Hardware/Software Co-Exploration of Neural Architectures

Weiwen Jiang, Lei Yang, Edwin Hsing-Mean Sha, Qingfeng Zhuge, Shouzhen Gu, Sakyasingha Dasgupta, Yiyu Shi, Jingtong Hu

wjiang8@gmu.edu | https://jqub.ece.gmu.edu **(Slides Available at Here!)**

Speaker

Weiwen Jiang
Assistant Professor
George Mason University

Lei Yang
Assistant Professor
Univ. of New Mexico

Yiyu Shi
Professor
Univ. of Notre Dame

Jingtong Hu
Associate Professor
Univ. of Pittsburgh

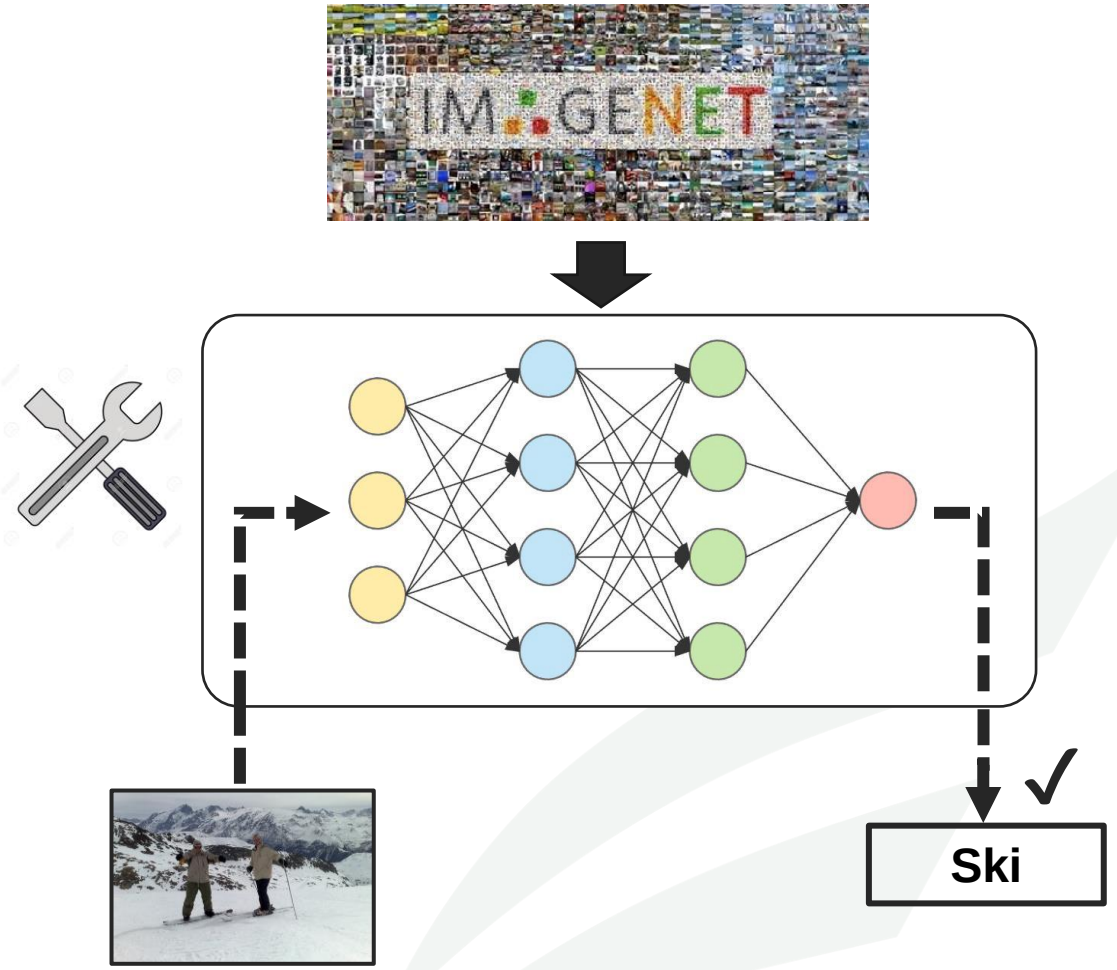
Bottleneck in Applying ML for Specific Applications

- **Manually Design Neural Network:**
  - Requires expertise from **different domains**
  - **Collaboration** of these experts is difficult
  - Large **human labor**
  - **Long launch time**…

Computer Vision Experts

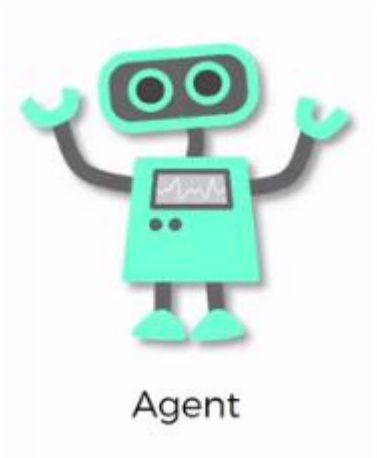
Computer Scientist

Data Scientist



AutoML is a Promising Solution

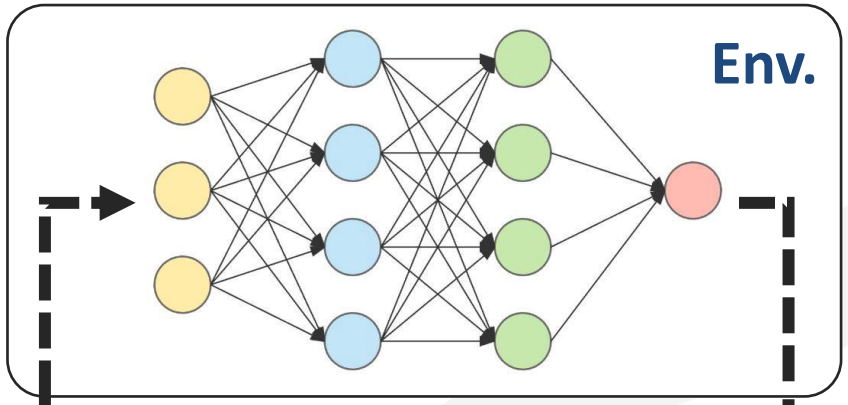
- **Automated Design Neural Network:**
  - ✓ Replace exerts by using **automated optimization approaches**
  - ✓ **Release the labor** from experts
  - ✓ **Short launch time**



Action

e.g. Reinforcement Learning

Reward



Ski

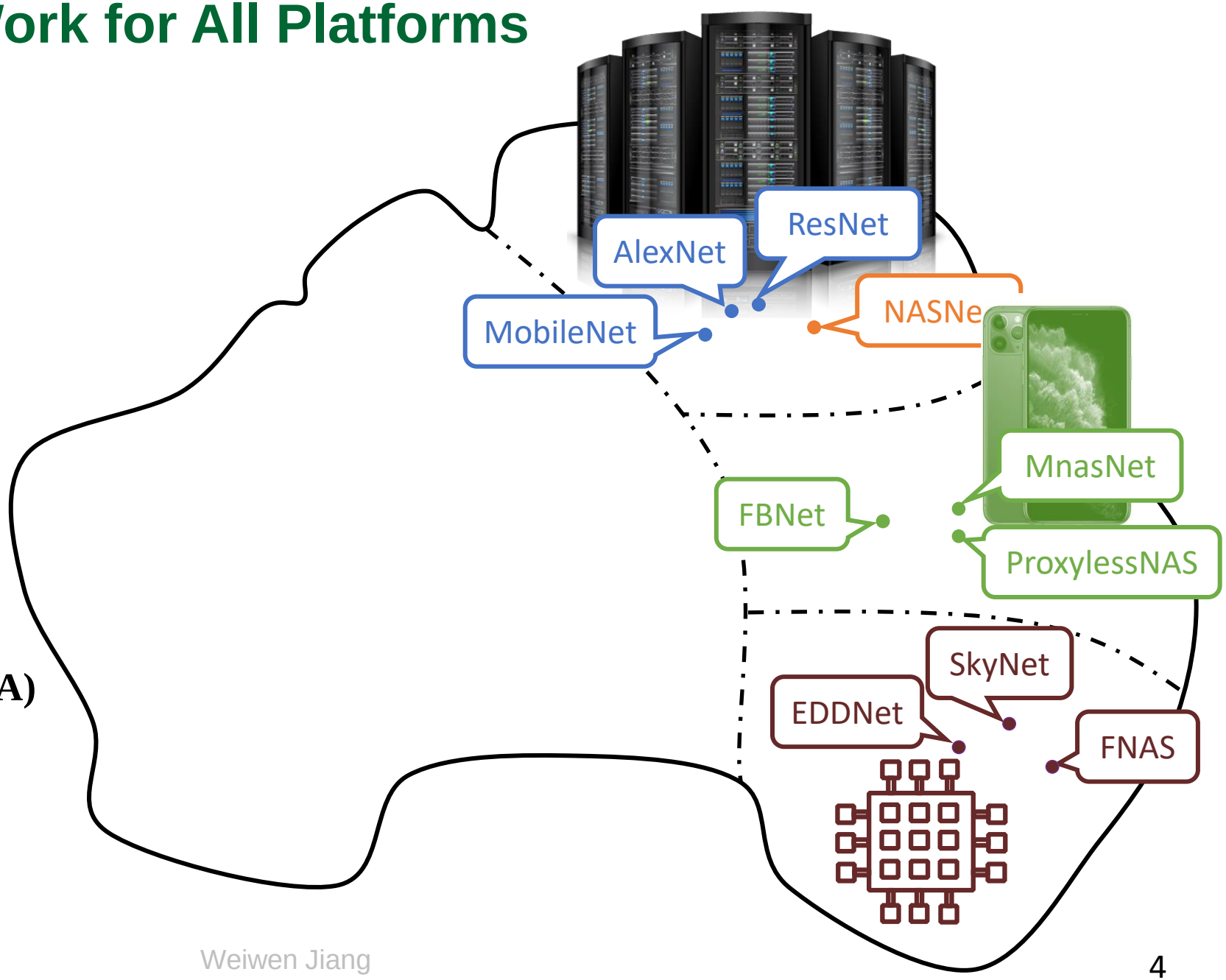
Objective:

