**Crossroads Student Workshop, Summer 2021**
**11:30AM ~ 6:30PM US Eastern**
**Friday August 27th**

**After connecting, please set your Zoom ID to your name and affiliation. Please don't be shy about turning on video when speaking. Please**

**11:30 AM** **Introduction**

**11:40 AM** **RV1: Zhipeng Zhao (CMU)**
**Title:** Pigasus: Efficient Handling of Input-Dependent Streaming on FPGAs

**Abstract:** FPGAs have well-demonstrated success in many networking applications but failed in accelerating Intrusion Detection and Prevention Systems(IDS/IPS). The root cause is the mismatch of the traditional static, fixed-performance FPGA design and input-dependent behaviors of IDS/IPS. As a result, the design is provisioned to handle worst-case, losing the opportunity to utilize the resource allocated for worst-case to improve the common-case performance.

In this talk, I will present an FPGA based IDS/IPS called Pigasus which is tailored to the common-case, thus using minimal resources to extract maximum performance. Pigasus can achieve 100Gbps using 1 FPGA and on average 5 CPU cores, 100x faster than CPU-only baseline and 50x faster than existing FPGA designs. A natural objection to this design is that it will suffer from shifting workloads. In the second part, I will show how to use a disaggregated architecture and spillover mechanism to scale subcomponents of the system on demand to address changes in the traffic profile at both compile time and runtime.

**Bio:** Zhipeng Zhao is a Ph.D. candidate in Electrical and Computer Engineering at Carnegie Mellon University, advised by Prof. James C. Hoe. His research interests broadly lie at the intersection of FPGA and networking. Prior to CMU, he received a BS and an MS in Electrical Engineering, both from Beihang University, China.

**12:05 PM** **Invited Talk: Eric Chung (Microsoft)**

**12:45 PM** **RV1: Hugo Sadok (CMU)**
**Title:** We Need Kernel Interposition over the Network Dataplane

**Abstract:** Kernel-bypass networking, which allows applications to circumvent the kernel and interface directly with NIC hardware, is one of the main tools for improving application network performance. However, allowing applications to circumvent the kernel makes it impossible to use tools (e.g., tcpdump) or impose policies (e.g., QoS and filters) that need to interpose on traffic sent by different applications running on a host. This makes maintainability and manageability a challenge for kernel-bypass applications. In response, we propose Kernel On-Path Interposition (KOPI), in which traditional kernel dataplane functionality is retained but implemented in a fully programmable SmartNIC. We hypothesize that KOPI can support the same tools and policies as the kernel stack while retaining the performance benefits of kernel bypass.

**Bio:** Hugo Sadok is a third-year PhD student in Computer Science at CMU advised by Prof. Justine Sherry and part of the SNAP Lab. His research interests are broadly in computer networks and computer systems. Prior to CMU, he received a BS in Electronic and Computer Engineering and an MS in Electrical Engineering, both from UFRJ.

**1:10 PM** **5 min breakout + 10 min Q&A (Moderator Justine Sherry)**

*Everyone in attendance will be randomly assigned into small-group breakout rooms. Please spend 5 minutes together to discuss questions or comments within the small group. When ready, please post in Chat any questions you like to raise with the speakers. A moderator will help the speakers to navigate as many questions as possible; left-over questions will be answered offline in writing.*

**1:25 PM** **RV1: Mohammad Ibrahim (Toronto)**

**Title:** Sparsity-Aware, Hyper Pipelined Architecture for CNN Inference Acceleration on FPGAs (HPIPE)

**Abstract:** HPIPE is a state-of-the-art CNN accelerator for FPGAs. Through building deeply-pipelined, customized hardware for every layer in the CNN and modeling the physical device characteristics, HPIPE can achieve very high compute density while maintaining high operating frequency. Although HPIPE achieves high performance, it requires all CNN parameters to be on chip (no off-chip memory), limiting HPIPE to CNNs that can completely fit on one chip. Moreover, the compute resources on a single chip are limited and cannot fulfill the compute requirements of large models such as Resnet-200 (15 GFLOPS).

