EE 599: Accelerated Computing using Field Programmable Gate Arrays

Units: 2

Term—Day—Time:

Spring 2021
Tue 5 – 620 (Lecture)
Thu 5 – 620 (Lab/Discussion)

Location: Zoom (online)
Lecture and discussion meeting link:
https://urldefense.com/v3/__https://usc.zoom.us/j/91027990240?pwd=L1NzalNYRWF4Rkt3Znl4RDJsMDhnUT09__;!!LIr3w8kk_Xxm!9NTtm4UNMfzsy1L0nnMmV5sfNhBbbMfk2XRAE8JYERLMwVPn9WPVJHXTW__OLeQIrA$
or type
Meeting ID: 910 2799 0240
Passcode: _A3hm1BaLQ

Instructor: Viktor Prasanna
Office: https://usc.zoom.us/s/97033355040
Office Hours: TBD
Contact Info: prasanna@usc.edu
(213)740-4483
sites.usc.edu/prasanna

Teaching Assistant: Yuan Meng
Office: https://usc.zoom.us/j/8629150353
Office Hours: TBD
Contact Info: ymeng643@usc.edu
IT Help: Group to contact for technological services, if applicable.

Course Description

Recently, Field Programmable Gate Arrays (FPGAs) 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 performance.

Learning Objectives

Identify what specific, measurable skills a student will obtain and be able to demonstrate by the end of the course. Learning objectives should be both taught and assessed in your course. They are aligned with your assignments, assessments and learning materials.

- Understand basics of FPGA technology
- Design parallel strategies to achieve high performance
- Understand modeling and system design
- Use high level tools for application development
- Understand tradeoffs in designing accelerators using FPGAs

Prerequisite(s): EE 457 or EE 451 or consent of the instructor. Students are encouraged to discuss their background with the instructor.

Recommended Preparation: Expected to have knowledge of basic algorithms, for example, at the level of CS 570.

Course Notes

Blackboard and Piazza will be used. Zoom will be used if needed.

Technological Proficiency and Hardware/Software Required

Students should have basic understanding of high-level programming such as C/C++.

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

Assignment	**% of Grade**
Lab Assignments (HW)	25
Project Proposal	25
Project Presentation	20
Project Final Report	30
TOTAL	100

Project Presentation: Each project will have 25 mins of presentation time followed by 10 mins for Q and A. The presentation should be Power Point slides distributed prior to the class. It should be approximately 10 slides addressing the following: Context, Problem definition, Project hypothesis, Approach, Initial results, Comparison with the state of the art, Details of final report, and Conclusion.

Project Final Report: This is a written document approximately 10 pages in IEEE or ACM conference format (single space, point size 10, 1 inch margins, 2 column) addressing the topics covered in the project presentation and in addition will include the final results, experiments conducted and comparisons.

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.

Assignment Submission Policy

Assignments to be submitted via Blackboard.

Grading Timeline

Graded homeworks will be returned within one week and solutions will be posted two days after the submission deadline.

Additional Policies

5% penalty for each day after the due date. Max of 2 days.

Course Schedule (Tentative): A Weekly Breakdown

Note: L and D refer to lecture and discussion sessions. Discussion sessions will be led by a TA over the first 8 weeks to cover the software tools and access to cloud resources as well as the FPGA resources of professor Prasanna's Lab (fpga.usc.edu). The instructor will be incharge of both the discussion session and the lab session during weeks 9-15.

	Topics/Daily Activities	**Readings and Homework**	**Deliverables/Due Dates**
Week 1	Course outline, focus, policies and expectations (L) Tools and software access, setup and background assessment (D)		
Week 2	Course outline, project requirements and examples (L) Introduction to FPGA design flow (D)		
Week 3	FPGA Introduction and technology evolution (L) Introductory session on Xilinx Vivado Tools and Intel Quartus Tools (D)	HW #1	HW # 1 due
Week 4	FPGA Devices and computational features (L) Introduction to Verilog coding (D)	HW #2	HW # 2 due
Week 5	Fine grained models of computation for FPGAs (L) Introductory session on Dev-cloud and OpenCL (D)		
Week 6	Accelerator Design for Networking (L) Introductory session on Vivado HLS and Xilinx Vitis tools (D)	HW #3	HW # 3 due
Week 7	Accelerator Design for Deep Learning (L) More Examples of Verilog designs (D)	HW #4	HW # 4 due, Project Proposal due
Week 8	Accelerator Design for Graph Analytics (L) More Examples of HLS designs (D)		
Week 9	FPGAs in the Cloud (1) (L) Interfacing HLS designs in the cloud (D)		
Week 10	FPGAs in the Cloud (2)—Accelerator interface Project Discussion (D)		
Week 11	Systolic Designs and Optimizations (L) Advanced Programming Models : possible guest lecture (Intel) (D)		
Week 12	Communication optimization and area time tradeoffs (L) Advanced Programming Models: possible guest lecture (Xilinx) (D)		
Week 13	Project Presentation (L) Project Presentation (D)		
Week 14	Project Presentation (L) Project Presentation (D)		
Week 15	Project Presentation (L) Project Presentation (D)		Final report due last day of classes
FINAL	No final		Refer to the final exam schedule in 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

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) - (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Campus Support and Intervention - (213) 821-4710
campussupport.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101

diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.