arm

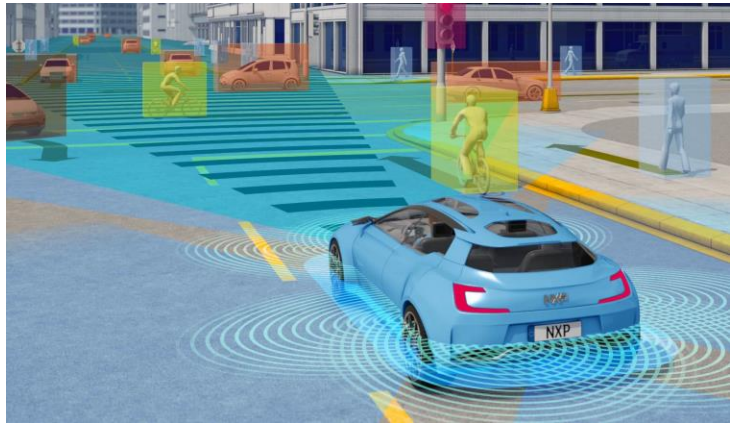
On-Chip Memory Technology Design Space Explorations for Mobile Deep Neural Network Accelerators

Haitong Li[1], Mudit Bhargava[2],
Paul N. Whatmough[2] and H.-S. Philip Wong[1]

[1]Stanford University [2]Arm Research

Machine Learning at Edge

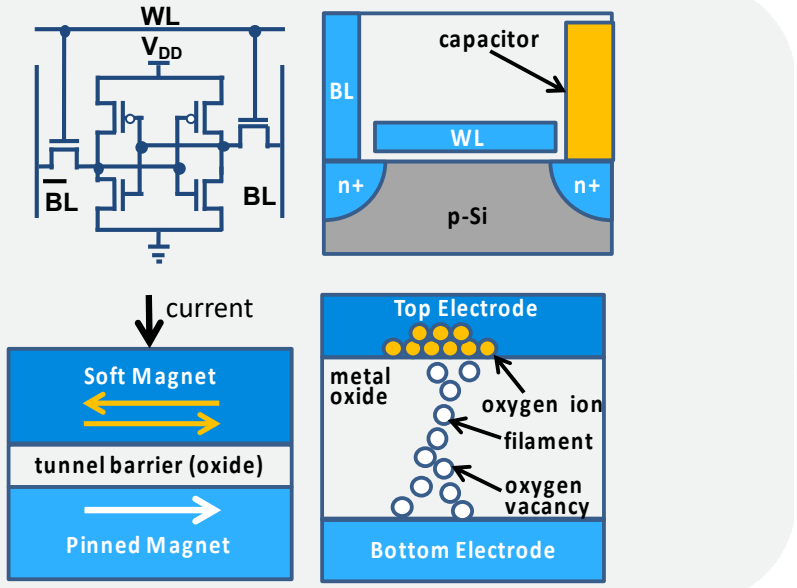
Efficiency & privacy



Industry efforts: Arm ML Processor, Google Edge TPU, Apple Neural Engine, various startups, etc.

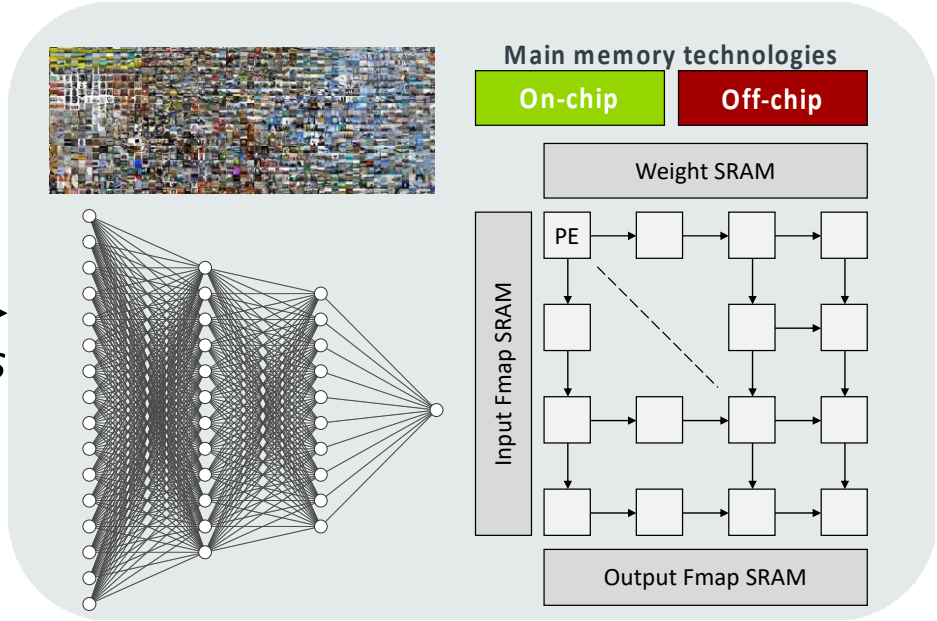
Machine Learning at Edge

- Mobile SoCs: heavily area-constrained, memory resources limited
- Data movement: a critical energy-efficiency bottleneck
- "Augment" DNN accelerators with on-chip NVMs → understand energy-area tradeoffs



