1. Benchmarks

1.1 TailBench: A Benchmark Suite and Evaluation Methodology for Latency-Critical Applications [IISWC 2016]

- https://people.csail.mit.edu/sanchez/papers/2016.tailbench.iiswc.pdf
- https://people.csail.mit.edu/sanchez/papers/2016.tailbench.iiswc.slides.pdf
- http://tailbench.csail.mit.edu/

1.2 GARDENIA: A Domain-specific Benchmark Suite for Next-generation Accelerators

- https://arxiv.org/pdf/1708.04567.pdf
- https://github.com/chenxuhao/gardenia

1.3 Pmem + WHISPER [ASPLOS 2018]

- http://research.cs.wisc.edu/multifacet/papers/asplos17_whisper.pdf
- http://pmem.io/about/

2. Branch Predictors

2.1 The Inner Most Loop Iteration counter: a new dimension in branch history [MICRO 2015]

- https://dl.acm.org/citation.cfm?id=2830831

2.2 Wormhole: Wisely Predicting Multidimensional Branches [MICRO 2014]

- https://dl.acm.org/citation.cfm?id=2742207

2.3 5th JILP Workshop on Computer Architecture Competitions (JWAC-5): Championship Branch Prediction (CBP-5)

- https://www.jilp.org/cbp2016/

2.4 Bias-Free Branch Predictor

- https://dl.acm.org/citation.cfm?id=2742208

3. Caches

3.1 Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement [ISCA 2016]

- https://www.cs.utexas.edu/~lin/papers/isca16.pdf

3.2 Compute Caches [HPCA 2017]

- http://web.eecs.umich.edu/~reetudas/papers/compute_cache.pdf

3.3 Perceptron Learning for Reuse Prediction [MICRO 2016]

- http://hpca23.cse.tamu.edu/pdfs/micro2016-perceptron.pdf
- https://github.com/nautilusPrime/Perceptron-learning-Cache-Reuse-Predictor

3.4 The Bunker Cache for Spatio-Value Approximation [MICRO 2016]

- http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7783746
- http://www.eecg.toronto.edu/~sanmigu2/sanmiguel-micro2016-presentation.pdf

3.5 Best-Offset Hardware Prefetching [HPCA 2016]

- https://hal.inria.fr/hal-01254863/document

3.6 Doppelgänger: A Cache for Approximate Computing [MICRO 2015]

- http://www.eecg.toronto.edu/~sanmigu2/sanmiguel-micro2015.pdf
- https://www.microarch.org/micro48/files/slides/A1-2.pdf

3.7 Efficiently Prefetching Complex Address Patterns [MICRO 2015]

- https://www.cs.utah.edu/~rajeev/pubs/micro15m.pdf

3.8 IMP: Indirect Memory Prefetcher [MICRO 2015]

- http://people.csail.mit.edu/yxy/pubs/imp.pdf
- https://www.microarch.org/micro48/files/slides/B1-4.pptx

3.9 Cache-Guided Scheduling: Exploiting Caches to Maximize Locality in Graph Processing [AGP 2017]

- https://people.csail.mit.edu/sanchez/papers/2017.cgs.agp.pdf

3.10 Translation-Triggered Prefetching [ASPLOS 2017]

- https://dl.acm.org/citation.cfm?id=3037705

3.11 Jenga: Software-Defined Cache Hierarchies [ISCA 2017]

- http://people.csail.mit.edu/poantsai/papers/2017.jenga.isca.pdf

4. Cache Partitioning

4.1 KPart: A Hybrid Cache Partitioning-Sharing Technique for Commodity Multicores [HPCA 2018]

- http://people.csail.mit.edu/sanchez/papers/2018.kpart.hpca.pdf
- https://github.com/Nosayba/kpart

4.2 Whirlpool: Improving Dynamic Cache Management with Static Data Classification [ASPLOS 2016]

- http://people.csail.mit.edu/beckmann/publications/papers/2016.asplos.whirlpool.pdf