In this talk, I will first present HPIPE and how it produces highly efficient accelerators for CNNs. I will then describe the ongoing work to overcome the constraints of HPIPE and allow more models to be accelerated. The first approach is to scale HPIPE to multiple-chip systems using an automated CAD flow, increasing the compute (DSPs) and the storage (on-chip memory) resources hence enabling more CNNs to be accelerated and increasing the performance compared to single-chip accelerators. The second approach is to reduce the model's memory requirements through quantization, and use more compute through enabling HPIPE to utilize the AI-dedicated Stratix 10 NX (143 INT-8 TOPs).

**Bio:** Mohamed Ibrahim is currently a M.A.Sc. student at department of Electrical and Computer Engineering (ECE) at the University of Toronto under the supervision of Prof. Vaughn Betz. He holds a BSc degree in Electronics and Communications Engineering from the American University in Cairo. His research focuses on Machine learning acceleration on FPGAs, more specifically systems of multiple FPGAs

**1:50 PM** **RV2: Alex Hsu (Austin)**

**Title:** Towards Automatically Generated Specialized Overlay Processors

**Abstract:** Time-to-solution is an often-overlooked but important metric that includes not only run-time but also architecture, implementation, compilation, debug, and deployment. Though FPGAs are well known to reduce run-time, they are traditionally programmed with RTL, which is not highly productive and compile and simulate slowly. FPGA soft

processor overlays potentially provide a balance between agility and performance that can reduce overall time-to-solution.

In this talk we will introduce the Primate project that is intended to prove out overlays and provide the necessary infrastructure to make overlays more practical. A Primate overlay is composed of an array of heterogeneous function units that are driven by wide instructions. We will present the current Primate overlay microarchitecture, applications that we have studied for implementation on Primate overlays and discuss how Primate can enable both rapid development and competitive performance for those applications.

**Bio:** Alex Hsu is a PhD student in the ECE department at UT Austin under the supervision of Derek Chiou. He completed his undergrad at UT Austin as well. His research interests include automatic generation and modeling of architectural designs and taking advantage of the flexibility of reconfigurable fabrics for acceleration.

**2:15 PM** **5 min breakout, 10 min Q&A (Moderator Derek Chiou)**
*Everyone in attendance will be randomly assigned into small-group breakout rooms. Please spend 5 minutes together to discuss questions or comments within the small group. When ready, please post in Chat any questions you like to raise with the speakers. A moderator will help the speakers to navigate as many questions as possible; left-over questions will be answered offline in writing.*

**2:30 PM** **Intermission**

**2:45 PM** **RV3: Andrew Boutros (Toronto)**
**Title:** System-level Simulator for 3D-FPGA Architecture Modeling & Exploration
Andrew Boutros, University of Toronto

**Abstract:** The Crossroads FPGA exploits new 3D chip integration technologies to stack an FPGA fabric on top of a base die. This base die can potentially contain heterogeneous accelerator blocks, embedded memories, external memory controllers, and more. It will also contain a high-performance packet-switched network-on-chip (NoC) that handles all communication between different system components. The integration of all these components results in a significantly larger design space and more complex beyond-FPGA platforms, opening up many interesting research questions: How do we best architect these devices? How should a NoC be embedded in such a platform, and what abstraction should be presented to designers? What accelerator blocks and how many of them to add? What is the effect of these design decisions on key applications?

In this talk, I will present our ongoing work on an architecture modeling and exploration framework that can be used to answer these questions. Our framework performs SystemC simulation of the complete system to produce NoC traffic statistics and performance results of the given application design on the described Crossroads architecture. This allows rapid what-if analysis of architecture ideas that incorporate different NoCs, embedded memory and accelerators in a Crossroads 3D-FPGA.

**Bio:** Andrew Boutros received his B.Sc. degree in electronics engineering from the German University in Cairo in 2016, and his M.A.Sc. degree in electrical and computer engineering from the University of Toronto in 2018. He was a research scientist at Intel's

Accelerator Architecture Lab before he returned to the University of Toronto where he is currently pursuing his Ph.D. degree under the supervision of Prof. Vaughn Betz. His research interests include FPGA architecture and CAD, deep learning acceleration, and domain-specific architectures. He is also a postgraduate affiliate of the Vector Institute for Artificial Intelligence and the Center for Spatial Computational Learning.