Co-evaluate

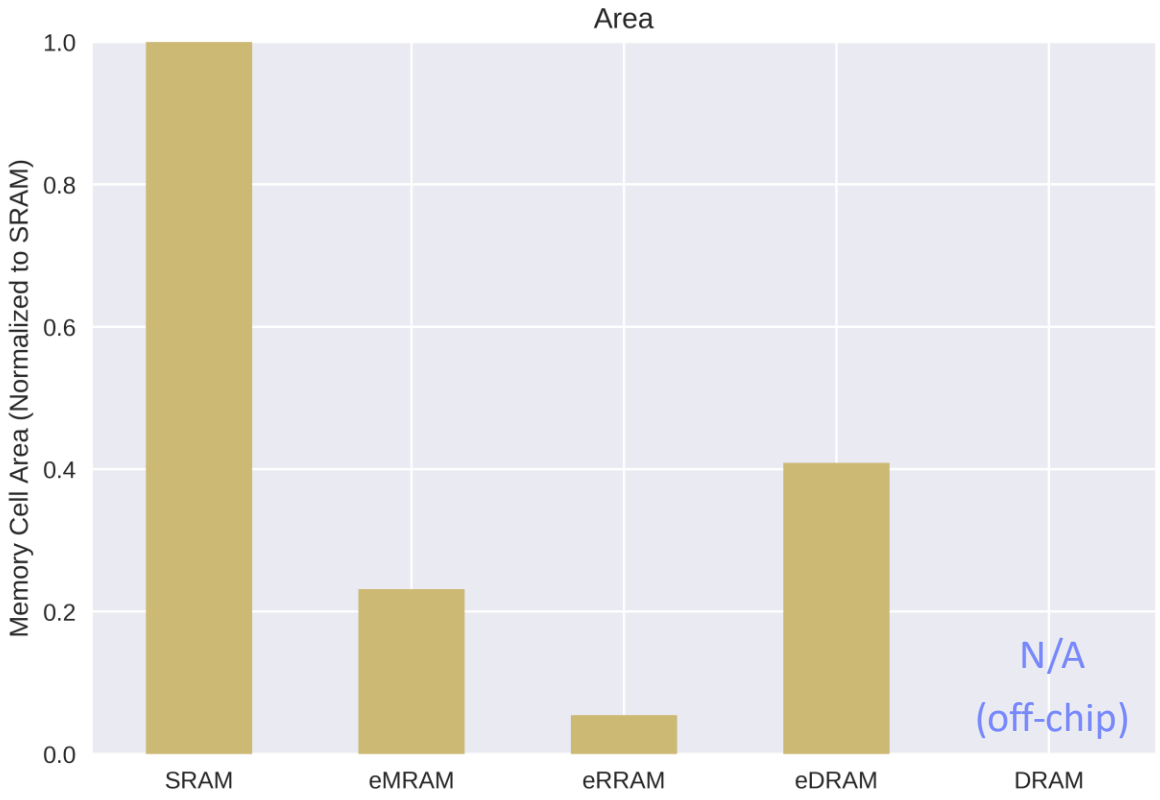
previously disjoint spaces

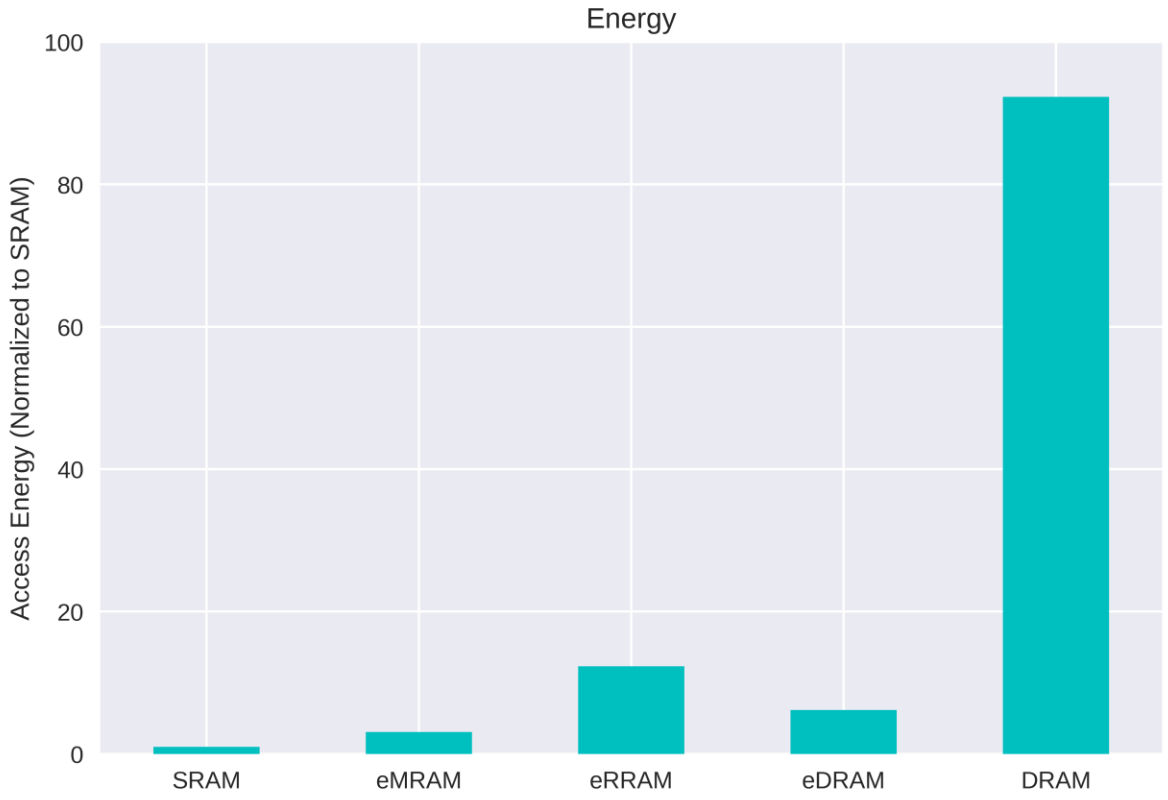


H. Li, M. Bhargava, P. N. Whatmough, H.-S. P. Wong, *DAC*, 2019

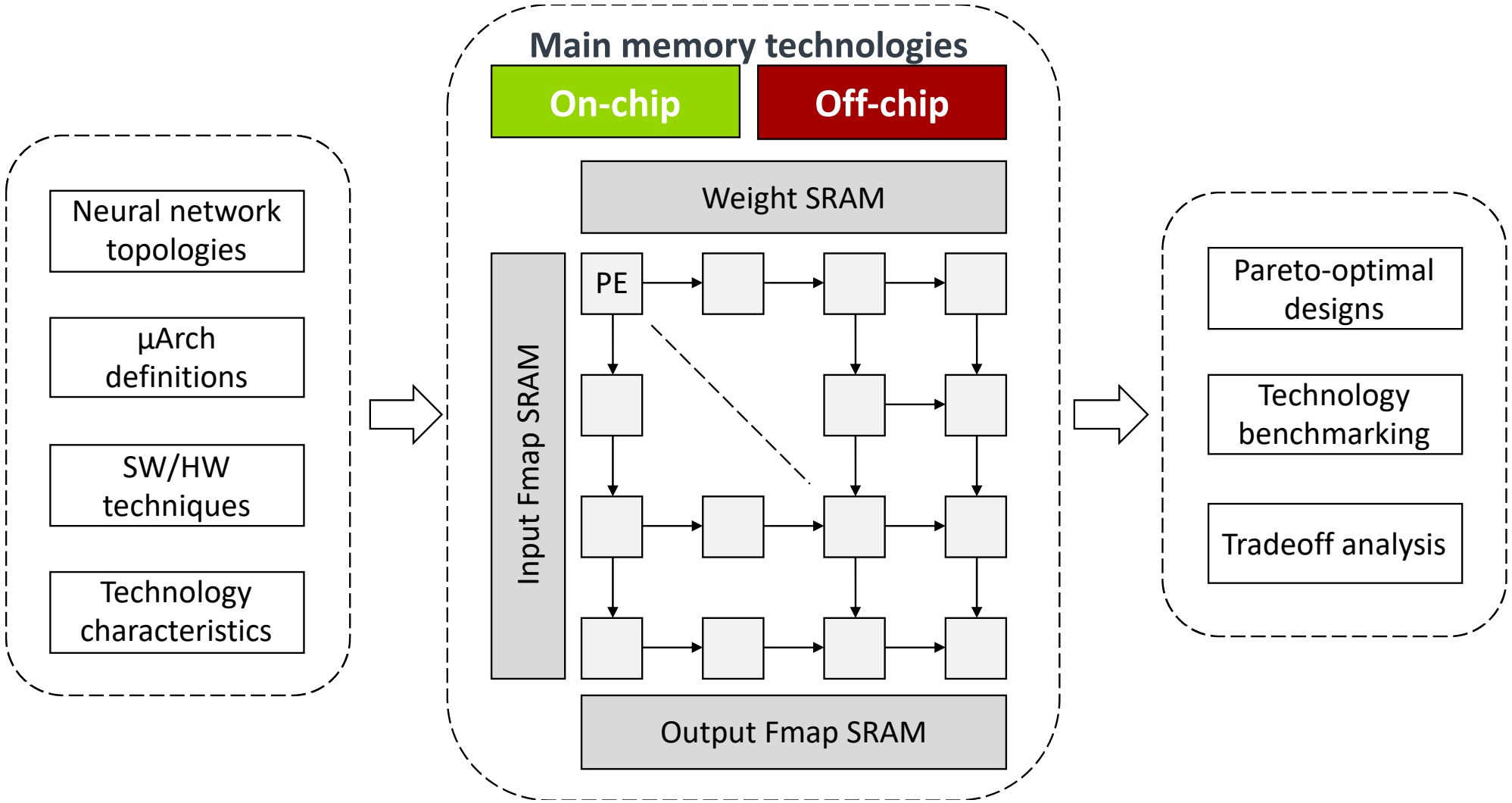
Memory Technology Landscape

- DRAM & eDRAM: destructive read; high refresh power; scalability
- Emerging NVM (e.g., MRAM, RRAM): non-volatile; reads more efficient than writes; 3D integration
- SRAM: bleeding edge, most expensive





Modeling and Evaluation Methodology

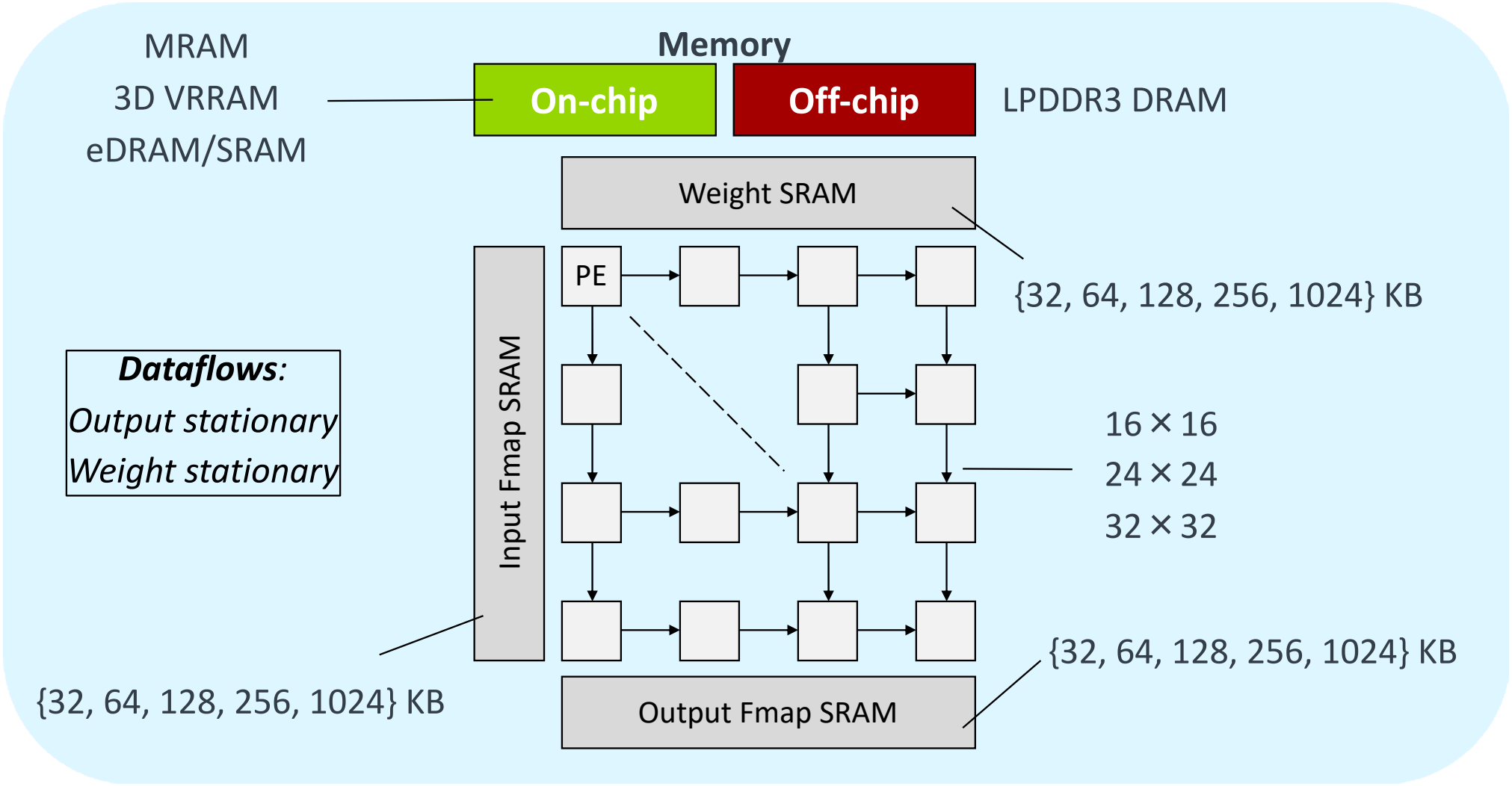


Evaluation framework built upon open-source ***SCALE-Sim*** from Arm Research

https://github.com/ARM-software/SCALE-Sim

Design Space Explorations (DSE)

