Energy Aware Computing Research Group

Cross-Layer Design Space Exploration of NVM-based Caches for Deep Learning

Ahmet Inci[1], Mehmet M Isgenc[1], Diana Marculescu[1,2]

[1] Carnegie Mellon University

[2] TEXAS The University of Texas at Austin

12th Annual Non-Volatile Memories Workshop (NVMW)

March 9, 2021

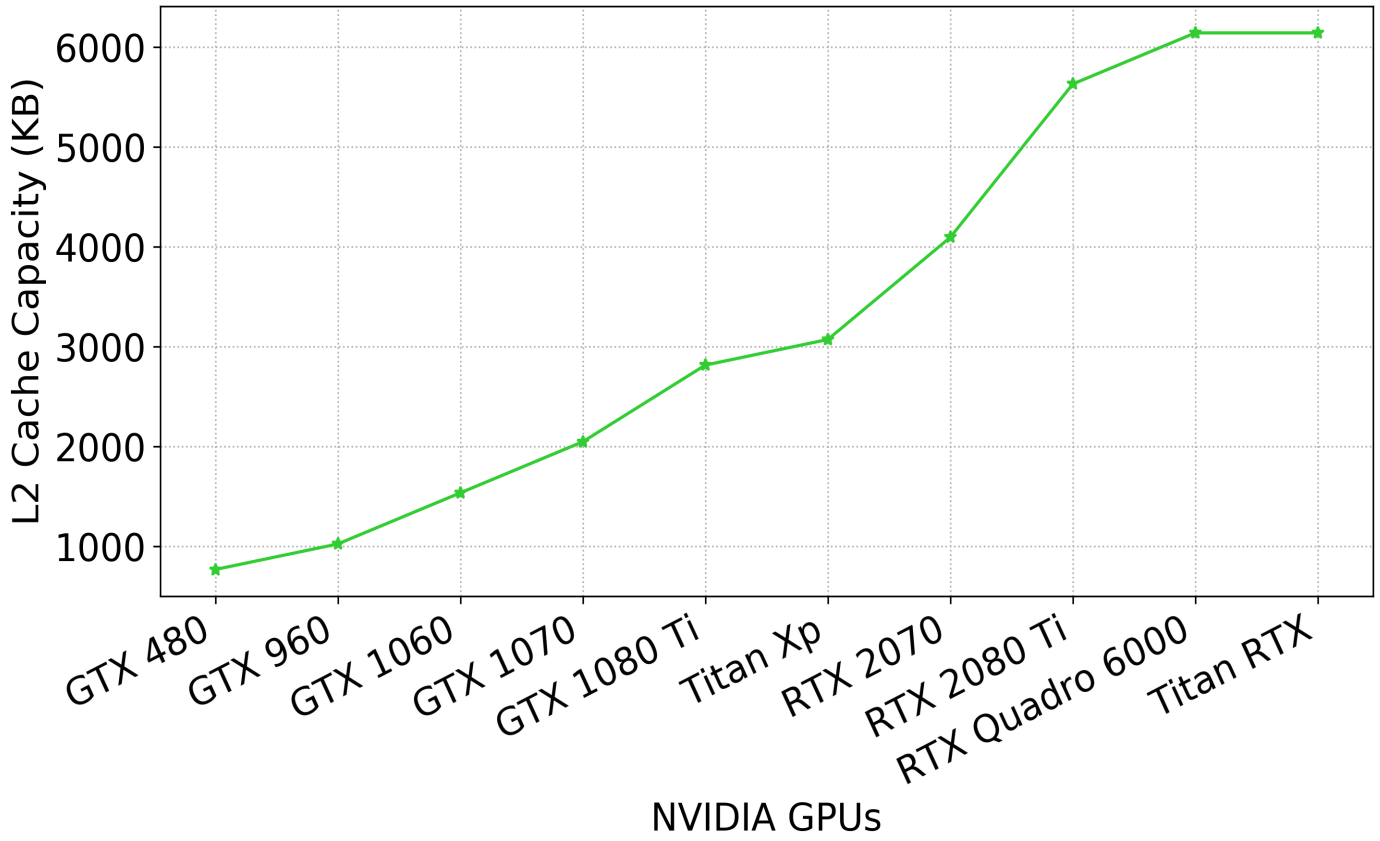
Breaking the memory wall

- **"Hitting the memory wall"**
  - [Wulf *et al.*, CAL'95]
- **Breaking the wall**
  - Caches can help
- **SRAM caches consume large on-chip area and leakage energy**
  - Exacerbated for deep learning workloads

[Source: iStock]

Breaking the memory wall

- **"Hitting the memory wall"**
  - [Wulf *et al.*, CAL'95]
- **Breaking the wall**
  - Caches can help
- **SRAM caches consume large on-chip area and leakage energy**
  - Exacerbated for deep learning workloads
- **What's next?**
  - **Non-volatile memory (NVM)**
    - [Chi *et al.*, ASP-DAC'16]



[Source: NVIDIA]

Should we use NVM for deep learning?

- NVM technologies have significant advantages compared to conventional SRAM due to their non-volatility, high cell density, and scalability features
- NVM is not commercially available for on-chip storage
- Integrating NVM technology into cycle-accurate simulators for current architectures is non-trivial
- Need a design space exploration framework for NVMs for DL workloads

Our work

- **We present *DeepNVM++*, a framework to characterize, model, and analyze NVM-based caches in GPU architectures for deep learning workloads**

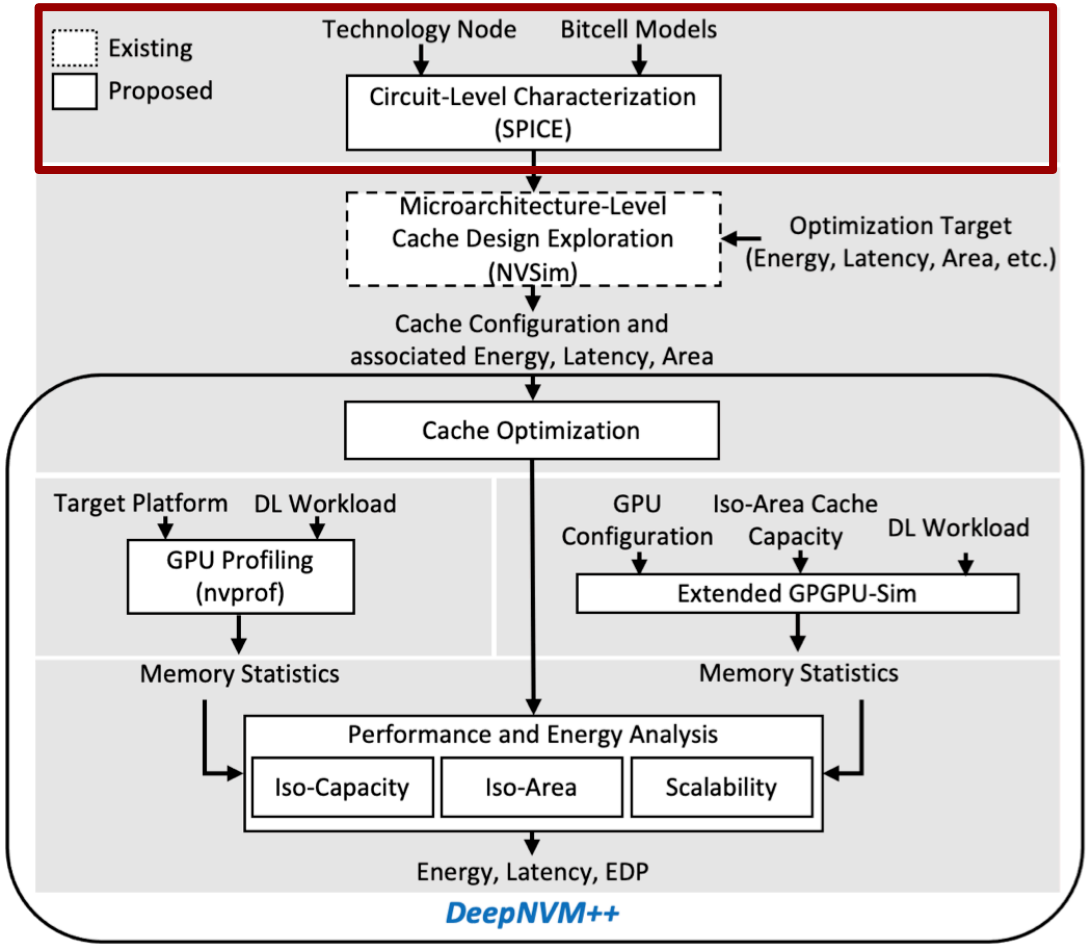
- **We perform design space exploration for conventional SRAM, STT-MRAM, and SOT-MRAM caches for DL workloads**

