

Towards AI Democratization

— Linking Software and Hardware Designs

Weiwen Jiang, Ph.D.

Postdoc Research Associate

Department of Computer Science and Engineering

University of Notre Dame

wjiang2@nd.edu | <https://wjiang.nd.edu>

This project is supported in part by the EdgeCortix Inc. via National Science Foundation I/UCRC center under grants CNS-1822099, and in part by Facebook and IBM.

Embedded Computing Hardware Has Been in Every Corner



Agriculture



Military



Power System



Manufacture



Education



Medical Operation



Finance

.....

Today, AI is Going to Every Embedded Computing Hardware



Agriculture



Military



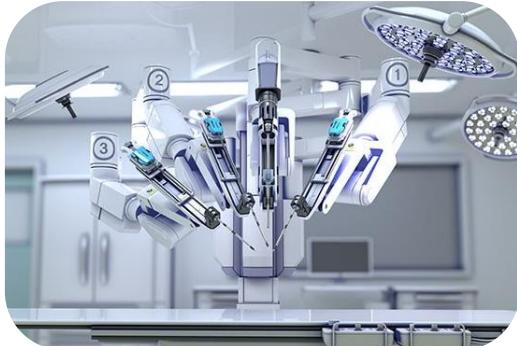
Power System



Manufacture



Education



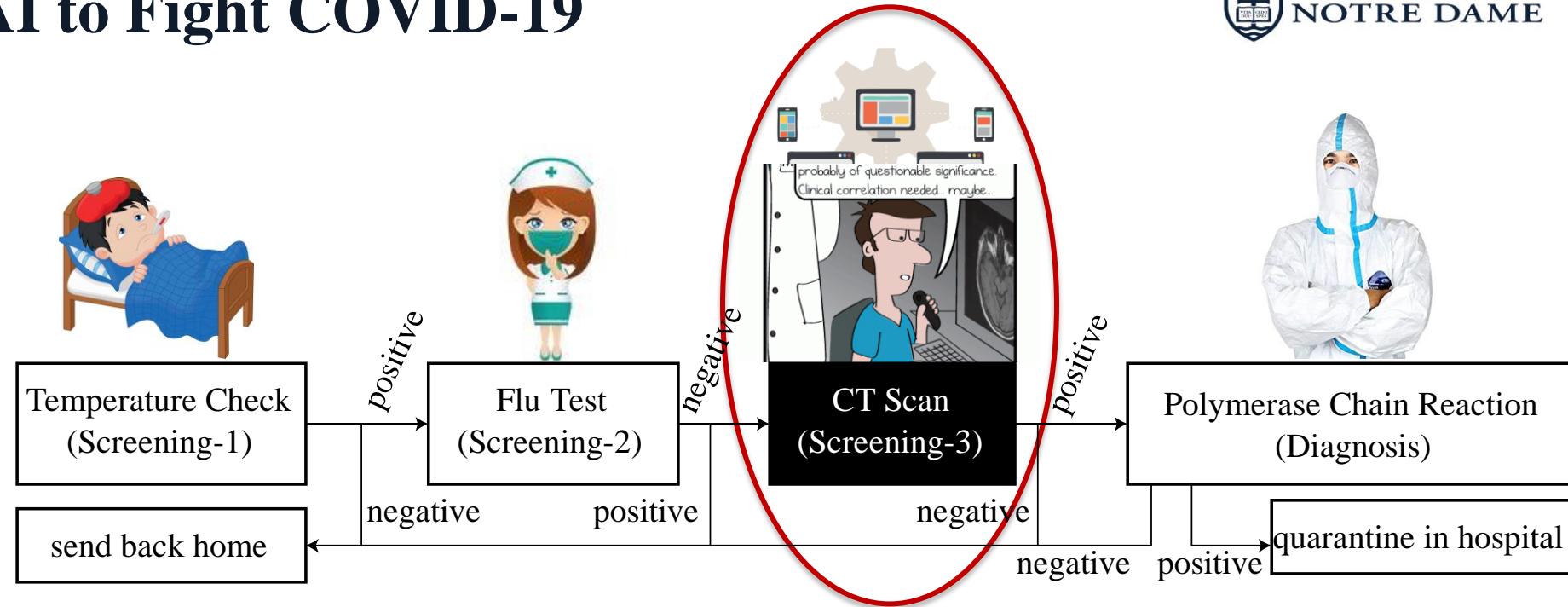
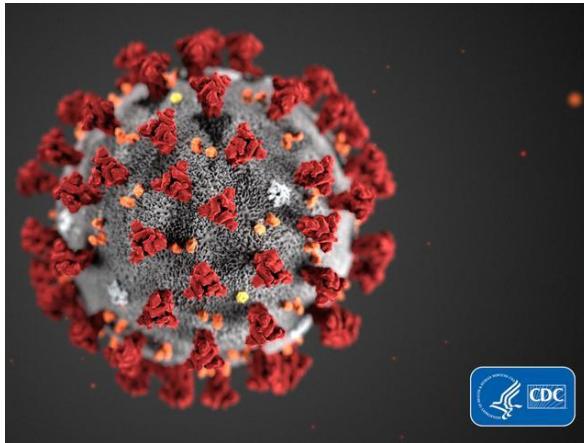
Medical Operation



Finance

.....

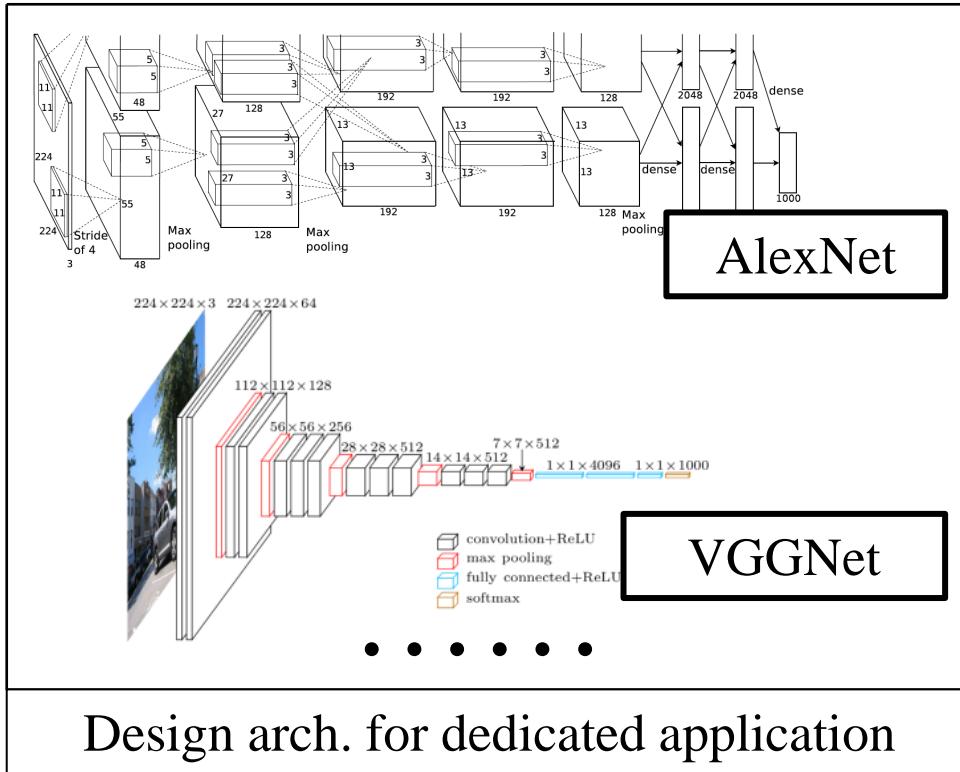
Example: Equip AI to Fight COVID-19



Challenge	Response
Shortage of rRT-PCR test kits	Accurate screening
Burden on radiologists in reading CT scan results	AI judgement to reduce burden
Days of deployment is intolerant	Plug-and-play in clinics within Hours

[ref] How a country serious about coronavirus does testing and quarantine. <https://www.youtube.com/watch?v=e3gCbkeARbY>. [Online; accessed 03/17/2020]

Manual AI Design is **TOO** Expensive in Both Domain Knowledge and Time



1 year for only 1 application

Name	Year	Acc.
AlexNet	2012	83.4%
ZFNet	2013	88.3%
VGGNet	2014	92.7%
ResNet	2015	96.4%
GoogleNet	2016	96.9%

Problem

- Domain knowledge and excessive labor
- It takes too long to devise new architectures



Challenge1:

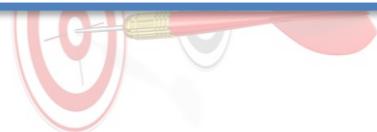
No Uniform AI Solution Works For All Scales



Low-Power

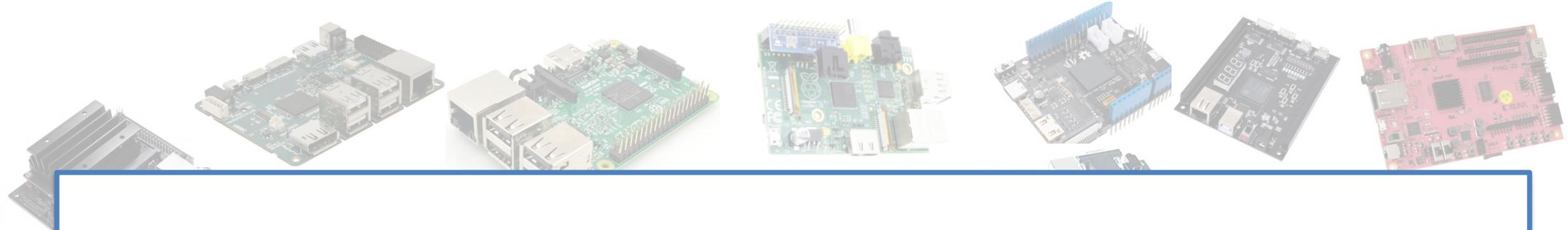


Real-Time



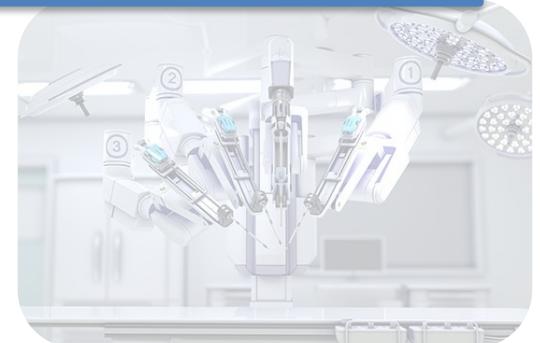
Accuracy

AI Democratization brings New Challenges



Challenge 2: in the same scale

Each application needs a specific AI Solution and Hardware Design

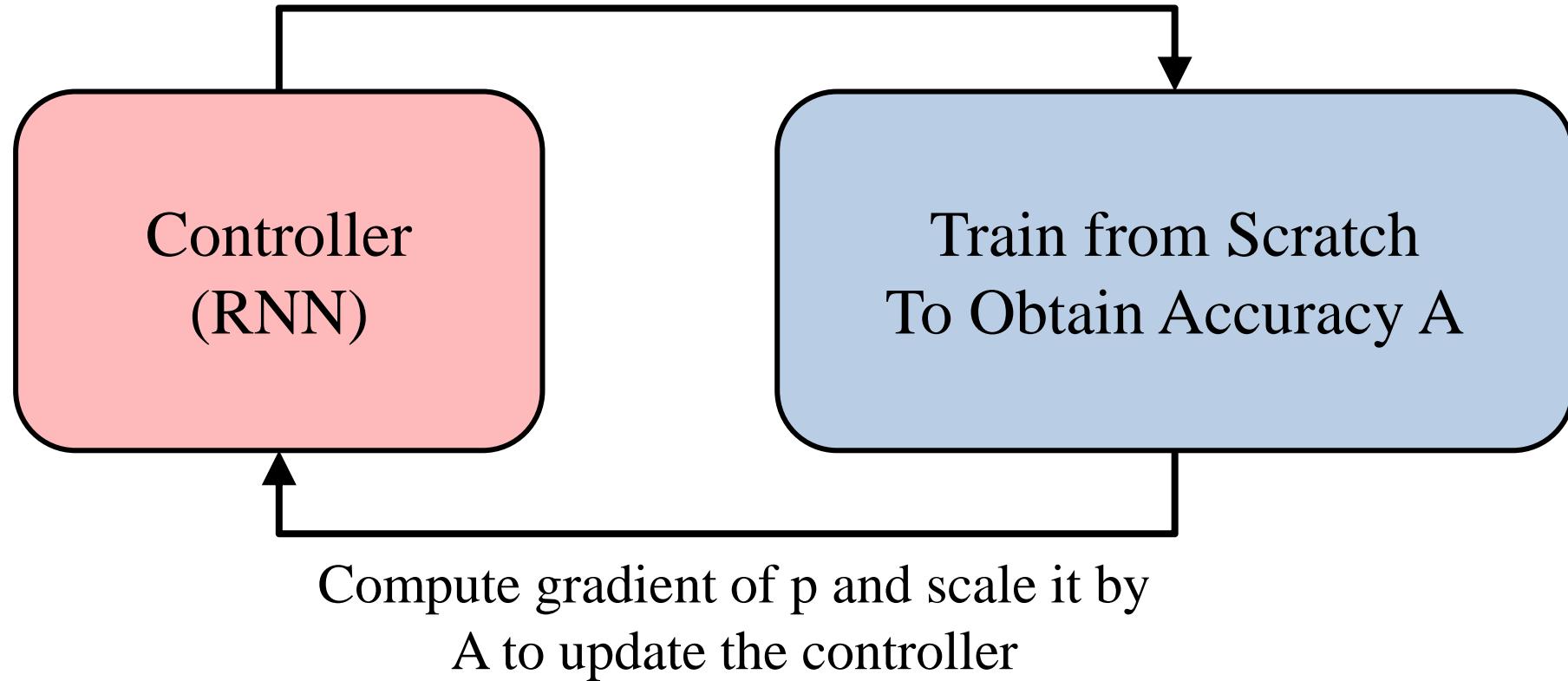


It's time to **automatically** explore good neural architectures
and tailored hardware design

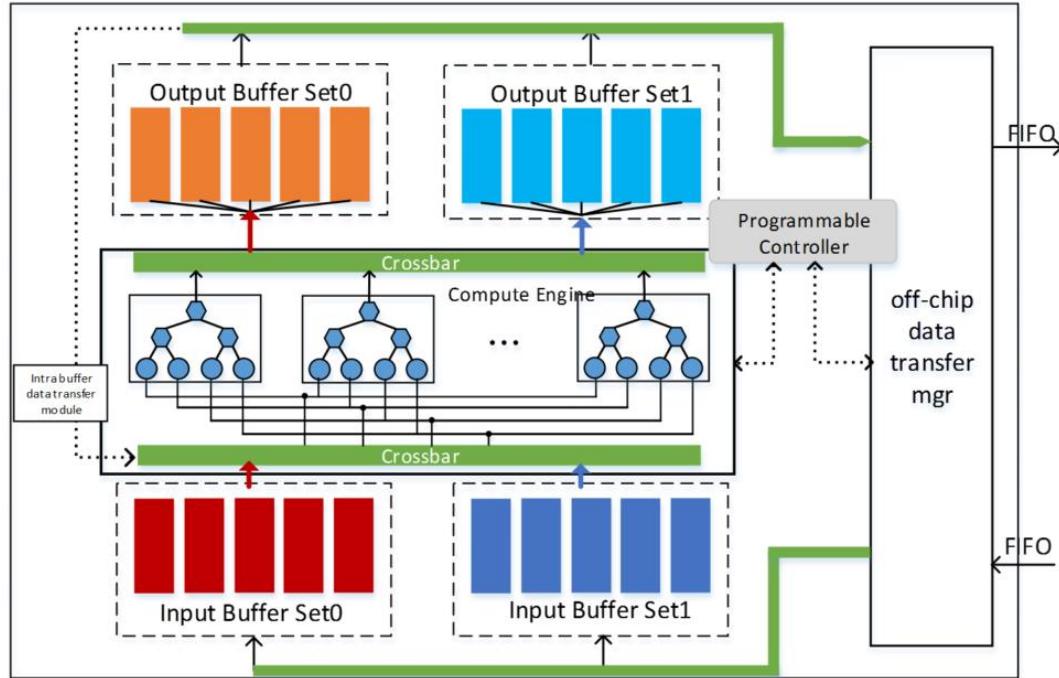
Software Automation: Neural Architecture Search (NAS)