Maximize ML Performance

(e.g., Prediction Accuracy)

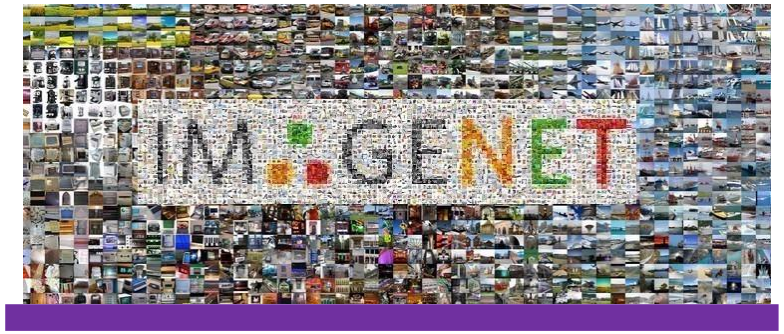
One Network Cannot Work for All Platforms

- **Cloud / Server**
  - Unlimited Resource
  - Maximizing Accuracy
  - AlexNet, VGGNet, ResNet, …
- **Mobile Phones**
  - Fixed Hardware
  - Accuracy v.s. Latency
  - MnasNet, ProxylessNAS, …
- **Hardware Accelerators (e.g., FPGA)**
  - Hardware Design Flexibility
  - Accuracy, Latency, and Energy
  - FNAS, SkyNet, EDDNet…



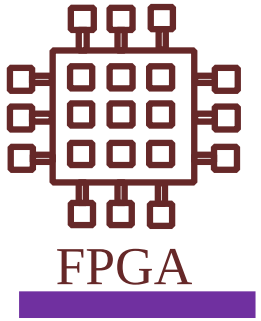
Problem: Datasets/Applications, Hardware, and Neural Networks

Datasets / Applications

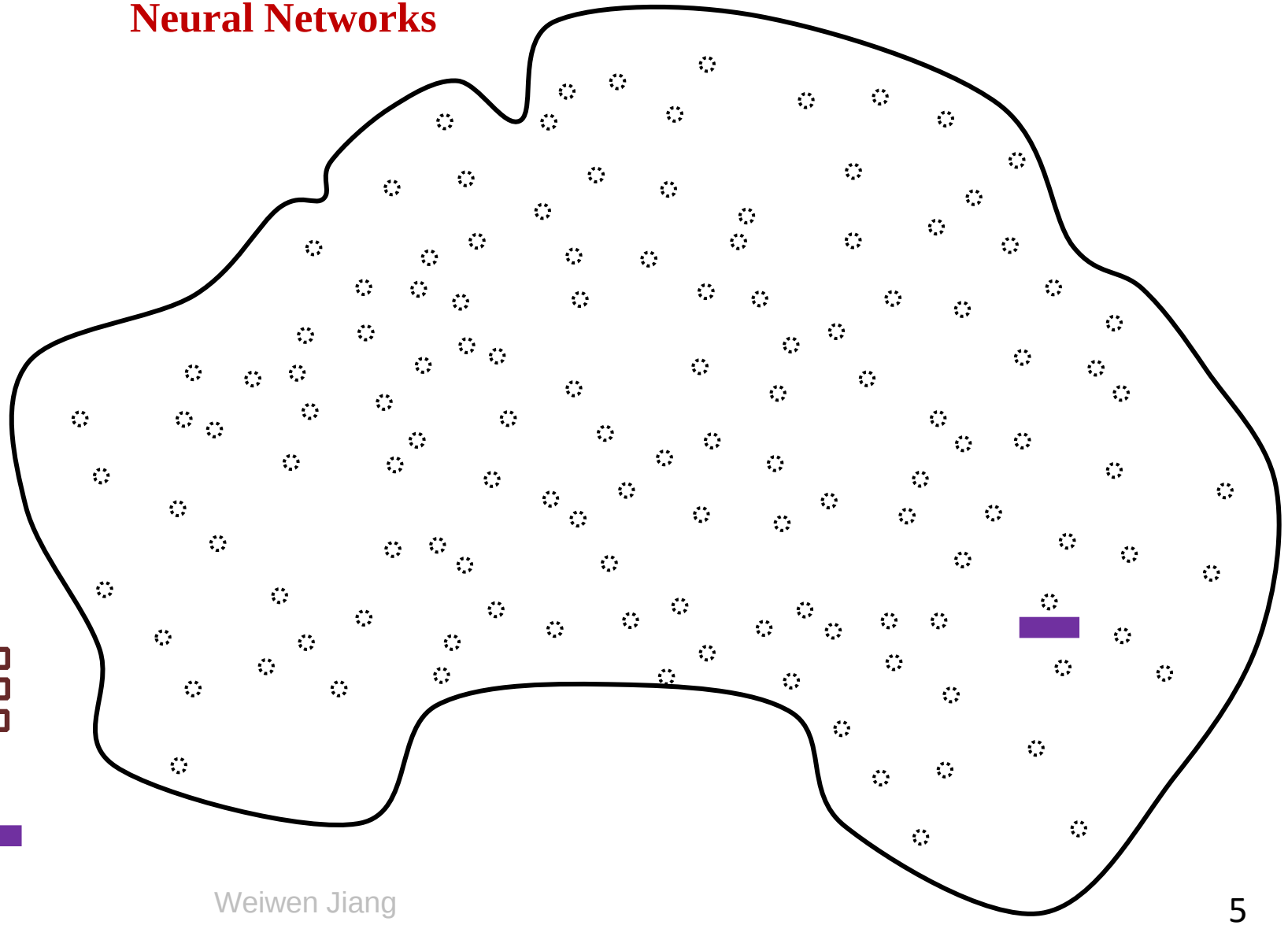


CITYSCAPES DATASET

Hardware Platforms



Neural Networks

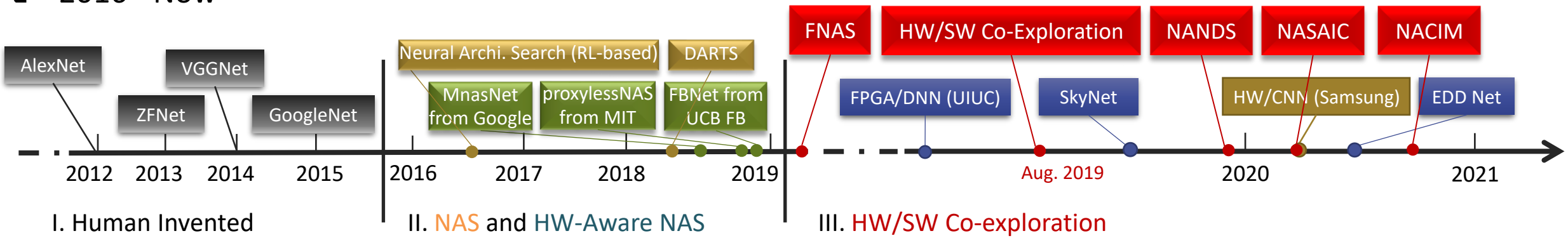


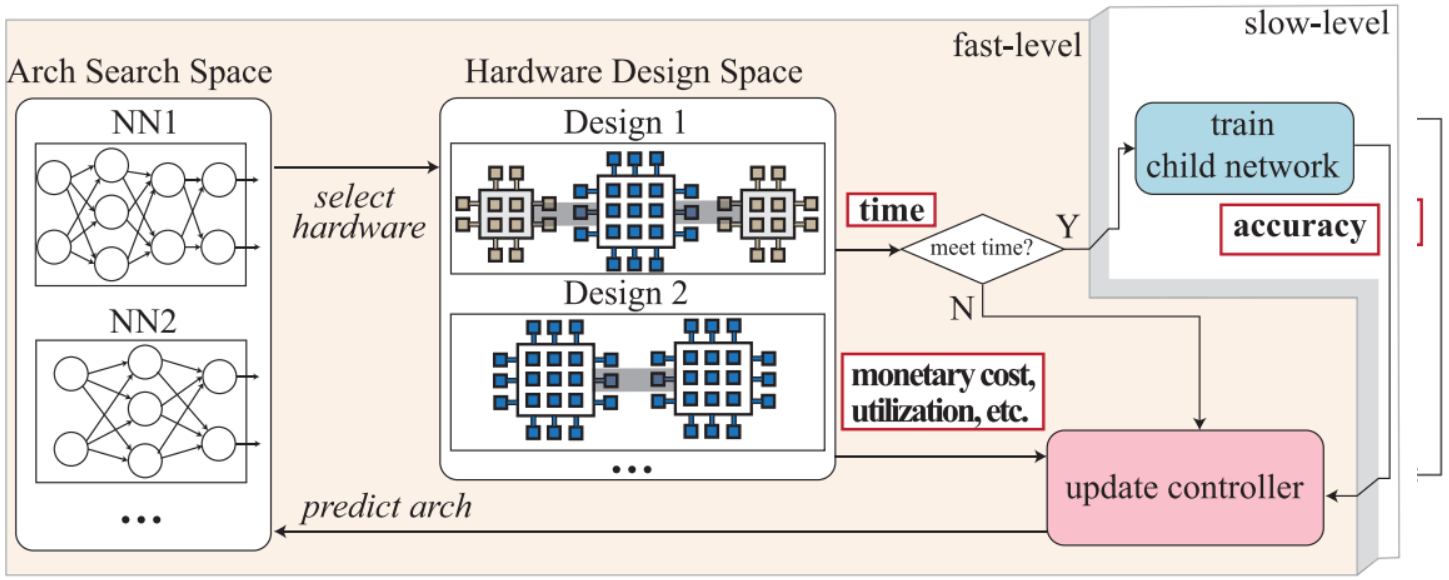
Outline

- Background
- **A Quick Overview of The Road From Manual Design to AutoML**
- HW/SW Co-Exploration Framework
  - Motivation
  - Framework Overview and Details
  - Results
- Follow-up Works and Conclusion

HW/SW Co-Exploration of Neural Architectures

- 2016 - Now





Neural Architecture Search (NAS) **HW-aware NAS** **HW/SW Co-Exploration of Neural Architectures**

Outline

- Background
- A Quick Overview of The Road From Manual Design to AutoML
- **HW/SW Co-Exploration Framework**
  - Motivation
  - Framework Overview and Details
  - Results
- Follow-up Works and Conclusion

Motivation

TABLE I
ON **CIFAR-10** AND **XILINX XC7Z015 FPGA**: COMPARISONS OF THREE NEURAL ARCHITECTURE AND HARDWARE DESIGN PAIRS IN ACCURACY, THROUGHPUT, AND ENERGY EFFICIENCY (E.-E): A) OPTIMAL ARCHITECTURE ON A FIXED HARDWARE IMPLEMENTATION THROUGH HARDWARE-AWARE NAS; B) THE SAME ARCHITECTURE BUT WITH FURTHER FPGA OPTIMIZATION; AND C) A JOINTLY OPTIMIZED NEURAL ARCHITECTURE AND FPGA IMPLEMENTATION THROUGH OUR CO-EXPLORATION.

