

# **Coyote** Abstractions for Modern Heterogeneous Hardware

Dario Korolija, Timothy Roscoe, Gustavo Alonso

Systems Group, Dept. of Computer Science, ETH Zurich



# Introduction...

- Scpecialized hardware becoming a reality
  - Amazon, Microsoft, Google, Alibaba, Intel, AMD ...







Google TPU

2 x ThunderX 40Gb/s Ethernet

Baseboard

Management

Controller

4 x FPGA QSFP28

(16 x 25Gb/s)

FPGA PCle 16x

ThunderX PCIe 8x

FPGA FMC connector



Xilinx XCVU9P FPGA

2 x FPGA DDR4 DIMMs

✤ One built within System group at ETH:





2

# Complex to program (interact with)

- ✤ I/O interaction ...
- Low level APIs
- No standard execution environment
- Lack of portability (HLS ...)
- Lack of traditional abstractions



*FPGA layout PCle, memory, storage, network...* 



# Microkernel for FPGAs

Hybrid computing system

- Plenty of research, focused on individual functionalities only
- Coyote provides a complete minimal core set of essential features on which further services can be used

Interdependence





<text><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></text>	Justicing of	Abel Mulugeta Eneyew <sup>2</sup> Zhengwei Ql <sup>1</sup> University of Michigan <sup>9</sup> Hong Kong University <sup>1</sup> Shunghai Jiao Tong University <sup>3</sup> Addis Ababa	Baris Kasikci* of Science and Technology Institute of Technology		
<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	act roviders	A Hyperviso	r for Shared-Monory FPGA Platforms. In J	mondingi	
<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	ators for				
<section-header><section-header><section-header><section-header><section-header><section-header><section-header><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	runnie				
<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	s confi				
<section-header><section-header><section-header><section-header><section-header><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></section-header></section-header></section-header></section-header></section-header>	mem				
<section-header><section-header><section-header><section-header><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></section-header></section-header></section-header></section-header>	fs rueti Thus.				
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	rogram				
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	limite	Sharing, Protection, a	and Compatibility for I	Reconfigurable Fabric with	
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	aper p		AMORPHOS		
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	rs both				
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	wide ef GA. To	Ahmed Kha	waja <sup>1</sup> , Joshua Landgraf <sup>1</sup> ,	Rohith Prakash <sup>1</sup> ,	
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	quenc	Michael Wei	<sup>2</sup> , Eric Schkufza <sup>2</sup> , Chris	topher J. Rossbach <sup>3</sup>	
<text><section-header><section-header><section-header><section-header><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></section-header></section-header></section-header></section-header></text>	e each nique o	The Universi	ty of Texas at Austin <sup>2</sup> VMv	ware Research Group	
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	n. To se	<sup>3</sup> The Universit	ty of Texas at Austin and VM	ware Research Group	
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	TIMUS				
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	A and a fd benc	Abstra			
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	ncepts	Cloud providers such as An			
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	s: • Se	begun to support on-demand I			
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	tr OP	croud, and hardware vendors will processors. At the same time, t	100 12 - 10 - P	CONTRACTOR AND CONTRACTOR CONTRACTOR	
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	ference	such as 3D stacking, through-	The Feniks	FPGA Operating System for Cloud Co	mputing
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	Ma, Gel Mulup	sive parallelism of current FP0	E	anal Voneniane Vional Nimori Vid Das Ch. 17 D.	in 1 ifi
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	n to mak	extremely large applications,	mansong Zh	Peng Cheng <sup>§</sup> Guo Chen <sup>§</sup> Thomas Moscibroda <sup>§</sup>	n. s.e
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	r classroo istributed	System support for FPGAs.		# Microsoft Research Tsinghua University #USTC	
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	and the 1 k owned!	Unlike software, where resour	(jiazhang	yqx,ningyixu,v-ranshu,v-bojli,pengc,guoche,moscitho)@microsoft.co	m
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	us permit stip to Ini	FPGAs provide a multi-dime	ABSTRACT	EC2 instances mounting multiple FPG	As to cloud users [2]:
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	is from p N. Masci	known as the FPGA fabric: 1	Driven by explosive d	The Law Avenue A Programmer and a second sec	at CCD is is-
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	gyright b	ing to spatial constraints on F	FPGAs into datacente		
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>	\$ 978-1- org/10.1	stacks either support only a sing	eration. In this paper,		
