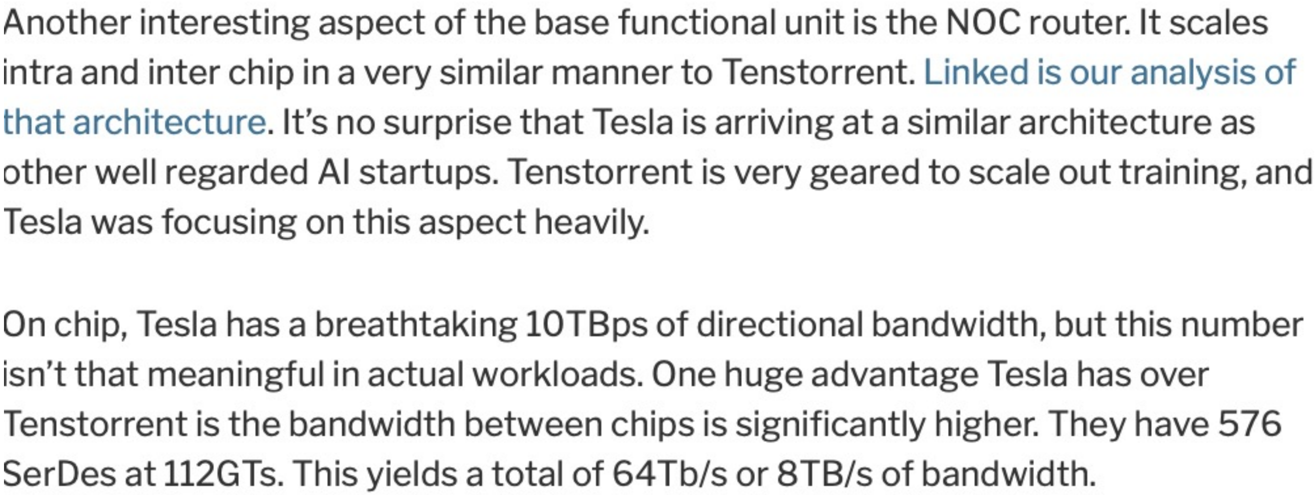
10TBps/dir. On-Chip Bandwidth
4TBps/edge. Off-Chip Bandwidth

400W TDP

645mm²
7nm Technology

50 Billion
Transistors

11+ Miles
Of Wires

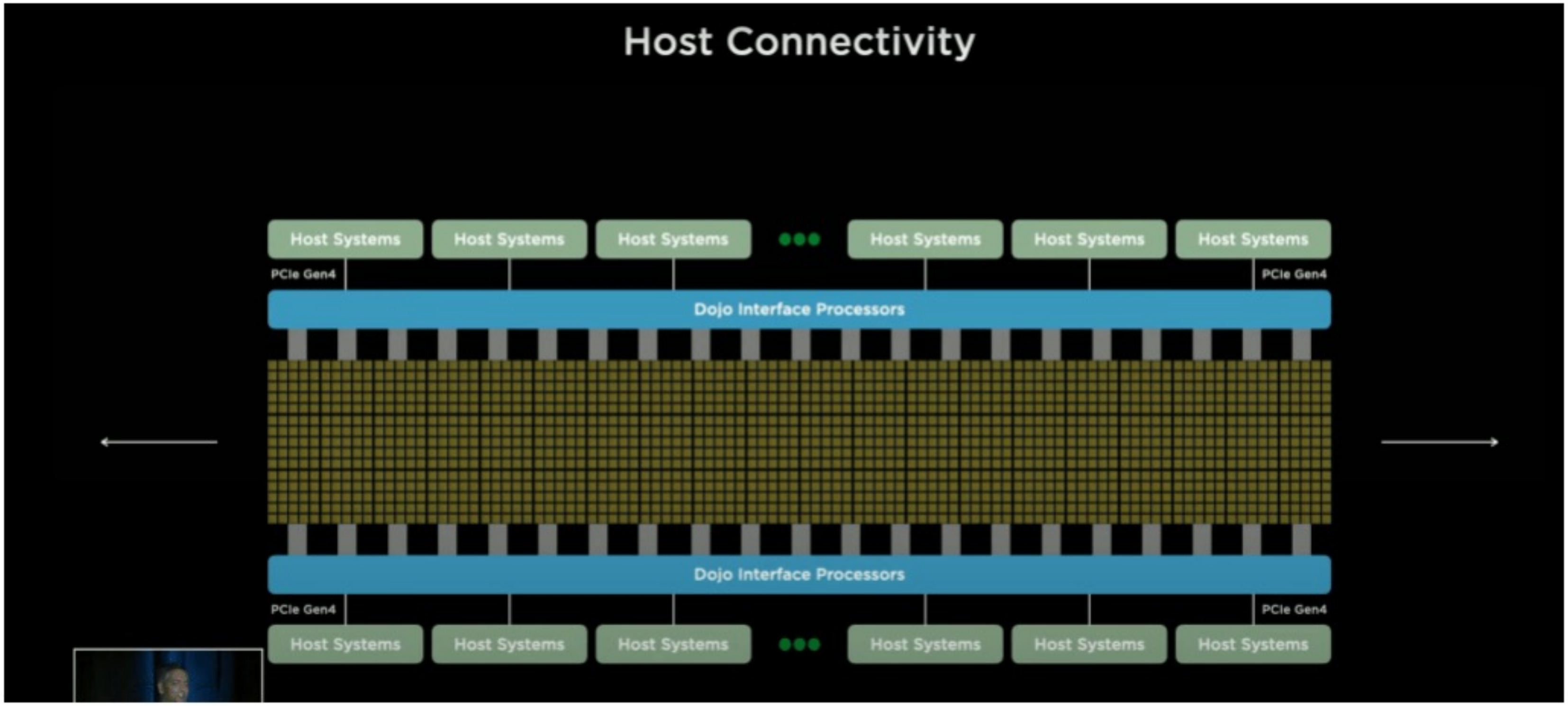


Another interesting aspect of the base functional unit is the NOC router. It scales intra and inter chip in a very similar manner to Tenstorrent. Linked is our analysis of that architecture. It's no surprise that Tesla is arriving at a similar architecture as other well regarded AI startups. Tenstorrent is very geared to scale out training, and Tesla was focusing on this aspect heavily.

