CSUN | CALIFORNIA STATE UNIVERSITY NORTHRIDGE

DR JEFF SOFTWARE
INDIE APP DEVELOPER

COMP222

COMP 222

Spring 2023 Rev 1-22-23

Computer Organization (Architecture)

Lecture 4

Dr Jeff Drobman

website → *drjeffsoftware.com/classroom.html*

email → *jeffrey.drobman@csun.edu*

Index (Vol. 4)

Section

Chipsets

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

COMP222

Intel 10th Gen Chips

May 2020

PC

By Michael Sexton Updated May 21, 2020

Mega-Guide: All the Intel Z490 Motherboards for 10th Generation 'Comet Lake-S' CPUs

Intel announced its new 10th Generation "Comet Lake-S" desktop CPUs today, along with a new 400 series of system-level chipsets. This is an all-new platform, and the launch is triggering a wave of new motherboards built on these new chipsets. The Intel Z490 is the high-end chipset designed to support these new CPUs, while boards based on B-series (B460, mainstream) and H-series (H410, basic/budget) chipsets will roll out a smidge later.

What Z490 Is All About

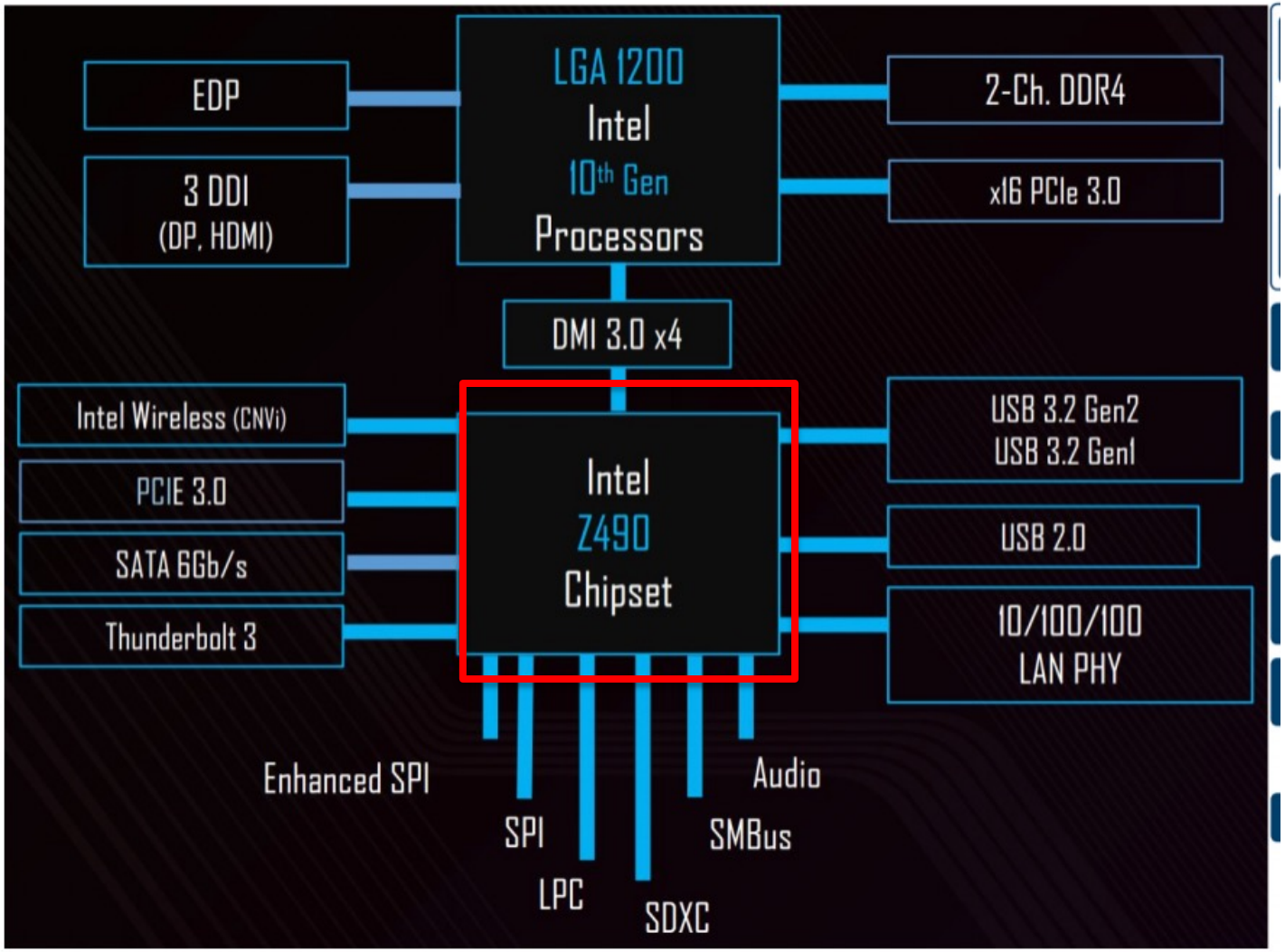
First things first: With these new boards and chipsets comes a new CPU socket for the 10th Generation chips, dubbed "LGA 1200." So, to be clear: Last-generation processors will not work on this platform. Don't try to install an LGA 1151 CPU (the socket used in the last few chip generations) on one of these new LGA 1200 boards, or vice versa. If you do, the best you can hope for is that nothing happens.