- **We evaluate the power, performance, and area (PPA) implications of magnetic non-volatile memories on last-level (L2) caches for GPUs in two scenarios:**
  - **Iso-Capacity**
    - We carry out memory profiling of various DL workloads on an existing GPU platform
  - **Iso-Area**
    - We rely on architecture-level simulators to quantify and better understand last-level cache capacity and off-chip memory accesses

Outline

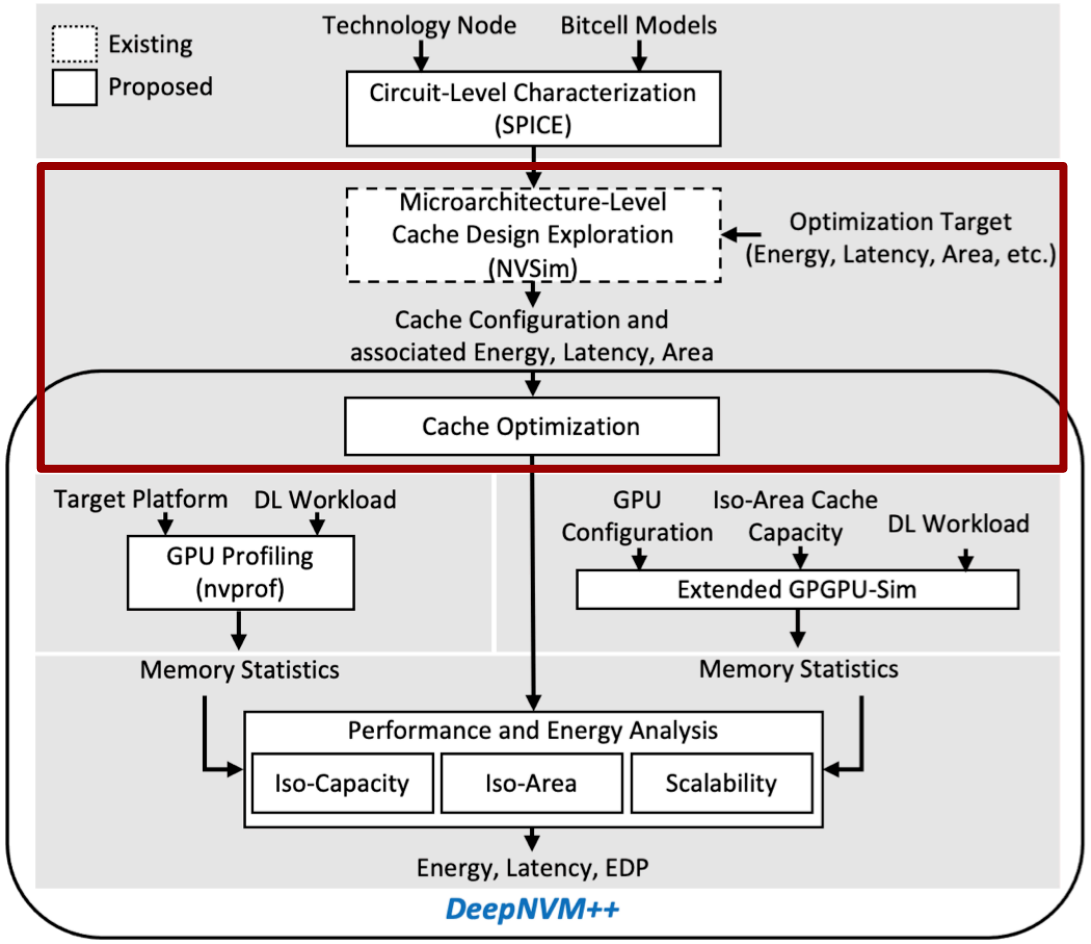
- ✓ Motivation
- **Methodology**
- **Performance and Energy Results**
- **Conclusion**

Overview of the cross-layer analysis flow



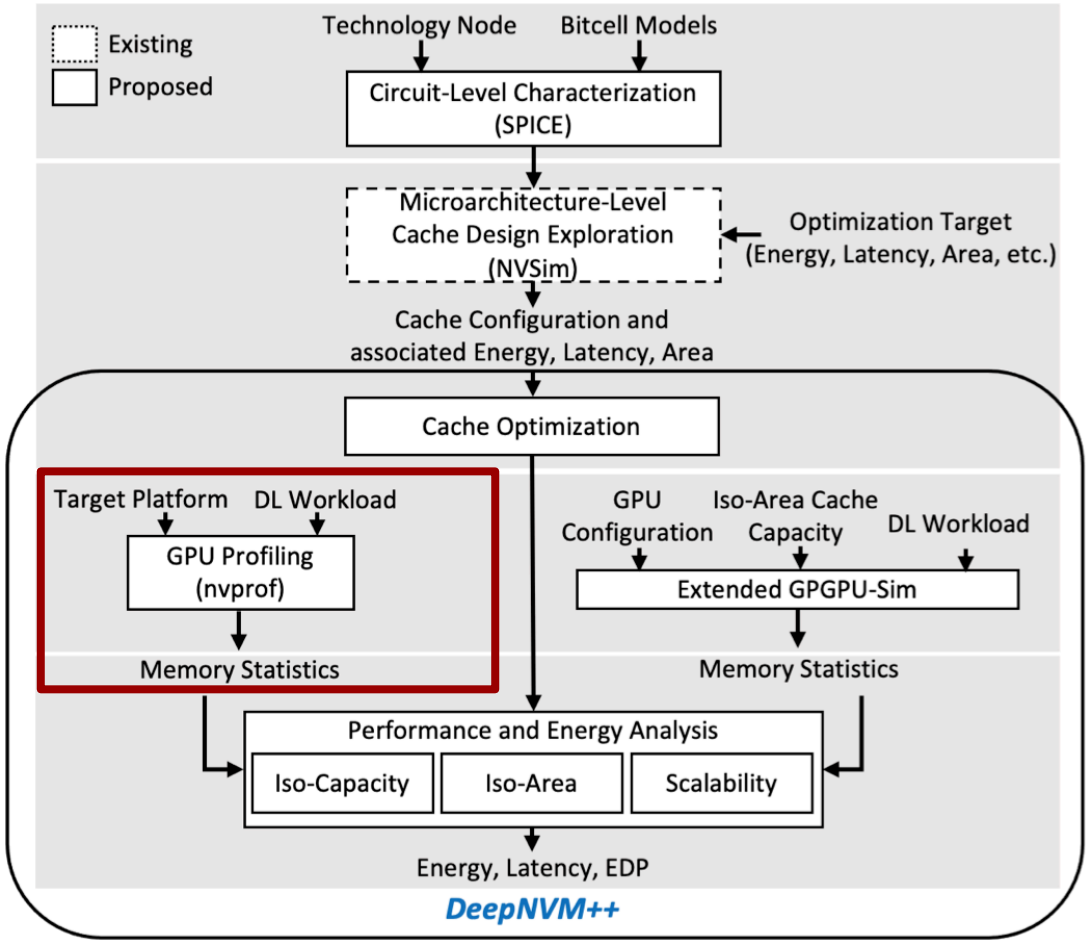
- **Circuit-level characterization with SPICE**
  - Bitcell parameters obtained by using a commercial 16nm technology node

