18-742 Research in Parallel Computer Architecture, Fall 2014
Paper for In-Class Discussion

Overview - memory

- Onur Mutlu. Memory scaling: A systems architecture perspective. In *MemCon*, 2013

Cache

- George Kurian, Srinivas Devadas, and Omer Khan. Locality-aware data replication in the last-level cache. In *HPCA*, pages 1–12, 2014
- Vivek Seshadri, Abhishek Bhowmick, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. The dirty-block index. In *ISCA*, pages 157–168, 2014
- Akanksha Jain and Calvin Lin. Linearizing irregular memory accesses for improved correlated prefetching. In *MICRO*, pages 247–259, 2013
- Abhisek Pan and Vijay S. Pai. Imbalanced cache partitioning for balanced data-parallel programs. In *MICRO*, pages 297–309, 2013
- Arthur Perais and André Seznec. Practical data value speculation for future high-end processors. In *HPCA*, pages 428–439, 2014

Memory system

- Gennady Pekhimenko, Vivek Seshadri, Yoongu Kim, Hongyi Xin, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. Linearly compressed pages: a low-complexity, low-latency main memory compression framework. In *MICRO*, pages 172–184, 2013
- Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Yixin Luo, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. Rowclone: fast and energy-efficient in-dram bulk data copy and initialization. In *MICRO*, pages 185–197, 2013
- Ali Shafiee, Meysam Taassori, Rajeev Balasubramonian, and Al Davis. Memzip: Exploring unconventional benefits from memory compression. In *HPCA*, pages 638–649, 2014
- Binh Pham, Abhishek Bhattacharjee, Yasuko Eckert, and Gabriel H. Loh. Increasing tlb reach by exploiting clustering in page translations. In *HPCA*, pages 558–567, 2014

DRAM and DRAM architecture

- Uksong Kang, Hak-soo Yu, Churoo Park, Hongzhong Zheng, John Halbert, Kuljit Bains, SeongJin Jang, and Joo Sun Choi. Co-architecting controllers and dram to enhance dram process scaling. 2014
- Yoongu Kim, Vivek Seshadri, Donghyuk Lee, Jamie Liu, and Onur Mutlu. A case for exploiting subarray-level parallelism (salp) in dram. In *ISCA*, 2012
- Yoongu Kim, Ross Daly, Jeremie Kim, Chris Fallin, Ji-Hye Lee, Donghyuk Lee, Chris Wilkerson, Konrad Lai, and Onur Mutlu. Flipping bits in memory without accessing them: An experimental study of dram disturbance errors. In *ISCA*, 2014
- Tao Zhang, Ke Chen, Cong Xu, Guangyu Sun, Tao Wang, and Yuan Xie. Half-dram: A high-bandwidth and low-power dram architecture from the rethinking of fine-grained activation. In *ISCA*, 2014

- Kevin Kai-Wei Chang, Donghyuk Lee, Zeshan Chishti, Alaa R. Alameldeen, Chris Wilkerson, Yoongu Kim, and Onur Mutlu. Improving dram performance by parallelizing refreshes with accesses. In *HPCA*, pages 356–367, 2014
- Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, Lavanya Subramanian, and Onur Mutlu. Tiered-latency dram: A low latency and low cost dram architecture. In *HPCA*, pages 615–626, 2013
- Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Yixin Luo, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. Rowclone: fast and energy-efficient in-dram bulk data copy and initialization. In *MICRO*, pages 185–197, 2013
- Samira Manabi Khan, Donghyuk Lee, Yoongu Kim, Alaa R. Alameldeen, Chris Wilkerson, and Onur Mutlu. The efficacy of error mitigation techniques for dram retention failures: a comparative experimental study. In *SIGMETRICS*, pages 519–532, 2014

Flash memory

- Yu Cai, Gulay Yalcin, Onur Mutlu, Erich F. Haratsch, Osman S. Unsal, Adrián Cristal, and Ken Mai. Neighbor-cell assisted error correction for mlc nand flash memories. In *SIGMETRICS*, 2014
- Jian Ouyang, Shiding Lin, Jiang Song, Zhenyu Hou, Yong Wang, and Yuanzheng Wang. Sdf: software-defined flash for web-scale internet storage systems. In *ASPLOS*, pages 471–484, 2014
- Kai Zhao, Kalyana S. Venkataraman, Xuebin Zhang, Jiangpeng Li, Ning Zheng, and Tong Zhang. Over-clocked ssd: Safely running beyond flash memory chip i/o clock specs. In *HPCA*, pages 536–545, 2014

