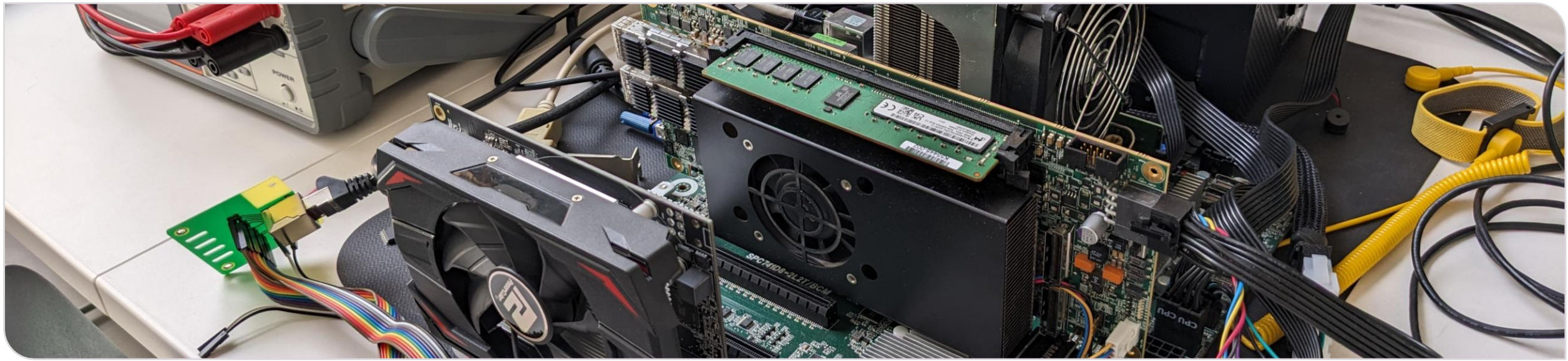
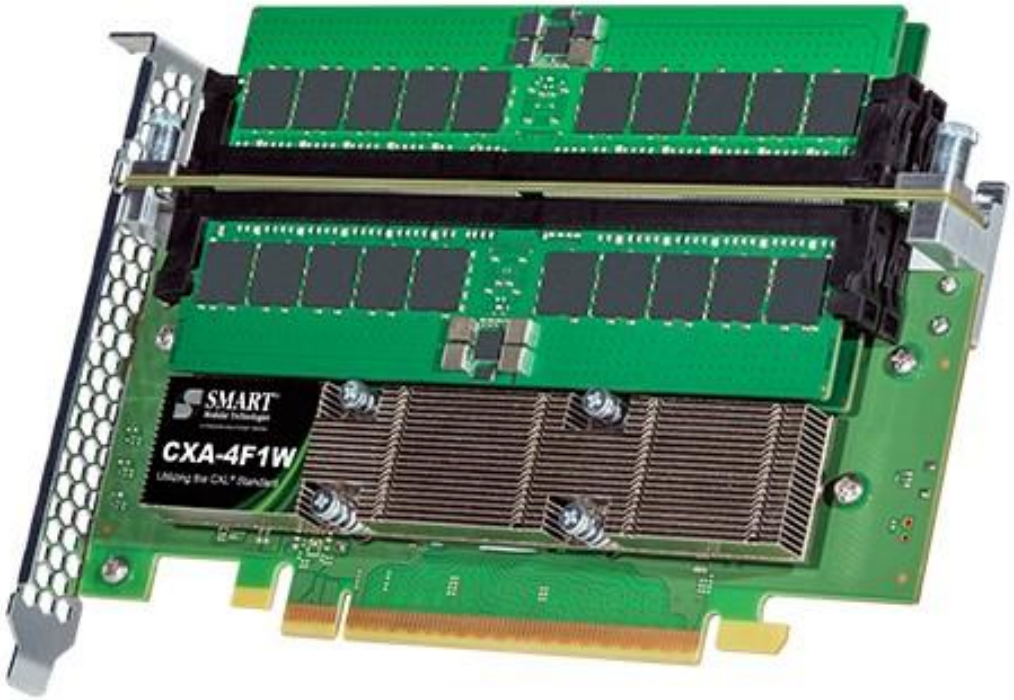
Karlsruhe Institute of Technology

Experiences with Real CXL Hardware

Yussuf Khalil



Our Hardware

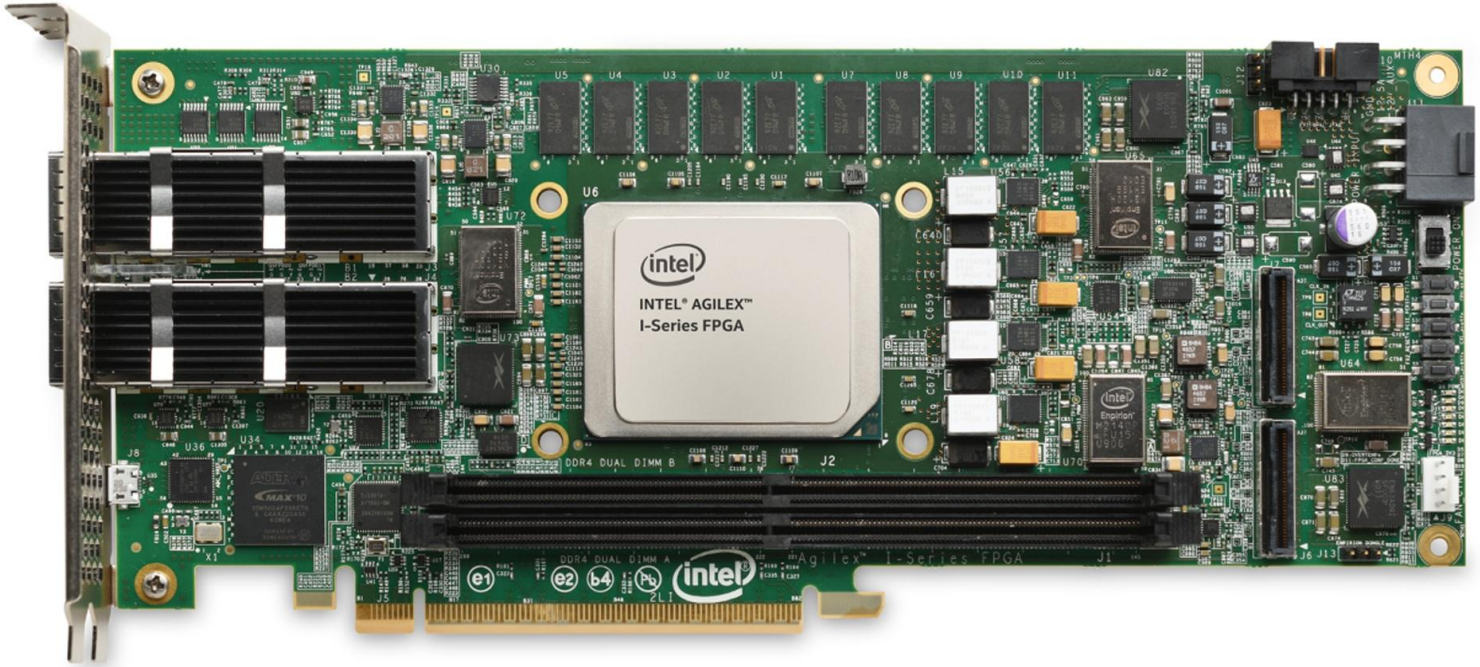


SMART Modular CXA-4F1W

- CXL 2.0 x16
- Simple memory expander with 256 GB DDR5
- €3k via Mouser

https://www.smartm.com/product/cxl-aic-cxa-4f1w

Our Hardware

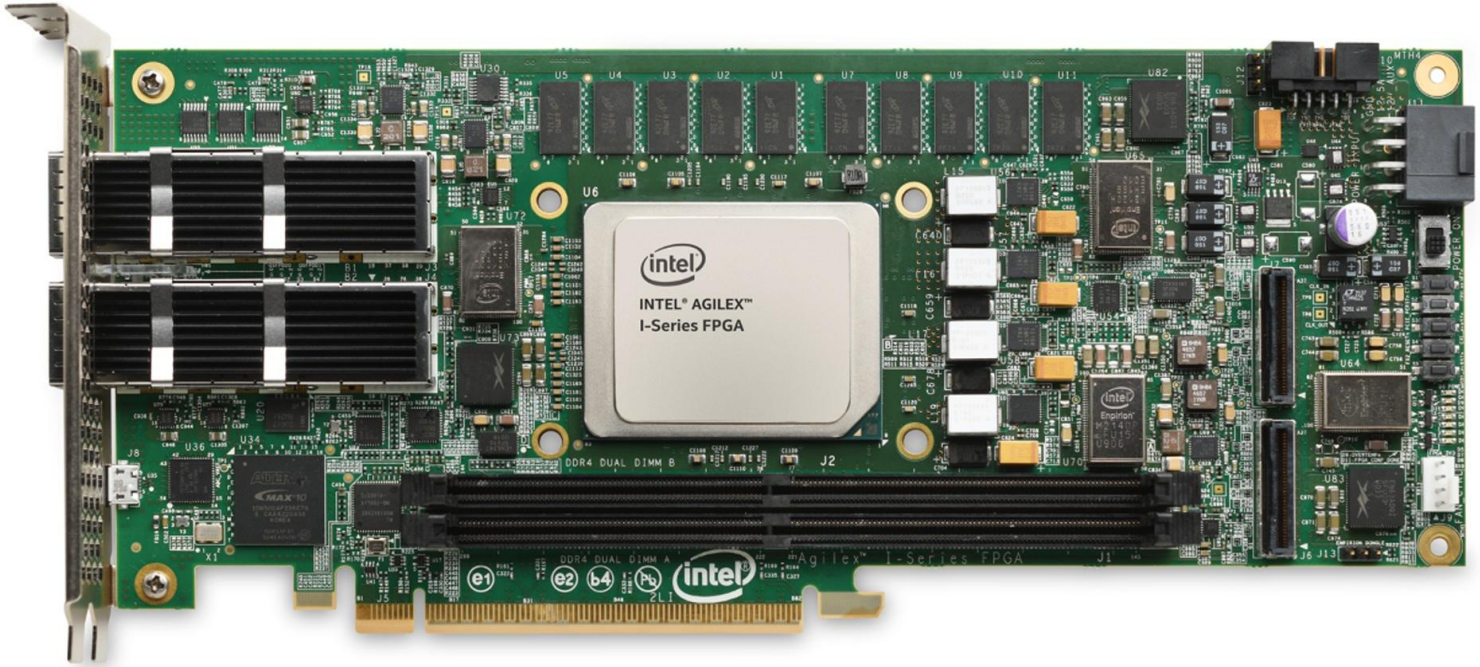


Altera Agilex 7 I-Series Development Kit

- CXL-capable FPGA
- €9k via Mouser

https://www.intel.com/content/www/us/en/products/details/fpga/development-kits/agilex/agi027.html

