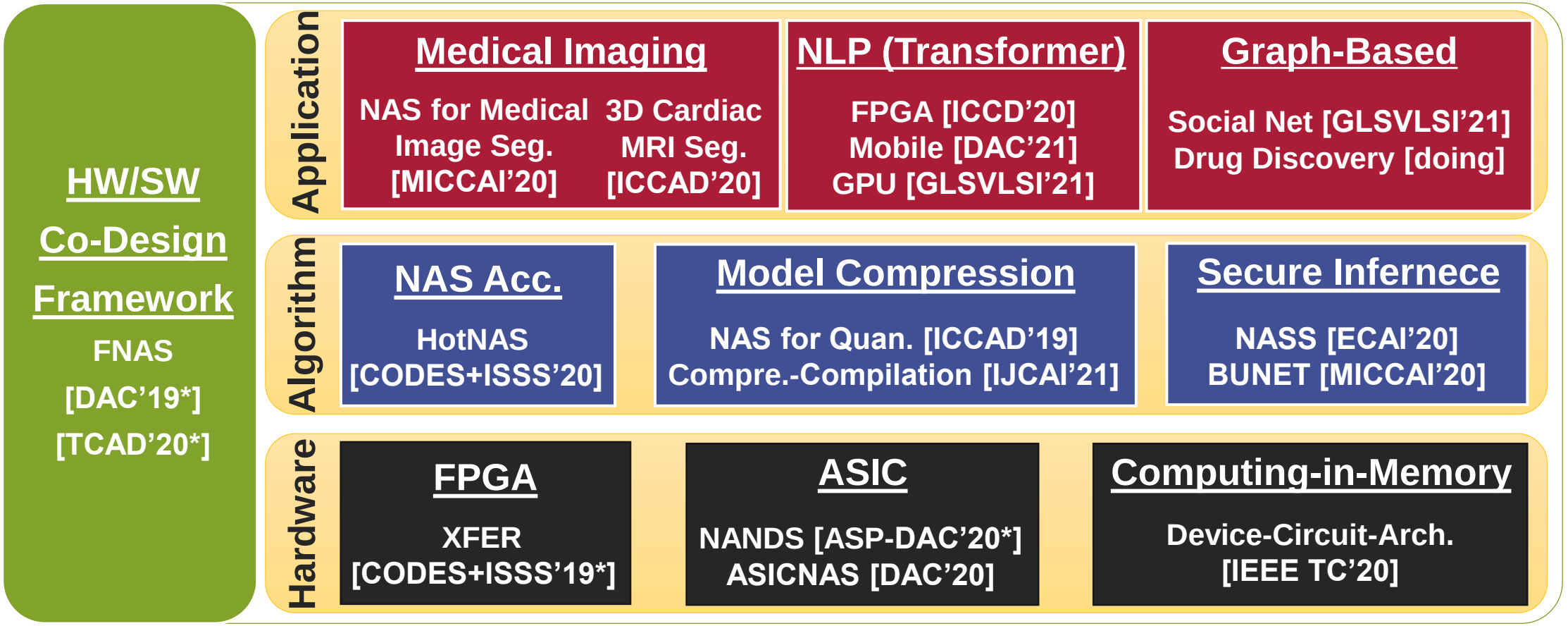
Full-Stack Classical AutoML Projects



* Best Paper Nomination

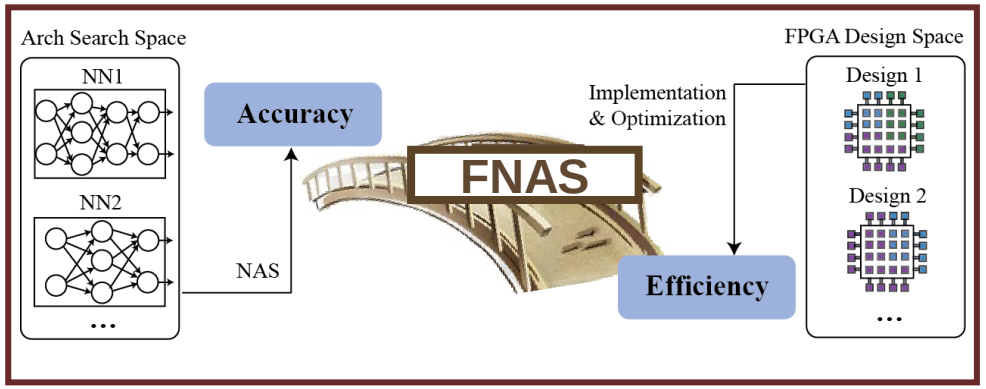
Full-Stack Classical AutoML Projects

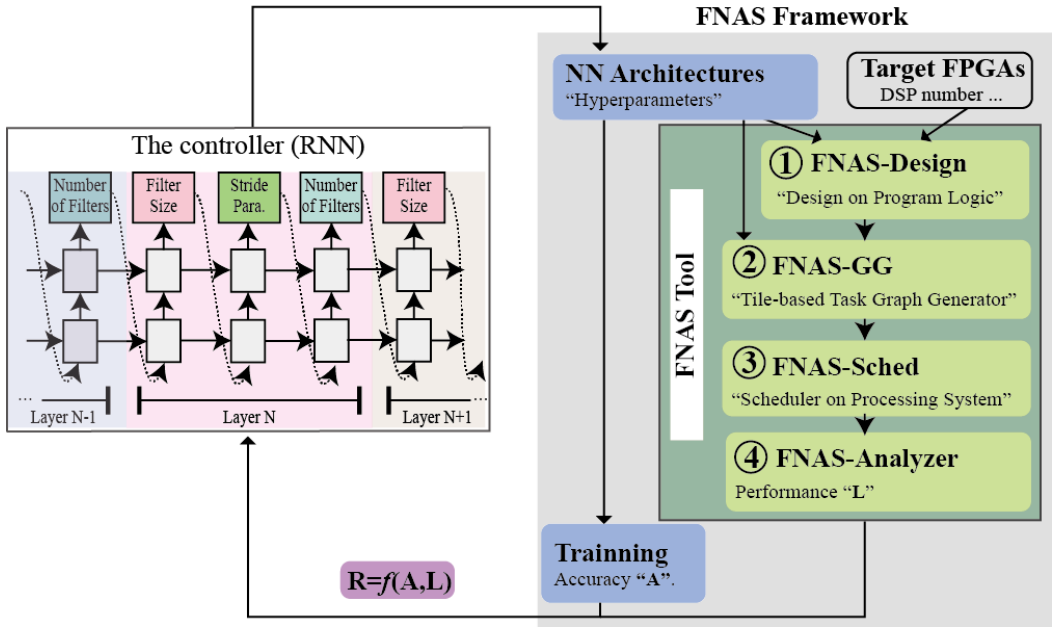
HW/SW Co-Design Framework

FNAS

[DAC'19*]

