[5] Mellanox Innova-2 Flex Open Programmable SmartNIC. https://www.mellanox.com/products/smartnics/innova-2-flex/.

[6] Netronome datapath programming tools. https://www.netronome.com/products/datapath-programming-tools/.

[7] The P4 language repositories. https://github.com/p4lang.

[8] SmartNIC Overview - Netronome. https://www.netronome.com/products/smartnic/overview/.

[9] IEEE P802.3bs 400 GbE Task Force. Adopted Timeline. http://www.ieee802.org/3/bs/, 2018.

[10] A. Abel and J. Reineke. Uops. info: Characterizing latency, throughput, and port usage of instructions on intel microarchitectures. In *Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019.

[11] J. R. Allen, B. M. Bass, C. Basso, R. H. Boivie, J. L. Calvignac, G. T. Davis, L. Frelechoux, M. Heddes, A. Herkersdorf, A. Kind, J. F. Logan, M. Peyravian, M. A. Rinaldi, R. K. Sabhikhi, M. S. Siegel, and M. Waldvogel. IBM PowerNP network processor: Hardware, software, and applications. *IBM Journal of Research and Development*, 47(2.3):177–193, 2003.

[12] M. S. B. Altaf and D. A. Wood. LogCA: A performance model for hardware accelerators. *IEEE Computer Architecture Letters*, 14(2):132–135, 2014.

[13] N. Ardalani, C. Lestourgeon, K. Sankaralingam, and X. Zhu. Cross-architecture performance prediction (XAPP) using CPU code to predict GPU performance. In *Proceedings of the 48th International Symposium on Microarchitecture*, 2015.

[14] C. Cadar, D. Dunbar, D. R. Engler, et al. Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs. In *8th USENIX Symposium on Operating Systems Design and Implementation*, 2008.

[15] P. D. O. Castro, C. Akel, E. Petit, M. Popov, and W. Jalby. CERE: LLVM-based codelet extractor and replayer for piecewise benchmarking and optimization. *ACM Transactions on Architecture and Code Optimization*, 12(1), 2015.

[16] M. K. Chen, X. F. Li, R. Lian, J. H. Lin, L. Liu, T. Liu, and R. Ju. Shangri-La: Achieving high performance from compiled network applications while enabling ease of programming. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2005.

[17] S. Choi, M. Shahbaz, B. Prabhakar, and M. Rosenblum. λ-nic: Interactive serverless compute on programmable smartnics. In *IEEE International Conference on Distributed Computing Systems*, 2020.

[18] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger. Dark silicon and the end of multicore scaling. In *Proceedings of the 38th Annual International Symposium on Computer Architecture*, 2011.

[19] D. Firestone, A. Putnam, S. Mundkur, D. Chiou, A. Dabagh, M. Andrewartha, H. Angepat, V. Bhanu, A. Caulfield, E. Chung, H. K. Chandrappa, S. Chaturmohta, M. Humphrey, J. Lavier, N. Lam, F. Liu, K. Ovtcharov, J. Padhye, G. Popuri, S. Raindel, T. Sapre, M. Shaw, G. Silva, M. Sivakumar, N. Srivastava, A. Verma, Q. Zuhair, D. Bansal, D. Burger, K. Vaid, D. A. Maltz, and A. Greenberg. Azure Accelerated Networking: SmartNICs in the Public Cloud. In *15th USENIX Symposium on Networked Systems Design and Implementation*, 2018.

[20] D. Firestone, A. Putnam, S. Mundkur, D. Chiou, A. Dabagh, M. Andrewartha, H. Angepat, V. Bhanu, A. Caulfield, E. Chung, et al. Azure accelerated networking: Smartnics in the public cloud. In *15th USENIX Symposium on Networked Systems Design and Implementation*, 2018.

[21] L. George and M. Blume. Taming the IXP network processor. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2003.

[22] J. L. Hennessy and D. A. Patterson. A new golden age for computer architecture. *Communications of the ACM*, 62(2):48–60, 2019.

[23] M. Hill and V. J. Reddi. Gables: A roofline model for mobile SoCs. In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*. IEEE, 2019.

[24] S. Hong and H. Kim. An analytical model for a GPU architecture with memory-level and thread-level parallelism awareness. In *Proceedings of the 36th Annual International Symposium on Computer Architecture*, 2009.

[25] S. Hong and H. Kim. An integrated GPU power and performance model. In *Proceedings of the 37th Annual International Symposium on Computer Architecture*, 2010.

[26] R. Iyer, L. Pedrosa, A. Zaostrovnykh, S. Pirelli, K. Argyraki, and G. Candea. Performance contracts for software network functions. In *16th USENIX Symposium on Networked Systems Design and Implementation*, 2019.

[27] G. P. Katsikas, T. Barbette, D. Kostic, R. Steinert, and G. Q. Maguire Jr. Metron:NFV service chains at the true speed of the underlying hardware. In *15th USENIX Symposium on Networked Systems Design and Implementation*, 2018.

[28] A. Kaufmann, S. Peter, N. K. Sharma, T. Anderson, and A. Krishnamurthy. High performance packet processing with FlexNIC. In *Proceedings of the 21st International Conference on Architectural Support for Programming Languages and Operating Systems*, 2016.

[29] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek. The Click modular router. *ACM Trans. Comput. Syst.*, 18(3):263âĂŞ297, 2000.

