Saman Biookaghazadeh, Ming Zhao, Fengbo Ren
Arizona State University

Are FPGAs Suitable For Edge Computing?

Arizona State University

VISA

Outline

- Introduction
- Background
- Methodology
- Experimental Results

Introduction

- Future of Internet-of-Things (IoT) by **2020**
  - IoT will connect **50 billion devices**
  - It is expected to generate **400 Zetta Bytes** of data annually
- Cloud infrastructure is falling short!
  - Cannot handle such large and distributed amount of data
  - Mainly designed for **time-insensitive** applications, **end-users,** processing **batches** of data
- Solution?
  - A new paradigm called **edge computing**
  - Serve **time-sensitive** IoT applications, and support various streaming I/O channels

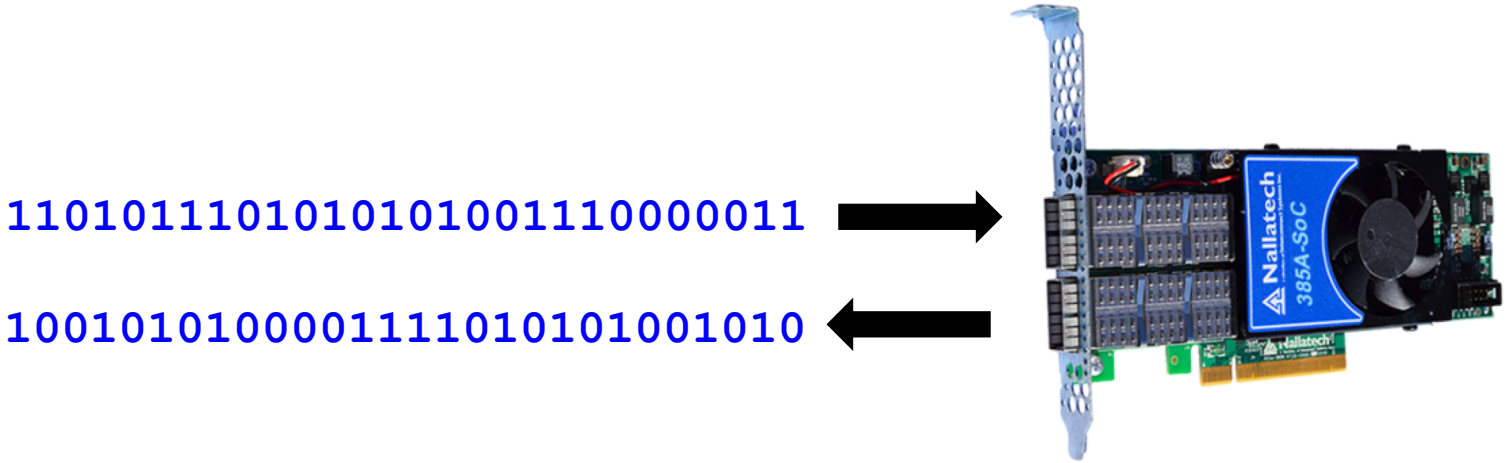
Limitations of Existing Solutions

- **How about existing cloud and edge servers?**
  - Simply a **miniature** version of cloud servers
  - Architectured for using CPUs and GPUs
    - For **batches of data**, **power hungry**, and unpredictable performance

- **What do we need?**
  - New hardware for the new paradigm

Background of FPGA

- **Reconfigurable** Farm of logic
- Opportunity to program using C, C++ and **OpenCL**
- Inherently efficient for **streaming** applications
- Suitable for a wide range of applications
  - **Spatial** parallelism, parallelism in *space*
  - **Temporal** parallelism, parallelism in *time*
- Power Efficient
  - Improve thermal stability and reduce cooling cost

