MEGA.mini: A NPU with Novel Heterogeneous AI Processing Architecture Balancing Efficiency, Performance, and Intelligence for the Era of Generative AI

Donghyeon Han[1,2] and Anantha P. Chandrakasan[1]
[1]Massachusetts Institute of Technology, Cambridge, MA, USA, [2]Chung-ang University, Seoul, Korea

Lab Website

LinkedIn

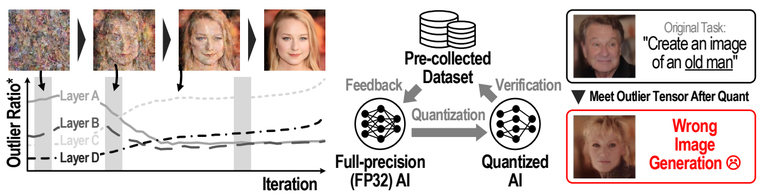
Abstract

- **NPU with a Novel big.LITTLE Core Architecture to Balance 3 Key Aspects of AI Acceleration**
  - Efficiency: > 95% computations w/ Low-precision FXP
  - Performance: 3 hierarchical solutions @ MEGA+mini
  - Intelligence: Hybrid IA (FP for < 5% outlier data)

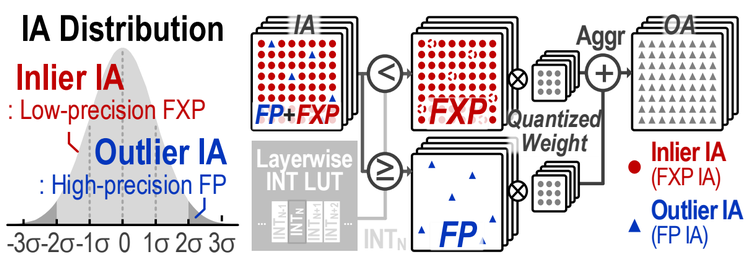
Algorithm	Architecture	Circuit
FXP-FP Hybrid IA	Weight Bit-slice Scalable Arch.	Sign-Magnitude (SM) MAC w/ Dual REG
+ Long / Short-term INT Update	MEGA.mini Arch.	Dynamic FP12 MULT
Delayed Statistics	Cross-shaped Memory Arch.	MEGA+mini Output Synchronizer
Conditional Polynomial Approx.	Tensor Streaming Arch.	Unified PE for Tensor Post-processing

Motivation

- **FXP Quantization: Expected for Efficiency Improvement but Causing Accuracy Drop**
  - Diffusion: Varying integer length for every iterations
  - Non-diffusion: Wrong quantization → **No recovery ☹**



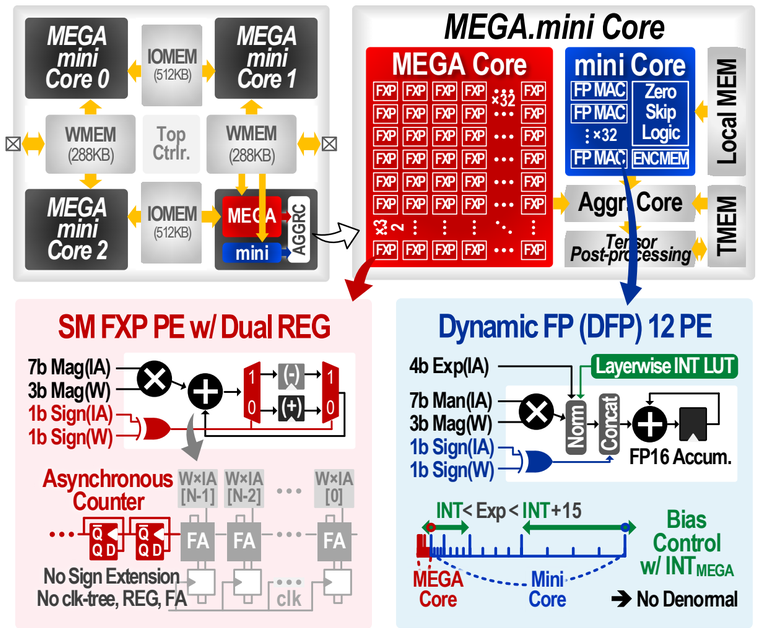
- **Adopting FXP-FP Hybrid IA Representation**
  - Inlier IA w/ FXP + Outlier IA w/ FP → **Accuracy ▲ ☺**



MEGA.mini Core Architecture

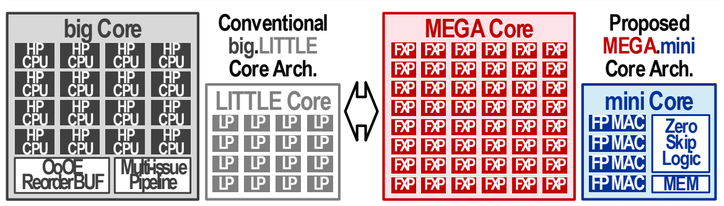
- **Dense FXP Inlier IA → MEGA Core (Date-reuse ▲)**
- **Sparse FP Outlier IA → mini Core (Zero-skip ▲)**