Sample architecture NN with
probability p

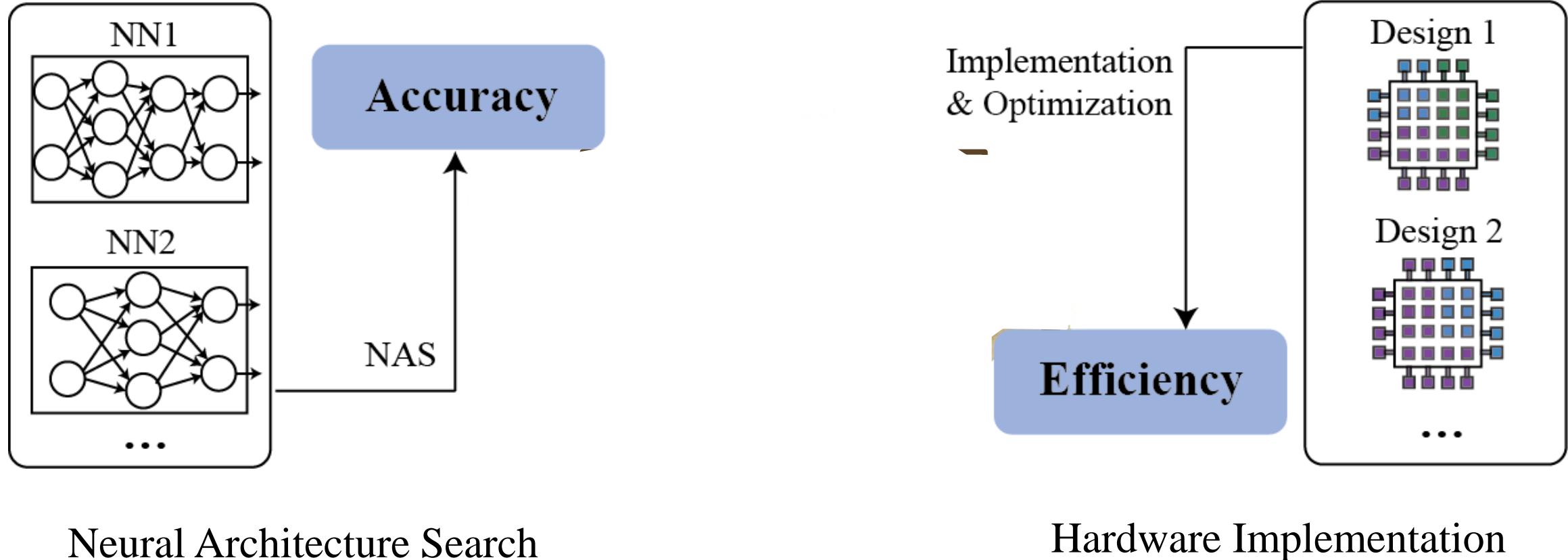


Hardware Automation: Standard Accelerator Design



Ref. C. Zhang et al. Optimizing fpga-based accelerator design for deep convolutional neural networks. In Proc. of FPGA, pages 161–170. ACM, 2015.

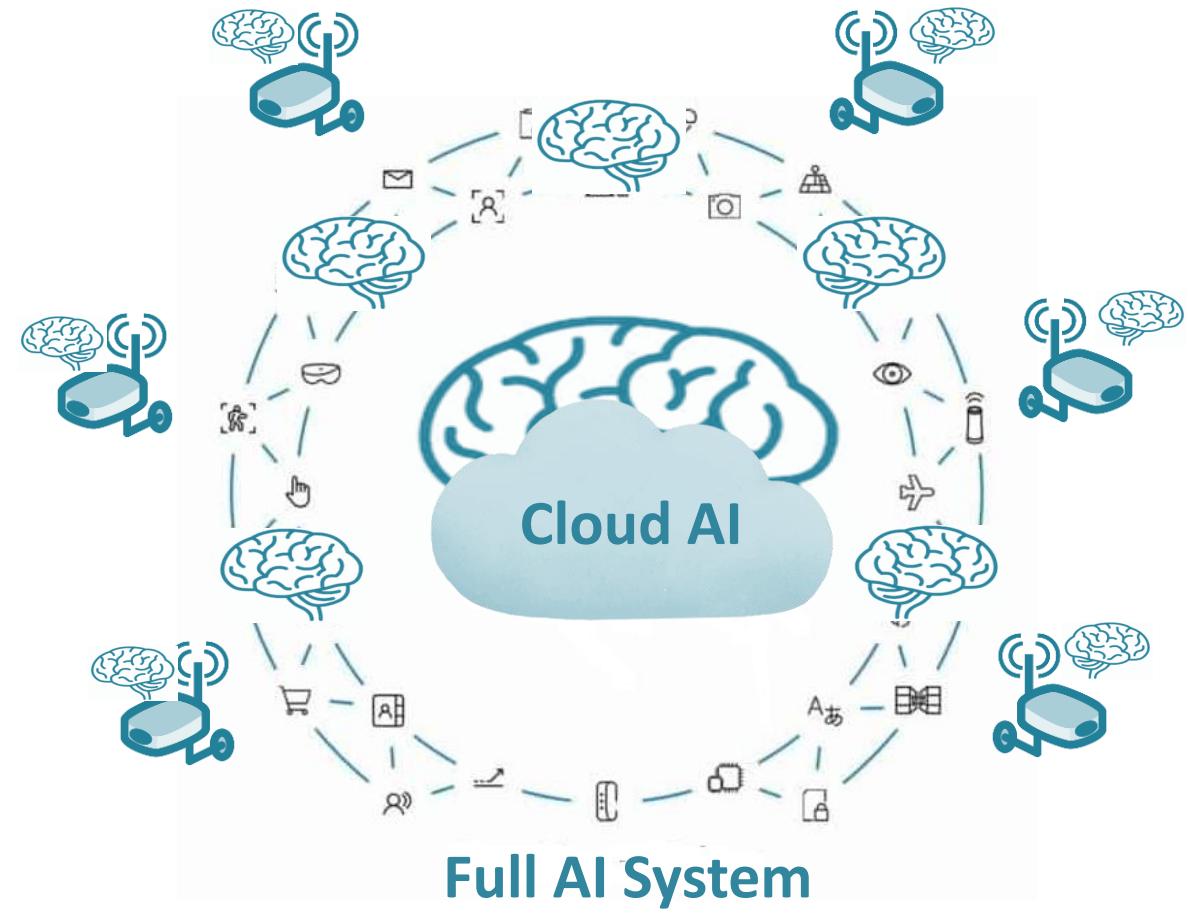
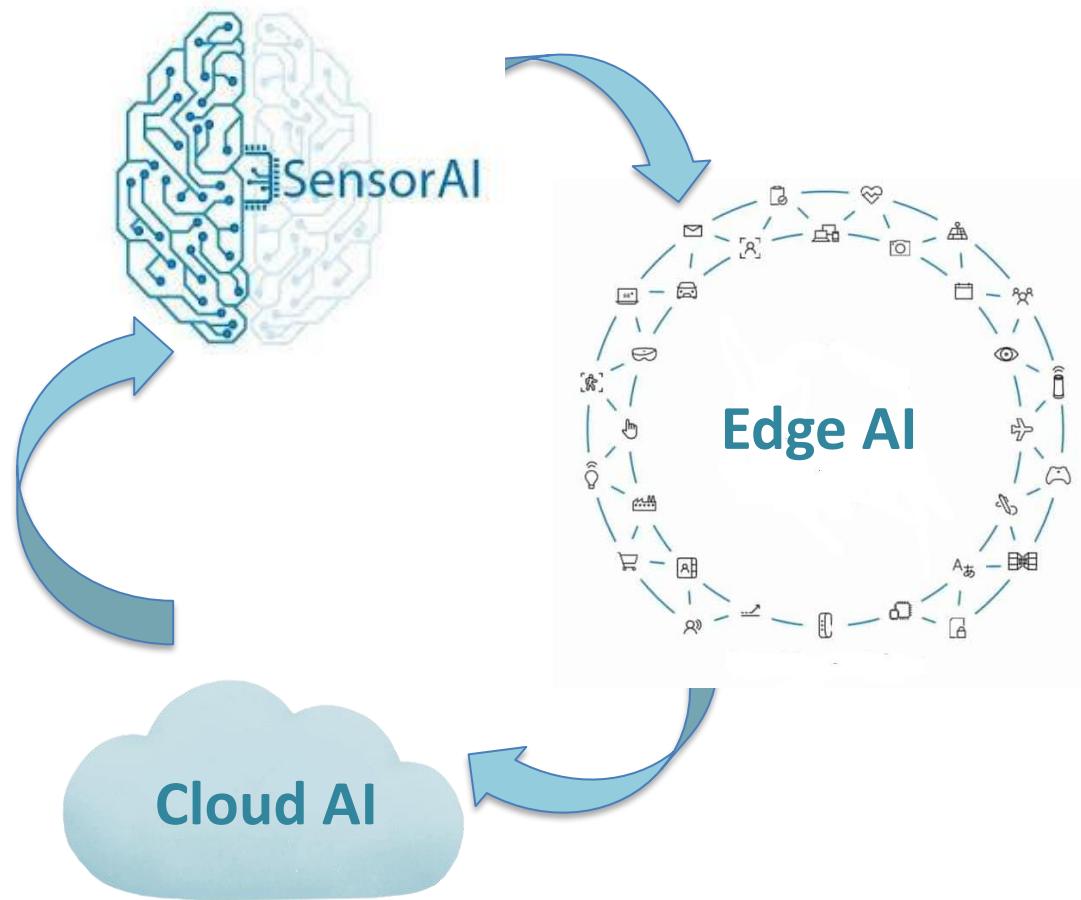
A Missing Link between Two Design Spaces



