

Evaluation of Non-Volatile Memory Based Last Level Cache Given Modern Use Case Behavior



Alex Hankin*, Tomer Shapira*, Karthik Sangaiah[†], Mike Lui[†], Mark Hempstead*

*TCAL, Tufts University [†]VANDAL, Drexel University

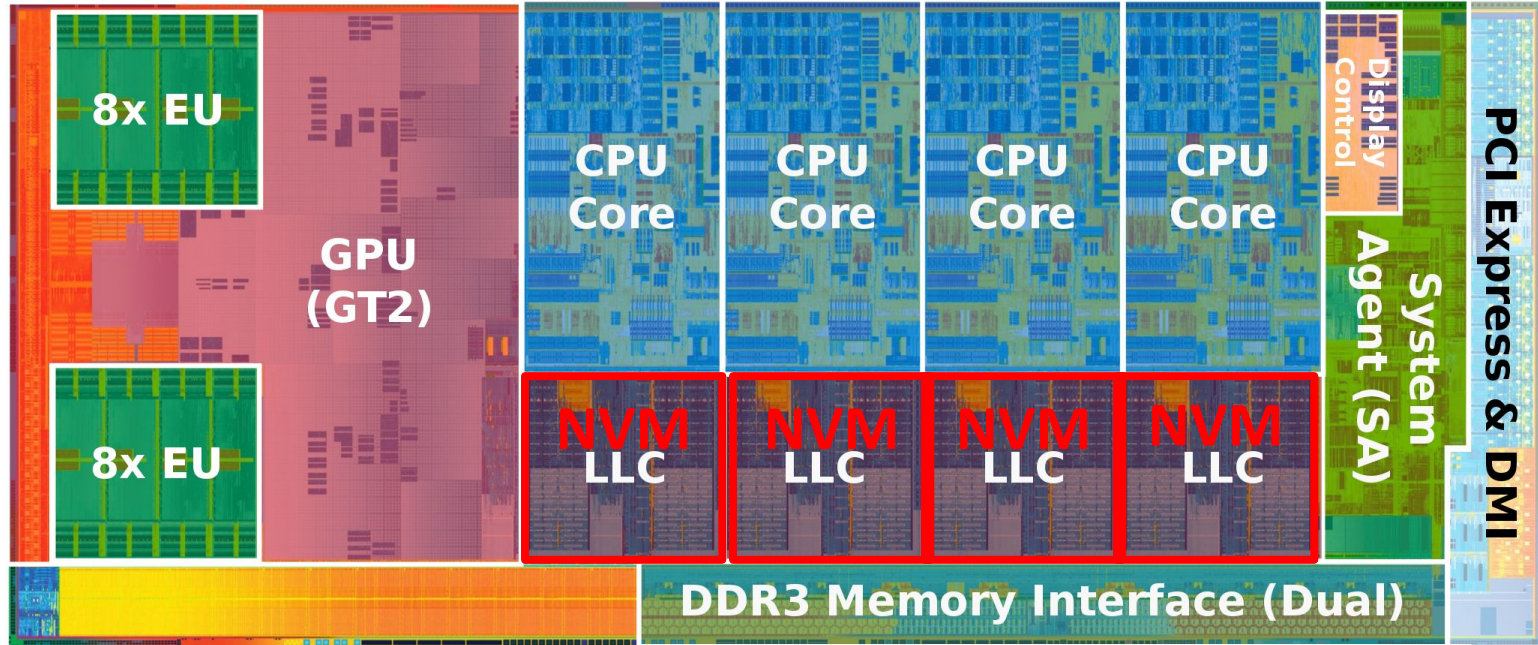
IISWC 2019, Orlando, Fl., November 4th, 2019

Motivation

- Working set sizes of applications increasing
 - E.g., ML/AI workloads
- Traditional LLC tech (SRAM) is density-limited and leaky

Benefits of NVMs:	Drawbacks:
<ul style="list-style-type: none">- Smaller cell size - Lower energy 	<ul style="list-style-type: none">- Write latency- Lifetime

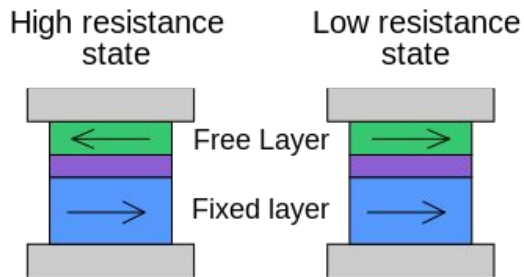
NVM-Based LLC



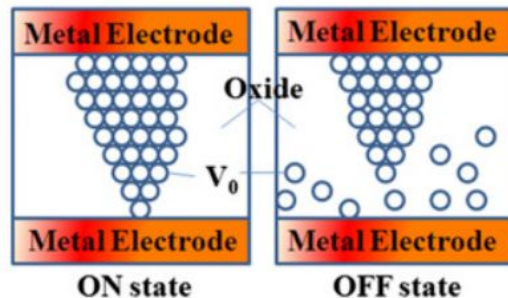
For a workload which NVM technology is best?