On chip, Tesla has a breathtaking 10TBps of directional bandwidth, but this number isn't that meaningful in actual workloads. One huge advantage Tesla has over Tenstorrent is the bandwidth between chips is significantly higher. They have 576 SerDes at 112GTs. This yields a total of 64Tb/s or 8TB/s of bandwidth.

Tesla Dojo

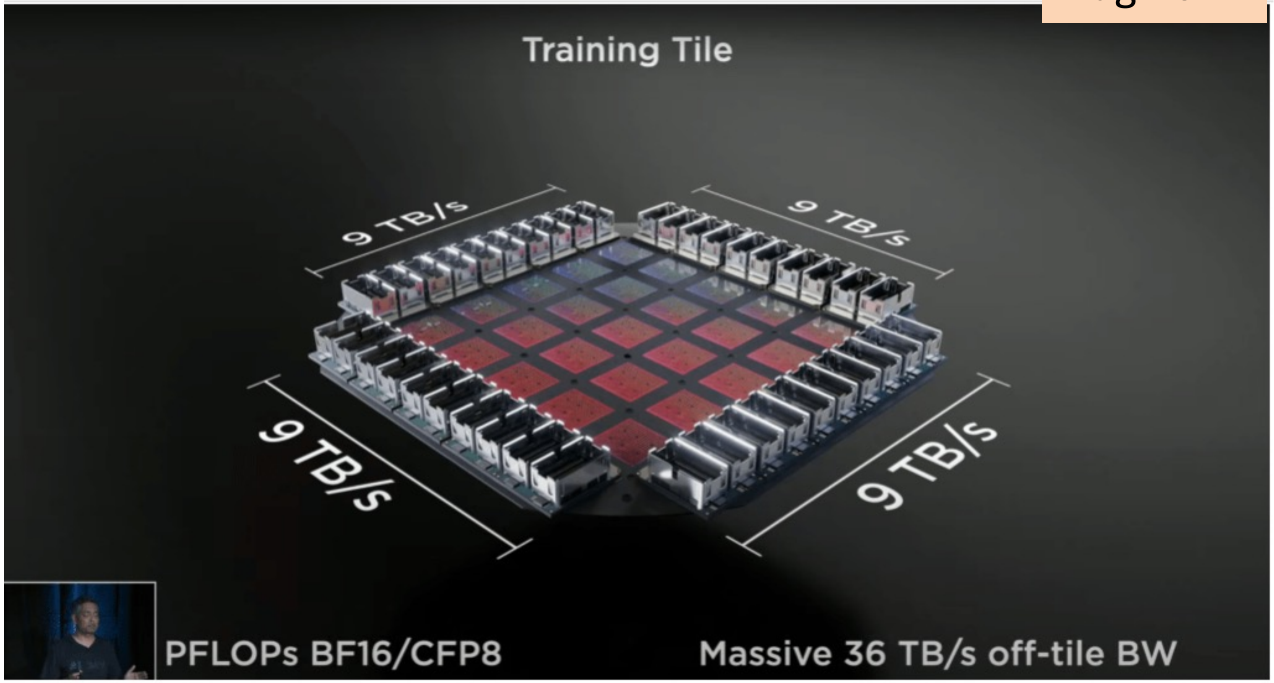
Aug 2021

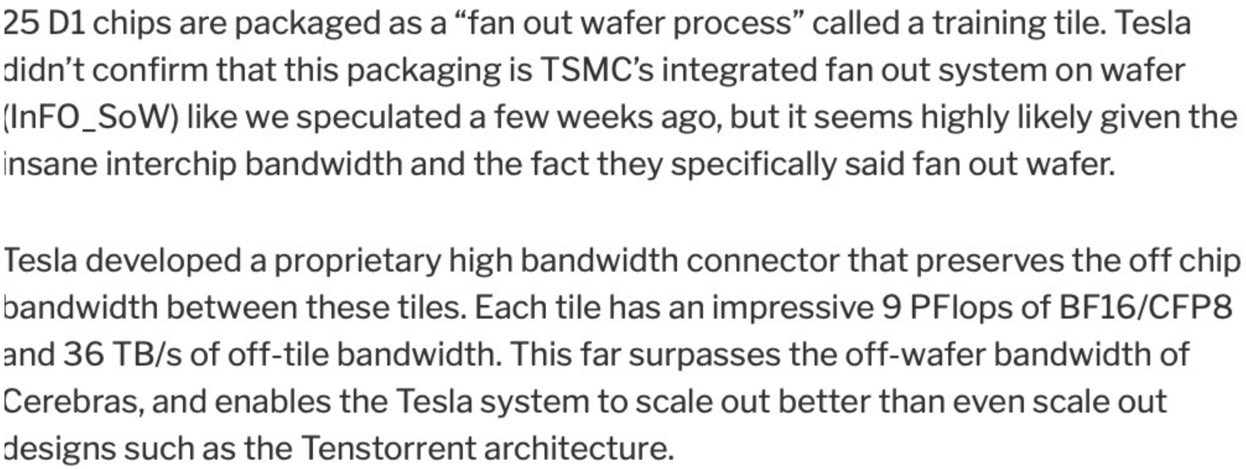


Tesla connects the compute plane of Dojo chips to interface processors which connect to host systems with PCIe 4.0. These interface processors also enable higher radix network connections that supplement the existing compute plane mesh.