**3:10 PM** **RV4: Kate Thurmer, (Toronto) & Rachel Selina Rajarathnam (Austin)**
**Title:** Towards scalable FPGA placement - Integration of VTR and elfPlace++

**Abstract:** As FPGAs evolve, flexible and scalable CAD tools are essential to the exploration of innovative 2.5- and 3D-FPGA architectures. Existing CAD tools require substantial enhancements in order to target 3D architectures like the proposed CrossRoads FPGA. To ensure flexible and scalable CAD tools, we propose to integrate the adaptable 'VTR' CAD flow with the extensible 'elfPlace++' placer. As the leading open-source FPGA CAD flow, VTR is capable of modeling and targeting any arbitrary FPGA architecture. Its flexible and extensible modeling language facilitates the extension of an FPGA architecture to multiple dies and heterogeneous interconnects among them. However, VTR's connectivity-based, pre-placement clustering can yield suboptimal results, and its simulated annealing based placer is prohibitively slow for very large designs. As 3D-FPGAs enable extremely large (10 million+ LE) devices, tool scalability is a key concern. elfPlace is a state of the art electrostatics-based placement engine that operates on a flat design to obtain a rough legal global placement. The instances are then clustered and further optimized to meet legality requirements as well as minimize wirelength. We refer to the accelerated version of elfPlace as 'elfPlace++' that is built on a deep-learning toolkit, similar to the open-source highly scalable DREAMPlace framework for ASICs.

As the first step towards this integration, we evaluate the performance of both VTR's placer and elfPlace on the ISPD'2016 contest benchmarks, which are based on the Xilinx Ultrascale Architecture. Next, we augment VTR to be able to send netlist and architecture information to elfPlace++ and to interpret elfPlace++ placement solutions. On the elfPlace++ side, we are currently working on making the tool generic to support different architectures, and later include support for timing optimizations.

**Bio: Kate Thurmer:** Kate is a second year PhD student at the University of Toronto advised by Prof. Vaughn Betz. Prior to joining U of T, she earned her MS degree in Computer Engineering at Boston University and developed CAD tools for application-specific configurable hardware at MIT Lincoln Laboratory. **Rachel Selina Rajarathnam:** Rachel Selina Rajarathnam is a PhD student in the Integrated Circuits and Systems (ICS) track at the University of Texas at Austin, supervised by Prof. David Z. Pan. Her research interests include Physical Design Automation and Hardware Security. She did a summer internship at Intel (2021) focusing on FPGA placement in Quartus compiler. Previously, She has worked on several projects related to ASIC physical design automation at Nvidia. Rachel Selina holds a bachelor's degree in Electronics and Communication Engineering, and a master's degree in Applied Electronics from Anna University, Chennai, India.

**3:30 PM** **RV4: Srivatsan Srinivasan (Toronto)**

**Title:** A method which allow designs with custom hard blocks to pass through Quartus synthesis and then be placed and route with VPR

**Abstract:** VTR is a configurable open-source CAD software that allows designers to target designs onto arbitrary FPGA architectures. VTR allows for a variety of synthesis flows; one such flow is capable of leveraging Quartus synthesis with the physical implementation done by the VPR place and route system. This flow benefits from the broad HDL language and IP support provided through Quartus, but has had the historical limitation of only allowing device primitives found in current Intel FPGAs to be used within designs. This has meant that this flow could not be used to evaluate FPGA architectures with new hard blocks, such as NoC (Network-on-Chip) routers.

In this talk, I will detail a flow and a new netlist conversion tool that can overcome the previously detailed limitations, allowing a design to pass through Quartus synthesis and then into VPR, even though the design can contain a new primitive block which does not exist in current Intel FPGAs. Finally, I will show how this new feature can be used to pass a design with a custom hard NoC router through Quartus synthesis and then into VPR for placement and routing. Although the sample use demonstration will focus on a NoC router, this feature is not router specific, we can use this for any arbitrary primitive blocks.

**Bio:** Srivatsan Srinivasan is a current M.A.Sc candidate in Electrical and Computer Engineering at the University of Toronto, under the supervision of Prof. Vaughn Betz. Before his M.A.Sc, Srivatsan completed his B.E.Sc in Electrical Engineering and B.Sc in Computer Science at Western University. His research interests include FPGA CAD, NoCs and CAD tools targeting FPGA architectures with integrated NoCs.

