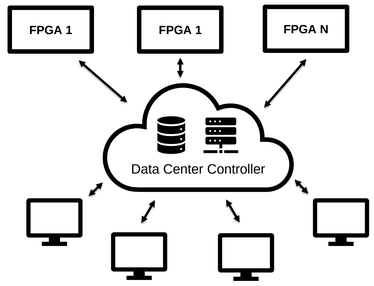
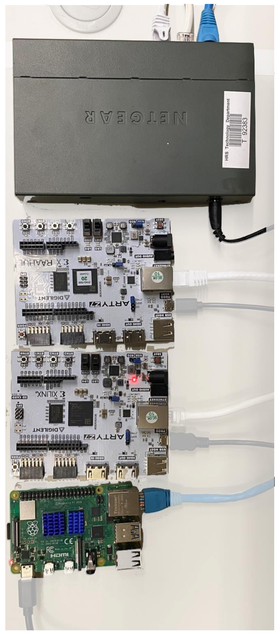
FlexFPGA: A Case for Multi-Tenant Disaggregated Cloud FPGA Architectures

STUDENTS: MISAEL LOPEZ, RAAHUL POTLURI

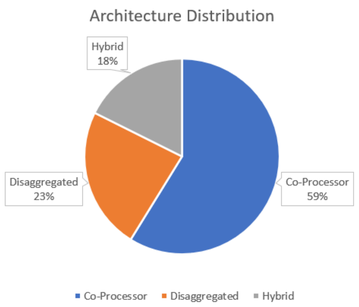
The Problem

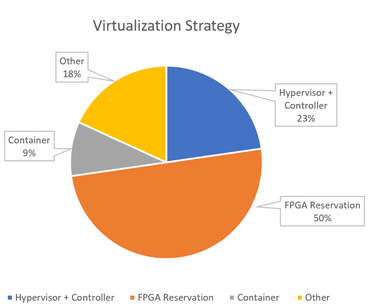
- Traditional data centers restrict FPGA access to a single user, leading to resource underutilization and power wastage.
- ***FlexFPGA*** presents FPGAs as individual compute nodes integrated into data centers, enhancing efficiency, flexibility, and real-time resource allocation for diverse AI and cloud computing applications.





State of Research and Industry

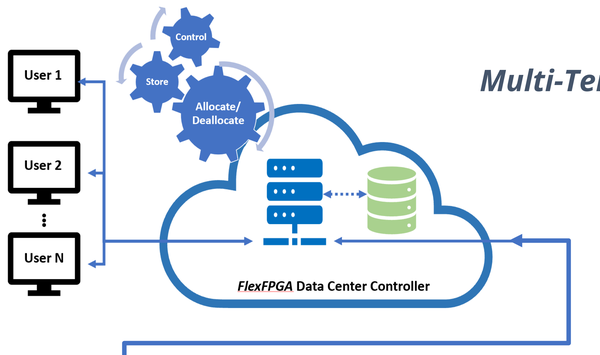




- Industry leaders like AWS, Alibaba, and Microsoft employ FPGA Co-Processor architectures, connecting FPGAs to hosts via PCIe, but still reserve them for single users, leading to potential resource and energy wastage.

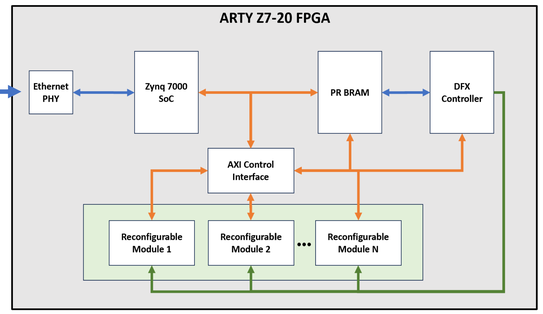
FlexFPGA

- ***FlexFPGA*** is a research cloud FPGA architecture leveraging Zynq 7000 SoC and partial reconfiguration to treat FPGAs as stand-alone compute nodes, integrating them into data centers via network attachment.
- Multi-tenancy is enabled by allowing users to upload partial BIT streams and push them onto reconfigurable modules, presenting a unified FPGA space to users while utilizing multiple reconfigurable regions in the background.