Tesla Dojo

Aug 2021



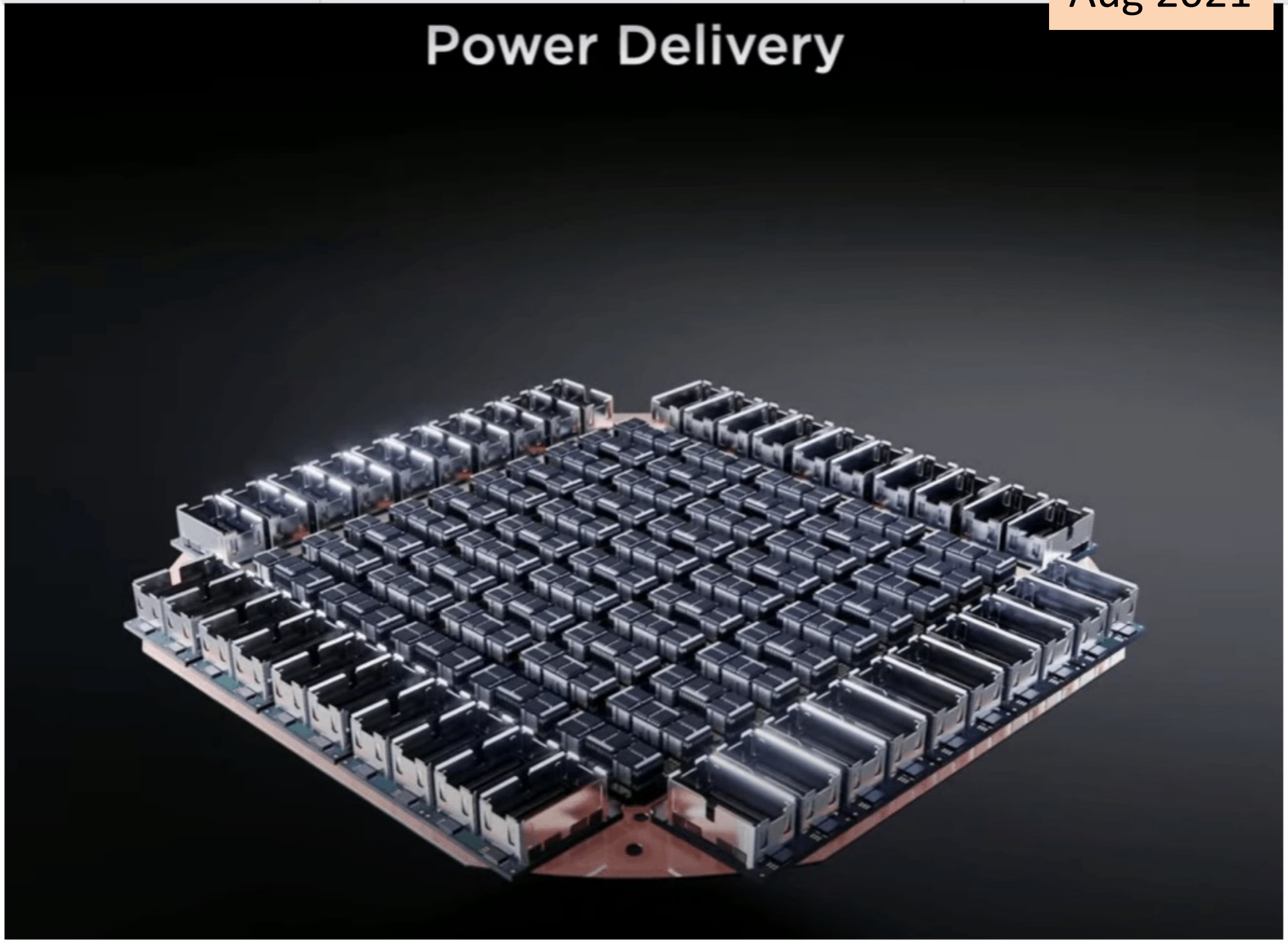


25 D1 chips are packaged as a "fan out wafer process" called a training tile. Tesla didn't confirm that this packaging is TSMC's integrated fan out system on wafer (InFO_SoW) like we speculated a few weeks ago, but it seems highly likely given the insane interchip bandwidth and the fact they specifically said fan out wafer.

Tesla developed a proprietary high bandwidth connector that preserves the off chip bandwidth between these tiles. Each tile has an impressive 9 PFlops of BF16/CFP8 and 36 TB/s of off-tile bandwidth. This far surpasses the off-wafer bandwidth of Cerebras, and enables the Tesla system to scale out better than even scale out designs such as the Tenstorrent architecture.

