EE 599: Accelerated Computing using FPGAs

Units: 2
Spring 2020
Thu 330-450 (Lecture)
Fri 4-520 (Lab, Discussion)

Instructor: Viktor Prasanna
Office: EEB 200C
Contact Info: prasanna@usc.edu
(213) 740 4483
sites.usc.edu/prasanna

Course Description

Recently, Field Programmable Gate Arrays have become a key computing platform to accelerate applications at data center, cloud and at the "edge". This course will review the technology and software tools from application acceleration perspective and discuss (application-specific) architectural, software and algorithmic innovations to realize the potential of this technology to optimize latency, throughput and energy efficiency.

Prerequisite(s): EE 457 and CS 570
Recommended Preparation: EE 451

Required Readings and Supplementary Materials

Course will be based on recent research publications and survey articles and vendor data sheets and tools. Details will be provided in the lectures as well as in the discussion sessions. A sample of relevant literature is appended to this.

Description and Assessment of Assignments

The course will be project oriented. Project proposal, presentation and final report are required.

Grading Breakdown

Including the above detailed assignments, how will students be graded overall? Participation should be no more than 15%, unless justified for a higher amount. All must total 100%.

Assignment	Points	% of Grade
Participation		10
Lab Assignments		20
Project Proposal		20
Project Presentation		20
Project Final Report		30
TOTAL		100

Project: The focus of the course is in designing accelerators using FPGAs. The project will be focused on specific application areas of interest to the students to identify a problem

that needs acceleration, design an application specific architecture, develop scalable parallel algorithm and map it onto a target FPGA device. The project will consist of literature survey, problem definition, solution idea, hardware design and use of software tools to map the design to a FPGA. It will consist of proposal preparation, discussions with the instructor and the TA, present details of the design and implement it and report the resulting acceleration.

Sample project:
Parallelizing LSTM models on FPGAs with coherent memory.
Identifying opportunities for parallelism, survey of state of the art techniques for kernels and primitives, performance modeling and projected performance. Implementation in VHDL or Verilog, synthesis, place and route results. Summary of latency and throughput performance and energy dissipation.

Course Schedule: A Weekly Breakdown

Note: L and D refers to lecture and discussion sessions. Discussion sessions will be led by a TA over the first 10 weeks.

	Topics/Daily Activities	Readings and Homework	Deliverable/ Due Dates
Week 1	Introduction (L) Computing platforms and technology evolution FPGA design flow (D)		
Week 2	FPGA basics, architectural characteristics (L) Example design flow, account set up (D)		
Week 3	FPGA abstractions and Computational models (L) Example design flow, practice designs (D)		
Week 4	Accelerating Dense Algebra (L) HW #1 solution strategies (D)	HW #1	
Week 5	Accelerating FFT (L) FFT design optimization (D)		
Week 6	Accelerating Networking (SDN) (L) HW #2 discussion (D)	HW #2	HW # 1 due
Week 7	Accelerating Networking (NFV) (L) IP look up design (D)		
Week 8	Accelerating ML Kernels (L) Reg Ex Matching design (D)		HW # 2 due
Week 9	Accelerating ML Kernels (L) Tools for FPGA resource management (D)		Project Proposal due
Week 10	FPGAs in the Cloud (L) Project discussion, guidelines (D)		
Week 11	Project Presentation (L, D)		
Week 12	Project Presentation (L, D)		
Week 13	Project Presentation (L, D)		
Week 14	Project Presentation (L, D)		
Week 15	Project Presentation (L, D)		Final report due last day of classes
FINAL	No final		Date: For the date and time of the final for this class, consult the USC *Schedule of Classes* at classes.usc.edu/.