[TCAD'20*]

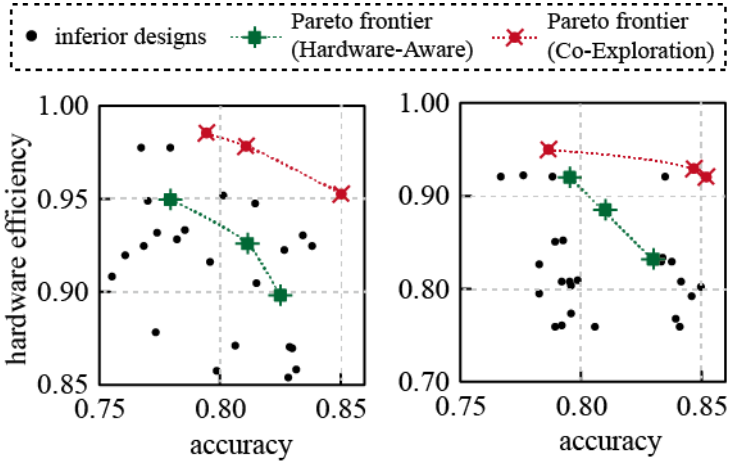




First HW/SW Co-Design Framework

FNAS:

- Neural Architecture Search
- RNN-based RL Framework
- HW/SW Co-Design
- FPGA Optimization



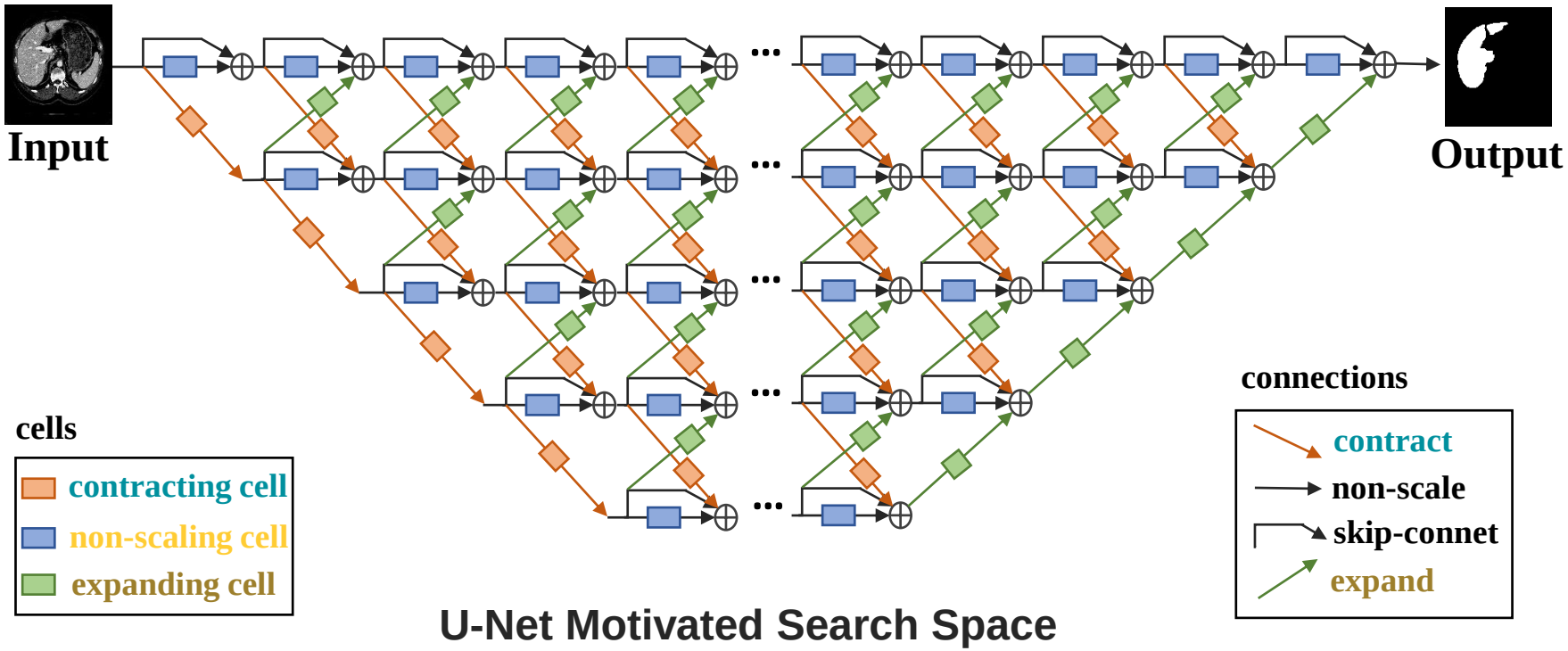
Full-Stack Classical AutoML Projects

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]
3D Cardiac MRI Seg. [ICCAD'20]	Mobile [DAC'21]	Drug Discovery [doing]
	GPU [GLSVLSI'21]	