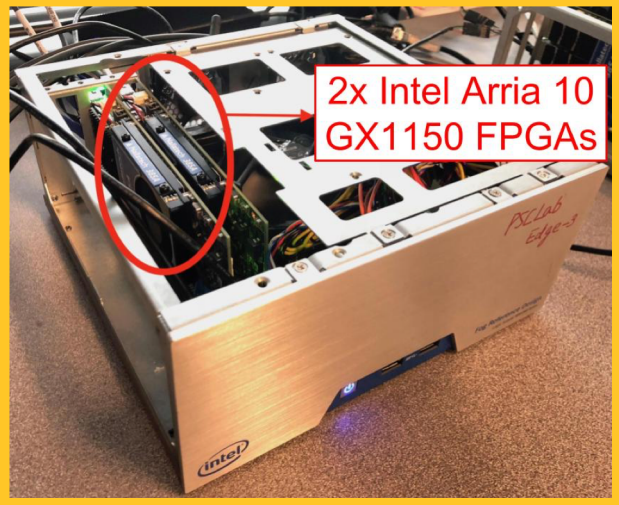


Motivations for FPGA-based Edge Computing

- **Edge computing's most important requirements**
  - **Predictable** performance for IoT service providers
  - Operational in locations with **limited power supply**
  - **Accelerate** a wide variety of service applications

- We study suitability of FPGAs with respect to:
  - **Sensitivity** of processing throughput to the workload size
  - **Adaptiveness** to algorithm concurrency and dependency
  - **Energy efficiency**

Testbench



Nallatech 385A
(Intel Arria A10)