Overview of the cross-layer analysis flow



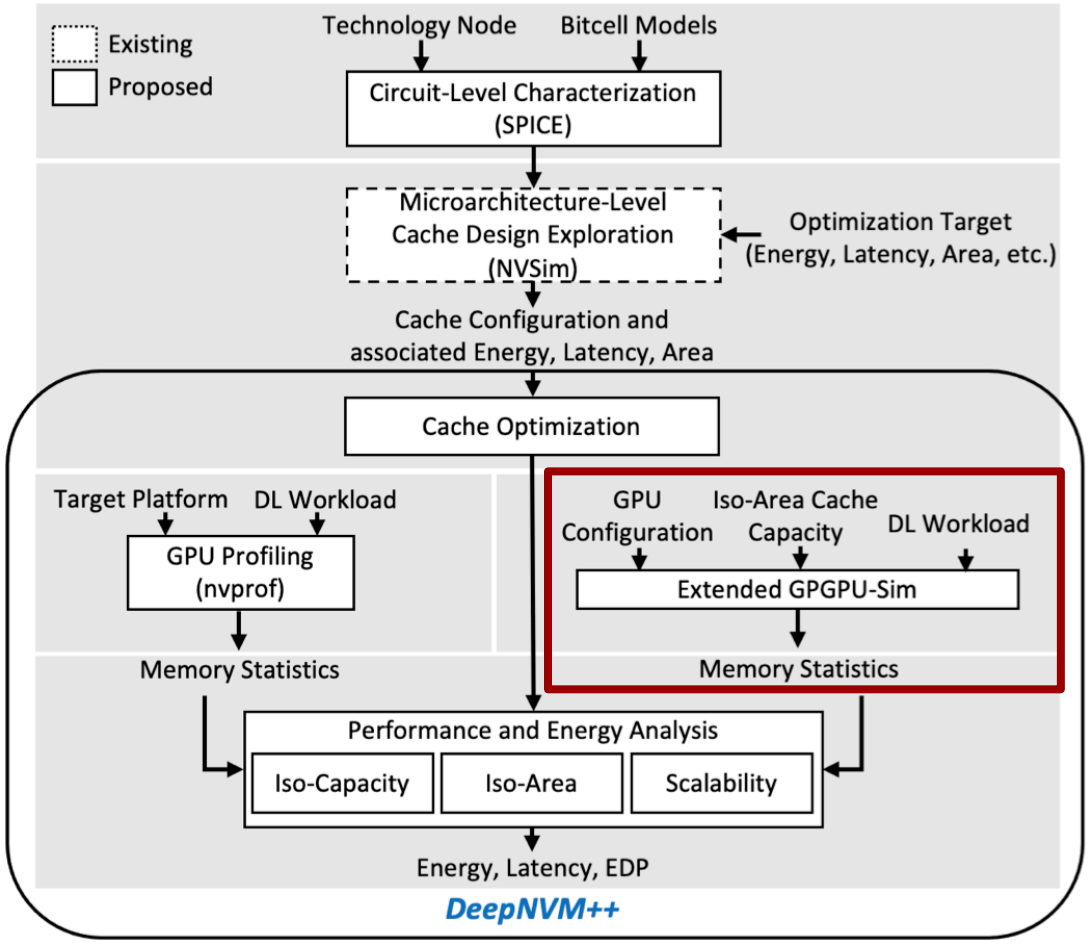
- **Cache design exploration with NVSim**
  - PPA results for different memory technologies at various capacities

Overview of the cross-layer analysis flow



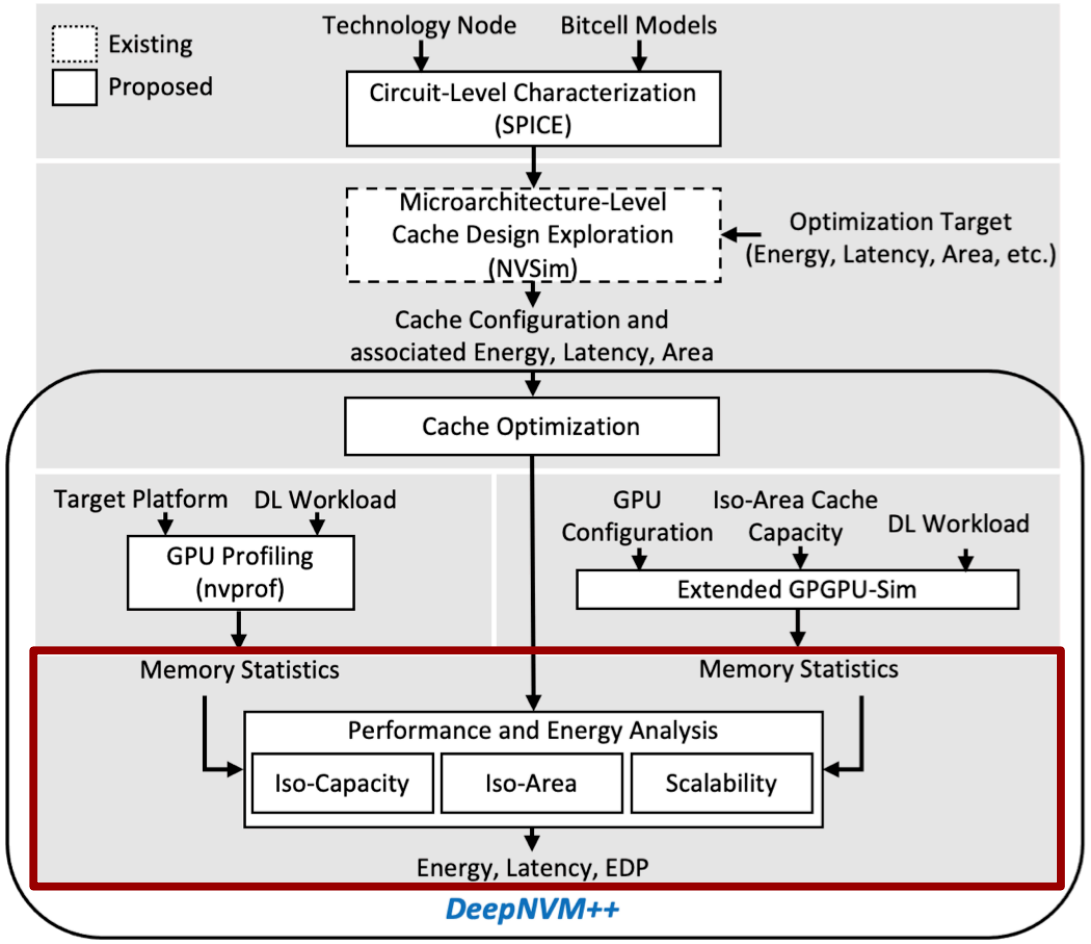
- **Architecture-level iso-capacity analysis**
  - Profiling actual platform memory statistics for various DNNs for both training and inference using NVIDIA GTX 1080 Ti GPU

Overview of the cross-layer analysis flow



- **Architecture-level iso-area analysis**
  - Extending GPGPU-Sim to explore performance and energy implications of having larger L2 caches for deep learning workloads

Overview of the cross-layer analysis flow



- **Performance and energy analysis**
  - Energy, latency, and energy-delay product (EDP) results for iso-capacity, iso-area, and scalability analysis

Implications in architecture-level analysis

- **Iso-Capacity**
  - Replace SRAM L2 cache in a GPU with the same cache capacity MRAM
  - Reduce overall chip area
- **Iso-Area**
  - Replace SRAM for the same area budget and increase the on-chip L2 cache capacity using MRAM equivalents
  - Reduce costly off-chip memory accesses

PPA comparison of SRAM and MRAM caches

	SRAM	STT-MRAM		SOT-MRAM	
		Iso-Capacity	Iso-Area	Iso-Capacity	Iso-Area
Capacity (MB)	3	3	7	3	10
Read Latency (ns)	2.91	2.98	4.58	3.71	6.69
Write Latency (ns)	1.53	9.31	10.06	1.38	2.47
Read Energy (nJ)	0.35	0.81	0.93	0.49	0.51
Write Energy (nJ)	0.32	0.31	0.43	0.22	0.40
Leakage Power (mW)	6442	748	1706	527	1434
Area (mm^2)	5.53	2.34	5.12	1.95	5.64

- For iso-capacity scenario, SRAM occupies significant on-chip area
  - STT-MRAM and SOT-MRAM are 2.4x and 2.8x more area-efficient compared to SRAM, respectively