<text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>		signs cannot efficiently support	ment in datacenters. X	A Claud Carla Area	In the Architecture
<text><text><text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text></text></text>		size of the largest slot places a	FPGA accelerators, s	A Cloud-Scale Acce	leration Architecture
A the properties About a properties About		cation size, and oversized slot resources and reduced concurrs	(1) development and		
<text><text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text></text>		This paper presents AMORP	to share an FPGA chi	Adrian M. Caulfield Eric S. Chung Andrew Putnam	
Interacting process, building properties in structure in the structure		phlets provide isolation and pr	(3) an FPGA resource	Han Angepat Jeremy Fowers Michael Haselman Mephen Heil Matt Humphrey Puneet Kaur Joo Yomer Kim Daniel Lo. Todd Masseneill. Kalin Owyhanov	
<text><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></text>		distrustful protection domains,	acenter. We implement Cotorool: Shall and Alu	Michael Papamichael Lisa Woods Site	aram Lanka Derek Chiou Doug Burger
<text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text>		of software processes. Morphlet altering their deployed form b	show that device-to-d	2.00000000	
<text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text>		ments and availability. To bui	sible and efficient. A a	Microsoft	Corporation
extend resource requirement is indicated. Subjective in analysis. Morphetes on the subjective Morphetes on the		with AMORPHOS, along with	L DEDODUCE		
<text><text><text><text><text><text></text></text></text></text></text></text>		external resource requirement	Driven by explosion d	Aluraci-Hyperscale datacenter providers have struggled to halance the grouning need for specialized hardware (efficiency)	number of configurations, or whether it is most cost-effective
minipless Mophless on the lingt interactive teach is the production of the produ		the parameter space, generatit of varying size and resource re-	down of Moore's law,	with the economic benefits of homogeneity (manageability). In	with the target workloads through its deployment lifetime (e.g.
<ul> <li>Index is more power</li> <li>Without Street Pick A and Street Pick A street Pick A</li></ul>		multiplexes Morphlets on the	huge interests recent service providers are	reconfigurable logic to accelerate both network plane func-	six years: two years to design and deploy the accelerator and
FVG. 2014 OML, rob ording the system. FVG. 2014 OML, ro			loads to more power	tions and applications. This Configurable Cloud architecture places a layer of reconfigurable logic (FPGAs) between the	is a challenge given both the diversity of cloud workloads
<ul> <li>citizity worklaad, A. Gar provide the height provide height provide the height provide the height provide the hei</li></ul>			PPGA and ASIC, so a deep learning inference	network switches and the servers, enabling network flows to be	and the rapid rate at which they change (weekly or monthly).
The second set is the product of the set of the second second set of the second secon			existing workloads. A	of local applications running on the server, and enabling the	It is thus highly desirable that accelerators incorporated into
<ul> <li>base dealed to signify the stand by the sta</li></ul>			can provide the higher power/cost efficiency (	FPGAs to communicate directly, at datacenter scale, to harvest remote FPGAs manuel by their local servers. We deployed this	examples being FPGAs and GPUs.
Between the states of the states in the state of the state is an interaction of the states in the state is a state of the sta			have decided to deple	design over a production server bod, and show how it can be	Both GPUs and FPGAs have been deployed in datacenter
<ul> <li>and other workloads.</li> <li>Tommosti de lange statistica de la construction de la constr</li></ul>			Microsoft has started to accelerate Bing ra	network acceleration (encryption of data in transit at high-	infrastructure at reasonable scale without direct connectivity between accelerators [1], [2], [3]. Our recent publication
Transmiss to task data in the process of the solution of second seco			and other workloads;	speeds). This architecture is much more scalable than prior work which used secondary rack-scale networks for inter-FPGA	described a medium-scale FPGA deployment in a production
<ul> <li>The structure of the late of</li></ul>			Permission to make digital	communication. By coupling to the network plane, direct FPGA-	datacenter to accelerate Bing web search ranking using multi-
<ul> <li>an antice and the bit rest in the rest i</li></ul>			personal or classroom use is made or distributed for prof	previous work, without the secondary network. Additionally, the	pre-directory-connected accelerators [4]. That design consisted of a tack-scale fabric of 48 FPGAs connected by a accordary
<ul> <li>undia permuta fiorpi permuta to permuta derivativa de la construction de la</li></ul>			this notice and the full citat of this work owned by other	scale of direct inter-FPGA messaging is much larger. The average round-trin latencies observed in our measurements are also	network. While effective at accelerating search ranking, our