Emerging memories

- Ren-Shuo Liu, De-Yu Shen, Chia-Lin Yang, Shun-Chih Yu, and Cheng-Yuan Michael Wang. Nvm duet: unified working memory and persistent store architecture. In *ASPLOS*, 2014
- Jishen Zhao, Sheng Li, Doe Hyun Yoon, Yuan Xie, and Norman P Jouppi. Kiln: closing the performance gap between systems with and without persistence support. In *ISCA*, 2013
- Steven Pelley, Peter M. Chen, and Thomas F. Wenisch. Memory persistency. In *ISCA*, pages 265–276, 2014

Parallelism

- Todd Mytkowicz, Madanlal Musuvathi, and Wolfram Schulte. Data-parallel finite-state machines. In *ASPLOS*, pages 529–542, 2014
- Amos Waterland, Elaine Angelino, Ryan P. Adams, Jonathan Appavoo, and Margo I. Seltzer. Asc: automatically scalable computation. In *ASPLOS*, pages 575–590, 2014
- Stijn Eyerman and Lieven Eeckhout. The benefit of smt in the multi-core era: flexibility towards degrees of thread-level parallelism. In *ASPLOS*, pages 591–606, 2014

GPU

- Marc S. Orr, Bradford M. Beckmann, Steven K. Reinhardt, and David A. Wood. Fine-grain task aggregation and coordination on gpus. In *ISCA*, pages 181–192, 2014
- Syed Zohaib Gilani, Nam Sung Kim, and Michael J. Schulte. Exploiting gpu peak-power and performance tradeoffs through reduced effective pipeline latency. In *MICRO*, pages 74–85, 2013
- Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt. Divergence-aware warp scheduling. In *MICRO*, pages 99–110, 2013
- Cedric Nugteren, Gert-Jan van den Braak, Henk Corporaal, and Henri Bal. A detailed gpu cache model based on reuse distance theory. In *HPCA*, pages 37–48, 2014

Heterogeneous multicore

- José A. Joao, M. Aater Suleman, Onur Mutlu, and Yale N. Patt. Utility-based acceleration of multithreaded applications on asymmetric cmps. In *ISCA*, 2013

- Kenzo Van Craeynest, Shoaib Akram, Wim Heirman, Aamer Jaleel, and Lieven Eeckhout. Fairness-aware scheduling on single-isa heterogeneous multi-cores. In *PACT*, 2013

- Shruti Padmanabha, Andrew Lukefahr, Reetuparna Das, and Scott A. Mahlke. Trace based phase prediction for tightly-coupled heterogeneous cores. In *MICRO*, pages 445–456, 2013

Accelerators

- Renée St. Amant, Amir Yazdanbakhsh, Jongse Park, Bradley Thwaites, Hadi Esmaeilzadeh, Arjang Hassibi, Luis Ceze, and Doug Burger. General-purpose code acceleration with limited-precision analog computation. In *ISCA*, pages 505–516, 2014

- Andrew Putnam, Adrian M. Caulfield, Eric S. Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Jan Gray, Michael Haselman, Scott Hauck, Stephen Heil, Amir Hormati, Joo-Young Kim, Sitaram Lanka, James R. Larus, Eric Peterson, Simon Pope, Aaron Smith, Jason Thong, Phillip Yi Xiao, and Doug Burger. A reconfigurable fabric for accelerating large-scale datacenter services. In *ISCA*, pages 13–24, 2014

- Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. Diannao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. In *ASPLOS*, pages 269–284, 2014

- Yusuf Onur Koçberber, Boris Grot, Javier Picorel, Babak Falsafi, Kevin T. Lim, and Parthasarathy Ranganathan. Meet the walkers: accelerating index traversals for in-memory databases. In *MICRO*, pages 468–479, 2013

Coherence and consistency

- Heiner Litz, David R. Cheriton, Amin Firoozshahian, Omid Azizi, and John P. Stevenson. Si-tm: reducing transactional memory abort rates through snapshot isolation. In *ASPLOS*, pages 383–398, 2014

- Blake A. Hechtman and Daniel J. Sorin. Exploring memory consistency for massively-threaded throughput-oriented processors. In *ISCA*, pages 201–212, 2013

Quality of service

- Harshad Kasture and Daniel Sanchez. Ubik: efficient cache sharing with strict qos for latency-critical workloads. In *ASPLOS*, pages 729–742, 2014

- Lavanya Subramanian, Vivek Seshadri, Yoongu Kim, Ben Jaiyen, and Onur Mutlu. Mise: Providing performance predictability and improving fairness in shared main memory systems. In *HPCA*, pages 639–650, 2013