Tesla Dojo

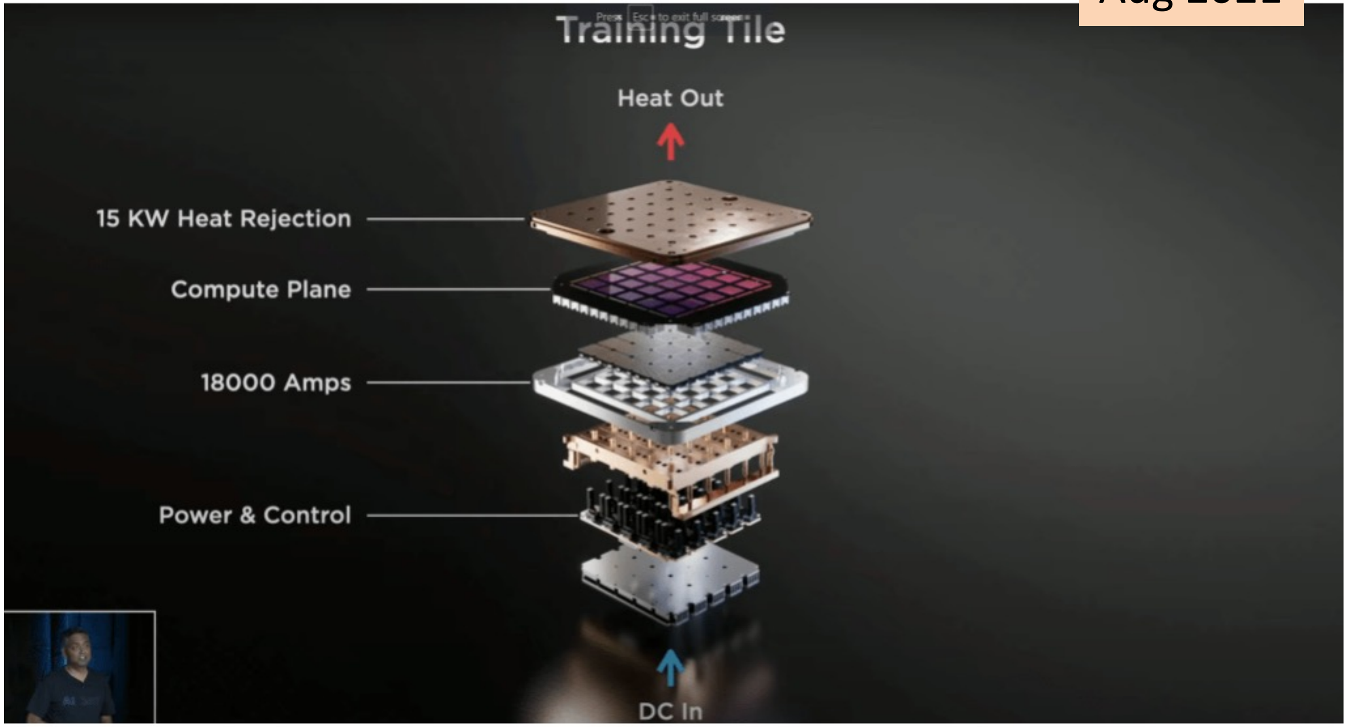
Aug 2021



The power delivery is unique, custom, and extremely impressive as well. With so much bandwidth and over 10KW of power consumption on the package, Tesla innovated on power delivery and feeds it vertically. The custom voltage regulator modulator is reflowed directly onto the fan out wafer. Power, thermal, and mechanical are all interfacing directly with the tile.