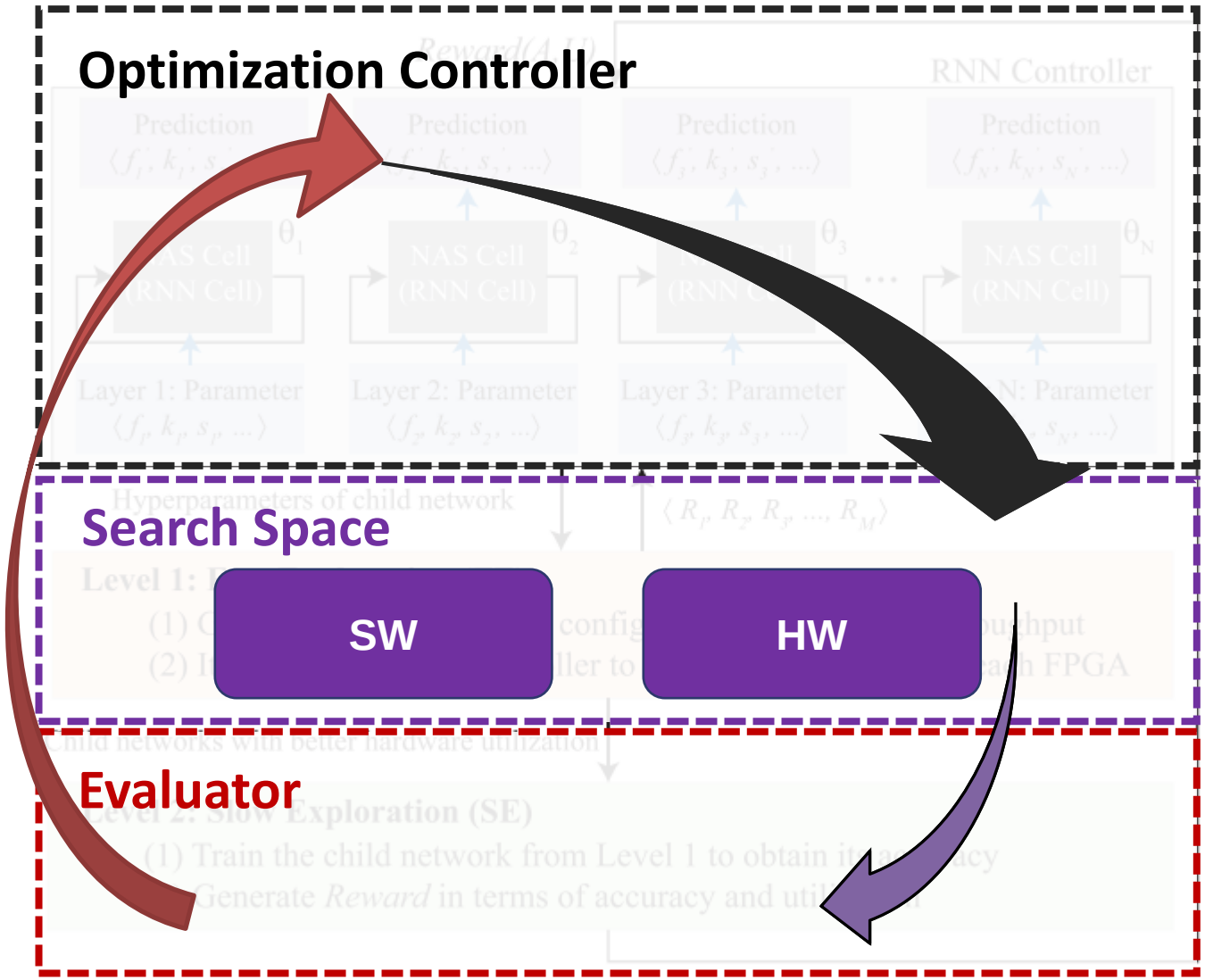
ID	Approach	Accuracy	Throughput (FPS)	
A	Hardware-Aware NAS	84.53%	16.2	
B	Sequential Optimization	84.53%	29.7	
C	**Co-Exploration**	**85.19%**	**35.5**	**1.91**

How to Co-Explore?

- Hardware-Aware NAS (Fixed HW)
  - Find network ***N*** under fixed HW ***H***
- Sequential Optimization:
  - Find network ***N*** under fixed HW ***H***
  - Optimize HW ***H*** for ***N***
- Co-Exploration:
  - Optimize ***N*** and ***H*** in one loop
  - **0.66%** Accuracy Gain
  - **2.19X** and **1.20X** Throughput Gain
  - **2.27X** and **1.40X** Energy Efficiency

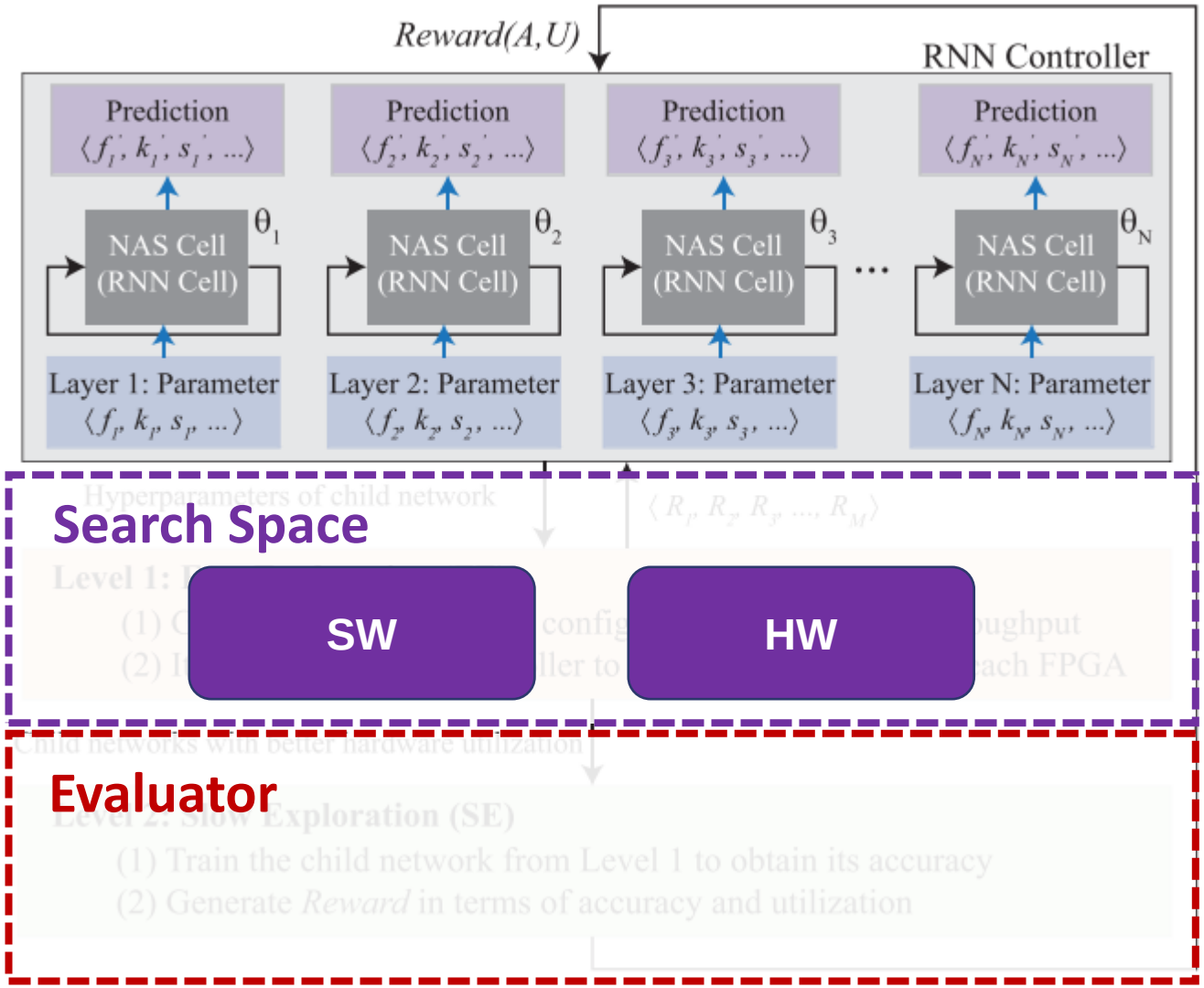
HW and Network Optimization are Coupled with Each Other