Our Hardware

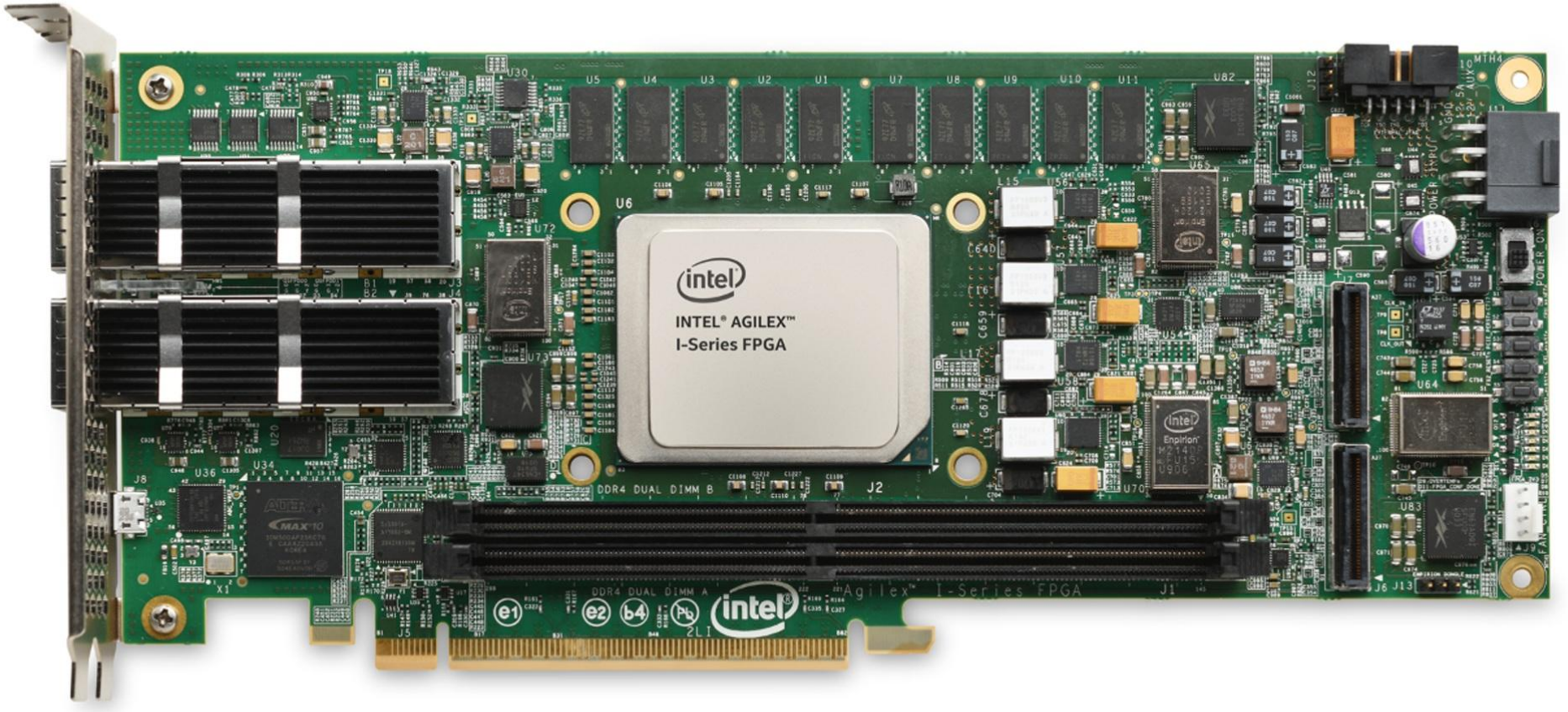


Altera Agilex 7 I-Series Development Kit

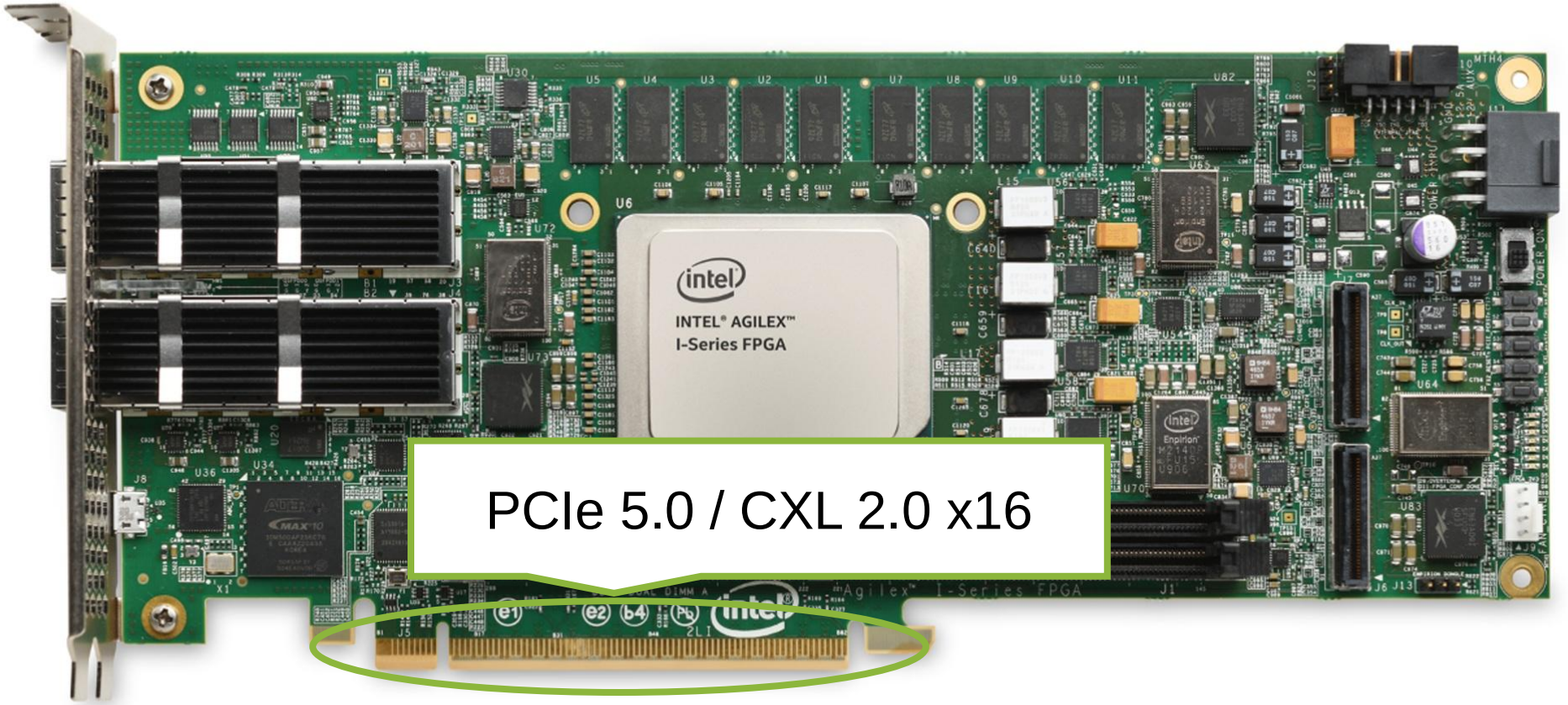
- CXL-capable FPGA
- €9k via Mouser
  - €0 via Altera FPGA Academic Program ☺

https://www.intel.com/content/www/us/en/products/details/fpga/development-kits/agilex/agi027.html