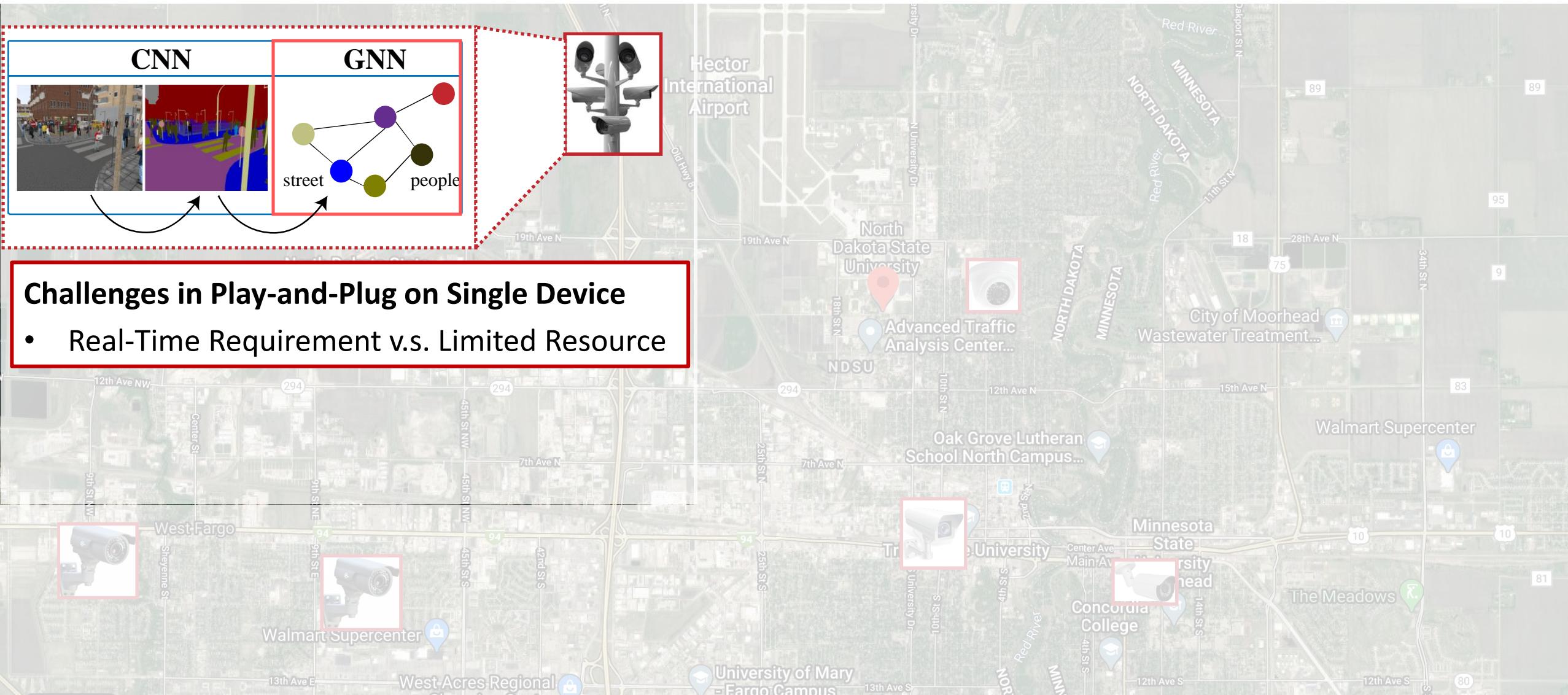
Neural Architecture Search

Hardware Implementation

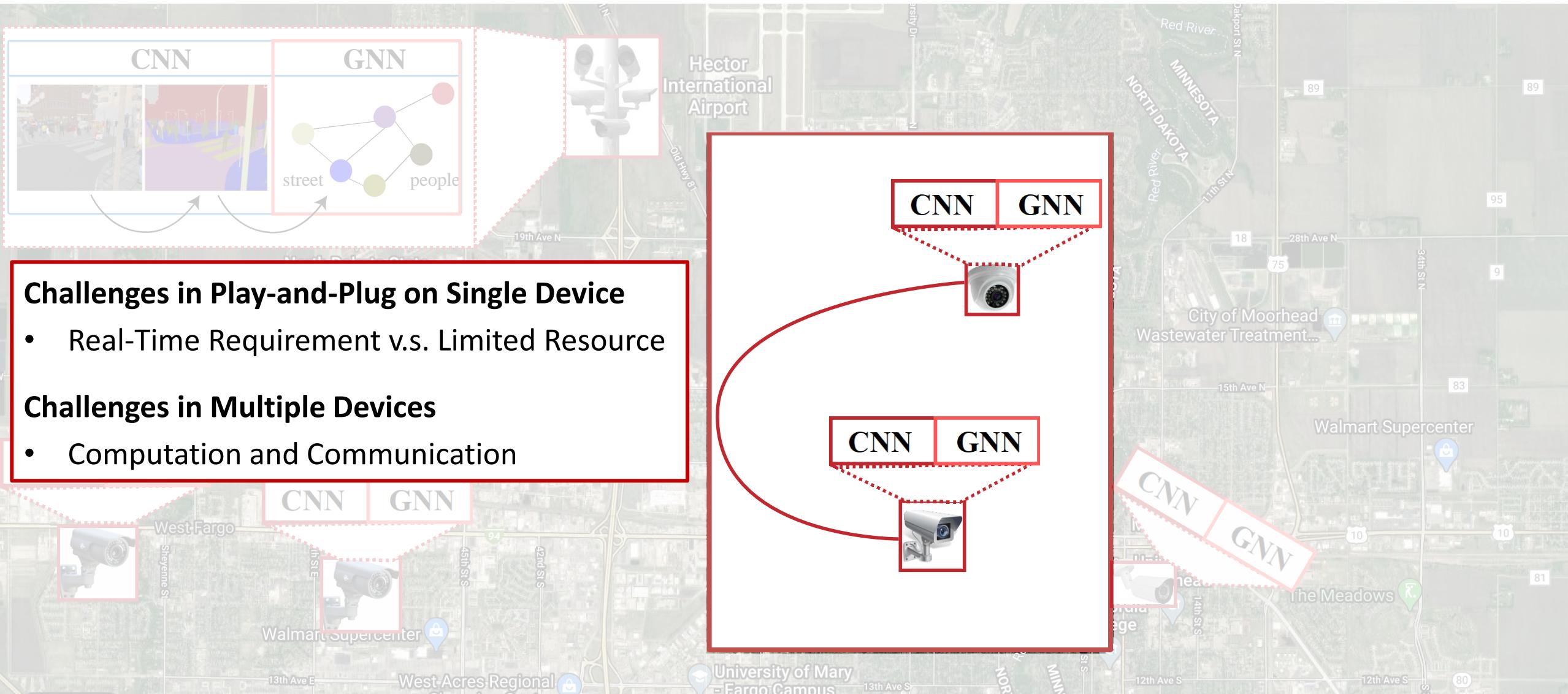
A Demands of Full System Co-Design



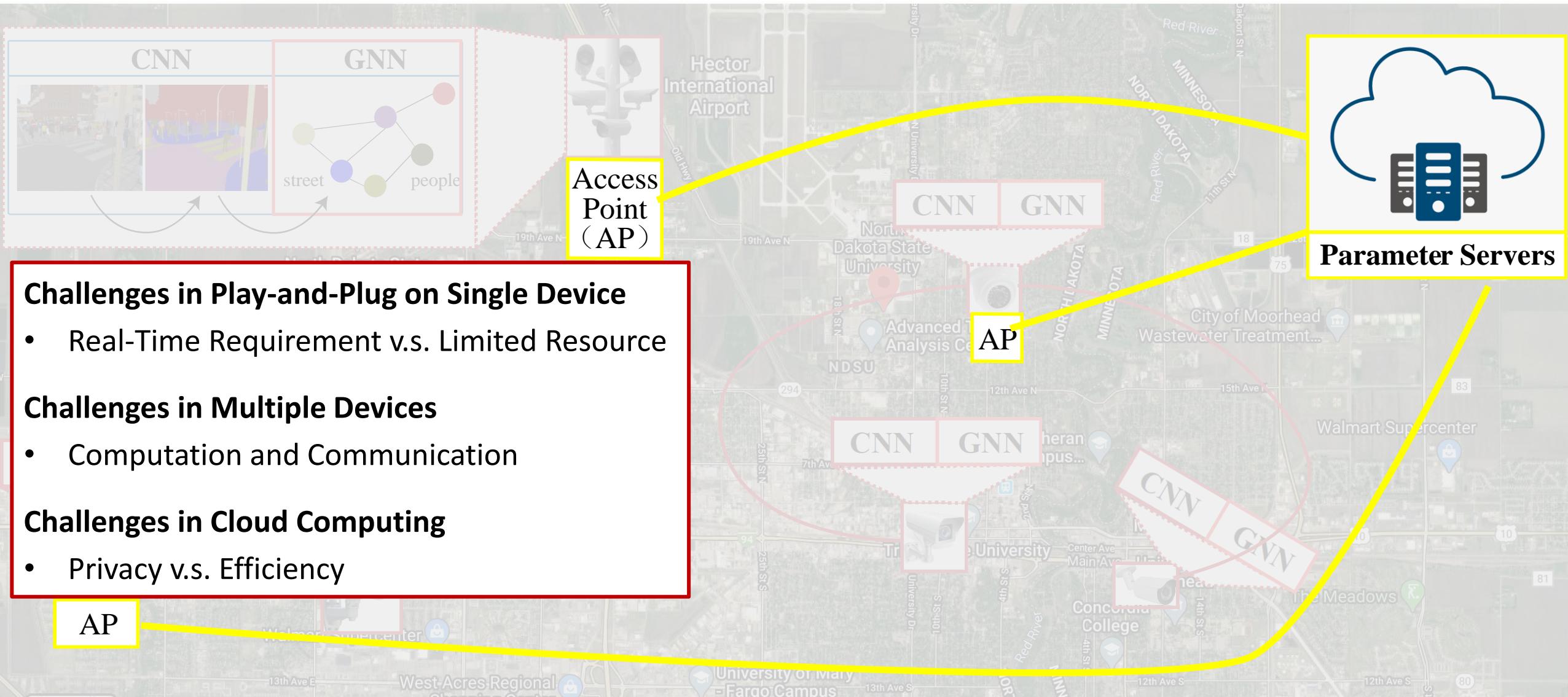
Challenges in Real Scenario: Taking Surveillance System as An Example



Challenges in Real Scenario: Taking Surveillance System as An Example



Challenges in Real Scenario: Taking Surveillance System as An Example



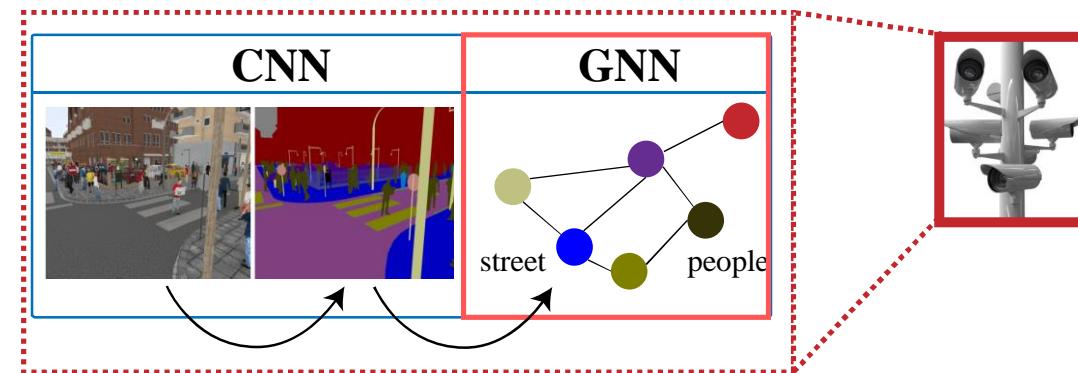
Overview of Our Co-Exploration Works:



- **FNAS (Single Device + NAS)**: DAC'19 Best Paper Nomination
- **XFER (Multiple Devices)**: CODES+ISSS'19 Best Paper Nomination
- **NASS (Secure NAS)**: ECAI'20
- HW/SW Co-Exploration: IEEE TCAD
- NANDS (NoC+NAS): ASP-DAC'20 Best Paper Nomination
- ASICNAS (ASIC+NAS): DAC'20
- NACIM (Computing-in-Memory+NAS): IEEE TC (major revision)

FNAS: Co-Design of Neural Architecture and FPGA

(DAC'19 Best Paper Nomination)



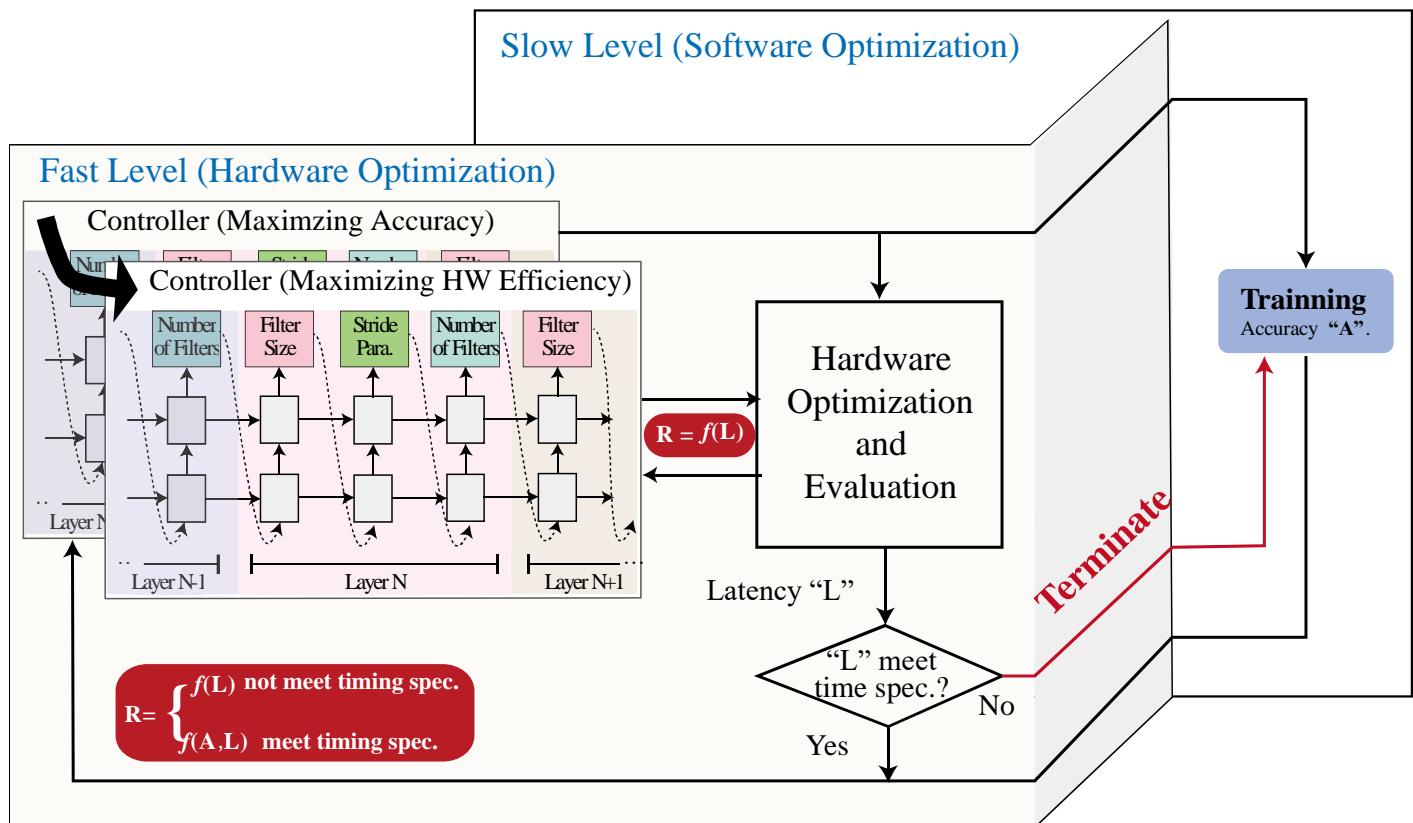
Problem Formulation

Input:

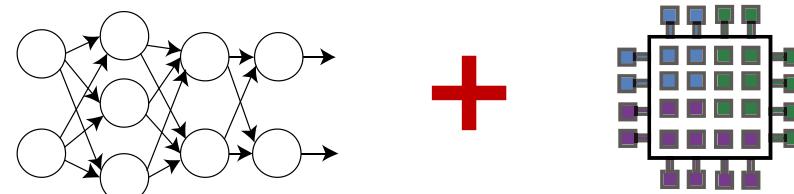
- Hardware (e.g., FPGA)
 - BRAM
 - DSPs
- Datasets (e.g. ImageNet)



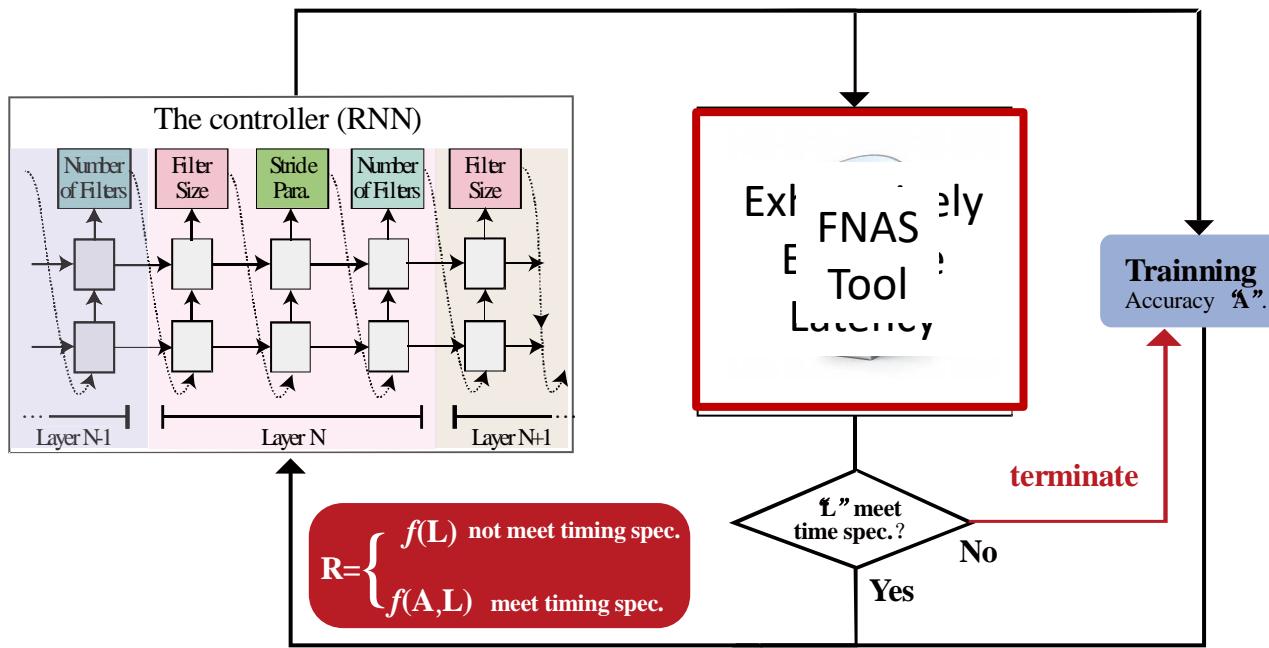
- Timing Constraints
 - e.g., 5ms



Output: A pair of neural architecture and hardware design



Possible Solutions and Challenge



Naïve Solution: HW-Aware + Exhaustively Evaluate Lat.

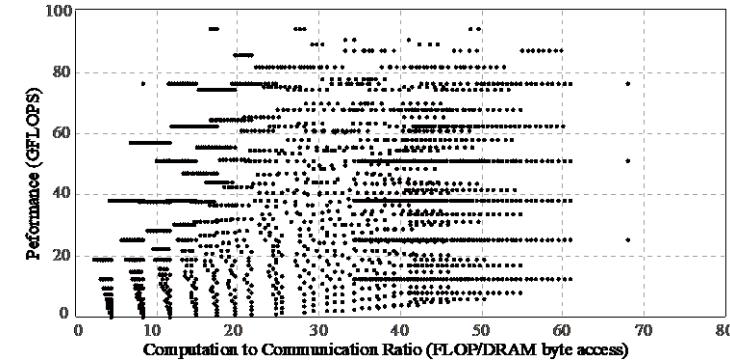


Fig1. Possible designs for Layer 5 of AlexNet on ZCU102

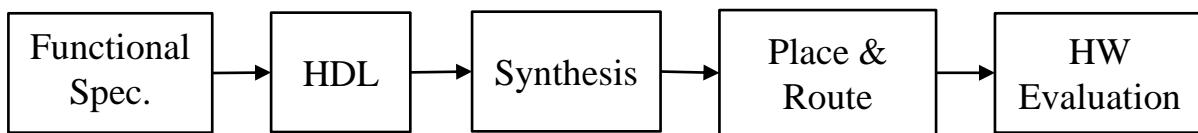


Fig2. Procedure of performance evaluation

Our Solution: FNAS tools to response to challenges

FNAS-Design C1
“Design on Program Logic”

FNAS-GG C2
“Tile-based Task Graph Generator”

FNAS-Sched C2
“Scheduler on Processing System”

FNAS-Analyzer C3
Estimate Performance “L”

Challenges:

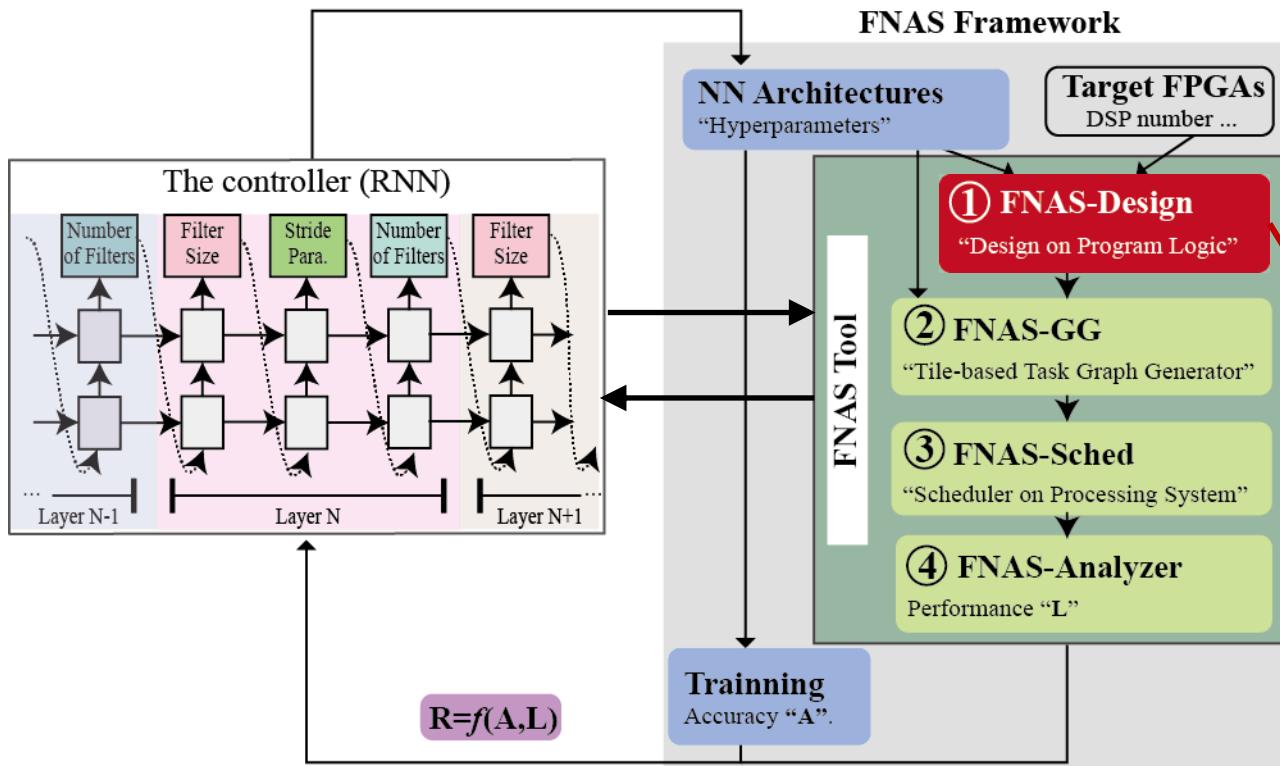
C1: Huge design space!

