

Coyote V2: An Open-Source FPGA Shell for Modern Distributed Data Processing



Dr. Shien Zhu

Systems Group, ETH Zurich, Switzerland

Email: shien.zhu@inf.ethz.ch

Introduction

FPGAs have proven to be excellent prototyping platforms for next-generation systems thanks to their stream-like nature and configurability. However, practically integrating them into realistic systems remains challenging.

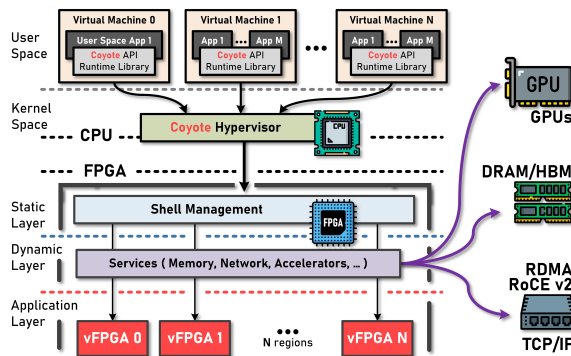


Figure 1: Overview of Coyote V2

In this 3-hour tutorial, we introduce Coyote V2, an open-source infrastructure for implementing SmartNICs on FPGAs as highly versatile, network-enabled acceleration platforms. Coyote V2 provides abstractions of FPGA resources and RDMA interfaces for in-network processing and advanced network management, in addition to DPU-like local acceleration. We will introduce our infrastructure and showcase how its high-level abstractions benefit the FPGA deployment, including accelerating quantized models from Python, offloading various functions (encryption, compression, recommender model pre-processing) to the network datapath, and distributed database transaction processing.



Acknowledgment

This work was supported in part by AMD and ETH Zurich, who provided access to the Heterogeneous Accelerated Compute Cluster (HACC) at ETHZ.

Learning Objectives

Attendees will achieve the following learning objectives:

- List the benefits of FPGA abstraction with shells
- Know the architecture and interfaces of Coyote V2
- Build customized FPGA applications with Coyote V2

Additional learning objectives:

- Integrate high-level programming languages like SpinalHDL/Scala and Python with Coyote V2
- Automate the application deployment workflow

Outline

This tutorial will take ~3 hours. The schedule is as follows:

10-minute Introduction: Presenting an overview of recent cloud and data center trends, hardware specialization and infrastructure for the AI/ML era. We will also introduce the external page AMD Heterogeneous Accelerated Compute Cluster (HACC) at ETH Zurich as a platform to support research in computer systems, architecture and networking.

50-minute Coyote V2: Presenting Coyote V2 basics and interfaces. Showing how to seamlessly deploy an accelerated application on an FPGA in a few lines of C++. Additionally, examples and live demos of hybrid computer systems with FPGAs and GPUs will be presented.

15-minute Q/A

30-minute Network Applications: Giving an overview and live demo of our RDMA stack, RoCE BALBOA. It will showcase how to perform 100G RDMA networking in a few lines of C++ code between two FPGAs as well as a FPGA and a commodity NIC (Mellanox-5).

30-minute AI Applications: Showcasing how to use FPGA-based SmartNICs to deploy ML pre-processing onto the network datapath and deploy low-latency quantized neural networks by utilizing the memory interfaces.

30-minute Database Applications: Presenting another example on FPGA-based distributed transaction processing with Coyote V2. It features efficient development with SpinalHDL and simulation-based design space exploration.

15-minute Q/A