QAPPA: Quantization-Aware Power, Performance, and Area Modeling of DNN Accelerators

Ahmet Inci[1], Siri Garudanagiri Virupaksha[1], Aman Jain[1], Venkata Vivek Thallam[1], Ruizhou Ding[1], Diana Marculescu[1,2]

[1] Carnegie Mellon University

[2] TEXAS The University of Texas at Austin

2nd On-Device Intelligence Workshop

April 9, 2021

MLSys'21

On-device intelligence pushes hardware to its limits

- **Increasing model size and computational cost of ML models**

Autonomous Vehicles

[Source: Wayve]

[Source: Google]

Drones

[Source: CNN]

IoT – Smart Home & Security

[Source: Nest]

Mobile Platforms

[Source: Fritz AI]

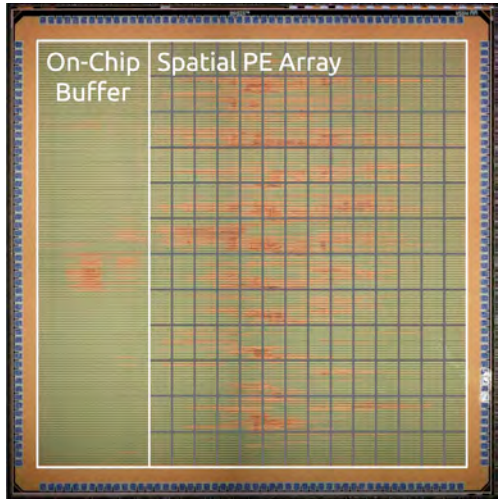
AR/VR

[Source: Facebook]

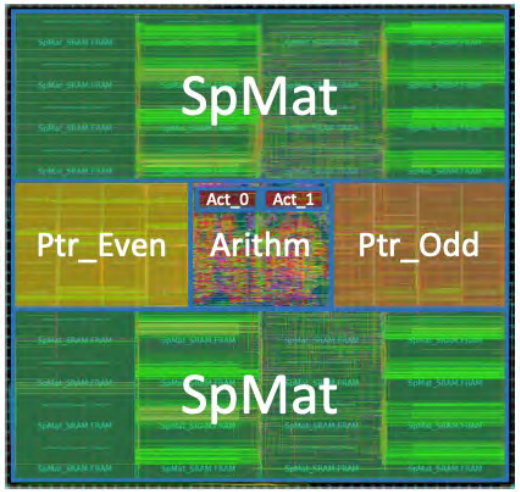
- **Hardware constraints are a key limiting factor for ML on edge devices**
  - Stringent performance per area and energy-efficiency constraints
  - Chip area is one of the most expensive real estates
  - On-device inference time (latency) constraints

Enabling deployment of DNNs onto edge devices

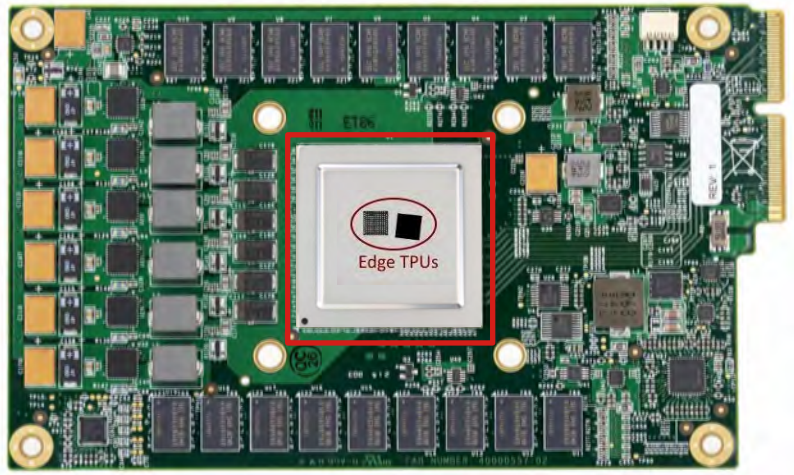
- **DNN accelerators**



[Source: Eyeriss]