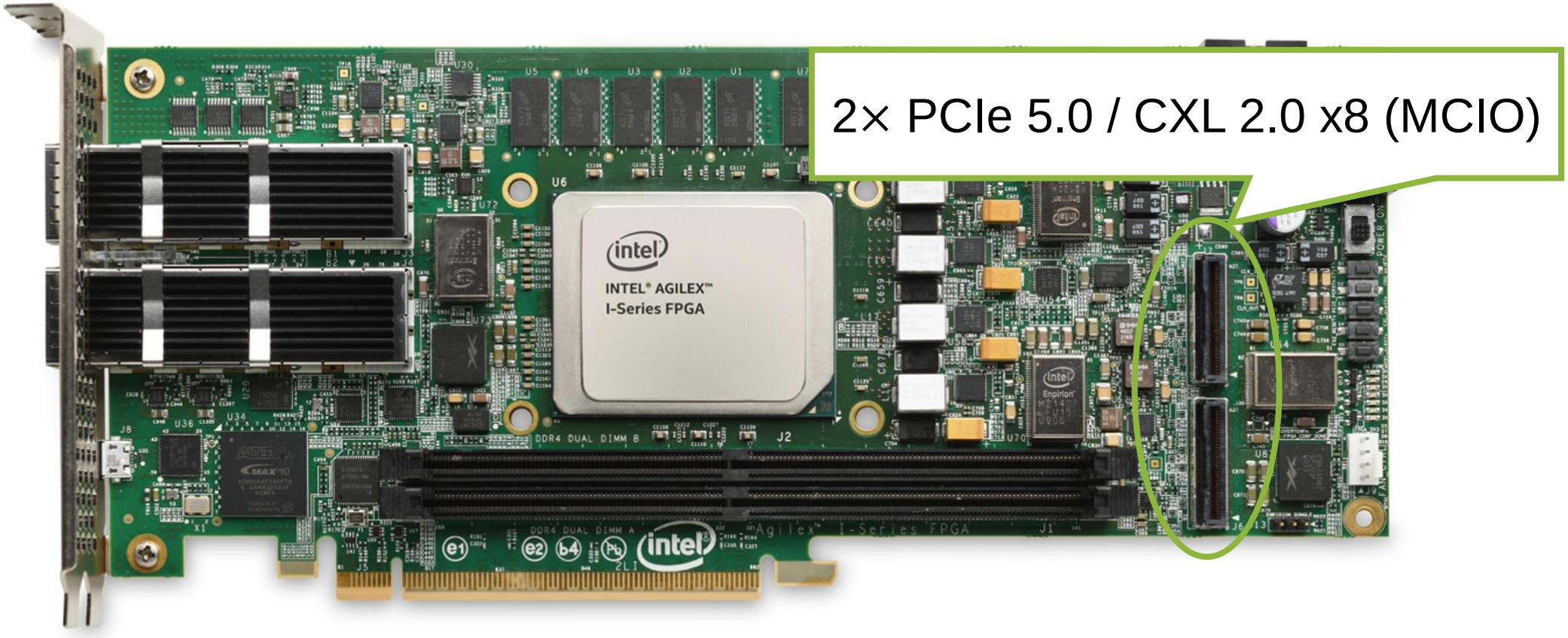
Our Hardware



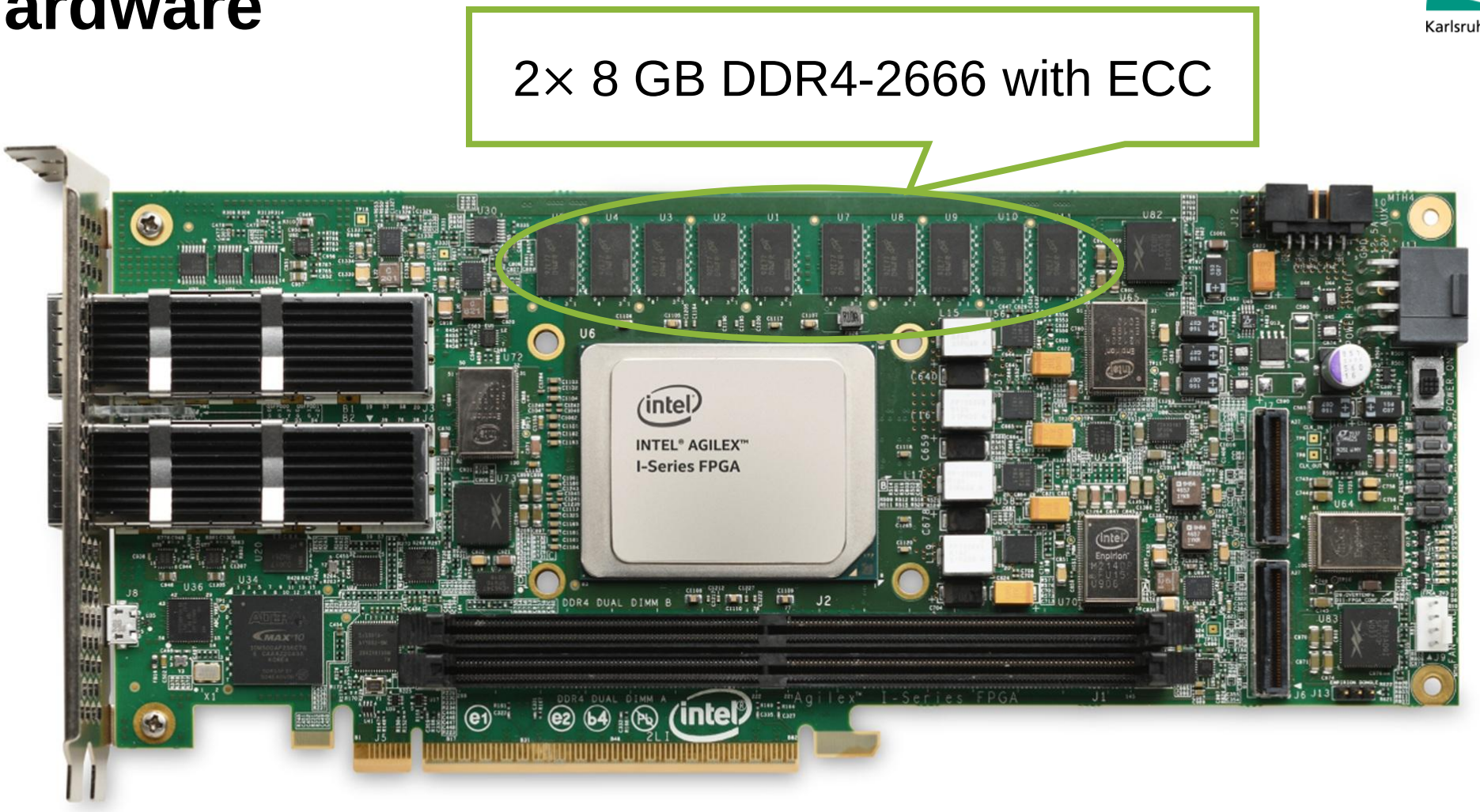
Our Hardware



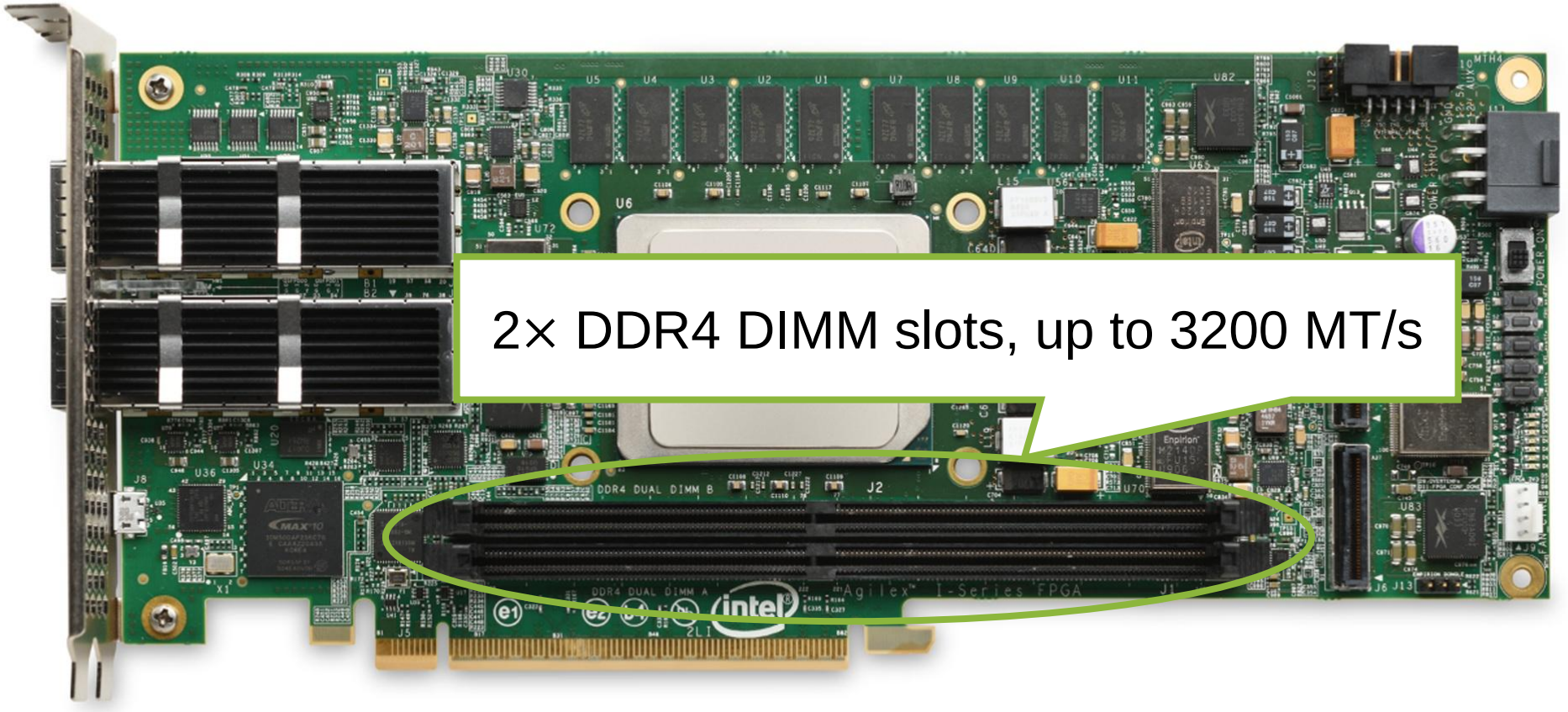
Our Hardware



Our Hardware

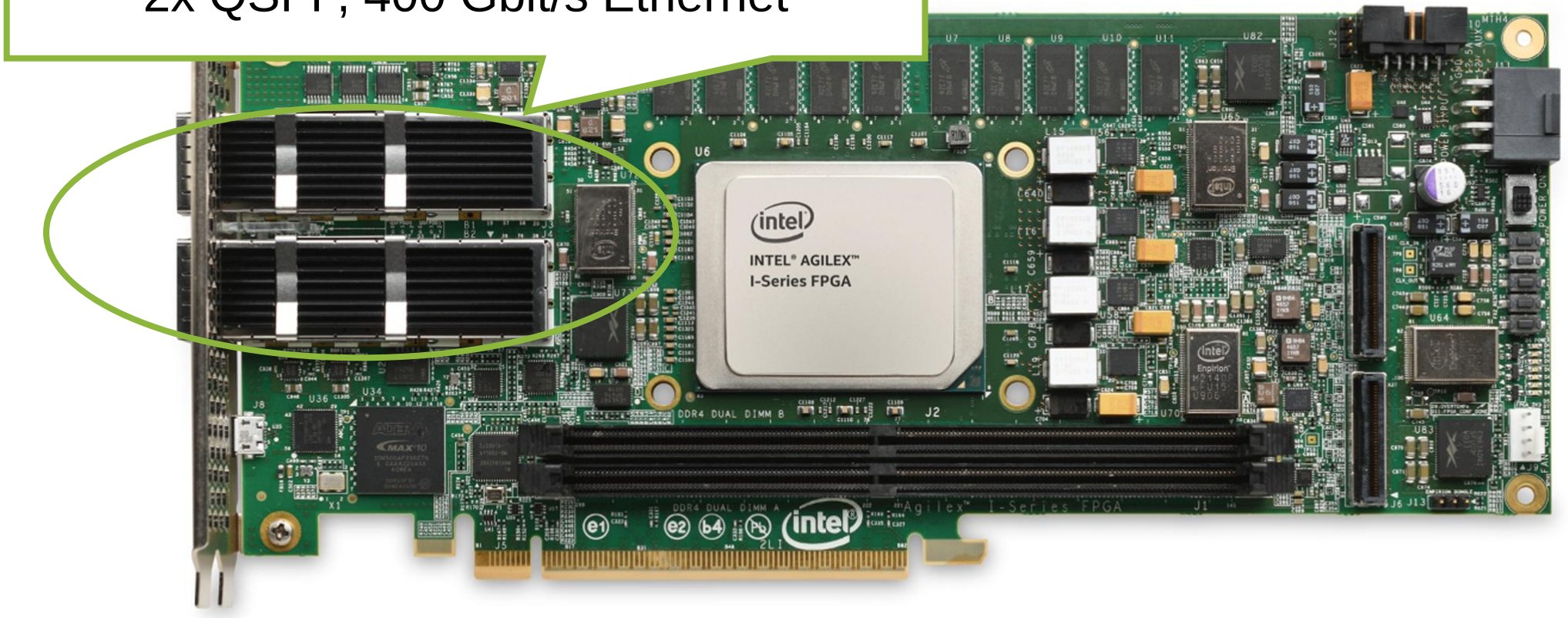


Our Hardware

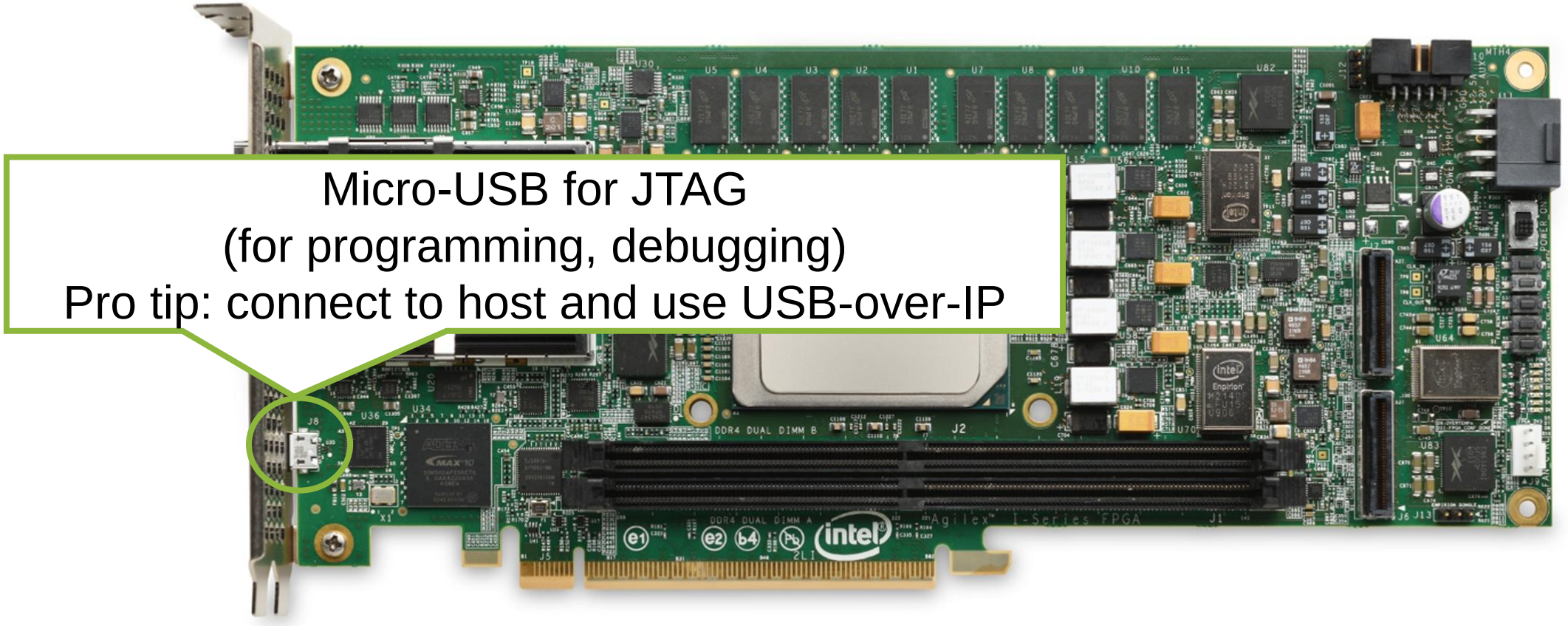


Our Hardware

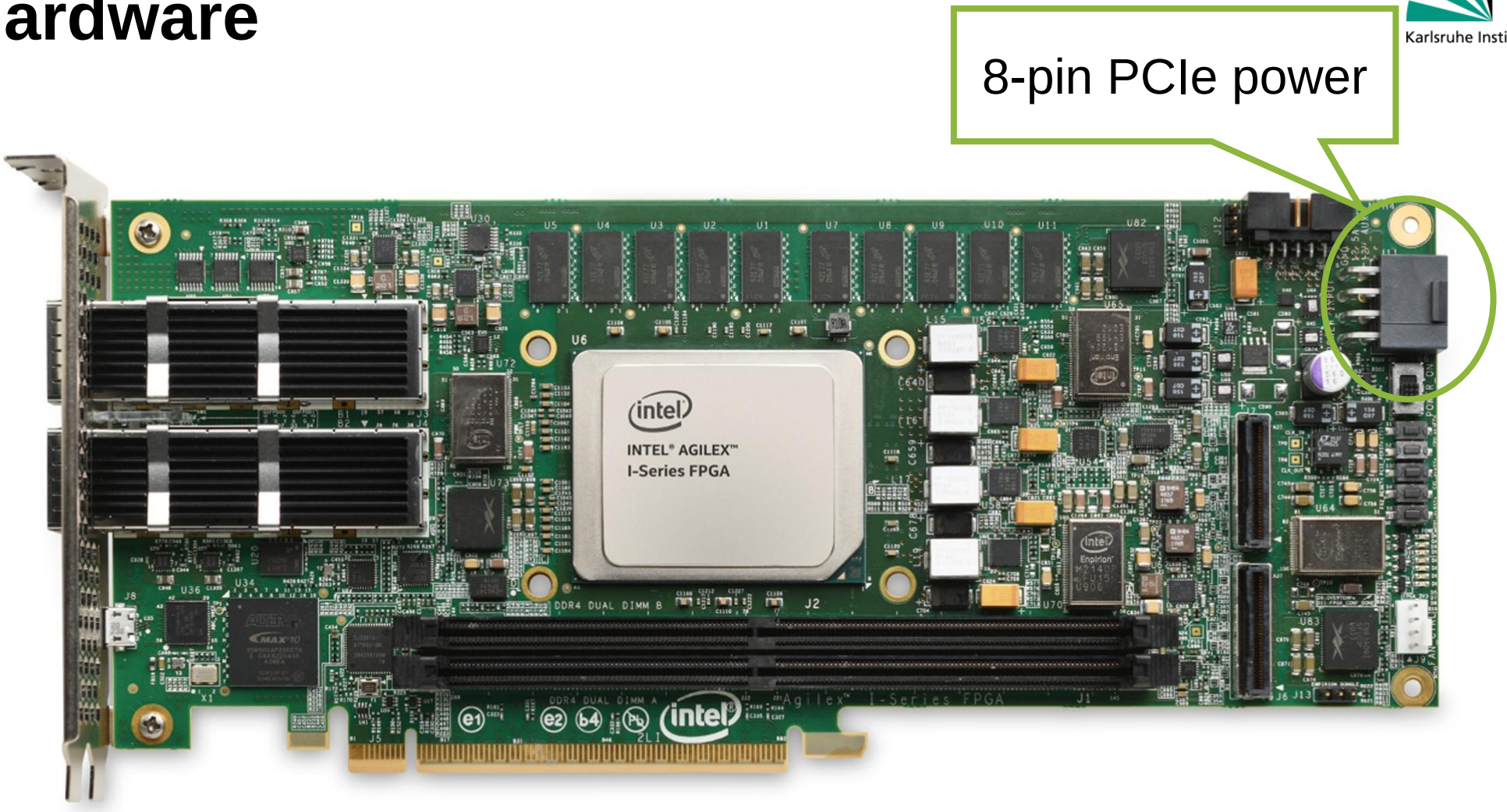
2x QSFP, 400 Gbit/s Ethernet



Our Hardware



Our Hardware



Our Hardware

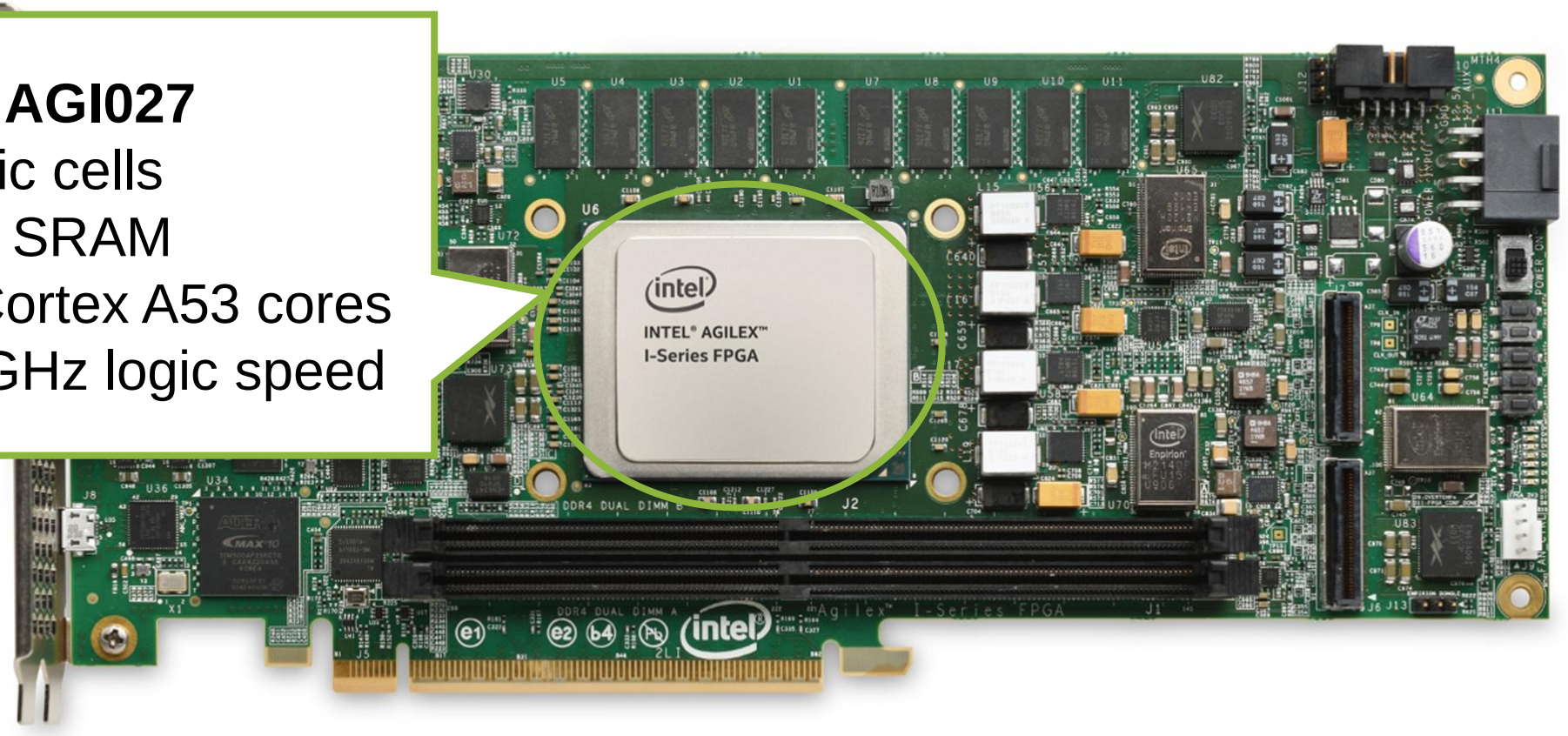
AGI027

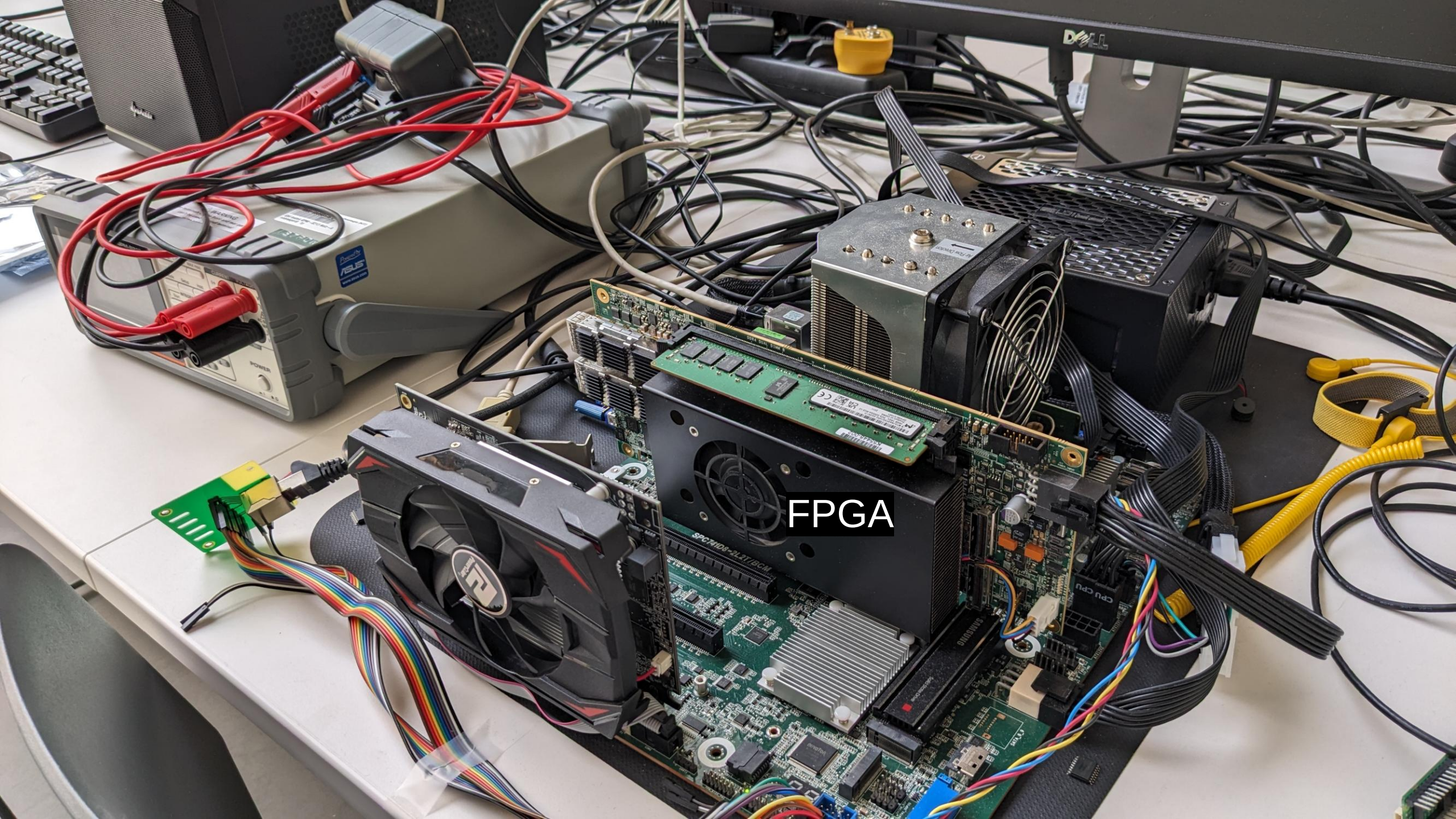
2.7M logic cells

287 Mbit SRAM

4 ARM Cortex A53 cores

Up to 1 GHz logic speed




FPGA

I Have Never Used an FPGA Before!

I Have Never Used an FPGA Before!

- You should have a basic understanding of digital circuitry
  - You know what a NAND gate is and understand the difference between SRAM and DRAM? **Perfect!**

I Have Never Used an FPGA Before!

- You should have a basic understanding of digital circuitry
  - You know what a NAND gate is and understand the difference between SRAM and DRAM? **Perfect!**
- I only had prior experience with small Lattice FPGAs
  - ⇒ I did know some Verilog

I Have Never Used an FPGA Before!

- You should have a basic understanding of digital circuitry
  - You know what a NAND gate is and understand the difference between SRAM and DRAM? **Perfect!**
- I only had prior experience with small Lattice FPGAs
  - ⇒ I did know some Verilog
- I worked with an Altera (Intel) FPGA for the first time in my master's thesis
  - All Altera FPGA IP is written in Verilog
  - However, those are *very* different beasts
  - Perhaps start with a small FPGA, e.g., cheap Lattice or old Xilinx

I Have Never Used an FPGA Before!

- You should have a basic understanding of digital circuitry
  - You know what a NAND gate is and understand the difference between SRAM and DRAM? **Perfect!**
- I only had prior experience with small Lattice FPGAs
  - ⇒ I did know some Verilog
- I worked with an Altera (Intel) FPGA for the first time in my master's thesis
