

# Do OS abstractions make sense in FPGAs?

**Dario Korolija** Timothy Roscoe Gustavo Alonso

Systems Group, Dept. of Computer Science, ETH Zurich



#### Introduction

 Hybrid computing systems (Amazon F1, Microsoft Catapult, Intel HARP, Alibaba F3...)



• At ETH we have built one of these:







# Hybrid computing systems are complex to program

- CPU-FPGA interaction
- Difficult to program
- No standard execution environment
- Lack of portability
- ✤ Lack of proper abstractions



FPGA layout

PCIe, memory, storage, network...



## OS functionality is starting to appear

- Shells for hybrid systems:
  - Catapult, HARP, SDAccel, Vitis
- Memory Management:
  - CoRAM
- ✤ Virtualization and scheduling :
  - Vital, Optimus, AmorphOS...

Focus on particular subsets of functionality only!

```
A Hypervisor for Shared-Memory FPGA Platforms
                                   Jiacheng Ma* Gefci Zuo* Kevin Loughlin* Xiache Chengl* Yanqiang Liu*
Abel Mulugeate Eneyes* Zhengwei Q1* Baris Katika*
"University of Michage - Hong Kong University of Konea and Tohonooy
"Shanghai Jiao Tong University * Addis Ababa Institute of Technology
Cloud providers
accelerators for
plex their FFGA
reducing runni
support is conti
centric program
to direct mem-
incurs high num
chasing. Thus,
memory progra-
DMAs. However
FFGAs is limin
This paper 7
supports scalab
Mtts offers bodi
ing to provide c
on an FPGA. To
clock frequency
To isolate each
the technique of
                                                                                                                                          Sharing, Protection, and Compatibility for Reconfigurable Fabric with
                                                                                                                                                                                                                                                                                                                 AMORPHOS
                                                                                                                                                                                                                  Ahmed Khawaia<sup>1</sup>, Joshua Landeraf<sup>1</sup>, Rohith Prakash<sup>1</sup>
                                                                                                                                                                                                           Michael Wei<sup>2</sup>, Eric Schkufza<sup>2</sup>, Christopher J. Rossbach<sup>3</sup>

<sup>1</sup>The University of Texas at Austin <sup>2</sup>VMware Research Group
    the technique
co-design. To s
TIMUS provide
that OPTIMUS
gle FPGA and
real-world ber
                                                                                                                                                                                                        <sup>3</sup>The University of Texas at Austin and VMware Research Group
  CCS Concepts
FPGAs: • Se
machines
                                                                                                                                        Cloud providers such as Am
                                                                                                                                   egun to support on-demand I
  Keywords OP
ACM Reference
Jiacheng Ma, Ge
Lin, Abel Malog
                                                                                                                                                        sors. At the same time,
                                                                                                                                                                                                                                                                                                                 The Feniks FPGA Operating System for Cloud Computing
                                                                                                                                 such as 3D stacking, through
FinFETs have greatly increase
                                                                                                                                 sive parallelism of current FF
                                                                                                                                                                                                                                                                                                                                                     Jiansong Zhang<sup>§</sup> Yongqiang Xiong<sup>§</sup> Ningyi Xu<sup>§</sup> Ran Shu<sup>§†</sup>
Peng Cheng<sup>§</sup> Guo Chen<sup>§</sup> Thomas Moscibroda<sup>§</sup>
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        Boiie Li
  En, Afei Malag
Permission to ma
personal or classer
made or distribute
this notice and the
of this work owned
with could is perm
to redistribute to lis
permissions firm,
to redistribute to lis
permissions firm,
to 2020 Copyright
ACM.
ACM ISBN 978-1
https://doi.org/10
                                                                                                                                 extremely large applications,
simultaneously as well.
                                                                                                                                                                                                                                                                                                                                                                                                                 <sup>1</sup> Microsoft Research <sup>†</sup>Tsinghua University <sup>‡</sup>USTC
                                                                                                                                   System support for FPGAs,
Unlike software, where resour
                                                                                                                                                                                                                                                                                                                                                                           (jiazhang,yqx,ningyixu,v-ranshu,v-bojli,penge,guoche,moscitho)@mia
                                                                                                                                 ited to simple dime
                                                                                                                                                                                                                                                                                                   ABSTRACT
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       ces mounting multiple FPGAs t
                                                                                                                              FPGAs provide a multi-dim
                                                                                                                                 known as the FPGA fabric:
                                                                                                                                                                                                                                                                                                   Driven by explosi
down of Moore's
                                                                                                                                 units, memories, and I/O can
                                                                                                                                 ing to spatial constraints on
                                                                                                                                                                                                                                                                                                   FPGAs into data
                                                                                                                                   tacks either support only a sin
                                                                                                                                                                                                                                                                                                   eration. In this pape
FPGA, called Fenik
                                                                                                                                 partition the FPGA fabric into
                                                                                                                                   igns cannot efficiently suppo
                                                                                                                                                                                                                                                                                                     ment in datacenters.
FPGA accelerators,
underlying hardware
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           A Cloud-Scale Acceleration Architecture
                                                                                                                                   ize of the largest slot places
                                                                                                                                 cation size, and oversized slot
                                                                                                                                               surces and reduced concu
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  Adrian M. Caulfield Eric S. Chung: Andrew Putnam
Hari Angepat Jeremy Fowers Michael Haselman Stephen Heil Matt Humphre
Puneet Kaur Joo-Young Kim Daniel Lo Todd Massengill Kalin Ovtcharov
                                                                                                                                        This paper presents AMOR
                                                                                                                                   aser FPGA logic in morphabl