- Challenges:
 - Lack of consistent models/modeling methodology for NVM LLCs in the literature
 - Leads to apples-to-oranges comparisons
 - Effects of NVM read-write asymmetry unknown
 - Not modeled in current architecture performance simulators
- Thought there would be one perfect NVM -- more complicated!
- Depends on workload -- predict best NVM without simulation
- Predict for workloads of the future!

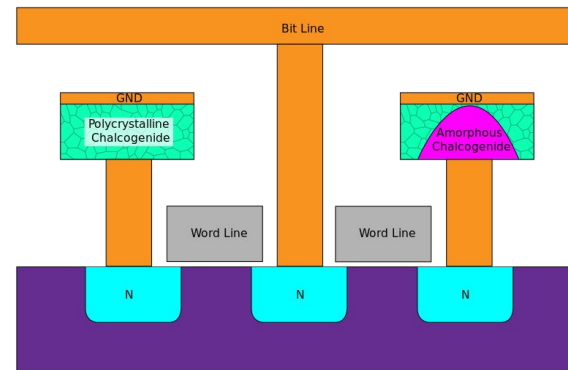
Emerging NVMs



Spin-Torque Transfer RAM (STTRAM) ²



Resistive RAM (RRAM) ³



Phase Change RAM (PCRAM) ⁴

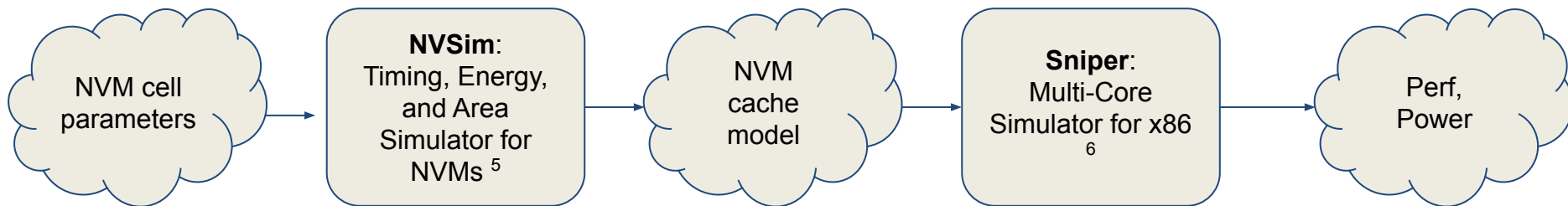
- Physical mechanisms affect energy, latency, and area
- Some technologies are already CMOS compatible
- All can be accessed by CMOS device -- easily interface with CMOS

² STTRAM diagram: https://en.wikipedia.org/wiki/Spin-transfer_torque#/media/File:Spin_valve_schematic.svg

³ RRAM diagram: Meena, Jagan & Sze, Simon & Chand, Umesh & Tseng, Tseung-Yuen. (2014). Overview of Emerging Non-volatile Memory Technologies. Nanoscale Research Letters. 9. 1-33. 10.1186/1556-276X-9-526

⁴ PCRAM diagram: https://en.wikipedia.org/wiki/Phase-change_memory#/media/File:PRAM_cell_structure.svg

Modeling

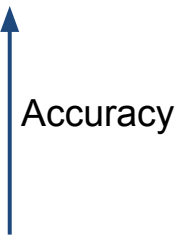


Two problems:

- Specifying NVM cell for NVM simulator
 - Some necessary parameters missing from VLSI/Circuits papers that introduces an NVM
- Modeling NVM cache in full-system simulator
 - Expects SRAM, i.e., symmetric read/write latency

NVM Cells

- Compiled about a dozen popular NVM **cells** from the VLSI and Architecture literature
 - A few from each class (PCM, STTRAM, ReRAM)
- Developed some **modeling heuristics** for determining unknown parameters:
 - Electrical Properties
 - Interpolation
 - Similarity



NVM Cells

	Oh	Chen	Kang	Close	Chung	Jan	Umeki	Xue	Hayakawa	Zhang
Class	PCRAM	PCRAM	PCRAM	PCRAM	STTRAM	STTRAM	STTRAM	STTRAM	RRAM	RRAM
Year	2005	2006	2006	2013	2010	2014	2015	2016	2015	2016
Access Device	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
Process Node[nm]	120	60*	100	90	54	90	65	45	40	22
Cell Size [F ²]	16.6*	10*	16.6	25	14	50	48 ⁺	63	4*	4*
Cell Levels	1	1	1	2	1	1	1	2	1	1