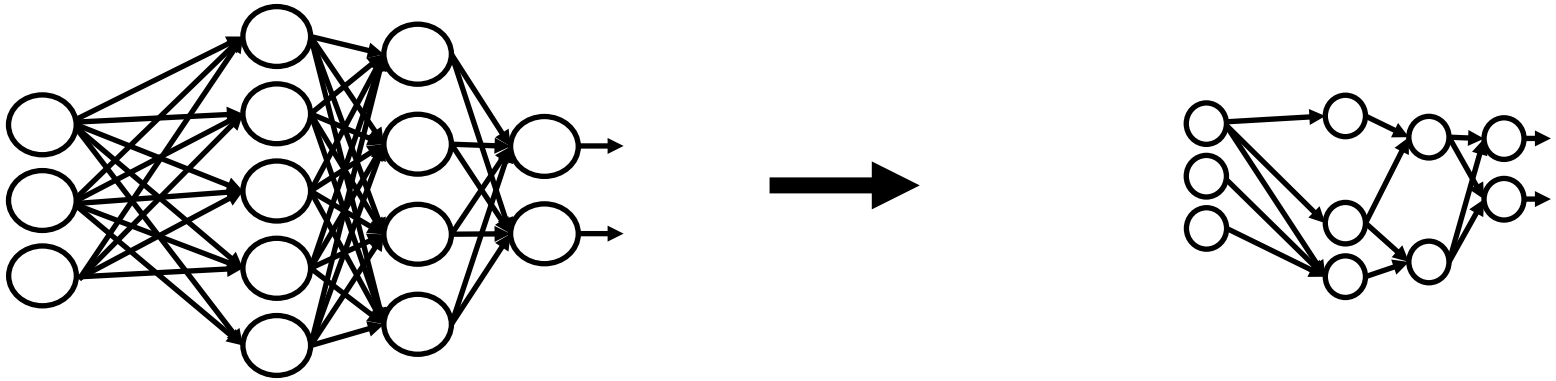


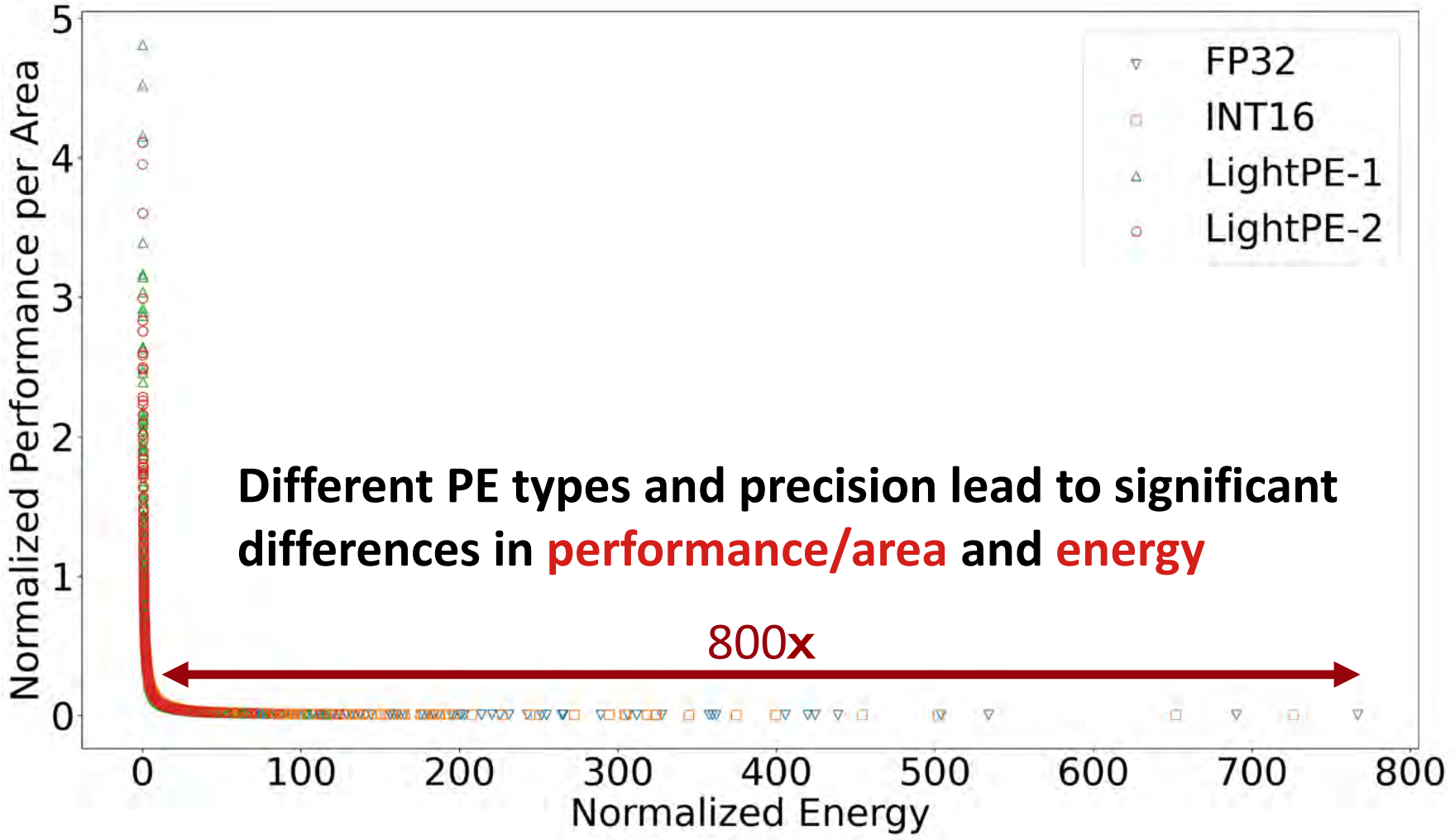
[Source: EIE]