C2: Multi-FPGA design!

C3: Time-consuming evaluation!

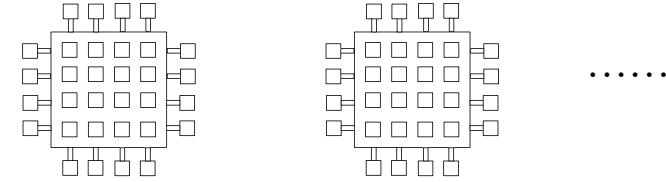
Infeasible

FNAS: Design Optimization



Given:

1. FPGAs with attributes including LUTs, DSPs, BRAM, etc.



2. A neural architecture with determined hyperparameters

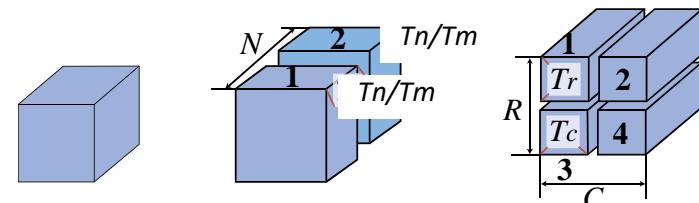
On-chip accelerator design:

Determine:

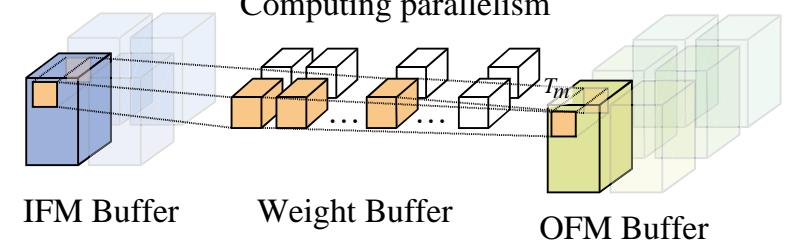
1. On-chip buffer allocation; 2. Accelerator size for computing

(note: both are determined by tiling parameters, T_m , T_n , T_r , T_c)

One layer:

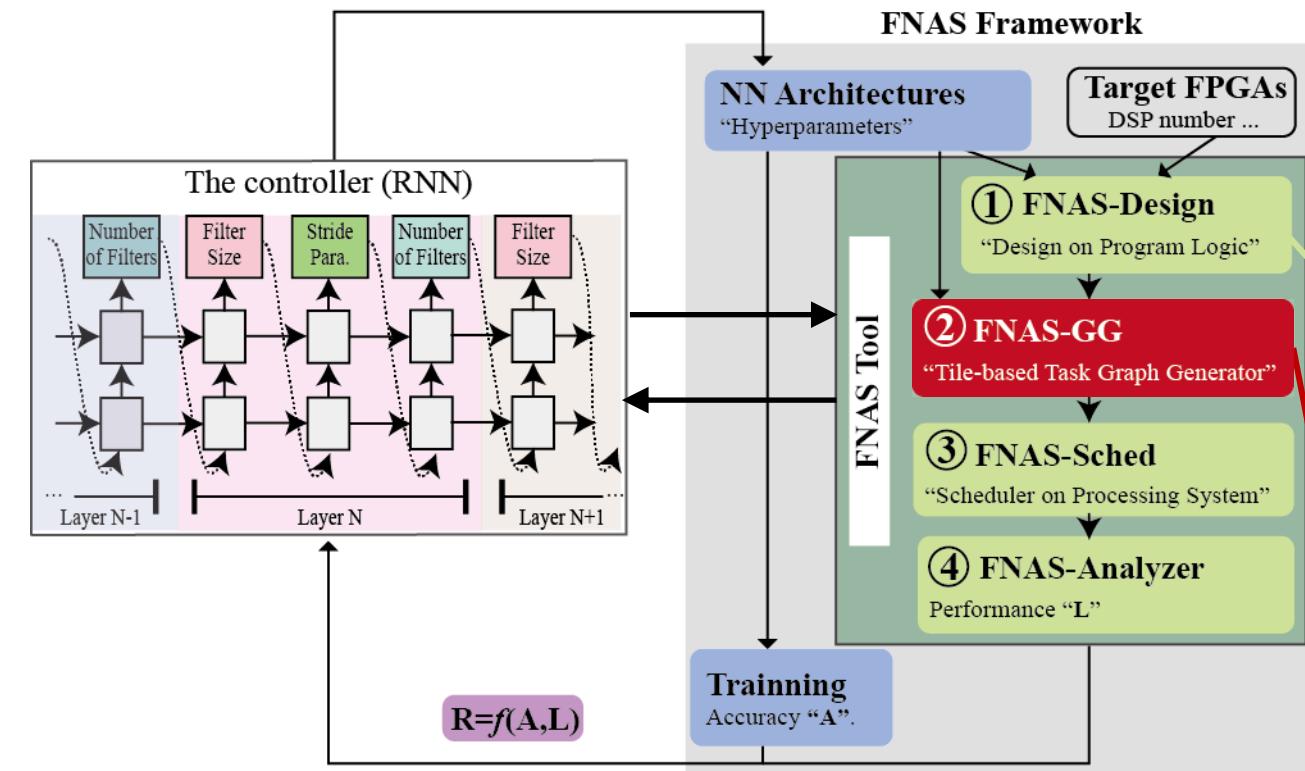


Computing parallelism



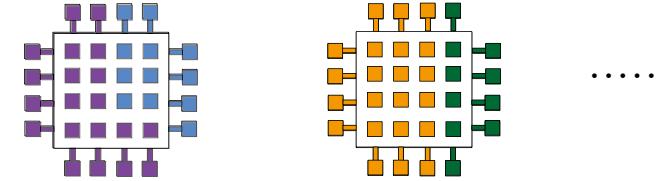
Multiple layers:

FNAS: Graph Generator

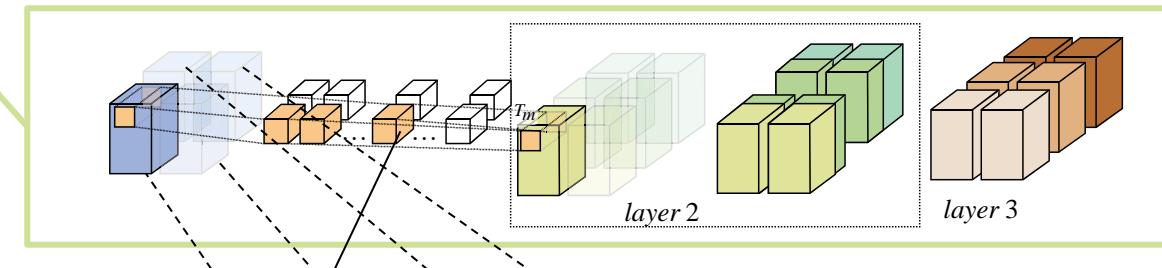


Given:

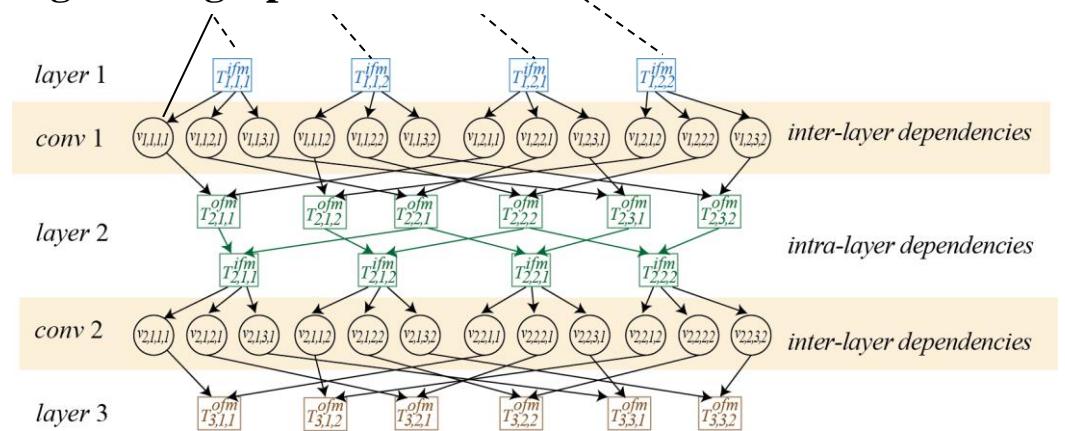
1. FPGAs with attributes including LUTs, DSPs, BRAM, etc.



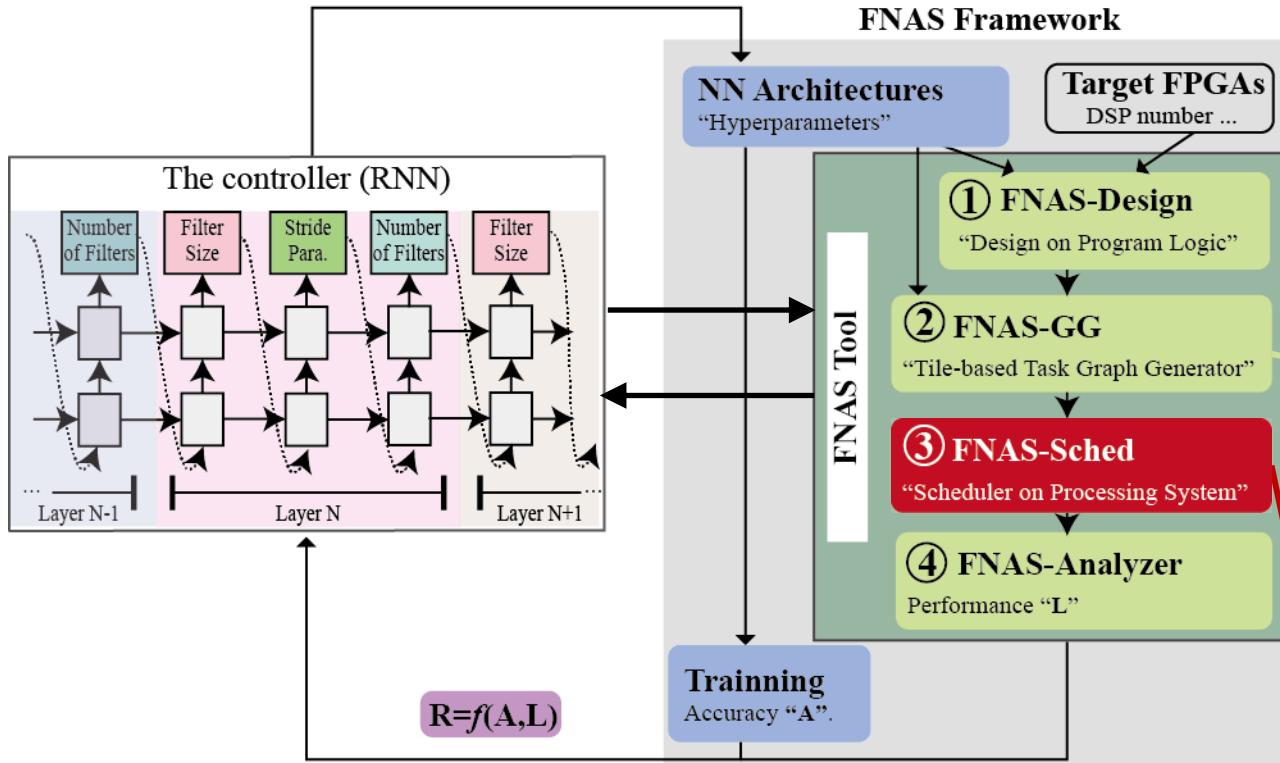
2. A neural architecture with determined hyperparameters



High-level graph abstraction

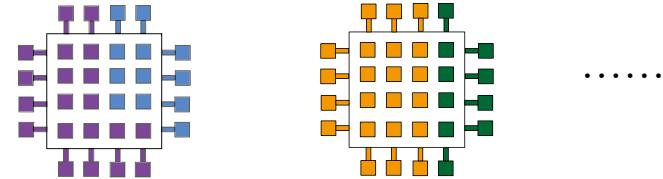


FNAS: Schedule

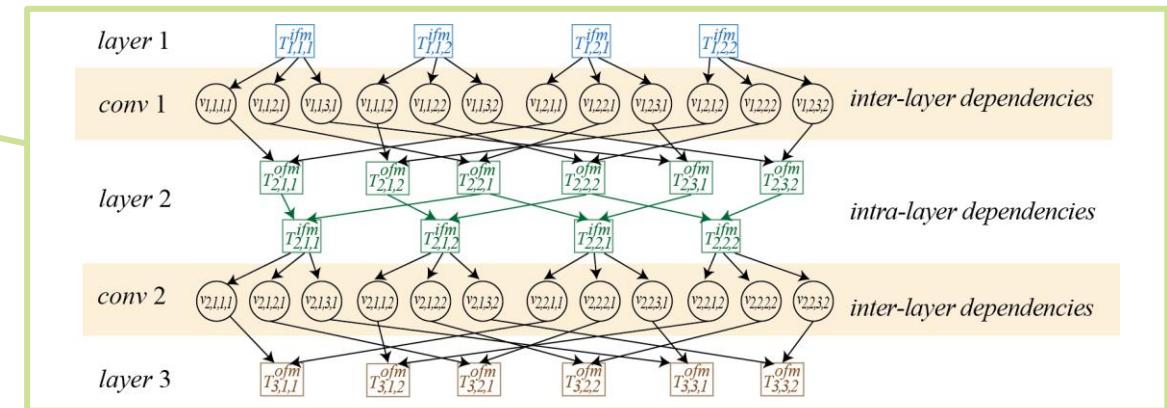


Given :

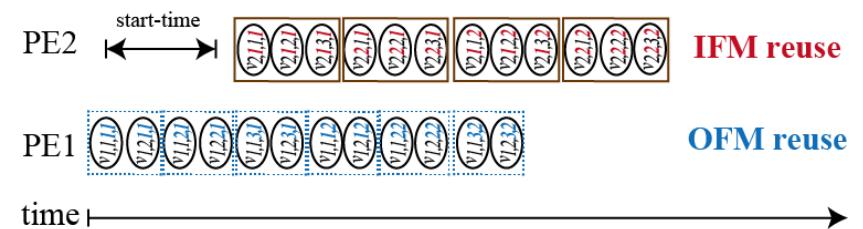
1. FPGAs with attributes including LUTs, DSPs, BRAM, etc.