- RRAM cell size is small
 - Can build dense LLC => expect good performance
- Baseline SRAM tech is 45 nm
- * and ⁺ indicate parameters not reported in the literature

NVM Cells

	Oh	Chen	Kang	Close	Chung	Jan	Umeki	Xue	Hayakawa	Zhang
Class	PCRAM	PCRAM	PCRAM	PCRAM	STTRAM	STTRAM	STTRAM	STTRAM	RRAM	RRAM
Reset Current [μA]	600	90	600	400	80	52	255 ⁺	150		
Set Current [μA]	200	55	200*	400	100 ⁺	38	255 ⁺	150		

- PCRAM current is high => expect high energy
- Grayed out boxes -> NVSim requires different parameters for different classes
- This is a subset of the data -- the rest of the NVM cell data is in the paper!
- All data is publicly available at: <http://sites.tufts.edu/tcal/nvm-models!>

NVM Cache Models

- **Two different sets** of NVM-based LLC models:
 - Iso-capacity: same capacity as SRAM-based LLC
 - Iso-area: same area as a 2 MB SRAM-based LLC

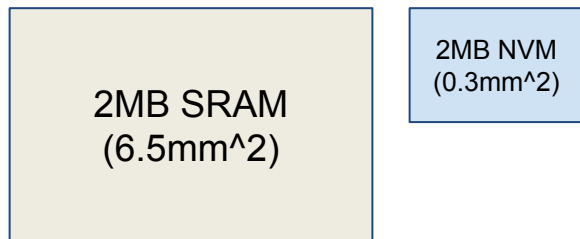


Figure 1: Iso-capacity SRAM and NVM LLC

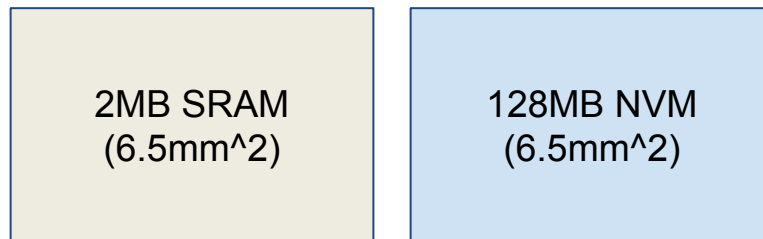


Figure 2: Iso-area SRAM and NVM LLC

- SRAM-based cache model baseline

Iso-capacity and Iso-area Models

	Ohp	Chenp	Kangp	Closep	Chung _s	Jang _s	Umeki _s	Xue _s	Hayakawa _g	Zhang _g	SRAM
Area [mm^2]	6.847	4.104	4.591	2.855	1.452	9.171	4.348	1.585	0.915	0.307	6.548
Tag Access Latency [ns]	0.74	0.604	0.656	0.582	1.240	1.423	1.208	1.156	1.396	1.722	0.439
Data Read Latency, t_{read} [ns]	1.907	0.607	1.497	0.82	1.763	3.072	2.715	2.878	1.722	2.16	1.234
Data Write Latency, t_{write} [ns] (set/ reset)	181.206 /11.206	80.491 /60.491	301.018 /51.018	20.681 /20.681	11.751	7.878	11.916	4.038	20.716	300.834	0.515
Cache Hit Dynamic Energy, $E_{dyn,hit}$ [nJ]	0.840	0.421	0.678	0.437	0.209	0.188	0.173	0.251	0.263	0.217	0.565
Cache Miss Dynamic Energy, $E_{dyn,miss}$ [nJ]	0.042	0.025	0.033	0.023	0.082	0.077	0.058	0.121	0.078	0.086	0.011
Cache Write Dynamic Energy, $E_{dyn,write}$ [nJ]	225.413	34.108	375.073	51.116	1.332	2.305	1.644	0.597	0.952	0.523	0.537
Cache Total Leakage Power [W]	0.062	0.100	0.061	0.137	0.661	0.025	0.295	0.828	3.896	9.000	3.438
Capacity [MB]	2	4	2	4	8	1	2	8	32	128	2
Tag Access Latency [ns]	0.740	0.607	0.656	0.581	1.283	1.288	1.208	1.229	1.690	2.392	0.439
Data Read Latency, t_{read} [ns]	1.909	1.428	1.497	0.789	3.262	2.074	2.715	3.378	2.536	9.537	1.234
Data Write Latency, t_{write} [ns] (set/ reset)	181.206 11.206	81.17/ 61.17	301.018/ 51.018	20.46/ 20.46	13.088	6.17	11.916	3.928	20.735	304.936	0.515
Cache Hit Dynamic Energy, $E_{dyn,hit}$ [nJ]	0.840	0.496	0.678	1.003	0.457	0.187	0.173	0.683	0.715	0.605	0.565
Cache Miss Dynamic Energy, $E_{dyn,miss}$ [nJ]	0.042	0.030	0.033	0.029	0.083	0.080	0.058	0.123	0.088	0.089	0.011
Cache Write Dynamic Energy, $E_{dyn,write}$ [nJ]	225.413	33.599	375.073	50.912	1.656	1.780	1.644	0.912	1.458	0.921	0.537
Cache Total Leakage Power [W]	0.062	0.100	0.061	0.137	0.661	0.025	0.295	0.828	3.896	9.000	3.438