  - All Altera FPGA IP is written in Verilog
  - However, those are *very* different beasts
  - Perhaps start with a small FPGA, e.g., cheap Lattice or old Xilinx

If I can do it, so can you. ☺

Altera FPGA CXL IP

- You can implement all your dreams and desires on the FPGA

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/cxl-ip.html
https://www.intel.com/content/www/us/en/developer/topic-technology/fpga-academic/overview.html

Altera FPGA CXL IP

- You can implement all your dreams and desires on the FPGA
  - …if you have the necessary licenses

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/cxl-ip.html
https://www.intel.com/content/www/us/en/developer/topic-technology/fpga-academic/overview.html

Altera FPGA CXL IP

- You can implement all your dreams and desires on the FPGA
  - …if you have the necessary licenses
- Altera Quartus Prime Pro (design tooling and compiler)
  - €0 via Altera FPGA Academic Program ☺

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/cxl-ip.html
https://www.intel.com/content/www/us/en/developer/topic-technology/fpga-academic/overview.html

Altera FPGA CXL IP

- You can implement all your dreams and desires on the FPGA
  - …if you have the necessary licenses
- Altera Quartus Prime Pro (design tooling and compiler)
  - €0 via Altera FPGA Academic Program ☺
- Altera CXL IP
  - IP-CXLBASEHIP (raw CXL) – €95k via DigiKey ☹
  - IP-CXLTYP1 (Type 1 device) – €120k via DigiKey ☹
  - IP-CXLTYP2 (Type 2 device) – €120k via DigiKey ☹
  - IP-CXLTYP3 (Type 3 device) – €192k via DigiKey ☹

Type 1 = .cache
Type 2 = .cache + .mem
Type 3 = .mem

} we have those

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/cxl-ip.html
https://www.intel.com/content/www/us/en/developer/topic-technology/fpga-academic/overview.html

Altera FPGA CXL IP

- You can implement all your dreams and desires on the FPGA
  - …if you have the necessary licenses
- Altera Quartus Prime Pro (design tooling and compiler)
  - €0 via Altera FPGA Academic Program ☺
- Altera CXL IP