5. TLBs

5.1 Hybrid TLB Coalescing: Improving TLB Translation Coverage under Diverse Fragmented Memory Allocations [ISCA 2017]

- http://calab.kaist.ac.kr:8080/~jhuh/papers/park_isca17.pdf

5.2 Devirtualizing Memory in Heterogeneous Systems [ASPLOS 2018]

- http://research.cs.wisc.edu/multifacet/papers/asplos18_dvm.pdf

5.3 CSALT: Context Switch Aware Large TLB [MICRO 2017]

- https://lca.ece.utexas.edu/pubs/csalt.pdf

5.4 Supporting Address Translation for Accelerator-Centric Architectures [HPCA 2017]

- http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7920812
- https://github.com/cdsc-github/parade-ara-simulator

5.5. Agile Paging: Exceeding the Best of Nested and Shadow Paging [ISCA 2016]

- http://research.cs.wisc.edu/multifacet/papers/isca16_agile_paging.pdf

5.6 Increasing TLB Reach by Exploiting Clustering in Page Translations [HPCA 2014]

- https://www.cs.rutgers.edu/~abhib/binhpham-hpca14.pdf

5.7 CoLT: Coalesced Large-Reach TLBs [MICRO 2012]

- https://www.cs.rutgers.edu/~abhib/binhpham-micro12.pdf

5.8 Efficient Address Translation for Architectures with Multiple Page Sizes [ASPLOS 2017]

- https://guilhermecox.github.io/dw/gcox-asplos17.pdf

5.9 Rethinking TLB Designs in Virtualized Environments: A Very Large Part-of-Memory TLB [ISCA 2017]

- https://dl.acm.org/citation.cfm?id=3080210

5.10 Efficient Synonym Filtering and Scalable Delayed Translation for Hybrid Virtual Caching [ISCA 2016]

https://iamchanghyunpark.github.io/papers/hvc-isca2016.pdf

6. Non-Volatile / Hybrid Memories

6.1 An Analysis of Persistent Memory Use with WHISPER [ASPLOS 2017]

- http://research.cs.wisc.edu/multifacet/papers/asplos17_whisper.pdf

6.2 Heterogeneous memory architectures: A HW/SW approach for mixing die-stacked and off-package memories [2015 HPCA]

- http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7056027

6.3 Utility-Based Hybrid Memory Management [Cluster 2017]

- https://people.inf.ethz.ch/omutlu/pub/utility-based-hybrid-memory-management_cluster17.pdf

6.4 DudeTM: Building Durable Transactions with Decoupling for Persistent Memory [ASPLOS 2017]

- https://dl.acm.org/citation.cfm?id=3037714

6.5 TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory [ASPLOS 2017]

- https://web.stanford.edu/~mgao12/pubs/tetris.asplos17.pdf

6.6 memif: Towards Programming Heterogeneous Memory Asynchronously [ASPLOS 2016]

- https://engineering.purdue.edu/~xzl/xsel/papers/asplos16.pdf

7. DRAM Memories

7.1 Accelerating Pointer Chasing in 3D-Stacked Memory: Challenges, Mechanisms, Evaluation [ICCD 2016]

- https://people.inf.ethz.ch/omutlu/pub/in-memory-pointer-chasing-accelerator_iccd16.pdf
- https://github.com/CMU-SAFARI/IMPICA

7.2 Banshee: Bandwidth-Efficient DRAM Caching via Software/Hardware Cooperation [MICRO 2017]

- https://people.inf.ethz.ch/omutlu/pub/banshee-bandwidth-efficient-DRAM-cache_micro17.pdf
- https://github.com/yxymit/banshee

7.3 Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology" [MICRO 2017]