Tesla Dojo

Aug 2021

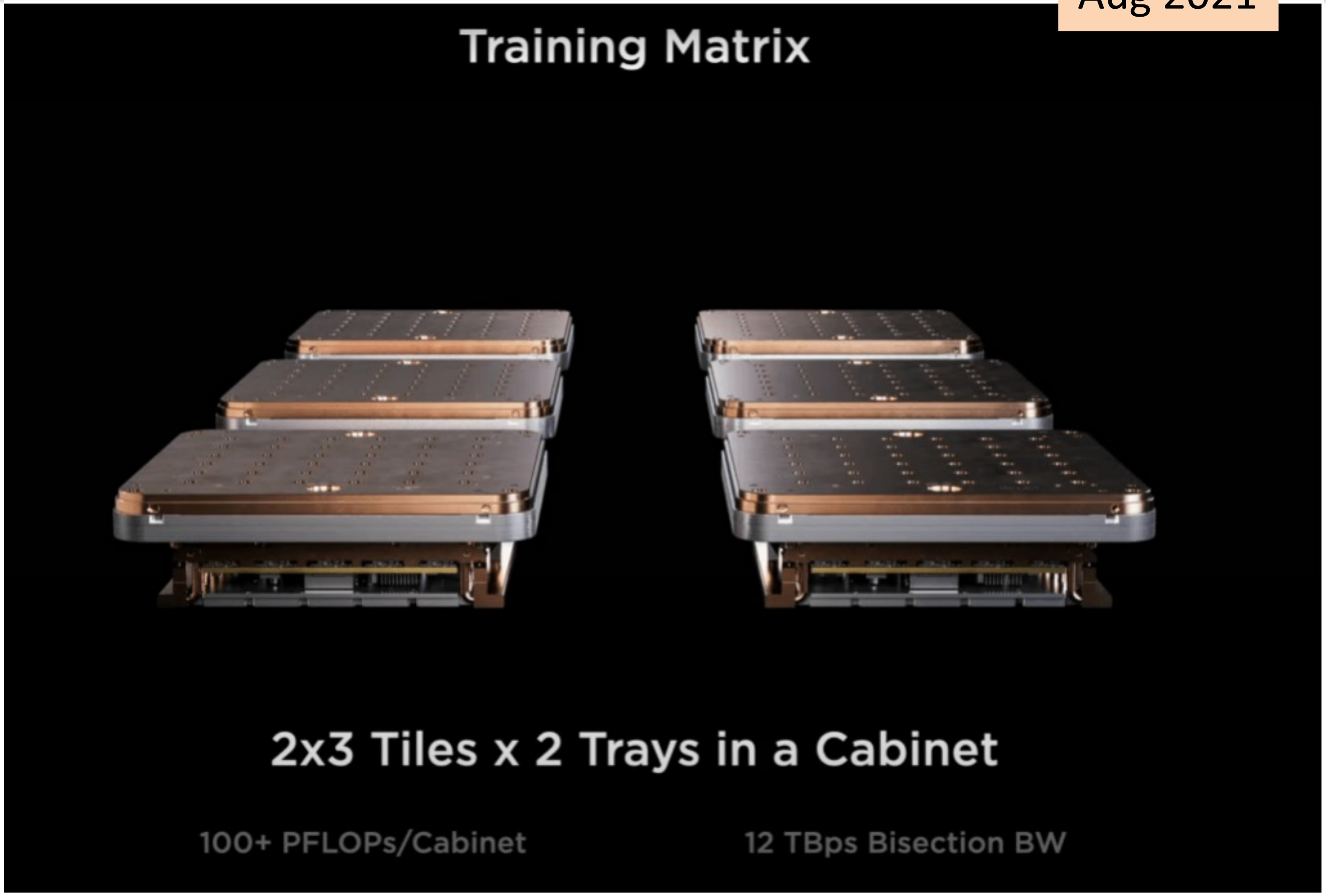


It appears the total tile is 15KW of power even if the chips themselves are only 10KW total. Power delivery, IO, and wafer wires are drawing a ton of power as well. Power comes in from the bottom while the heat comes out the top. Chips are not the unit of scale for Tesla, the 25 chip tiles are. This tile far surpasses anything from Nvidia, Graphcore, Cerebras, Groq, Tenstorrent, SambaNova, or any other AI training geared start up in per unit performance and scale up capabilities.

All of this seems like far out technology, but Tesla claims they already have tiles running at 2 GHz on real AI networks in their labs.

Tesla Dojo

Aug 2021



Training Matrix

2x3 Tiles x 2 Trays in a Cabinet

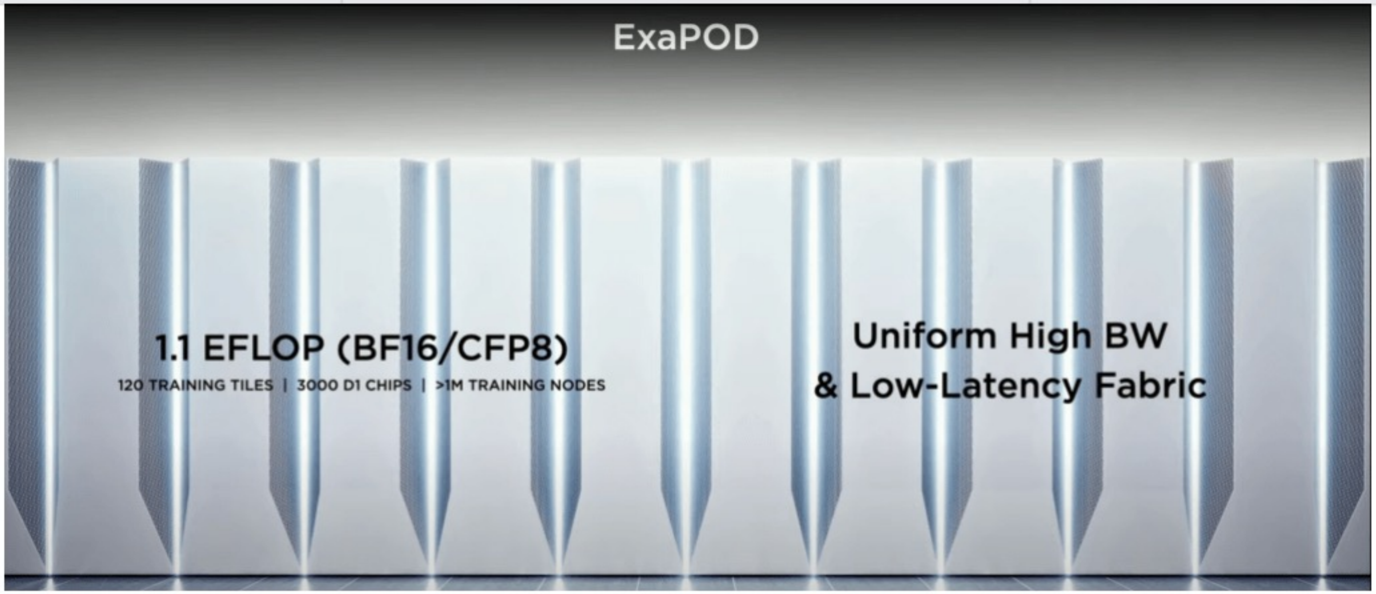
100+ PFLOPs/Cabinet

12 TBps Bisection BW

The next step up to scaling to thousands of chips is the server level. Dojo is scaled up to 2 by 3 tile configs and there are two of these configurations in a server cabinet. For those counting at home, there are 12 total tiles per cabinet for a total of 108 PFlops per cabinet. Over 100,000 functional units, 400,000 custom cores, and 132GB of SRAM per server cabinet is mind blowing numbers.

Tesla Dojo

Aug 2021



ExaPOD

1.1 EFLOP (BF16/CFP8)

120 TRAINING TILES | 3000 D1 CHIPS | >1M TRAINING NODES

Uniform High BW
& Low-Latency Fabric

Tesla keeps scaling up further beyond the cabinet level in their mesh. There is no breakdown of bandwidth between chips. It is one homogenous mesh of chips with insane amounts of bandwidth. They plan to scale to 10 cabinets and 1.1 Exaflops. 1,062,000 functional units, 4,248,000 cores, and 1.33TB of SRAM.

Tesla Dojo

With the scaling that Tesla may hope to achieve in their Dojo supercomputer design, there will be an immense amount of heat. InFO_SoW is capable of 7,000W of power. This is compared to Nvidia's datacenter A100 GPU which comes in configurations as high as 500W. With this requires immense consideration for cooling, and the TSMC paper on InFO_SoW offers a solution.