Approximate computing

- Mehrzad Samadi, Davoud Anoushe Jamshidi, Janghaeng Lee, and Scott A. Mahlke. Paraprox: pattern-based approximation for data parallel applications. In *ASPLOS*, pages 35–50, 2014

- Mehrzad Samadi, Janghaeng Lee, Davoud Anoushe Jamshidi, Amir Hormati, and Scott A. Mahlke. Sage: self-tuning approximation for graphics engines. In *MICRO*, pages 13–24, 2013

- Adrian Sampson, Jacob Nelson, Karin Strauss, and Luis Ceze. Approximate storage in solid-state memories. In *MICRO*, pages 25–36, 2013

Resilience

- Long Chen and Zhao Zhang. Memguard: A low cost and energy efficient design to support and enhance memory system reliability. In *ISCA*, pages 49–60, 2014

- David J. Palframan, Nam Sung Kim, and Mikko H. Lipasti. Precision-aware soft error protection for gpus. In *HPCA*, pages 49–59, 2014

Security

- Lluís Vilanova, Muli Ben-Yehuda, Nacho Navarro, Yoav Etsion, and Mateo Valero. Codoms: Protecting software with code-centric memory domains. In *ISCA*, pages 469–480, 2014

Virtualization

- Nadav Amit, Dan Tsafrir, and Assaf Schuster. Vswapper: a memory swapper for virtualized environments. In *ASPLOS*, pages 349–366, 2014
- Canturk Isci, Suzanne McIntosh, Jeffrey O. Kephart, Rajarshi Das, James E. Hanson, Scott Piper, Robert R. Wolford, Thomas Brey, Robert Kantner, Allen Ng, James Norris, Abdoulaye Traore, and Michael Frissora. Agile, efficient virtualization power management with low-latency server power states. In *ISCA*, pages 96–107, 2013
- Xiaotao Chang, Hubertus Franke, Yi Ge, Tao Liu, Kun Wang, Jimi Xenidis, Fei Chen, and Yu Zhang. Improving virtualization in the presence of software managed translation lookaside buffers. In *ISCA*, pages 120–129, 2013
- Ardalan Amiri Sani, Kevin Boos, Shaopu Qin, and Lin Zhong. I/o paravirtualization at the device file boundary. In *ASPLOS*, pages 319–332, 2014

References

[1] Renée St. Amant, Amir Yazdanbakhsh, Jongse Park, Bradley Thwaites, Hadi Esmaeilzadeh, Arjang Hassibi, Luis Ceze, and Doug Burger. General-purpose code acceleration with limited-precision analog computation. In *ISCA*, pages 505–516, 2014.

[2] Nadav Amit, Dan Tsafrir, and Assaf Schuster. Vswapper: a memory swapper for virtualized environments. In *ASPLOS*, pages 349–366, 2014.

[3] Yu Cai, Gulay Yalcin, Onur Mutlu, Erich F. Haratsch, Osman S. Unsal, Adrián Cristal, and Ken Mai. Neighbor-cell assisted error correction for mlc nand flash memories. In *SIGMETRICS*, 2014.

[4] Kevin Kai-Wei Chang, Donghyuk Lee, Zeshan Chishti, Alaa R. Alameldeen, Chris Wilkerson, Yoongu Kim, and Onur Mutlu. Improving dram performance by parallelizing refreshes with accesses. In *HPCA*, pages 356–367, 2014.

[5] Xiaotao Chang, Hubertus Franke, Yi Ge, Tao Liu, Kun Wang, Jimi Xenidis, Fei Chen, and Yu Zhang. Improving virtualization in the presence of software managed translation lookaside buffers. In *ISCA*, pages 120–129, 2013.

[6] Long Chen and Zhao Zhang. Memguard: A low cost and energy efficient design to support and enhance memory system reliability. In *ISCA*, pages 49–60, 2014.

[7] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. Diannao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. In *ASPLOS*, pages 269–284, 2014.

[8] Kenzo Van Craeynest, Shoaib Akram, Wim Heirman, Aamer Jaleel, and Lieven Eeckhout. Fairness-aware scheduling on single-isa heterogeneous multi-cores. In *PACT*, 2013.

[9] Stijn Eyerman and Lieven Eeckhout. The benefit of smt in the multi-core era: flexibility towards degrees of thread-level parallelism. In *ASPLOS*, pages 591–606, 2014.

[10] Syed Zohaib Gilani, Nam Sung Kim, and Michael J. Schulte. Exploiting gpu peak-power and performance tradeoffs through reduced effective pipeline latency. In *MICRO*, pages 74–85, 2013.