Sample reading materials

1. Weerasinghe, Jagath, et al., Enabling FPGAs in hyperscale data centers, 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom). IEEE, 2015.
2. Pena, Maria Dolores Valdes, Juan J. Rodriguez-Andina, and Milos Manic, The internet of things: The role of reconfigurable platforms, IEEE Industrial Electronics Magazine 11.3 (2017): 6-19.
3. Putnam, Andrew, FPGAs at HyperScale--The Past, Present, and Future of the Reconfigurable Cloud, FPGAs Keynote, ReConfig, 2018.
4. Stamelos, Ioannis, et al, A Novel Framework for the Seamless Integration of FPGA Accelerators with Big Data Analytics Frameworks in Heterogeneous Data Centers, 2018 International Conference on High Performance Computing & Simulation (HPCS). IEEE, 2018.
5. Mbongue, Joel Mandebi, et al, FPGA Virtualization in Cloud-Based Infrastructures Over Virtio, IEEE 36th International Conference on Computer Design (ICCD), IEEE, 2018.
6. Chen, Y., He, J., Zhang, X., Hao, C. and Chen, D., Cloud-DNN: An Open Framework for Mapping DNN Models to Cloud FPGAs. Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (pp. 73-82), ACM, 2019.
7. X. Wei, H. Yu, P. Zhang, Y. Chen, Y. Wang, H. Hu, Y. Liang and J. Cong, Automated Systolic Array Architecture Synthesis for High Throughput CNN Inference on FPGAs, 54th ACM/EDAC/IEEE Design Automation, 2017.
8. S. Wang, Z. Li, C. Ding, B. Yuan, Q. Qiu, Y. Wang and Y. Liang, C-LSTM: Enabling Efficient LSTM using Structured Compression Techniques on FPGAs, Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, CA, 2018.
9. S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. Horowitz and W. Dally, EIE: Efficient Inference Engine on Compressed Deep Neural Network, ACM/IEEE 43th Annual International Symposium on Computer Architecture (ISCA), 2016.
10. Y. Umuroglu, N. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre and K. Vissers, FINN: A Framework for Fast, Scalable Binarized Neural Network Inference, Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, CA, 2017.
11. Yang, Y., Huang, Q., Wu, B., Zhang, T., Ma, L., Gambardella, G., Blott, M., Lavagno, L., Vissers, K., Wawrzynek, J. and Keutzer, K., Synetgy: Algorithm-hardware co-design for convnet accelerators on embedded fpgas. Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (pp. 23-32), ACM, 2019.
12. Zhang, C., Sun, G., Fang, Z., Zhou, P., Pan, P. and Cong, J.,. Caffeine: Towards uniformed representation and acceleration for deep convolutional neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.
13. Asiatici, M. and Ienne, P., Stop Crying Over Your Cache Miss Rate: Handling Efficiently Thousands of Outstanding Misses in FPGAs. Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (pp. 310-319), ACM, 2019.
14. Moreau, T., Chen, T., Jiang, Z., Ceze, L., Guestrin, C. and Krishnamurthy, A., 2018. VTA: An Open Hardware-Software Stack for Deep Learning, arXiv preprint arXiv:1807.04188.
15. Xilinx, Xilinx AI Engines and Their Applications, Xilinx White Paper WP506, 2018.

16. Xilinx, Accelerating DNNs with Xilinx Alveo Accelerator Cards, Xilinx White Paper WP504, 2018.
17. Vissers, Kees, Versal: The Xilinx Adaptive Compute Acceleration Platform (ACAP), Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM, 2019.
18. Intel Altera, AgilexTM FPGAs Deliver a Game-Changing Combination of Flexibility and Agility for the Data-Centric World, Intel White Paper, 2019.
19. Zeng, Hanqing; Zhang, Chi; Prasanna, Viktor K., *Fast Generation of High Throughput Customized Deep Learning Accelerators on FPGAs*, International Conference on ReConFigurable Computing and FPGAs (ReConFig), pp. 1–8, 2017
20. Zhou, Shijie; Kannan, Rajgopal; Zeng, Hanqing; Prasanna, Viktor K., *An FPGA Framework for Edge-Centric Graph Processing*, Proceedings of the 15th ACM International Conference on Computing Frontiers, pp 69–77, 2018
21. Tong, Da; Prasanna, Viktor K., *Sketch Acceleration on FPGA and its Applications in Network Anomaly Detection*, IEEE Transactions on Parallel & Distributed Systems, Vol 29, Issue 4, pp 929–942, 2018
22. Zhou, Shijie; Kannan, Rajgopal; Yu, Min; Prasanna, Viktor K., *FASTCF: FPGA-based Accelerator for Stochastic-Gradient-Descent-based Collaborative Filtering*, ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp 259–268, 2018
23. Zeng, Hanqing; Chen, Ren; Zhang, Chi; Prasanna, Viktor K., *A Framework for Generating High Throughput CNN Implementations on FPGAs*, ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 117–126, 2018
24. Qu, Yun R.; Prasanna, Viktor K., *Fast Online Set Intersection for Network Processing on FPGA*, IEEE Transactions on Parallel and Distributed Systems, 2016
25. Qu, Yun R.; Prasanna, Viktor K., *High-performance and Dynamically Updatable Packet Classification Engine on FPGA*, IEEE Transactions on Parallel and Distributed Systems, Vol. 27, No. 1, pp. 197–209, 2016
26. Qu, Yun R.; Zhang, Hao H.; Zhou, Shijie; Prasanna, Vikor K., Optimizing Many-field Packet Classification on FPGA, multi-core General Purpose Processor, and GPU, ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), 2015
27. Jiang, Weirong; Prasanna, Viktor K., *Scalable Packet Classification on FPGA*, IEEE Transactions on Very Large Scale Integration Systems (TVLSI), 2012
28. Yang, Yi-Hua E.; Prasanna, Viktor K., *High-Performance and Compact Architecture for Regular Expression Matching on FPGA*, IEEE Transactions on Computers, Vol. 61, No. 7, pp. 1013–1025, 2012