Framework: Optimizing Network Architecture and HW Design in One Loop



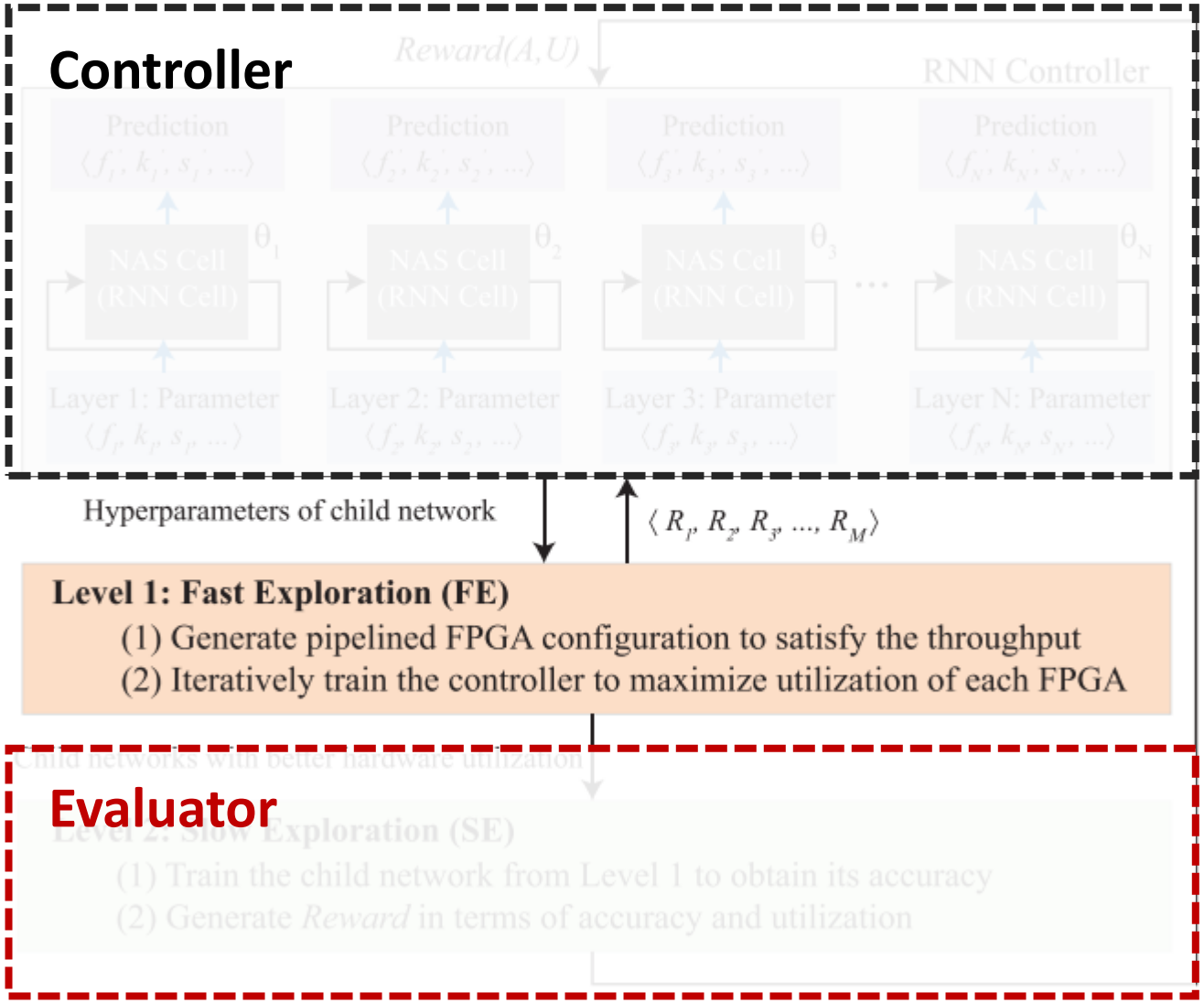
- **Controller** iteratively selects solution from the **search space** for **evaluation**
- **Controller** is evolved using the **evaluation results** from previous iteration.

Framework: Optimizing Network Architecture and HW Design in One Loop

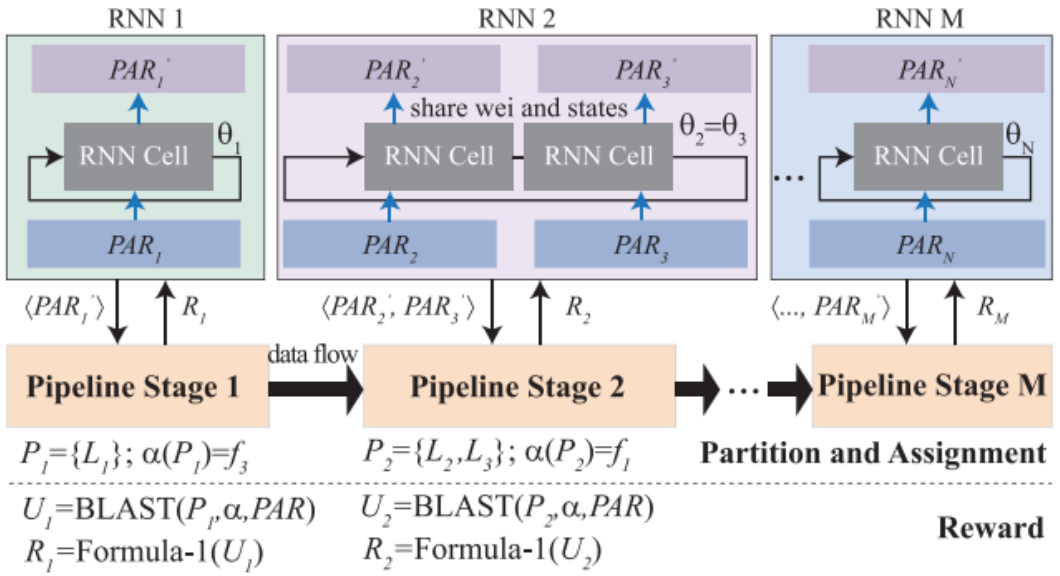


- **Controller**: Reinforcement learning based optimizer

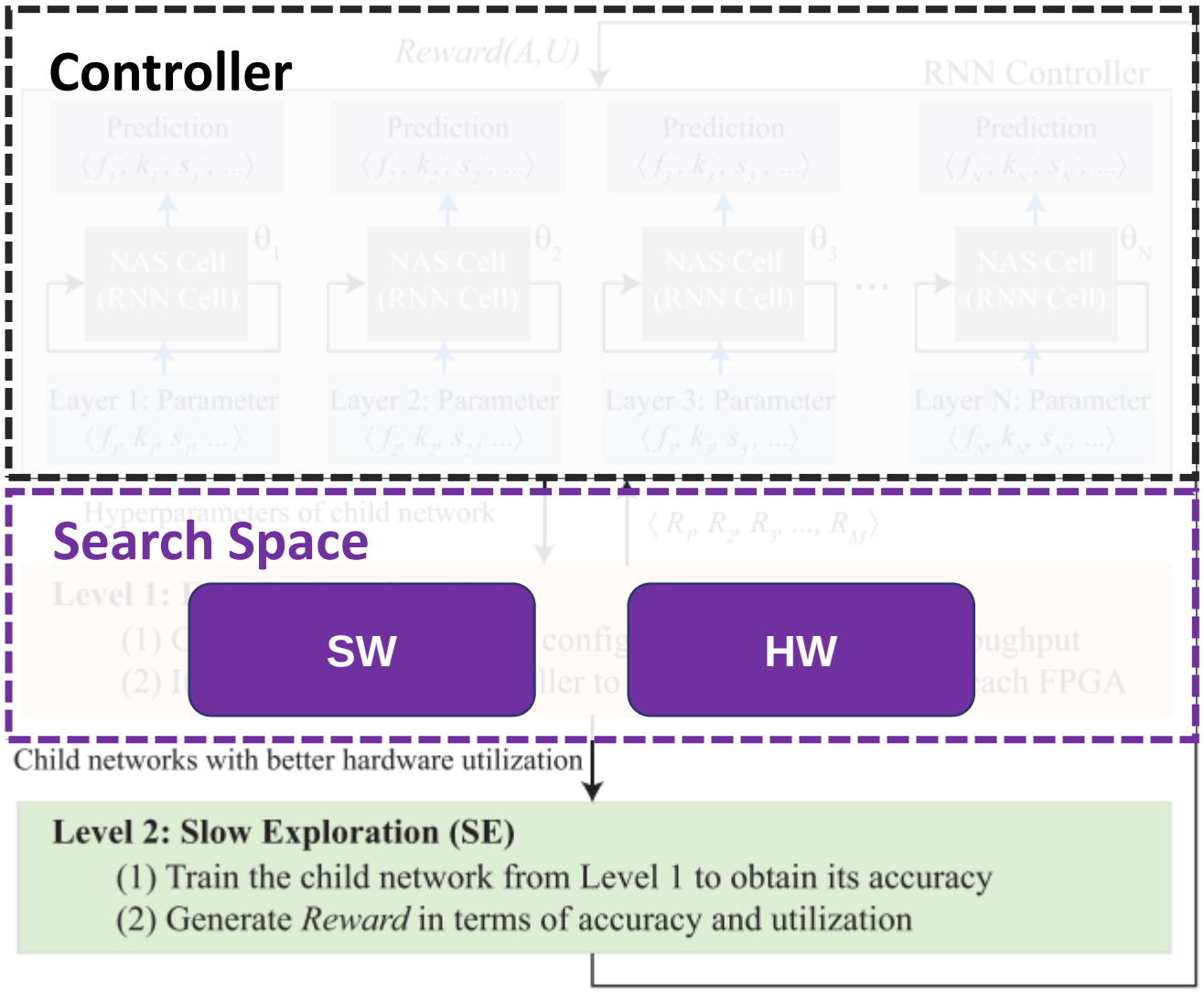
Framework: Optimizing Network Architecture and HW Design in One Loop