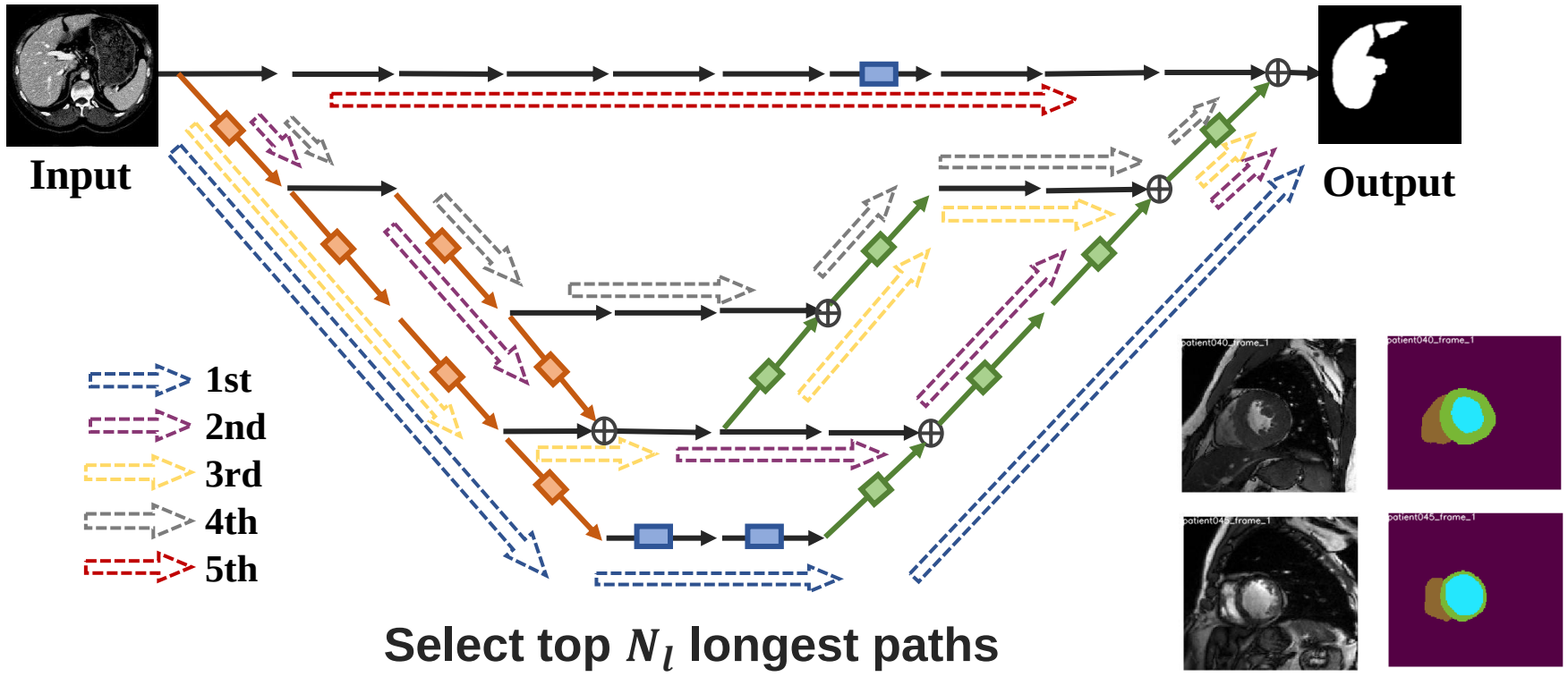
Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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

Application

Medical Imaging	NLP (Transformer)	Graph-Based
NAS for Medical Image Seg. [MICCAI'20]	3D Cardiac MRI Seg. [ICCAD'20]	FPGA [ICCD'20] Mobile [DAC'21] GPU [GLSVLSI'21]	Social Net [GLSVLSI'21] Drug Discovery [doing]



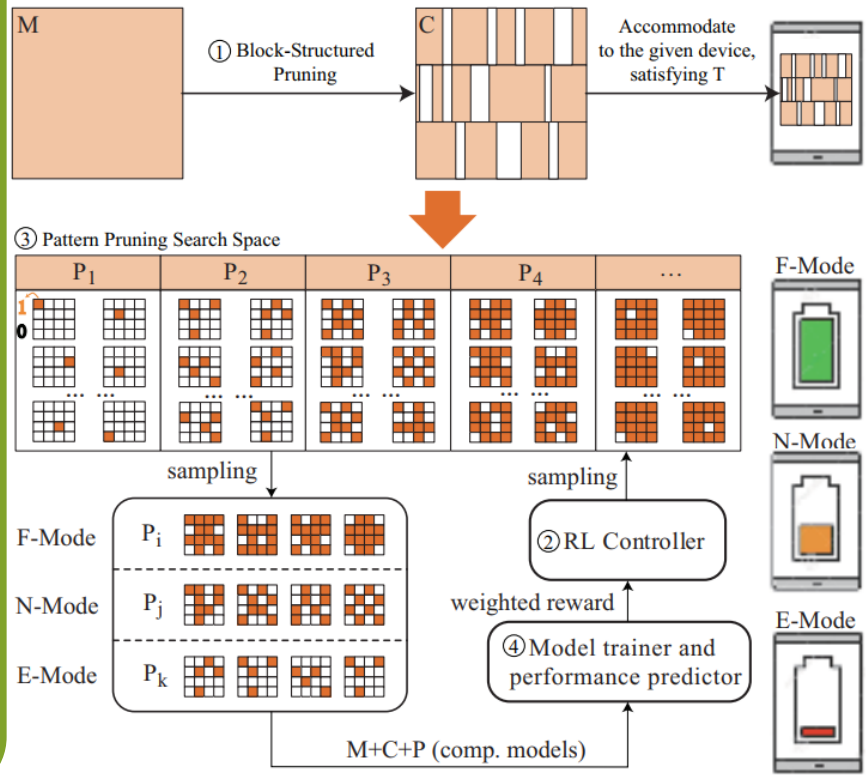
Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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