Intel 10th Gen Chips

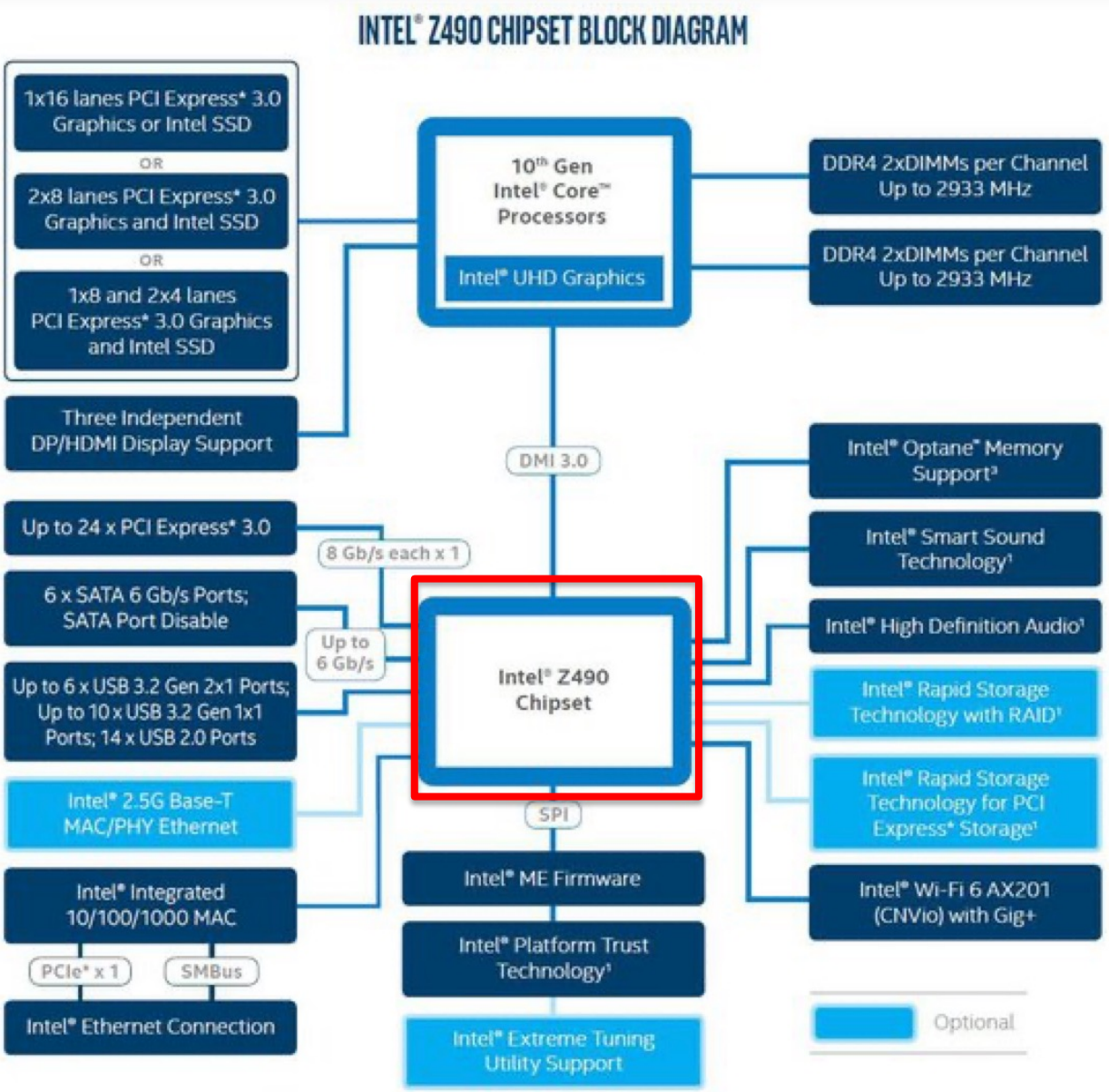
PC

By Michael Sexton Updated May 21, 2020

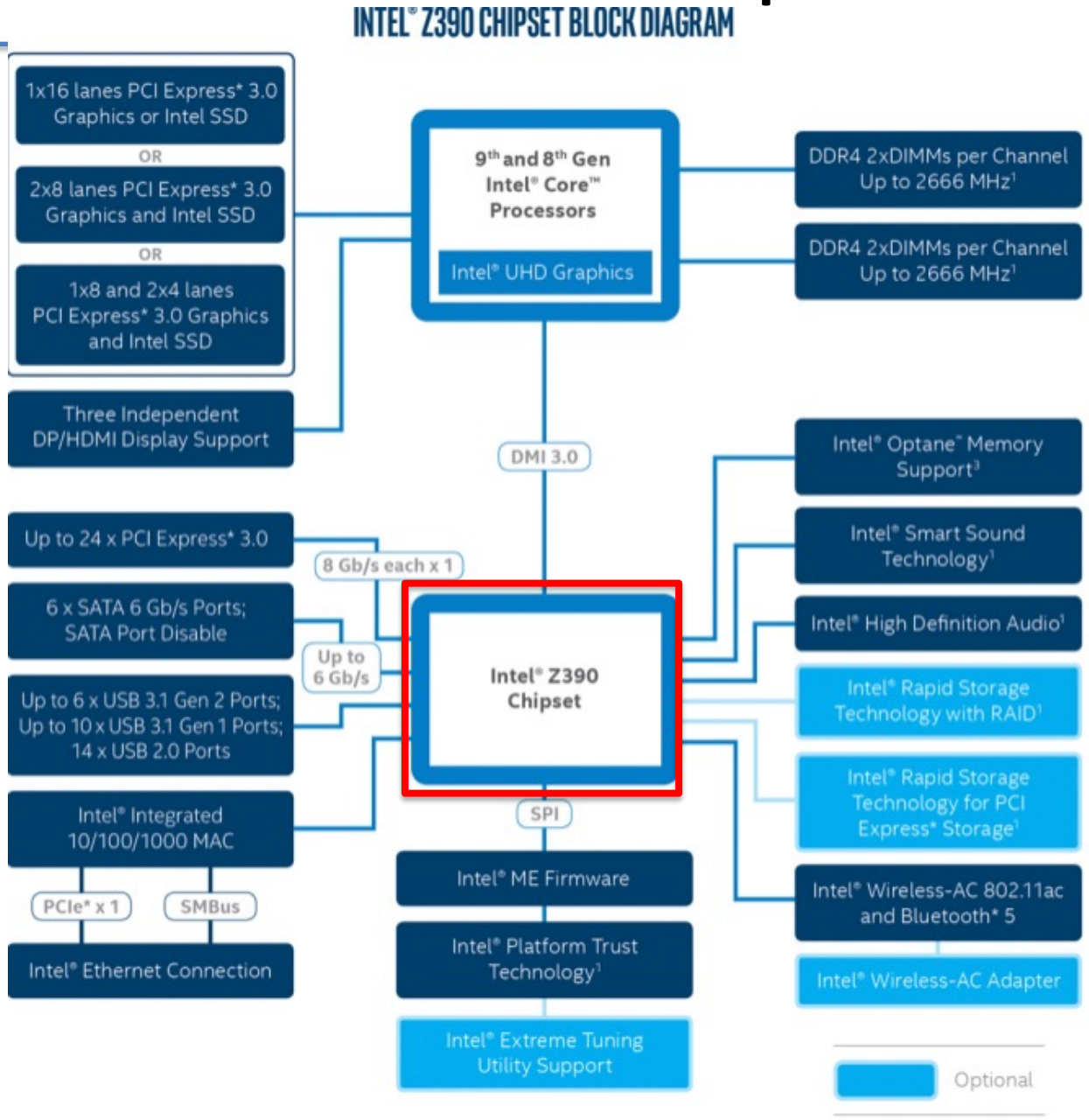
May 2020



Intel Z490 Chipset



Intel Z390 Chipset



CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP222

DR JEFF SOFTWARE
INDIE APP DEVELOPER

Intel 10th Gen PCB

PC

By Michael Sexton Updated May 21, 2020

May 2020

CPU socket

Intel 12th Gen P/E Cores

Quora

Pawel Kraszewski, Developer, researcher, teacher

Answered 5h ago

Some 12th gen Intel processors (Adler Lake) have 2 types of cores.

- P-cores (performance), which *may* support HT
- E-cores (power-efficient), which don't support HT.

For example, i9-12900 has 8 P-cores with HT and 8 E-cores, giving the total of 24 threads.

Intel 12th Gen P/E Cores

Quora

Drazen Zoric, lives in Cork, Ireland

Answered 4h ago

You already got good answers, I will like to extend.

As answered, Intel 12th gen is first x86 having two different core types, P and E. And this was big problem for Microsoft - bugs in Windows 11, and for some software - DRM in games.

But mobile CPUs for long time not only have 2 core types but even 3 or 4 different core types. Apple CPUs have two core types while some Android CPUs have 3 or more. Eg Samsung S20 has this cores:

CPU	Octa-core (2x2.73 GHz Mongoose M5 & 2x2.50 GHz Cortex-A76 & 4x2.0 GHz Cortex-A55) - Global
	Octa-core (1x2.84 GHz Kryo 585 & 3x2.42 GHz Kryo 585 & 4x1.8 GHz Kryo 585) - USA

Intel 12th Gen P/E Cores

Quora

Drazen Zoric, lives in Cork, Ireland

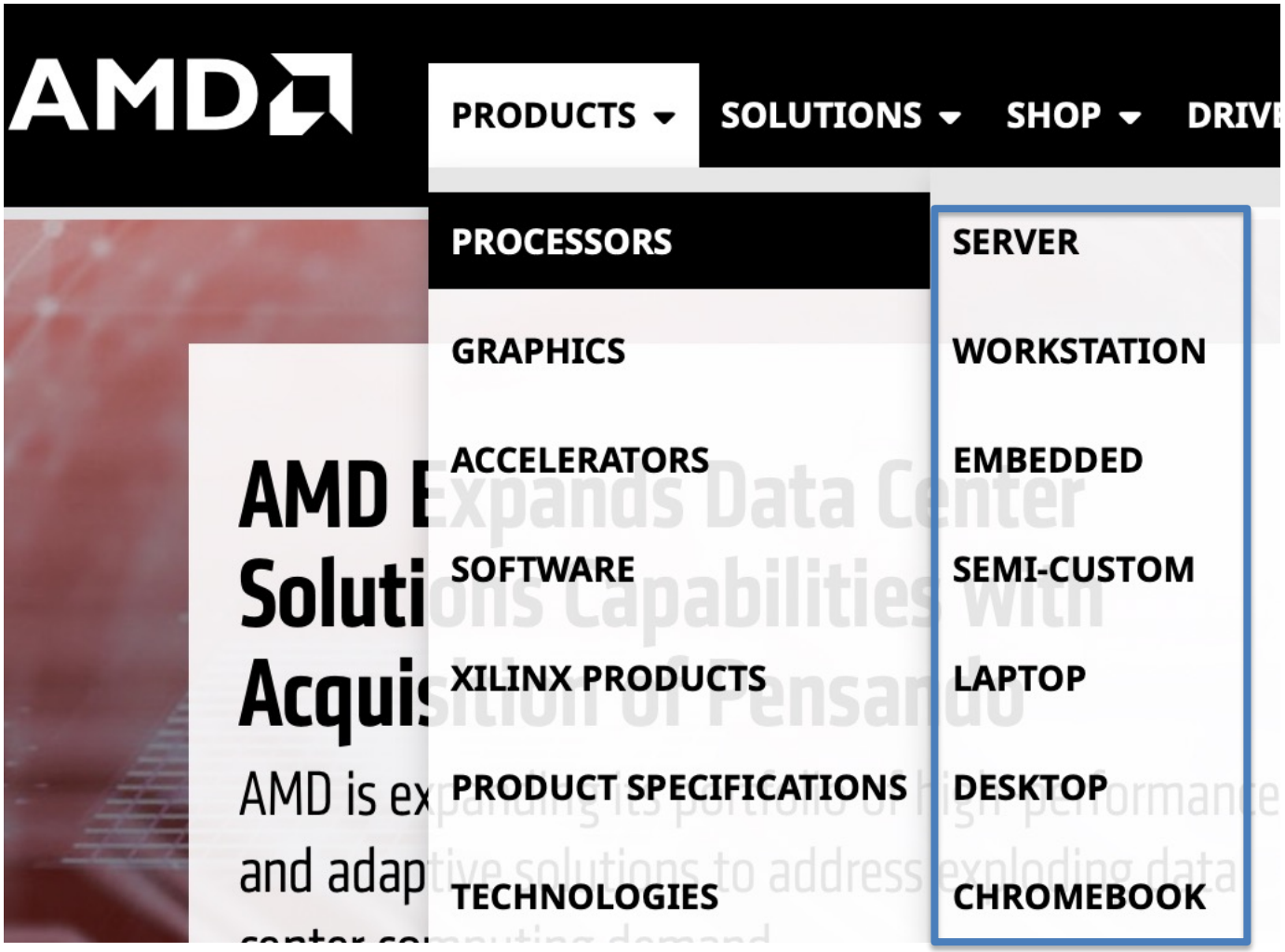
Answered 4h ago

Something else. Previous answers mentioned how 12900K is 8 + 8 core CPU. Remember this when you will read about benchmarking against Apple M1 series where M1 Max is 10 + 2 and M1 Ultra is 20 + 4!

In Intel case E cores are cca 50% performance of P cores. And Hyperthreading is not as a core, it is something like 20 - 30%. So, x86 core with HT is like 1.2 cores. And vary very with type of load, Adobe is even slower with HT enabled.