  - IP-CXLBASEHIP (raw CXL) – €95k via DigiKey ☹
  - IP-CXLTYP1 (Type 1 device) – €120k via DigiKey ☹
  - IP-CXLTYP2 (Type 2 device) – €120k via DigiKey ☹
  - IP-CXLTYP3 (Type 3 device) – €192k via DigiKey ☹
  - Ask your local Intel Sales Representative for individual prices ☺

Type 1 = .cache
Type 2 = .cache + .mem
Type 3 = .mem

} we have those

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/cxl-ip.html
https://www.intel.com/content/www/us/en/developer/topic-technology/fpga-academic/overview.html

Altera FPGA CXL IP

- You can implement all your dreams and desires on the FPGA
  - …if you have the necessary licenses
- Altera Quartus Prime Pro (design tooling and compiler)
  - €0 via Altera FPGA Academic Program ☺
- Altera CXL IP
  - IP-CXLBASEHIP (raw CXL) – €95k via DigiKey ☹
  - IP-CXLTYP1 (Type 1 device) – €120k via DigiKey ☹
  - IP-CXLTYP2 (Type 2 device) – €120k via DigiKey ☹
  - IP-CXLTYP3 (Type 3 device) – €192k via DigiKey ☹
  - Ask your local Intel Sales Representative for individual prices ☺
  - **IP documentation is under NDA, licenses valid for 1 year only**

Type 1 = .cache
Type 2 = .cache + .mem
Type 3 = .mem

} we have those

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/cxl-ip.html
https://www.intel.com/content/www/us/en/developer/topic-technology/fpga-academic/overview.html

Want DMA on top of CXL.io?

- Would need another IP: IP-PCIEMCDMA-AXI
- Still relatively new, can not be bought anywhere
  - We do not have a license either
- This kind of IP used to be included with Quartus in the past

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/multichannel-dma-mcdma.html

Want DMA on top of CXL.io?

- Would need another IP: IP-PCIEMCDMA-AXI
- Still relatively new, can not be bought anywhere
  - We do not have a license either
- This kind of IP used to be included with Quartus in the past

- …or just do it yourself ☺

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/multichannel-dma-mcdma.html

Want DMA on top of CXL.io?