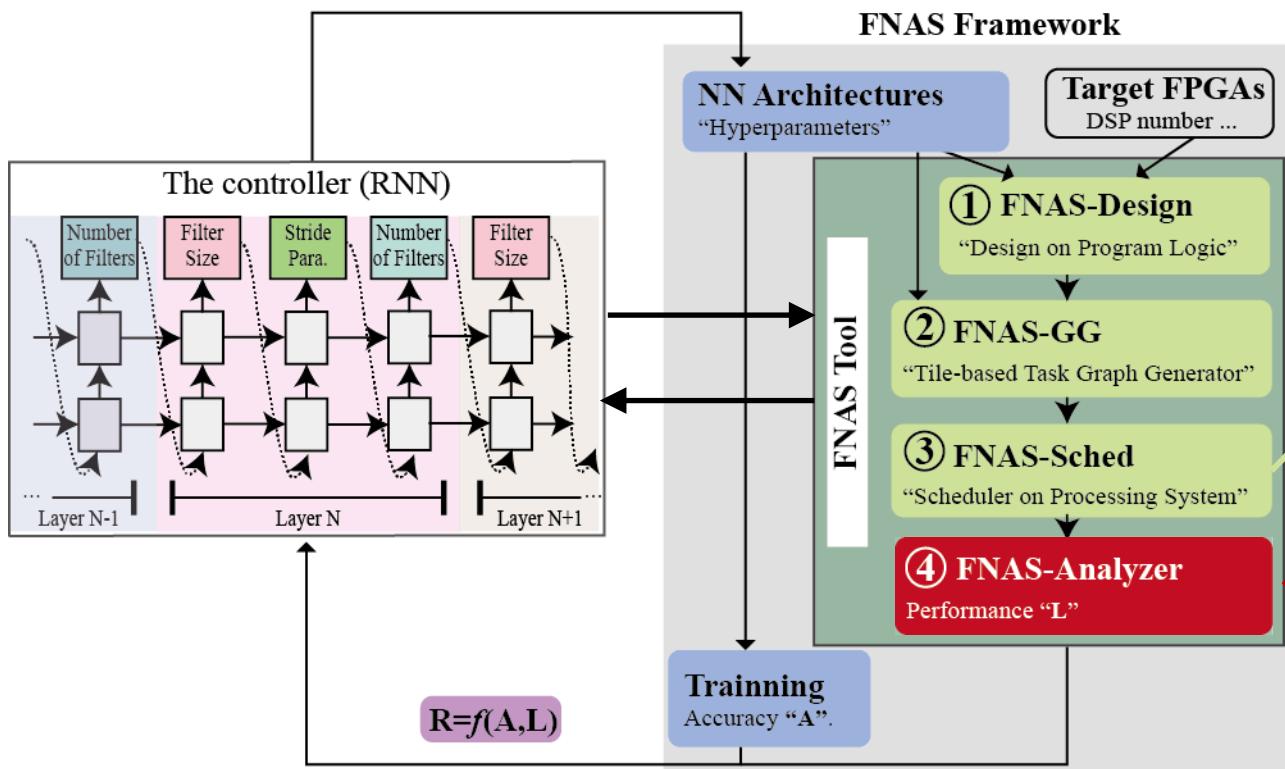
2. A neural architecture with determined hyperparameters



Schedule of tasks in graph on multiple accelerators

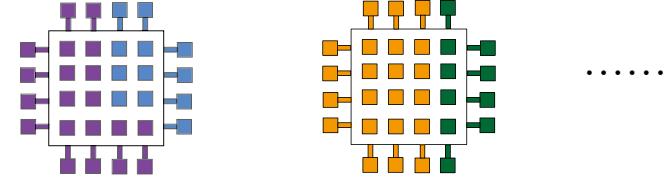


FNAS: Analyzer



Given:

1. FPGAs with attributes including LUTs, DSPs, BRAM, etc.



2. A neural architecture with determined hyperparameters



Latency = pipeline start time + processing time

Output:

1. A tailored FPGA Design
2. The system latency

Experimental Setting

FPGAs



Xilinx 7A50T



Xilinx 7Z020

Datasets

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

MNIST



CIFAR-10

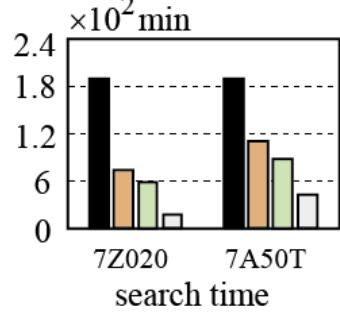
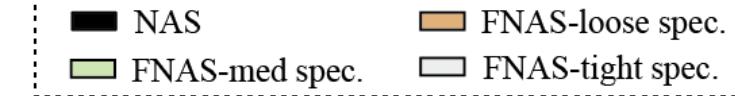


ImageNet

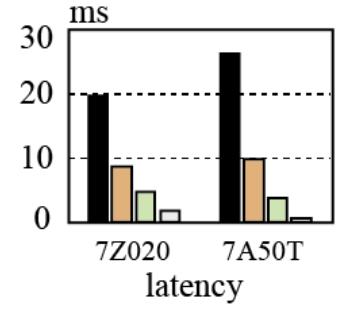
	Layer Num.	up to 5	up to 10	up to 15
NAS Search Space	Filter Size	[5, 7, 14]	[1, 3, 5, 7]	[1, 3, 5, 7]
	Filter Num.	[9, 18, 36]	[24, 36, 48, 64]	[16, 32, 64, 128]
HW Search Space	Channel Tiling Para. (Tm,Tn); Row Tiling Para. (Tr); Col Tiling Para. (Tc); Schedule			
Timing Spec. (ms)	[2, 5, 10, 20]		[1.5, 2, 2.5, 10]	[2.5, 5, 7.5, 10]

Experimental Results

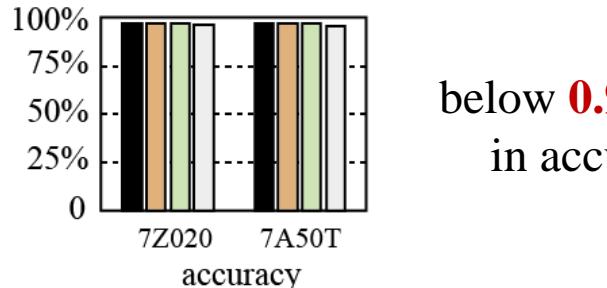
Different Hardware (MNIST)



up to **11.13X** reduction in search time



up to **7.81X** reduction in inference latency



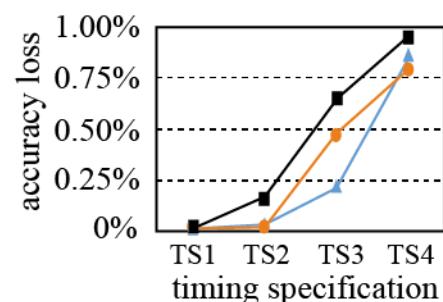
below **0.9%** loss in accuracy

Different Datasets (7Z020)



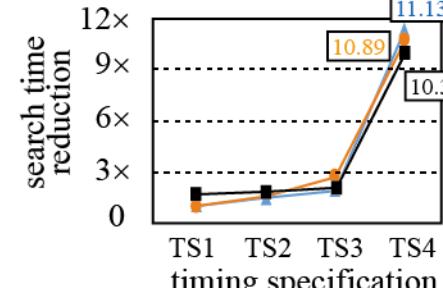
tightness of timing specification

TS1 TS2 TS3 TS4
loose → tight



Baseline: NAS

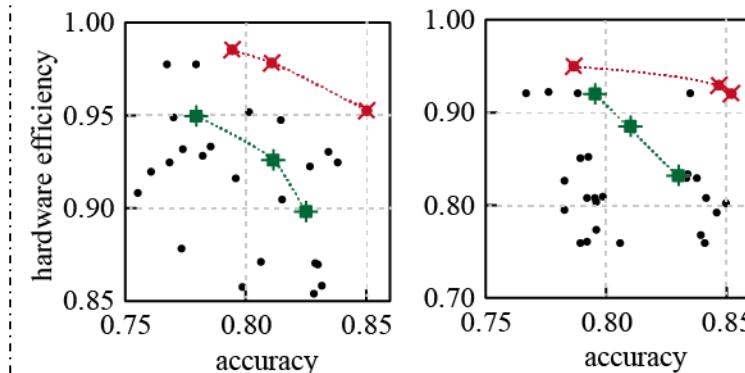
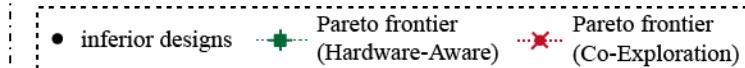
below **1%** loss in accuracy



up to **10X** reduction in inference latency

Weiwen Jiang

Compare to HW-Aware NAS (CIFAR-10 + 7Z020)



FNAS can significantly **push forward** the Pareto frontiers between **accuracy and efficiency** tradeoff

XFER: Achieving Super-Linear Speedup via Multiple FPGAs

(CODES+ISSS'19 Best Paper Nomination)



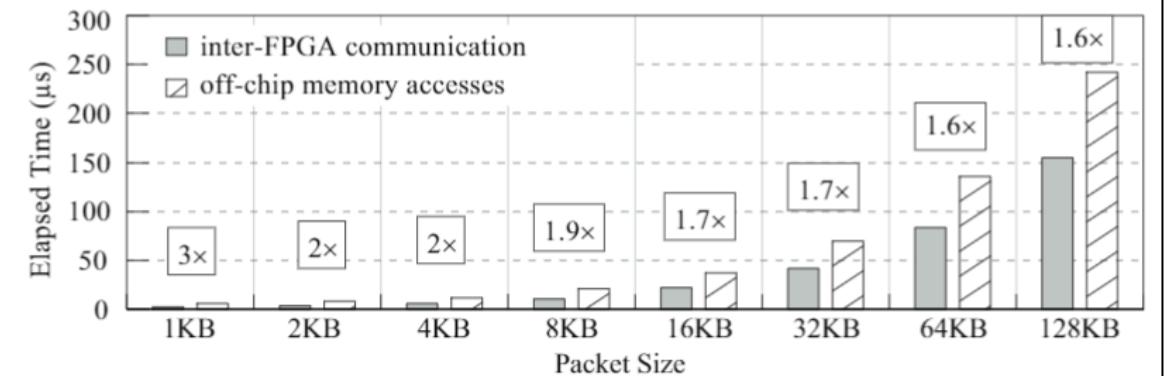
Architecture and Motivation



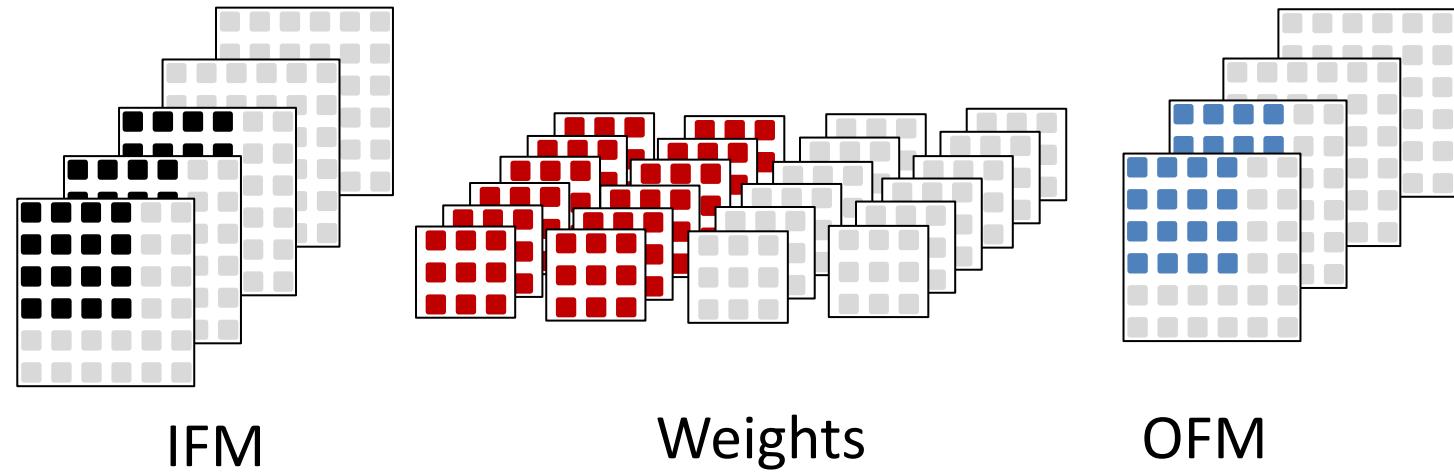
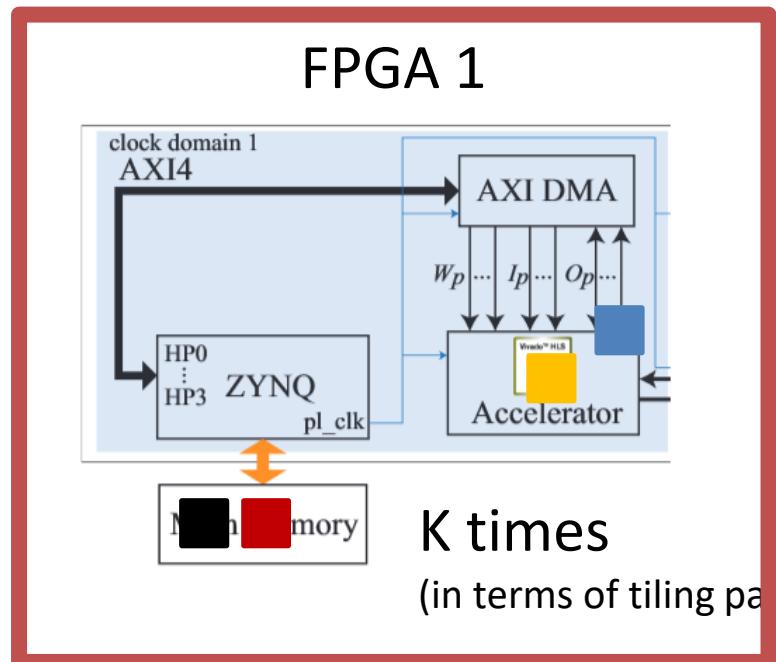
Feature

- Different **clock domains** for computation (**low**) and communication (**high**)
- Communication **not go through Off-Chip Memory**, but **directly switch between on-chip buffers**