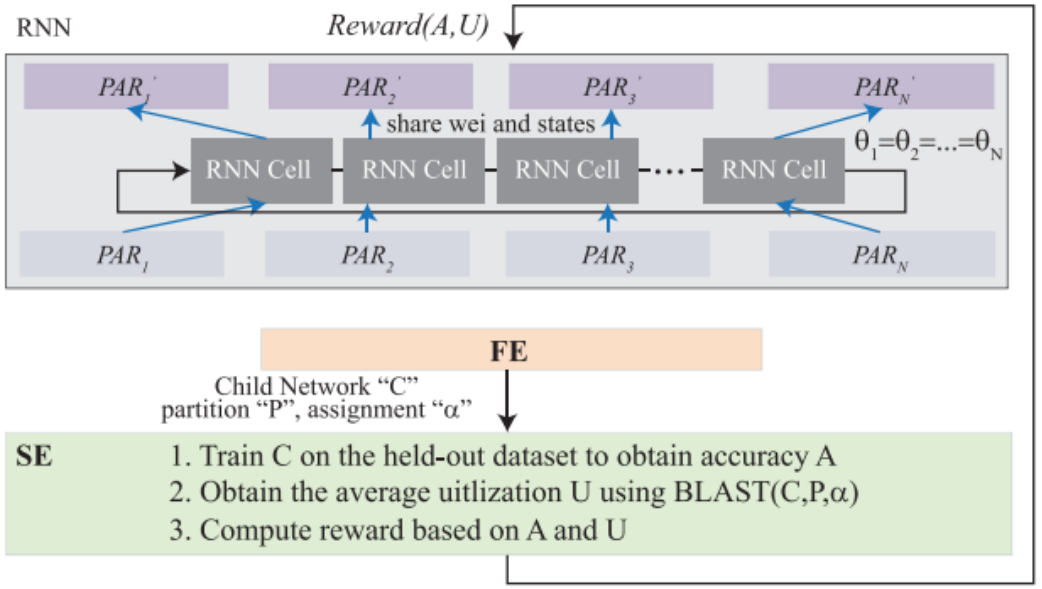
- **Search Space Exploration**:
  - Network Hyperparameters (SW)
  - Partition and Assignment (HW)



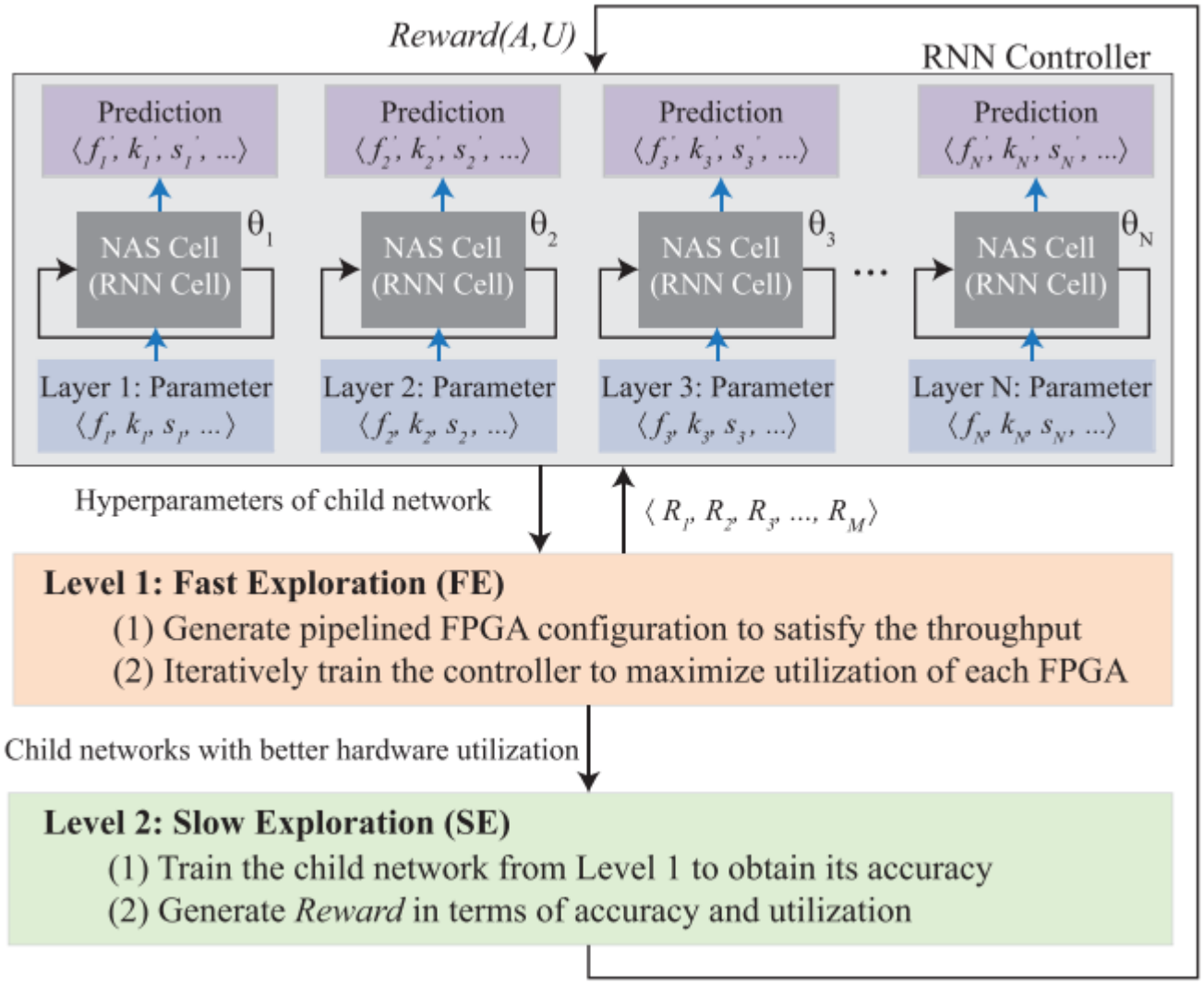
Framework: Optimizing Network Architecture and HW Design in One Loop



- **Evaluator**:
  - Obtain Accuracy **A** and HW Utilization ***U***
  - Generate Reward with ***A*** & ***U***