- Would need another IP: IP-PCIEMCDMA-AXI
- Still relatively new, can not be bought anywhere
  - We do not have a license either
- This kind of IP used to be included with Quartus in the past

- …or just do it yourself ☺
  - Need to implement the respective PCIe transactions

https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/multichannel-dma-mcdma.html

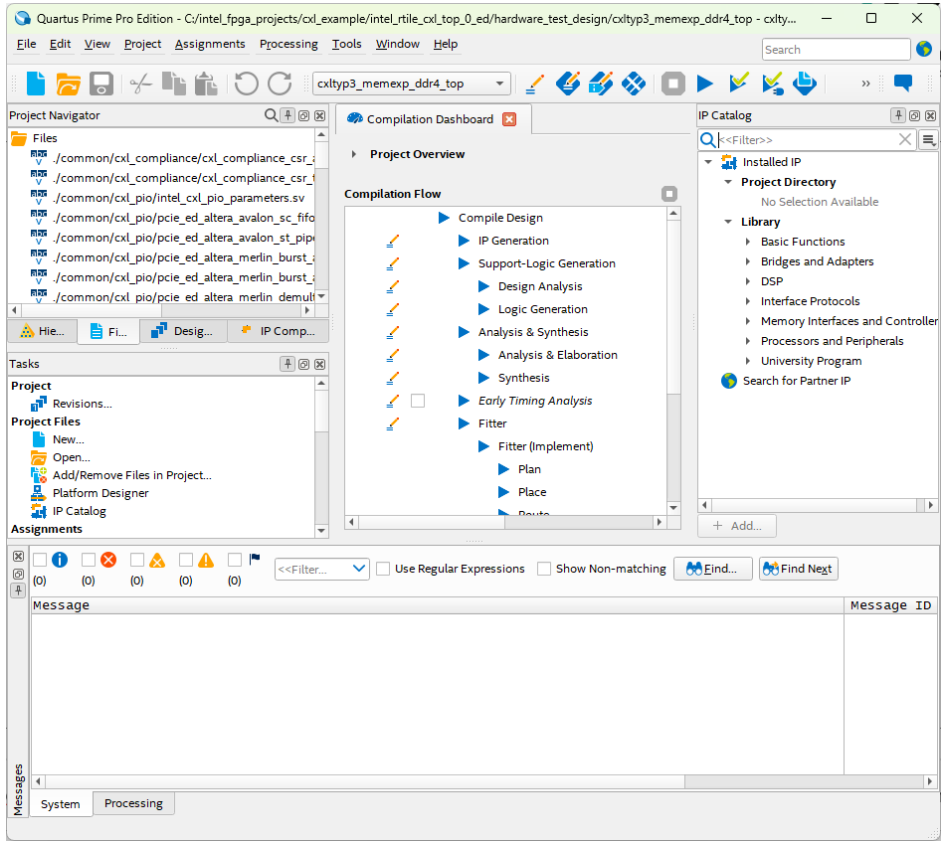
Want DMA on top of CXL.io?

- Would need another IP: IP-PCIEMCDMA-AXI
- Still relatively new, can not be bought anywhere
  - We do not have a license either
- This kind of IP used to be included with Quartus in the past

- …or just do it yourself ☺
  - Need to implement the respective PCIe transactions
  - Maybe parts of that can be done in software on the ARM cores?
    - Performance may be limited, but perhaps it is fast enough

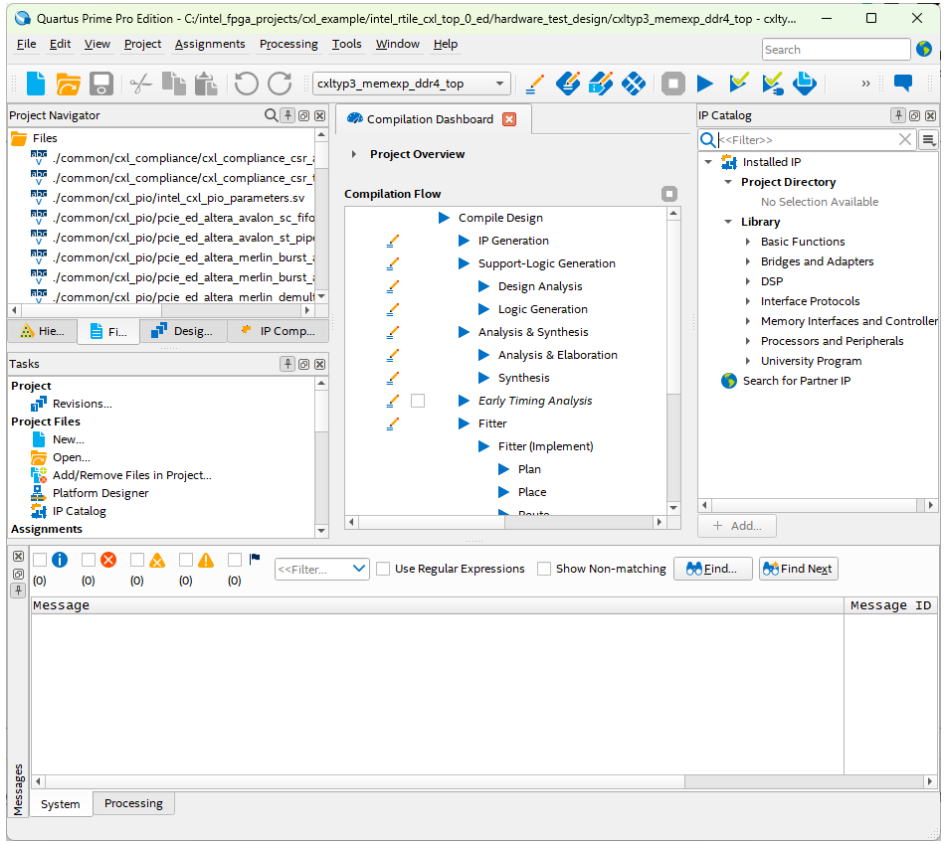
https://www.intel.com/content/www/us/en/products/details/fpga/intellectual-property/interface-protocols/multichannel-dma-mcdma.html

Altera Quartus Prime Pro



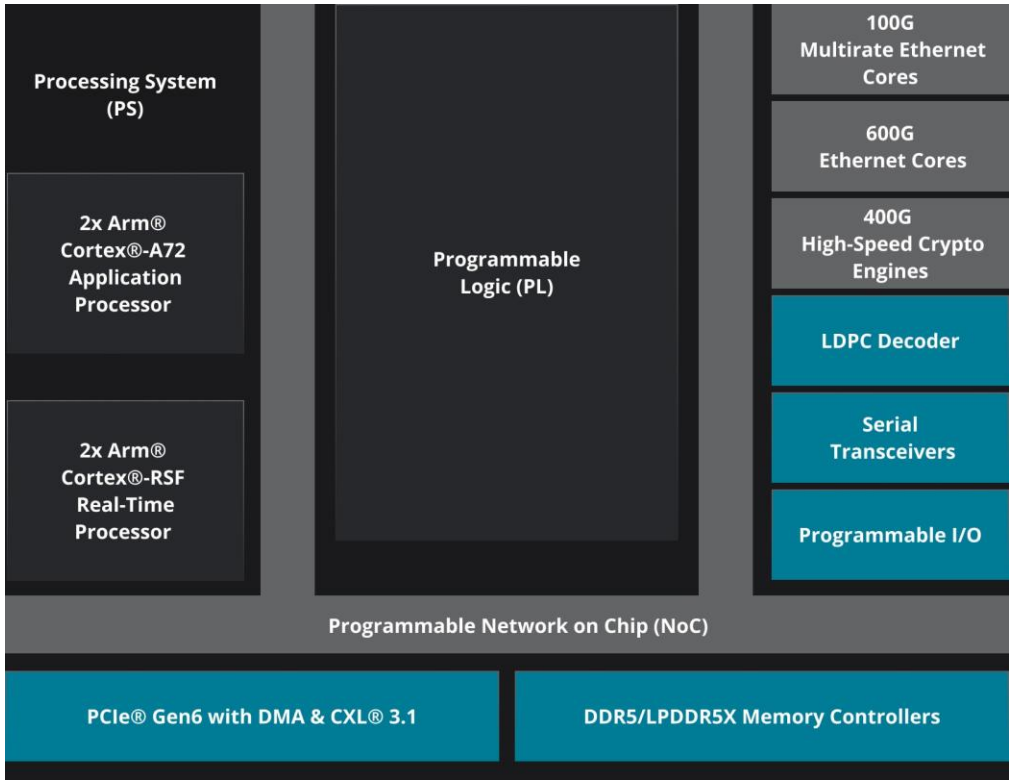
Altera Quartus Prime Pro

- Strongly recommend using a machine specifically built for Quartus
  - At least 16 cores (do **not** use SMT)
  - Fast DRAM (and at least 128 GB)
  - Single-thread performance is most important
  - Expect compilation times ≥ 45 minutes



An Alternative?

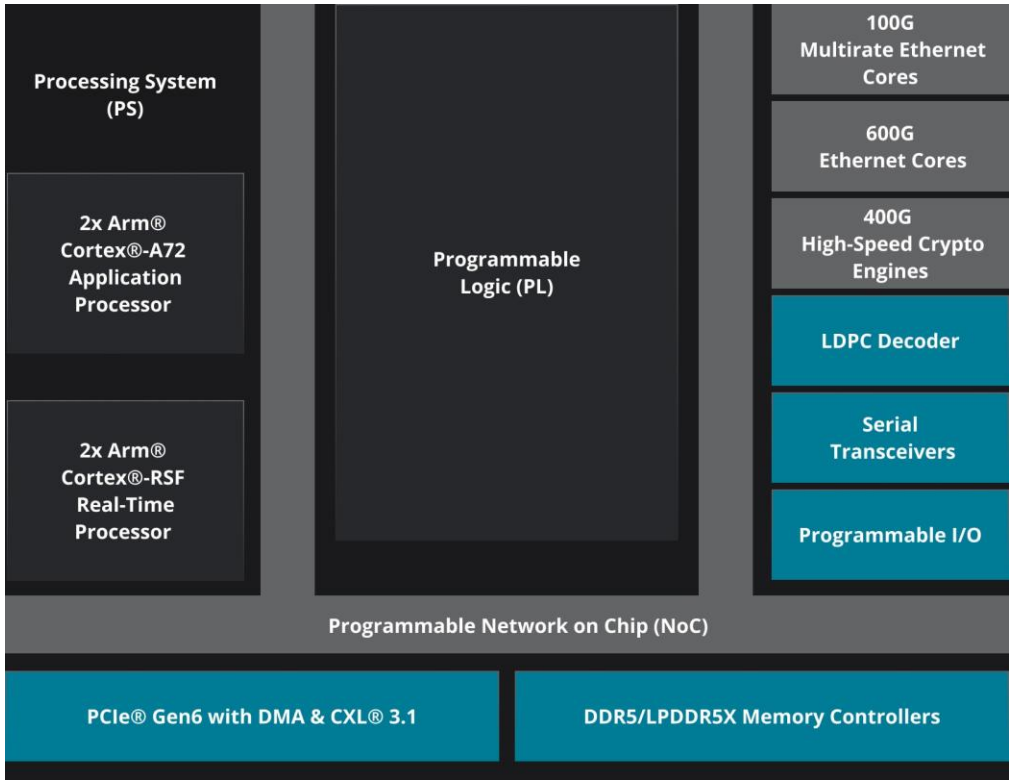
- AMD Versal Premium Series Gen 2
  - To be released soon
  - PCIe 6.0 / CXL 3.1 2× x8



https://www.amd.com/en/products/adaptive-socs-and-fpgas/versal/gen2/premium-series.html

An Alternative?