Intel Xeon E5-1275

32GB Main Memory

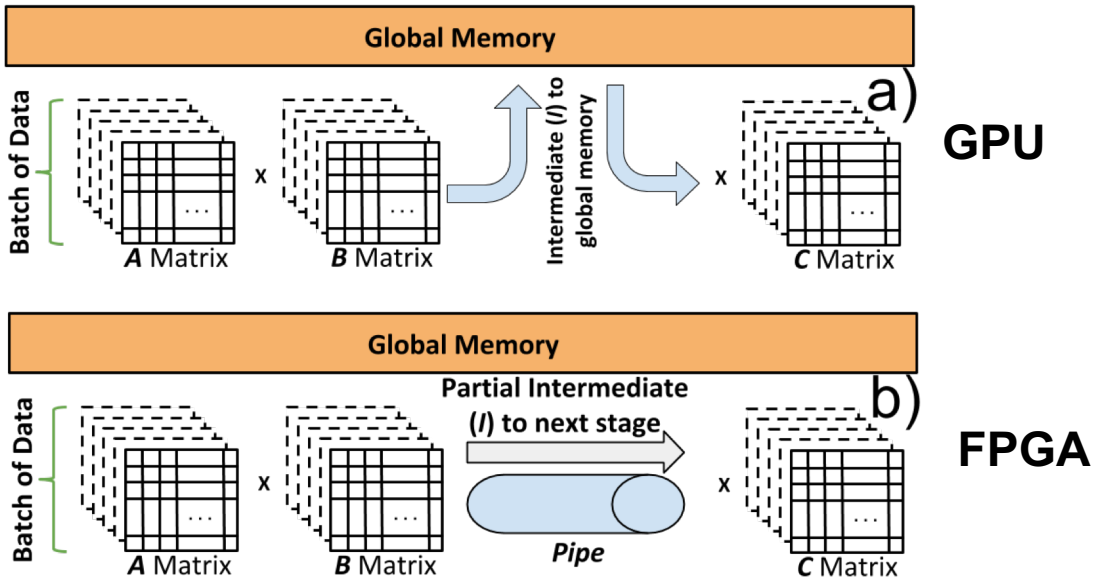
Tesla K40m

Intel Xeon E5-2637

64GB Main Memory

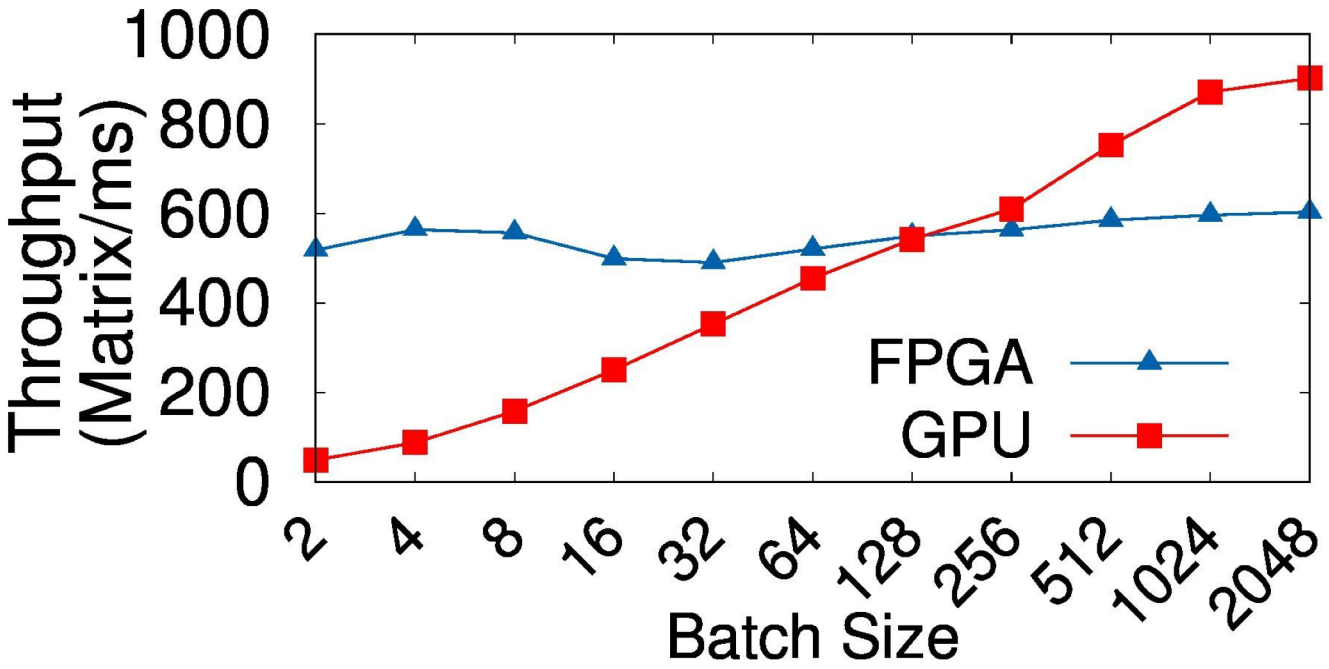
Sensitivity to Workload Size

- Two stage matrix multiplication (*A x B x C*) as a benchmark
  - Widely used in linear algebraic algorithms
- **32 x 32** matrices, with single-precision floating-point random numbers
- Varying batch size between **2** to **2048**



Sensitivity to Workload Size

- FPGA reads input from the **Ethernet I/O**
- GPU reads input from the card main memory
- Unlike GPU, FPGA can provide **consistent throughput**
  - By jointly exploiting spatial and temporal parallelism