[Source: Google TPU]

- **Model compression**

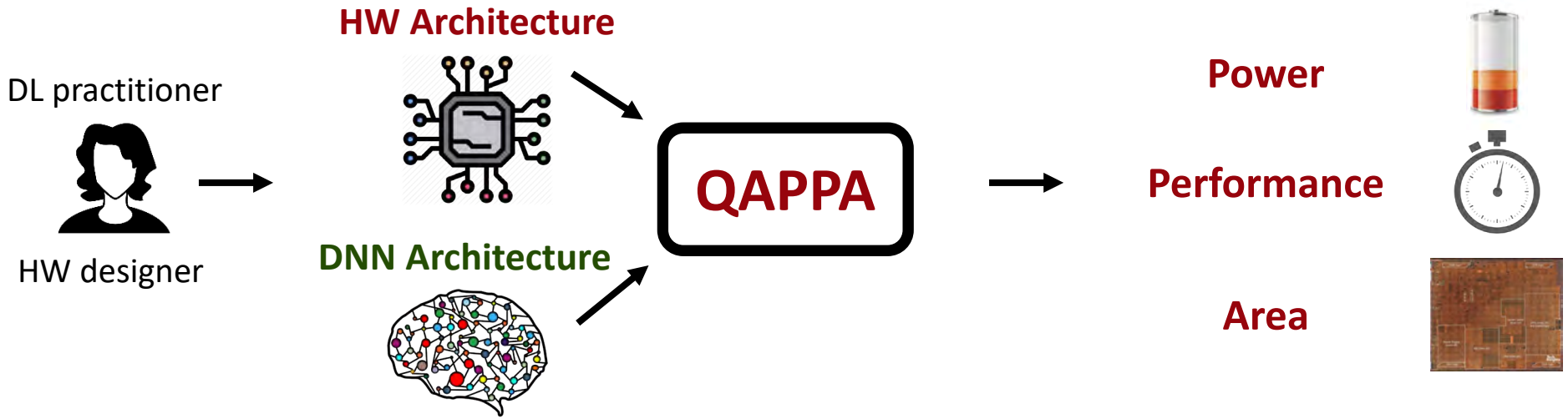




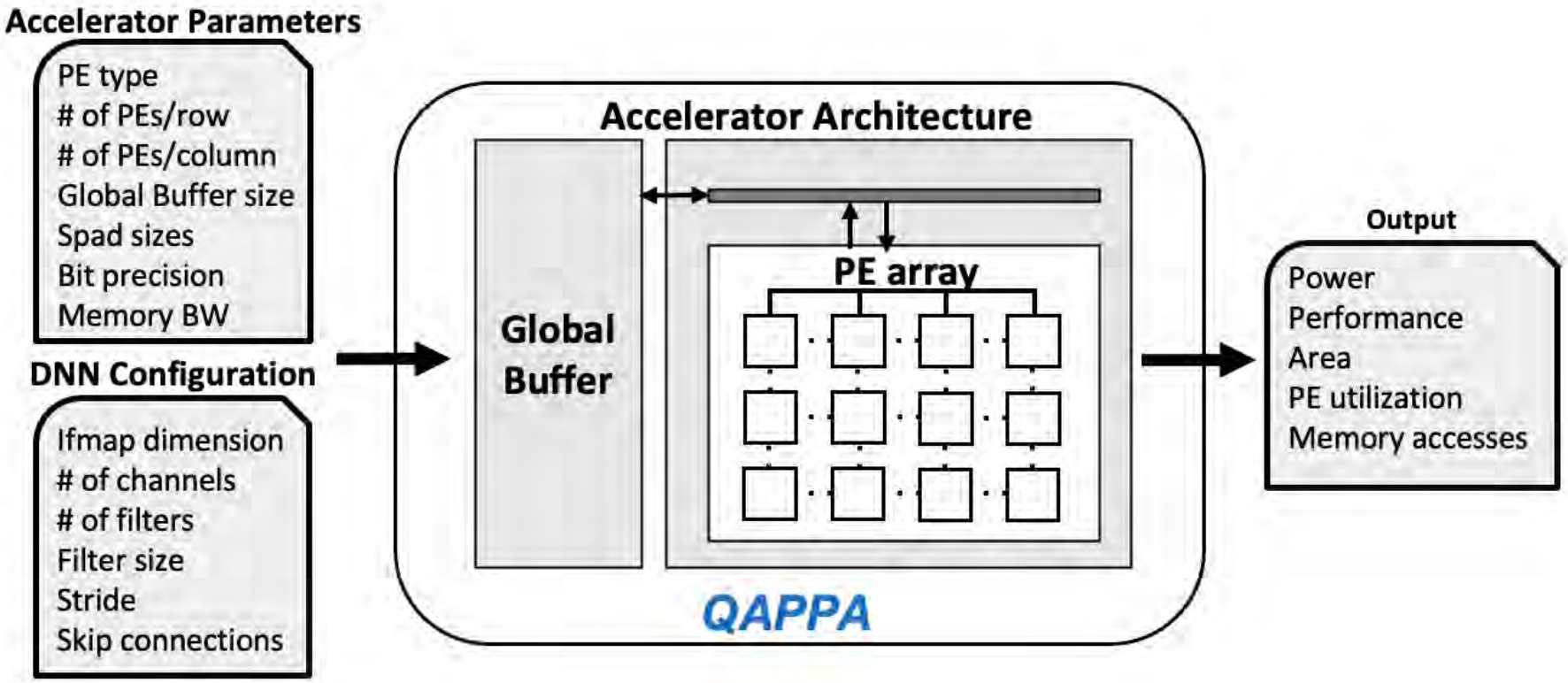
There is a need for a design space exploration framework that incorporates quantization-aware hardware and rapidly iterate over various designs

Our work: *QAPPA*

- **We present *QAPPA*, a quantization-aware power, performance, and area (PPA) modeling framework for DNN accelerators**
  - Highly parameterized framework implemented in RTL for spatial array accelerators
    - No need to have HW expertise to perform design space exploration
    - More optimized and more control in the design as opposed to HLS flow
  - Foster future research on HW/ML model co-design with lightweight processing element (PE)



