Edge Acceleration-as-a-Service

Jason Cong

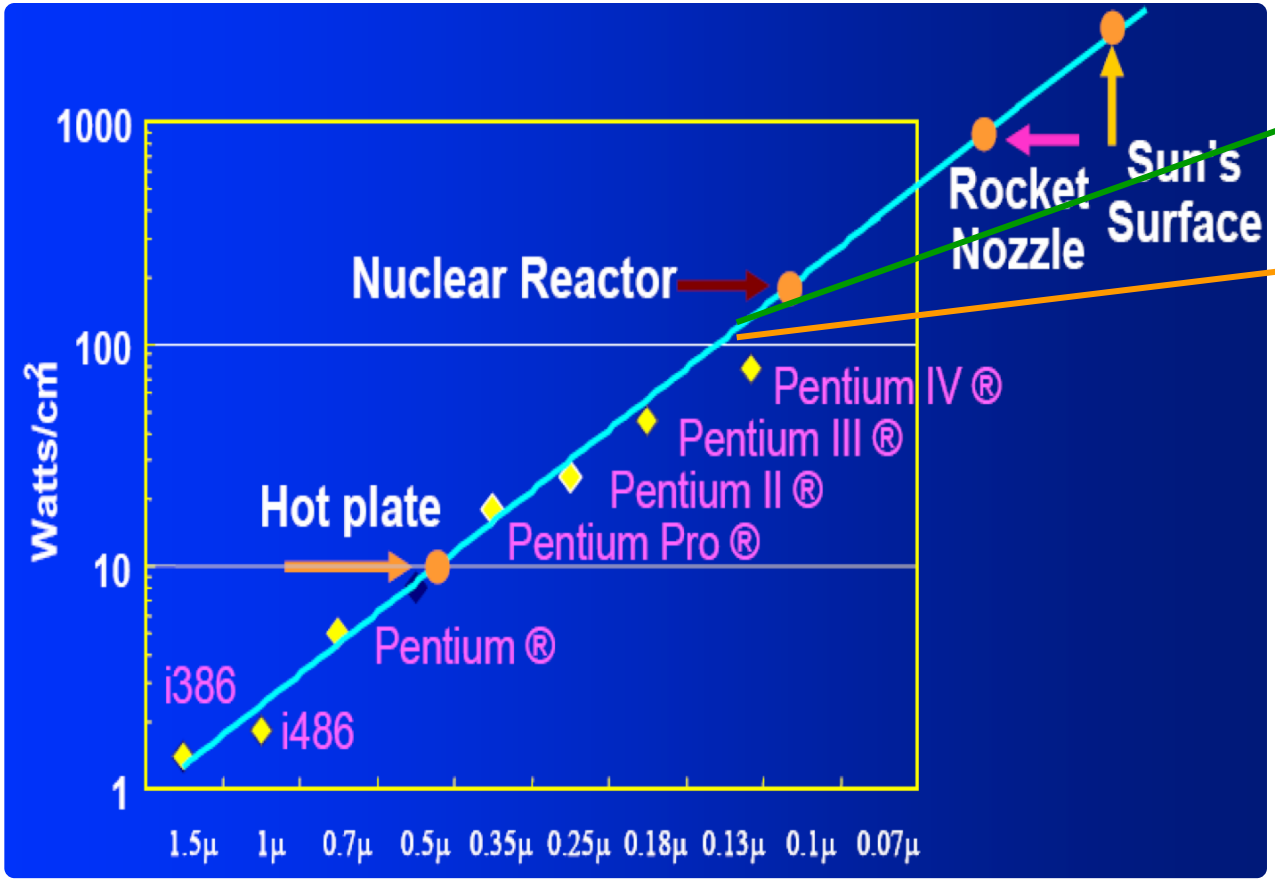
Distinguished Chancellor's Professor, UCLA

Director, Center for Domain-Specific Computing

cong@cs.ucla.edu

http://cadlab.cs.ucla.edu/~cong

Center for Customizable Domain-Specific Computing – Focus on Energy Efficient Computing *[2009 CDSC Proposal]*