Results & Motivations

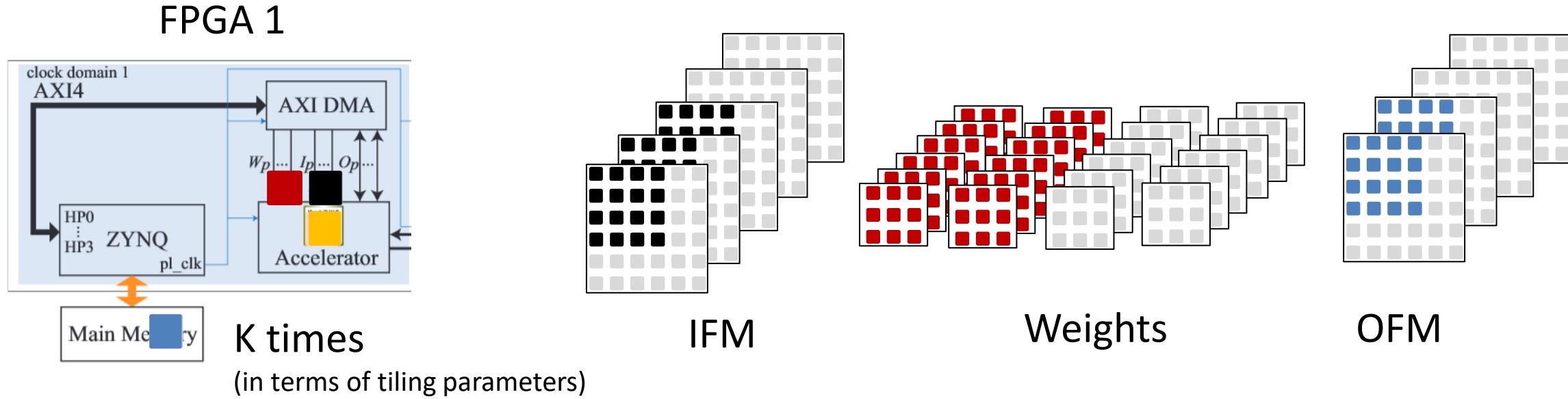


Performance on Single FPGA bounded by Limited Resource

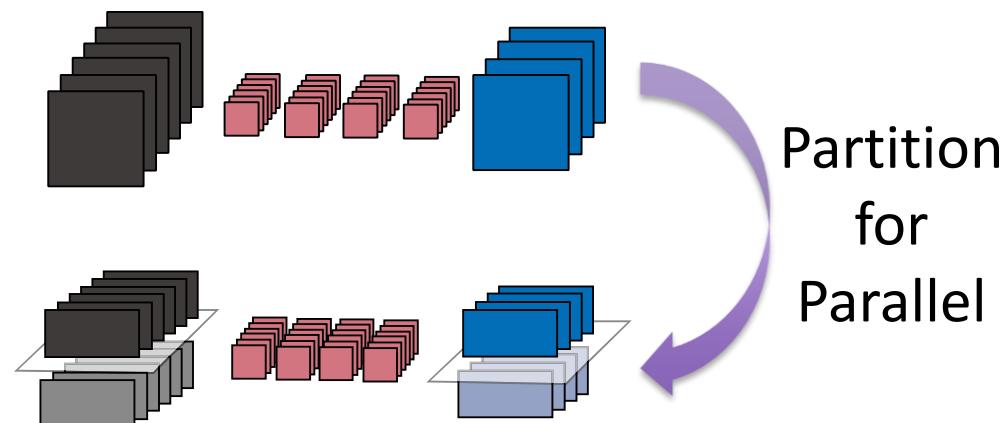


NN	Operation	Cycles	Note
AlexNet Layer 5	Comm_IFM	2,612	Performance Dominated by Comm_Weights Latency is 5,658
	Comm_Weights	5,658	
	Comm_OFM	368	
	Computation	3,326	

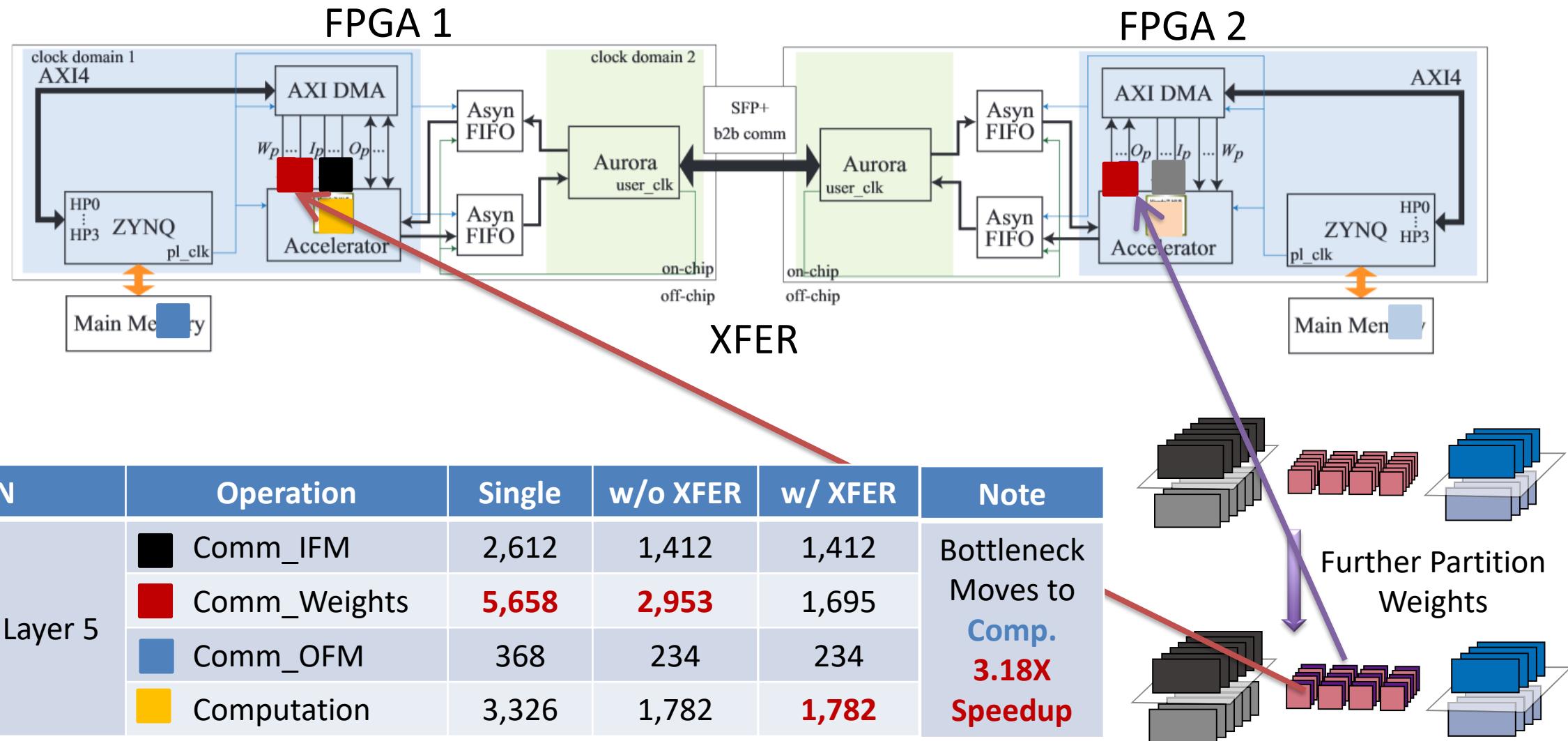
Double Computation Resource cannot Double Performance



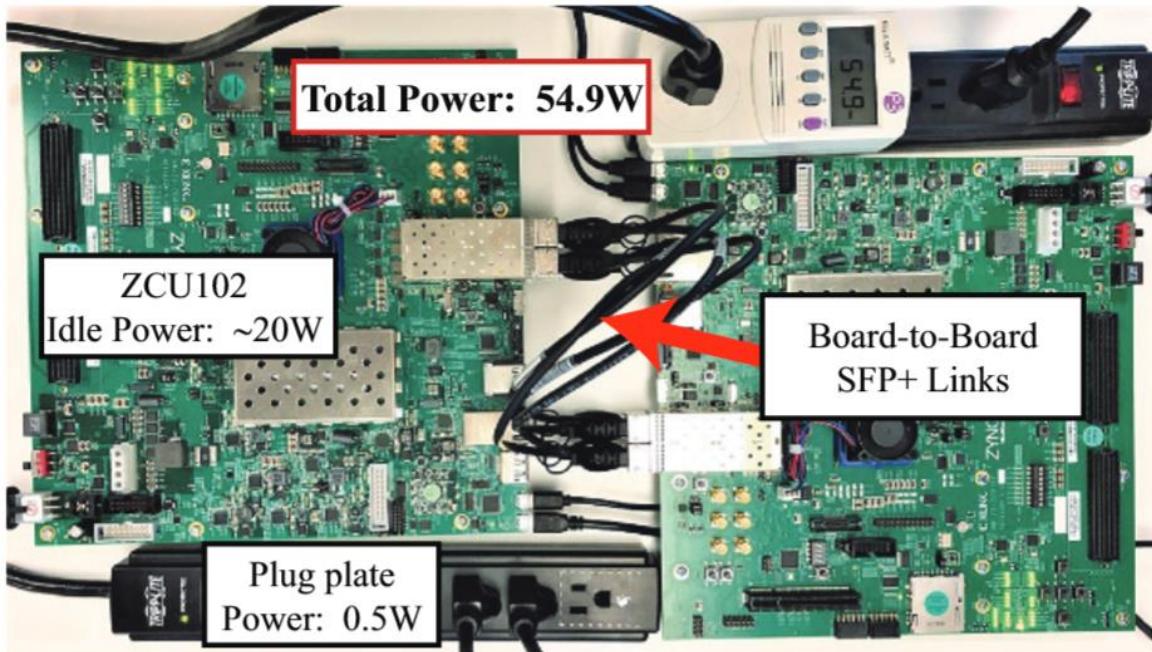
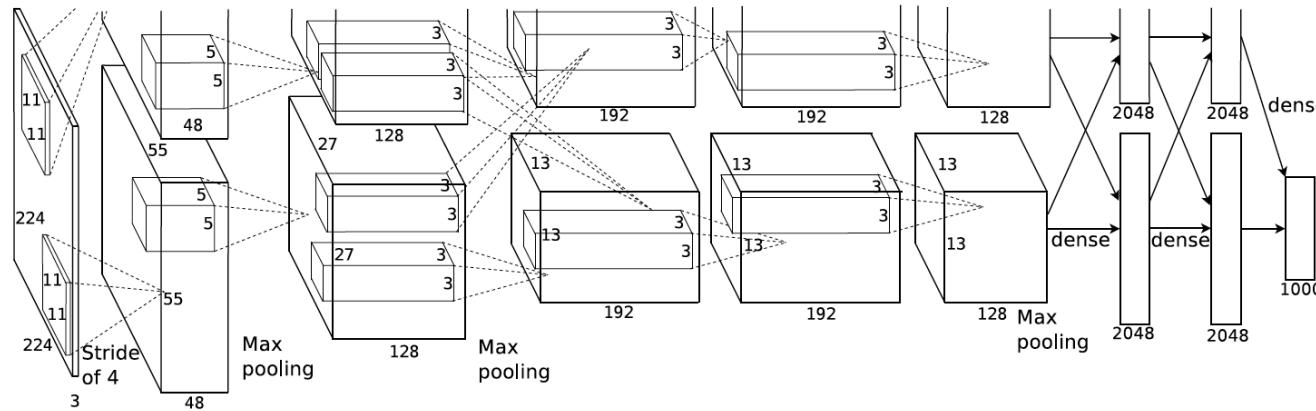
NN	Operation	Single	w/o XFER	Note
AlexNet Layer 5	Comm_IFM	2,612	1,412	1.91X Speedup
	Comm_Weights	5,658	2,953	
	Comm_OFM	368	234	
	Computation	3,326	1,782	



XFER: Achieve Super-Linear Speedup by Transferring Accesses from Off-Chip Memory to Inter-FPGA Links



Experimental Setting



Evaluation Neural Networks:
AlexNet on ImageNet

Evaluation Platform:
Xilinx ZUC102 FPGAs
connected by SFP+ Links

Experimental Results —— Achieving Super-Linear Performance

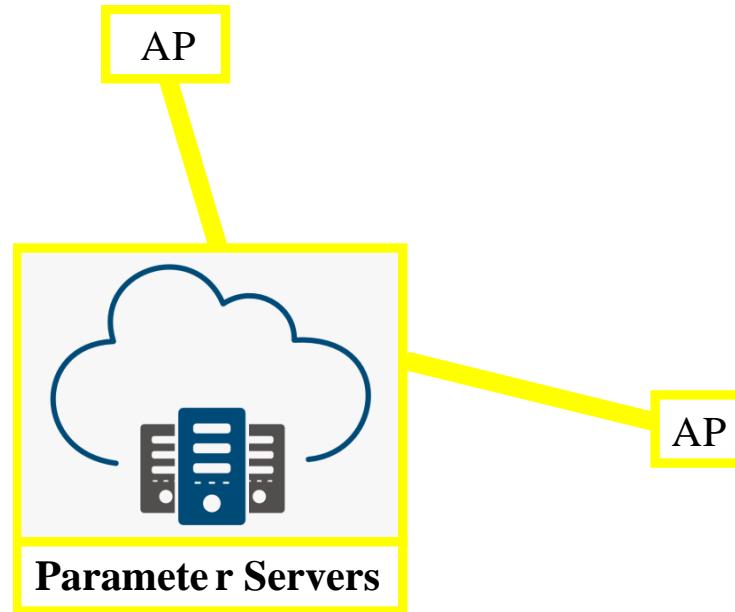
Comparison results of **XFER** with comparisons to **GPUs** and the **existing FPGA designs**

Design	mGPU		GPU		FPGA15		ISCA17		ISLPED16		XFER			
Precision	32bits float		32bits float		32bits float		32bits float		16bits fixed		32bits float		16bits fixed	
Device	Jetson TX2		Titan X		VX485T		VX485T		4×VX690t		2×ZCU102		2×ZCU102	
Freq (MHz)	1300MHz		1139MHz		100MHz		100MHz		150MHz		100MHz		200MHz	
Power (Watt)	16.00		162.00		18.61		-		126.00		52.40		54.40	
DSP Uti.	-		-		80%		80%		-		90.79%		55.87%	
BRAM Uti.	-		-		49.71%		43.25%		-		72.92%		92.43%	
Overall Perf.	Lat. <i>ms</i>	Thr. <i>GOPS</i>												
	11.1 - 13.2	110.75	5.1 - 6.4	235.55	21.62	69.09	60.13	85.47	30.6	128.8	10.13	149.54	2.27	679.04
E.-E. (GOPS/W)	6.88		1.45		3.71		-		1.02		2.85		12.48	

**Lowest
Latency
among all
competitors**

NASS: Secure Inference via NAS

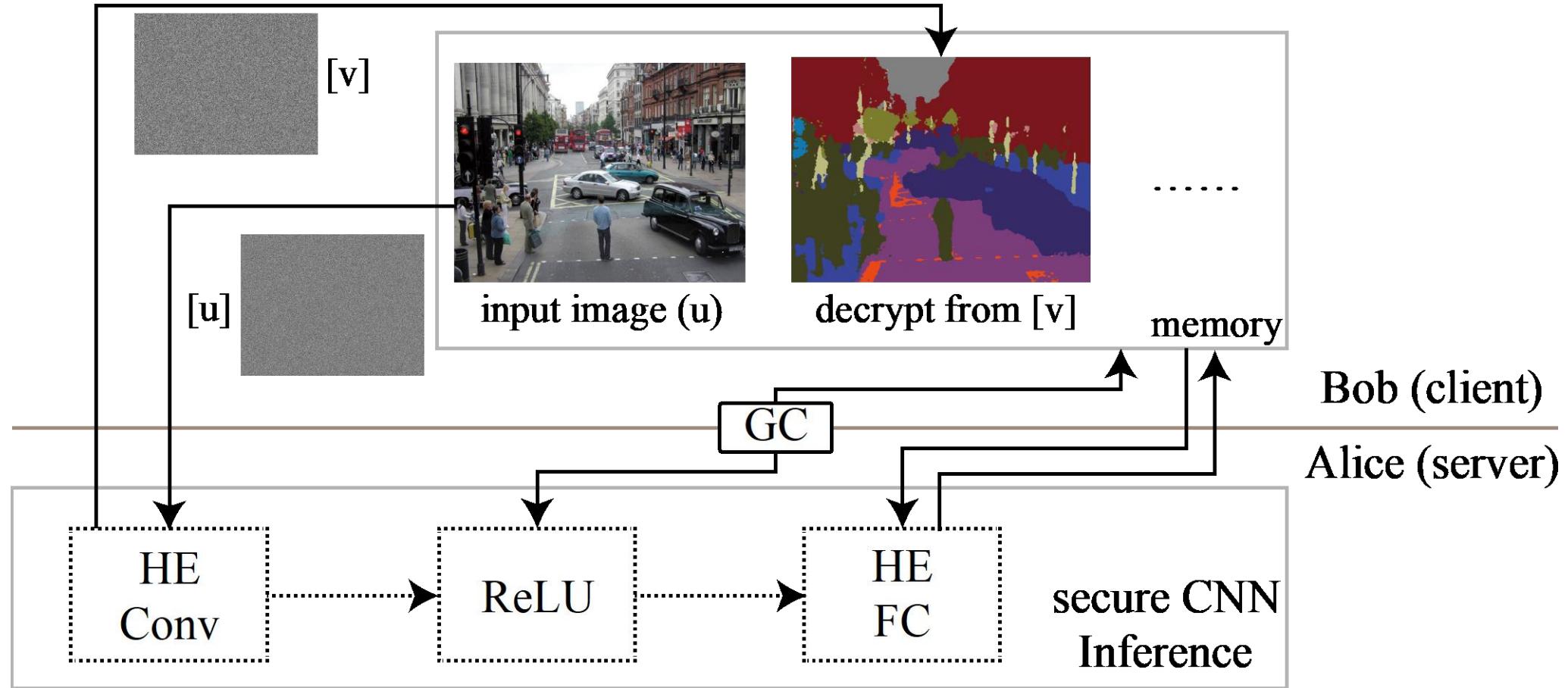
(ECAI'20)