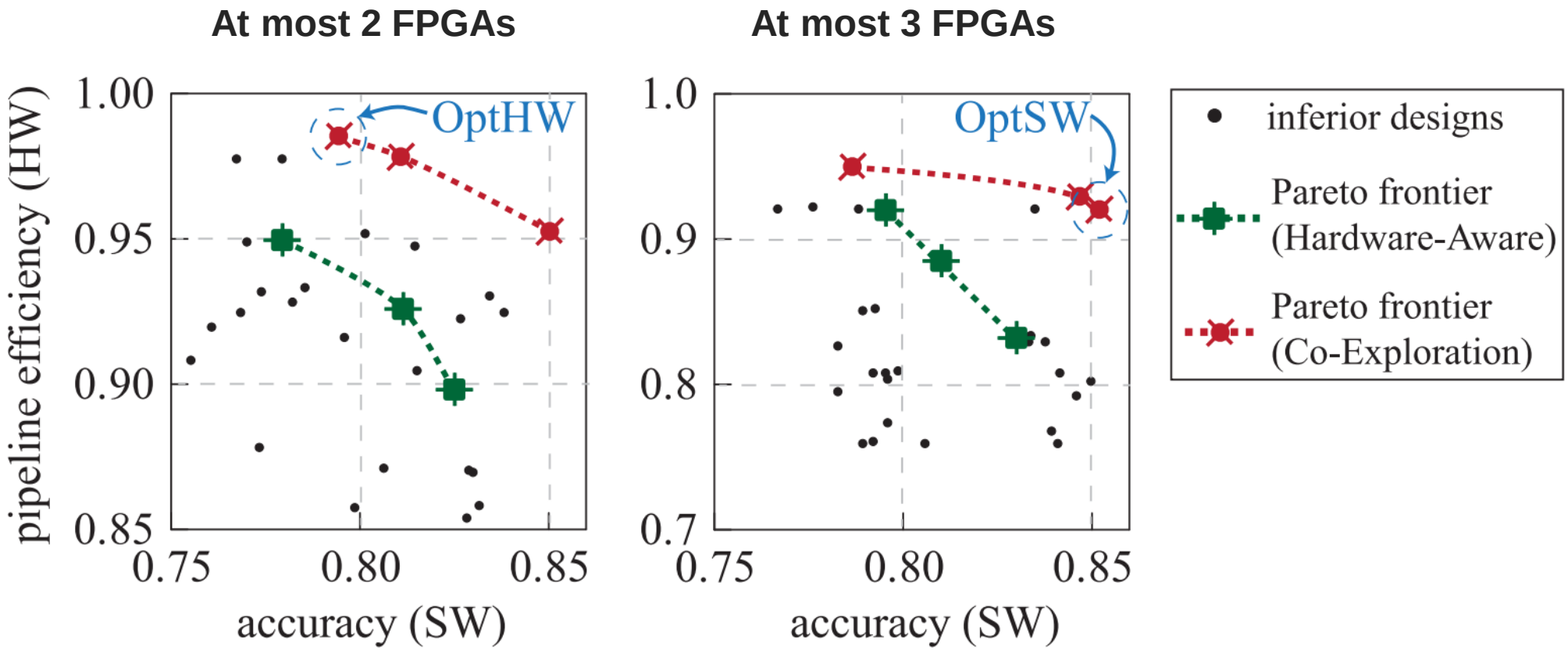


Framework: Optimizing Network Architecture and HW Design in One Loop



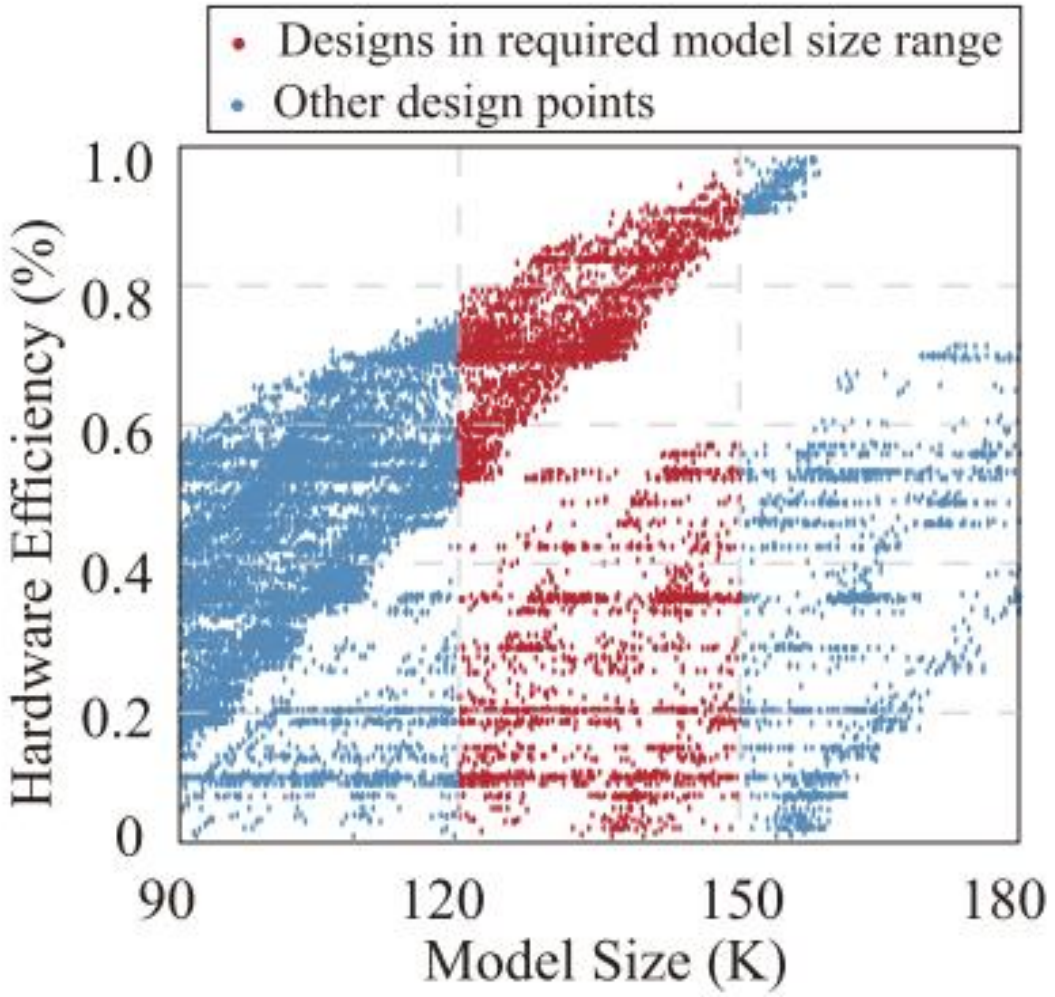
- **Controller** iteratively selects solution from the **search space** for **evaluation**
- **Controller** is evolved using the **evaluation results** from previous iteration.

Results



Pareto Frontier between Accuracy and HW Efficiency can be Significantly Pushed Forward

Results



Minimizing model size may not achieve the highest hardware efficiency

Results

Dataset	Models	Depth	Parameters	Accuracy (Top1)	Accuracy (Top5)	Pipeline Eff.	FPS	Energy Eff. GOPS/W
CIFAR-10	Hardware-Aware NAS	13	0.53M	84.53%	-	73.27%	16.2	0.84
	Sequential Optimization	13	0.53M	84.53%	-	92.20%	29.7	1.36
	Co-Exploration (OptHW)	10	0.29M	80.18%	-	**99.69%**	**35.5**	**2.55**
	Co-Exploration (OptSW)	14	0.61M	**85.19%**	-	92.15%	**35.5**	1.91
ImageNet	Hardware-Aware NAS	15	0.44M	68.40%	89.84%	81.07%	6.8	0.34
	Sequential Optimization	15	0.44M	68.40%	89.84%	86.75%	10.4	0.46
	Co-Exploration (OptHW)	17	0.54M	68.00%	89.60%	**96.15%**	**12.1**	**1.01**
	Co-Exploration (OptSW)	15	0.48M	**70.24%**	**90.53%**	93.89%	10.5	0.74

Co-Exploration Outperform Competitors on Different Datasets

Outline

- Background
- A Quick Overview of The Road From Manual Design to AutoML
- HW/SW Co-Exploration Framework
  - Motivation
  - Framework Overview and Details
  - Results
- **Follow-up Works and Conclusion**

Design Stack Built By the Team

- **HW/SW Co-Design**
- **Quantum Machine Learning**

HW/SW Co-Design Stack using NAS

HW/SW Co-Design Framework
FNAS [DAC'19*] [TCAD'20*]

Application

Medical Imaging	NLP (Transformer)	Graph-Based
NAS for Medical Image Seg. [MICCAI'20]; FaHaNA Fairness [DAC'22]	FPGA [ICCD'20] Mobile [DAC'21] GPU [GLSVLSI'21]	Social Net [GLSVLSI'21] Drug Discovery [ICCAD'21]

Algorithm

NAS Acc.	Model Compression	Secure Infernece
HotNAS [CODES+ISSS'20]	NAS for Quan. [ICCAD'19] Compre.-Compilation [IJCAI'21]	NASS [ECAI'20] BUNET [MICCAI'20]

Hardware

FPGA	ASIC	Computing-in-Memory
XFER [CODES+ISSS'19*]	NANDS [ASP-DAC'20*] ASICNAS [DAC'20]	Device-Circuit-Arch. [IEEE TC'20]

Best Paper Award:

IEEE

IEEE Council on Electronic Design Automation

hereby presents the

2021 IEEE Transactions on Computer-Aided Design Donald O. Pederson Best Paper Award

to

Weiwen Jiang, Lei Yang, Edwin Hsing-Mean Sha, Qingfeng Zhuge, Shouzhen Gu, Sakyasingha Dasgupta, Yiyu Shi, Jingtong Hu

for the paper entitled

"Hardware/Software Co-Exploration of Neural Architectures"

Yao-Wen Chang, President, IEEE Council on Electronic Design Automation

Rajesh Gupta, Editor-in-Chief, IEEE Transactions on Computer-Aided Design

CEDA

Best Paper Nominations:

58 DESIGN AUTOMATION CONFERENCE
FROM CHIPS TO SYSTEMS — LEARN TODAY, CREATE TOMORROW

EMBEDDED SYSTEMS WEEK

ASIA SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE
DAC

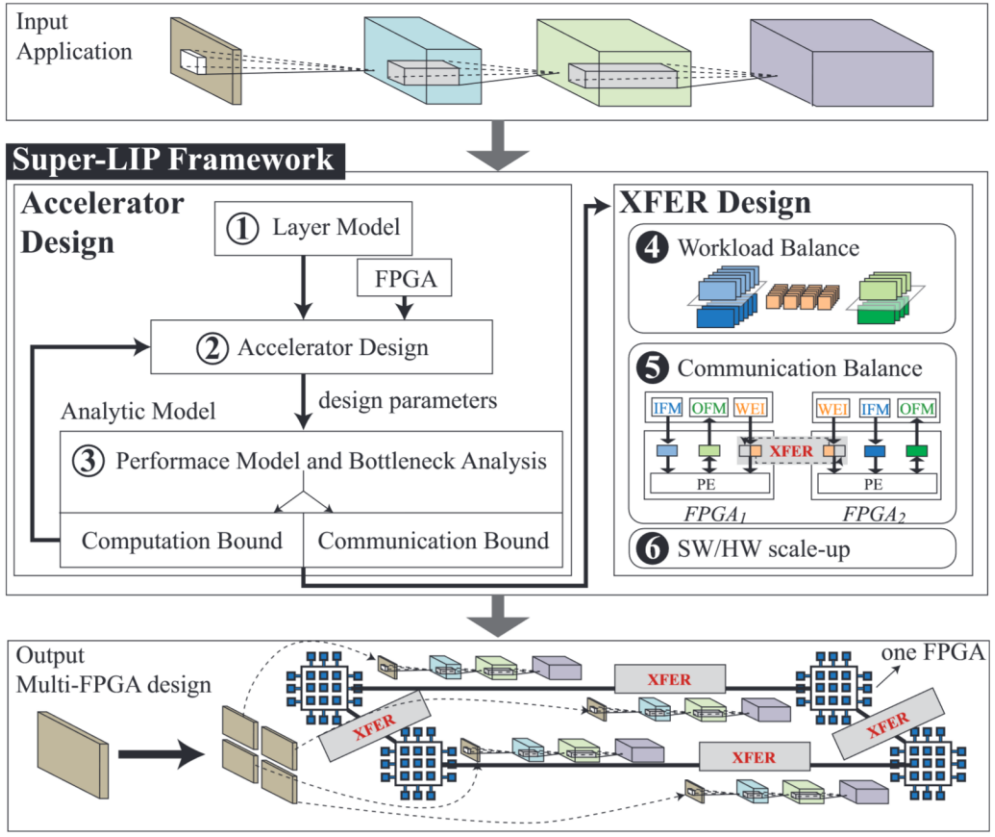
Co-Design NAS Full-Stack

HW/SW Co-Design Framework

FNAS

[DAC'19*]

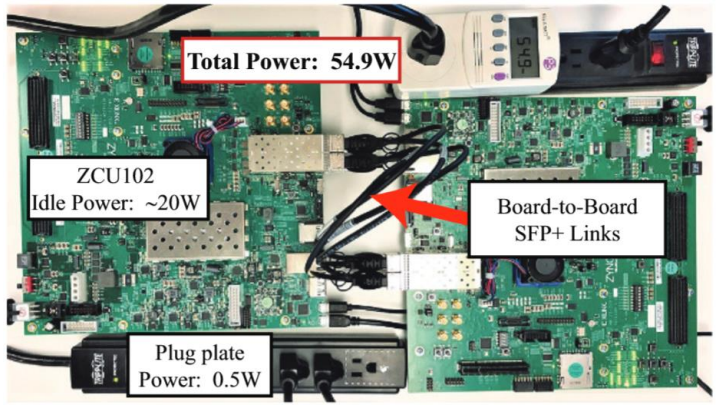
[TCAD'20*]



Best Paper Nomination

XFER:

- Neural Network Partition
- Performance Model
- Multiple FPGA
- Load Balance



Hardware

FPGA

XFER [CODES+ISSS'19*]

ASIC

NANDS [ASP-DAC'20*]
ASICNAS [DAC'20]