ResNet-50, GoogLeNet, MobileNet, FasterRCNN, YOLO-tiny



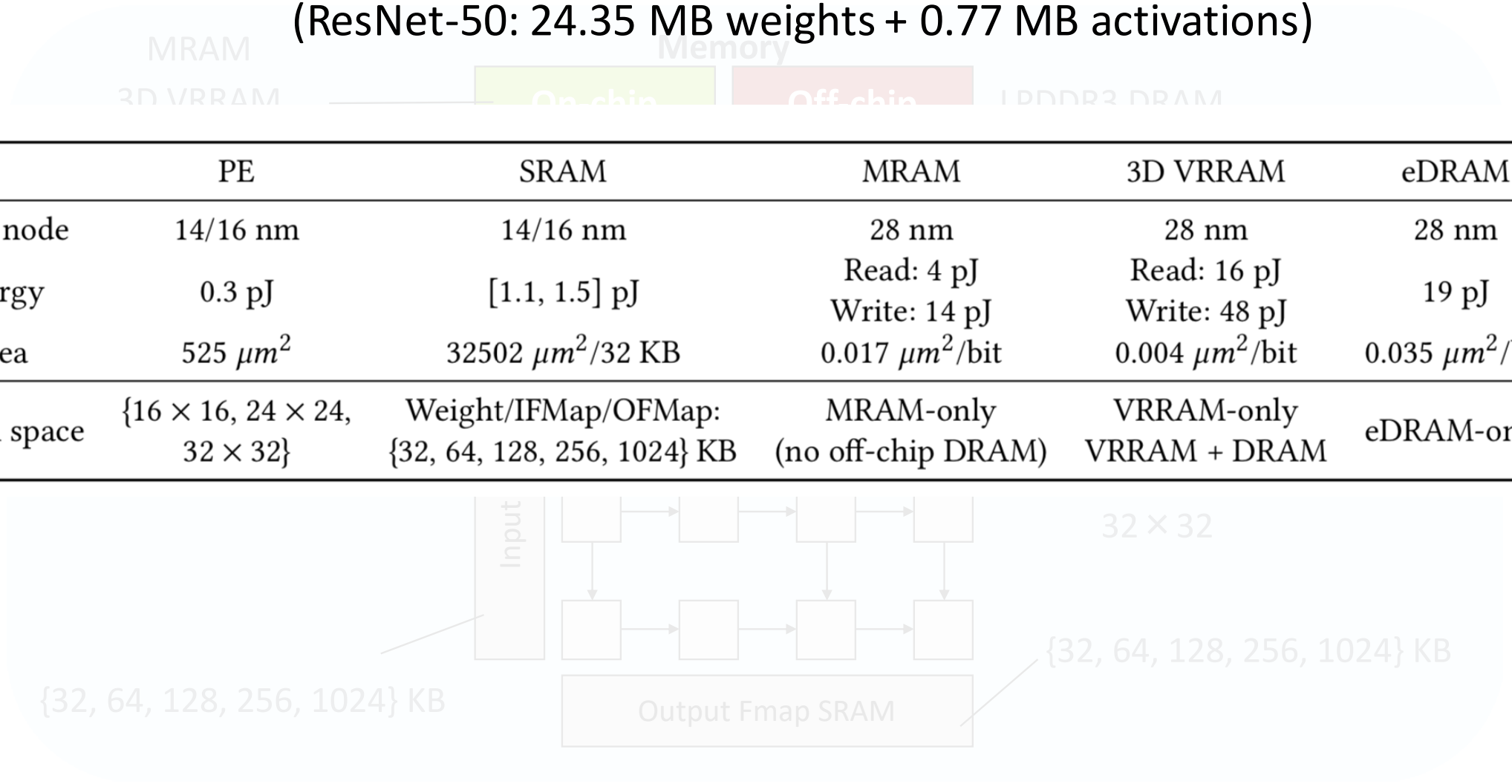
Precision = 8 bit; Clk = 1 GHz; DRAM latency hidden (pipelining)

Design Space Explorations (DSE)

ResNet-50, GoogLeNet, **MobileNet**, **FasterRCNN**, YOLO-tiny

(ResNet-50: 24.35 MB weights + 0.77 MB activations)

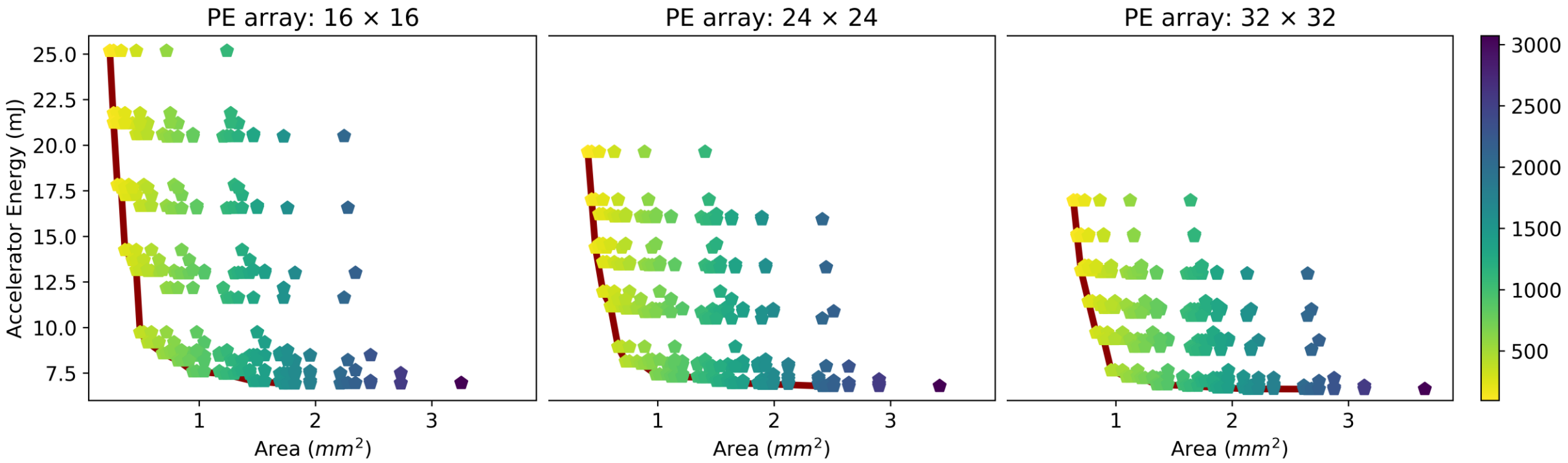
	PE	SRAM	MRAM	3D VRRAM	eDRAM	DRAM
Tech. node	14/16 nm	14/16 nm	28 nm	28 nm	28 nm	28 nm
Energy	0.3 pJ	[1.1, 1.5] pJ	Read: 4 pJ Write: 14 pJ	Read: 16 pJ Write: 48 pJ	19 pJ	120 pJ
Area	525 μm^2	32502 μm^2/32 KB	0.017 μm^2/bit	0.004 μm^2/bit	0.035 μm^2/bit	N/A
Design space	{16 × 16, 24 × 24, 32 × 32}	Weight/IFMap/OFMap: {32, 64, 128, 256, 1024} KB	MRAM-only (no off-chip DRAM)	VRRAM-only VRRAM + DRAM	eDRAM-only	LPDDR3



Details can be found in: H. Li, M. Bhargava, P. N. Whatmough, H.-S. P. Wong, *DAC*, 2019

Accelerator Energy-Area Pareto Frontiers

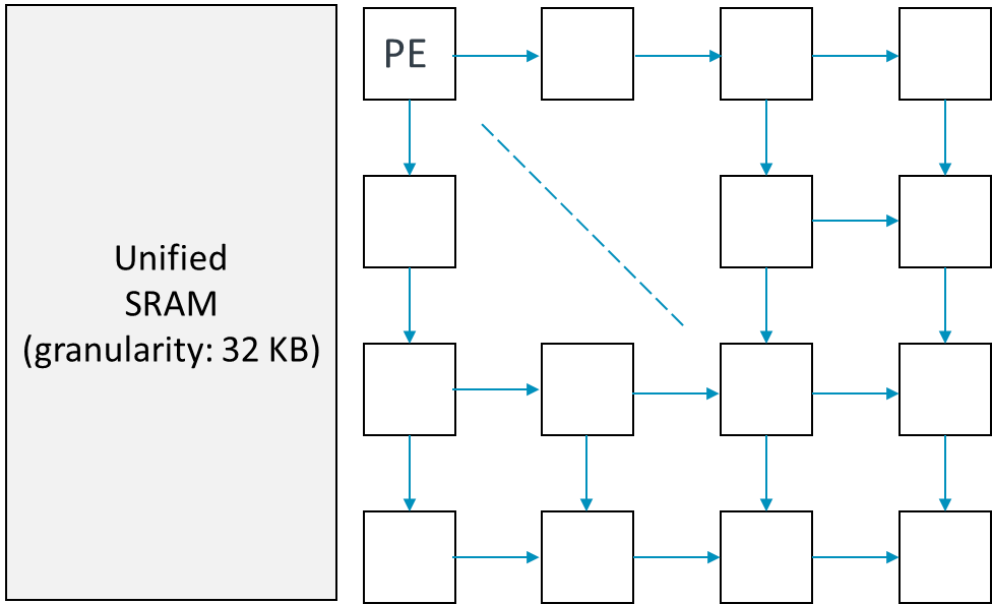
ResNet-50, output stationary, DRAM baselines