Water cooled

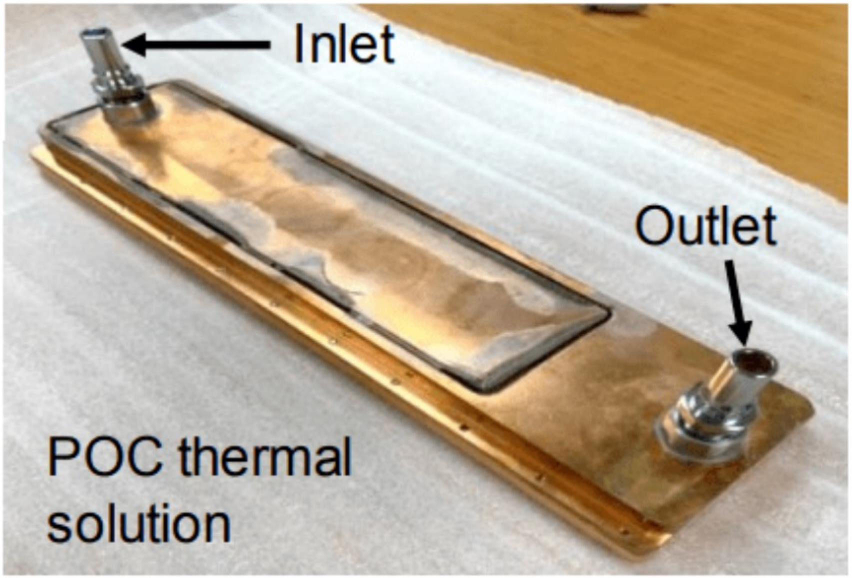
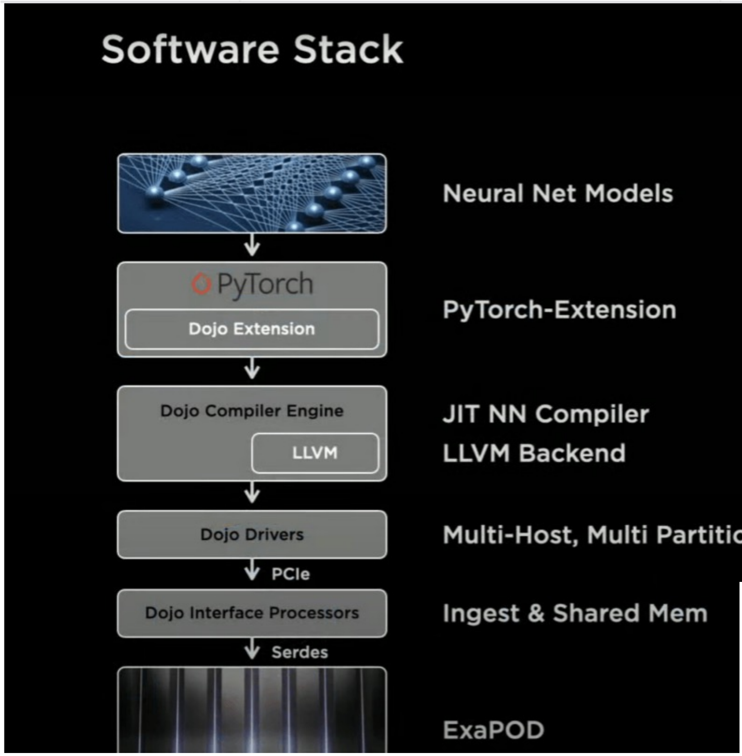


Fig. 7. The 2-row POC thermal solution

This image is quite crude, but the cold plate in the Tesla image looks similar in that it has many inlets and outlets. Water-cooling is a necessity with this level of power and heat density.

Tesla Dojo

Aug 2021



Software Stack

Neural Net Models

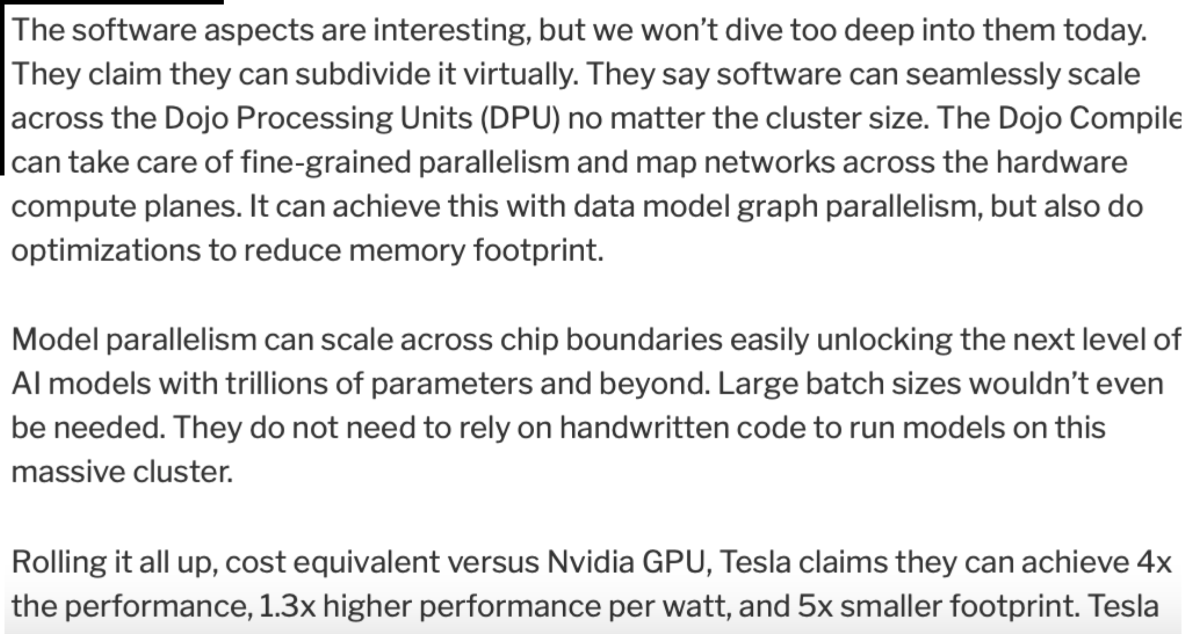
PyTorch-Extension

JIT NN Compiler
LLVM Backend

Multi-Host, Multi Partition Management

Ingest & Shared Mem

ExaPOD



The software aspects are interesting, but we won't dive too deep into them today. They claim they can subdivide it virtually. They say software can seamlessly scale across the Dojo Processing Units (DPU) no matter the cluster size. The Dojo Compile can take care of fine-grained parallelism and map networks across the hardware compute planes. It can achieve this with data model graph parallelism, but also do optimizations to reduce memory footprint.

