1. “Trace Cache: a Low Latency Approach to High Bandwidth Instruction Fetching”
2. Jimenez and Lin, “Dynamic branch prediction with perceptrons”
3. “Hawkeye: Leveraging Belady’s Algorithm for Improved Cache Replacement”
4. Pierre-Yves Péneau, et al “Performance and Energy Assessment of Last-Level Cache Replacement Policies”. EDIS: Embedded Distributed Systems, 2018
5. Spectrum: Classifying, Replicating and Mitigating Spectre Attacks on a Speculating RISC-V Microarchitecture, 2018
6. Sparsh Mittal, “A Survey of Techniques for Dynamic Branch Prediction”
7. “A new case for the TAGE branch predictor”, 2006. https://comparch.net/2013/06/30/why-tage-is-the-best/
8. Improving Branch Prediction By Modeling Global History with Convolutional Neural Networks,Intel 2019
9. Seznec, Branch Prediction and the Performance of Interpreters -Don’t Trust Folklore, 2015
10. “High Performance Cache Replacement Using Re-reference Interval Prediction (RRIP)”, ISCA’10
11. NVIDIA Tesla: A unified graphics and computing architecture, 2008
12. “Improving GPU Performance via large warps and two-level WARP Scheduling”, MICRO 2011
13. Pierre Michaud, Best-Offset Hardware Prefetching, HPCA, 2016
14. JinChun KIm et al, “Kill the Program Counter: Reconstructing Program Behavior in the Processor Cache Hierarchy”, ASPLOS 2017
15. J. Doweck et al., "Inside 6th-Generation Intel Core: New Microarchitecture Code-Named Skylake," in IEEE Micro, vol. 37, no. 2, pp. 52-62, Mar.-Apr. 2017
16. E. Qin et al., "SIGMA: A Sparse and Irregular GEMM Accelerator with Flexible Interconnects for DNN Training," 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA), San Diego, CA, USA, 2020, pp. 58-70
17. Boosting Store Buffer Efficiency with Store-Prefetch Bursts, 2020 Micro
18. Lee et al, “UNPU: An Energy-Efficient Deep Neural Network Accelerator With Fully Variable Weight Bit Precision”, 2019, IEEE JOURNAL OF SOLID-STATE CIRCUITS

Approx computing, Vector processors and in-memory computing:

1. Mittal, “A Survey of Techniques for Approximate Computing” , ACM 2016
2. SNNAP: Approximate Computing on Programmable SoCs via Neural Acceleration, 2015
3. S. Venkataramani, V. K. Chippa, S. T. Chakradhar, K. Roy and A. Raghunathan, "Quality programmable vector processors for approximate computing," 2013 46th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), Davis, CA, 2013, pp. 1-12.

4. H. Jia, H. Valavi, Y. Tang, J. Zhang and N. Verma, "A Programmable Heterogeneous Microprocessor Based on Bit-Scalable In-Memory Computing," in IEEE Journal of Solid-State Circuits, vol. 55, no. 9, pp. 2609-2621, Sept. 2020, doi: 10.1109/JSSC.2020.2987714.
5. G. Singh et al., "A Review of Near-Memory Computing Architectures: Opportunities and Challenges," 2018 21st Euromicro Conference on Digital System Design (DSD), Prague, Czech Republic, 2018, pp. 608-617, doi: 10.1109/DSD.2018.00106.
6. S. F. Yitbarek, T. Yang, R. Das and T. Austin, "Exploring specialized near-memory processing for data intensive operations," 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, Germany, 2016, pp. 1449-1452.