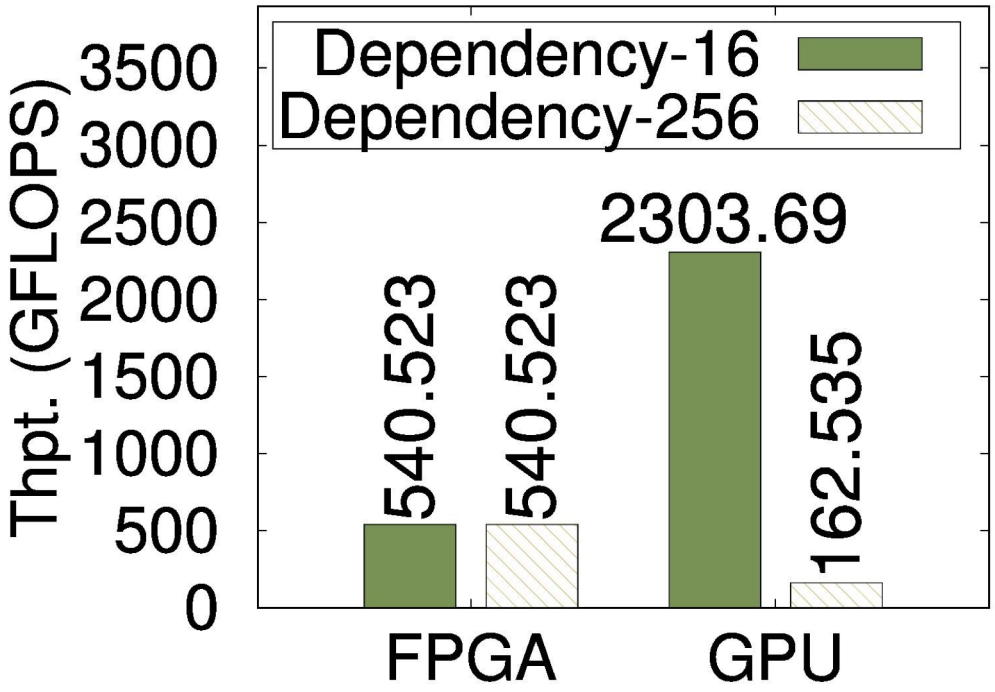


Adaptiveness

- How well FPGAs and GPUs **adapt** to algorithm **characteristics?**
- ***Data Dependency***: Dependency across different iterations of a loop
- ***Conditional Dependency***: Dependency on conditional statements with each iteration of the loop
- Benchmark
  - Simple iterative block (*for-loop*)
  - Each iteration performs certain number of operations
  - **Generic enough** to model large set of computationally intensive applications

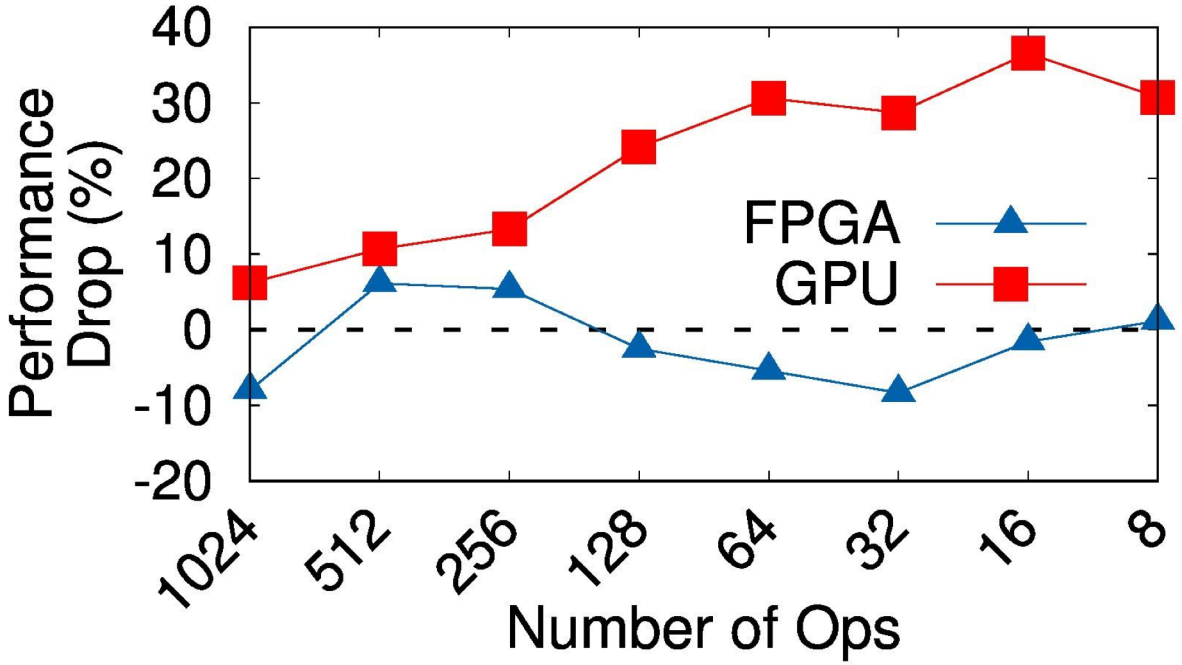
Data Dependency

- Introducing dependency among different iterations
- Varying the data dependency **degree**
  - Changing the size of the **group**
- GPUs performance closely depends on available data parallelism
- FPGAs can **exploit pipelining** and execute iterations, regardless of dependency