➔ **High Accuracy + 81% Power Reduction**



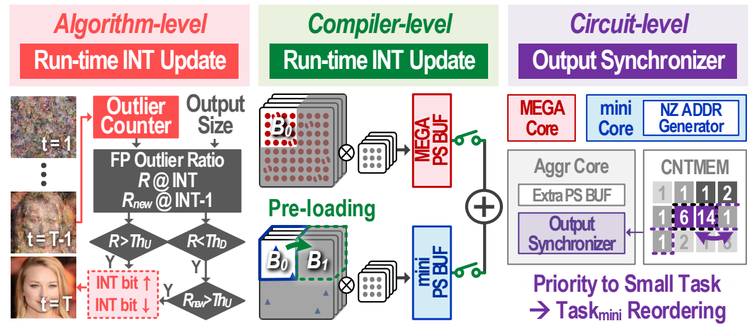
big.LITTLE vs MEGA.mini Core

	big.LITTLE Core		MEGA.mini Core	
Core Type	big	LITTLE	MEGA	mini
Target	CPU		NPU	
Data Format	FP32 / FP64		DFXP8	DFP12/FP16
Feature	Out-or-order Execution	In-order Execution	In-order Execution	Zero-IA Skipping
Complexity	High	Low	Low	High
OPs/cycle*	High	Low	High	Low
J/Ops*	Low	High	High	Low

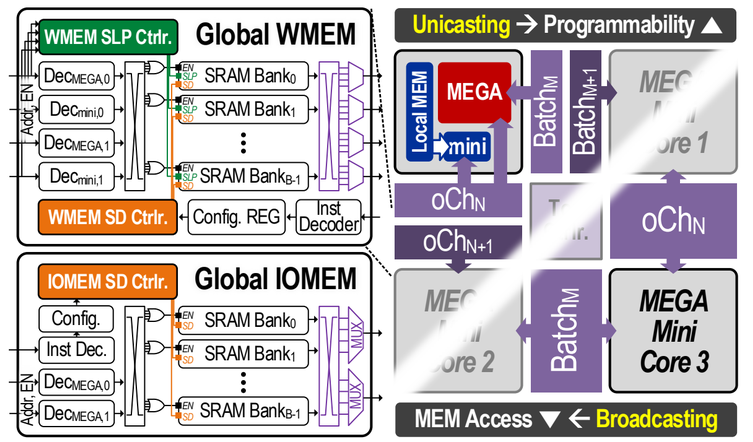


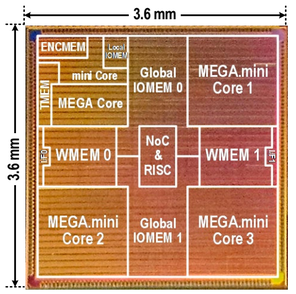
*Computing of Nonzero Operation (Excluding Zero-skipping Effect)

MEGA+mini Workload Balancing



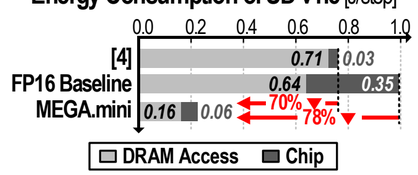
Cross-shaped Memory Architecture





	[1]	[2]	[3]	[4]	[5]	[6]	This Work
Process [nm]	65	12	28	28	22	3	**28**
Generative AI Type	GAN, VAE	NLP	NLP	DM	DM	GAN, VAE, DM	**GAN, VAE, NLP, DM**
Power [mW]	647	122	469	171	279	50.8	**153**
Avg Performance @ Non Diffusion Model (GAN, VAE, NLP)							
TOPS/W	0.83	3.01	7.27	X	X	12.4	**13.4**
TOPS	0.54	0.37	3.41	X	X	0.63	**2.05**
Avg Performance @ Diffusion Model (Stable Diffusion v1.5)							
mJ/step	X	X	X	760	912	435	**224**
ms/step	X	X	X	609	552	1503	**267**

Energy Consumption of SD v1.5 [J/step]



[1] S. Kang et al., ISSCC'20
[2] T. Tambe et al., ISSCC'23
[3] S. Kim et al., ISSCC'24
[4] R. Guo et al., ISSCC'24
[5] Y. Qin et al., ISSCC'25
[6] S.-W. Hsieh et al., ISSCC'25

Advisor (PI): Anantha P. Chandrakasan (anantha@mit.edu)
Presenter: Donghyeon Han (hdh4797@cau.ac.kr)