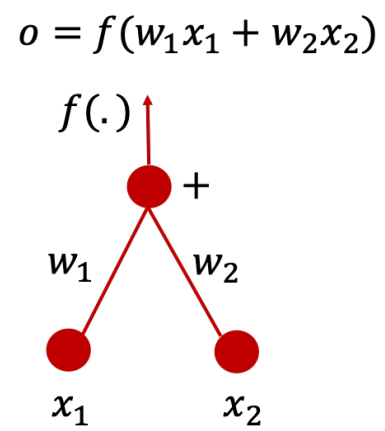
Overview of *QAPPA* framework



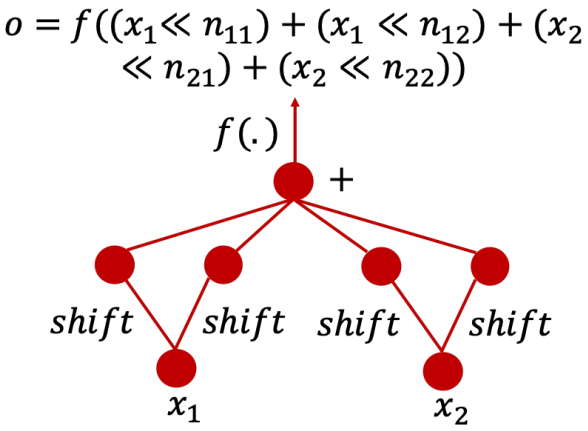
- **We present *QAPPA*, a highly parameterized quantization-aware PPA modeling framework for DNN accelerators**

Methodology

- **Lightweight Processing Elements (LightPE)**
  - $w \cdot x = sign(w)(2^{n_1} + 2^{n_2} + \cdots + 2^{n_K}) \cdot x = sign(w)(x \ll n_1 + \cdots + x \ll n_K)$
  - LightPE-1
    - 4W8A
    - 1 shift and add
  - LightPE-2
    - 8W8A
    - 2 shifts and add