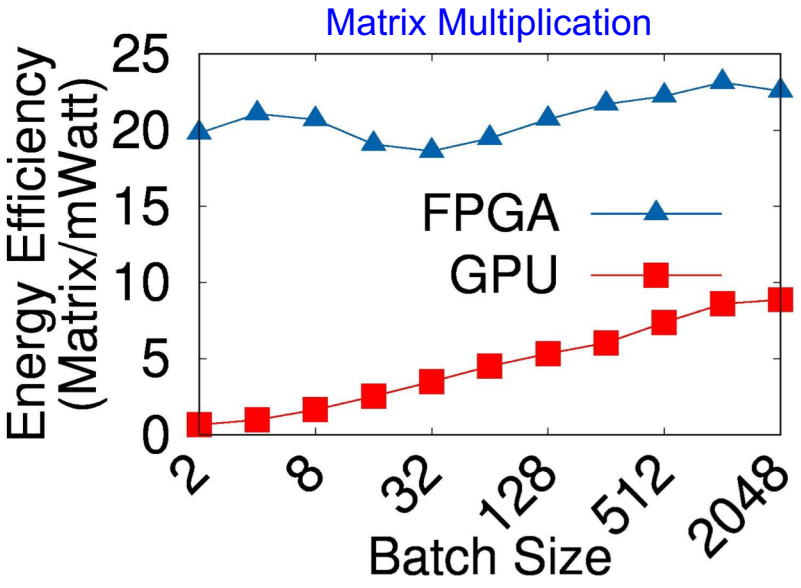
Conditional Dependency

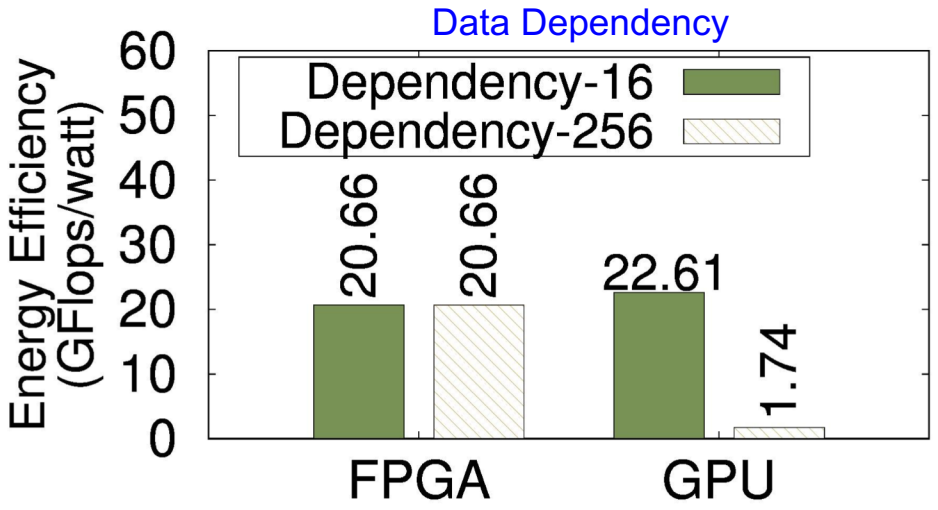
- Adding *if-else* statements into the loop
- Each branch contains **half** of the operations
- Varying number of operations in each *if* and *else* block
- Different devices show different behaviours:
  - GPU is highly sensitive to conditional statement
  - FPGA can utilize a **look-up table**



Energy Efficiency

- Collecting energy consumption on both devices
  - *Nvidia-smi* on the GPU
  - Nallatech *MMD Layer API* on FPGA.
- Varying workload input size





Conclusions and Future Works

- FPGAs can handle unique edge requirements
- FPGA can be considered as a core computational accelerator in the emerging edge systems
- FPGAs can provide **predictive throughput**, Algorithm **adaptiveness**, and **energy-efficiency**

- Future Directions?
  - Studying edge workloads
  - Studying other algorithms characteristics and suitability of different hardware architectures
  - Scalability (Up & Out) of FPGAs compared to GPUs

Acknowledgement

- **Sponsors**
  - National Science Foundation (CNS-1629888)
  - Intel FPGA university program

- **VISA LAB**: http://visa.lab.asu.edu

Question?

Supported by NSF CNS-1629888 and Intel FPGA University Program