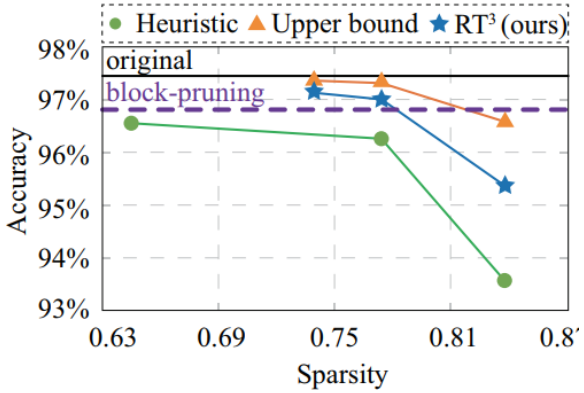
Application

Medical Imaging	NLP (Transformer)	Graph-Based
NAS for Medical Image Seg. [MICCAI'20] 3D Cardiac MRI Seg. [ICCAD'20]	FPGA [ICCD'20] Mobile [DAC'21] GPU [GLSVLSI'21]	Social Net [GLSVLSI'21] Drug Discovery [doing]



Mobile:

- Patten Pruning
- RL-Based NAS
- DVFS
- Reconfigurable



Dataset/Task		WikiText-2 (T: 94ms)		
Models		Transformer		
		M1	M2	M3
Sparsity		70.80%	80.61%	87.32%
Latency (ms)		93.55	86.78	70.72
UB	Accuracy	97.27%	96.29%	93.03%
	Interrupt	51.82 seconds		
RT³	**Accuracy**	**95.40%**	**95.37%**	**90.04%**
	Interrupt	**8.75 milliseconds**		
	Accuracy gap	**1.87%**	**0.92%**	**2.99%**

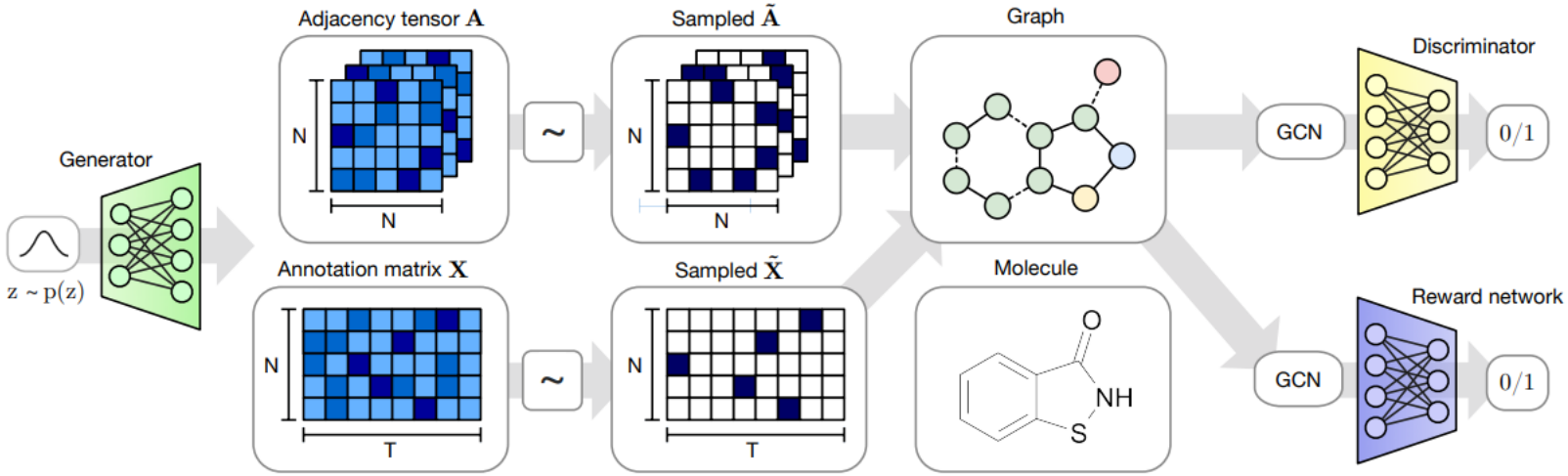
Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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

Application

Medical Imaging	NLP (Transformer)	Graph-Based
NAS for Medical Image Seg. [MICCAI'20] 3D Cardiac MRI Seg. [ICCAD'20]	FPGA [ICCD'20] Mobile [DAC'21] GPU [GLSVLSI'21]	Social Net [GLSVLSI'21] Drug Discovery [doing]



Drug Discovery [Conducting Project]:

- Graph Neural Network
- Generative Model

NAS?

Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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

NAS Acc.

HotNAS [CODES+ISSS'20]