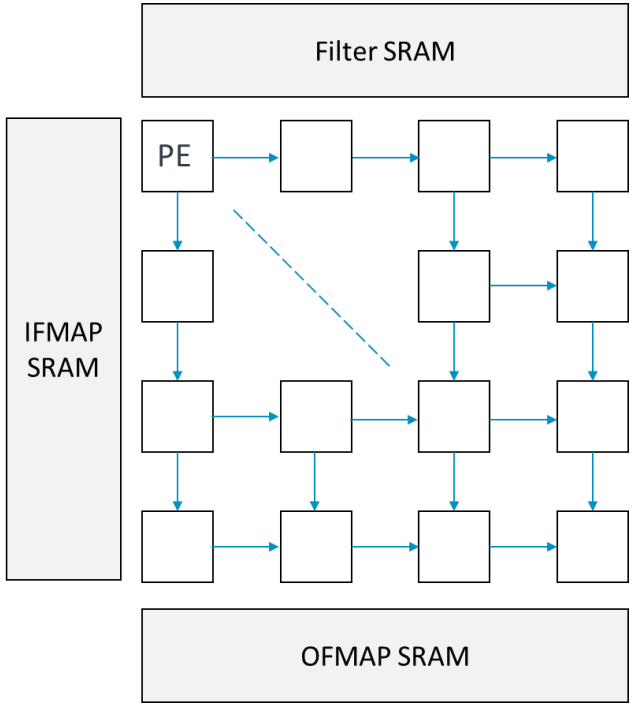


colormap represents total SRAM capacity in a specific design w/ off-chip DRAM

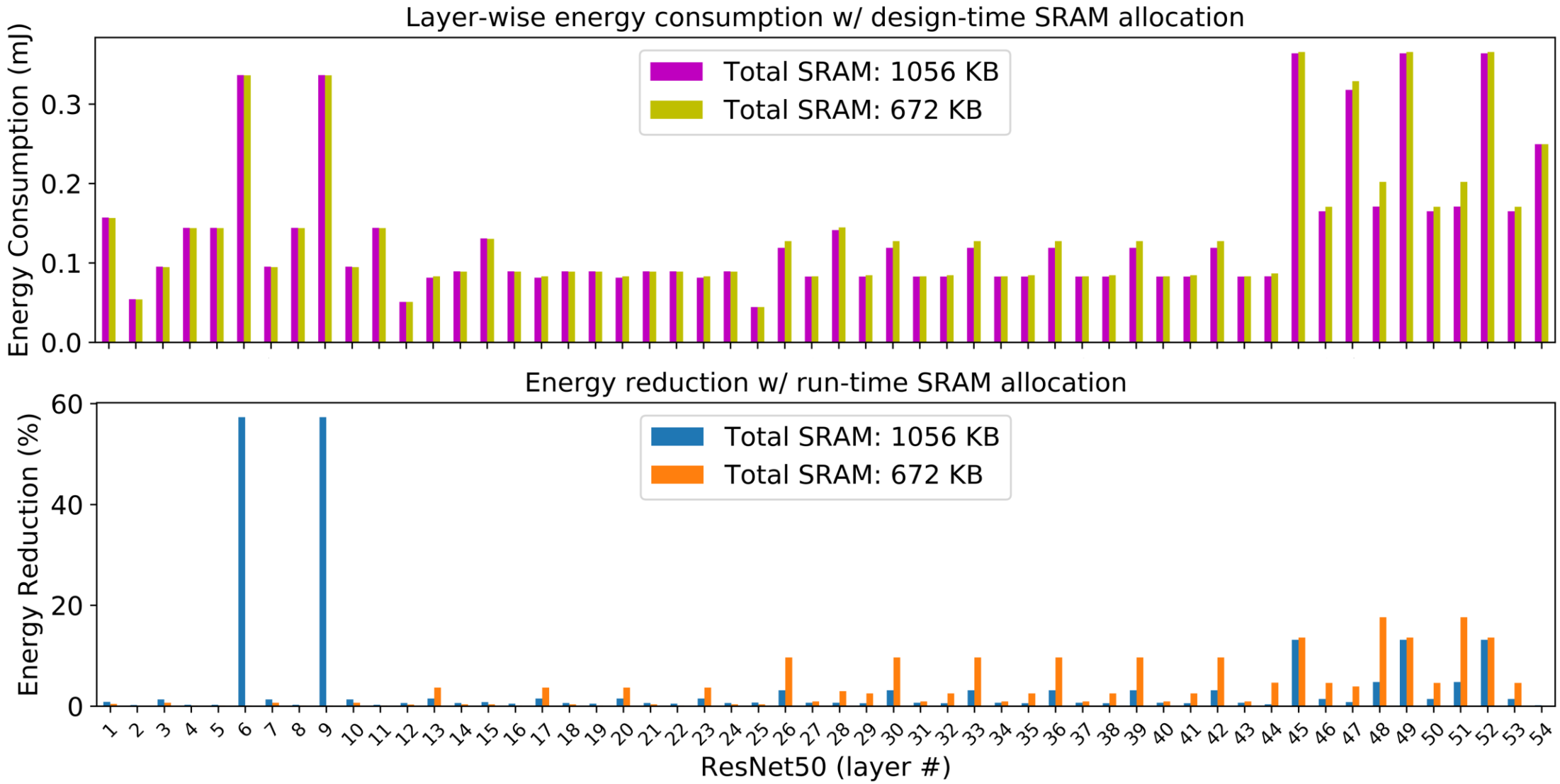
Design-Time vs. Run-Time SRAM Allocation

- Compare representative pareto-optimal designs (same SRAM capacity)
  - Layer-wise *vs.* network-wise optimization & allocation
- Granularity = 32 KB
  - Minimum block size to allocate for IFMap, OFMap, and filter weights

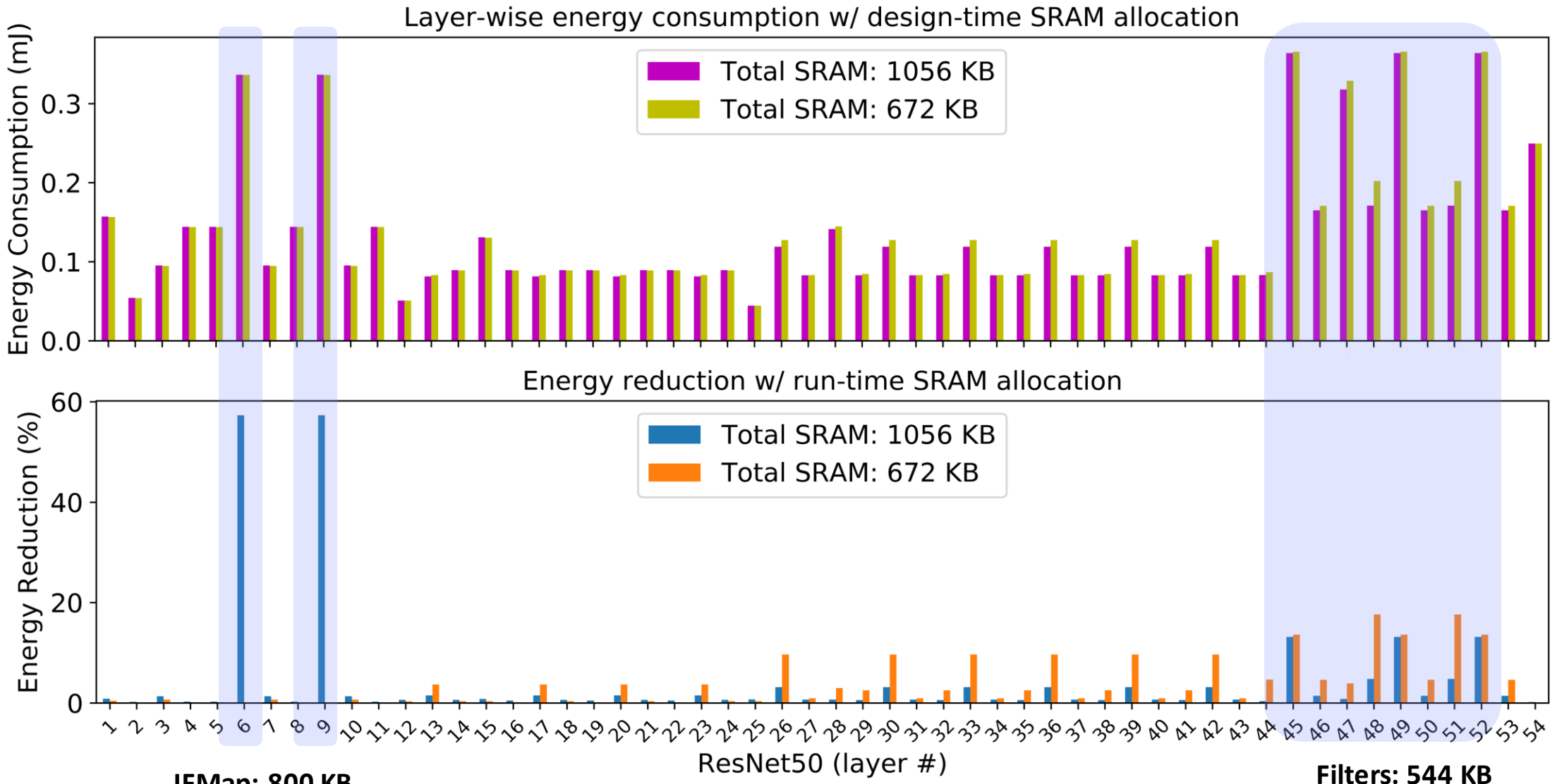




Design-Time vs. Run-Time SRAM Allocation



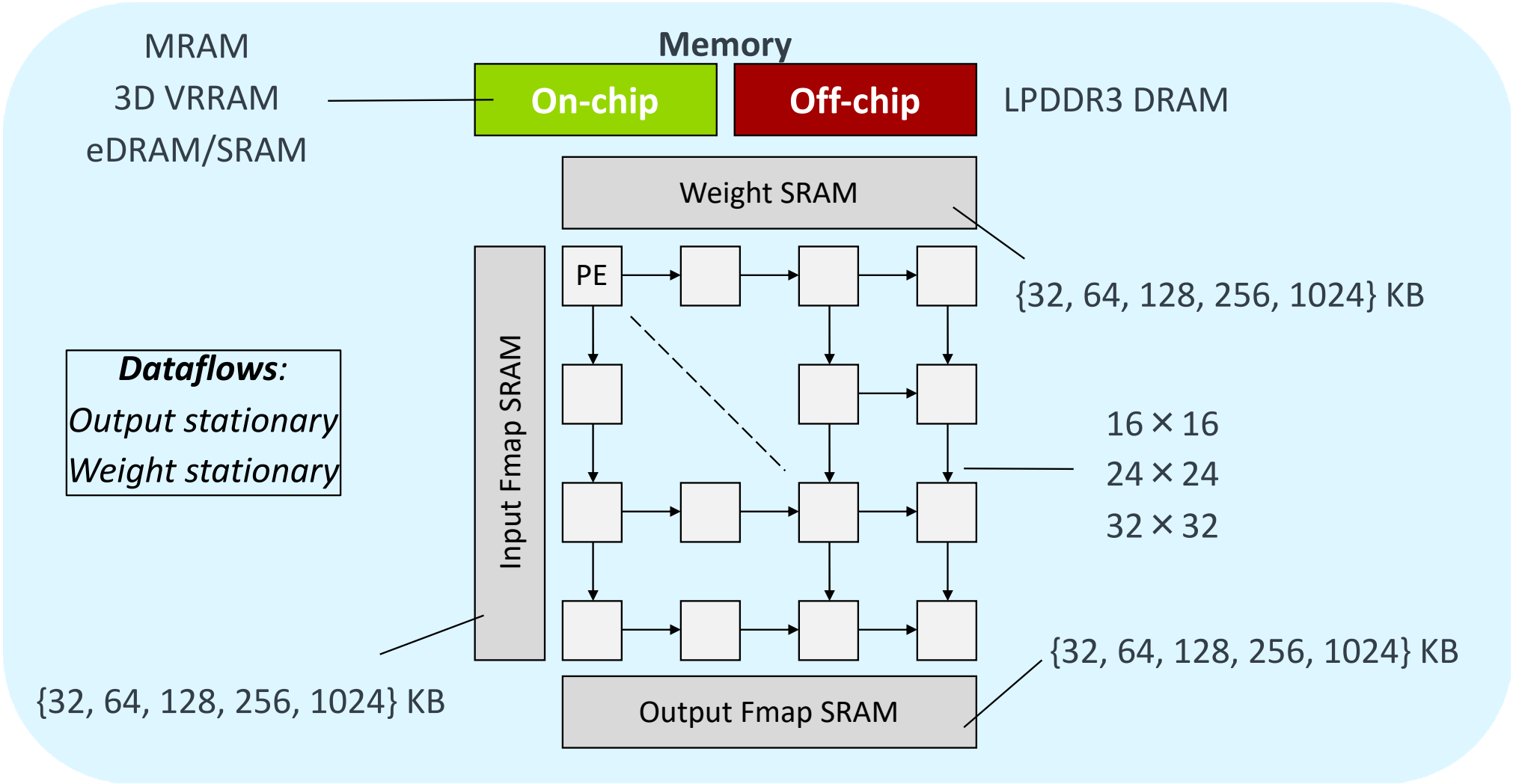
Design-Time vs. Run-Time SRAM Allocation