[11] Blake A. Hechtman and Daniel J. Sorin. Exploring memory consistency for massively-threaded throughput-oriented processors. In *ISCA*, pages 201–212, 2013.

[12] Canturk Isci, Suzanne McIntosh, Jeffrey O. Kephart, Rajarshi Das, James E. Hanson, Scott Piper, Robert R. Wolford, Thomas Brey, Robert Kantner, Allen Ng, James Norris, Abdoulaye Traore, and Michael Frissora. Agile, efficient virtualization power management with low-latency server power states. In *ISCA*, pages 96–107, 2013.

[13] Akanksha Jain and Calvin Lin. Linearizing irregular memory accesses for improved correlated prefetching. In *MICRO*, pages 247–259, 2013.

[14] José A. Joao, M. Aater Suleman, Onur Mutlu, and Yale N. Patt. Utility-based acceleration of multithreaded applications on asymmetric cmps. In *ISCA*, 2013.

[15] Uksong Kang, Hak-soo Yu, Churoo Park, Hongzhong Zheng, John Halbert, Kuljit Bains, SeongJin Jang, and Joo Sun Choi. Co-architecting controllers and dram to enhance dram process scaling. 2014.

[16] Harshad Kasture and Daniel Sanchez. Ubik: efficient cache sharing with strict qos for latency-critical workloads. In *ASPLOS*, pages 729–742, 2014.

[17] Samira Manabi Khan, Donghyuk Lee, Yoongu Kim, Alaa R. Alameldeen, Chris Wilkerson, and Onur Mutlu. The efficacy of error mitigation techniques for dram retention failures: a comparative experimental study. In *SIGMETRICS*, pages 519–532, 2014.

[18] Yoongu Kim, Ross Daly, Jeremie Kim, Chris Fallin, Ji-Hye Lee, Donghyuk Lee, Chris Wilkerson, Konrad Lai, and Onur Mutlu. Flipping bits in memory without accessing them: An experimental study of dram disturbance errors. In *ISCA*, 2014.

[19] Yoongu Kim, Vivek Seshadri, Donghyuk Lee, Jamie Liu, and Onur Mutlu. A case for exploiting subarray-level parallelism (salp) in dram. In *ISCA*, 2012.

[20] Yusuf Onur Koçberber, Boris Grot, Javier Picorel, Babak Falsafi, Kevin T. Lim, and Parthasarathy Ranganathan. Meet the walkers: accelerating index traversals for in-memory databases. In *MICRO*, pages 468–479, 2013.

[21] George Kurian, Srinivas Devadas, and Omer Khan. Locality-aware data replication in the last-level cache. In *HPCA*, pages 1–12, 2014.

[22] Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, Lavanya Subramanian, and Onur Mutlu. Tiered-latency dram: A low latency and low cost dram architecture. In *HPCA*, pages 615–626, 2013.

[23] Heiner Litz, David R. Cheriton, Amin Firoozshahian, Omid Azizi, and John P. Stevenson. Si-tm: reducing transactional memory abort rates through snapshot isolation. In *ASPLOS*, pages 383–398, 2014.

[24] Ren-Shuo Liu, De-Yu Shen, Chia-Lin Yang, Shun-Chih Yu, and Cheng-Yuan Michael Wang. Nvm duet: unified working memory and persistent store architecture. In *ASPLOS*, 2014.

[25] Onur Mutlu. Memory scaling: A systems architecture perspective. In *MemCon*, 2013.

[26] Todd Mytkowicz, Madanlal Musuvathi, and Wolfram Schulte. Data-parallel finite-state machines. In *ASPLOS*, pages 529–542, 2014.

[27] Cedric Nugteren, Gert-Jan van den Braak, Henk Corporaal, and Henri Bal. A detailed gpu cache model based on reuse distance theory. In *HPCA*, pages 37–48, 2014.

[28] Marc S. Orr, Bradford M. Beckmann, Steven K. Reinhardt, and David A. Wood. Fine-grain task aggregation and coordination on gpus. In *ISCA*, pages 181–192, 2014.

[29] Jian Ouyang, Shiding Lin, Jiang Song, Zhenyu Hou, Yong Wang, and Yuanzheng Wang. Sdf: software-defined flash for web-scale internet storage systems. In *ASPLOS*, pages 471–484, 2014.

[30] Shruti Padmanabha, Andrew Lukefahr, Reetuparna Das, and Scott A. Mahlke. Trace based phase prediction for tightly-coupled heterogeneous cores. In *MICRO*, pages 445–456, 2013.

