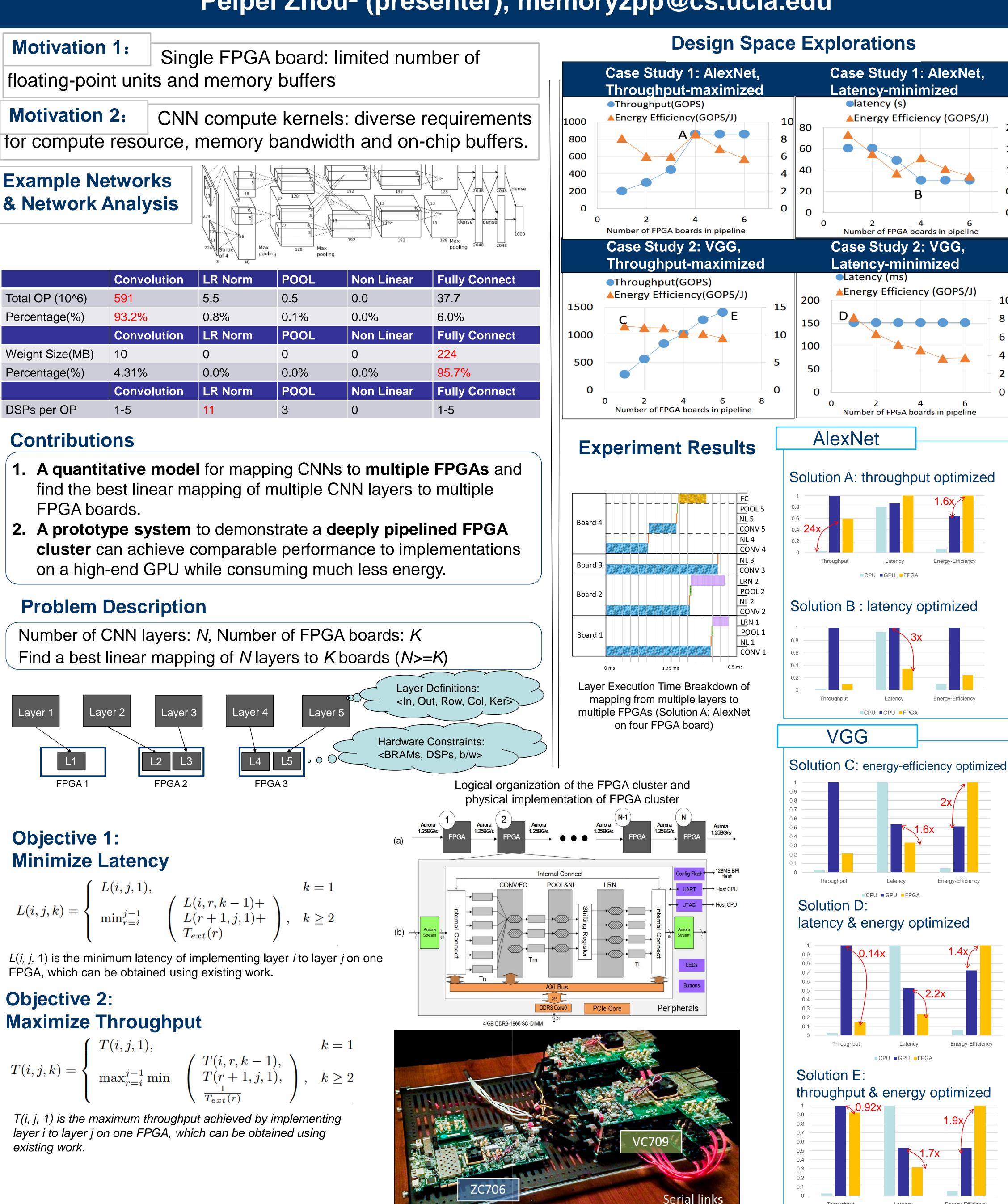# Energy-Efficient CNN Implementation on a Deeply Pipelined FPGA Cluster

Chen Zhang[1], Di Wu[2], Jiayu Sun[1], Guangyu Sun[1,3], Guojie Luo[1,3], Jason Cong[1,2,3]
Peipei Zhou[2] (presenter), memoryzpp@cs.ucla.edu

Motivation 1:
Single FPGA board: limited number of floating-point units and memory buffers

Motivation 2:
CNN compute kernels: diverse requirements for compute resource, memory bandwidth and on-chip buffers.

Example Networks & Network Analysis

	Convolution	LR Norm	POOL	Non Linear	Fully Connect
Total OP (10^6)	591	5.5	0.5	0.0	37.7
Percentage(%)	93.2%	0.8%	0.1%	0.0%	6.0%
	Convolution	LR Norm	POOL	Non Linear	Fully Connect
Weight Size(MB)	10	0	0	0	224
Percentage(%)	4.31%	0.0%	0.0%	0.0%	95.7%
	Convolution	LR Norm	POOL	Non Linear	Fully Connect
DSPs per OP	1-5	11	3	0	1-5

Contributions

1. **A quantitative model** for mapping CNNs to **multiple FPGAs** and find the best linear mapping of multiple CNN layers to multiple FPGA boards.
2. **A prototype system** to demonstrate a **deeply pipelined FPGA cluster** can achieve comparable performance to implementations on a high-end GPU while consuming much less energy.

Problem Description

Number of CNN layers: N, Number of FPGA boards: K
Find a best linear mapping of N layers to K boards ($N>=K$)

Layer Definitions: <In, Out, Row, Col, Ker>

Hardware Constraints: <BRAMs, DSPs, b/w>

Objective 1: Minimize Latency

$$L(i,j,k) = \begin{cases} L(i,j,1), & k=1 \\ \min_{r=i}^{j-1} \begin{pmatrix} L(i,r,k-1)+ \\ L(r+1,j,1)+ \\ T_{ext}(r) \end{pmatrix}, & k \geq 2 \end{cases}$$

$L(i, j, 1)$ is the minimum latency of implementing layer i to layer j on one FPGA, which can be obtained using existing work.

Objective 2: Maximize Throughput

$$T(i,j,k) = \begin{cases} T(i,j,1), & k=1 \\ \max_{r=i}^{j-1} \min \begin{pmatrix} T(i,r,k-1), \\ T(r+1,j,1), \\ \frac{1}{T_{ext}(r)} \end{pmatrix}, & k \geq 2 \end{cases}$$

$T(i, j, 1)$ is the maximum throughput achieved by implementing layer i to layer j on one FPGA, which can be obtained using existing work.

Design Space Explorations

Case Study 1: AlexNet, Throughput-maximized

Case Study 1: AlexNet, Latency-minimized

Case Study 2: VGG, Throughput-maximized

Case Study 2: VGG, Latency-minimized

Experiment Results

Layer Execution Time Breakdown of mapping from multiple layers to multiple FPGAs (Solution A: AlexNet on four FPGA board)

AlexNet

Solution A: throughput optimized

Solution B : latency optimized

VGG

Solution C: energy-efficiency optimized

Solution D: latency & energy optimized

Solution E: throughput & energy optimized

Logical organization of the FPGA cluster and physical implementation of FPGA cluster

[1]Center for Energy-Efficient Computing and Applications, Peking University, Beijing, China
[2]Computer Science Department, University of California, Los Angeles, USA
[3]PKU/UCLA Joint Research Institute in Science and Engineering