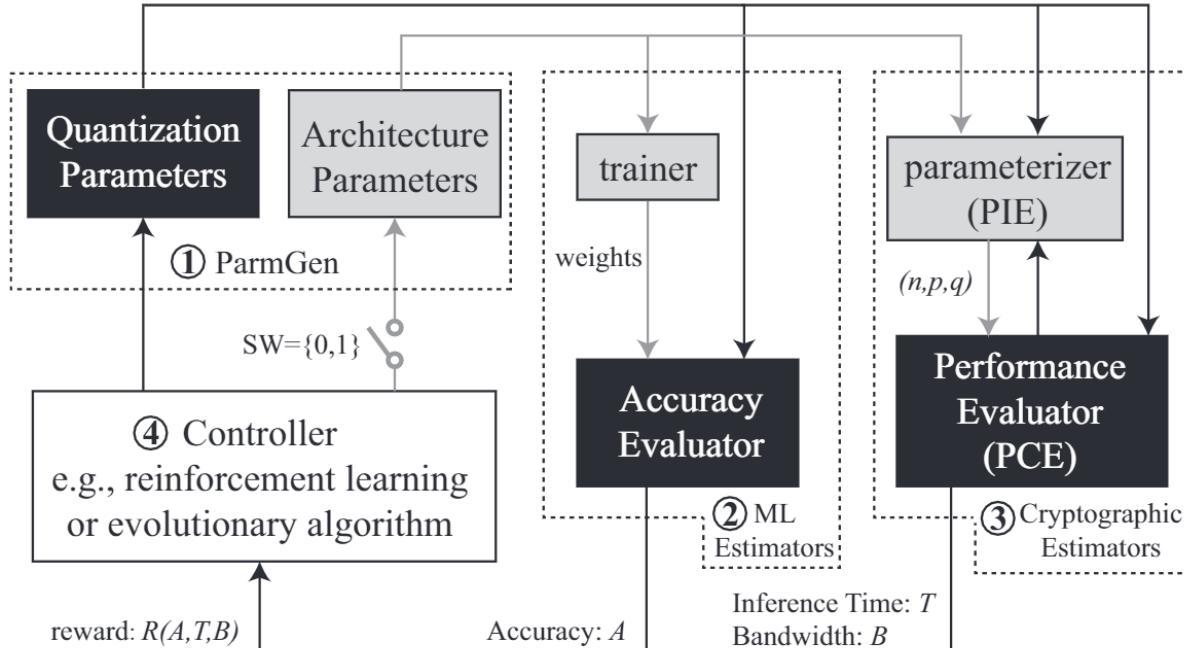
NASS: Identifying Secure Inference Architecture via NAS



Privacy and Security Problems: homomorphic encryption & garbled circuits



NASS: Framework and Results

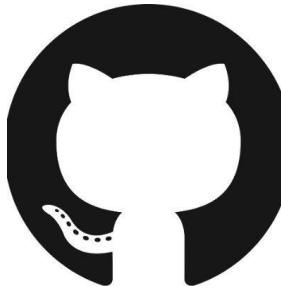


- Determination of hyper-parameters and quantization
- Performance Modeling

Gazelle			Best Searched by NASS		
Layer	Dimension	Quant.	Layer	Dimension	Quant.
CR	$(64 \times 3 \times 3)$	23	CR	$(24 \times 5 \times 3)$	(8, 8)
CR	$(64 \times 3 \times 3)$	23	CR	$(48 \times 3 \times 5)$	(6, 7)
PL	(2×2)	23	PL	(2×2)	(8, 8)
CR	$(64 \times 3 \times 3)$	23	CR	$(48 \times 5 \times 7)$	(7, 6)
CR	$(64 \times 3 \times 3)$	23	CR	$(36 \times 3 \times 3)$	(6, 5)
PL	(2×2)	23	PL	(2×2)	(8, 8)
CR	$(64 \times 3 \times 3)$	23	CR	$(24 \times 7 \times 1)$	(4, 6)
CR	$(64 \times 3 \times 3)$	23	FC	(1024×10)	(16, 16)
Accuracy: 81.6%			Accuracy: 84.6%		
Bandwidth: 1.815 GBytes			Bandwidth: 977 MB		
PAHE Time: 3.22 s			PAHE Time: 1.62 s		
GC Time: 13.2 s			GC Time: 6.38 s		
Total Time: 16.4 s			Total Time: 8.0 s		

- Improve accuracy by 3%
- Decrease 2X bandwidth requirement
- Decrease 2X computation time in server side

Open Source Projects



NAS + FPGA

<https://github.com/ND-SCL/NAQS>

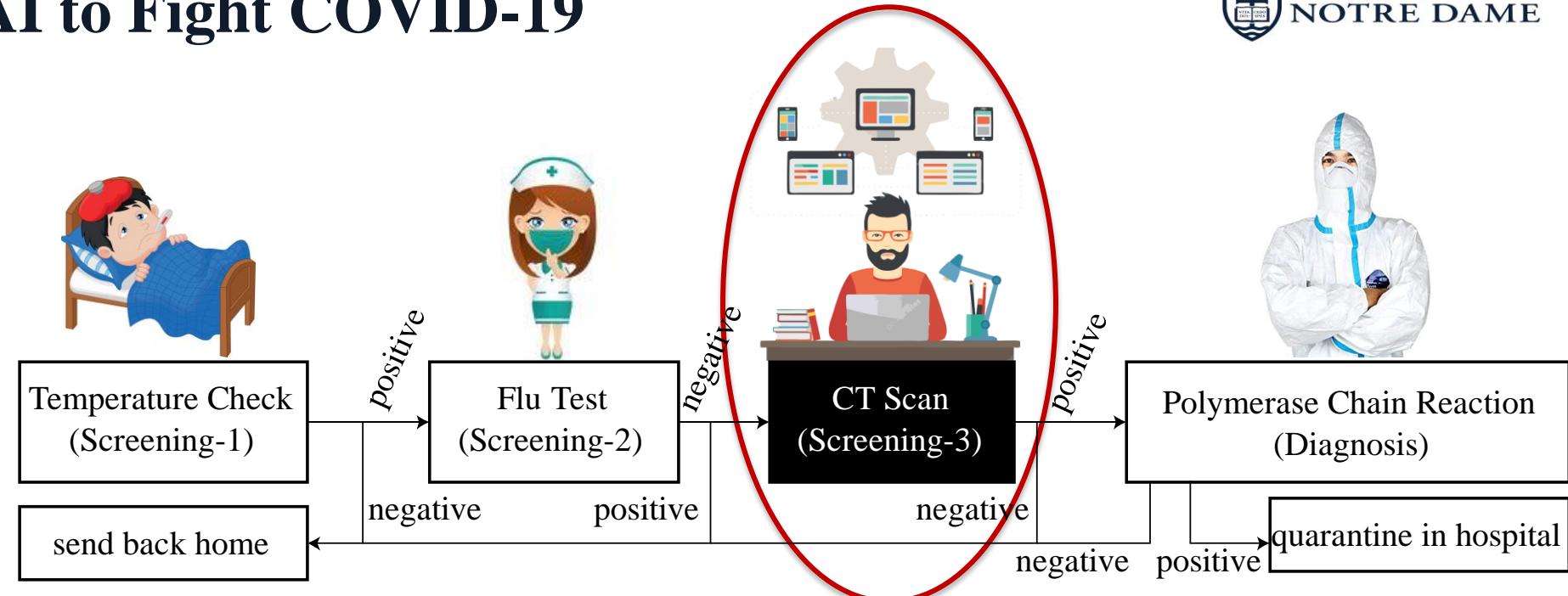
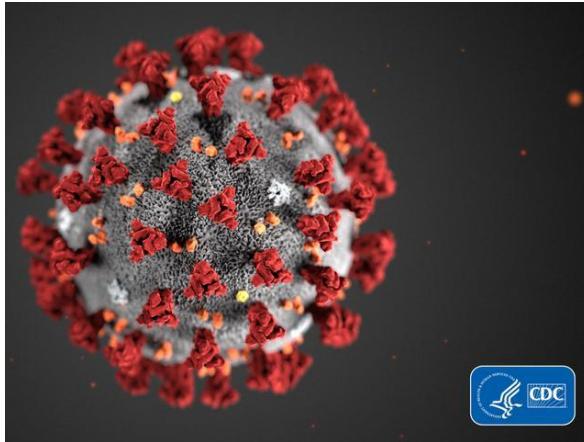
Multiple FPGAs

https://github.com/PITT-JZ-COOP/XFER_FPGA



Future Directions

Example: Equip AI to Fight COVID-19



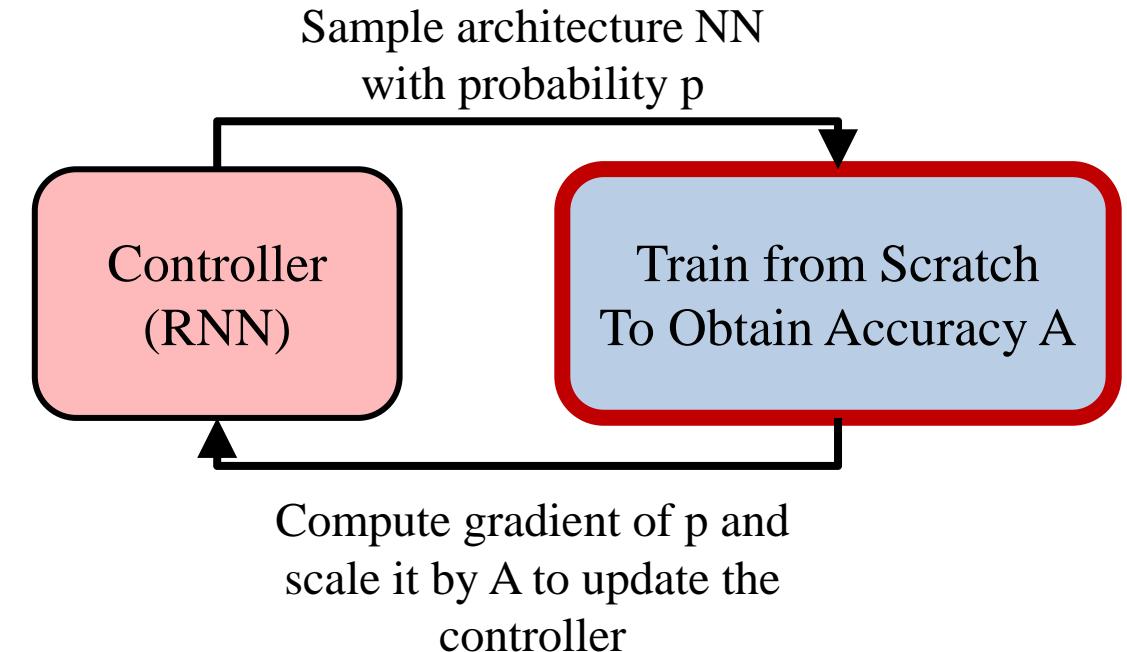
Challenge	Response
Shortage of rRT-PCR test kits	<u>Accurate screening</u>
Burden on radiologists in reading CT scan results	<u>AI judgement to reduce burden</u>
Days of deployment is intolerant	Plug-and-play in clinics within Hours

[ref] How a country serious about coronavirus does testing and quarantine. <https://www.youtube.com/watch?v=e3gCbkeARbY>. [Online; accessed 03/17/2020]

An Immediately Future Work

iNAS: From Hundreds GPU Hours to < 3

Approach for ImageNet	GPU Hours
eNASNet [1] (Google)	48,000
MnasNet [2] (Google)	40,000
SinglePathNAS [3] (Tsinghua)	312
FNAS [4] (ND+PITT, Ours)	267
FBNet [5] (FB+UCB)	216
ProxlessNAS [6] (MIT)	200
.....	
iNAS	< 3



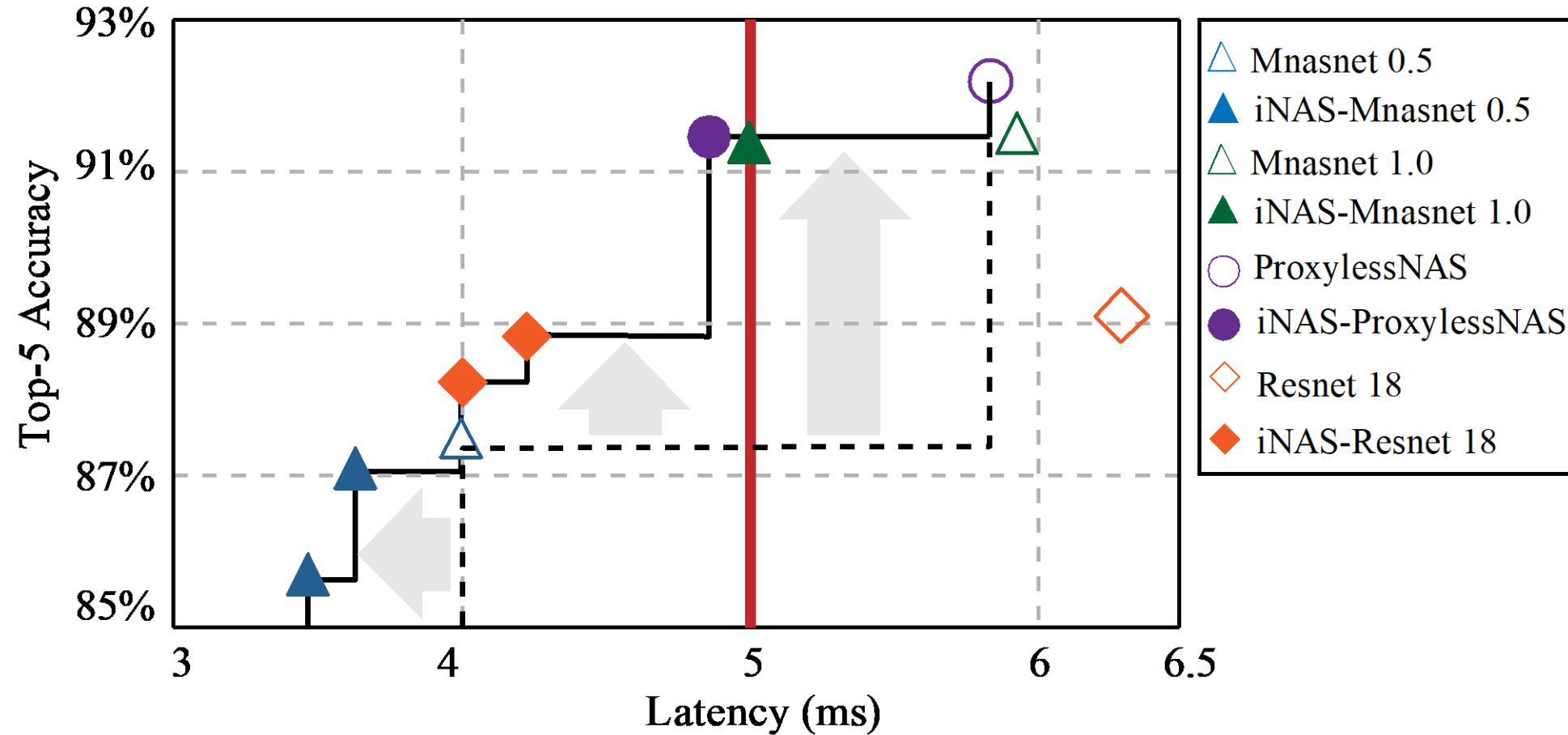
Experimental Results



Model	Type	Latency	Sat.	Param.	Param.	Top-1	Top-5	Top-1 Imp.	Top-5 Imp.	GPU Cost
AlexNet	manually	2.02	✓	61.1M	122.20MB	56.52%	79.07%	-	-	-
<i>MnasNet 0.5</i> *	<i>auto</i>	3.99	✓	2.22M	4.44MB	67.60%	87.50%	-	-	40,000H
SqueezeNet 1.0	manually	4.76	✓	1.25M	2.50MB	58.09%	80.42%	-	-	-
ProxylessNAS	auto	5.83	✗	4.08M	8.16MB	74.59%	92.20%	-	-	200H
MnasNet 1.0	auto	5.94	✗	4.38M	8.77MB	73.46%	91.51%	-	-	40,000H
Resnet18	manually	6.27	✗	11.69M	23.38MB	69.76%	89.08%	-	-	-
FBNet	auto	7.37	✗	5.57M	11.14MB	75.12%	92.39%	-	-	216H
iNAS-Resnet18(4ms)	auto	4.00	✓	10.99M	17.49MB	68.27%	88.21%	0.67%	0.71%	2H22M
iNAS-Resnet18	auto	4.22	✓	11.19M	17.90MB	69.14%	88.83%	1.54%	1.33%	2H01M
iNAS-ProxylessNAS	auto	4.86	✓	4.38M	8.31MB	73.39%	91.47%	5.79%	3.97%	2H37M
iNAS-Mnasnet 1.0	auto	4.99	✓	4.07M	6.56MB	73.24%	91.37%	5.64%	3.87%	1H50M