PPA comparison of SRAM and MRAM caches

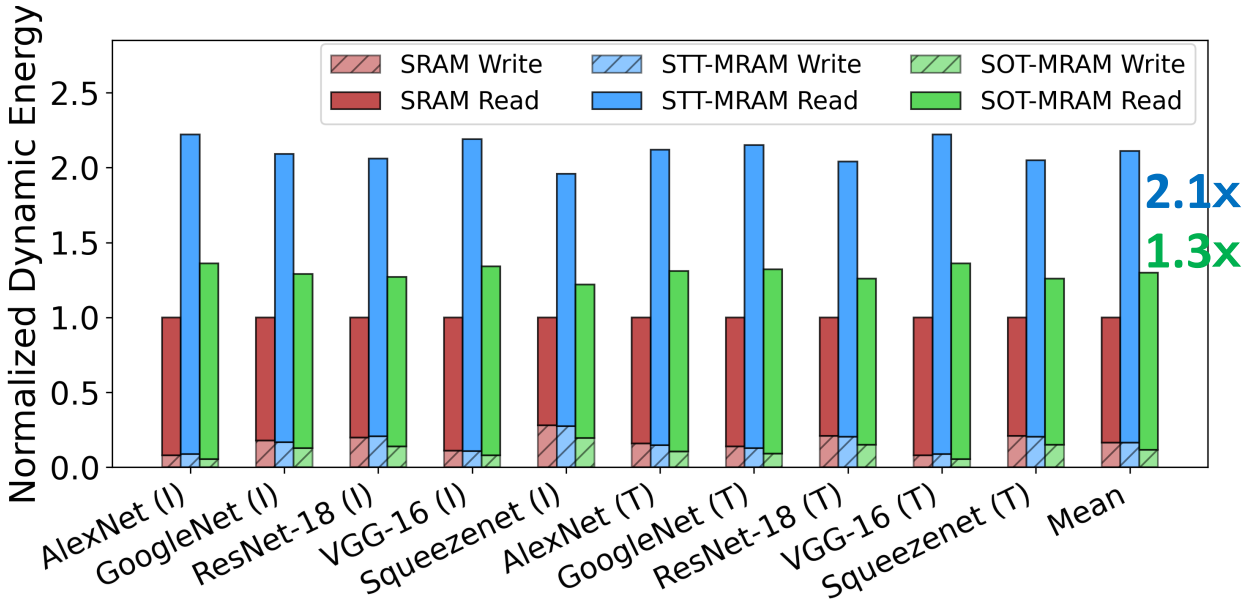
	SRAM	STT-MRAM		SOT-MRAM	
		Iso-Capacity	Iso-Area	Iso-Capacity	Iso-Area
Capacity (MB)	3	3	7	3	10
Read Latency (ns)	2.91	2.98	4.58	3.71	6.69
Write Latency (ns)	1.53	9.31	10.06	1.38	2.47
Read Energy (nJ)	0.35	0.81	0.93	0.49	0.51
Write Energy (nJ)	0.32	0.31	0.43	0.22	0.40
Leakage Power (mW)	6442	748	1706	527	1434
Area (mm^2)	5.53	2.34	5.12	1.95	5.64

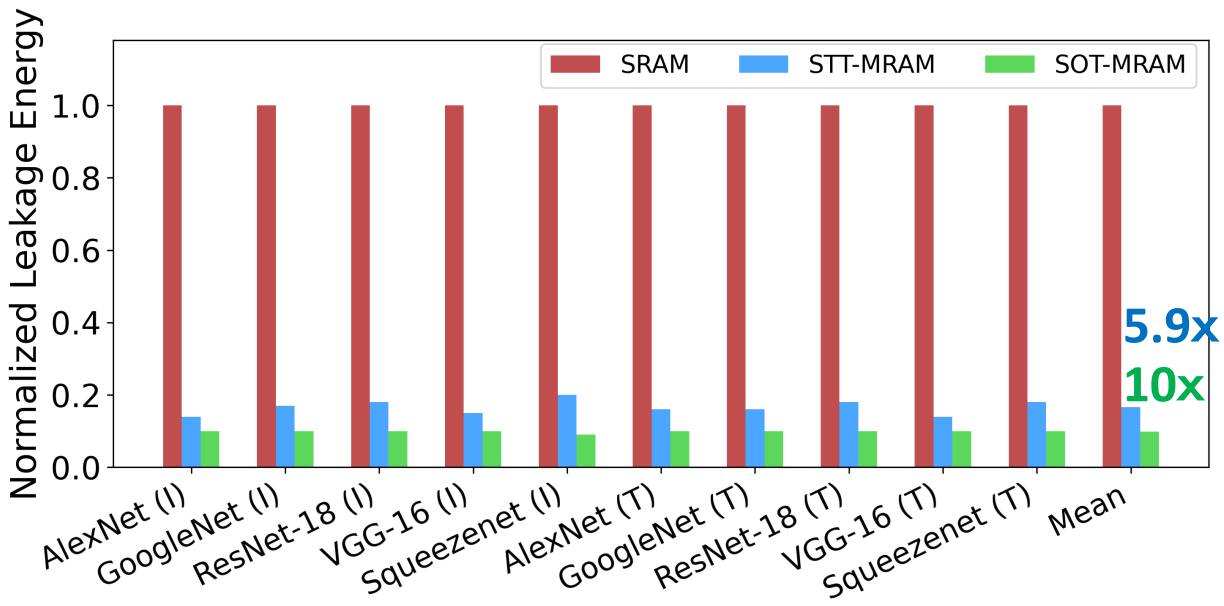
- For iso-area scenario, STT-MRAM and SOT-MRAM accommodate 2.3x and 3.3x cache capacity compared to SRAM, respectively

Outline

- ✓ Motivation
- ✓ Methodology
- **Performance and Energy Results**
- **Conclusion**