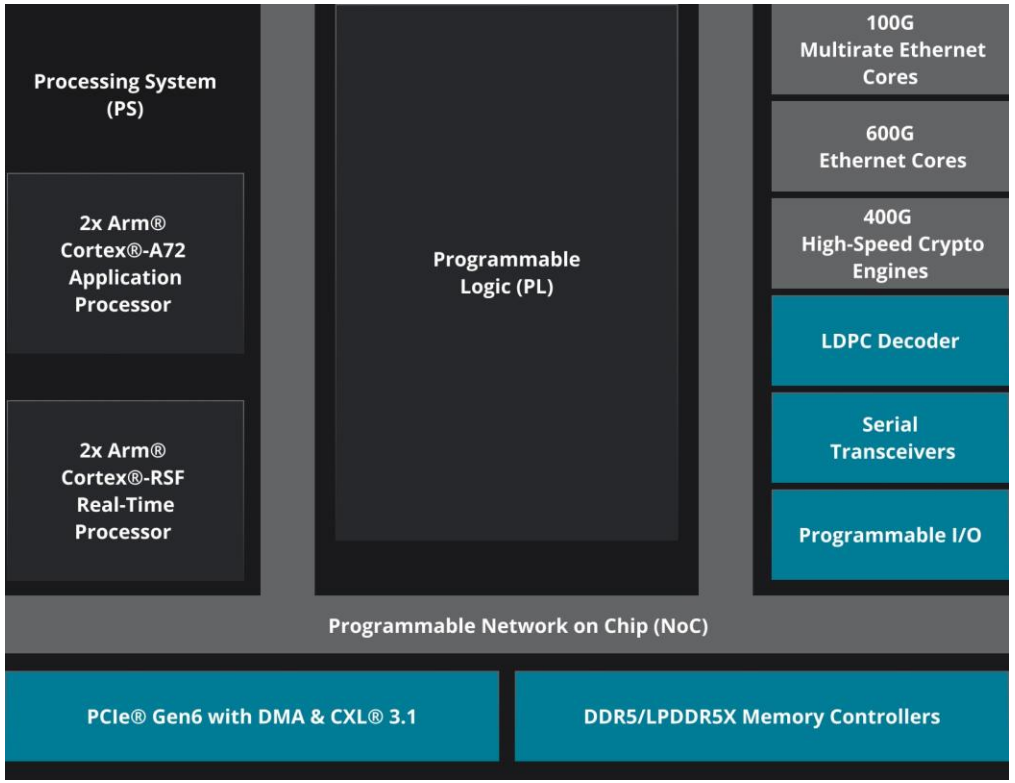
- AMD Versal Premium Series Gen 2
  - To be released soon
  - PCIe 6.0 / CXL 3.1 2× x8
- Licensing?
- Pricing?
- NDA required?



https://www.amd.com/en/products/adaptive-socs-and-fpgas/versal/gen2/premium-series.html

An Alternative?

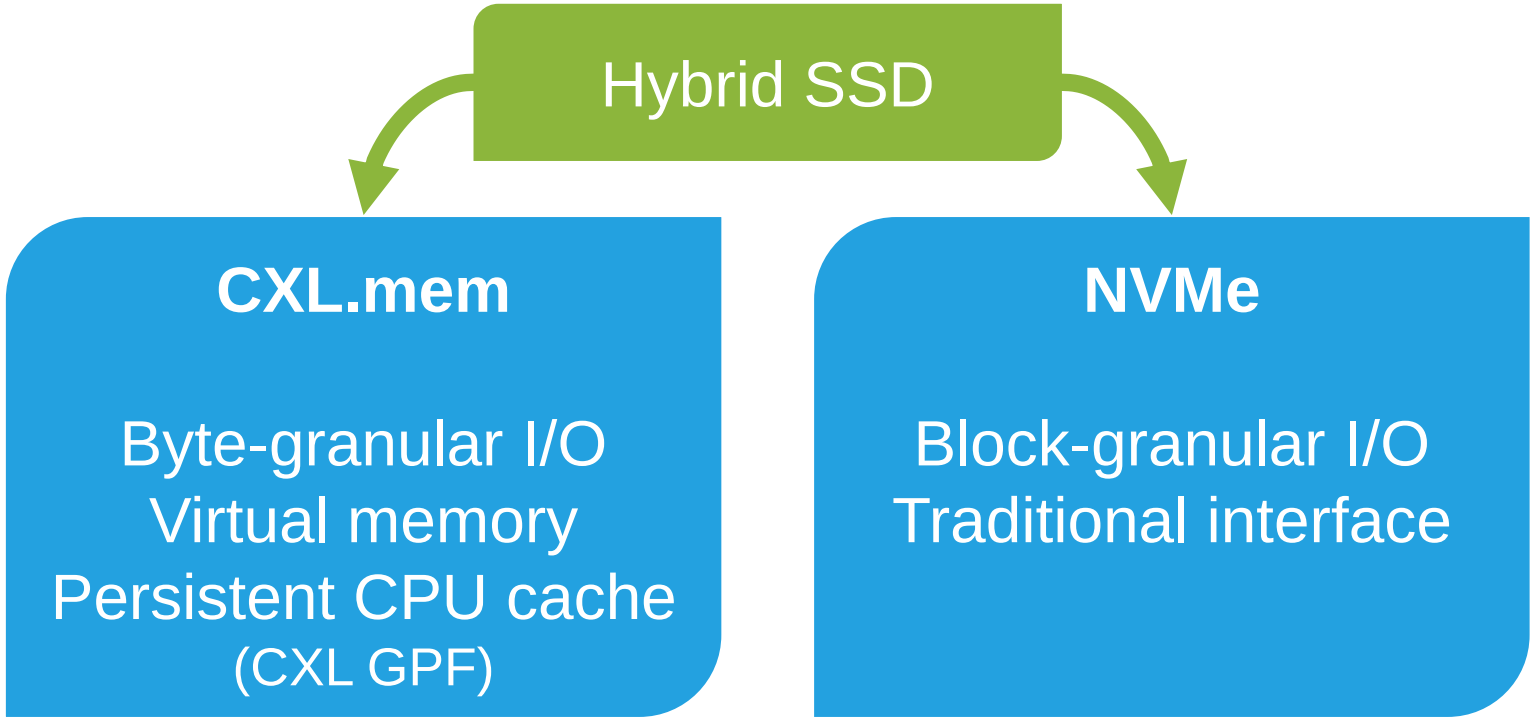
- AMD Versal Premium Series Gen 2
  - To be released soon
  - PCIe 6.0 / CXL 3.1 2× x8
- Licensing?
- Pricing?
- NDA required?
- Early Access Program linked on website
  - Perhaps give it a shot?



https://www.amd.com/en/products/adaptive-socs-and-fpgas/versal/gen2/premium-series.html

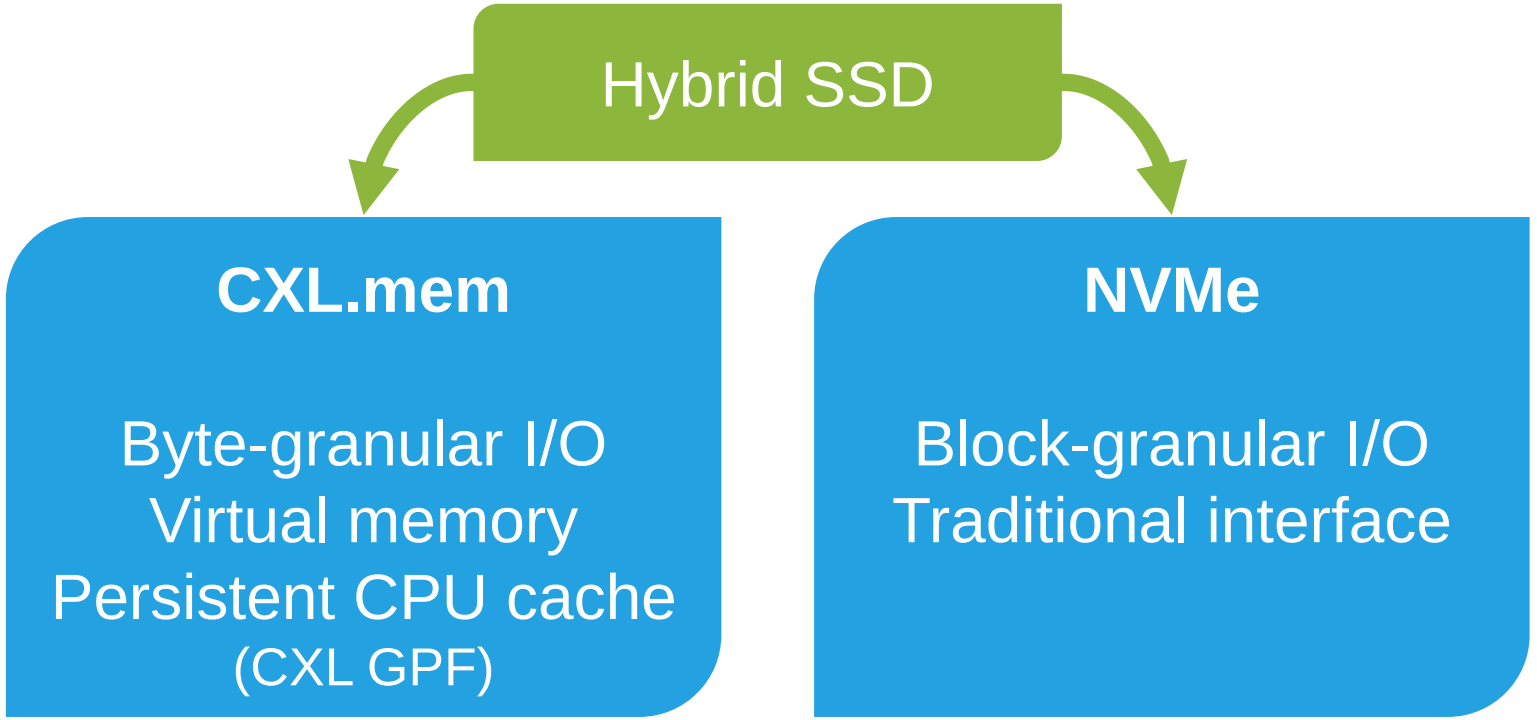
Current CXL-Related Research Projects

Resource Management for Hybrid SSDs



Habicht et al. *Fundamental OS Design Considerations for CXL-based Hybrid SSDs* (DIMES'24)