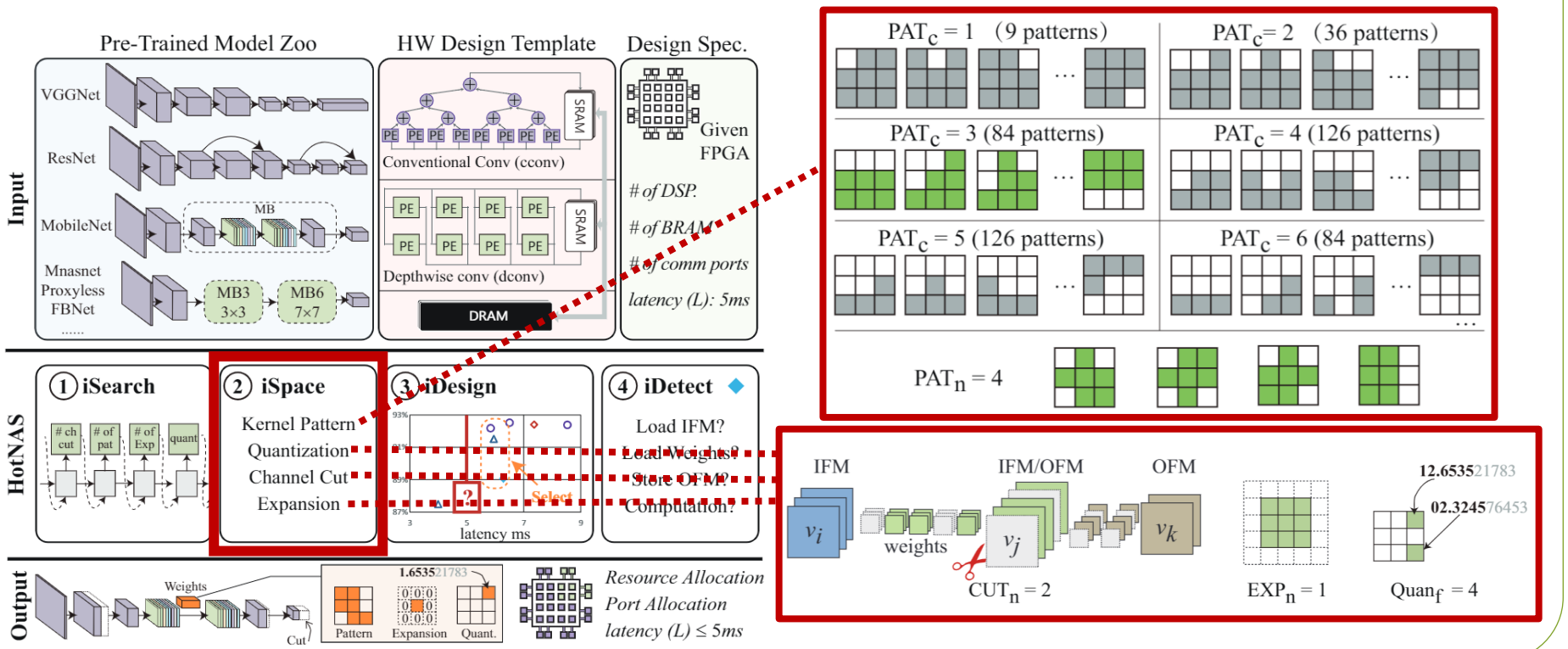
Model Compression

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



Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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

NAS Acc.

HotNAS [CODES+ISSS'20]

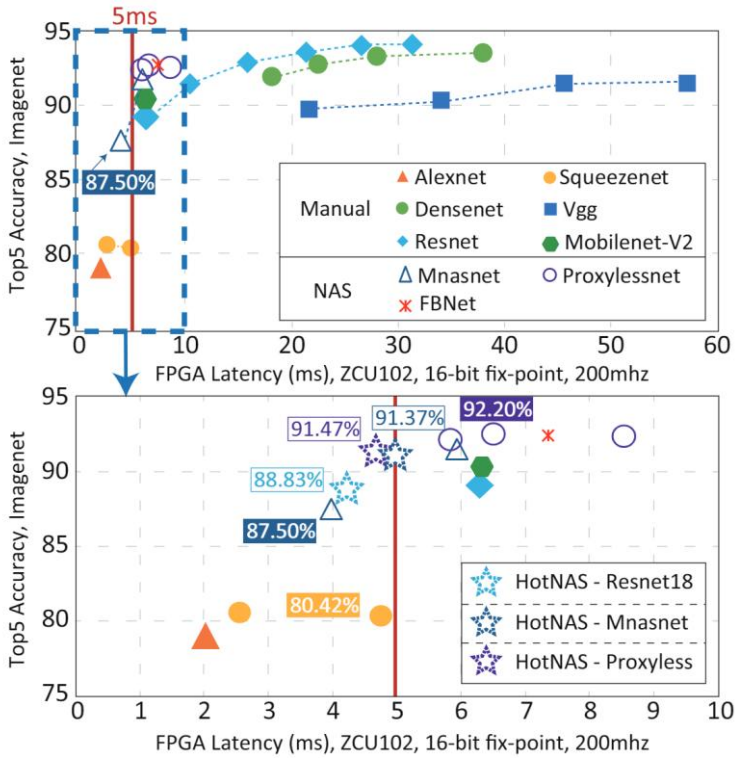
Model Compression

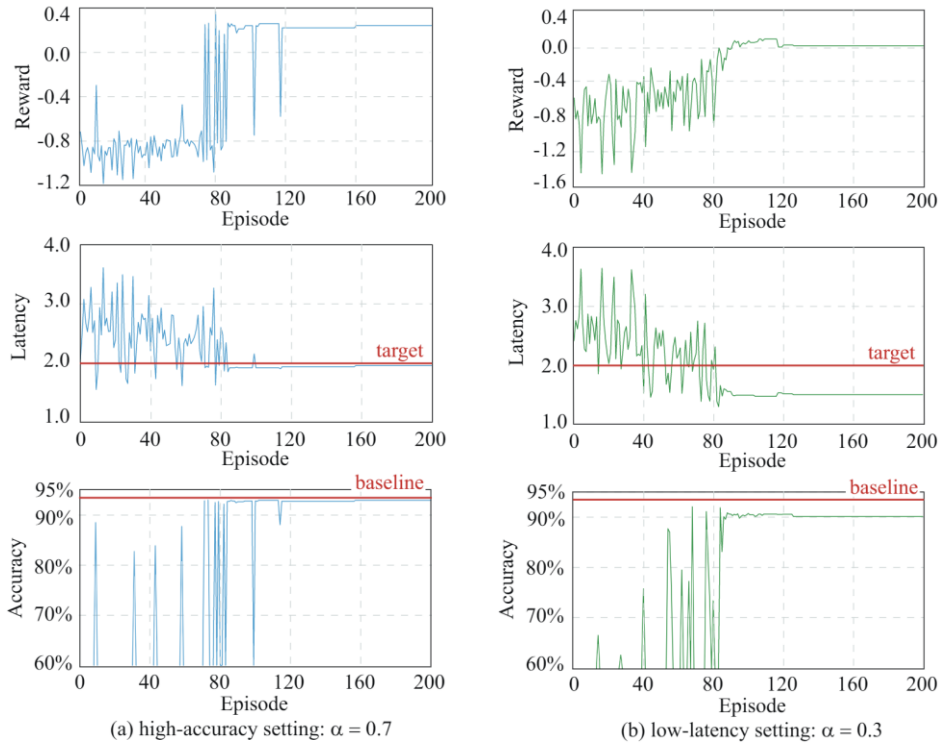
NAS for Quan. [ICCAD'19]
Compre.-Compilation [IJCAI'21]

Secure Infernece

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

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





Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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

NAS Acc.