[30] M. Kotsifakou, P. Srivastava, M. D. Sinclair, R. Komuravelli, V. Adve, and S. Adve. HPVM: Heterogeneous parallel virtual machine. In *Proceedings of ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 2017.

[31] M. Kudlur and S. Mahlke. Orchestrating the Execution of Stream Programs on Multicore Platforms. In *Proceedings of the 29th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2008.

[32] C. Lattner and V. Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *Proceedings of CGO*, 2004.

[33] B. Li, Z. Ruan, W. Xiao, Y. Lu, Y. Xiong, A. Putnam, E. Chen, and L. Zhang. Kv-direct: High-performance in-memory key-value store with programmable nic. In *Proceedings of the 26th Symposium on Operating Systems Principles*, 2017.

[34] B. Li, K. Tan, L. Luo, Y. Peng, R. Luo, N. Xu, Y. Xiong, P. Cheng, and E. Chen. ClickNP: Highly flexible and high performance network processing with reconfigurable hardware. In *Proceedings of the 2016 ACM SIGCOMM Conference*, 2016.

[35] M. Liu, S. Peter, A. Krishnamurthy, and P. M. Phothilimthana. E3: Energy-efficient microservices on smartnic-accelerated servers. In *Proceedings of ther 2019 USENIX Annual Technical Conference*, 2019.

[36] C. Mendis, A. Renda, S. P. Amarasinghe, and M. Carbin. IIthemal: Accurate, portable and fast basic block throughput estimation using deep neural networks. In *Proceedings of International Conference on Machine Learning*, 2019.

[37] J. Meng, V. A. Morozov, K. Kumaran, V. Vishwanath, and T. D. Uram. Grophecy: GPU performance projection from cpu code skeletons. In *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2011.

[38] T. Nowatzki, M. Sartin-Tarm, L. De Carli, K. Sankaralingam, C. Estan, and B. Robatmili. A General Constraint-Centric Scheduling Framework for Spatial Architectures. In *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2013.

[39] C. Partridge, P. P. Carvey, E. Burgess, I. Castineyra, T. Clarke, L. Graham, M. Hathaway, P. Herman, A. King, S. Kohalmi, et al. A 50-Gb/s IP router. *IEEE/ACM Transactions on Networking*, 6(3):237–248, 1998.

[40] N. M. Patel. Half-latency rule for finding the knee of the latency curve. *ACM Performance Evaluation Review*, 43:28–29, 2014.

[41] L. Pedrosa, R. Iyer, A. Zaostrovnykh, J. Fietz, and K. Argyraki. Automated synthesis of adversarial workloads for network functions. In *Proceedings of the 2018 ACM SIGCOMM Conference*, 2018.

[42] L. T. X. Phan, M. Xu, and I. Lee. Cache-aware interfaces for compositional real-time systems. In *Proceedings of Workshop on Compositional Theory and Technology for Real-Time Embedded Systems*, 2015.

[43] P. M. Phothilimthana, M. Liu, A. Kaufmann, S. Peter, R. Bodik, and T. Anderson. Floem: A programming system for NIC-accelerated network applications. In *13th USENIX Symposium on Operating Systems Design and Implementation*, 2018.

[44] S. Pontarelli, R. Bifulco, M. Bonola, C. Cascone, M. Spaziani, V. Bruschi, D. Sanvito, G. Siracusano, A. Capone, M. Honda, et al. Flowblaze: Stateful packet processing in hardware. In *16th USENIX Symposium on Networked Systems Design and Implementation*, 2019.

[45] S. Sarda and M. Pandey. *LLVM Essentials*. O'Reilly, 2015.

[46] J. Sim, A. Dasgupta, H. Kim, and R. Vuduc. A performance analysis framework for identifying potential benefits in GPU applications. In *Proceedings of the 17th ACM SIGPLAN symposium on Principles and Practice of Parallel Programming*, 2012.

[47] A. Snavely, L. Carrington, N. Wolter, and J. Labarta. A framework for performance modeling and prediction. In *Proceedings of the ACM/IEEE Conference on Supercomputing*, 2002.

[48] A. Sriraman and A. Dhanotia. Accelerometer: Understanding acceleration opportunities for data center overheads at hyperscale. In *Proceedings of the 25th International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020.

[49] Z. Wang, B. He, W. Zhang, and S. Jiang. A performance analysis framework for optimizing opencl applications on fpgas. In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*. IEEE, 2016.

[50] M. Xu, L. T. X. Phan, I. Lee, O. Sokolsky, S. Xi, C. Lu, and C. Gill. Cache-aware compositional analysis of real-time multicore virtualization platforms. In *Proceedings of IEEE Real-Time Systems Symposium*, 2013.

[51] L. T. Yang, X. Ma, and F. Mueller. Cross-platform performance prediction of parallel applications using partial execution. In *Proceedings of the ACM/IEEE Conference on Supercomputing*. IEEE, 2005.

[52] J. Zhai, W. Chen, and W. Zheng. Phantom: Predicting performance of parallel applications on large-scale parallel machines using a single node. *ACM Sigplan Notices*, 45(5):305–314, 2010.

[53] R. Zhang, Z. Budimlić, and K. Kennedy. Performance modeling and prediction for scientific Java applications. In *Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software*, 2006.

[54] W. Zhang, M. Hao, and M. Snir. Predicting HPC parallel program performance based on LLVM compiler. *Cluster Computing*, 20, 2017.