[31] David J. Palframan, Nam Sung Kim, and Mikko H. Lipasti. Precision-aware soft error protection for gpus. In *HPCA*, pages 49–59, 2014.

[32] Abhisek Pan and Vijay S. Pai. Imbalanced cache partitioning for balanced data-parallel programs. In *MICRO*, pages 297–309, 2013.

[33] Gennady Pekhimenko, Vivek Seshadri, Yoongu Kim, Hongyi Xin, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. Linearly compressed pages: a low-complexity, low-latency main memory compression framework. In *MICRO*, pages 172–184, 2013.

[34] Steven Pelley, Peter M. Chen, and Thomas F. Wenisch. Memory persistency. In *ISCA*, pages 265–276, 2014.

[35] Arthur Perais and André Seznec. Practical data value speculation for future high-end processors. In *HPCA*, pages 428–439, 2014.

[36] Binh Pham, Abhishek Bhattacharjee, Yasuko Eckert, and Gabriel H. Loh. Increasing tlb reach by exploiting clustering in page translations. In *HPCA*, pages 558–567, 2014.

[37] Andrew Putnam, Adrian M. Caulfield, Eric S. Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Jan Gray, Michael Haselman, Scott Hauck, Stephen Heil, Amir Hormati, Joo-Young Kim, Sitaram Lanka, James R. Larus, Eric Peterson, Simon Pope, Aaron Smith, Jason Thong, Phillip Yi Xiao, and Doug Burger. A reconfigurable fabric for accelerating large-scale datacenter services. In *ISCA*, pages 13–24, 2014.

[38] Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt. Divergence-aware warp scheduling. In *MICRO*, pages 99–110, 2013.

[39] Mehrzad Samadi, Davoud Anoushe Jamshidi, Janghaeng Lee, and Scott A. Mahlke. Paraprox: pattern-based approximation for data parallel applications. In *ASPLOS*, pages 35–50, 2014.

[40] Mehrzad Samadi, Janghaeng Lee, Davoud Anoushe Jamshidi, Amir Hormati, and Scott A. Mahlke. Sage: self-tuning approximation for graphics engines. In *MICRO*, pages 13–24, 2013.

[41] Adrian Sampson, Jacob Nelson, Karin Strauss, and Luis Ceze. Approximate storage in solid-state memories. In *MICRO*, pages 25–36, 2013.

[42] Ardalan Amiri Sani, Kevin Boos, Shaopu Qin, and Lin Zhong. I/o paravirtualization at the device file boundary. In *ASPLOS*, pages 319–332, 2014.

[43] Vivek Seshadri, Abhishek Bhowmick, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. The dirty-block index. In *ISCA*, pages 157–168, 2014.

[44] Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Yixin Luo, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, and Todd C. Mowry. Rowclone: fast and energy-efficient in-dram bulk data copy and initialization. In *MICRO*, pages 185–197, 2013.

[45] Ali Shafiee, Meysam Taassori, Rajeev Balasubramonian, and Al Davis. Memzip: Exploring unconventional benefits from memory compression. In *HPCA*, pages 638–649, 2014.

[46] Lavanya Subramanian, Vivek Seshadri, Yoongu Kim, Ben Jaiyen, and Onur Mutlu. Mise: Providing performance predictability and improving fairness in shared main memory systems. In *HPCA*, pages 639–650, 2013.

[47] Lluís Vilanova, Muli Ben-Yehuda, Nacho Navarro, Yoav Etsion, and Mateo Valero. Codoms: Protecting software with code-centric memory domains. In *ISCA*, pages 469–480, 2014.

[48] Amos Waterland, Elaine Angelino, Ryan P. Adams, Jonathan Appavoo, and Margo I. Seltzer. Asc: automatically scalable computation. In *ASPLOS*, pages 575–590, 2014.

[49] Tao Zhang, Ke Chen, Cong Xu, Guangyu Sun, Tao Wang, and Yuan Xie. Half-dram: A high-bandwidth and low-power dram architecture from the rethinking of fine-grained activation. In *ISCA*, 2014.

[50] Jishen Zhao, Sheng Li, Doe Hyun Yoon, Yuan Xie, and Norman P Jouppi. Kiln: closing the performance gap between systems with and without persistence support. In *ISCA*, 2013.

[51] Kai Zhao, Kalyana S. Venkataraman, Xuebin Zhang, Jiangpeng Li, Ning Zheng, and Tong Zhang. Over-clocked ssd: Safely running beyond flash memory chip i/o clock specs. In *HPCA*, pages 536–545, 2014.