IFMap: 800 KB
Filters: 32 KB; OFMap: 224 KB

Filters: 544 KB
IFMap: 32 KB; OFMap: 480 KB

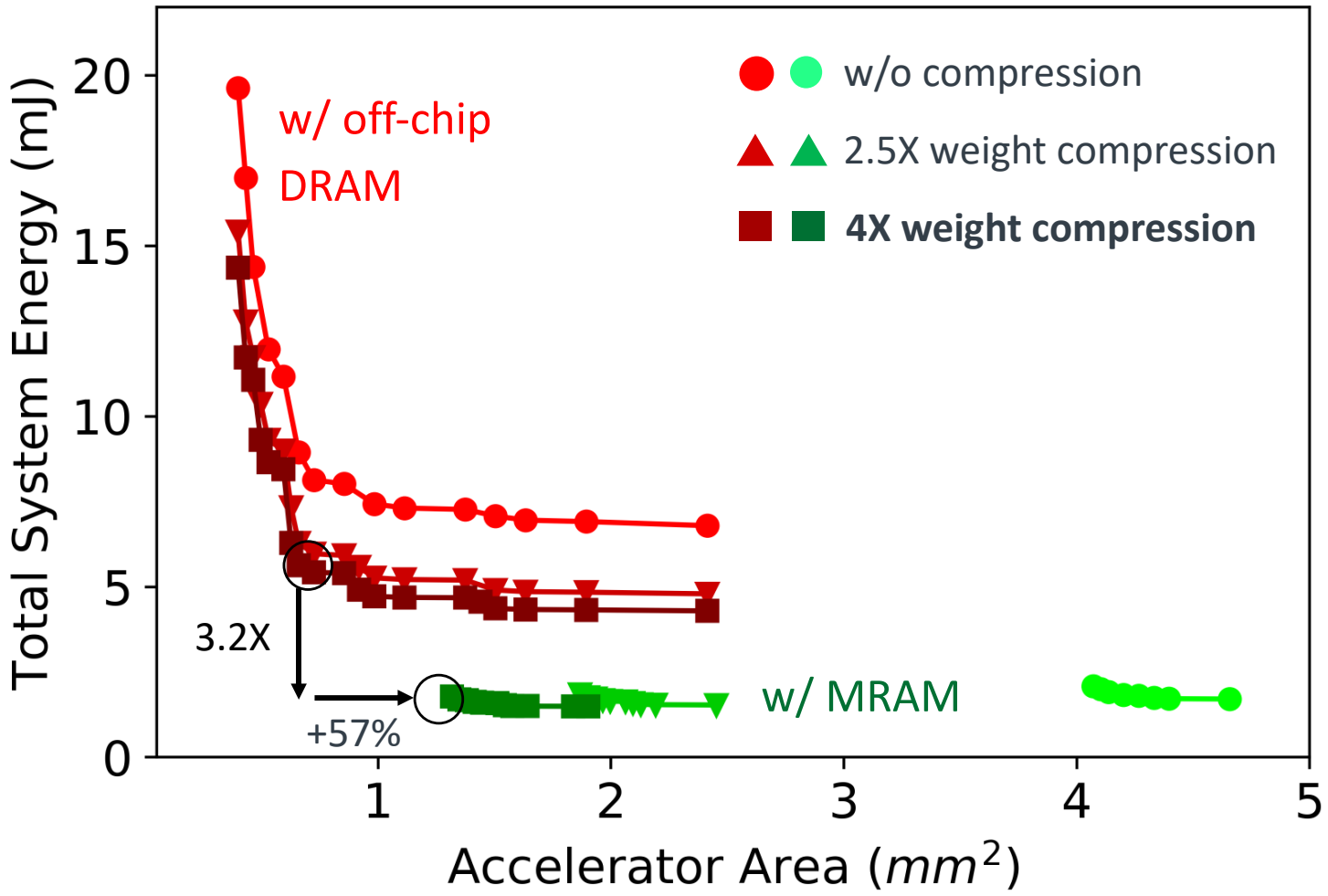
Recap



Precision = 8 bit; Clk = 1 GHz; DRAM latency hidden (pipelining)