Conventional MAC

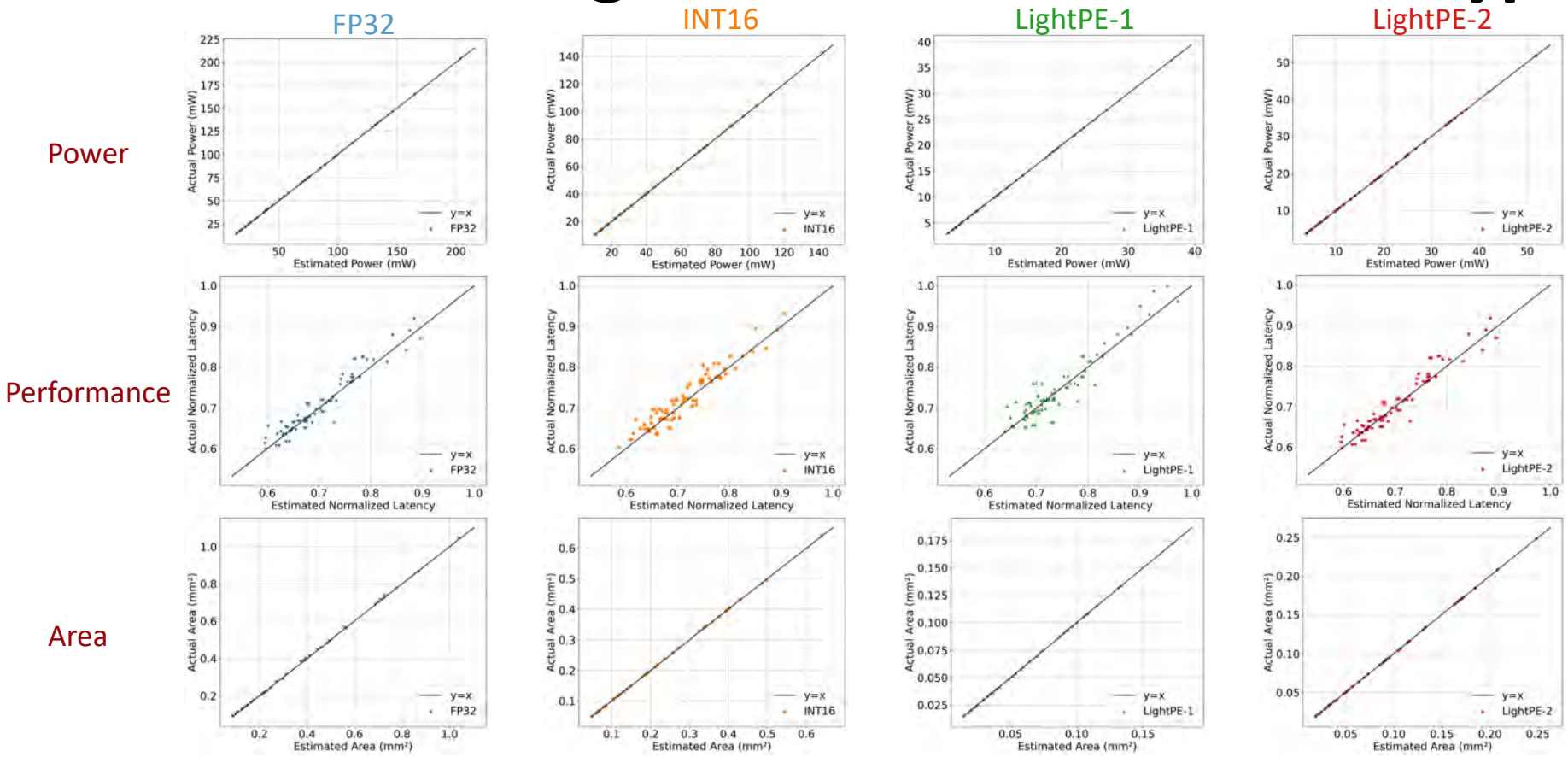


LightPE Shift & Add

[Ding *et al.*, GVLSI'17]

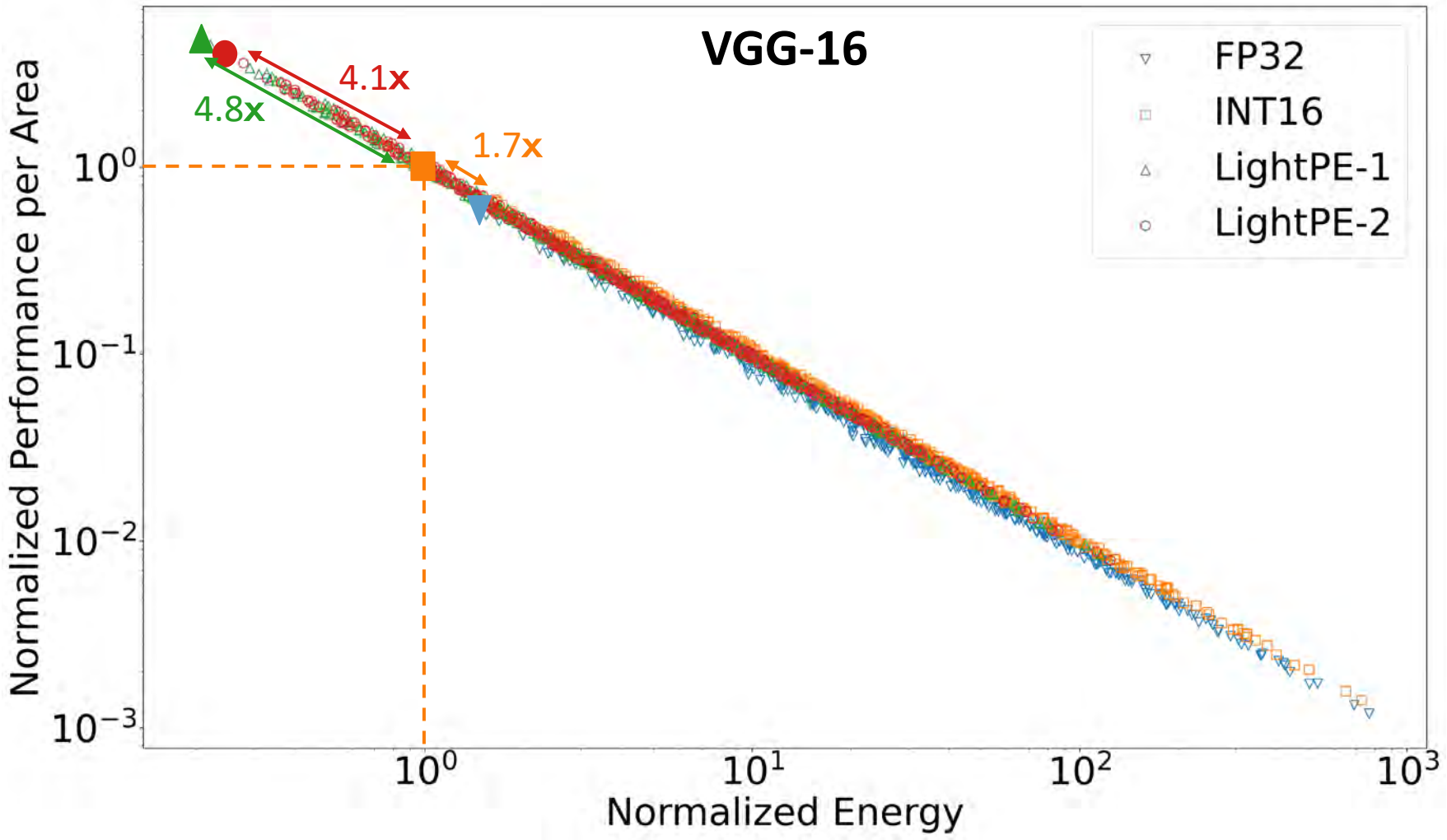
- **PPA Modeling**
  - Synopsys Design Compiler & FreePDK45
  - Synopsys VCS Simulator
  - Synthesizing and determining PPA is expensive
    - Polynomial regression models using *k*-fold cross validation