- https://people.inf.ethz.ch/omutlu/pub/ambit-bulk-bitwise-dram_micro17.pdf

7.4 DICE: Compressing DRAM Caches for Bandwidth and Capacity [ISCA 2017]

- https://dl.acm.org/citation.cfm?id=3080243

---

8. Side-Channel Attacks

8.1 Last-Level-Cache Side-Channel attacks

- https://en.wikipedia.org/wiki/Side-channel_attack
- https://cyber.wtf/2016/06/16/cache-side-channel-attacks-cpu-design-as-a-security-problem/
- http://palms.ee.princeton.edu/system/files/SP_vfinal.pdf
- https://www.blackhat.com/docs/asia-17/materials/asia-17-Irazoqui-Cache-Side-Channel-Attack-Exploitability-And-Countermeasures.pdf
- https://conference.hitb.org/hitbsecconf2016ams/materials/D2T1%20-%20Anders%20Fogh%20-%20Cache%20Side%20Channel%20Attacks.pdf

8.2 Meltdown

- https://meltdownattack.com/
- https://meltdownattack.com/meltdown.pdf

8.3 Spectre

- https://spectreattack.com/spectre.pdf

8.4 MeltdownPrime and SpectrePrime: Automatically-Synthesized Attacks Exploiting Invalidation-Based Coherence Protocols

- https://arxiv.org/pdf/1802.03802.pdf
- Appendix with code

8.5 DrK: Breaking Kernel Address Space Layout Randomization with Intel TSX

- http://people.oregonstate.edu/~jangye/assets/papers/2016/jang:drk-ccs.pdf
- https://www.youtube.com/watch?v=rtuXG28g0CU
- https://www.blackhat.com/docs/us-16/materials/us-16-Jang-Breaking-Kernel-Address-Space-Layout-Randomization-KASLR-With-Intel-TSX.pdf
- https://github.com/sslab-gatech/DrK

8.6 ASLR on the Line: Practical Cache Attacks on the MMU

- http://www.cs.vu.nl/~giuffrida/papers/anc-ndss-2017.pdf

8.7 Cache Side-Channel Attacks and the case of Rowhammer

- https://gruss.cc/files/cache_and_rowhammer_ruhrsec.pdf

8.8 Microarchitectural Side-Channel Attacks

- https://cs.adelaide.edu.au/~yval/CHES16/
- https://cs.adelaide.edu.au/~yval/CHES16/CHES16-tutorial.pptx
- https://cs.adelaide.edu.au/~yval/CHES16/CHES16-tutorial2.pptx

- https://cs.adelaide.edu.au/~yval/CHES16/Mastik.tgz

8.9 Jump Over ASLR: Attacking Branch Predictors to Bypass ASLR [MICRO 2016]

- http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7783743

8.10 Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems [CRYPTO 1996]

- http://web.cse.msstate.edu/~ramkumar/TimingAttacks.pdf

---

9. GPUs

9.1 Page Placement Strategies for GPUs within Heterogeneous Memory Systems [ASPLOS 2015]

- https://dl.acm.org/citation.cfm?id=2694381

9.2 Access Pattern-Aware Cache Management for Improving Data Utilization in GPU [ISCA 2017]

- https://dl.acm.org/citation.cfm?id=3080239

9.3 Mosaic: A GPU Memory Manager with Application-Transparent Support for Multiple Page Sizes [MICRO - 2017]

- https://people.inf.ethz.ch/omutlu/pub/mosaic-application-transparent-multiple-page-sizes-for-GPUs_micro17.pdf
- https://github.com/CMU-SAFARI/Mosaic

---

10. Binary Translation

10.1 Low Overhead Dynamic Binary Translation on ARM [PLDI 2016]

- https://www.research.manchester.ac.uk/portal/files/56078084/pldi_16.pdf

---

11. Graphs

11.1 Locality Exists in Graph Processing: Workload Characterization on an Ivy Bridge Server [IISCW'15]

- https://cloudfront.escholarship.org/dist/prd/content/qt8gd2p3qs/qt8gd2p3qs.pdf

11.2 Software Prefetching for Indirect Memory Accesses [CGO'17]

- https://www.cl.cam.ac.uk/~sa614/papers/Software-Prefetching-CGO2017.pdf
- https://github.com/SamAinsworth/reproduce-cgo2017-paper