All of the details in paper (and released models)!

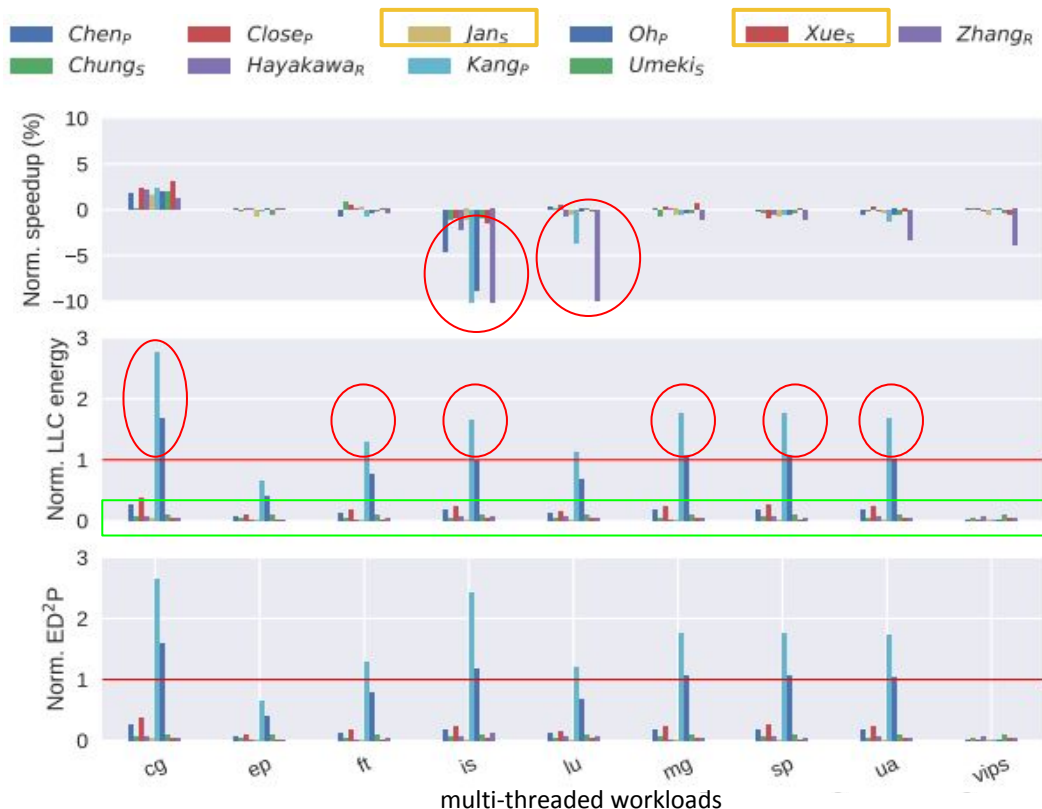
TABLE I: Gainestown LLC models generated by NVSim for *fixed-capacity* and *iso-area* models. The data write latency format is: set/ reset. Heatmap on per-row basis.

Comparison of NVM-based LLCs with SRAM Baseline

- Measured speedup and LLC energy for both iso-area and iso-capacity configs
- Workloads:
 - SPEC 2006, SPEC 2017 AI, PARSEC 3, Nasa Parallel Benchmarks
 - Includes single-threaded and multi-threaded workloads
 - Only the LLC intensive workloads