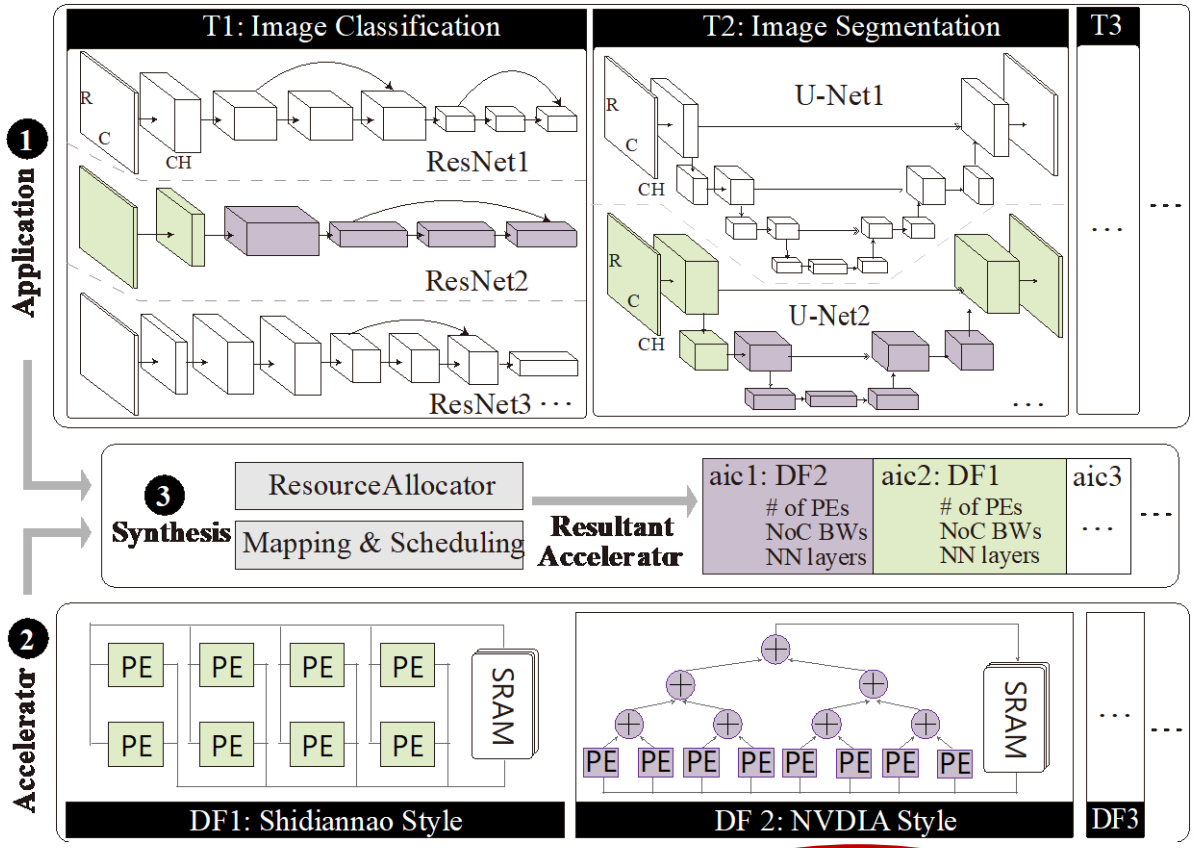
Computing-in-Memory

Device-Circuit-Arch. [IEEE TC'20]

Co-Design NAS Full-Stack

HW/SW Co-Design Framework

FNAS
[DAC'19*]
[TCAD'20*]



DESIGN AUTOMATION CONFERENCE 57

facebook

First HW/SW Co-Design For ASICs with Huge HW Design Space

ASINAS:

- Multi-Tasks
- Template-Based NAS
- Heterogenous ASICs

FPGA

XFER
[CODES+ISSS'19*]

ASIC

NANDS [ASP-DAC'20*]
ASICNAS [DAC'20]

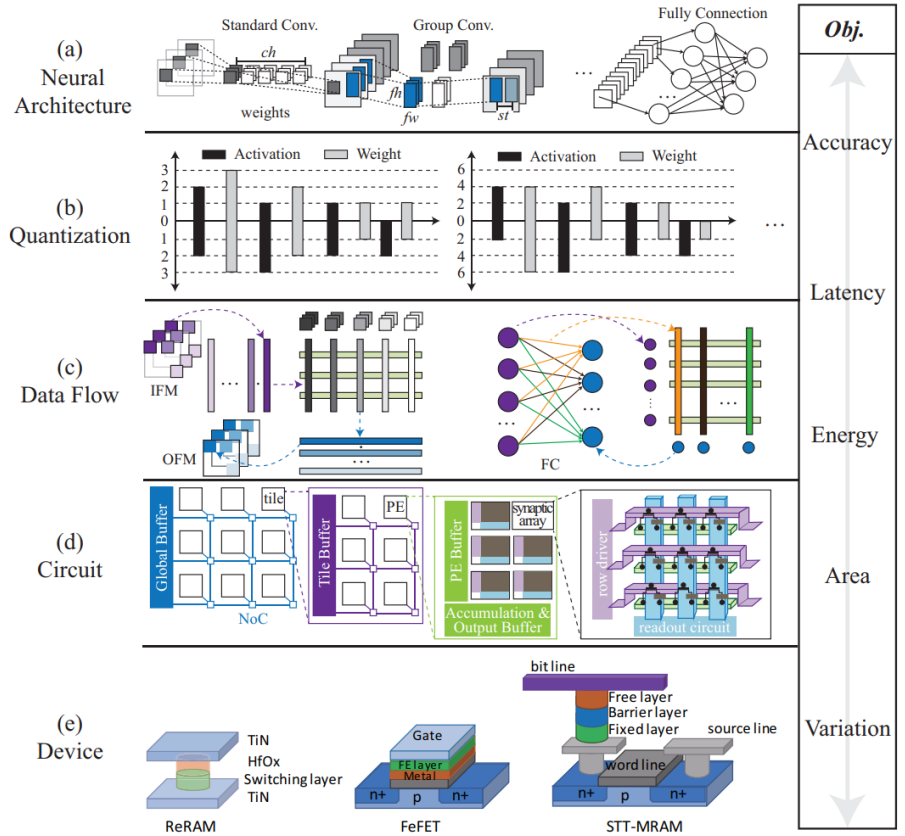
Computing-in-Memory

Device-Circuit-Arch.
[IEEE TC'20]

Co-Design NAS Full-Stack

HW/SW Co-Design Framework

FNAS [DAC'19*] [TCAD'20*]



First HW/SW Co-Design For Computing-in-Memory Accelerators with Device Variation

NACIM:

- Cross-layer Optimization
- Multi-Object Optimization
- CiM Accelerator
- Device Variation

FPGA

XFER [CODES+ISSS'19*]

ASIC

NANDS [ASP-DAC'20*]
ASICNAS [DAC'20]

Computing-in-Memory

Device-Circuit-Arch. [IEEE TC'20]

Co-Design NAS Full-Stack

HW/SW Co-Design Framework

FNAS
[DAC'19*]
[TCAD'20*]

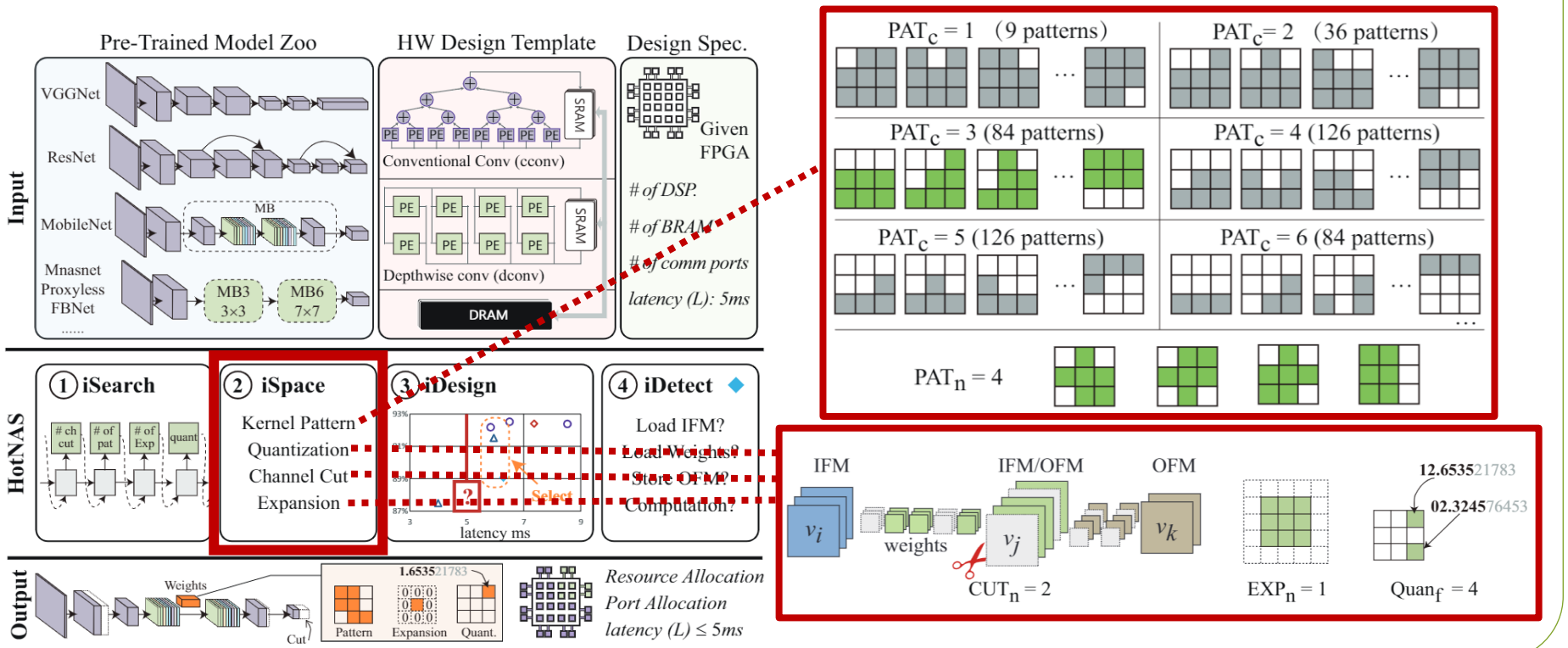
Algorithm

NAS Acc.
HotNAS
[CODES+ISSS'20]