AMD

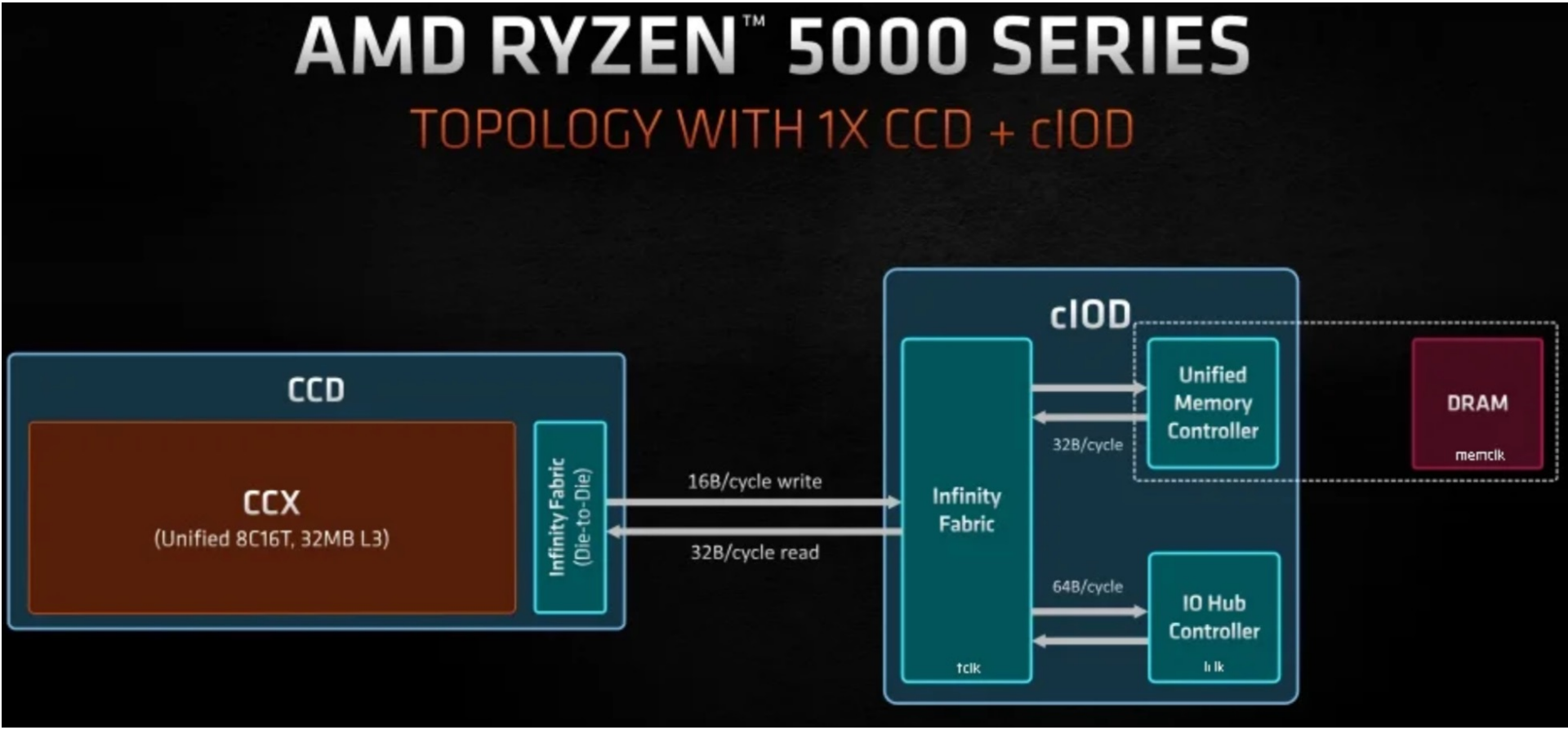
April 2022



AMD

AMD RYZEN™ 5000 SERIES

April 2022

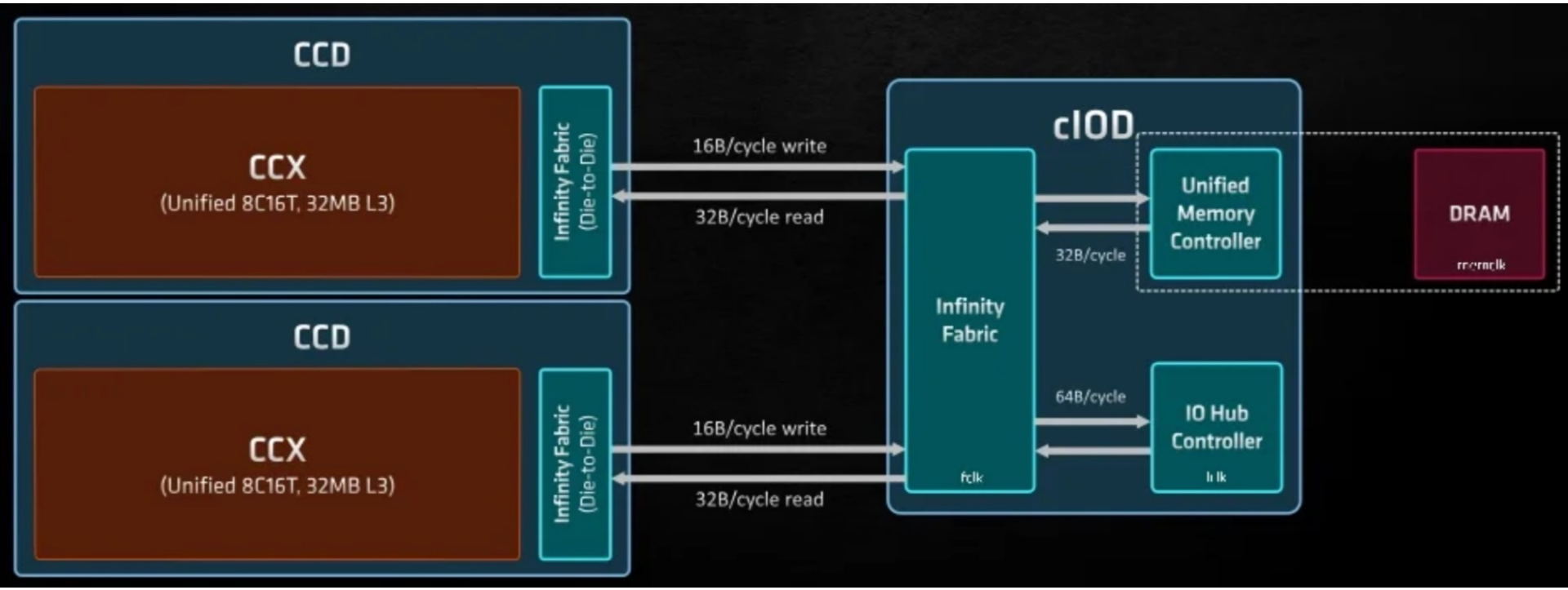


AMD

AMD RYZEN™ 5000 SERIES

April 2022

TOPOLOGY WITH 2X CCD + cIOD



AMD

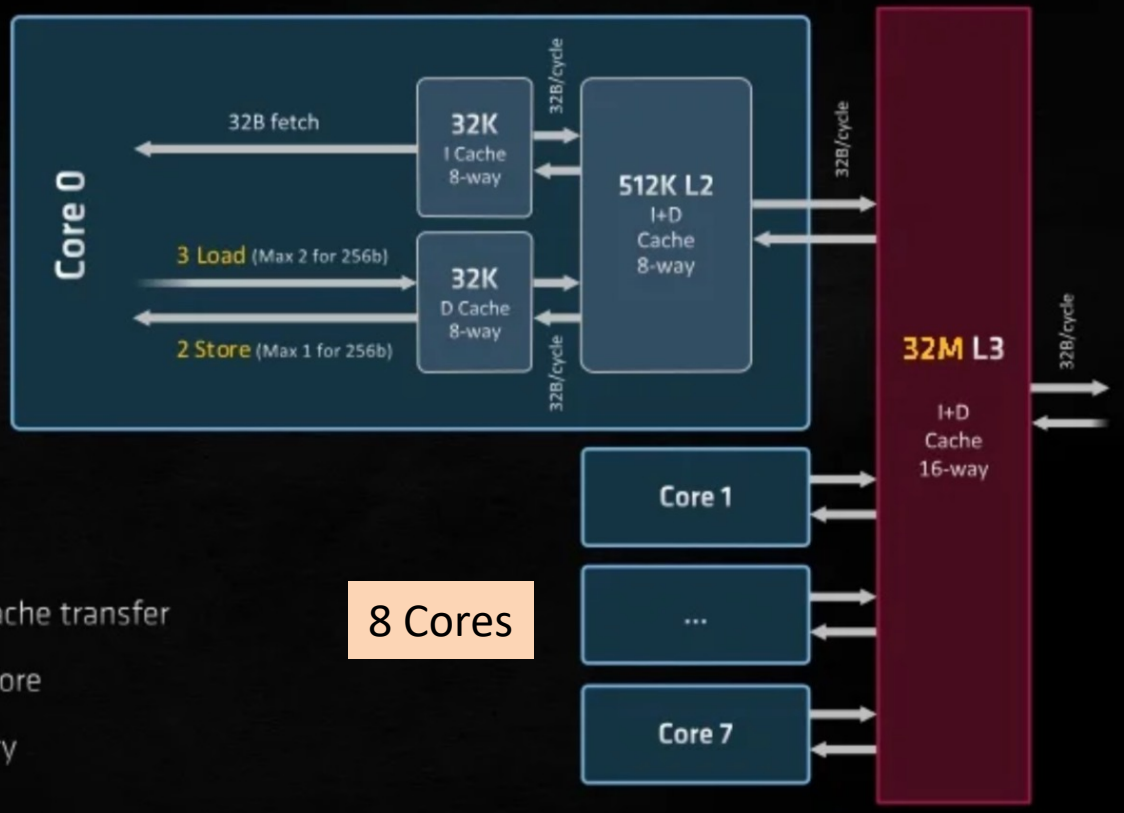
AMD RYZEN™ 5000 SERIES

April 2022

"ZEN 3" CACHE HIERARCHY

(8-CORE)

- Fast private 512K L2 cache
- High bandwidth enables prefetch improvements
- L3 is filled from L2 victims (i.e. mostly exclusive)
- L2 tags duplicated in L3 for probe filtering and fast cache transfer
- 64 outstanding misses supported from L2 to L3 per core
- 192 outstanding misses supported from L3 to memory
- L3 shared among all 8 cores in the complex



Section

PCIe

PCIe

WIKIPEDIA
The Free Encyclopedia

Peripheral Component Interconnect Express

PCI Express logo

Year created	2003; 18 years ago
Created by	Intel · Dell · HP · IBM
Supersedes	AGP · PCI · PCI-X
Width in bits	1 per lane (up to 16 lanes)
No. of devices	1 on each endpoint of each connection.[a]
Speed	Dual simplex (in each direction); examples in single-lane (×1) and 16-lane (×16): **Version 1.x:** 2.5 GT/s **×1:** 250 MB/s **×16:** 4 GB/s **Version 2.x:** 5 GT/s **×1:** 500 MB/s **×16:** 8 GB/s **Version 3.x:** 8 GT/s **×1:** 985 MB/s **×16:** 15.75 GB/s **Version 4.0:** 16 GT/s **×1:** 1.97 GB/s **×16:** 31.5 GB/s **Version 5.0:** 32 GT/s **×1:** 3.94 GB/s **×16:** 63 GB/s **Version 6.0:** 64 GT/s **×1:** 7.88 GB/s

PCIe

PCI Express (**Peripheral Component Interconnect Express**), officially abbreviated as **PCIe** or **PCI-e**,[1] is a high-speed serial computer expansion bus standard, designed to replace the older PCI, PCI-X and AGP bus standards. It is the common motherboard interface for personal computers' graphics cards, hard disk drive host adapters, SSDs, Wi-Fi and Ethernet hardware connections.[2] PCIe has numerous improvements over the older standards, including higher maximum system bus throughput, lower I/O pin count and smaller physical footprint, better performance scaling for bus devices, a more detailed error detection and reporting mechanism (Advanced Error Reporting, AER),[3] and native hot-swap functionality. More recent revisions of the PCIe standard provide hardware support for I/O virtualization.

Defined by its number of lanes,[4] (the number of simultaneous sending and receiving lines of data as in a highway which features traffic in both directions) the PCI Express electrical interface is also used in a variety of other standards, most notably the laptop expansion card interface ExpressCard and computer storage interfaces SATA Express, U.2 (SFF-8639) and M.2.

Format specifications are maintained and developed by the PCI-SIG (PCI Special Interest Group), a group of more than 900 companies that also maintain the conventional PCI specifications.

PCIe

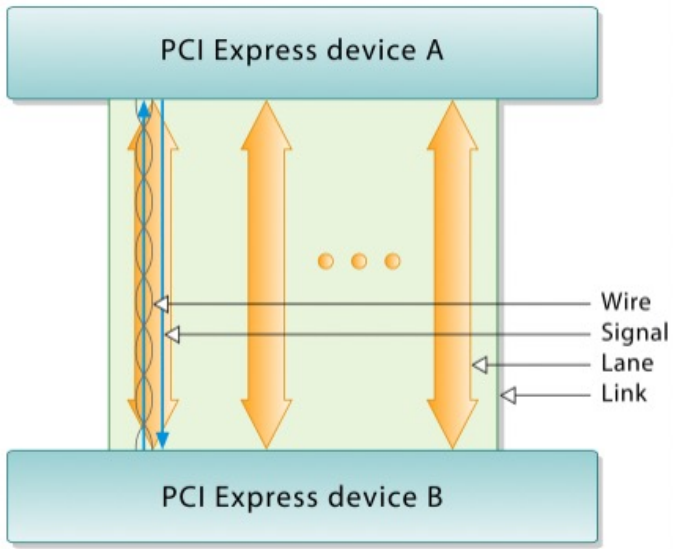
CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

PCIe Lanes

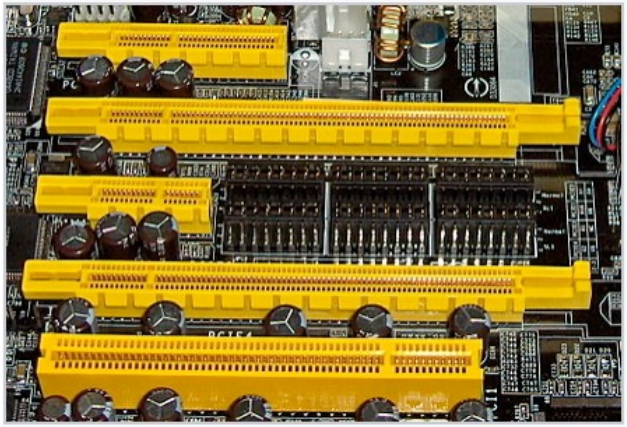
WIKIPEDIA
The Free Encyclopedia

Lane [edit

A lane is composed of two differential signaling pairs, with one pair for receiving data and the other for transmitting. Thus, each lane is composed of four wires or signal traces. Conceptually, each lane is used as a full-duplex byte stream, transporting data packets in eight-bit "byte" format simultaneously in both directions between endpoints of a link.[10] Physical PCI Express links may contain 1, 4, 8 or 16 lanes.[11][5]:4,5[9] Lane counts are written with an "×" prefix (for example, "×8" represents an eight-lane card or slot), with ×16 being the largest size in common use.[12] Lane sizes are also referred to via the terms "width" or "by" e.g., an eight-lane slot could be referred to as a "by 8" or as "8 lanes wide."



