Bridging Software and Hardware: Stateless Remote Accelerator Calls

Ziyi Yang, Krishnan B. Iyer, Yixi Chen, Ran Shu^, Zsolt István*, Marco Canini, Suhaib A. Fahmy

*KAUST, ^Microsoft Research, *TU Darmstadt*

Background

Non-expressive switching ASIC

- Shallow program stack depth
- No Loop
- No FP support
- Coupled mem. access w/ stage

Partial offload for common semantics

Flexible FPGA enables complex varied accelerators

Host-controlled FPGA offload overhead

Throughput-sensitive Application

Latency-sensitive Application

Batching hides data movement costs only works for throughput-sensitive applications!

Request Reassembly

Request data size exceeds MTU, like tensor (3KB~3MB)

Request fragmented when reaching FPGA

Request reassembly crucial for virtualizing accelerator service across multiple clients for our run-to-completion model

Accelerator Slot Abstraction

Top-k

Slot #1

Slot #2

Accel. Slots

Compiled accelerators stored in FPGA DRAM for fast configuration

Loading accelerators into slots within single-digit milliseconds

System Design

OffRAC hosts multiple accelerator slots on FPGA, accessible using remote accelerator calls

Network Stack

Request Assembler

Top-k

CNN

Accel. Slots

OffRAC

FPGA

Client sends Remote Accelerators Calls

Requests fragmented over network transport

Request fragments arrive at FPGA

OffRAC assembles fragments into requests (virtual queues)

Slots dynamically load accelerators as needed to service requests

Accelerator execute request calls (run to completion)

Reroute request for further process

Response sent to network stack

Preliminary Results

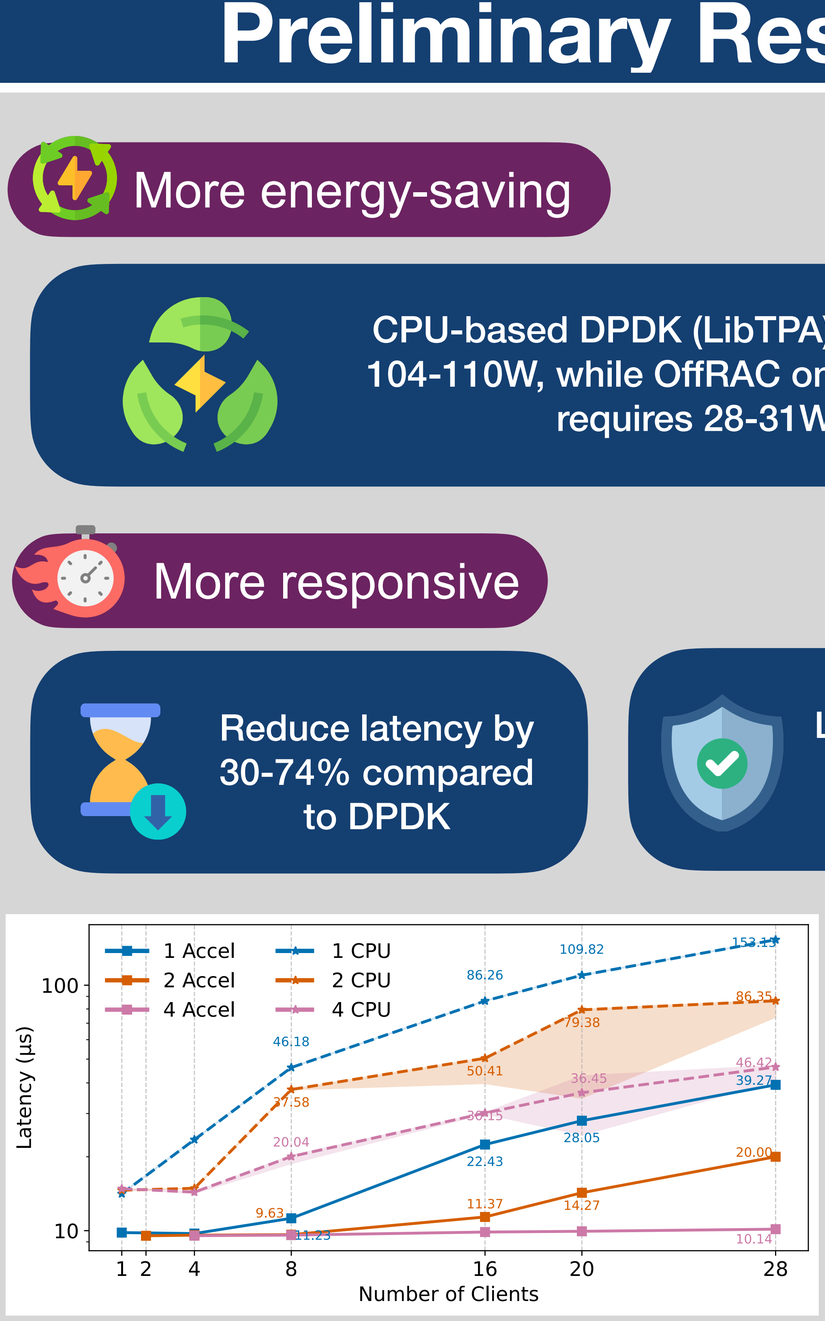
More energy-saving

CPU-based DPDK (LibTPA) consumes 104-110W, while OffRAC on FPGA only requires 28-31W

More responsive

Reduce latency by 30-74% compared to DPDK

Latency consistent until accelerator saturates



Digital Version

Ziyi Yang <ziyi.yang@kaust.edu.sa> Krishnan B. Iyer <krishnan.iyer@kaust.edu.sa> Yixi Chen <yixi.chen@kaust.edu.sa>
Marco Canini <marco@kaust.edu.sa> Suhaib A. Fahmy <suhaib.fahmy@kaust.edu.sa>