<ul> <li>protection for production of the Configuration Canadi and th</li></ul>			credit is permitted. To copy redistribute to lists, requires	1000, and 250,000 machines are under 3, 9, and 20 microseconds,	hist architecture had several significant limitations:
<ul> <li>CONTRACT TO A TRANSPORT</li> <li>C INTEGORCTION</li> <li>A INTEG</li></ul>			permissions from permissio APSer '17, Mambai, Juffer	deployed at hyperscale in Microsoft's production datacenters	and complex cabling, and required awareness of the physical
1. INTRODUCTION 1. INTRODUCTION 1. INTRODUCTION 1. INTRODUCTION 1. INTRODUCTION 3. Address in the presence of the strength of of the stren			0 2017 ACM. 978-1-4503	worldvide.	location of machines.
Modern bysperciarle distances have made large utilds with a sin obtaine of sole under conta thaten paterna improvements in interviewing without and the second sec			LOC 10.1145/3124040.312	1. INTRODUCTION	<ul> <li>rannee musting or the torus required complex re-source loss of traffic to neighboring nodes, causing both performance loss</li> </ul>
The sample of PFOA that could communicate and information in the second				Modern hyperscale datacenters have made huge strides with	and isolation of nodes under certain failure patterns,
ambuter as they have for years individual servers with multicover CPUs, DBAM as local senser, rescale and the servers in the PDE factors and accelerate applications but offeed little local acceleration of the servers in the servers in the individual servers with the servers in the individual servers in the servers in the magnetal finguidy across the infrarenteries, and management in this paper, we describe a new cloud scale. First, Bob et al. CPU validing and provide across the servers in the server servers in the individual servers in the infrarenteries in the server in the bowers. Lans have resulted in a growing teel for band- ware operation growing servers in the bowers pointing spectral acceleration in the servers in the bowers pointing servers in the part of servers in the bowers pointing servers in the part of servers in the bowers pointing servers in the part of servers in the server servers in the pointing servers in the server servers in the pointing servers in the bowers pointing in the part of servers in the servers in the servers in the servers in the servers in the server server server server server server servers in the servers in the servers in the server in the servers in the servers in the servers in the servers in the servers in the server in the servers in the servers in the servers in the servers in the servers in the servers in the serve				and infrastructure management, but still have the same basic	<ul> <li>The number of FPGAs that could communicate directly, without point through software, was limited to a sincle rack.</li> </ul>
ministers CPUs, DBAM and host wrapp, consected by put host factors and intervents. Adjustment of the strength lifetent within the strength constant interventing and migrated factors in maximizing homespectry, workloads can be an impacted factors in the strength constant interventing and migrated factors in the strength constant interventing and migrated factors in the strength constant interventing and the strength constant interventing and the strength constant interventing and the strength constant interventing and the strength constant interventing and the strength constant interventing and the strength constant interventing and the strength constant interventing and the strength constant interventing and the strength constant interventing and add homespectry. The strength constant is and the strength and interventing accounts in a strength constant in the strength constant is a strength constant interventing and the strength and interventing accounts in a strength constant in the strength constant is a strength constant in the strength constant in a strength constant in the strength constant in the strength constant in the strength constant in the strength constant in the strength constant in the strength constant in the strength constant in the strength constant in the				structure as they have for years: individual servers with	(i.e. 48 nodes).
Chambands of the dissociation multimose of services it, there are signated in care to be multimose of services it, there are signated in any ingrated finally services the services in the service intervention of the service intervention in this properties with the service intervention of the service intervent				manicore CPUs, DRAM, and local storage, connected by the NRC through Ethernet switches to other servery. At honerscale	<ul> <li>The fabric was a limited-scale "holt on" accelerator, which could accelerate applications but offered little for otherwise</li> </ul>
icard leartists to maximizing homogeneity: workloads can be fines. migrated fuelphy, zower to be atfinisteric, and management is simplified, reflexing nois and configuration errors. Is simplified reflexing to an additional simplified reflexing the simplified reflexing the simplified reflexing to a simplified reflexing the simplified reflexi				(hundreds of thousands to millions of servers), there are signif-	the datacemer infrastructure, such as networking and storage