A PCI Express link between two devices consists of one or more lanes, which are dual simplex channels using two differential signaling pairs.[5]:3



Various slots on a computer motherboard, from top to bottom:

- PCI Express ×4
- PCI Express ×16
- PCI Express ×1
- PCI Express ×16
- Conventional PCI (32-bit, 5 V)

PCIe Pins – Lane 0

WIKIPEDIA
The Free Encyclopedia

Express connector pinout (×1, ×4, ×8 and ×16 variants)

Pin	Side B	Side A	Description
1	+12 V	PRSNT1#	Must connect to farthest PRSNT2# pin
2	+12 V	+12 V	Main power pins
3	+12 V	+12 V	
4	Ground	Ground	
5	SMCLK	TCK	SMBus and JTAG port pins
6	SMDAT	TDI	
7	Ground	TDO	
8	+3.3 V	TMS	
9	TRST#	+3.3 V	
10	+3.3 V aux	+3.3 V	Standby power
11	WAKE#	PERST#	Link reactivation; fundamental reset [23]
Key notch			
12	CLKREQ#[24]	Ground	Clock Request Signal
13	Ground	REFCLK+	Reference clock differential pair
14	HSOp(0)	REFCLK–	Lane 0 transmit data, + and –
15	HSOn(0)	Ground	
16	Ground	HSIp(0)	Lane 0 receive data, + and –
17	PRSNT2#	HSIn(0)	
18	Ground	Ground	
PCI Express ×1 cards end at pin 18			

PCIe Pins – Lanes 1-4

19	HSOp(1)	Reserved	Lane 1 transmit data, + and –
20	HSOn(1)	Ground	
21	Ground	HSIp(1)	Lane 1 receive data, + and –
22	Ground	HSIn(1)	
23	HSOp(2)	Ground	Lane 2 transmit data, + and –
24	HSOn(2)	Ground	
25	Ground	HSIp(2)	Lane 2 receive data, + and –
26	Ground	HSIn(2)	
27	HSOp(3)	Ground	Lane 3 transmit data, + and –
28	HSOn(3)	Ground	
29	Ground	HSIp(3)	Lane 3 receive data, + and –
30	PWRBRK#[25]	HSIn(3)	
31	PRSNT2#	Ground	
32	Ground	Reserved	
PCI Express ×4 cards end at pin 32			
33	HSOp(4)	Reserved	Lane 4 transmit data, + and –
34	HSOn(4)	Ground	
35	Ground	HSIp(4)	Lane 4 receive data, + and –
36	Ground	HSIn(4)	

WIKIPEDIA
The Free Encyclopedia

PCIe Alternates

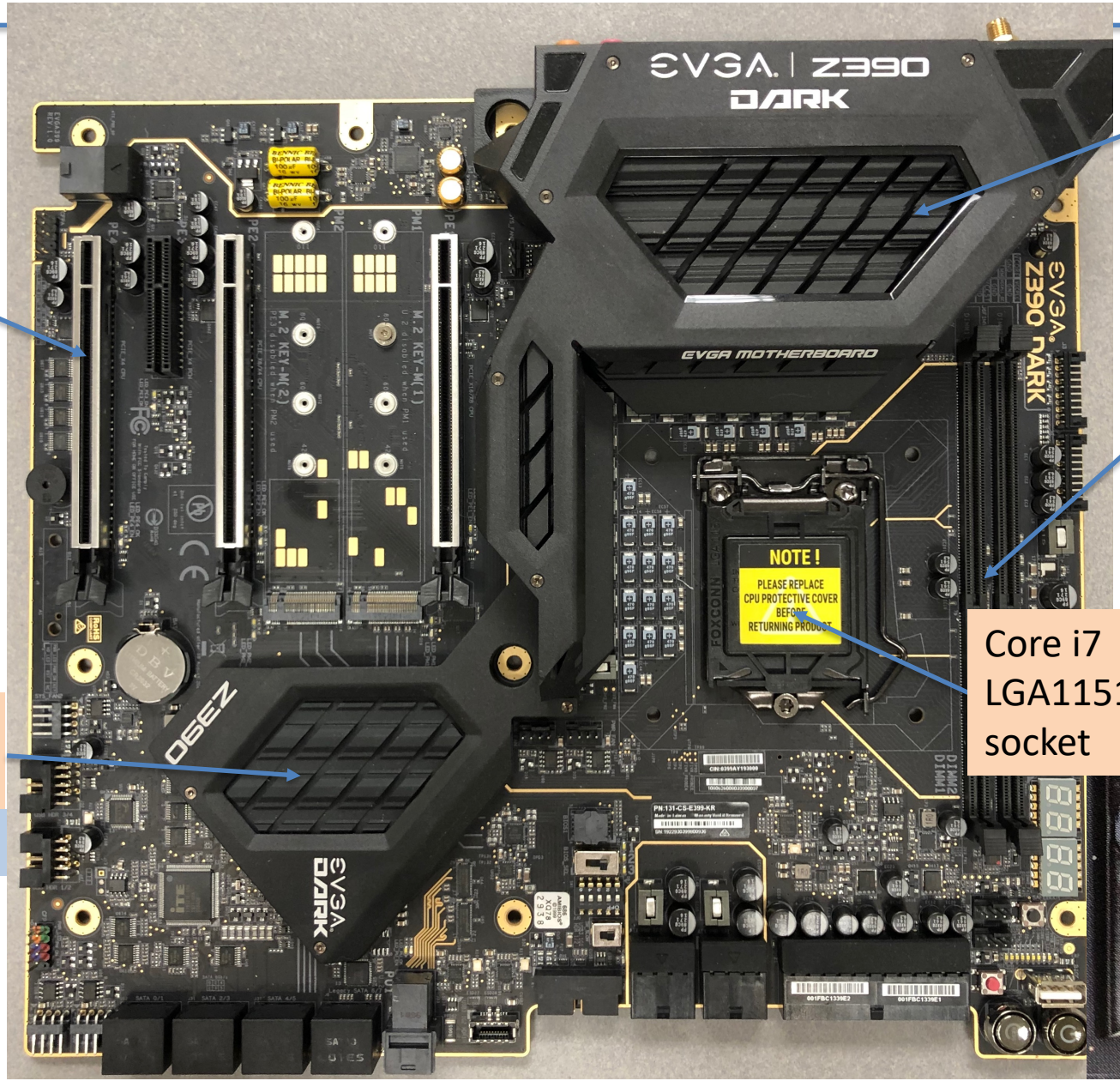
Other communications standards based on high bandwidth serial architectures include InfiniBand, RapidIO, HyperTransport, Intel QuickPath Interconnect, and the Mobile Industry Processor Interface (MIPI).

On 11 March 2019, Intel presented Compute Express Link (CXL), a new interconnect bus, based on the PCI Express 5.0 physical layer infrastructure. The initial promoters of the CXL specification included: Alibaba, Cisco, Dell EMC, Facebook, Google, HPE, Huawei, Intel and Microsoft.[

Section

PCB Sockets

PCB: Intel 8-9th Gen



Power supply + cooling

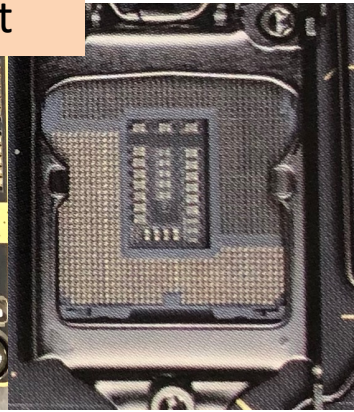
PCIex4

Dual Channel DDR i/f

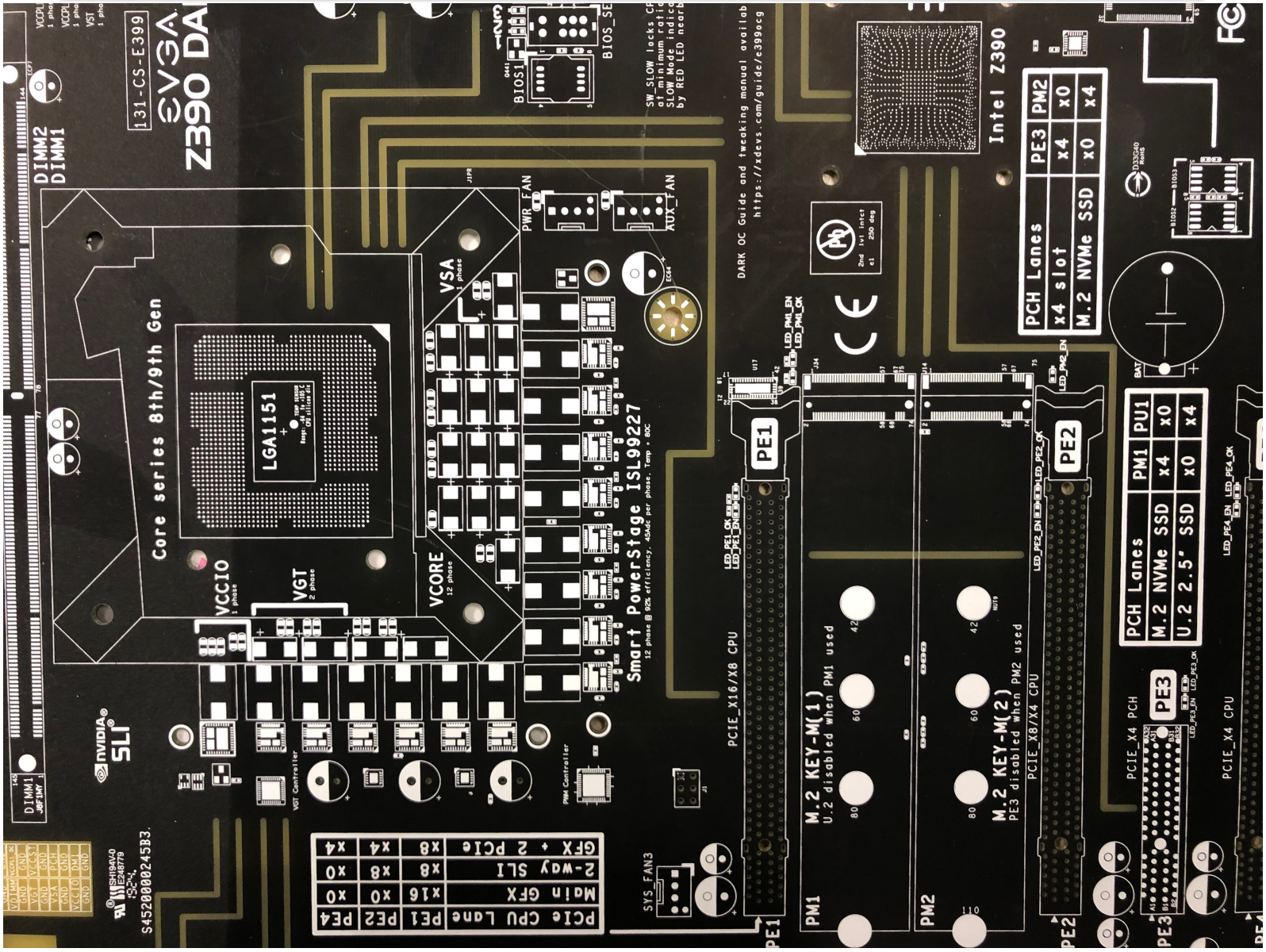
Core i7 LGA1151 socket

Z390 chipset

cooling



Bare PCB: 8-9th Gen



PCB Sockets

Quora

Brett Bergan
Building PC's for 25 years · Feb 23

Why can AMD keep the same CPU socket across so many generations but Intel can't?