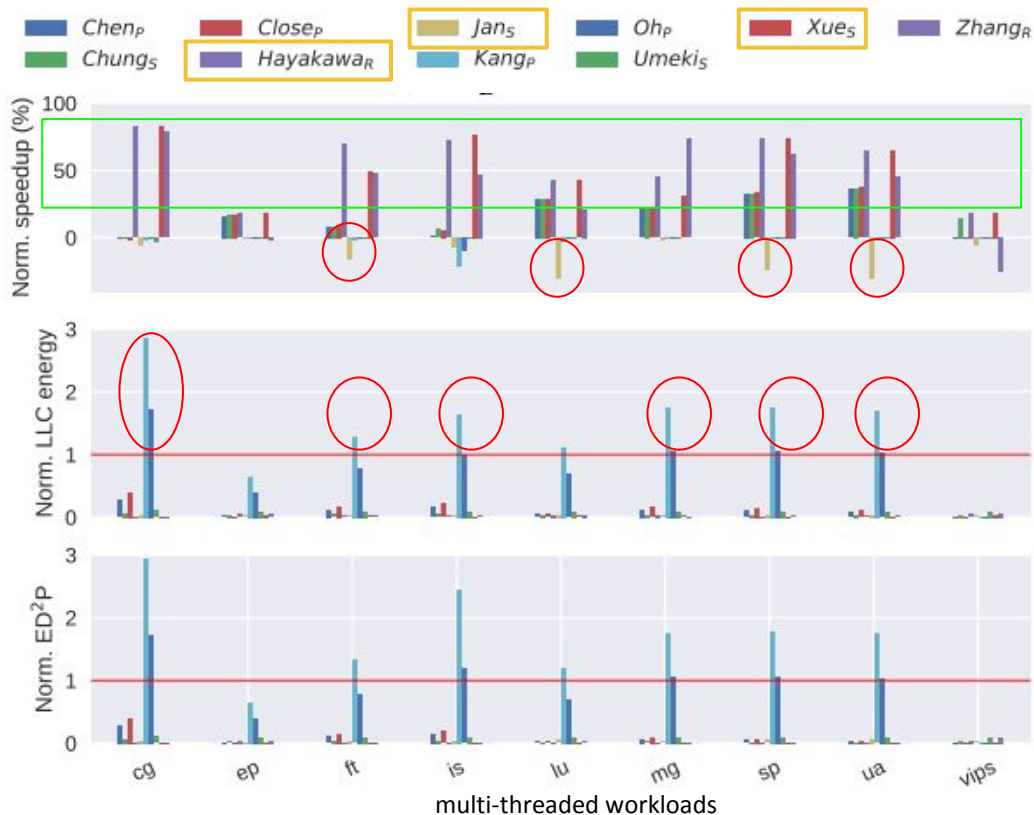
Simulation Results - Iso-capacity

- Multi-threaded workloads
- **Performance** can be a lot **worse** (up to 10%)
- **Energy** is generally a lot **better** (up to ~10X)
 - Except for PCRAM
- A couple NVMs are top-performing: Jan (STTRAM) and Xue (STTRAM)



Simulation Results - Iso-area

- Multi-threaded workloads
- **Big performance** enhancement (up to ~80%)!
- **Energy** about the **same** as **iso-capacity**
- Top performers: Hayakawa (ReRAM), Jan (STTTRAM), Xue (STTTRAM)



Simulation Results Takeaways

- Generally, for all workloads there are a **few** top performers:
 - Jan (STTRAM), Xue (STTRAM), Hayakawa (ReRAM)
 - Better energy efficiency
 - Positive speedup (for iso-area)
- **No perfect correlation** between cache parameters (**access latency/energy**) and **LLC energy/speedup...**

Workload Characterization

- Profiled workloads by capturing virtual memory addresses accessed in the execution of a workload ⁷
- Conducted ISA-independent characterization of workload memory behavior ⁸
- Proxy for cache sensitivity
- **Delineated metrics** based on memory access type (**read** or **write**)

⁷ M. Lui, K. Sangaiah, M. Hempstead, and B. Taskin, "Towards cross-framework workload analysis via flexible event-driven interfaces," in 2018 ISPASS, pp. 169–178.

⁸ Y. S. Shao and D. Brooks, "Isa-independent workload characterization and its implications for specialized architectures," in 2013 ISPASS, April 2013, pp. 245–255.

Memory Behavior Metrics

- **Global Address Entropy (H_{wg}, H_{rg})**
 - Randomness of accessed memory addresses
- **Local Address Entropy (H_{wl}, H_{rl})**
 - Spatial locality of regions of the memory address space
 - Skip M lowest order bits of address in entropy calculation (M=10)
- **90% Address Footprint ($90\%ft_w, 90\%ft_r$)**
 - Estimate of the working set
- **Unique Address Footprint (w_{unig}, r_{unig})**
- **Total Address Footprint (w_{total}, r_{total})**