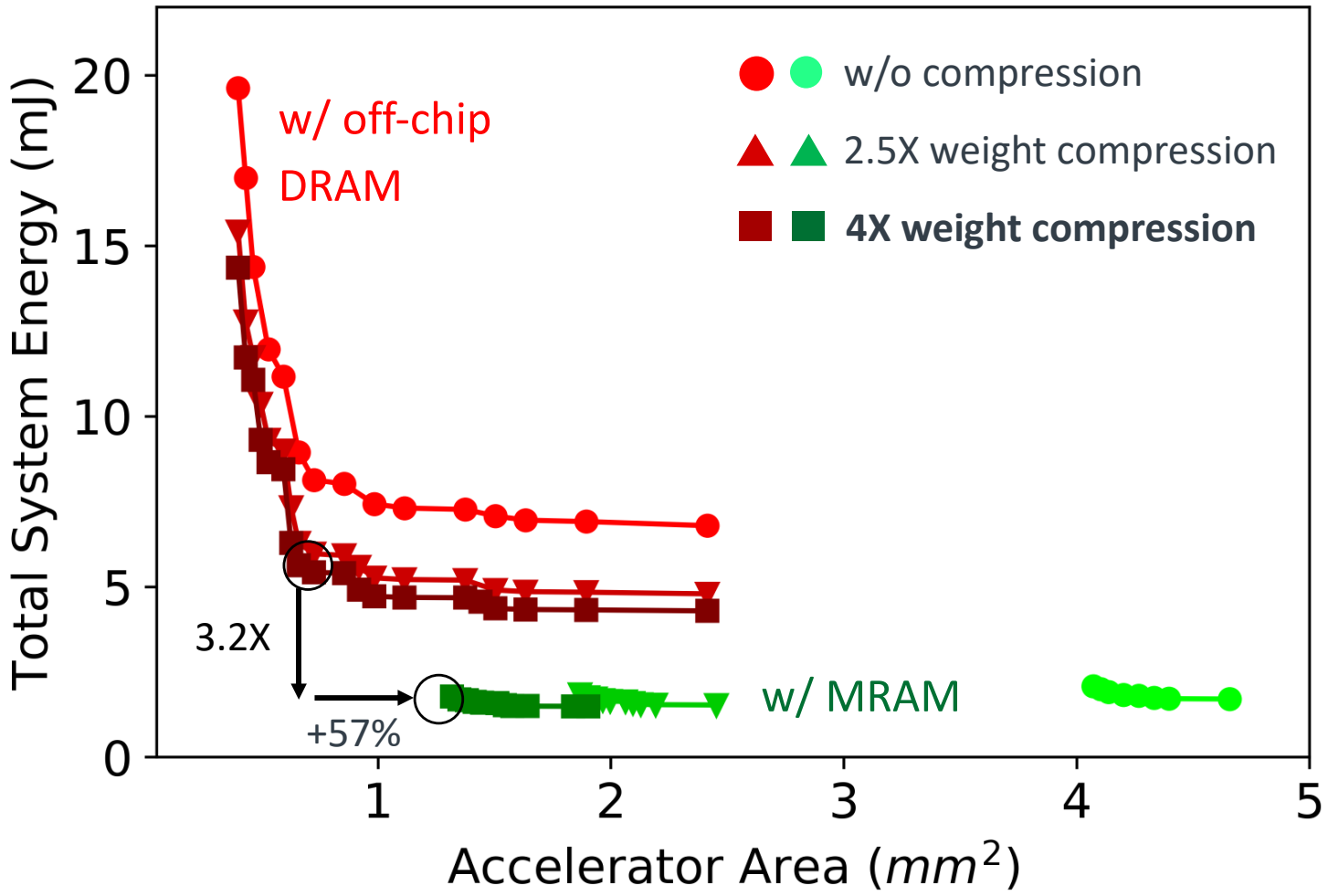
Energy-Area Tradeoffs with MRAM

ResNet-50 | output stationary | PE size = 24 × 24



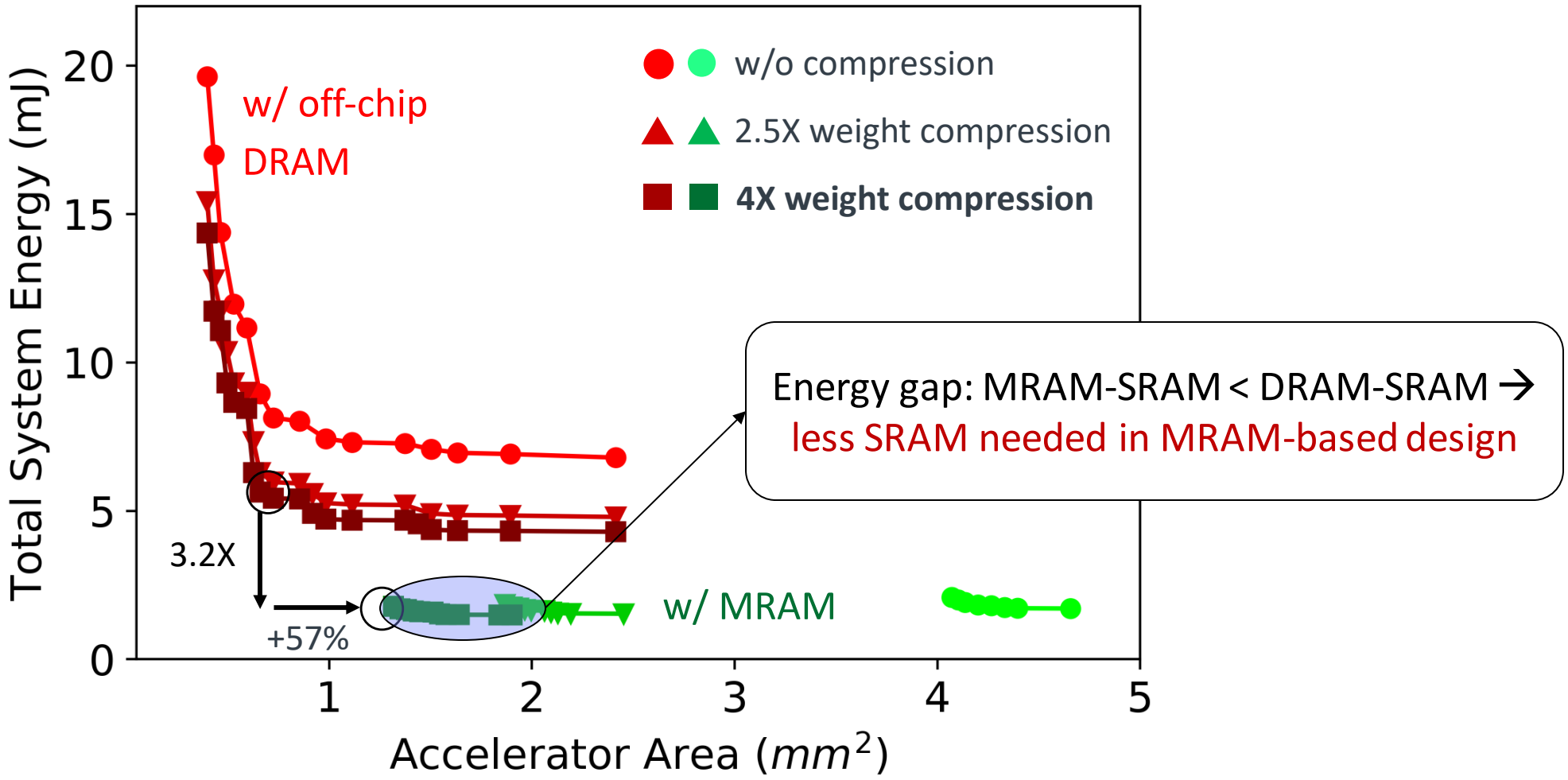
Energy-Area Tradeoffs with MRAM

3.2X energy benefits | 57% area increase



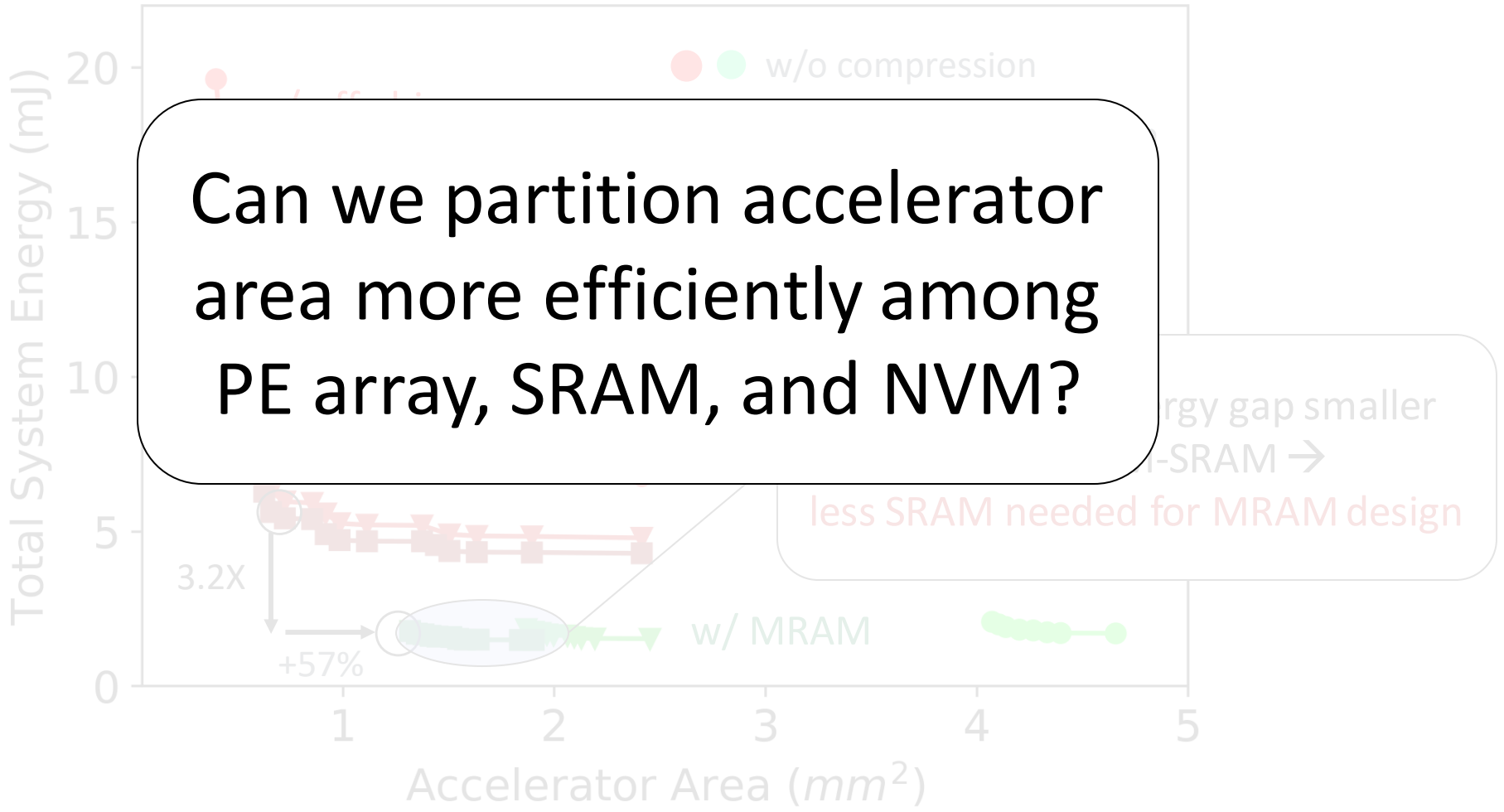
Energy-Area Tradeoffs with MRAM

3.2X energy benefits | 57% area increase



Energy-Area Tradeoffs with MRAM

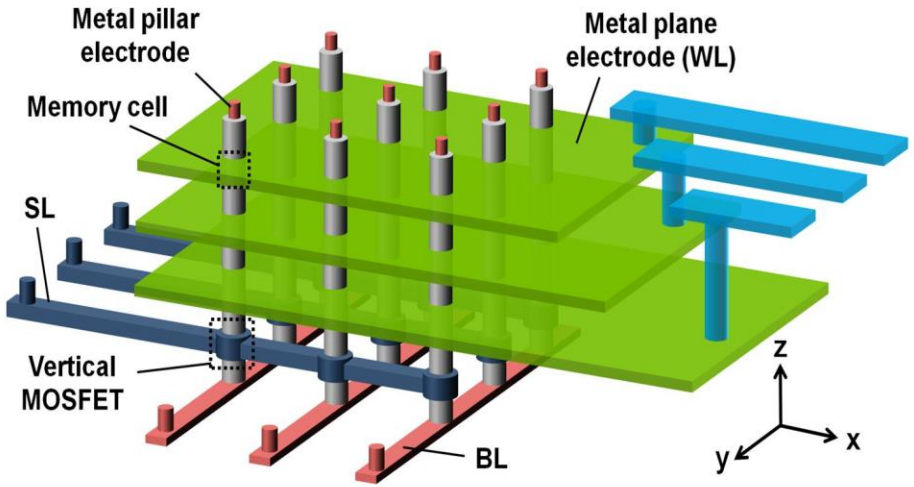
3.2X energy gains | 57% area increase

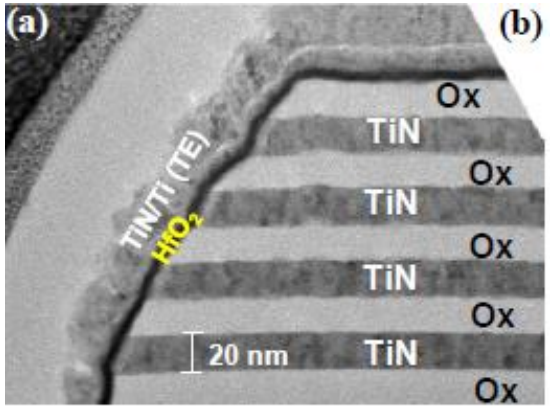


Can we partition accelerator area more efficiently among PE array, SRAM, and NVM?

rgy gap smaller
-SRAM →
less SRAM needed for MRAM design

Towards Higher Density: 3D Vertical RRAM



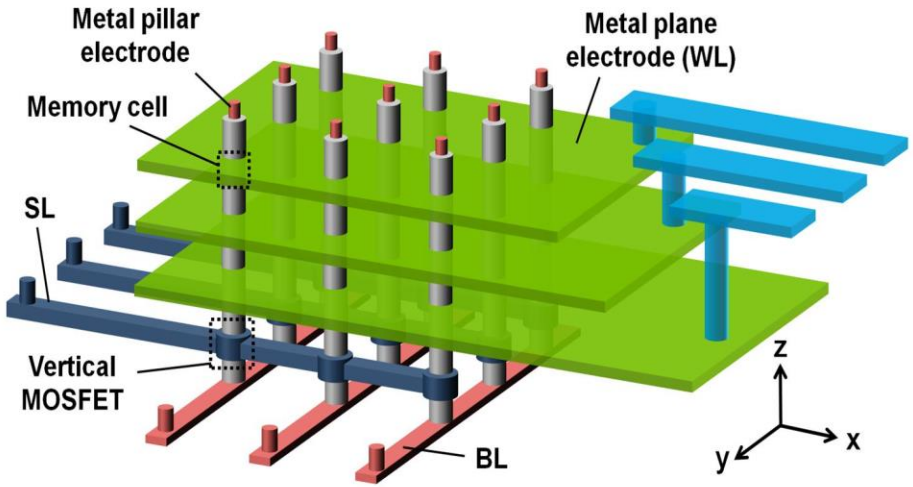


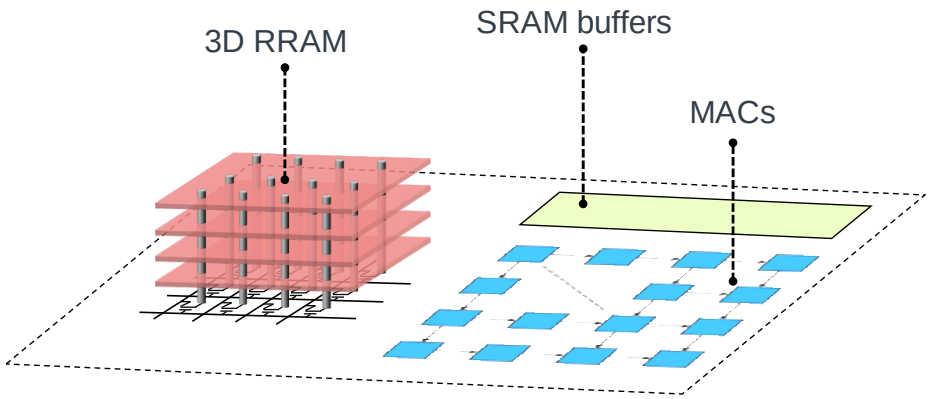
- Array structure similar to 3D V-NAND
- more layers → bit area ↓ & bit cost ↓
- Today: energy/bit higher than MRAM