AMD is moving to DDR5 and CPU packages without pins.

Unfortunately AMD was unable to keep one AM4 platform since the birth of Ryzen. It had to make The Ryzen 5000 series incompatible with the 300 series motherboards and earlier. So AMD has at least tried to keep consumers with good hardware capable of upgrading across several generations. Unfortunately it just isn't possible to forestall obsolescence indefinitely when so many new features are added with every update.

Intel could have at least made an effort in this direction, but they really haven't—essentially requiring a new motherboard with every CPU upgrade. Intel could easily have put 10th gen processors on Coffee Lake motherboards, but they opted to pull the wool over us and offer "Socket 1200" with future promises of PCIe 4.0 and the hodgepodge of other nonexistent perks—all in a fruitless attempt to steer you away from buying AMD.

PCB Sockets

Quora

Brett Bergan
Building PC's for 25 years · Feb 23

Why can AMD keep the same CPU socket across so many generations but Intel can't?

Both AMD and Intel are starting with a clean slate now with a new socket and DDR5. It is my hope that Intel will curb its shenanigans and stick with mutually compatible Socket 1700 processors for at least three years. Let motherboard manufacturers get a few solid designs that sell well into the millions so that they can make a profit before being plunged into the chaos of the next redesign schedule.

Section

FPGA

FPGA

Quora

Drazen Zoric · Follow
Lives in Cork, Ireland · Sep 20

LUT is short for Look Up Table where user can program 3 to 1 bit mapping, FA is full adder cell used to add two bits, MUX is a switch and DFF is D type Flip-Flop, like 1 bit memory.

FPGA compiler take program written in VHDL or Verilog and converts into format suitable for specific FPGA which is a bitmap file used to program logic cells and how they are connected to interconnect bus.

FPGA has two main problems:

- Speed - clocks are not high cause of complexity, eg 1 GHz is very fast FPGA, cause of complex interconnect clock delays are huge and design which works in FPGA might/will not work in ASIC
- Price - 10,000 USD is cheap for complex FPGA. But look at this one from DigiKey,

Next answer

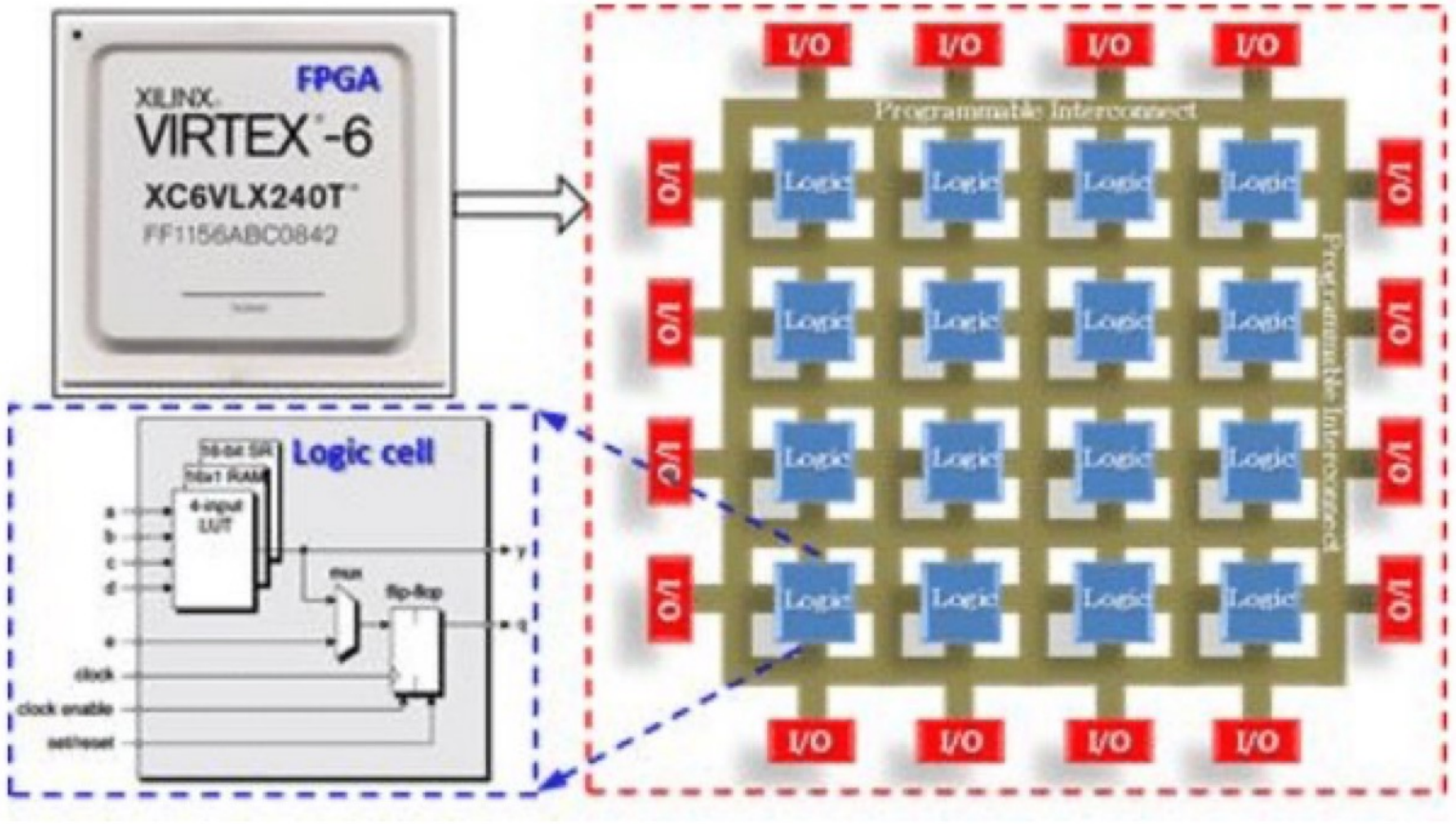
FPGA

Drazen Zoric · Follow
Lives in Cork, Ireland · Sep 20

Quora

Related Why not use an FPGA and update the circuitry instead of making a new chip for every generation of CPU or GPU?

FPGA is programmable digital hardware and as such it has very specific architecture. It looks like this:



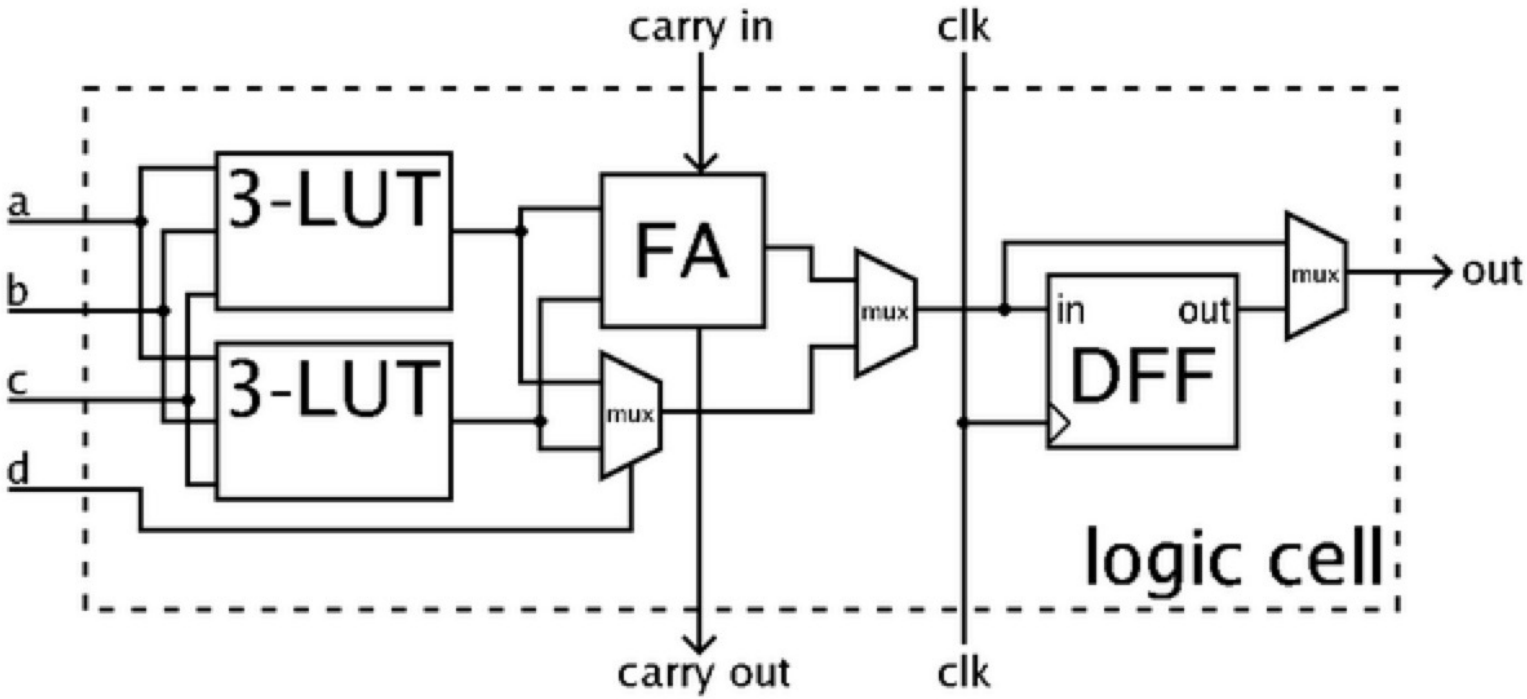
It has many, really many, logic cells which are programmable and interconnect as above.

FPGA

Quora

Drazen Zoric · Follow
Lives in Cork, Ireland · Sep 20

Logic cells are not complex, something like this:



Section

Chip Design

ASIC vs FPGA

Quora

Jeff Drobman
Lecturer at California State University, Northridge (2016–present) · Just now ·

an ASIC can be any digital or even mixed-signal System-on-a-Chip (SoC) that is designed as "semi-custom". that is, a chip not designed entirely from scratch (fully custom), but uses pre-designed building blocks (hardware macros) and other IP supplied by ASIC vendors such as LSI Logic (who invented it under founding CEO Wilf Corrigan in 1980.)

an FPGA, as its name implies, is a Field Programmable digital only array of small pre-fab logic blocks of mostly NAND gates.

both are functionally designed at the RTL level by a hardware description language (HDL) such as Verilog or VHDL. All chip designs use common tool sets and simulation flows, including low-level Spice analog simulations.