The simplified radiating even and comfiguration across. Both the shordware is 1CPL using and the radiag of the short is 1CPL soling and the radiag of the short is 1CPL soling and the radiag of the radiation of the short is 1CPL soling and the radiag of the radiation of the radiation of the radiation of the radiation of the radiation of the radiation of the physical inframeworks a servers realizes the lightly doub able howsparenci. The application is nonly one of eccentration is the radiation of the rad				scant benefits to maximizing homogeneity; workloads can be migrated famelide across the infrastructure and monotone	fows.
Both the shore-four is CPU scaling and the ending of "Code which demaness of order instantoms based be Moort's Law have readed in a growing end of hands are specialization to siccrease performance and efficienty. However, piccing specialization controls in subher of a proportion dimension of the strength of the strength of the strength of the strength of proportion dimension of the strength of the strength of the strength of the strength of the strength of the strength of the strength of the ended is in cost-efficient to deploy an accelerate in every new strengt, whether it is better to specialize a subset of an information of the strength of th				is simplified, reducing costs and configuration errors.	in this paper, we describe a new cloud-scale, FPGA-based acceleration architecture, which we call the Confoundate
boost the same treatment of a proving most for black and the second se				Both the slowdown in CPU scaling and the ending of	Cloud, which eliminates all of the limitations listed above with
However, placing specialized accelerators in a subset of z → dension in the majority of nue servers in Man 15 count hyperscele informations servers reduce which highly douts. The biomegarity is the special of the server is the server in the server is the server is the whether it is constrainty in doping an acceleration is every new server, whether it is better to specialize a subset of z hardware. The disorders are server and maintain an every provide applications. The server is a server is an accelerator is every to a subset of z = 10000000000000000000000000000000000				stoore's Law have resulted in a growing need for had- ware specialization to increase performance and efficiency	a single design. This architecture has been and is being
byperscale infratorieures's servers robues the lightly done's constants. A configurable Cloud allows the data adult homogenesity. The agreetion is mody user occommission is the accordenation in the accordenation in the proceeding with program whether it is construction to according to a scattering of the advances. This disquite non-include networking for an infrattructure's new servers and mutania an every provincias, and distributed (multi eris i non-integram servers) and mutania an every provincias. and distributed for advances of the i non-integram servers and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias. and distributed (multi revision sites integram servers) and mutania and every provincias and distributed (multi revision servers) and mutania and every provincias and distributed (multi revision servers) and mutania and every provincias and distributed (multi revision servers) and mutantia and every provincias and distributed (multi revision servers) and mutania and every provincias and distributed (multi revision servers) and mutania and every provincias and distributed (multi revision servers) and mutantia and every provincias and distributed (multi revision servers) and mutantia and every provincias and distributed (multi revision servers) and mutantia and every provincias and distributed (multi revision servers) and mutantia and every provision servers and mutantia revision ser				However, placing specialized accelerators in a subset of a	- depuyee in the majority of new servers in Microsoft's production datacenters across more than 15 countries and
and comparison to a spectra of the spectra acceleration in every hardware. These accelerated with progra- subledner it is coefficient to deptoy an acceleration in every hardware. This charapta can include meteorising for new server, whether it is benter to specialize a subset of gate from, security operations, and distributed (multi an inframeturity from servers and muntitian an every growing applications. with a smootime 5000 5000 00 2000 antic.				hyperscale infrastructure's servers reduces the highly desir-	5 continents. A Configurable Cloud allows the datapath of
new screte, whether it is better to specialize a subset of gata flass. The diapped can also also also better distribution (sub- ma informativer's wavers and mutataia as ever growing applications. and distribution (sub- rus 1 smoother 50031070 ≥ 2164 arXii.				whether it is cost-effective to deploy an accelerator in every	cloud communication to be accelerated with programmable
an infrastructure's new servers and maintain an every growing applications. www.s.sono.3500.5500.5500.5500.5500.5500.5500.550				new server, whether it is better to specialize a subset of	age flows, security operations, and distributed (multi-FPGA)
978-1 5890 3506 3516751-09 @ 2016 HEE. The key difference over previous work is that the a				an infrastructure's new servers and maintain an ever-growing	applications.
				978-1 5880 2508-3/16/E31.00 (2) 2016 HEE	The key difference over previous work is that the accelera-