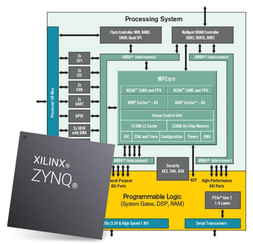
Multi-Tenancy

Disaggregated Reconfigurable Hardware



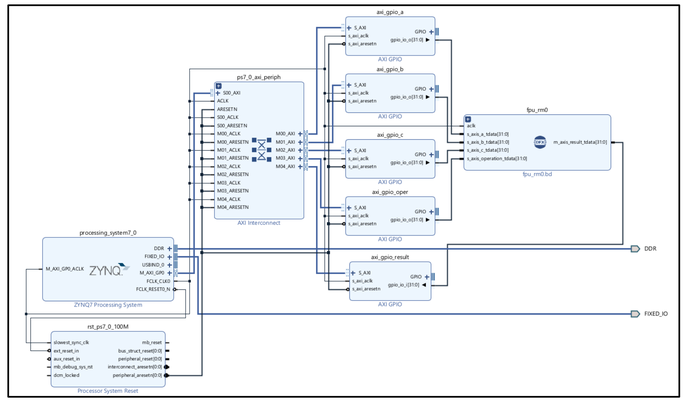
Leveraging Zynq SoC and Dynamic Function Exchange

- Partial Reconfiguration preserves FPGA integrity while modifying specific areas with BIT files.
- FPGA shells can be created in Vivado and offer customizable infrastructure.
- Zynq 7000 devices integrate ARM Cortex-A9 processors with FPGAs, offering high performance and flexibility, suitable for seamless integration into data center and cloud environments with rich peripherals, facilitated by Xilinx's Vitis software platform.



Methodology

- To evaluate the feasibility of our proposed architecture, we plan to measure processing speed and power consumption across disaggregated FPGAs and compare against traditional cloud computing architectures.



Disaggregated Architecture Prototype

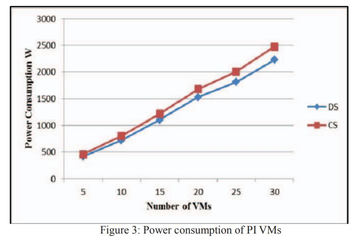
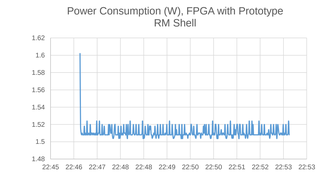
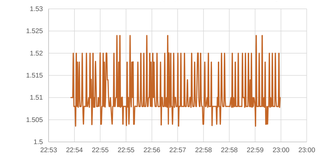


Figure 3: Power consumption of PI VMs

[1]





Future Work, References, and Acknowledgments

- Develop data center software architecture for multi-tenant support.
- Gather additional power consumption data for architecture analysis, comparing against traditional setups.
- Evaluate processing speed and scalability of the architecture and comparing with other alternatives.

Faculty: Sep Makhsous
Graduate Students: Misael Lopez Granados
Undergraduate Students: Raahul Potluri, Kevin S. Lu

[1] H. M. Mohammad Ali, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient disaggregated servers for future data centers," in 2015 20th European Conference on Networks and Optical Communications – (NOC), London: IEEE, Jun. 2015
[2] B. Ringlein, F. Abel, A. Ditter, B. Weiss, C. Hagleitner, and D. Fey, "System Architecture for Network-Attached FPGAs in the Cloud using Partial Reconfiguration," in 2019 29th International Conference on Field Programmable Logic and Applications (FPL), Barcelona, Spain: IEEE, Sep. 2019.
[3] F. Steinert, N. Schelten, A. Schulte, and B. Stabernack, "Hardware and Software Components towards the Integration of Network-Attached Accelerators into Data Centers," in 2020 23rd Euromicro Conference on Digital System Design (DSD), Kranj, Slovenia: IEEE, Aug. 2020.
[4] A. Vaishnav, K. D. Pham, and D. Koch, "A Survey on FPGA Virtualization," in 2018 28th International Conference on Field Programmable Logic and Applications (FPL), Dublin, Ireland: IEEE, Aug. 2018, pp. 131–1317. doi: 10.1109/FPL.2018.00031.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISERS: SEP MAKHSOUS