F. K. Hsueh *et al.*, *IEDM,* 2017

H. Li *et al.*, *VLSI*, 2016

Architecture Implications



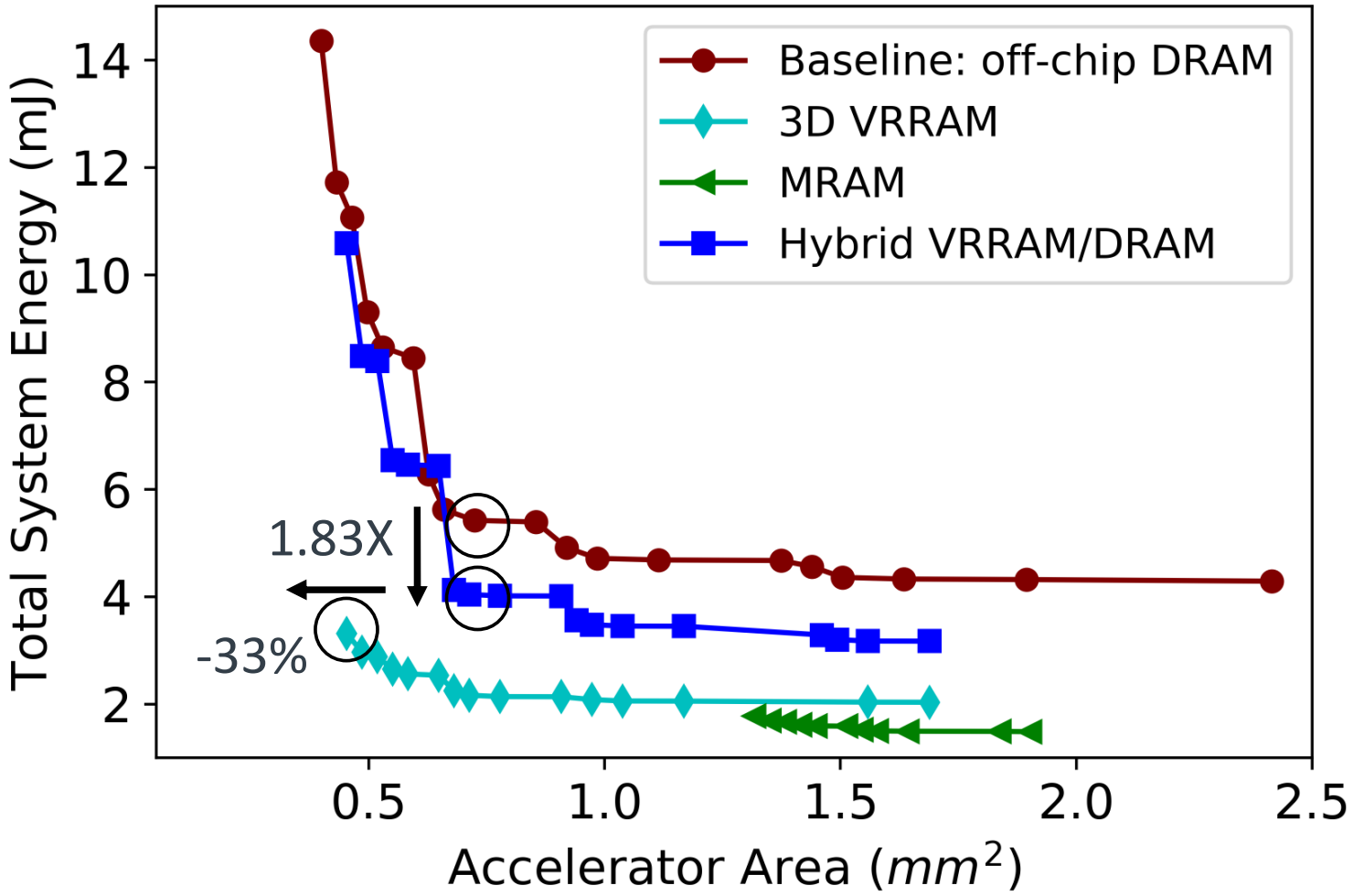


- IFMap/OFMap/Weights → VRRAM
  - Endurance resilience required [1], [2]
- Weights → VRRAM; IFMap/OFMap → DRAM
  - NVM for weight storage only
- Ultra dense due to 3D, then what?

[1] M. M. A. Sabry, et al., Proc. IEEE, 2019
[2] T. Wu, et al., ISSCC, 2019

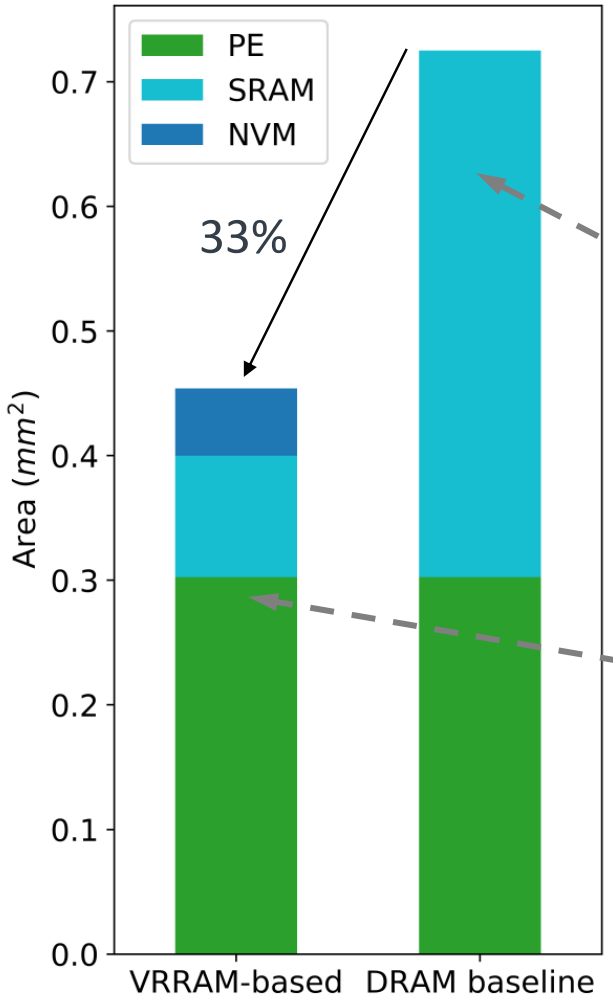
Energy-Area Tradeoffs with 3D VRRAM

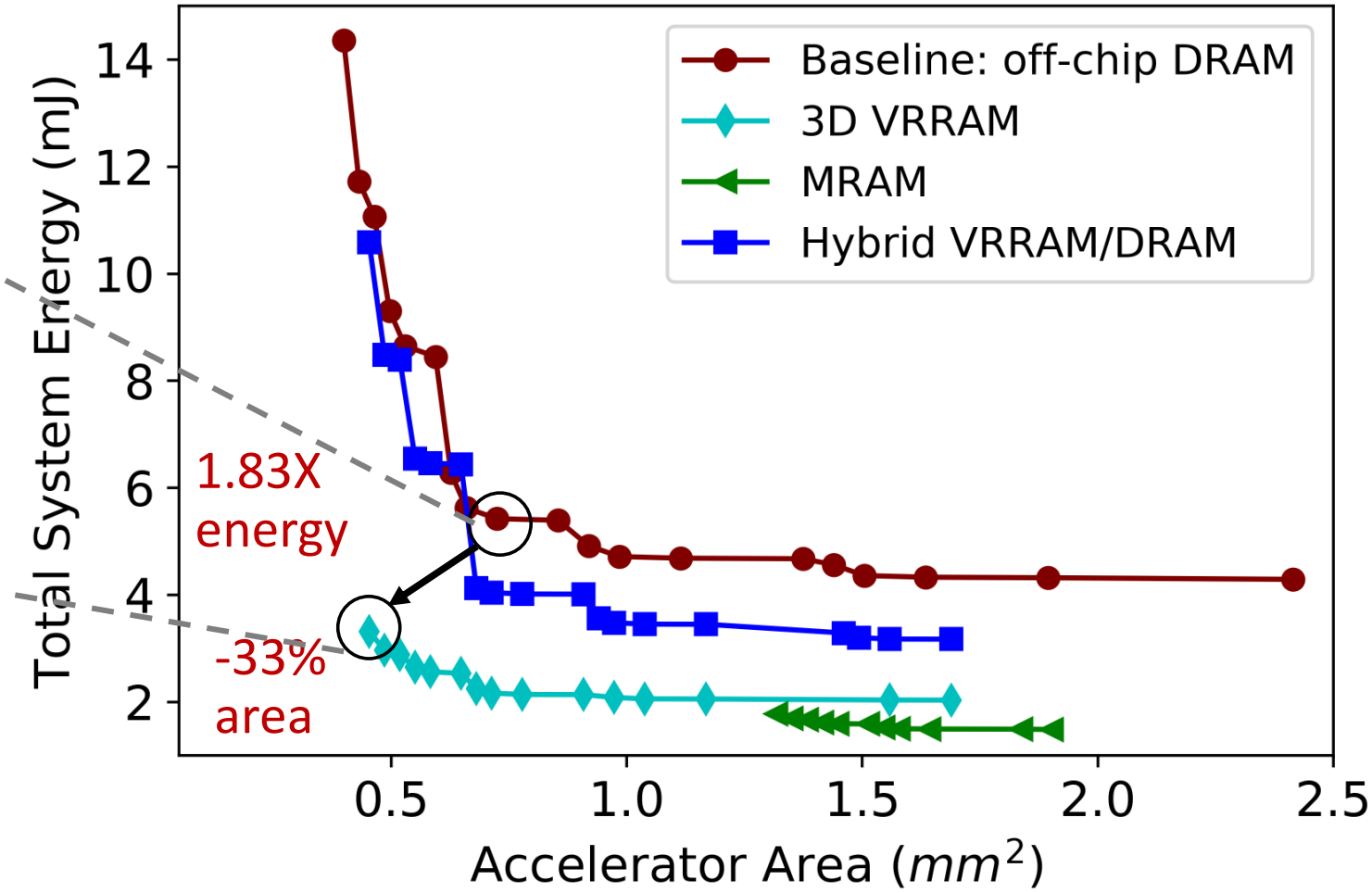
1.83X energy benefits | 33% area savings