#### Coyote HW Architecture

- Microkernel for FPGAs
  - Hardware split into 3 regions:
    - Static region
    - Service (shell) region
    - Application region (further split into trusted wrapper and untrusted user app)
- Host software consists of:
  - Kernel driver (hypervisor)
  - Runtime scheduler
  - High level API
- Runs on Alveo, Enzian





# Host <-> FPGA connection

- PCIe connection (cards still at gen3 x16 ...)
- XDMA IP<sup>[1]</sup> core used with Alveo cards (trial runs with QDMA core as well)
- Very little dependency on the actual DMA core used
- Port to open source cores



[1]: https://docs.xilinx.com/r/en-US/pg195-pcie-dma/Introduction



6

# **ECI** – Enzian Coherency Interface

- Cache coherent interface<sup>[2]</sup>
- Coherency between CPU and FPGA side memories
- Our own DMA wrapper built on top which provides integration for Coyote interfaces



7

[2]: Enzian: An Open, General, CPU/FPGA Platform for Systems Software Research, ASPLOS 2022



# ECI vs XDMA



- Throughput PCIe:~12 GB/s
- Throughput ECI:~21 GB/s

![](_page_7_Picture_4.jpeg)

#### Coyote Processes, Threads, Tasks ...

- FPGAs are fundamentally different
  - No CPUs or cores
  - Spatial and temporal sharing
- Coyote multi-tenancy:
  - Combines both approaches
  - Multitasking abstraction for a set of independent, isolated vFPGAs
  - Applications within vFPGAs are untrusted

![](_page_8_Figure_8.jpeg)

#### Coyote Virtual Memory

- Important abstraction
  - Provides the illusion of unlimited memory, simplifies compilation, provide protection
- Typically missing in FPGAs !
- Dedicated MMU layer in Coyote
  - Same virtual address space between HW SW
  - Applications can master transactions (pointer chasing example ...)
  - RDMA ...
  - Virtualization of FPGA side memory
  - Additional memory models on top

![](_page_9_Picture_10.jpeg)

![](_page_9_Picture_11.jpeg)

# Memory organization (DRAM)

- Dynamic allocation across all available channels
- Striping access pattern
- ✤ Same interface (DRAM, HBM, ...)
- Optimizes bandwidth distribution across vFPGAs

![](_page_10_Figure_5.jpeg)

![](_page_10_Picture_6.jpeg)

# Memory Organization (HBM)

- Single channel has access to the whole memory
- RAMA IP<sup>[3]</sup> core used for striping (same idea ...)
- Current HBM in Alveo cards:
   32 AXIM channels at 450 MHz
- Data width conversion from
   32B to 64B, max thr: ~14 GB/s
- Coyote typically runs at ~300 MHz, max thr: ~19 GB/s

![](_page_11_Figure_6.jpeg)

[3]: https://www.xilinx.com/products/intellectual-property/rama.html

![](_page_11_Picture_8.jpeg)

#### **Coyote** Shared Virtual Memory (Unified Memory)

- Programming model which allows multiple devices with independent memories to be programmed as if they share memory
- Seamless sharing of host and device side user space pointers
- Provides memory consistency between host and device accesses (Heterogeneous Memory Management API)
- Used extensively in modern GPUs

![](_page_12_Figure_5.jpeg)

![](_page_12_Picture_6.jpeg)

# Unified Memory Example

// NO UNIFIED MEMORY

// WITH UNIFIED MEMORY

IRQ handling overhead:

vm: mean: 2430.298 medain: 2396.0 std: 644.358263077304 host: mean: 2470.978 medain: 2402.0 std: 1608.3578157599134

![](_page_13_Picture_5.jpeg)

# FPGAs are not GPUs ...

- Direct access and corresponding interfaces are needed (streaming interfaces)
- Especially if network and/or other I/O comes into play ...
- Bump-in-the-wire processors
- How to integrate this within HMM(SVM)?