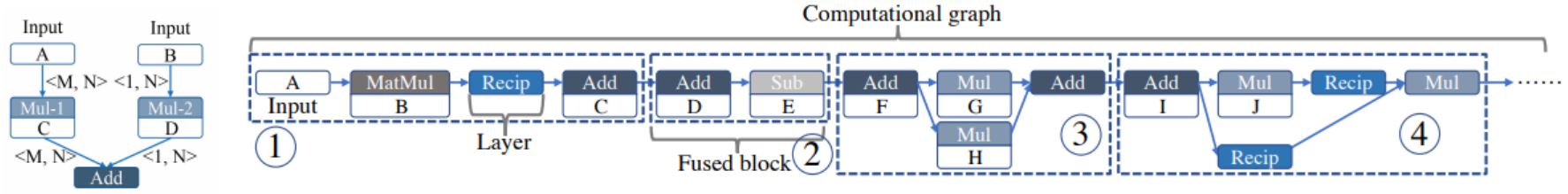
HotNAS
[CODES+ISSS'20]

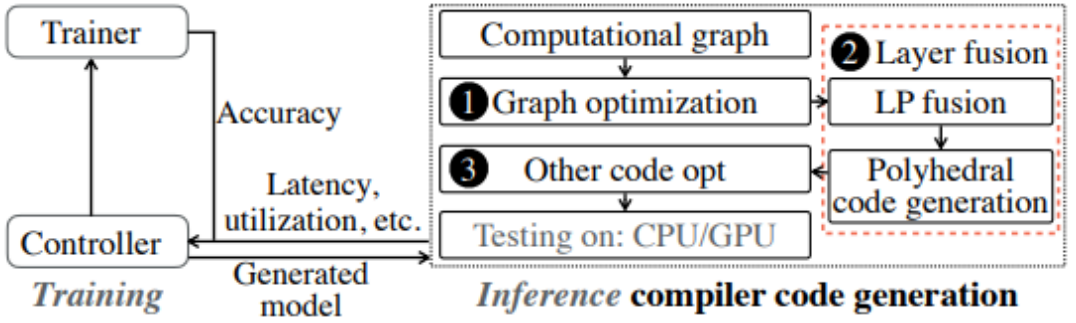
Model Compression

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

Secure Infernece

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





7.8× speedup compared with TensorFlow-Lite with only minor accuracy loss.

Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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

NAS Acc.

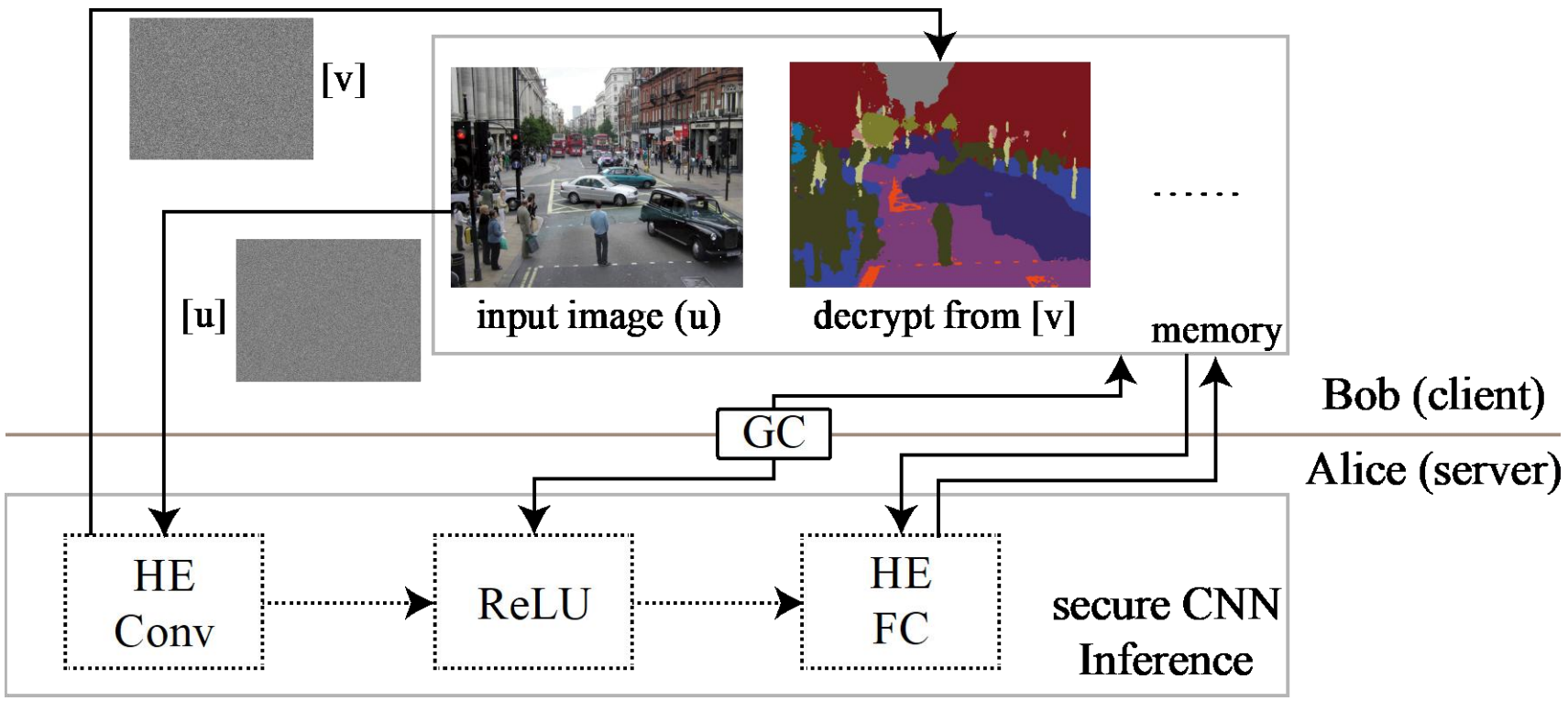
HotNAS
[CODES+ISSS'20]