Shannon Entropy Equation:

$$H = - \sum_i^N p(x_i) \cdot \log_2 p(x_i)$$

where N is total number of addresses and x_i is the i th address

Memory Behavior Data

**SPEC
2006
(s.t.)**

**SPEC
2017 AI
(s.t.)**

**PARSEC
3.0
(m.t.)**

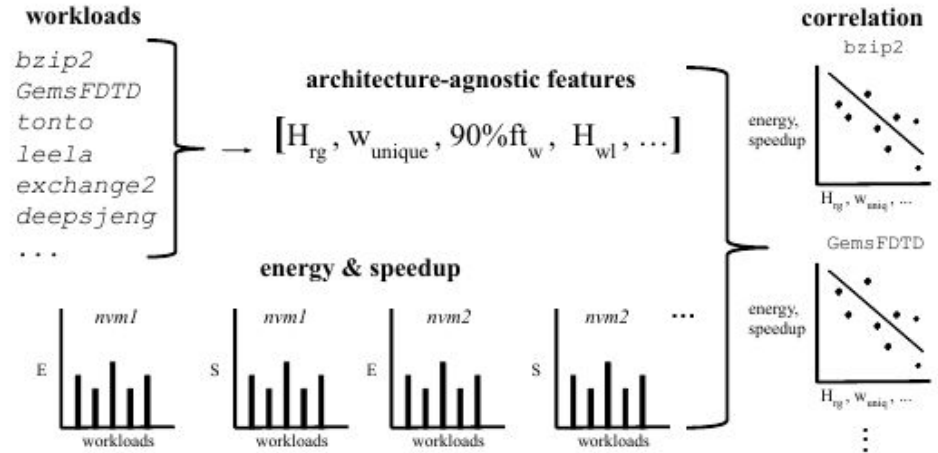
**NPB
3.3.1
(m.t.)**

	H_{rg}	H_{rl}	H_{wg}	H_{wl}	r_{unig} (E6)	w_{unig} (E6)	90%ft _r (E3)	90%ft _w (E3)	r_{total} (E9)	w_{total} (E9)
bzip2	18.0	10.2	11.7	5.9	6.0	5.9	2,505.4	750.9	4.3	1.5
GemsFDTD	19.9	13.6	22.3	15.0	116.9	143.6	76,576.6	113,183.5	1.3	0.7
leela	10.1	4.1	9.0	3.0	2.3	5.1	1.6	1.3	6.0	2.4
exchange2	8.8	3.5	8.6	3.5	0.03	0.02	0.6	0.6	62.3	42.9
deepsjeng	11.3	5.7	11.9	5.9	58.9	68.3	4.8	4.3	9.4	4.4
vips	15.2	10.3	17.8	11.6	12.0	6.3	1,107.2	1,325.3	1.9	0.7
x264	16.1	7.4	11.8	4.0	11.4	9.3	1,585.5	3.6	18.1	2.8
cg	19.0	11.7	18.9	12.0	2.3	2.4	1,015.4	819.2	0.7	0.04
ep	8.0	4.8	8.1	4.7	0.6	1.5	0.8	113.2	1.3	0.5

The rest of the memory behavior data is in the paper!

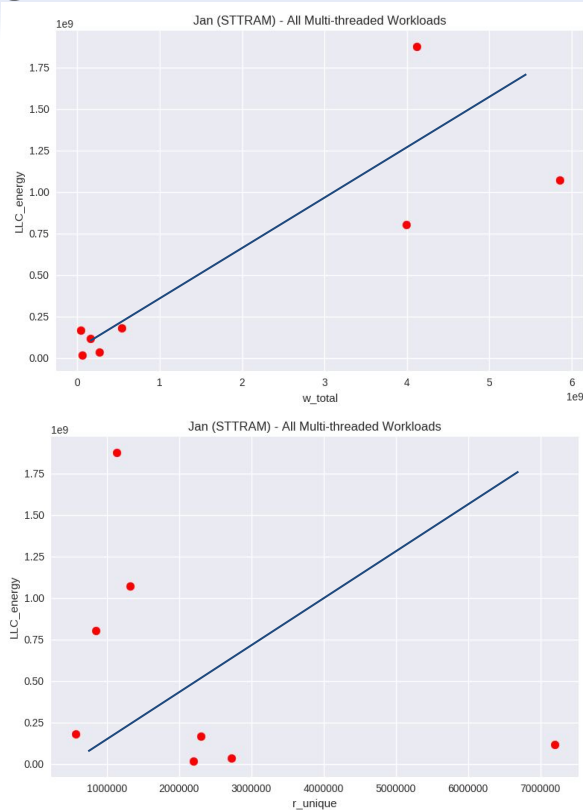
Correlation

- Observed correlation between architecture-agnostic memory behavior and energy/speedup
- For a “general purpose” system and “specialized” system



Correlation Results

- Performance and energy almost perfectly correlated with **total reads and writes** for **all multi-threaded workloads**
- However, SPEC 2017 AI benchmarks are different...
 - Performance and energy correlated with **other architecture-agnostic features**:
 - Unique accesses, entropy, and 90% footprint
 - Total reads and writes has almost **no correlation**
 - SPEC 2017 AI may not be an indicator of AI in general