![](_page_14_Figure_5.jpeg)

![](_page_14_Picture_6.jpeg)

# HMM operation

![](_page_15_Figure_1.jpeg)

![](_page_15_Picture_2.jpeg)

# **MMU Notifiers**

- Traditionally pages pinned prior to device access
- Callback functions within MMU notifiers can keep the TLBs up to date without pinning
- Invalidations need to be handled within properly

![](_page_16_Figure_4.jpeg)

![](_page_16_Picture_5.jpeg)

# HMM Obstacles ...

- Coverage with hugePages
   (hugetlbfs), variable TLB page size
- Transparent HugePages (khugepaged thread)

✤ API being changed ...

![](_page_17_Figure_4.jpeg)

![](_page_17_Picture_5.jpeg)

![](_page_18_Picture_0.jpeg)

- 100G TCP/IP and RDMA (RoCE v2) network stacks<sup>[4]</sup>
- Reliable protocols (use single channel of on-board memory)
- ✤ ~20% of overall FPGA resources
- Interleaved, accessible by all tenants (vFPGAs)

[4]: https://github.com/fpgasystems/fpga-network-stack

![](_page_18_Picture_6.jpeg)

## TCP stack

- Shared across vFPGAs
- Multiple accelerators serving clients on different ports
- AXI stream interface (opening and closing ports and connections,
- Lower level of abstraction than RDMA !

![](_page_19_Figure_5.jpeg)

[5]: https://github.com/fpgasystems/Vitis\_with\_100Gbps\_TCP-IP/blob/vitis\_2022\_1/img/interface.png

![](_page_19_Picture_7.jpeg)

## **RDMA** stack

- ✤ Open source RDMA stack (UC, RC)<sup>[5]</sup>
- RDMA over Converged Ethernet (RoCE v2)
- Implemented on top of UDP/IPv4/IPv6 (far lower overhead than iWARP)
- InfiniBand (IB) transport packets over Ethernet (READ, WRITE, SEND)

![](_page_20_Figure_5.jpeg)

https://docs.nvidia.com/networking/display/WINOFv55053000/RoCEv2

[5] **StRoM**: Smart Remote Memory, EuroSys '20 David Sidler, Zeke Wang, Monica Chiosa, Amit Kulkarni, Gustavo Alonso

#### **Coyote** Unified Logic Interface

- Don't really exist across FPGA platforms ...
  - ✤ Vitis HLS
  - Much higher flexibility than other devices
- User Logic Interface
  - AXI streams (descriptor interface)

![](_page_21_Figure_6.jpeg)

![](_page_21_Picture_7.jpeg)

# **Unified Logic Interface**

![](_page_22_Figure_1.jpeg)

![](_page_22_Picture_2.jpeg)

## Can this be simplified?

![](_page_23_Figure_1.jpeg)

![](_page_23_Picture_2.jpeg)

# SQ Interface

All local(host, card, ...) and remote(RDMA, TCP/IP) operations can be invoked through SQ interface

- OPCODE Operation to be invoked
- CPID Host process identifier
- VFID vFPGA (reserved)
- DEST Destination queue
- SID Connection ID
- VADDR User space shared pointer
- LEN Transfer size
- RQ Received remote requests (RDMA and TCP/IP)
- CQ Completion events

![](_page_24_Picture_12.jpeg)

## **Destination queues**

![](_page_25_Figure_1.jpeg)

![](_page_25_Picture_2.jpeg)

# **Untrusted environment**

- Credit system for each destination queue:
  - Write requests are issued only if accompanying data is provided
  - Read requests are issued only if queues are ready to accept the data
- Credit system for all local and remote requests
- Can quickly overwhelm the resources ...
- For RQ, RDMA accepts packets only if vFPGA is able to accept

![](_page_26_Figure_7.jpeg)

![](_page_26_Picture_8.jpeg)

## Coyote

Dynamic Reconfiguration and Scheduling

- Basic mechanisms to capture the state of the FPGA don't exist
- Non-preemptive task based approach
- Preemption?
  - User application trust
  - Requires a form of cooperation
  - Additional application complexity

![](_page_27_Figure_8.jpeg)

![](_page_27_Picture_9.jpeg)