**3:38 PM** **RV4: Kimia Talaei (Toronto)**

**Title:** Compilation of Titan benchmarks for Stratix 10 and Agilex devices and using Quartus to generate node-level netlists

**Abstract:** Open-source CAD tools such as VTR have limited support of HDL languages and IP cores which results in their inability to make use of large modern benchmarks. Titan is a hybrid CAD flow that enables VTR to handle complex benchmarks through utilization of Quartus to perform HDL synthesis and generate a node-level netlist (.vqm file). By taking an architecture file as input, Titan flow will then convert the VQM file to BLIF format which is supported by most open-source CAD tools. Titan flow enables comparison of commercial and academic CAD tools by providing a comprehensive benchmark suite consisting of 74 heterogeneous designs and an architecture capture of Stratix IV allowing VPR to map the design to a Stratix-IV like architecture. The more recent families of Intel FPGAs have evolved and include new features to better accommodate emerging FPGA application designs. Our goal is to come up with an architecture capture of two newer families of FPGAs (Stratix 10 and Agilex) and enable compilation of Titan benchmarks for the two new architectures. In this talk, I will give a brief overview of titan flow and our roadmap for architecture capture of the two new FPGA families. I will also demonstrate the results from the first step, which is compilation of Titan benchmarks on Quartus for the two new families and producing the node-level netlists.

**Bio:** Kimia Talaei is a second-year M.A.Sc. student in Electrical and Computer Engineering at the University of Toronto and under the supervision of Prof. Vaughn Betz. Prior to University of Toronto, she received her B.Sc. degree in Computer Engineering from Sharif University of Technology, Iran. Her research interests include open-source FPGA CAD tools and potential applications of machine learning in this area.

**3:45 PM** **5 min breakout, 10 min Q&A (Moderator Vaughn Betz)**
*Everyone in attendance will be randomly assigned into small-group breakout rooms. Please spend 5 minutes together to discuss questions or comments within the small group. When ready, please post in Chat any questions you like to raise with the speakers. A moderator will help the speakers to navigate as many questions as possible; left-over questions will be answered offline in writing.*

**4:00 PM** **Lightning talks All students (not presenting otherwise)**

*Each student gets to summarize/overview what they are working on in 2 minutes with 1 slide only.*

**4:30 PM** **RV 5: Shashank Obla (CMU)**
**Title:** Extracting Design Tradeoff Slack at Runtime with Dynamically Composable Designs

**Abstract:** Field programmability is the hallmark, differentiating FPGAs from ASICs and Partial Reconfiguration (PR) takes it a step further adding support for runtime reprogrammability. But PR is still under-utilized, with the significant use-case being the role-and-shell approach only to provide security and always-on connectivity in servers. State-of-the-art PR can still do much better ,as shown in prior work [Nguyen, FPL2019, Nguyen, FPL 2020] but this still falls short of the full potential of future FPGAs at the center of data flows to be truly programmable. The Crossroads FPGA needs to efficiently cater to the immense demand for applications to be near and compute on data, exceeding static capacity.

PR is essential, and the only mode of operation in the Crossroads FPGA. This talk will motivate and propose a new design framework that allows FPGA designs to be more dynamic and responsive to their environment, including inputs and other applications. The talk will present a proof-of-concept, using BFS and Pigasus as examples, on how such a framework could extract data-dependent slack to improve efficiency. Lastly, the talk will discuss the latest work on generalizing this approach to effectively utilize slack using application analysis, starting with Pigasus 2.0, and the need for runtime analysis.

**Bio:** Shashank Obla is a third year PhD student at the Electrical and Computer Engineering (ECE) department at the Carnegie Mellon University, advised by Prof. James C. Hoe. His research interests broadly include reconfigurable computing and computer architecture with focus on emerging applications of Partial Reconfiguration. He received his BTech and MTech degrees in Electrical Engineering with focus on Microelectronics and VLSI from the Indian Institute of Technology, Bombay in 2019.

**4:55 PM** **5 min breakout, 10 min Q&A (Moderator James Hoe)**

*Everyone in attendance will be randomly assigned into small-group breakout rooms. Please spend 5 minutes together to discuss questions or comments within the small group. When ready, please post in Chat any questions you like to raise with the speakers. A moderator will help the speakers to navigate as many questions as possible; left-over questions will be answered offline in writing.*

**5:10 PM** **External Open-Mic**

*Anyone not from the center can speak their mind. PIs will not answer questions asked during this time; the PIs will take note and follow up.*

**5:30 PM** **Adjourn**

Siddharth Sahay