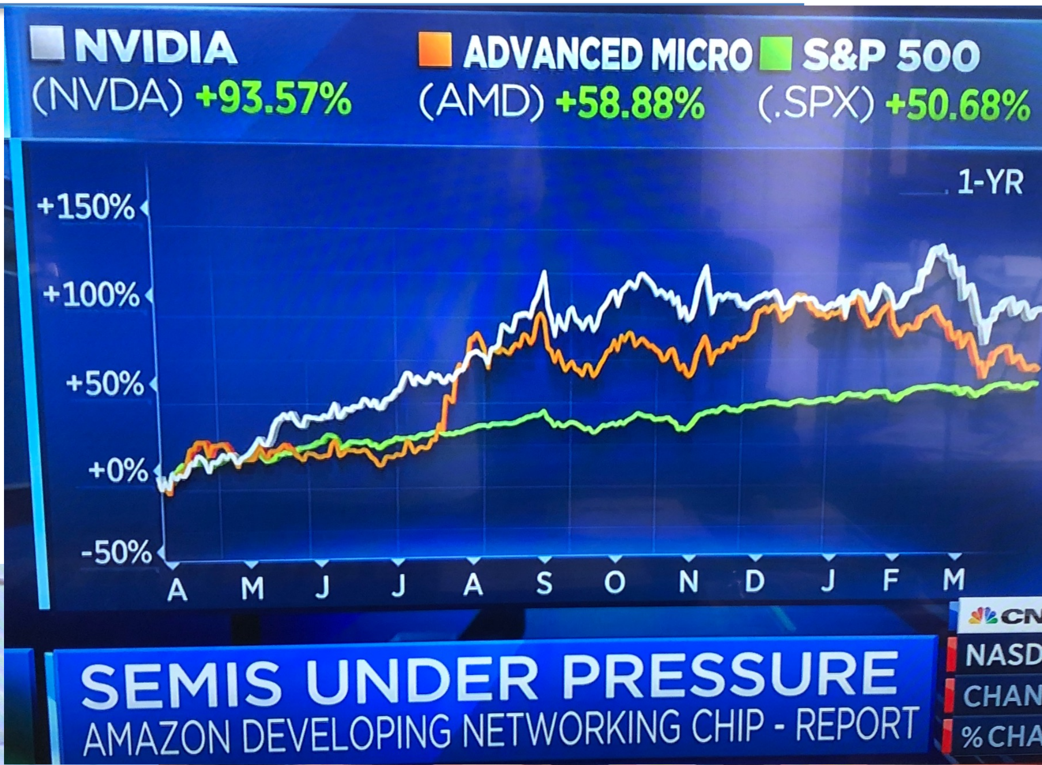
Model parallelism can scale across chip boundaries easily unlocking the next level of AI models with trillions of parameters and beyond. Large batch sizes wouldn't even be needed. They do not need to rely on handwritten code to run models on this massive cluster.

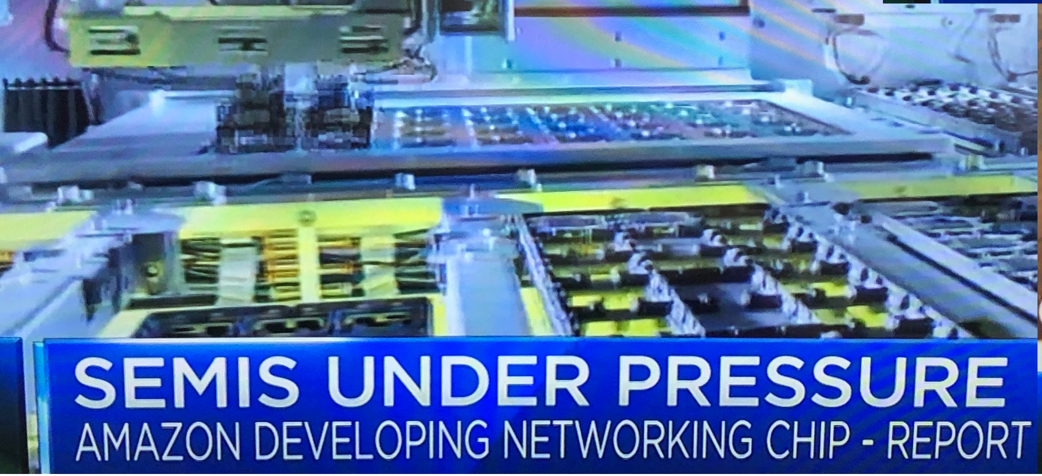
Rolling it all up, cost equivalent versus Nvidia GPU, Tesla claims they can achieve 4x the performance, 1.3x higher performance per watt, and 5x smaller footprint. Tesla