# Reconfiguration overhead and scheduling efficiency

Penalty of partial reconfiguration is high

Modified priority queue based scheme reduces overall execution time

![](_page_28_Figure_3.jpeg)

#### Scheduling algorithm

![](_page_28_Figure_5.jpeg)

![](_page_28_Picture_6.jpeg)

**Reconfiguration times** 

# **Hierarchical Reconfiguration**

- Nested reconfiguration (Nested-DFX)
  - □ <u>Static</u> layer
  - □ <u>Service</u> layer
  - Dynamic layer
  - □ <u>Application</u> layer

![](_page_29_Figure_6.jpeg)

![](_page_29_Figure_7.jpeg)

![](_page_29_Figure_8.jpeg)

![](_page_29_Picture_9.jpeg)

![](_page_30_Picture_0.jpeg)

- Virtualization –vFPGAs «pulled up» all the way to different VMs
  - Virtual Function I/O (VFIO) Mediated devices

![](_page_30_Figure_3.jpeg)

![](_page_30_Picture_4.jpeg)

# **Virtualization Architecture**

- Hypervisor running on top of Coyote driver within Linux kernel
- Virtualization of CPU and memory with KVM
- For emulation of other components QEMU
- Passthrough through IOCTL interface and emulation of PCI device with VFIO + MDEV

![](_page_31_Figure_5.jpeg)

![](_page_31_Picture_6.jpeg)

## **VM Performance**

![](_page_32_Figure_1.jpeg)

![](_page_32_Picture_2.jpeg)

#### Coyote SW Architecture

- Layered parallelization
- User space abstractions
  - □ <u>cSched</u> Coyote scheduler, reconfiguration controller
  - CProc Coyote process, multiple can run within a single vFPGA
  - CThread Coyote thread, multiple can run within a single *cProc*. Task level parallelism
  - CTask Coyote task, arbitrary user variadic function executed by *cThreads*
  - <u>cService</u> <u>Coyote</u> library daemon, background service, UDS for IPC

```
/**
 * Open a Unix Domain Socket and send a decrypt_and_compress task request
 * This is the only place of interaction with Coyote
 */
cLib clib("/tmp/coyote-daemon-vfid-0);
clib.task({opDecryptAndCompress, {mem, size, key}); // blocking
```

![](_page_33_Picture_9.jpeg)

#### Coyote Build System

- CMake incremental builds
  - Revamped for hierarchical flow
- Cores can be both RTL and HLS

```
hw/aes
cyt_user_top.sv
hdl
*.sv, *.svh, *.vhd
hls
*.cpp, *.hpp
```

- <u>make shell</u> (shell partial bitstream)
- <u>make app</u> (application partial bitstreams)

cmake\_minimum\_required(VERSION 3.0)
project(test)

```
set(CYT_DIR ${CMAKE_SOURCE_DIR})
set(CMAKE_MODULE_PATH ${CMAKE_MODULE_PATH} ${CYT_DIR}/cmake)
```

find\_package(CoyoteHW REQUIRED)

```
# Configuration
set(SHELL_PROBE 1)
set(FDEV_NAME "u55c")
set(COMP_CORES 80)
set(N_REGIONS 2)
set(EN_STRM 1)
set(EN_MEM 1)
set(EN_PR 1)
set(N_CONFIG 3)
set(N_CARD_AXI 2)
set(N_HOST_AXI 2)
```

```
# Load applications
load_apps(
```

VFPGA\_C0\_0 "hw/adder" VFPGA\_C0\_1 "hw/adder" VFPGA\_C1\_0 "hw/aes" VFPGA\_C1\_1 "hw/aes" VFPGA\_C2\_0 "hw/sha" VFPGA\_C2\_1 "hw/sha"

create\_hw()

![](_page_34_Picture_14.jpeg)

### Coyote

What kind of stuff can we run on top?

- ACCL project (Accl: Fpga-accelerated collectives over 100 gbps tcp-ip, Z. He et al.)
  - Streaming interfaces
  - Kernel invocation overhead Vitis: ~50us, Coyote: ~1–1.5us
  - RDMA

![](_page_35_Figure_6.jpeg)

![](_page_35_Picture_7.jpeg)

### **Questions?**

![](_page_36_Picture_1.jpeg)