Model Compression
NAS for Quan. [ICCAD'19]
Compre.-Compilation [IJCAI'21]

Secure Infernece
NASS [ECAI'20]
BUNET [MICCAI'20]

More than 200 → Less than 3 (GPU Hours) on ImageNet Dataset



Co-Design NAS Full-Stack

HW/SW Co-Design Framework

FNAS [DAC'19*] [TCAD'20*]

Application

Medical Imaging	NLP (Transformer)	Graph-Based
NAS for Medical Image Seg. [MICCAI'20]	FPGA [ICCD'20]	Social Net [GLSVLSI'21]
FaHaNA Fairness [DAC'22]	Mobile [DAC'21]	Drug Discovery [ICCAD'21]
	GPU [GLSVLSI'21]	

Fairness

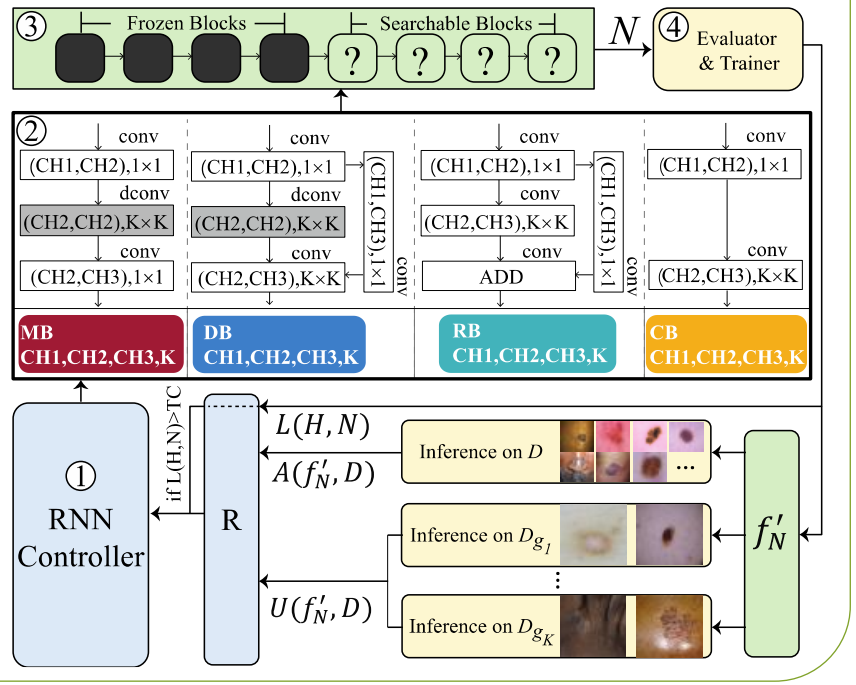
Vascular lesion | Melanoma | Basal cell carcinoma

- *Existing Networks*

Light Skin 76.80%

Dark Skin 49.38%

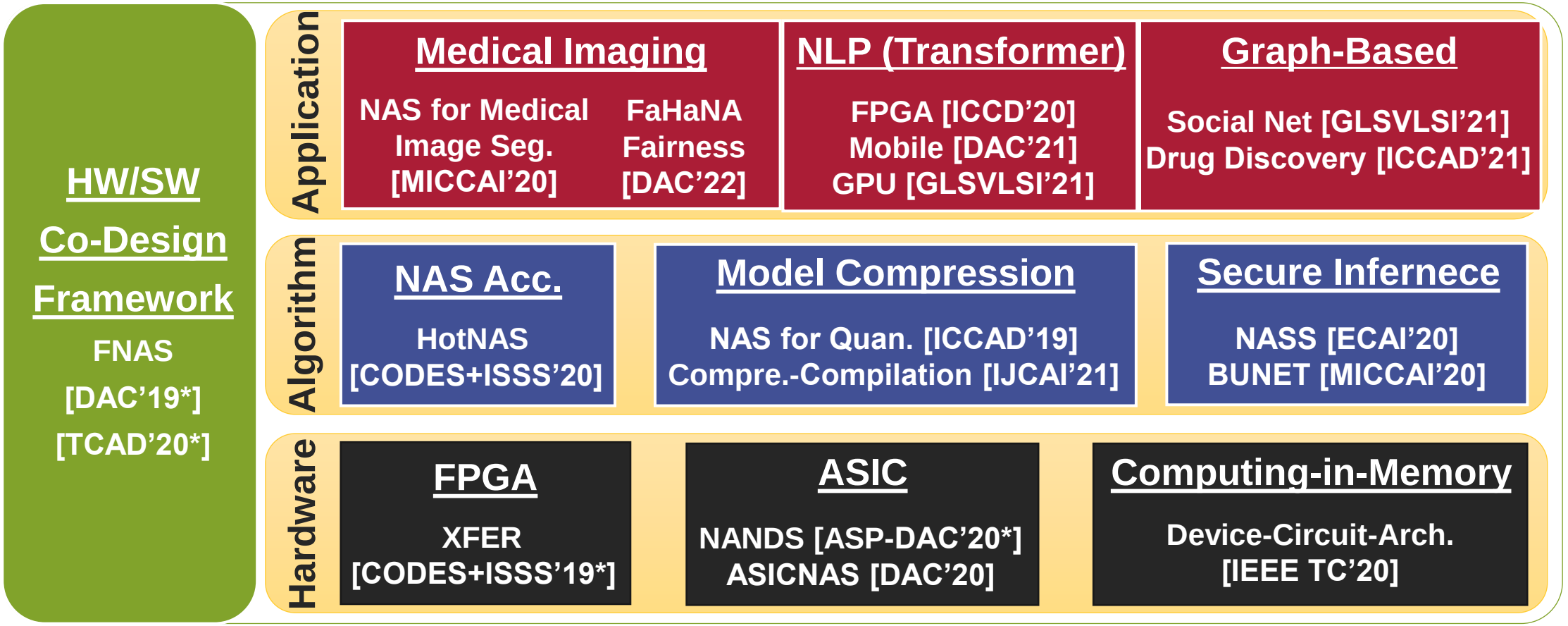
First Framework to Co-Optimize Accuracy and Fairness via NAS



DESIGN AUTOMATION CONFERENCE 59
FROM CHIPS TO SYSTEMS — LEARN TODAY, CREATE TOMORROW

DAC This Year!

Co-Design NAS Full-Stack



Conclusion

- The **very first framework** to conduct **HW/SW co-exploration**
- Co-exploration can **push forward** Pareto frontier of **accuracy vs. efficiency**
- Providing **fundamentals** of the **co-design stack**
- We even apply the **co-design philosophy** to **quantum machine learning**, the initial work was published at **Nature Communications**

Reference

[1] **W. Jiang,** J. Xiong, and Y. Shi, A Co-Design Framework of Neural Networks and Quantum Circuits Towards Quantum Advantage, ***Nature Communications,*** *Jan. 2021*

[2] **W. Jiang**, B. Xie, C-C Liu and Y. Shi, Integrating Memristors and CMOS for Better AI, ***Nature Electronics****, Sep. 2019*

[3] Y. Ding, **W. Jiang**, Q. Lou, J. Liu, J. Xiong, X. Sharon Hu, X. Xu, and Y. Shi, Hardware Design and the Competency Awareness of a Neural Network, ***Nature Electronics****, Aug. 2020*

[4] **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)**

[5] **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**)

[6] **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.*

[7] **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, *International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (****CASES****), also appears at IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (****TCAD****), Torino, Italy, Oct. 2018.*

[8] **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.*

Reference

[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] **W. Jiang**, L. Yang, S. Dasgupta, J. Hu and Y. Shi, Standing on the Shoulders of Giants: Hardware and Neural Architecture Co-Search with Hot Start, *International Conference on Hardware/Software Co-design and System Synthesis **CODE+ISSS**), also appears at IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**), Virtual, 2020.*

[11] **W. Jiang**, Q. Lou, Z. Yan, L. Yang, J. Hu, X. S. Hu and Y. Shi, Device-Circuit-Architecture Co-Exploration for Computing-in-Memory Neural Accelerators, *IEEE Transactions on Computers (**TC**), Accepted, 2020.*

[12] 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.*

[13] W. Liu, L. Yang, **W. Jiang**, L. Feng, N. Guan, W. Zhang, and N. Dutt, Thermal-aware Task Mapping on Dynamically Reconfigurable Network-on-Chip based Multiprocessor System-on-Chip, *IEEE Transactions on Computers (**TC**), Accepted, 2018.*

[14] L. Yang, W. Liu, **W. Jiang**, M. Li, P. Chen and E. H.-M. Sha, FoToNoC: A Folded Torus-Like Network-on-Chip based Many-Core Systems-on-Chip in the Dark Silicon Era, *IEEE Transactions on Parallel and Distributed Systems (**TPDS**), 28(7), 1905-1918, 2017.*

[15] E. H.-M. Sha, X. Chen, Q. Zhuge, L. Shi and **W. Jiang,** A New Design of In-Memory File System Based on File Virtual Address Framework, *IEEE Transactions on Computers (**TC**), 65(10), 2959-2972, Oct. 2016.* **(Editor's pick of the year)**

Reference

[16] **W. Jiang,** J. Xiong, and Y. Shi, When Machine Learning Meets Quantum Computers: A Case Study," in Proc. of IEEE/ACM Asia and South Pacific Design Automation Conference (ASP-DAC), 2021 **(Invited Paper)**

[17] **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)**

[18] 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**)

[19] 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.*

[20] Q. Lu, **W. Jiang**, X. Xiao, J. Hu and Y. Shi, On Neural Architecture Search for Resource-Constrained Hardware Platforms, *Proc. IEEE/ACM International Conference On Computer-Aided Design (**ICCAD**), Westminster, CO, 2019.*

[21] 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.*

[22] **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.*

[23] **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.*

wjiang8@gmu.edu

George Mason University

4400 University Drive
Fairfax, Virginia 22030

Tel: (703)993-1000