Energy-Area Tradeoffs with 3D VRRAM

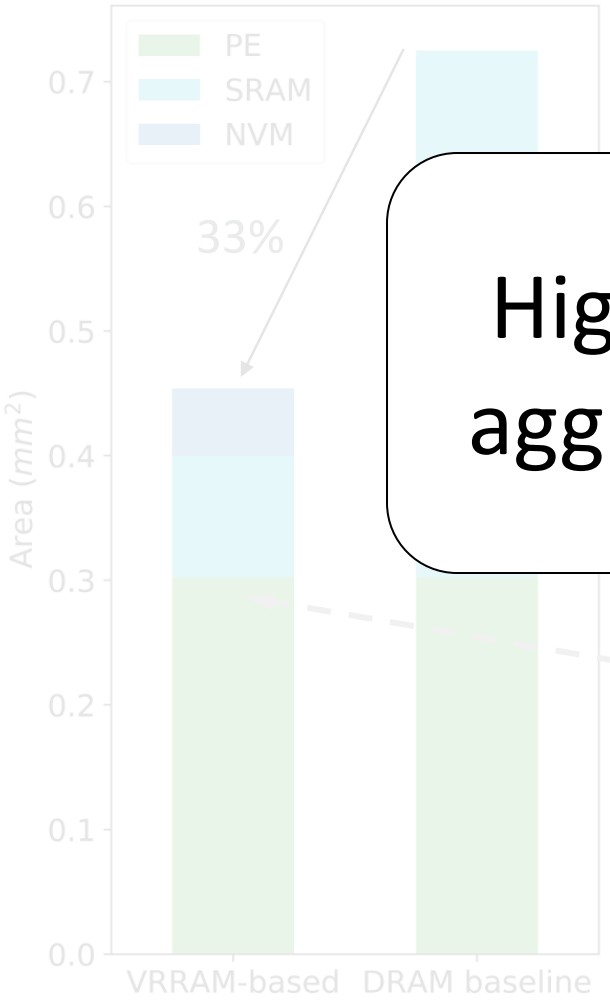
3D RRAM-based design vs. DRAM baseline:
Accelerator area savings due to 4X less SRAM required

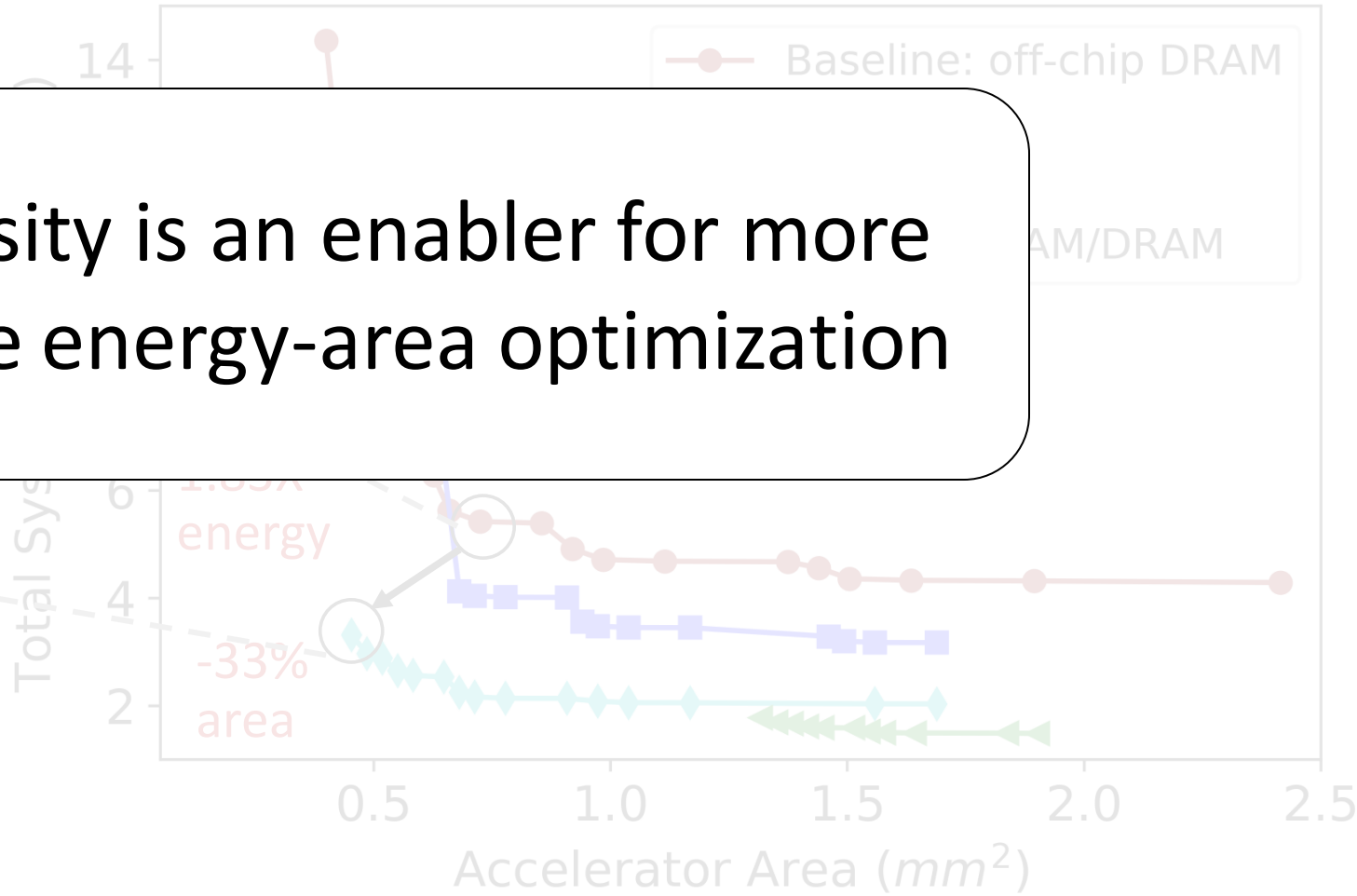




Energy-Area Tradeoffs with 3D VRRAM

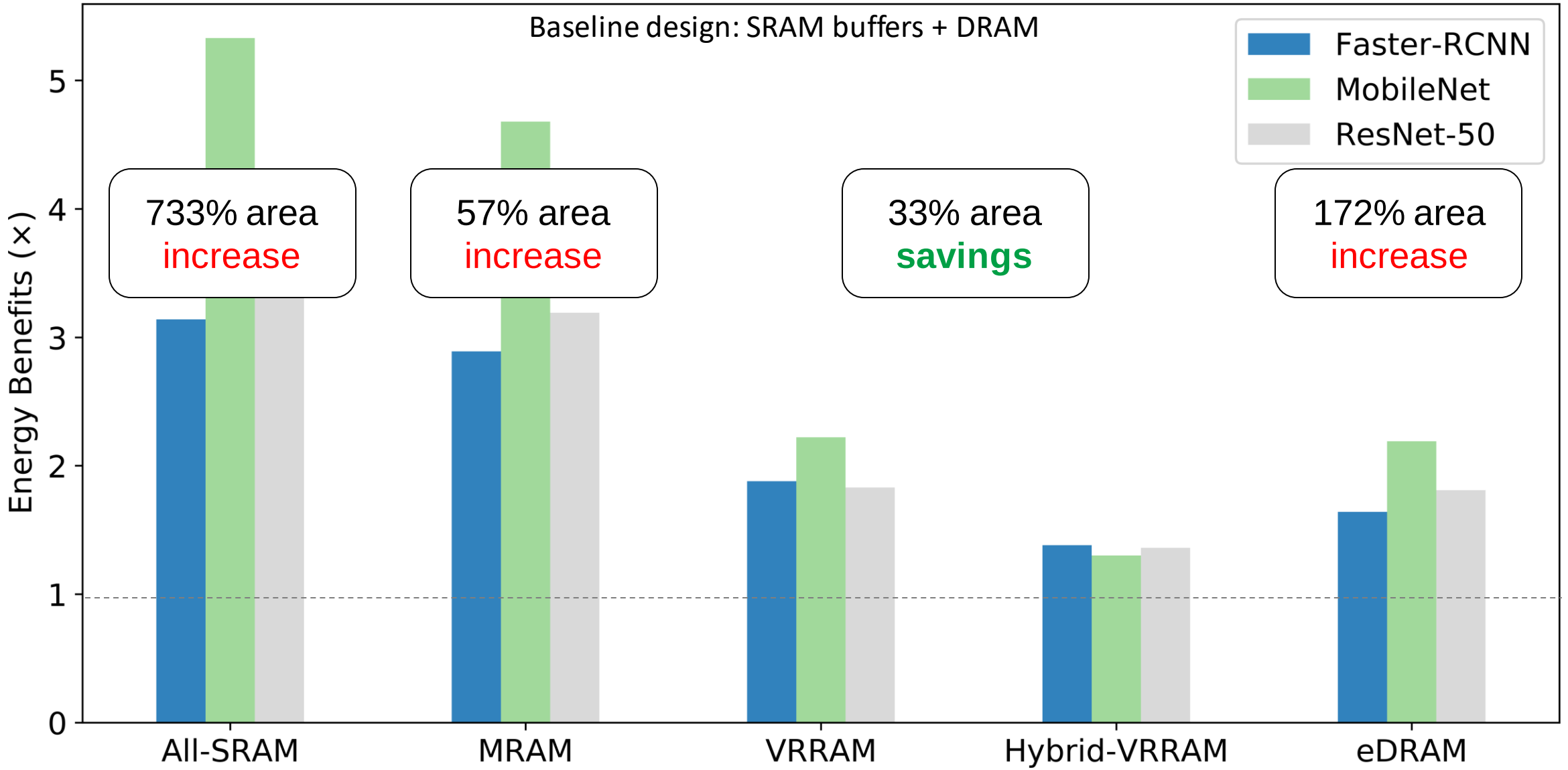
3D RRAM-based design vs. DRAM baseline:
Accelerator area savings due to 4X less SRAM required





High density is an enabler for more aggressive energy-area optimization

Energy-Area-Efficiency Landscape



Conclusions

- Extensive design space explorations for DNN accelerators with NVM
- Energy-area tradeoffs obtained w.r.t. pareto-optimal baselines
  - MRAM: 4.68X energy benefits & 57% area increase
  - 3D VRRAM: 2.22X energy benefits & 33% area savings
- Today's "technology node gap" between SRAM and NVM
  - Low-power SRAM and high-density NVM join forces

Acknowledgement

Brian Cline, Greg Yeric, Matthew Mattina, Naveen Suda, YK Chong (**Arm**)

Priyanka Raina, Subhasish Mitra (**Stanford University**)

E2CDA - ENIGMA

Stanford | SystemX Alliance

Stanford | Non-Volatile Memory Technology Research Initiative (NMTRI)

End of Talk

- Extensive design space explorations for DNN accelerators with NVM
- Energy-area tradeoffs obtained w.r.t. pareto-optimal baselines
  - MRAM: 4.68X energy benefits & 57% area increase
  - 3D VRRAM: 2.22X energy benefits & 33% area savings
- Today's "technology node gap" between SRAM and NVM
  - Low-power SRAM and high-density NVM join forces