Energy results for iso-capacity

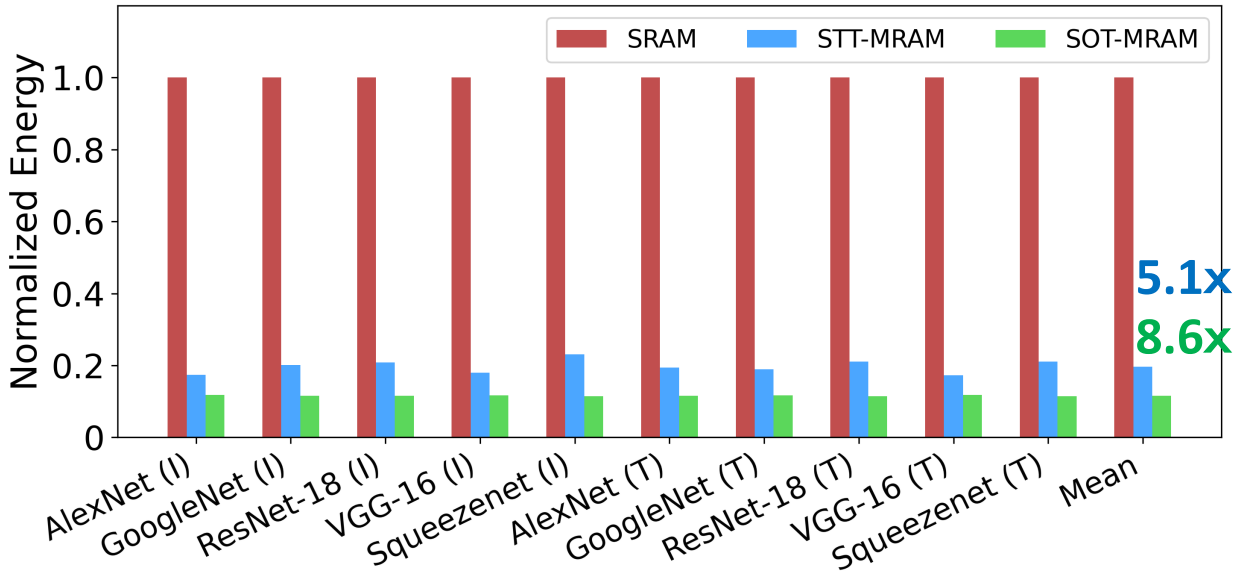


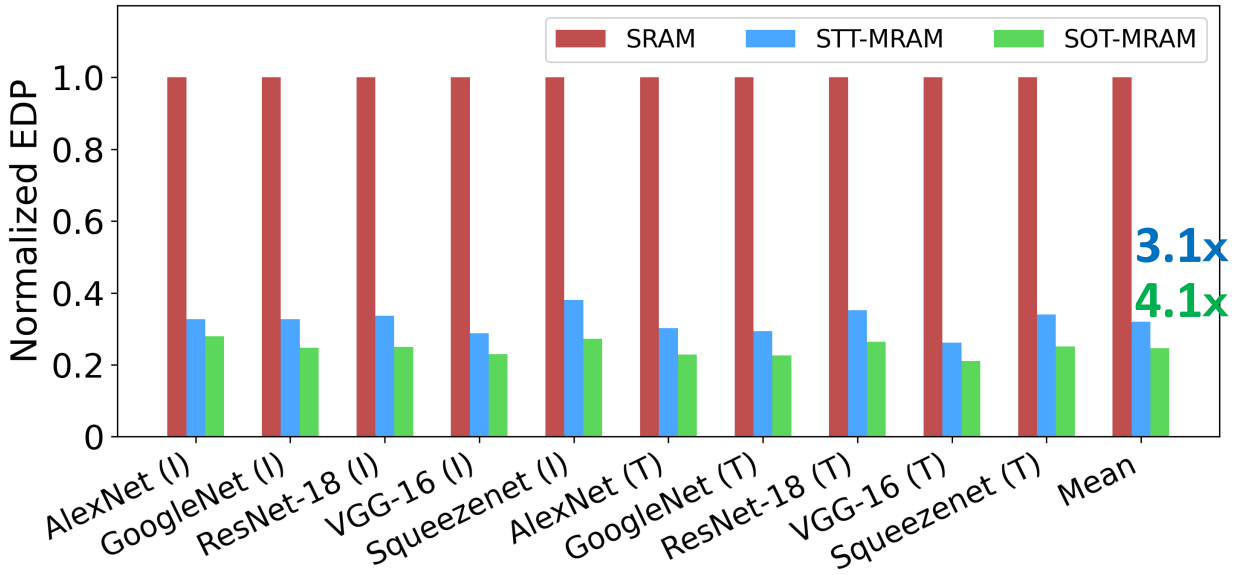


- **STT-MRAM has 2.1x more dynamic energy whereas SOT-MRAM has 1.3x more dynamic energy on average when compared to SRAM**
  - **83% of the total dynamic energy comes from read operations whereas write operations only make for 17% of all transactions on average**

- **STT-MRAM and SOT-MRAM provide 5.9x and 10x lower leakage energy on average when compared to SRAM, respectively**

Performance and energy results for iso-capacity

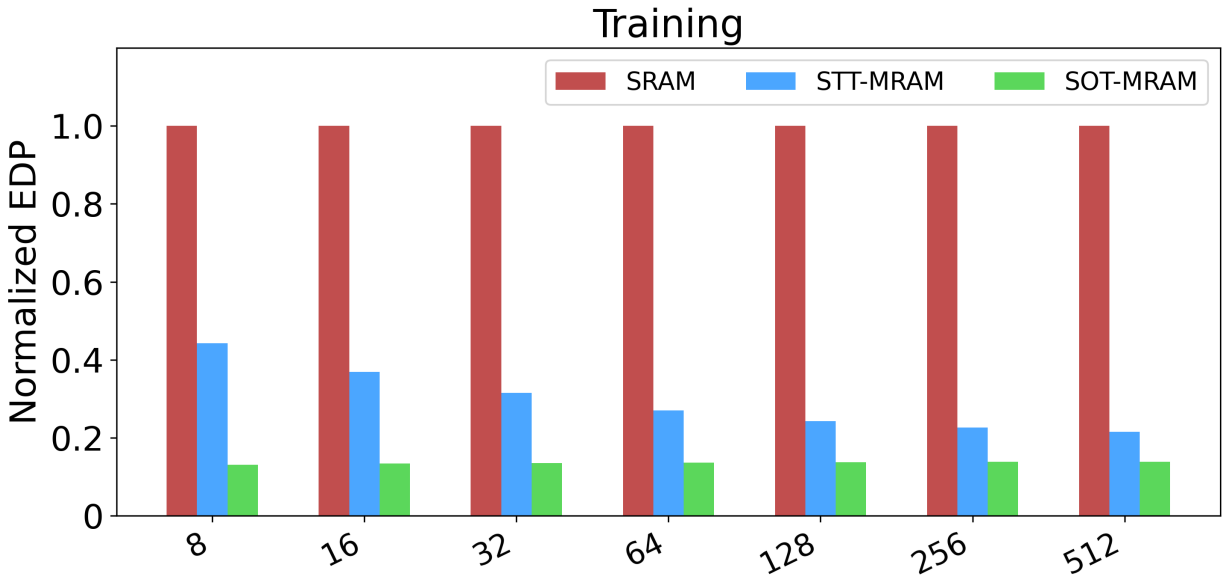


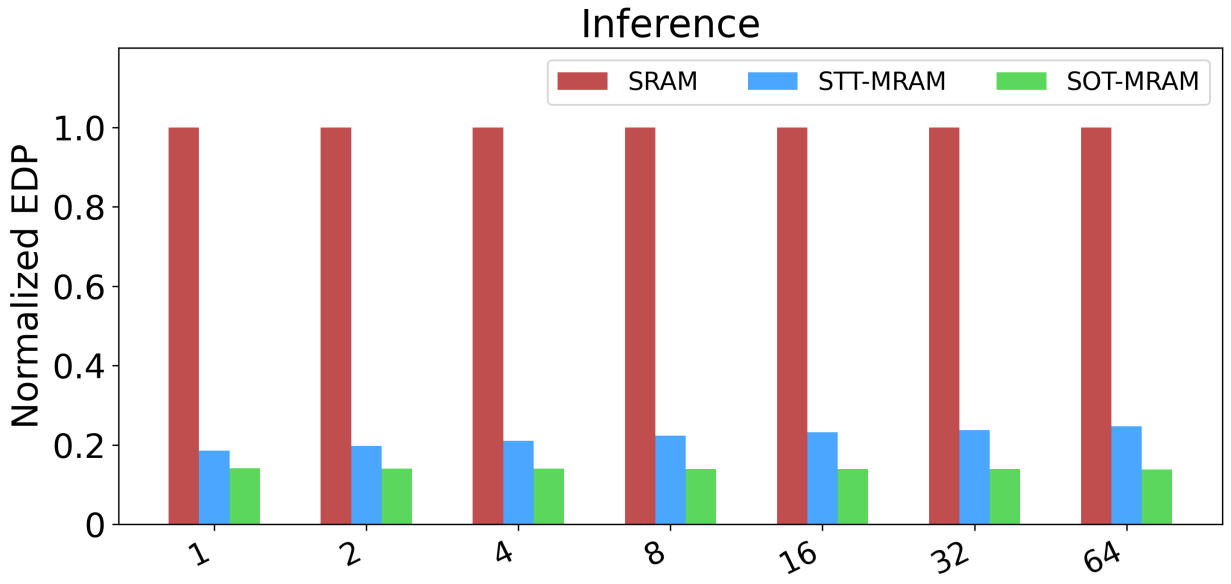


- **Compared to SRAM, STT-MRAM and SOT-MRAM achieve:**
  - ♦ **5.1x and 8.6x energy reduction**
  - ♦ **Leakage energy dominates the total energy**

- **Compared to SRAM, STT-MRAM and SOT-MRAM provide up to:**
  - ♦ **3.8x and 4.7x EDP reduction**
  - ♦ **2.4x and 2.8x area reduction**