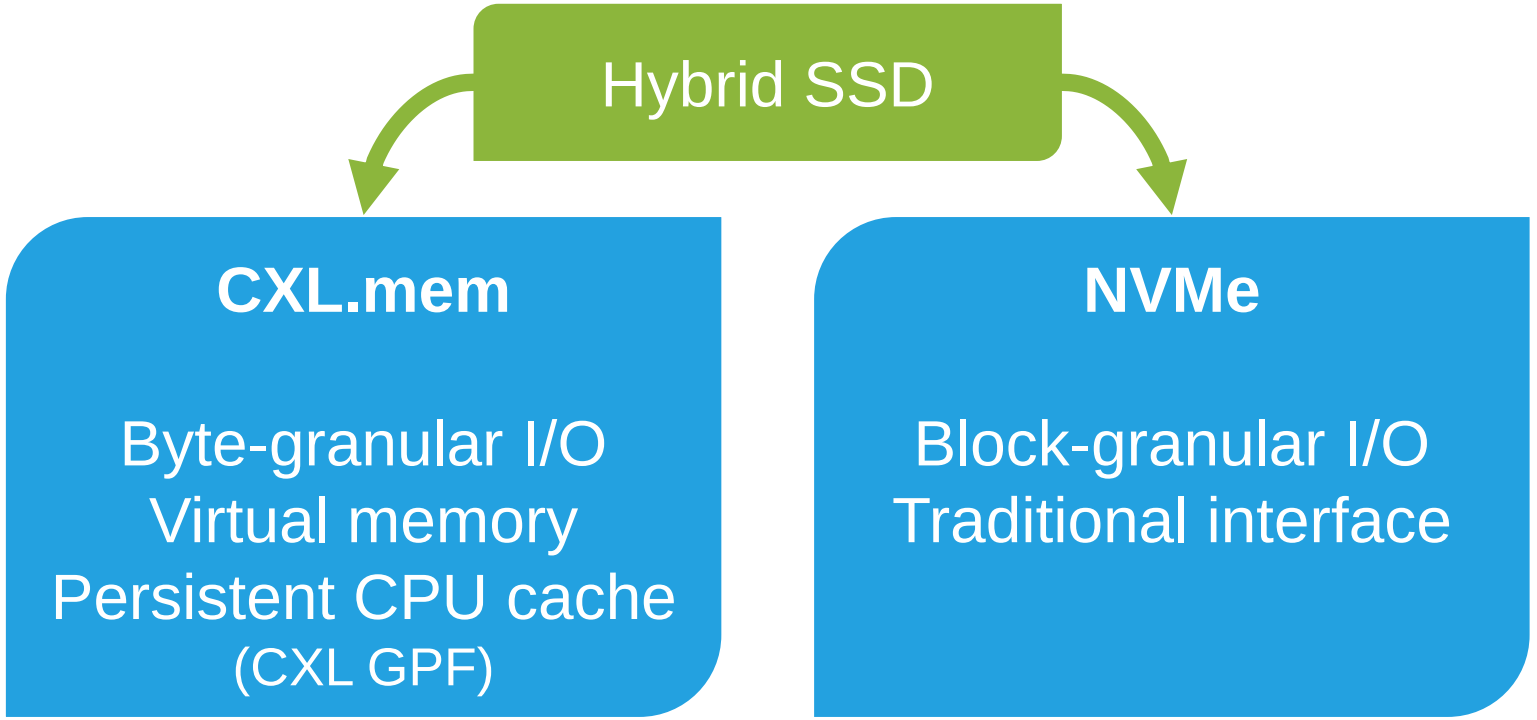
Resource Management for Hybrid SSDs



- How to manage hybrid SSDs in an operating system?
  - No uniform access to entire capacity ⇒ **Traditional DAX unsuitable**
  - Need better abstractions and management techniques

Habicht et al. *Fundamental OS Design Considerations for CXL-based Hybrid SSDs* (DIMES'24)

Resource Management for Hybrid SSDs



- How to manage hybrid SSDs in an operating system?
  - No uniform access to entire capacity ⇒ **Traditional DAX unsuitable**
  - Need better abstractions and management techniques
- Demonstrated up to 4.1× more throughput in *Valkey* in recent publication

Habicht et al. *Fundamental OS Design Considerations for CXL-based Hybrid SSDs* (DIMES'24)

Resource Management for Hybrid SSDs

- We want to experiment with our own hardware-level ideas

Resource Management for Hybrid SSDs

- We want to experiment with our own hardware-level ideas

Hybrid SSD Project

- Commercial hybrid SSDs are not yet available
- Commercial products may have shortcomings
  - Cache management is utterly important in a hybrid SSD
  - We would like to test various different approaches here

Resource Management for Hybrid SSDs

- We want to experiment with our own hardware-level ideas

Hybrid SSD Project

- Commercial hybrid SSDs are not yet available
- Commercial products may have shortcomings
  - Cache management is utterly important in a hybrid SSD
  - We would like to test various different approaches here

- OpenExpress provides open-source NVMe hardware implementation

Jung. *OpenExpress: Fully Hardware Automated Open Research Framework for Future Fast NVMe Devices* (ATC'20)
https://www.usenix.org/system/files/atc20-jung.pdf

Fast and Power-Efficient System Suspend

Firmware + OS Co-Design Approach with Hybrid SSDs.

- Goal: fast wakeup from fully powered off state without runtime impact
- Leverage hybrid SSDs to avoid page copies
- Let OS and firmware collaborate to minimize startup overhead

Khalil et al. *Towards Fast and Power-Efficient System Suspend* (SOSP'24 Poster)

Fast and Power-Efficient System Suspend

Firmware + OS Co-Design Approach with Hybrid SSDs.

- Goal: fast wakeup from fully powered off state without runtime impact
- Leverage hybrid SSDs to avoid page copies
- Let OS and firmware collaborate to minimize startup overhead
- Initial prototype delivers 5.8× faster wakeup vs. ACPI S4

Khalil et al. *Towards Fast and Power-Efficient System Suspend* (SOSP'24 Poster)

GPU4FS

- Problems:
  - File systems cause significant CPU overhead in large storage setups
  - Accelerator-based AI training and HPC applications require vast amounts of data from storage

Maucher et al. *Full-Scale File System Acceleration on GPU* (FGBS Spring'24)

GPU4FS

- Problems:
  - File systems cause significant CPU overhead in large storage setups
  - Accelerator-based AI training and HPC applications require vast amounts of data from storage

- Idea: Design *accelerator-first* file system that also runs on a CPU

Maucher et al. *Full-Scale File System Acceleration on GPU* (FGBS Spring'24)

GPU4FS

- Problems:
  - File systems cause significant CPU overhead in large storage setups
  - Accelerator-based AI training and HPC applications require vast amounts of data from storage
- Idea: Design *accelerator-first* file system that also runs on a CPU
- Challenge: NVMe unsuitable for implementation on GPUs
  - Current prototype based on Optane
  - CXL-based implementation with hybrid SSDs planned

Maucher et al. *Full-Scale File System Acceleration on GPU* (FGBS Spring'24)

File System Crash Consistency

- Vinter: Crash consistency testing for PM file systems
  - Trace program execution in virtual machine
  - Generate likely crash images with model and heuristic

Kalbfleisch et al. *Vinter: Automatic Non-Volatile Memory Crash Consistency Testing for Full Systems* (USENIX ATC'22)

File System Crash Consistency

- Vinter: Crash consistency testing for PM file systems
  - Trace program execution in virtual machine
  - Generate likely crash images with model and heuristic

- New opportunities with CXL on FPGA
  - Trace updates at CXL device
  - Verify crash consistency with real device states
  - Confirm correctness of crash image generation model

Kalbfleisch et al. *Vinter: Automatic Non-Volatile Memory Crash Consistency Testing for Full Systems* (USENIX ATC'22)

File System Crash Consistency

- Vinter: Crash consistency testing for PM file systems
  - Trace program execution in virtual machine
  - Generate likely crash images with model and heuristic

- New opportunities with CXL on FPGA
  - Trace updates at CXL device
  - Verify crash consistency with real device states
  - Confirm correctness of crash image generation model

- Hybrid SSD file systems may yield interesting new challenges

Kalbfleisch et al. *Vinter: Automatic Non-Volatile Memory Crash Consistency Testing for Full Systems* (USENIX ATC'22)

PM File System Performance

- Previous work with Optane PMem
  - Parallel access to PMem generates device contention
    - Mitigation required to avoid excessive power consumption
    - We proposed a new metric and OS-directed mitigation mechanisms

[1] Sun et al. *Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices* (MICRO'23)
Werling et al. *Analyzing and Improving CPU and Energy Efficiency of PM File Systems* (DIMES'23)

PM File System Performance

- Previous work with Optane PMem
  - Parallel access to PMem generates device contention
    - Mitigation required to avoid excessive power consumption
    - We proposed a new metric and OS-directed mitigation mechanisms
- Similar issues found with CXL-attached memory[1]

[1] Sun et al. *Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices* (MICRO'23)
Werling et al. *Analyzing and Improving CPU and Energy Efficiency of PM File Systems* (DIMES'23)

PM File System Performance

- Previous work with Optane PMem
  - Parallel access to PMem generates device contention
    - Mitigation required to avoid excessive power consumption
    - We proposed a new metric and OS-directed mitigation mechanisms
- Similar issues found with CXL-attached memory[1]

- Try to transfer our previous approaches to CXL-enabled systems
  - Goal: Improve performance and energy consumption

[1] Sun et al. *Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices* (MICRO'23)
Werling et al. *Analyzing and Improving CPU and Energy Efficiency of PM File Systems* (DIMES'23)

Conclusion

- CXL offers exciting new possibilities for memory-related research

Conclusion

- CXL offers exciting new possibilities for memory-related research
- Our group is working on various CXL-related projects
  - Hybrid SSD management, system suspend, GPU4FS, crash consistency
  - Some continued from previous Optane research

Conclusion

- CXL offers exciting new possibilities for memory-related research
- Our group is working on various CXL-related projects
  - Hybrid SSD management, system suspend, GPU4FS, crash consistency
  - Some continued from previous Optane research
- FPGAs offer a wide variety of research opportunities
  - …but they are challenging

Conclusion

- CXL offers exciting new possibilities for memory-related research
- Our group is working on various CXL-related projects
  - Hybrid SSD management, system suspend, GPU4FS, crash consistency
  - Some continued from previous Optane research
- FPGAs offer a wide variety of research opportunities
  - …but they are challenging
  - …and you can do it