11.3 Graph Prefetching Using Data Structure Knowledge [ICS'16]

- https://www.repository.cam.ac.uk/bitstream/handle/1810/260404/Ainsworth_et_al-2016-ICS16-AM.pdf?sequence=1

11.4 A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing [ISCA 2015]

- https://people.inf.ethz.ch/omutlu/pub/tesseract-pim-architecture-for-graph-processing_isca15.pdf

---

12. Vectorization

12.1 CVR: Efficient Vectorization of SpMV on X86 Processors [CGO'18]

- https://dl.acm.org/citation.cfm?id=3168818
- https://github.com/puckbee/CVR

---

13. Quantum Computing

13.1 Performing Quantum Computing Experiments in the Cloud

- https://quantumexperience.ng.bluemix.net/qx/community/question?questionId=984ab856f2044ab0bd118bab822d941f&channel=papers

13.2 A quantum teleportation experiment for undergraduate students

- https://quantumexperience.ng.bluemix.net/qx/community/question?questionId=88cfecc5724191d8b985bbac874df90b&channel=papers

13.3 Implementing a distance-based classifier with a quantum interference circuit

- https://quantumexperience.ng.bluemix.net/qx/community/question?questionId=c4ecbb33b09d8663f9bd6b4edf08de8d&channel=papers

13.4 Optimization and experimental realization of quantum permutation algorithm

- https://quantumexperience.ng.bluemix.net/qx/community/question?questionId=f7392b374bf0afcae9c3923fd4747ed9&channel=papers

13.5 Solving Linear Systems of Equations by Gaussian Elimination Method Using Grover's Search Algorithm: An IBM Quantum Experience

- https://quantumexperience.ng.bluemix.net/qx/community/question?questionId=7bc65ac0a8a53ab829b608a49bd7cc80&channel=papers

14. FPGAs

14.1 Evaluating and Optimizing OpenCL kernels for high-performance computing [SC'16]

- http://delivery.acm.org/10.1145/3020000/3014951/a35-zohouri.pdf?ip=147.102.3.198&id=3014951&acc=ACTIVE%20SERVICE&key=5641A0C343C36AC1%2E170A05475919F66C%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1520439606_15a36f27e6556996361bfeccc6458bbc
- https://github.com/fpga-opencl-benchmarks/rodinia_fpga

14.2 Efficient FPGA Implementation of OpenCL High-Performance Computing Applications via High-Level Synthesis

- http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7859319

15. Various

15.1 The Application Slowdown Model: Quantifying and Controlling the Impact of Inter-Application Interference at Shared Caches and Main Memory [MICRO 2015]

- https://people.inf.ethz.ch/omutlu/pub/application-slowdown-model_micro15.pdf
- https://github.com/CMU-SAFARI/ASMSim

15.2 "Continuous Runahead: Transparent Hardware Acceleration for Memory Intensive Workloads" [MICRO 2016]

- https://people.inf.ethz.ch/omutlu/pub/continuous-runahead-engine_micro16.pdf

15.3 Biscuit: A Framework for Near-Data Processing of Big Data Workloads (ISCA 2016)

- http://ieeexplore.ieee.org/document/7551390/

15.4 ActivePointers: a case for software address translation on GPUs [ISCA 2016]

- https://dl.acm.org/citation.cfm?id=3001200

15.5 Fusion: design tradeoffs in coherent cache hierarchies for accelerators [ISCA 2015]

- https://dl.acm.org/citation.cfm?id=2750421

15.6 libPRISM: An Intelligent Adaptation of Prefetch and SMT Levels [ICS17]

- https://dl.acm.org/citation.cfm?id=3079101
- https://upcommons.upc.edu/bitstream/handle/2117/107645/libPRISM+An+Intelligent+Adaptation+of+Prefetch+and+SMT+Levels.pdf
- https://github.com/kuleshov/prism/tree/master/libprism