Model Compression

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

Secure Infernece

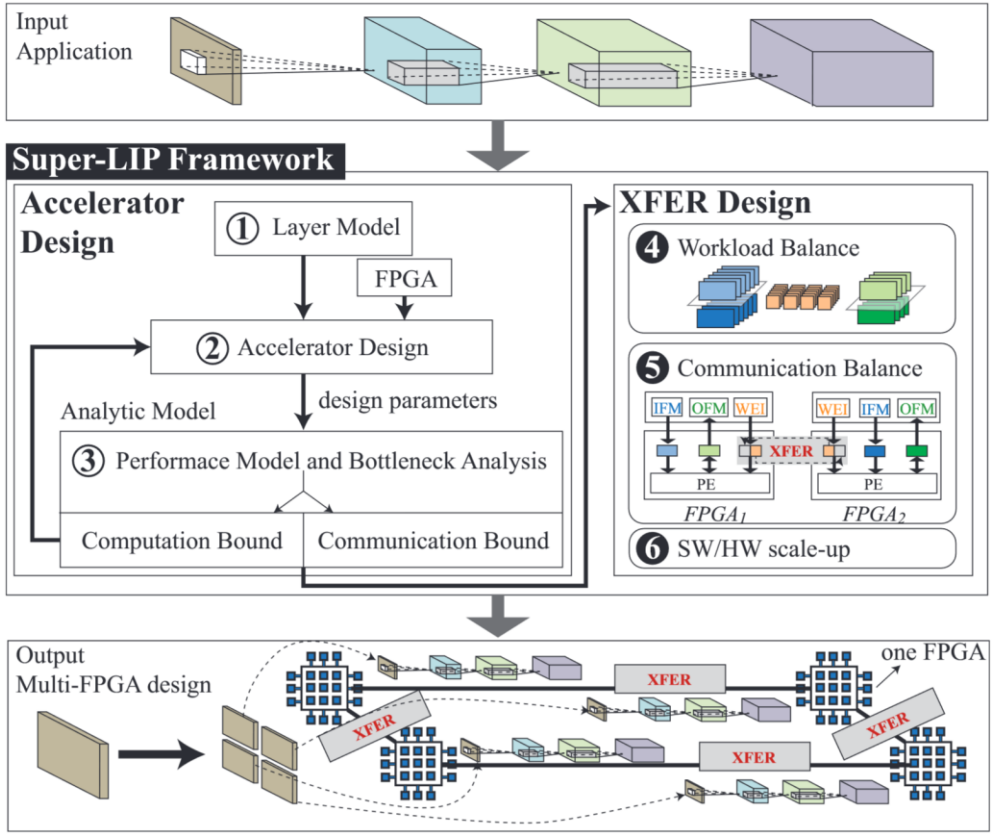
NASS [ECAI'20]
BUNET [MICCAI'20]



Full-Stack Classical AutoML Projects

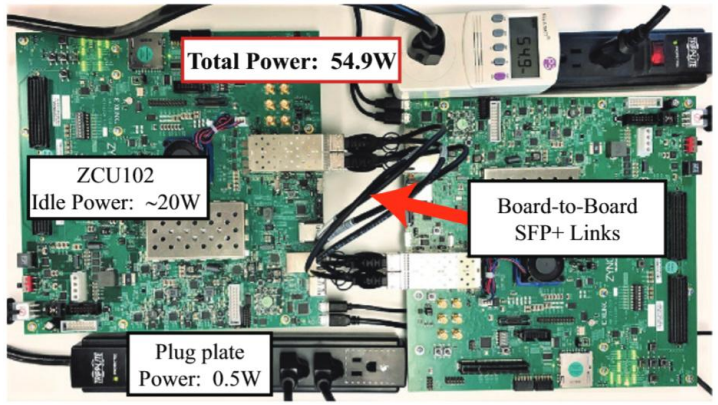
HW/SW Co-Design Framework

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



XFER:

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



FPGA

XFER
[CODES+ISSS'19*]