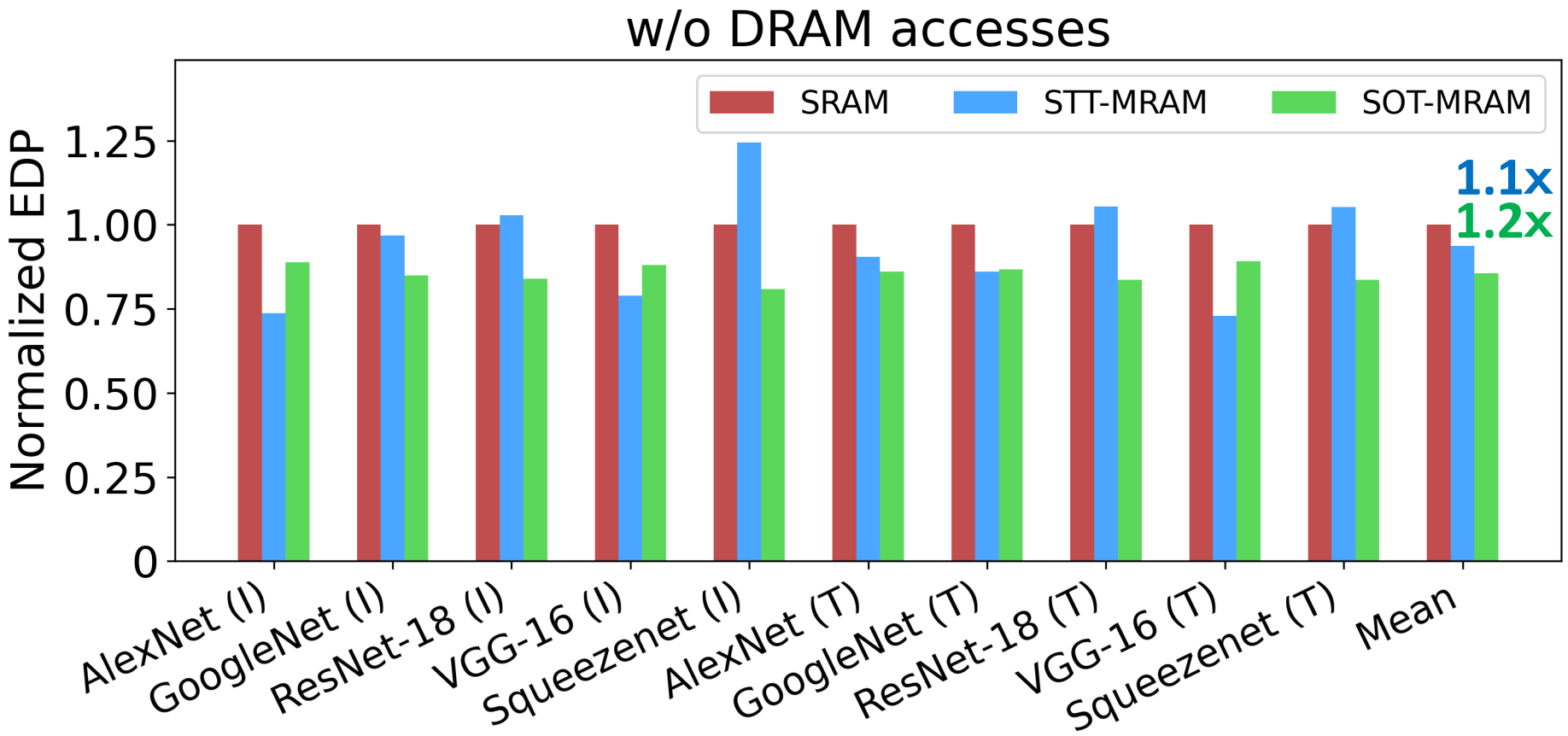
The impact of batch size on EDP





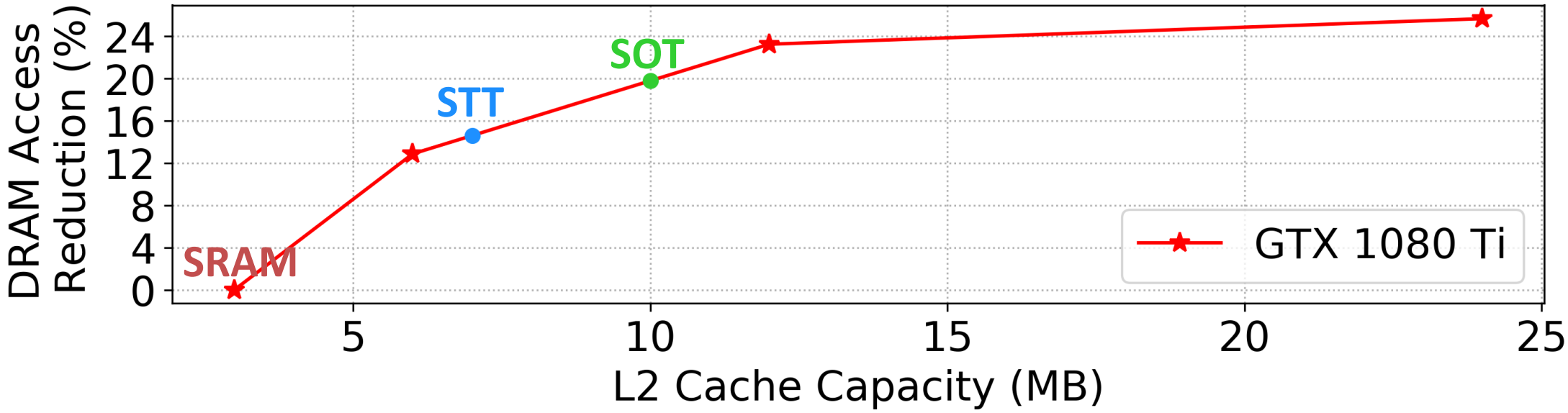
- **As batch size increases;**
  - **STT-MRAM provides 2.3x to 4.6x EDP reduction for *training* and 5.4x to 4.1x EDP reduction for *inference***
  - **SOT-MRAM provides 7.6x to 7.2x EDP reduction for *training* and 7.1x to 7.3x EDP reduction for *inference***

EDP results for iso-area (w/o DRAM accesses)



- **STT-MRAM and SOT-MRAM provide 1.1x and 1.2x EDP reduction on average when excluding DRAM accesses in EDP calculations, respectively**

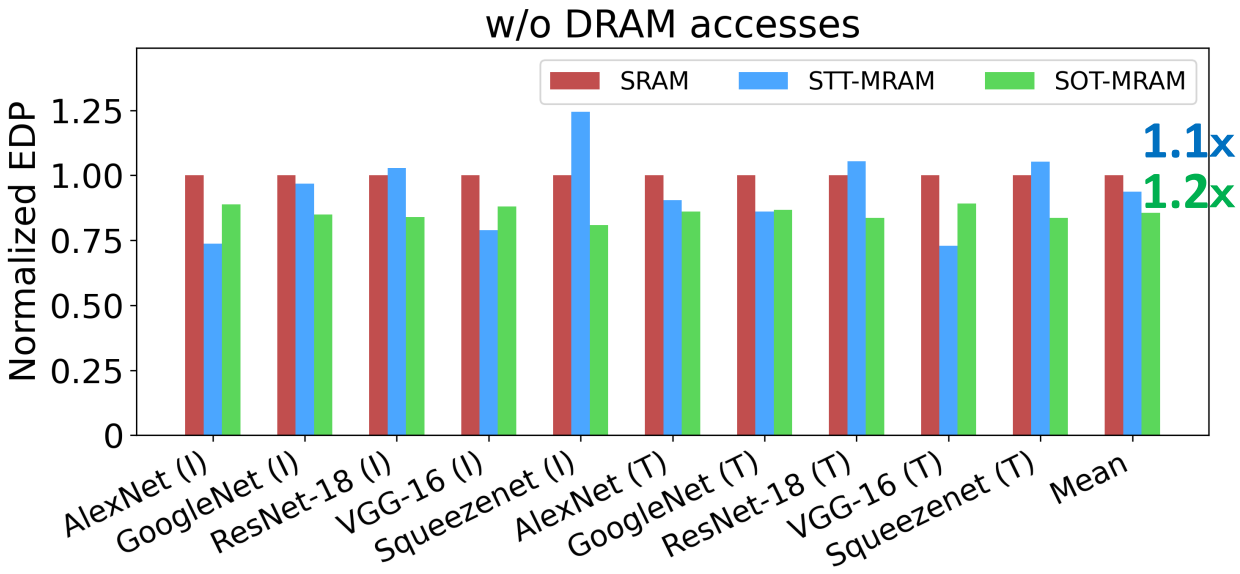
Quantifying the DRAM access reduction with GPGPU-Sim