PPA modeling results for different PE types

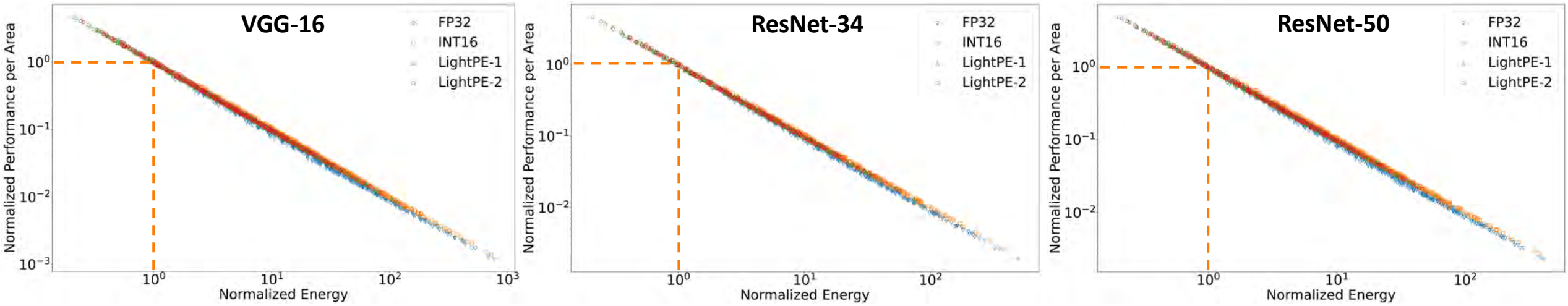


- ***QAPPA*'s PPA models achieve high correlation to the actual PPA values**
  - FP32 implementation has the highest area and power cost whereas LightPEs have the lowest area and power results which shows the hardware-efficiency of LightPEs