Summary

- Publicly available NVM models:
<http://sites.tufts.edu/tcal/nvm-models>
- Workload characterization of LLC intensive SPEC 2006, SPEC 2017 AI, PARSEC, and NPB workload memory behavior per access type (read, write)
- Correlation between memory behavior and best NVM technology for LLC

Evaluation of Non-Volatile Memory Based Last Level Cache Given Modern Use Case Behavior

Alex Hankin*, Tomer Shapira*, Karthik Sangaiah[†], Mike Lui[†], Mark Hempstead*

*TCAL, Tufts University [†]VANDAL, Drexel University

IISWC 2019, Orlando, Fl., November 4th, 2019

- Publicly available NVM models: <http://sites.tufts.edu/tcal/nvm-models>
- Workload characterization of LLC intensive SPEC 2006, SPEC 2017 AI, PARSEC, and NPB workload memory behavior per access type (read, write)
- Correlation between memory behavior and best NVM technology for LLC

NVM Cells

	Oh	Chen	Kang	Close	Chung	Jan	Umeki	Xue	Hayakawa	Zhang
class	PCRAM	PCRAM	PCRAM	PCRAM	STTRAM	STTRAM	STTRAM	STTRAM	RRAM	RRAM
year	2005	2006	2006	2013	2010	2014	2015	2016	2015	2016
access device	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
cell size [F^2]	16.6*	10*	16.6	25	14	50	48 [†]	63	4*	4*
cell levels	1	1	1	2	1	1	1	2	1	1
read current [μA]	40*	40*	60*	60*						
read voltage [V]					0.65	0.08	0.38	1.2	0.4*	0.2
read power [μW]					24.1 [†]	30*	1.70	65	0.16*	0.02
read energy [pJ]	2*	2*	2*	2*						
reset current [μA]	600	90	600	400	80	52	255 [†]	150		
reset voltage [V]									2*	1
reset pulse [ns]	10	60	50	20	10	4	10	2	10*	150
reset energy [pJ]					0.52 [†]	1*	1.12	0.36	0.6*	0.4
set current [μA]	200	55	200*	400	100 [†]	38	255 [†]	150		
set voltage [V]									2*	1
set pulse [ns]	180	80	300	20	10	4.5	10	2	10*	150
set energy [pJ]					0.75 [†]	1*	1.12	0.36	0.6*	0.4

TABLE I: NVM Parameters (* indicates parameters not found in cited paper; [†] indicates parameters derived from other known

Iso-capacity and Iso-area Models

Iso-
capacity

	Ohp	Chenp	Kangp	Closep	Chungs	Jan _s	Umeki _s	Xue _s	Hayakawa _r	Zhang _r	SRAM
Area [mm²]	6.847	4.104	4.591	2.855	1.452	9.171	4.348	1.585	0.915	0.307	6.548
Tag Access Latency [ns]	0.74	0.604	0.656	0.582	1.240	1.423	1.208	1.156	1.396	1.722	0.439
Data Read Latency, t_{read} [ns]	1.907	0.607	1.497	0.82	1.763	3.072	2.715	2.878	1.722	2.16	1.234
Data Write Latency, t_{write} [ns] (set/ reset)	181.206 /11.206	80.491 /60.491	301.018 /51.018	20.681 /20.681	11.751	7.878	11.916	4.038	20.716	300.834	0.515
Cache Hit Dynamic Energy, $E_{dyn,hit}$ [nJ]	0.840	0.421	0.678	0.437	0.209	0.188	0.173	0.251	0.263	0.217	0.565
Cache Miss Dynamic Energy, $E_{dyn,miss}$ [nJ]	0.042	0.025	0.033	0.023	0.082	0.077	0.058	0.121	0.078	0.086	0.011
Cache Write Dynamic Energy, $E_{dyn,write}$ [nJ]	225.413	34.108	375.073	51.116	1.332	2.305	1.644	0.597	0.952	0.523	0.537
Cache Total Leakage Power [W]	0.062	0.071	0.061	0.039	0.166	0.048	0.295	0.115	0.194	0.151	3.438