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP222

ASIC

WIKIPEDIA
The Free Encyclopedia

Application-specific integrated circuit

An **application-specific integrated circuit** (**ASIC** /ˈeɪsɪk/) is an integrated circuit (IC) chip customized for a particular use, rather than intended for general-purpose use. For example, a chip designed to run in a digital voice recorder or a high-efficiency video codec (e.g. AMD VCE) is an ASIC. Application-specific standard product (ASSP) chips are intermediate between ASICs and industry standard integrated circuits like the 7400 series or the 4000 series.[1] ASIC chips are typically fabricated using metal-oxide-semiconductor (MOS) technology, as MOS integrated circuit chips.[2]

As feature sizes have shrunk and design tools improved over the years, the maximum complexity (and hence functionality) possible in an ASIC has grown from 5,000 logic gates to over 100 million. Modern ASICs often include entire microprocessors, memory blocks including ROM, RAM, EEPROM, flash memory and other large building blocks. Such an ASIC is often termed a SoC (system-on-chip). Designers of digital ASICs often use a hardware description language (HDL), such as Verilog or VHDL, to describe the functionality of ASICs.[1]

Field-programmable gate arrays (FPGA) are the modern-day technology improvement on breadboards, meaning that they are not made to be application-specific as opposed to ASICs. Programmable logic blocks and programmable interconnects allow the same FPGA to be used in many different applications. For smaller designs or lower production volumes, FPGAs

- Dedicated single application vs. programmable logic

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE
COMP222

WIKIPEDIA
The Free Encyclopedia

ASIC Design Flow

By the late 1990s, logic synthesis tools became available. Such tools could compile HDL descriptions into a gate-level netlist. Standard-cell integrated circuits (ICs) are designed in the following conceptual stages referred to as electronics design flow, although these stages overlap significantly in practice:

1. **Requirements engineering**: A team of design engineers starts with a non-formal understanding of the required functions for a new ASIC, usually derived from requirements analysis.
2. **Register-transfer level (RTL) design**: The design team constructs a description of an ASIC to achieve these goals using a hardware description language. This process is similar to writing a computer program in a high-level language.
3. **Functional verification**: Suitability for purpose is verified by functional verification. This may include such techniques as logic simulation through test benches, formal verification, emulation, or creating and evaluating an equivalent pure software model, as in Simics. Each verification technique has advantages and disadvantages, and most often several methods are used together for ASIC verification. Unlike most FPGAs, ASICs cannot be reprogrammed once fabricated and therefore ASIC designs that are not completely correct are much more costly, increasing the need for full test coverage.
4. **Logic synthesis**: Logic synthesis transforms the RTL design into a large collection called of lower-level constructs called standard cells. These constructs are taken from a standard-cell library consisting of pre-characterized collections of logic gates performing specific functions. The standard cells are typically specific to the planned manufacturer of the ASIC. The resulting collection of standard cells and the needed electrical connections between them is called a gate-level netlist.
5. **Placement**: The gate-level netlist is next processed by a placement tool which places the standard cells onto a region of an integrated circuit die representing the final ASIC. The placement tool attempts to find an optimized placement of the standard cells, subject to a variety of specified constraints.
6. **Routing**: An electronics routing tool takes the physical placement of the standard cells and uses the netlist to create the electrical connections between them. Since the search space is large, this process will produce a "sufficient" rather than "globally optimal" solution. The output is a file which can be used to create a set of photomasks enabling a semiconductor fabrication facility, commonly called a 'fab' or 'foundry' to manufacture physical integrated circuits. Placement and routing are closely interrelated and are collectively called place and route in electronics design.
7. **Sign-off**: Given the final layout, circuit extraction computes the parasitic resistances and capacitances. In the case of a digital circuit, this will then be further mapped into delay information from which the circuit performance can be estimated, usually by static timing analysis. This, and other final tests such as design rule checking and power analysis collectively called signoff are intended to ensure that the device will function correctly over all extremes of the process, voltage and temperature. When this testing is complete the photomask information is released for chip fabrication.

Chip Design Flow

Quora

==========Front End==========

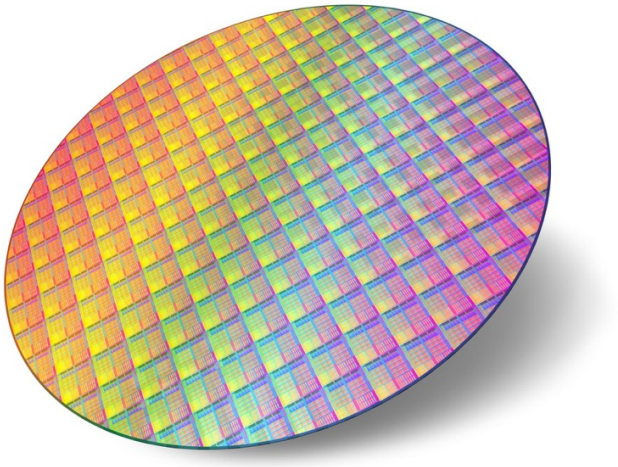
1. Design spec (at the same time start to think about test plan)
2. HDL design entry
3. Functional verification <=====it could take several iterations of steps 2 and 3
4. Logic synthesis
5. Testing
6. Gate level simulation
7. Pre layout simulation

==========Back End==========

8. floor plan
9. Placement
10. Routing
11. Parasitic extraction
12. Fab unit

It is important to think about test plan at the beginning of a design cycle.

Open Fab



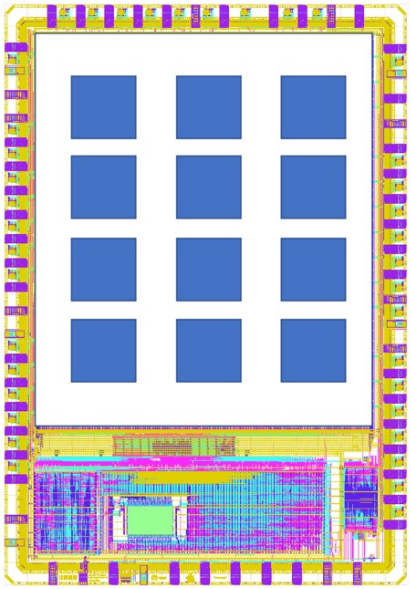
University-optimized shuttles

We have **chipIgnite** shuttles in **June** and **November** that are optimized for the University schedule.

These shuttle are scheduled to return silicon back to students in the next session (**October** and **March**).

Aggregate multiple projects into a single tile

Courses with smaller digital or analog student projects can be aggregated into a single project slot making fabrication very cost effective for larger class sizes.



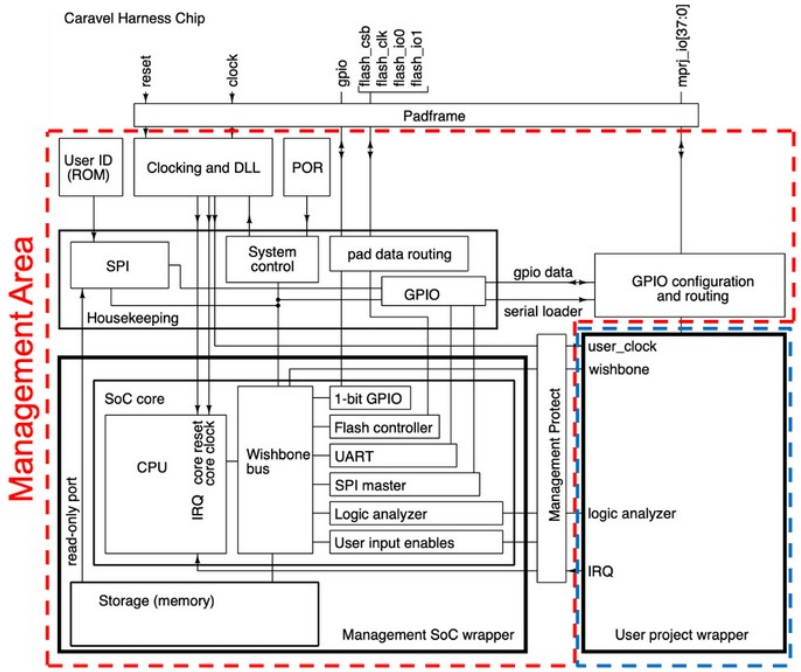
Open Fab

A full open-source reference design

...with easy access from our git **repo**.

Open-source designs enable you to focus the time and effort for design around specific goals.

Reference designs allow students to complete their own full design in a time-limited course session.



Section

U Wisc

Advanced Computer Architecture

U Wisc Slides

https://ece757.ece.wisc.edu/lect03-cores-multithread.pdf

ECE/CS 757: Advanced Computer Architecture II

Instructor:Mikko H Lipasti

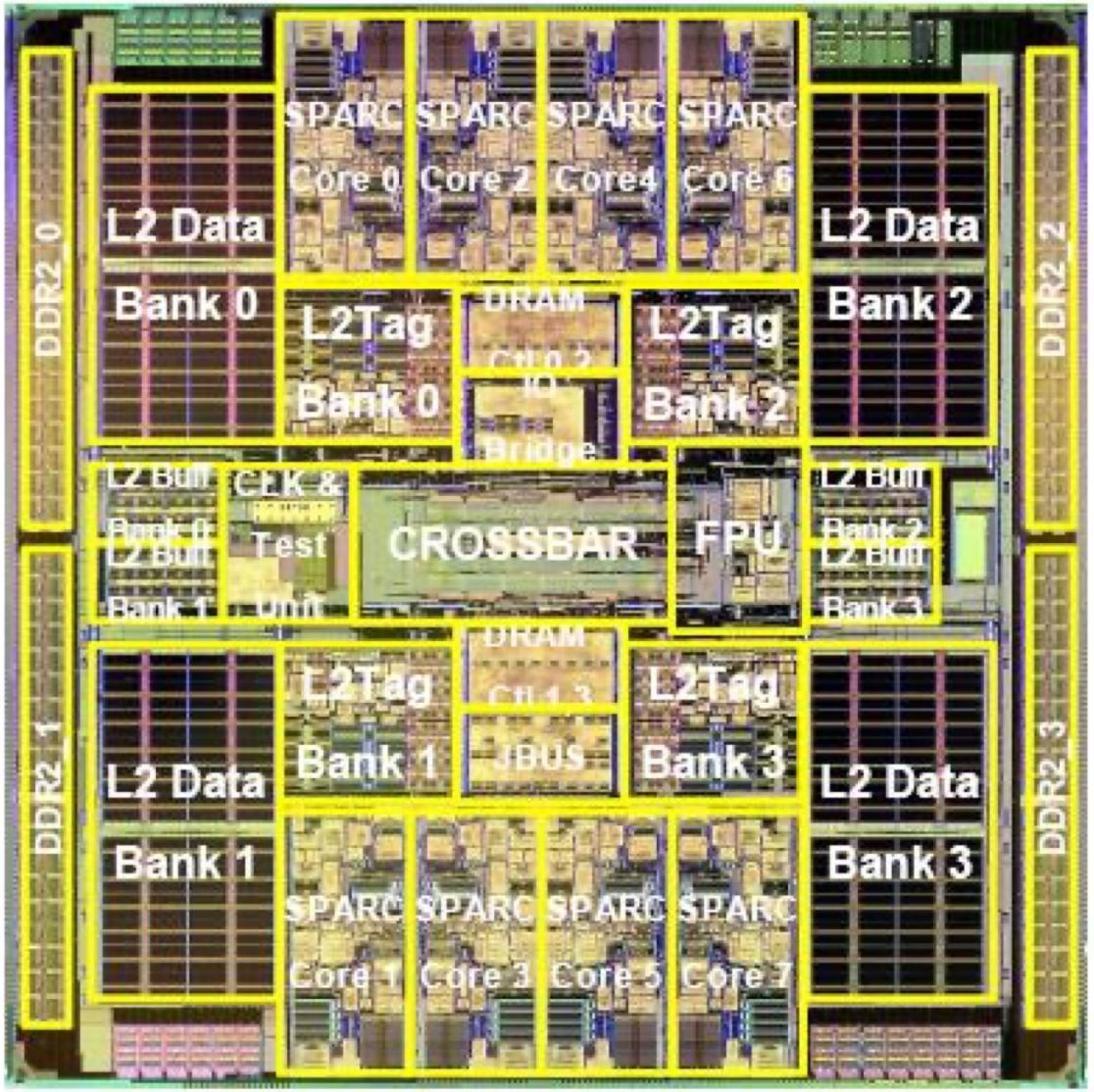
Spring 2017

University of Wisconsin-Madison

Lecture notes based on slides created by John Shen, Mark Hill, David Wood, Guri Sohi, Jim Smith, Natalie Enright Jerger, Michel Dubois, Murali Annavaram, Per Stenström and probably others