                                                                                                                                 phlets provide isolation and pr
                                                                                                                                 distrustful protection domain
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         Michael Papamichael Lisa Woods Sitaram Lanka Derek Chiou Dour Burger
                                                                                                                                                                                                                                                                                                     Catapult Shell and
                                                                                                                                 of software processes. Morphi
                                                                                                                                                                                                                                                                                                     show that device
                                                                                                                                 altering their deployed form
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           Microsoft Corporation
                                                                                                                                                                                                                                                                                                     sible and efficient.
                                                                                                                                   nents and availability. To bu
                                                                                                                         merns and availability. To bun
provide a parameterized hardw
with AMORPHOS, along with
external resource requirement
the parameter space, generatis
of varying size and resource re
multiplexes Morphilets on the
                                                                                                                                                                                                                                                                                                   can share an FPGA
                                                                                                                                                                                                                                                                                                                                                                                                                                    Advance-Expressed determines previous how a trangelob of an analytic of configuration, or whether is its none constraints
with the resonantic heather of propholical break with the target weakloads dready to a constraint of the
heat the gravity and analytic of the second second second second second second second
second second second second second second second second second second
second second second second second second second second second second
second second second second second second second second second second
second second second second second second second second second
second second second second second second second second second second
second second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second
second second second second second second second second second second second second
second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second second seco
                                                                                                                                                                                                                                                                                                 1 INTRODUCT
                                                                                                                                                                                                                                                                                                   Driven by explosiv
down of Moore's la
                                                                                                                                                                                                                                                                                                                                                                                                                                           used for both service acceleration (Web search ranking) and
network acceleration (encryption of data in transit at high-
speeds). This architecture is much more scalable than prior
work which used secondary rack-scale networks for inter-FPCA.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  infrastructure at reasonable scale without direct conn
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               tors [1], [2], [3]. Our recent publication
                                                                                                                                                                                                                                                                                                               accelerate Bing
                                                                                                                                                                                                                                                                                                             nd other worklos

    stark table and accounting reak-task interest her inter PT-R integrates and accounting reak-task interest in the starker at comparison largers in the starker at the sta
                                                                                                                                                                                                                                                                                                 redistribute to lists, r
permissions from per
APSys '17, Mambai,
0 2017 ACM, 978-1
DOI: 10.1145/01246
                                                                                                                                                                                                                                                                                                                                                                                                                                       respectively. In
deployed at hy-
waridwide.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                location of machines.

• Failure handling of the torus required complex re-
                                                                                                                                                                                                                                                                                                                                                                                                                             \label{eq:restrict} \textbf{i} = 1 \\ 
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        1. INTRODUCTION
                                                                                                                                                                                                                                                                                                                                                                                                                                                simplified, reducing costs and configuration errors.
Both the slowdown in CPU scaling and the ending of
Cloud, which eliminates all of the limitations listed above with
                                                                                                                                                                                                                                                                                                                                                                                                                                       Moore's Law have resulted in a growing need for hard-
ware specialization to increase performance and efficiency.
However, placing specialized accelerators in a subset of a
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             a single design. This architecture has been - and is being
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     - deployed in the majority of new servers in Microsoft'
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  production datacenters across more than 15 countries any
                                                                                                                                                                                                                                                                                                                                                                                                                                                  opercade infrastructure's servers reduces the highly desir
ble homogeneity. The question is mostly une of economics:
cload communication to be accelerated with programmable
                                                                                                                                                                                                                                                                                                                                                                                                                                    and additionation in monty with a secondary of the second
```