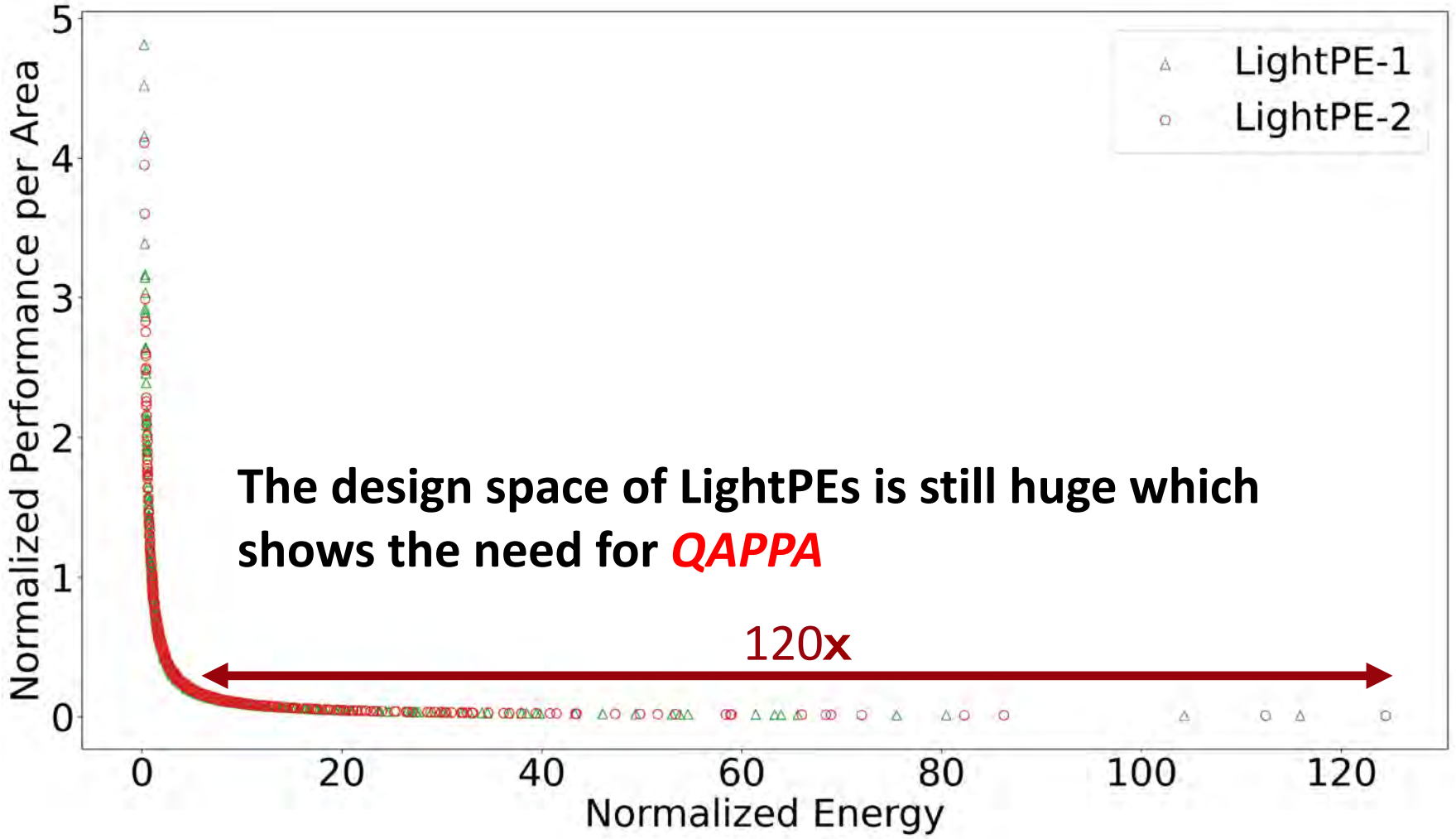
Design space exploration on various DNN models



Design space exploration on various DNN models



- **LightPEs consistently outperform conventional INT16/FP32-based designs**
  - LightPE-1 and LightPE-2 achieve 4.9x and 4.1x more performance/area and 4.9x and 4.2x energy improvement on average when compared to the best INT16 configuration
  - INT16 achieves 1.7x more performance/area and 1.4x energy improvement on average when compared to the best FP32 configuration


Normalized Performance per Area
Normalized Energy
LightPE-1
LightPE-2
0
1
2
3
4
5
0
20
40
60
80
100
120
The design space of LightPEs is still huge which shows the need for QAPPA
120x

Conclusion

- **We present *QAPPA*, a quantization-aware highly parameterized power, performance, and area modeling framework for DNN accelerators**

- **We show that different bit precisions and PE types lead to significant differences in terms of performance/area and energy**
  - ♦ LightPE-1 and LightPE-2 achieve 4.9x and 4.1x more performance/area and 4.9x and 4.2x energy improvement on average when compared to the best INT16 configuration

- **Our novel framework can foster the future research on design space exploration of DNN accelerators for various design choices including quantization-aware PE types, bit precision, and various microarchitectural design choices**

- **Future work:** Include accuracy as an additional exploration metric or optimization objective