U Wisc Slides

Ultrasparc T1 Die Photo [Source: J. Laudon]



Features:

- 8 64-bit Multithreaded SPARC Cores
- Shared 3 MB, 12-way 64B line writeback L2 Cache
- 16 KB, 4-way 32B line ICache per Core
- 8 KB, 4-way 16B line write-through DCache per Core
- 4 144-bit DDR-2 channels
- 3.2 GB/sec JBUS I/O

Technology:

- TI's 90nm CMOS Process
- 9LM Cu Interconnect
- 63 Watts @ 1.2GHz/1.2V
- Die Size: 379mm²
- 279M Transistors
- Flip-chip ceramic LGA

U Wisc Slides

CMP Examples

- Chip Multiprocessors (CMP)
- Becoming very popular

Processor	Cores/ chip	Multi-threaded?	Resources shared
IBM Power 4	2	No	L2/L3, system interface
IBM Power7	8	Yes (4T)	Core, L2/L3, system interface
Sun Ultrasparc	2	No	System interface
Sun Niagara	8	Yes (4T)	Everything
Intel Pentium D	2	Yes (2T)	Core, nothing else
AMD Opteron	2	No	System interface (socket)

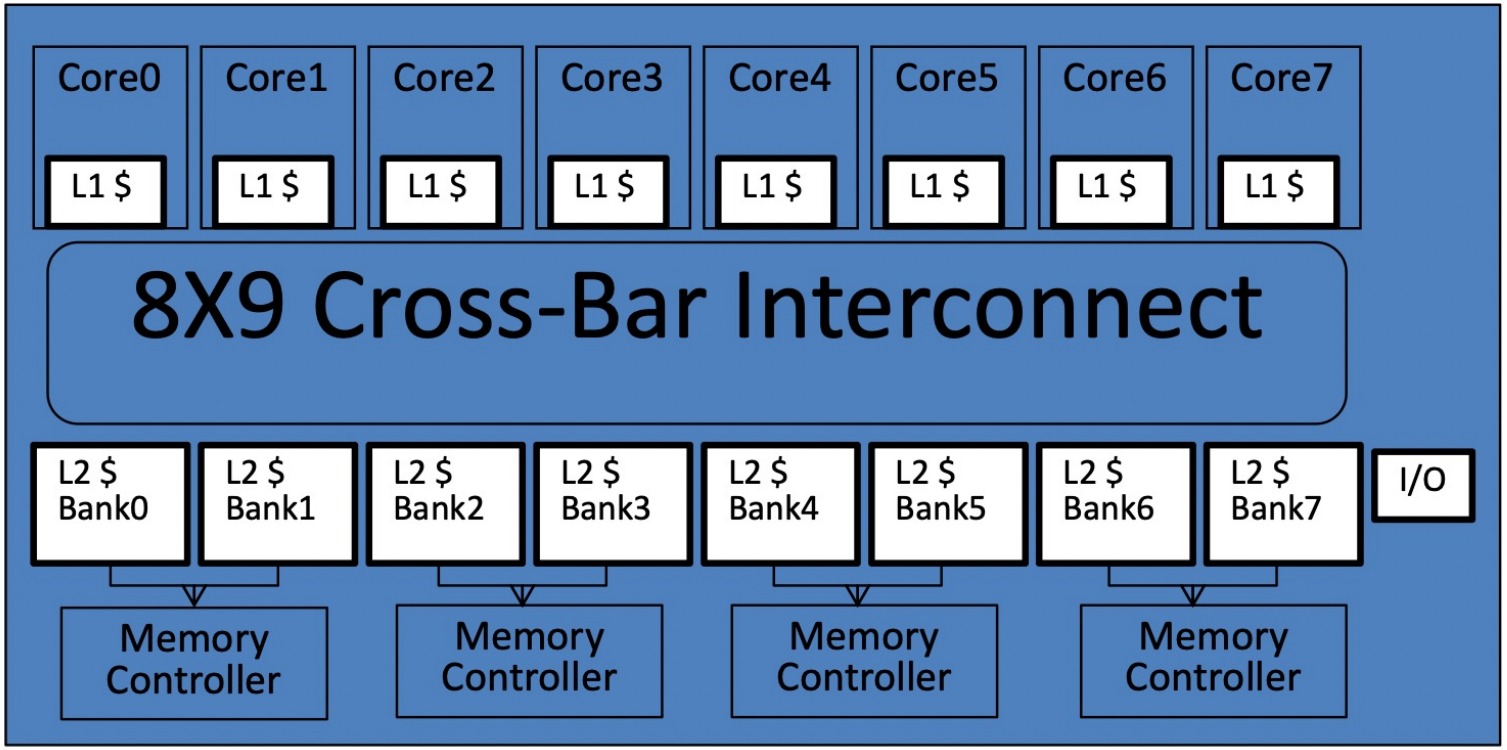
U Wisc Slides: L2 Buses

Multicore Interconnects

- Bus/crossbar - dismiss as short-term solutions?
- Point-to-point links, many possible topographies
  - 2D (suitable for planar realization)
    - Ring
    - Mesh
    - 2D torus
  - 3D - may become more interesting with 3D packaging (chip stacks)
    - Hypercube
    - 3D Mesh
    - 3D torus

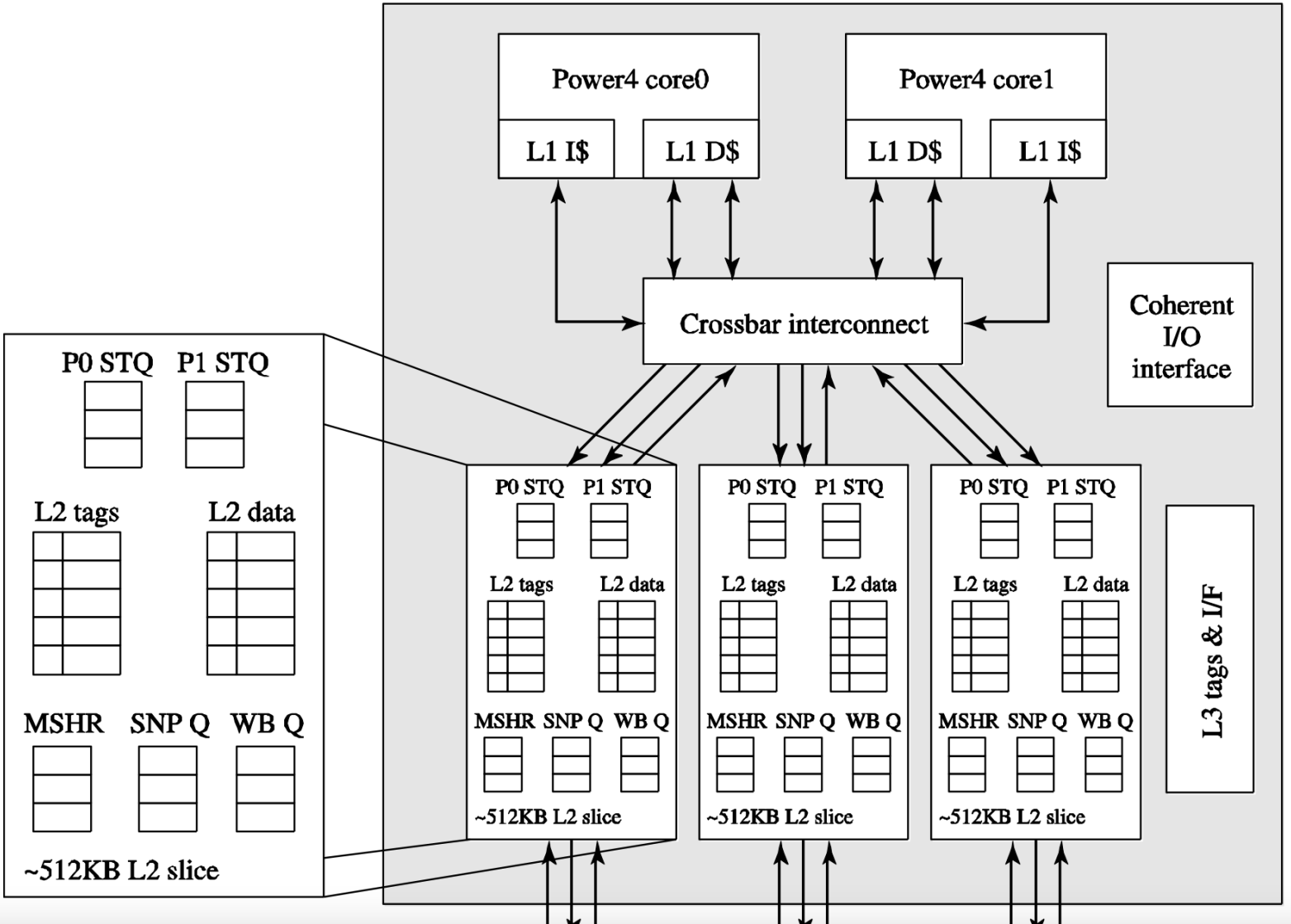
U Wisc Slides: L2 Buses

Cross-bar (IBM Power4-Power8)