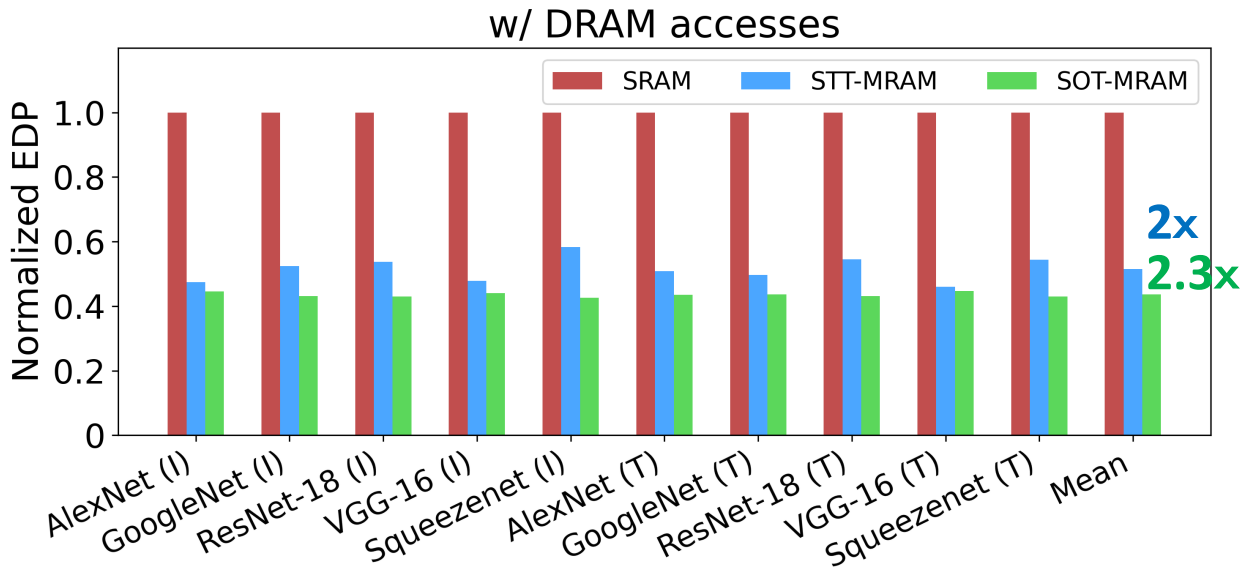


- We use AlexNet with the ImageNet dataset which is provided by *DarkNet* framework

- STT-MRAM and SOT-MRAM reduce 14.6% and 19.8% of the total DRAM accesses by using the same area budget with SRAM baseline, respectively

EDP results for iso-area (w/o vs w/ DRAM accesses)

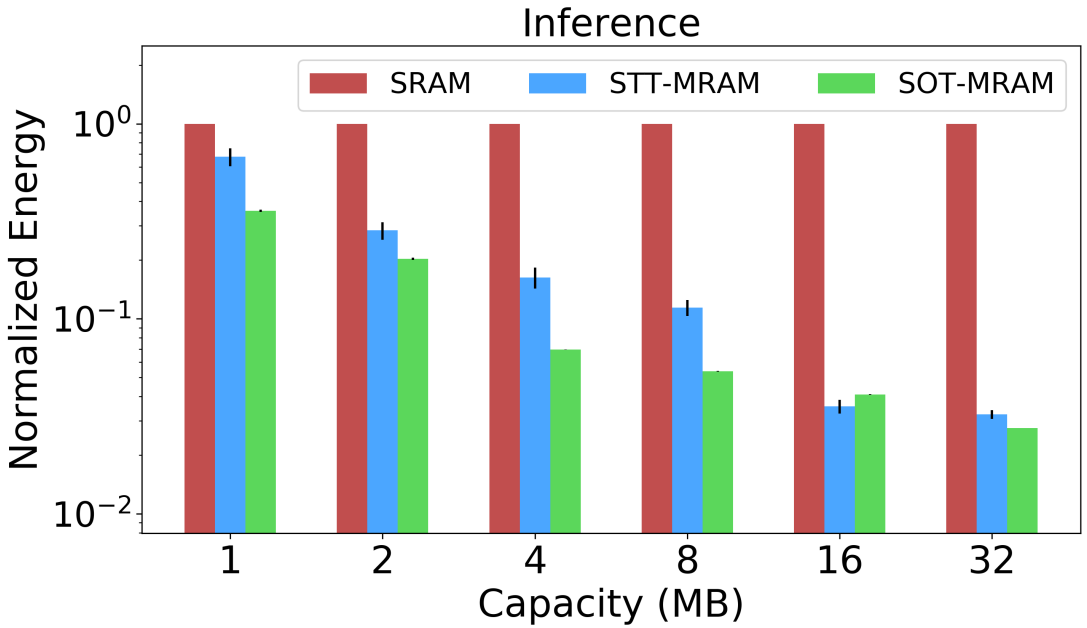


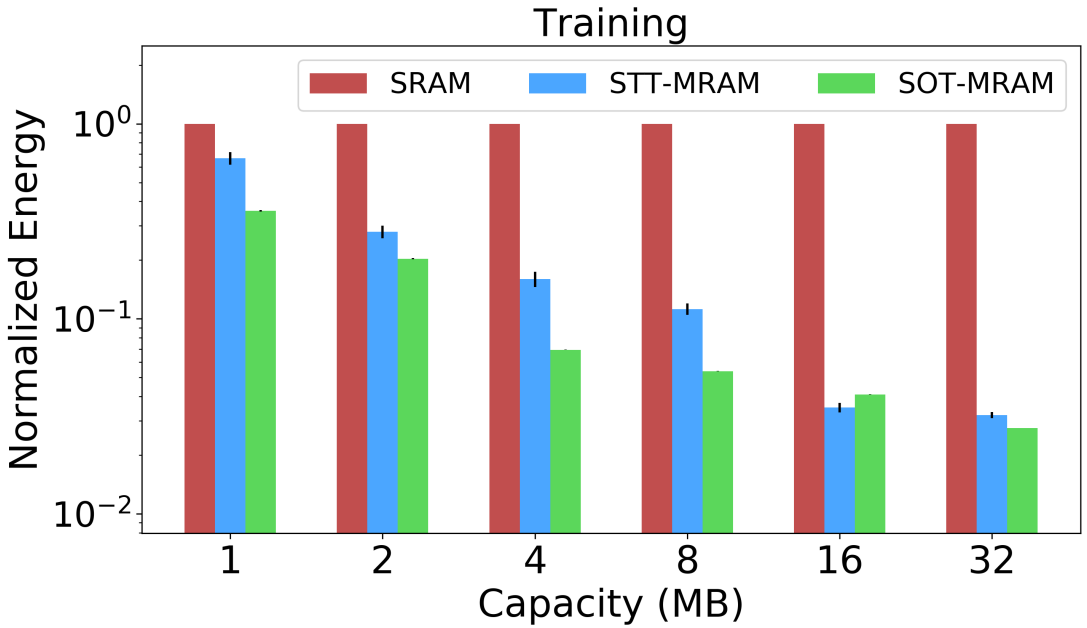


- **STT-MRAM and SOT-MRAM provide 2x and 2.3x EDP reduction on average when compared to SRAM, respectively**

- **Although the cache latency and energy results for MRAM caches do not outperform SRAM, they do outperform SRAM when DRAM accesses are also considered in EDP calculations**