Iso-
area

Capacity [MB]	2	4	2	4	8	1	2	8	32	128	2
Tag Access Latency [ns]	0.740	0.607	0.656	0.581	1.283	1.288	1.208	1.229	1.690	2.392	0.439
Data Read Latency, t_{read} [ns]	1.909	1.428	1.497	0.789	3.262	2.074	2.715	3.378	2.536	9.537	1.234
Data Write Latency, t_{write} [ns] (set/ reset)	181.206 11.206	81.17/ 61.17	301.018/ 51.018	20.46/ 20.46	13.088	6.17	11.916	3.928	20.735	304.936	0.515
Cache Hit Dynamic Energy, $E_{dyn,hit}$ [nJ]	0.840	0.496	0.678	1.003	0.457	0.187	0.173	0.683	0.715	0.605	0.565
Cache Miss Dynamic Energy, $E_{dyn,miss}$ [nJ]	0.042	0.030	0.033	0.029	0.083	0.080	0.058	0.123	0.088	0.089	0.011
Cache Write Dynamic Energy, $E_{dyn,write}$ [nJ]	225.413	33.599	375.073	50.912	1.656	1.780	1.644	0.912	1.458	0.921	0.537
Cache Total Leakage Power [W]	0.062	0.100	0.061	0.137	0.661	0.025	0.295	0.828	3.896	9.000	3.438

TABLE I: Gainestown LLC models generated by NVSim for *fixed-capacity* (top) LLC, *fixed-area* (bottom) LLC. For PCRAM data write latency format is: set/ reset. Heatmap on per-row basis.

Memory Behavior Data

	H_{r_g}	H_{r_l}	H_{w_g}	H_{w_l}	r_{uniq} (10^6)	w_{uniq} (10^6)	$90\%ft_r$ (10^3)	$90\%ft_w$ (10^3)	r_{total} (10^9)	w_{total} (10^9)
bzip2	18.03	10.23	11.72	5.90	5.99	5.88	2505.38	750.86	4.30	1.47
GemsFDTD	19.92	13.62	22.27	14.99	116.88	143.63	76576.59	113183.50	1.30	0.70
tonto	10.97	5.15	10.25	3.72	0.30	0.29	5.59	1.74	1.10	0.47
leela	10.13	4.07	8.95	3.01	2.26	5.06	1.59	1.29	6.01	2.35
exchange2	8.79	3.52	8.61	3.47	0.03	0.02	0.64	0.58	62.28	42.89
deepsjeng	11.31	5.69	11.86	5.93	58.89	68.28	4.79	4.33	9.36	4.43
vips	15.17	10.26	17.79	11.61	12.02	6.32	1107.19	1325.34	1.91	0.68
x264	16.14	7.43	11.84	4.04	11.40	9.28	1585.49	3.56	18.07	2.84
cg	19.01	11.71	18.88	11.96	2.30	2.36	1015.43	819.15	0.73	0.04
ep	8.00	4.81	8.05	4.74	0.563	1.47	0.84	113.18	1.25	0.54
ft	16.47	9.93	17.07	10.28	2.73	2.72	342.64	611.66	0.28	0.27
is	15.23	8.96	15.65	8.69	2.20	2.19	1228.86	794.26	0.12	0.06
lu	9.57	6.01	16.02	9.63	0.844	0.84	289.46	259.75	17.84	3.99
mg	17.97	11.80	16.93	10.18	7.20	7.29	4249.78	4767.97	0.76	0.16
sp	18.69	12.02	18.21	11.35	1.14	1.28	556.75	256.73	9.23	4.12
ua	13.95	8.17	11.23	5.69	1.32	1.57	362.45	106.25	9.97	5.85

TABLE IV: Workload Features. Heatmap on per-column basis. H_{r_g} =global read entropy, H_{r_l} =local read entropy, H_{w_g} =global write entropy, H_{w_l} =local write entropy, r_{unique} =number of unique reads, w_{unique} =number of unique writes, $90\%ft_r$ =90% read footprint, $90\%ft_w$ =90% write footprint, r_{total} =total number of reads, w_{total} =total number of writes

Correlation Results - AI Workloads

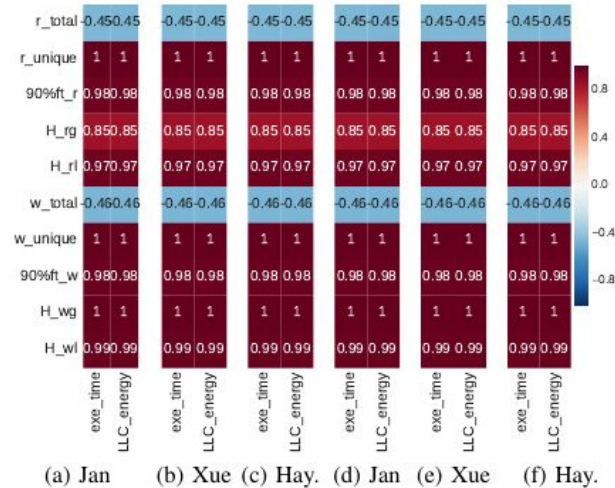


Fig. 4: Feature correlation with energy and speedup for AI benchmarks with *fixed-capacity* (a)-(c) and *fixed-area* (d)-(f) LLC