## Coyote

Hybrid computing system

Coyote provides a complete minimal core set of essential features above which other services can be based





### Coyote system foundations

Hardware split into *static* and *dynamic* regions

- Dynamic region split into multiple vFPGAs
  - **vFPGA** further split into the User Logic and the Dynamic Wrapper
- Functionality not on the critical path handled by the host OS
  - Kernel driver
  - Runtime scheduler
  - High level API





#### Basic multiplexing abstractions: Processes, threads and tasks

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



7





#### Execution environment: Application portability

- Don't really exist across the FPGA platforms
  - Effort by Xilinx with **Vitis**
  - FPGA not only a pure computational device
- Coyote:
  - Single User Logic Interface (ULI)
  - Interaction with the complete system
  - Access is low level
  - Portability and extensibility







### Scheduling: Non-preemptive

- Basic mechanisms to capture the state of the FPGA don't exist
- Coyote:
  - Non-preemptive task based approach
  - Avoid preemption
    - User applications can't always be trusted
    - Preemption requires cooperation
    - Preemption would impose additional application complexity
    - Additional problems of capturing the state of stateful services







#### Memory management: Access flexibility

- Virtual memory tends to be ignored in FPGAs
- Coyote implements a flexible approach giving us multiple ways to access both host and local FPGA memory
- Unified memory model built on top adds to the programmability







#### Services

- Network: TCP/IP and RDMA network stacks<sup>[1]</sup> shared between vFPGAs.
- **On-board FPGA memory:** Hiding the complexity of multiple local memory channels through *striping* abstraction
- Further services easily possible:
  - External storage
  - External accelerators (GPUs, ASICs)

[1] StRoM: Smart Remote Memory David Sidler, Zeke Wang, Monica Chiosa, Amit Kulkarni, Gustavo Alonso









## Evaluation

- Depending on the configuration Coyote shell uses 15-25% of current FPGA resources
- Compared with commercial systems Coyote achieves comparable or better performance for a real-world use case [1]
- Microbenchmarks:
  - Sharing of the resources is fair
  - Scheduling tactic reduces overall execution time
  - Abstraction performance penalties are negligible

#### **Decision trees:**



#### Striping:





Hybrid computing systems are difficult to program The need for proper abstractions is evident

#### To find the right abstractions for FPGAs, a complete set of functions has to be considered

https://github.com/fpgasystems

This work has been made possible through a generous equipment donation from Xilinx *Xilinx Adaptive Compute Cluster (XACC) Program: https://www.xilinx.com/support/university/XUP-XACC.html* 