Other CPU Chips

Amazon

Amazon Chips



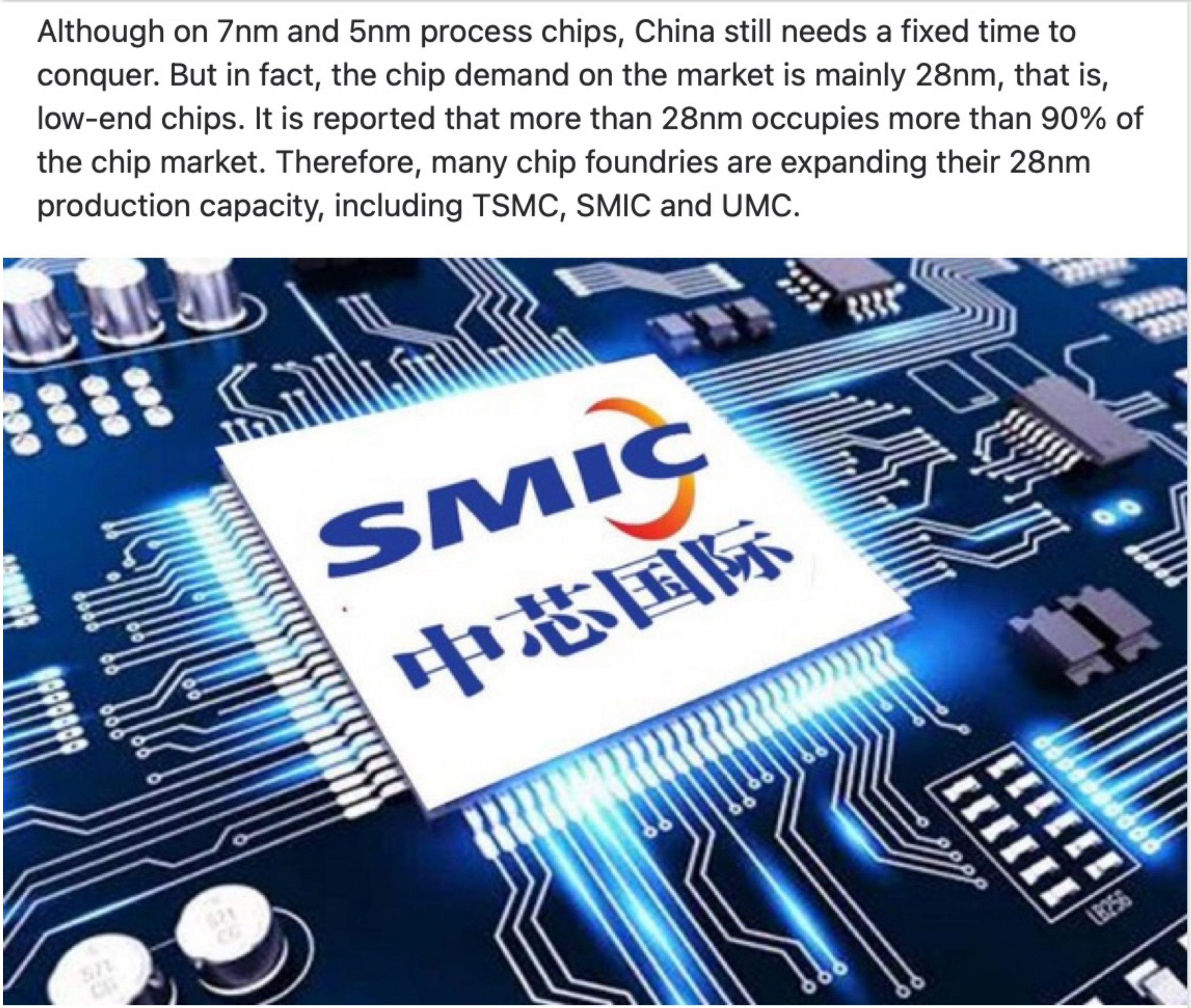


Chip Fab

ARM China
SMIC

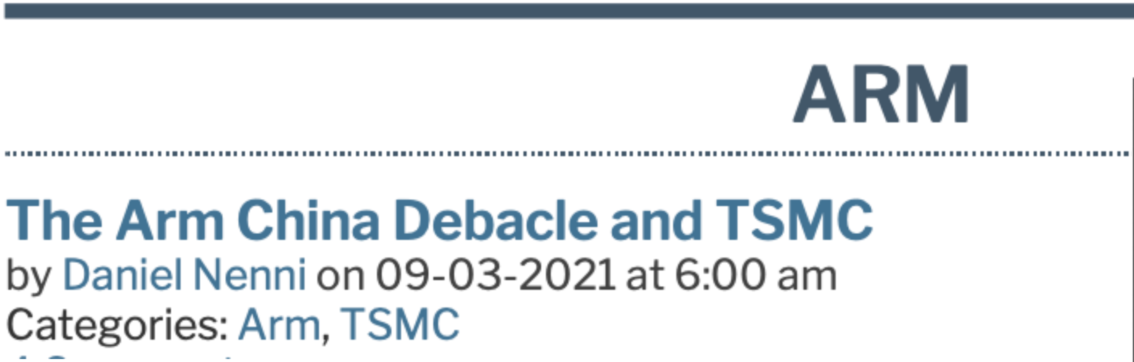
Chip Fab in China: SMIC

90% at 28nm



Although on 7nm and 5nm process chips, China still needs a fixed time to conquer. But in fact, the chip demand on the market is mainly 28nm, that is, low-end chips. It is reported that more than 28nm occupies more than 90% of the chip market. Therefore, many chip foundries are expanding their 28nm production capacity, including TSMC, SMIC and UMC.

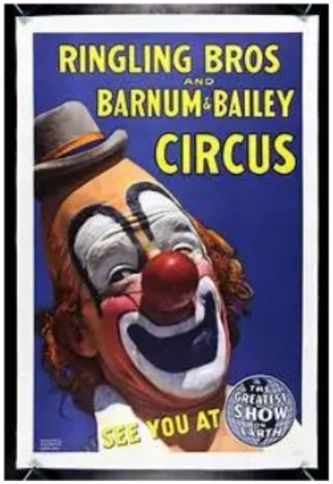
ARM vs China



ARM

The Arm China Debacle and TSMC

by Daniel Nenni on 09-03-2021 at 6:00 am
Categories: Arm, TSMC



ARM China Seizes IP, Relaunches as an 'Independent' Company [Updated]

ARM Refutes Accusations of IP Theft by Its ARM China Subsidiary