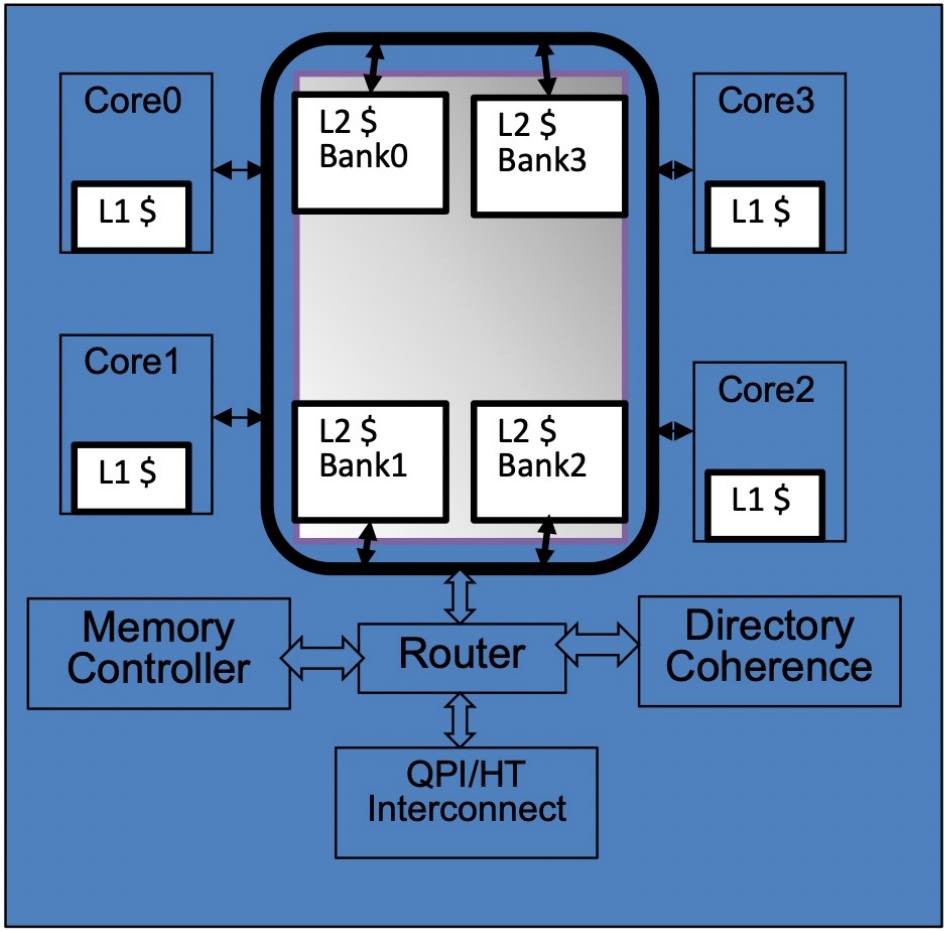
U Wisc Slides: L2 Buses

IBM Power4: Example CMP



U Wisc Slides: L2 Buses

On-Chip Ring (e.g. Intel)



Section

ACM

Special Purpose Computer Architecture

ACM Lecture

Jan 2022

Chip Measuring Contest

The benefits of purpose-built chips

Jessie Frazelle

Alan Kay once said, "People who are really serious about software should make their own hardware." We are now seeing product companies genuinely live up to this value. On August 19, 2021, Tesla showed off Dojo, its new chip used for training neural networks. You might imagine the lead of an article about this something along the lines of, "A company that is not in the business of making chips, made its own chip for its own specific use case, wat!" That part of the announcement was not so shocking because it was something seen before with Tesla and its FSD (full self-driving) computer, with Cisco and its network ASICs, and recently with Apple's M1 chip. In reality the shocking part of the Tesla announcement was not their chip but their humanoid robot, but we'll save that for another article.

ACM Lecture

Jessie Frazelle Jan 2022

Companies such as Tesla and Apple are so serious about their software (and hardware) that they bite off more and more challenging problems lower in the stack to give their customers better products. Additionally, with Moore's law slowing down, chip manufacturers are forced to get more and more creative in their approaches, resulting in diversification among chips. It is an exciting time to be alive when the incumbents known as the chip vendors are being outdone, in the very technology that is their bread and butter, by their previous customers.

ACM Lecture

Jessie Frazelle

Jan 2022

AI Chips

GPUs were originally designed for graphics, hence the name *graphics* processing unit. GPUs are not actually made for neural networks; however, they tend to be used for this solely because they outperform CPUs since they have lots of cores for running computations in parallel. In 2016, Google introduced the TPU (tensor processing unit), which is an ASIC (application-specific integrated circuit) made for neural networks. ASICs made explicitly for neural networks tend to be very good at matrix multiplication and floating-point operations since that is largely what training a neural network is all about. This is why you often see these types of chips advertised by comparing FLOPS (floating-point operations per second). Traditional GPUs focus on calculations for placing pixels; they are also capable of matrix multiplication and floating-point operations but not to the same scale as those made specifically for neural networks.

ACM Lecture

If you are doing any complex work with neural networks, you have only a few good options for compute. Traditionally, the champion in this space has been Nvidia's A100. A company like Tesla, that competes directly with Google's self-driving car experiments, likely does not want its data in Google's cloud. So the A100 is its only option. The A100 comes at a steep price, and Nvidia seems to take advantage of its domination in this space. Because of Nvidia's high margins, Tesla could get better unit economics and performance by making its own chips. Because of the cost of designing the chip, building the software, manufacturing, and maintenance, however, Tesla's strategy is likely less a result of economics and more because of vertical integration and the performance benefits of designing to its specific use case.

ACM Lecture

RISC architectures have fewer instructions but are more like Legos: They have all the building blocks for the complex instructions a CISC architecture provides, while also having the flexibility to build whatever the user wants. In a RISC-based system, since there are fewer instructions, more of them are required to do complex tasks; however, processing them can be more efficient. For a CISC-based architecture, it is harder to be as efficient because of the number of instructions and their complexity. (Intel started marketing its processors as RISC by adding a decoding stage to turn CISC instructions into RISC instructions [medium.com]. The advantages of RISC persist because of the fixed length; CISC still has to figure out the length of the instructions.) Using a RISC architecture leads to better power efficiency and performance.

One design detail of the M1 processor to point out is the large number of encoders and decoders. This can be accomplished only with a RISC-based architecture because of the fixed-length instructions. CISC-based architectures have variable-length instructions and lots of complex instructions. It is a bit of a meme that no one knows all the instructions available in x86 [twitter.com], but there are ways of discovering hidden instructions [github.com]. The fixed length of instructions means that RISC-based architectures require a simpler decode leading to less circuitry, heat, and power consumption.

ACM Lecture

The M1 takes advantage of OoOE (out-of-order execution) as a way to execute more instructions in parallel without exposing that capability as multiple threads. While you might be thinking, [yawn] "Intel and AMD do that as well," there is a core difference with the M1 chip. For OoOE to spread its wings and fly, a large buffer of micro-operations is needed; then the hardware can more easily find instructions to run in parallel. Decoders convert the machine-code instructions into micro-ops to pass off to the instruction buffer. Intel and AMD processors typically have four decoders. M1 has eight decoders and an instruction buffer three times larger than the industry norm. This means the M1 processor can more easily find instructions to run in parallel.

Now you might be thinking, Why don't AMD and Intel add more decoders? Because CISC-based architectures have variable-length instructions, it is nontrivial for the decoders to split up a stream of bytes into instructions because they have no idea where the next instruction starts. CISC decoders have to analyze each instruction to understand how long it is. AMD and Intel deal with this by brute force. They attempt to decode instructions at every possible starting point, making the decoder step too complex to add more decoders.

ACM Lecture

Jessie Frazelle

Jan 2022

Cloud Provider	AWS	Azure	GCP	GCP	GCP	GCP	GCP
Type	p4d.24xlarge[1]	Standard_ND96asr_v4[2]	v3-8[3]	v3-32[4]	v3-64[4]	a2-highgpu-8g[5]	a2-highgpu-16g[5]
Accelerator	8 NVIDIA A100s (40GB HBM2)	8 NVIDIA A100s (40GB HBM2)	4 TPU v3 (8 cores)	16 TPU v3 (32 cores)	32 TPU v3 (64 cores)	8 NVIDIA A100s (40GB HBM2)	16 NVIDIA A100s (40GB HBM2)
CPU	96 3.0 GHz 2nd Generation Intel Xeon Scalable (Cascade Lake)	96 2nd-generation AMD Epyc	96 2.0 GHz Intel Xeon	64 2.0 GHz Intel Xeon	128 2.0 GHz Intel Xeon	96 2.0 GHz Intel Xeon	96 2.0 GHz Intel Xeon
Accelerator Memory	320 GB HBM + 320 MB SRAM	320 GB HBM + 320 MB SRAM	137[6] GB	550[7] GB	1.10[8] TB	320 GB HBM + 320 MB SRAM	640 GB HBM + 640 MB SRAM
Host Memory	1237[9] GB	966[10] GB	256[11] GB	256[11] GB	256[11] GB	680 GB	680 GB
Cost per hour	$32.77	$28	$8 + cost of VM ($1.35) = $9.35	$32	$64	$23.47[12]	$46.94[13]
play_math time	Couldn't get quota for an instance	1m 47.854s	Too long to care	9m 19.873s	Couldn't get quota to try	1m 54.273s	3m 55.344s[14]
play_image time	Couldn't get quota for an instance	46m 0.339s	Too long to care	Consistently broke the cluster	Couldn't get quota to try	48m 43.917s	67m 54.672s
play_char time	Couldn't get quota for an instance	9m 45.164s	Too long to care	Consistently broke the cluster	Couldn't get quota to try	10m 21.712s	21m 25.199s

CSUN
CALIFORNIA STATE UNIVERSITY NORTHRIDGE

COMP222

ACM Lecture

DR JEFF SOFTWARE
INDIE APP DEVELOPER

Jessie Frazelle

Jan 2022

Name	Cerebras WSE-2[1]	Dojo Training Tile[2]	Dojo D1[2]	NVIDIA A100 80GB SXM>[3]	Google Cloud TPU v4i>[4]	Groq TSP[5]	Graphcore Colossus™ MK2 GC200 IPU[6]	Tenstorrent Grayskull e300 PCIe[7]
Size	46,225 mm^2	< 92,903 mm^2	645 mm^2	826 mm^2	< 400 mm^2		823 mm^2	
Cores	850,000	35,400	1,416[8]	6,912 CUDA + 432 Tensor	1	1	1,472	
BF16/CFP8[9]		9 PFLOPS	362 TFLOPS	312 TFLOPS[10]	138 TFLOPS[11]			
FP64				9.7 TFLOPS[12]				
FP32		565 TFLOPS	22.6 TFLOPS	19.5 TFLOPS			64 TFLOPS	
FP16				312 TFLOPS		250 TFLOPS	250 TFLOPS	
INT8				624 TOPS	138 TOPS	1 POPS		600 TOPS
On-chip memory (SRAM)	40 gigabytes	11 gigabytes	442.5 megabytes[13]	40 megabytes[14]	151 megabytes[15]	220 megabytes	900 megabytes	
DRAM				80 gigabytes[16] HBM	8 gibibytes HBM			16 gigabytes[16]
Memory bandwidth[17]	20 petabytes/sec	10 terabytes/sec	10 terabytes/sec	2.039 terabytes/sec	614 gigabytes/sec	80 terabytes/sec	47.5 terabytes/sec	200 gigabytes/sec
Fabric bandwidth	27.5 petabytes/sec[18]	36 terabytes/sec	4 terabytes/sec	600 gigabytes/sec[19]	100 gigabytes/sec	500 gigabytes/sec[20]	320 gigabytes/sec	
Max Thermal Design Power (TDP)	20kW / 15kW	15kW	400W	400W	175W			300W
Process	7nm	7nm	7nm	7nm	7 nm	14 nm	7nm	
Transistors	2.6 trillion	1.250 trillion	50 billion	54 billion	16 billion	26.8 billion	59.4 billion	
Made for	general-purpose	training	training	general-purpose	general-purpose	Inference	general-purpose	general-purpose
Price	$2-3 million+[21]			$20,000+				$2,000