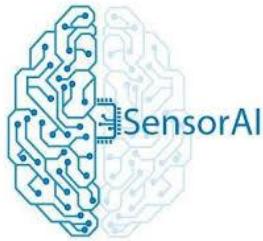
* : baseline

Push Forward Pareto Frontier

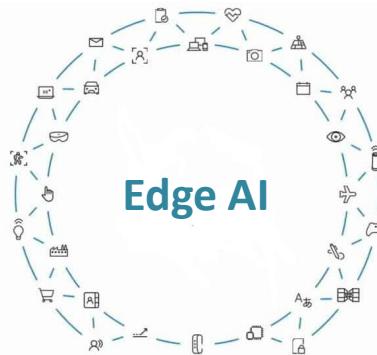


Short-Term Plan: 1-2 Years

Mid-Term Plan



Short-Term Plan



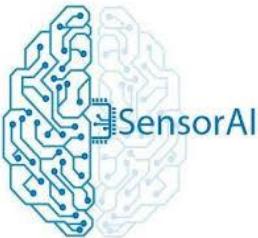
Long-Term Plan



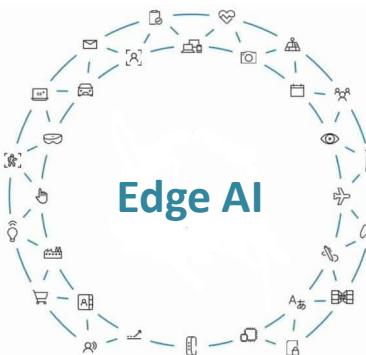
Short-Term Plan: 1-2 Years



Mid-Term Plan



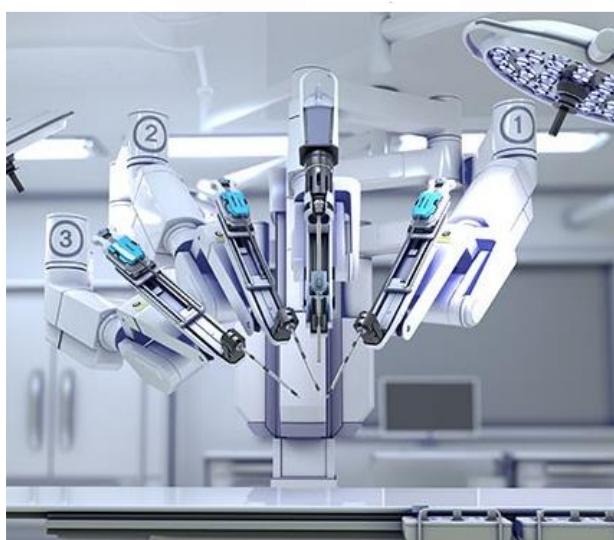
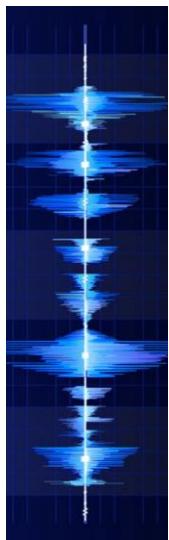
Short-Term Plan



Long-Term Plan



Algorithm Optimization



Robotics



Jacob



Sudarshan



Umamaheswara

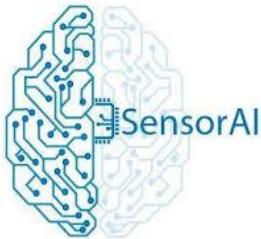


EDGE CORTIX

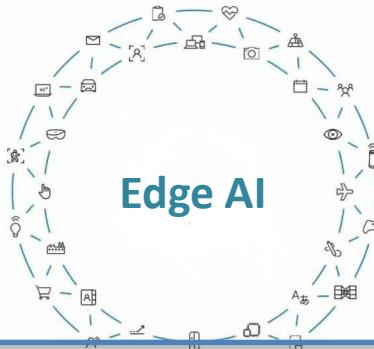
Mid-Term Plan: 3-5 Years



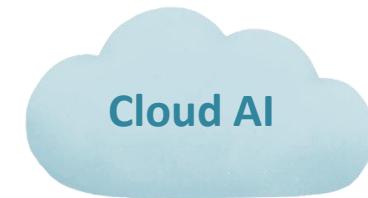
Mid-Term Plan



Short-Term Plan



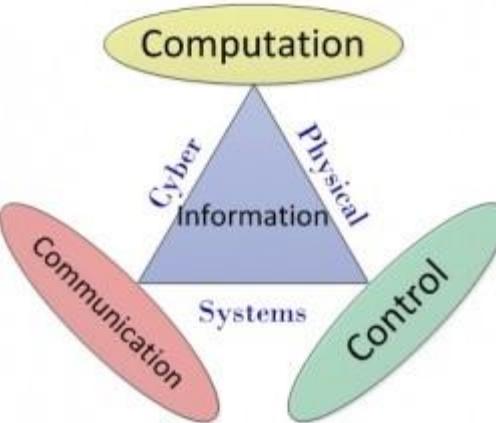
Long-Term Plan



Sensor Assistant Systems

CPS

Weiwen Jiang



Danling

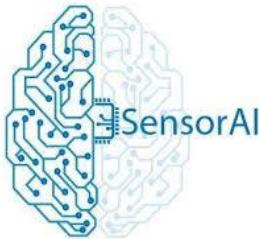


Umamaheswara

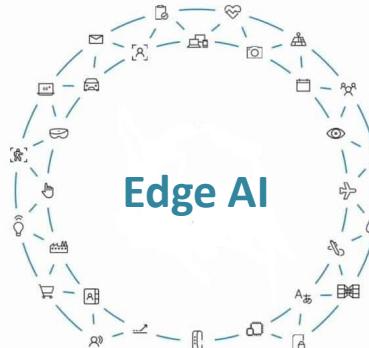


Long-Term Plan: 5+ Years

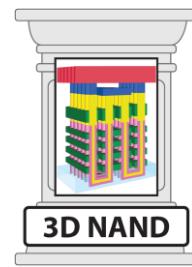
Mid-Term Plan



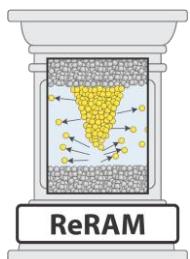
Short-Term Plan



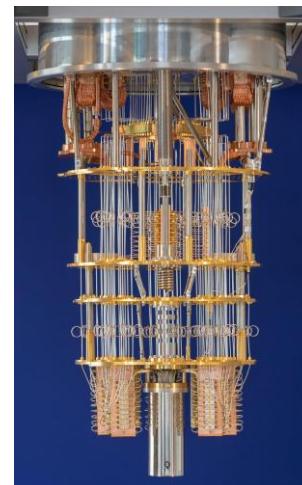
Long-Term Plan



Emerging Memory



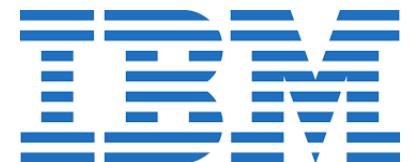
Emerging Computing



Sudarshan



Umamaheswara



Funded Projects



Awarded (Sole PI):

\$100K from *Edgecortix Inc* via NSF I/UCRC, (09/19-09/20)

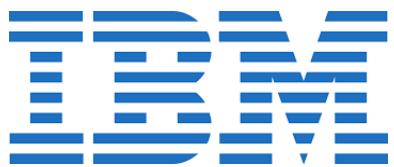
Ranked No.1 of I/UCRC ASIC projects voting in 2020



Awarded Last Week (Co-PI):

\$75K from *Facebook Research Award*, (05/20-05/21)

8 out of 160 proposals



Awarded (Co-PI):

Accessing to IBM Quantum Computer with 53 qubits

Teaching

- ***Experience***
 - **Univeristy of Notre Dame**
 - Logic Design and Sequential Circuits, CSE20221 (TA for Dr. Jay Brockman) Jan. 2020 – May 2020
 - Machine Learning for Embedded Systems, CSE60685 (Instructor) Jan. 2020 – May 2020
 - **Chongqing Univeristy**
 - High-Performance Parallel Computing (TA for Dr. Edwin Sha) Sep. 2014 – Jun. 2015
- ***Interested courses to teach in NDSU***
 - ECE 111. Introduction to Electrical and Computer Engineering.
 - ECE 173. Introduction to Computing.
 - ECE 275. Digital Design.
 - ECE 374 Computer Organization
 - ECE 474 Computer Architecture
 - ECE 376 Embedded Systems
- ***New Courses***
 - AutoML for embedded systems
 - Introduction to Quantum Computing



Selected Publication

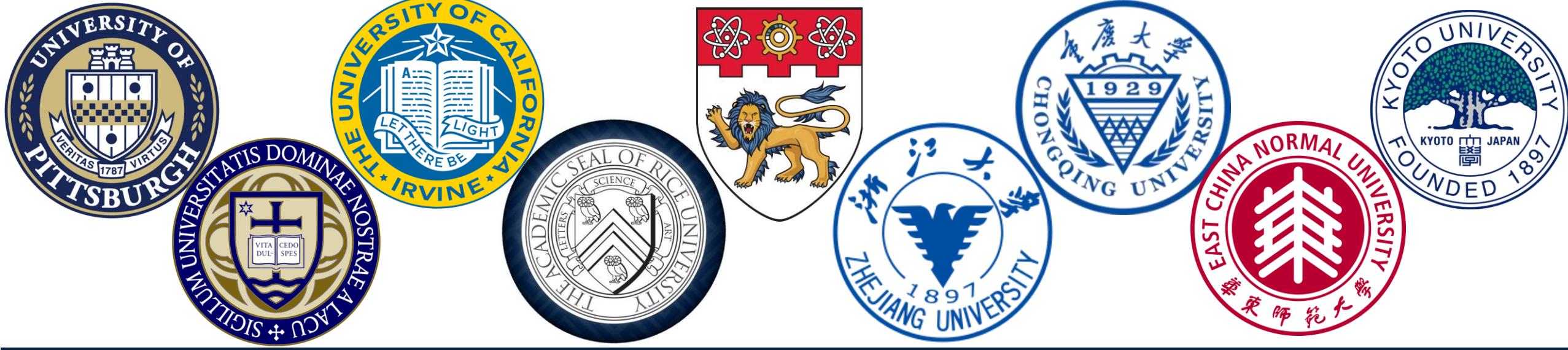


- [1] **W. Jiang, L. Yang, E. H.-M. Sha, Q. Zhuge, S. Gu, S. Dasgupta, Y. Shi and J. Hu**, Hardware/Software Co-Exploration of Neural Architectures, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Accepted, 2020.
- [2] **L. Yang, Z. Yan, M. Li, H. Kwon, L. Lai, T. Krishana, V. Chandra, W. Jiang, and Y. Shi**, Co-Exploration of Neural Architectures and Heterogeneous ASIC Accelerator Designs Targeting Multiple Tasks, *Design Automation Conference (DAC)*, 2020.
- [3] **B. Song, W. Jiang, Q. Lu, Y. Shi and T. Sato**, NASS: Optimizing Secure Inference via Neural Architecture Search, *Proc. European Conference on Artificial Intelligence (ECAI), Santiago de Compostela, June*. 2020.
- [4] **L. Yang, W. Jiang, W. Liu, E. H.-M. Sha, Y. Shi and J. Hu**, Co-Exploring Neural Architecture and Network-on-Chip Design for Real-Time Artificial Intelligence, *Proc. Asia and South Pacific Design Automation Conference (ASP-DAC), Beijing, Jan.* 2020. (**BEST PAPER NOMINATION**)
- [5] **W. Jiang, E. H.-M. Sha, X. Zhang, L. Yang, Q. Zhuge, Y. Shi and J. Hu**, Achieving Super-Linear Speedup across Multi-FPGA for Real-Time DNN Inference, *International Conference on Hardware/Software Co-design and System Synthesis CODE+ISSS*), also appears at *ACM Transactions on Embedded Computing Systems (TECS)*, NYC, New York, USA, Oct. 2019. (**BEST PAPER NOMINATION**)
- [6] **W. Jiang, B. Xie, C-C Liu and Y. Shi**, Integrating Memristors and CMOS for Better AI, *Nature Electronics (News and Views)*, Sep. 2019
- [7] **W. Jiang, X. Zhang, E. H.-M. Sha, L. Yang, Q. Zhuge, Y. Shi, and J. Hu**, Accuracy vs. Efficiency: Achieving Both through FPGA-Implementation Aware Neural Architecture Search, *Design Automation Conference (DAC)*, 2019 (**BEST PAPER NOMINATION**)

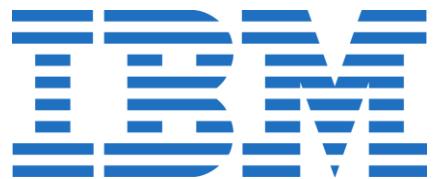
Selected Publication



- [8] **W. Jiang**, E. H.-M. Sha, Q. Zhuge, L. Yang, X. Chen, and J. Hu, Heterogeneous FPGA-based Cost-Optimal Design for Timing-Constrained CNNs, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Torino, Italy, Oct. 2018.
- [9] **W. Jiang**, E. H.-M. Sha, Q. Zhuge, L. Yang, X. Chen, and J. Hu, On the Design of Time-Constrained and Buffer-Optimal Self-Timed Pipelines, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Accepted, 2018.
- [10] E. H.-M. Sha, **W. Jiang**, H. Dong, Z. Ma, R. Zhang, X. Chen and Q. Zhuge, Towards the Design of Efficient and Consistent Index Structure with Minimal Write Activities for Non-Volatile Memory, *IEEE Transactions on Computers (TC)*, 67(3), 432-448, 2018.
- [11] **W. Jiang**, E. H.-M. Sha, Q. Zhuge, L. Yang, H. Dong and X. Chen, On the Design of Minimal-Cost Pipeline Systems Satisfying Hard/Soft Real-Time Constraints *IEEE International Conference on Computer Design (ICCD2017@BOSTON) & IEEE Transactions on Emerging Topics in Computing (TETC)*, Jan. 2018. (**BEST PAPER AWARD**)
- [12] **W. Jiang**, E. H.-M. Sha, Q. Zhuge, H. Dong and X. Chen, Optimal Functional Unit Assignment and Voltage Selection for Pipelined MPSoC with Guaranteed Probability on Time Performance, *Proc. Languages, Compilers, and Tools for Embedded Systems (LCTES)*, Barcelona, Spain, Jun. 2017.
- [13] **W. Jiang**, E. H.-M. Sha, X. Chen, L. Yang, L. Zhou and Q. Zhuge, Optimal Functional-Unit Assignment for Heterogeneous Systems under Timing Constraint, *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 28(9), 2567-2580, 2017.
- [14] **W. Jiang**, E. H.-M. Sha, Q. Zhuge and X. Chen, Optimal Functional-Unit Assignment and Buffer Placement for Probabilistic Pipelines, *Proc. International Conference on Hardware/Software Co-design and System Synthesis (CODES+ISSS)*, Pittsburgh, PA, USA, Oct. 2016.
-



Thank You!



facebook