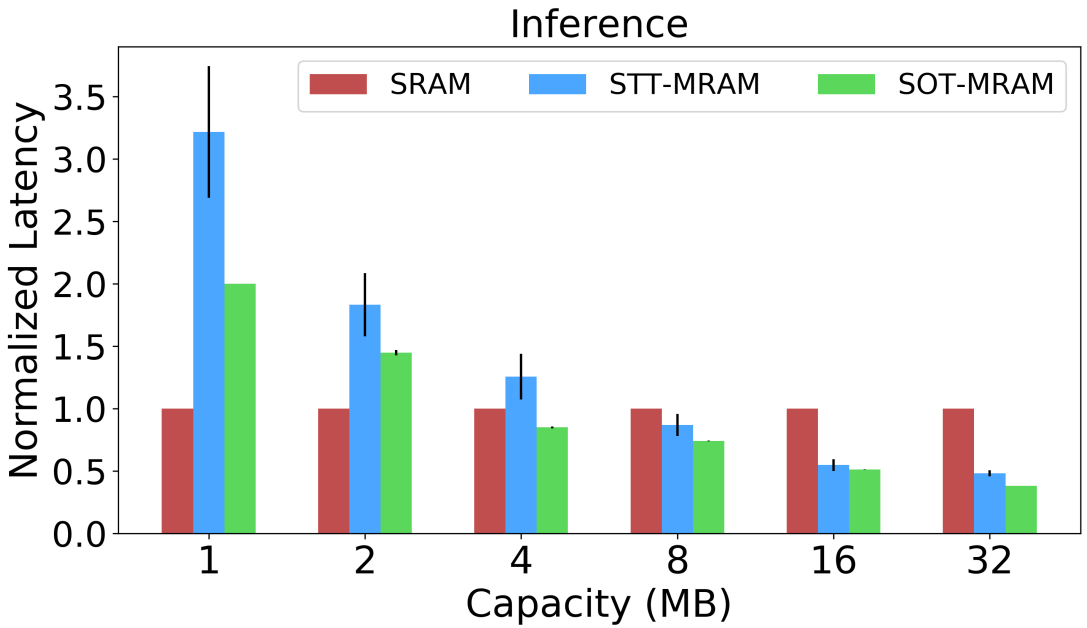
Scalability analysis for energy

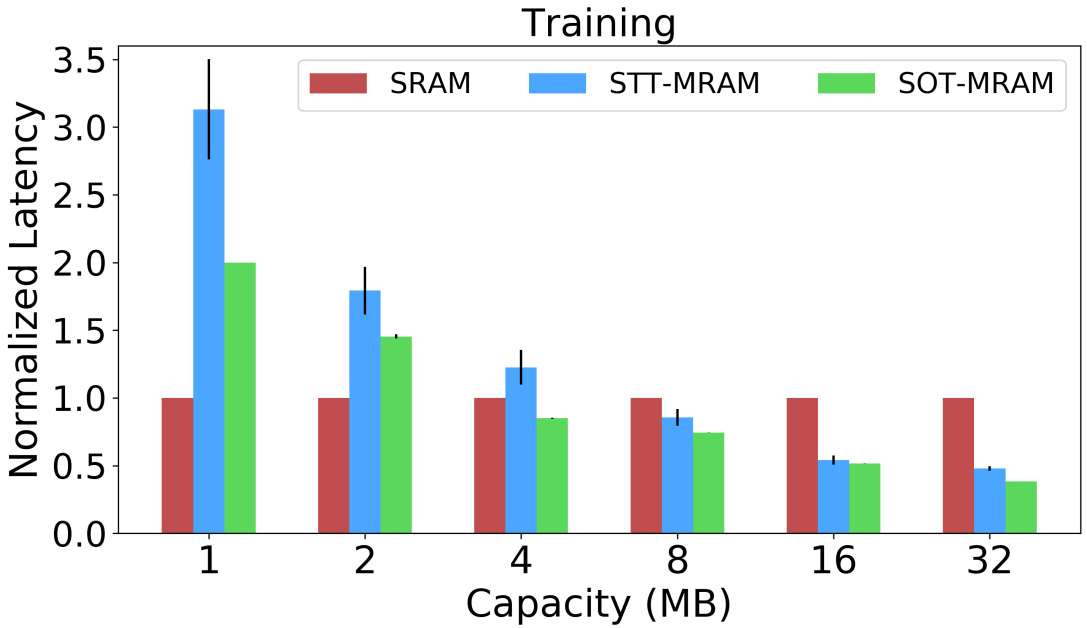




- **MRAM caches provide lower energy as cache capacity increases**
  - **STT-MRAM and SOT-MRAM achieve up to 31.2x and 36.4x energy reduction, respectively**

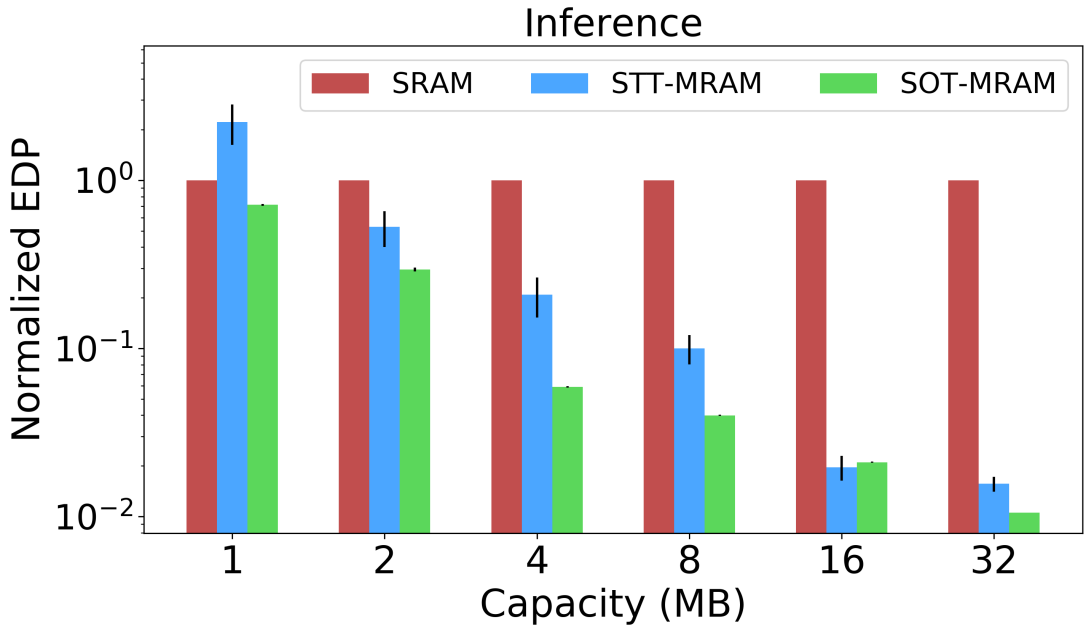
Scalability analysis for latency

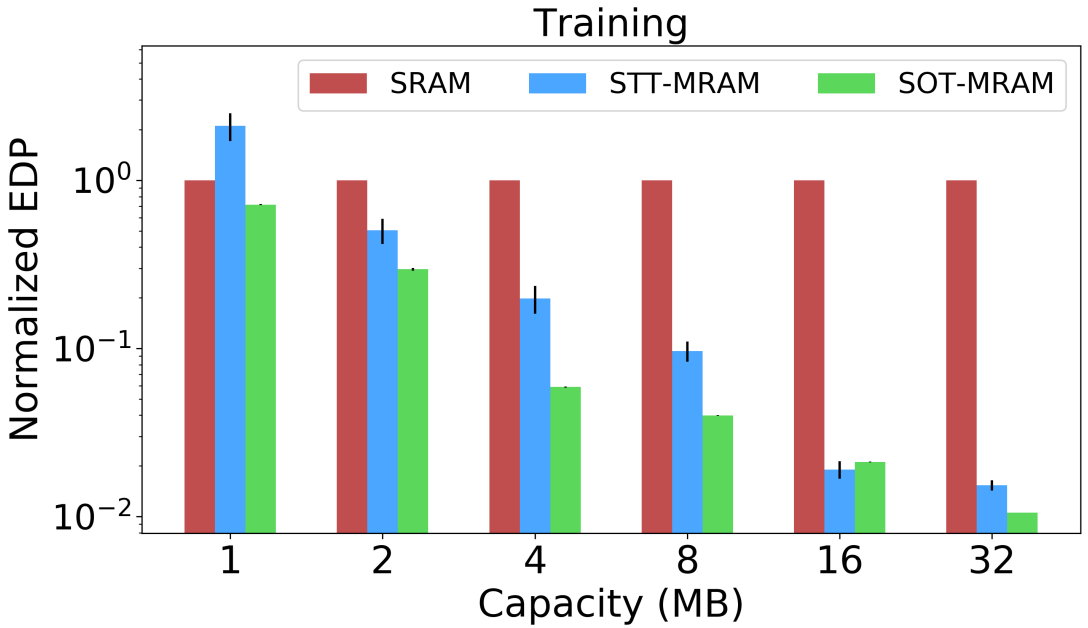




- **MRAM caches have higher latency results up to 4MB**
  - **SRAM provides up to 3.2x and 2x latency reduction for small cache capacities**
  - **STT-MRAM and SOT-MRAM achieve up to 2.1x and 2.6x latency reduction as cache capacity increases, respectively**

Scalability analysis for EDP





- **STT-MRAM and SOT-MRAM provide orders of magnitude improvement in terms of EDP reduction when compared to SRAM**

Conclusion

- **We present *DeepNVM++*, the first cross-layer analysis framework to characterize, model, and analyze various NVM technologies in GPU architectures for deep learning workloads**

- **We show that STT-MRAM and SOT-MRAM have promising results for deep learning workloads in terms of energy, latency, and energy-delay product**
  - In the iso-capacity case, STT-MRAM and SOT-MRAM provide up to 3.8x and 4.7x EDP reduction and 2.4x and 2.8x area reduction when compared to SRAM, respectively
  - In the iso-area case, STT-MRAM and SOT-MRAM achieve up to 2x and 2.3x EDP reduction and accommodate 2.3x and 3.3x larger cache capacity when compared to SRAM, respectively

- **Our novel framework can be used to further explore the feasibility of emerging NVM technologies for deep learning applications for different design choices**