ASIC

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

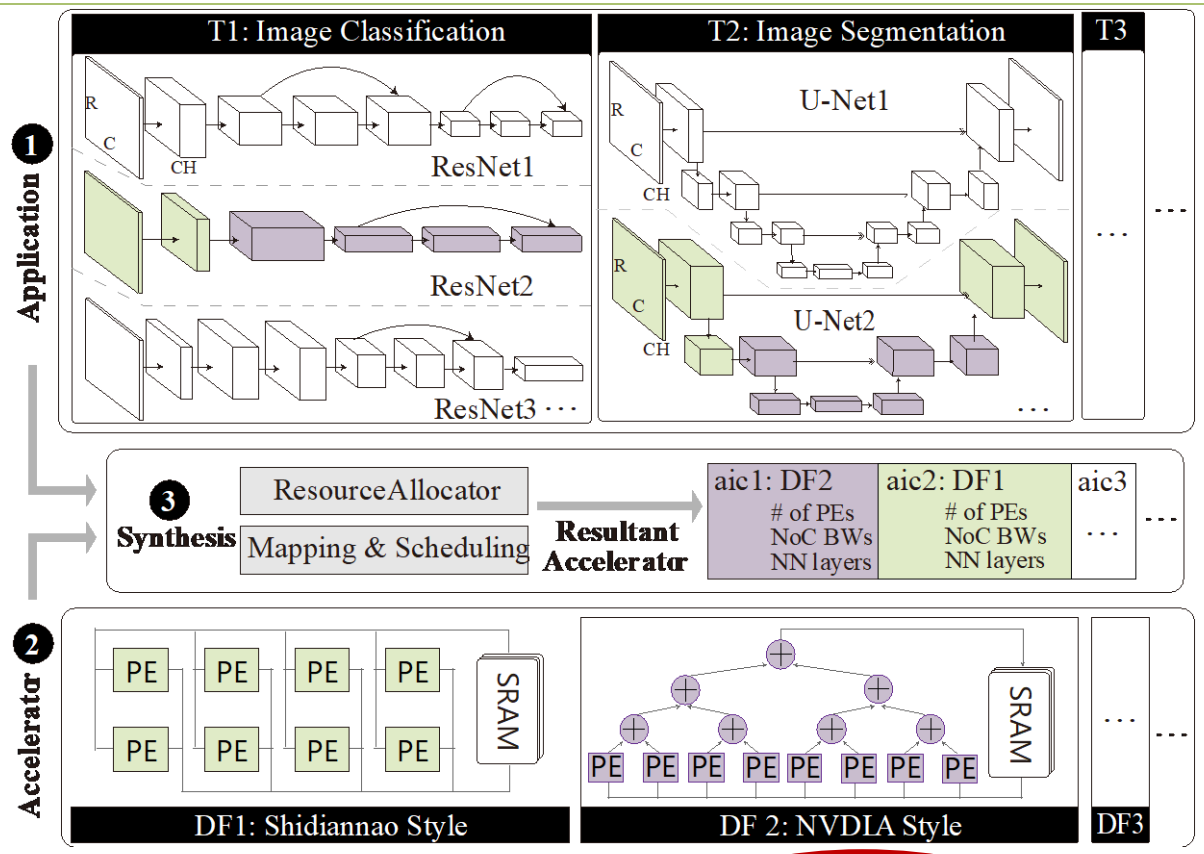
Computing-in-Memory

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

Full-Stack Classical AutoML Projects

HW/SW Co-Design Framework

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



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]

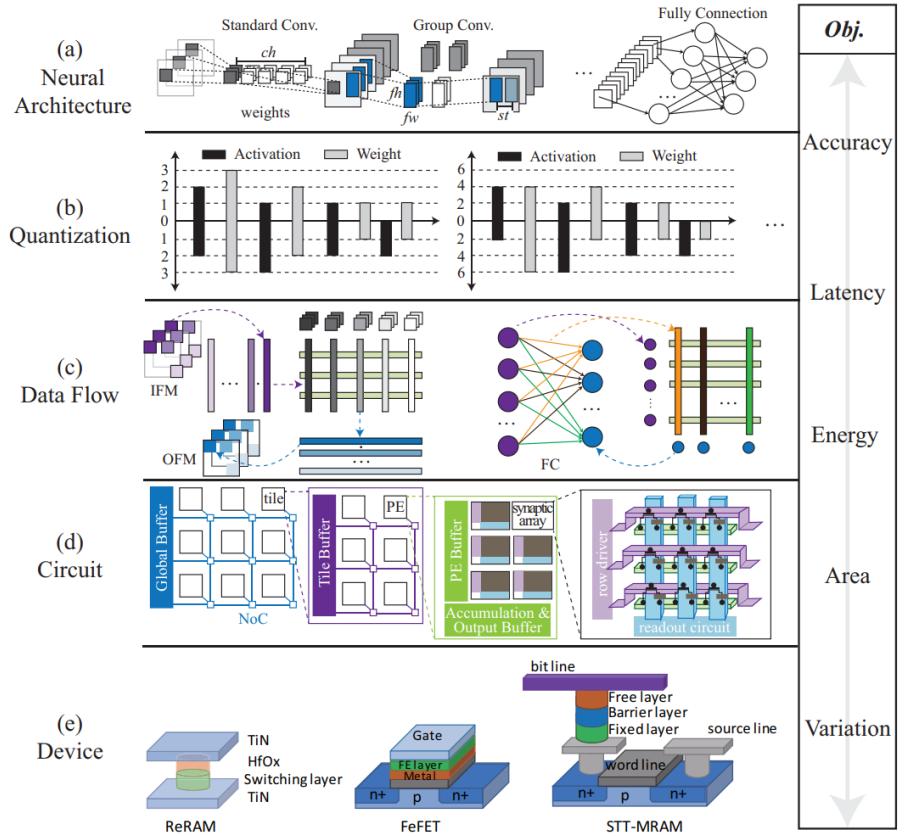
Computing-in-Memory

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

Full-Stack Classical AutoML Projects

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]

References

[GLSVLSI'21] **D. Manu**, S. Huang, C. Ding, **L. Yang**, Co-Exploration of Graph Neural Network and Network-on-Chip Design using AutoML

[IJCAI'21 Demonstration Track] W. Niu, Z. Kong, G. Yuan, **W. Jiang,** B. Ren, and Y. Wang, A Compression-Compilation Framework for On-mobile Real-time BERT Applications

[DAC'21] Y. Song, **W. Jiang**, B. Li, P. Qi, Q. Zhuge, E. H.-M. Sha, S. Dasgupta, Y. Shi, and C. Ding, Dancing along Battery: Enabling Transformer with Run-time Reconfigurability on Mobile Devices

[NCOMM'21] **W. Jiang**, J. Xiong, and Y. Shi, A Co-Design Framework of Neural Networks and Quantum Circuits Towards Quantum Advantage

[ASP-DAC'21] **W. Jiang**, J. Xiong, and Y. Shi, When Machine Learning Meets Quantum Computers: A Case Study

[ICCD'20] X. Zhang, **W. Jiang**, J. Hu, Achieving Full Parallelism in LSTM via a Unified Accelerator Design

[CODES+ISSS'20 & TCAD'20] **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

[IEEE TC'20] **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 TCAD'20] **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 **(Best Paper Nomination)**

[DAC'20] **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

[ECAI'20] B. Song, **W. Jiang**, Q. Lu, Y. Shi and T. Sato, NASS: Optimizing Secure Inference via Neural Architecture Search

[ASP-DAC'20] **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 **(Best Paper Nomination)**

[CODES+ISSS'19 & AMC TECS] **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 **(Best Paper Nomination)**

[DAC'19] **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