Parallelization

Customization

Adapt the architecture to Application domain

Source: Shekhar Borkar, Intel

NSF Expeditions in Computing (2009) & InTrans Award with Intel (2014)

UCLA Newsroom

Home

All Stories

All Stories
Featured News
News Releases
Advisories
Images
Multimedia

Research
Health Sciences
Arts & Humanities
Student Affairs
Academics & Faculty
Campus News
Media Contacts

Images
Video
Blogs

For the Media

Contacts
News releases
Advisories
About UCLA

UCLA Newsroom > All stories > News Releases

NSF awards UCLA $10 million to create customized computing technology

By Wileen Wong Kromhout| 8/11/2009 9:45:00 AM

The UCLA Henry Samueli School of Engineering and Applied Science has been
grant by the National Science Foundation's Expeditions in Computing program
performance, energy efficient, customizable computing that could revolutionize
used in health care and other important applications.

In particular, UCLA Engineering researchers will demonstrate how the new tech
domain-specific computing, could transform the role of medical imaging and he
providing more cost-effective and convenient solutions for preventive, diagnos
procedures and dramatically improving health care quality, efficiency and patie

"This significant award is another testament to the world-class faculty here at
push the envelope to solve society's most pressing issues," said UCLA Chancell
grateful to the NSF, which has repeatedly provided crucial funding to our facult
university among the nation's top five in research funding."

In an effort to meet ever-increasing computing needs in various fields, the com
entered an "era of parallelization," in which tens of thousands of computer serv
warehouse-scale data centers, said Jason Cong, the Chancellor's Professor of C
director of the new UCLA Center for Domain-Specific Computing (CDSC), which
research. But these parallel, general-purpose computing systems still face seri
of performance, energy, space and cost.

Domain-specific computing holds significant advantages, Cong said. While gen
relies on computer architecture and languages aimed at any type of applicatio
computing utilizes a customizable architecture and custom-oriented, high-level
tailored to a particular application area or domain — in this case, medical imag
modeling. This customization ultimately results in much less energy consumpti
costs and increased productivity.

The goal of the new UCLA center, Cong said, is to look beyond parallelization a
specific customization to bring significant power-performance efficiency improv
application domains.

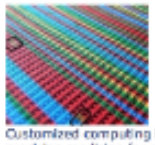
National Science Foundation
Directorate for Computer & Information Science & Engineering (CISE)

Press Release 14-086

TAKING GREAT IDEAS FROM THE LAB TO THE FAB

NSF and Intel support the development of domain-specific hardware to address health care needs

Real-time adaptive low-dose CT-scan enabled by customized computing.
Credit and Larger Version



Customized computing in search of precision medicine for cancer treatment.
Credit and Larger Version

Accelerator-rich architecture with composable and reconfigurable accelerators.
Credit and Larger Version

July 17, 2014

A "valley of death" is well-known to entrepreneurs--the lull between government funding for research and industry support for prototypes and products. To confront this problem, in 2013 the National Science Foundation (NSF) created a new program called InTrans to extend the life of the most high-impact NSF-funded research and help great ideas transition from lab to practice.

Today, in partnership with Intel Corporation, NSF announced the first InTrans award of $3 million to a team of researchers who are designing customizable, domain-specific computing technologies for use in healthcare.

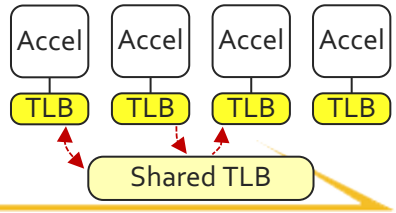
Why It Matters to This Project

- "5G is where computation and communication converge"
  - Geng Wu, Intel Fellow
- There is a great need for acceleration in the edge
- Proposed research – Acceleration-as-a-Service in NDN

What We have Learned So Far -- Levels of Customization

- **Single-chip level**
  - Require new processor designs, e.g. using composable accelerators [ISLPED' 12, DAC'14]
- **Server node level**
  - Host CPU + FPGA via PCI-e or QPI connections
- **Data center level**
  - Clusters of heterogeneous computing nodes

Chip-Level Customization: Accelerator-Rich Architectures (ARA)



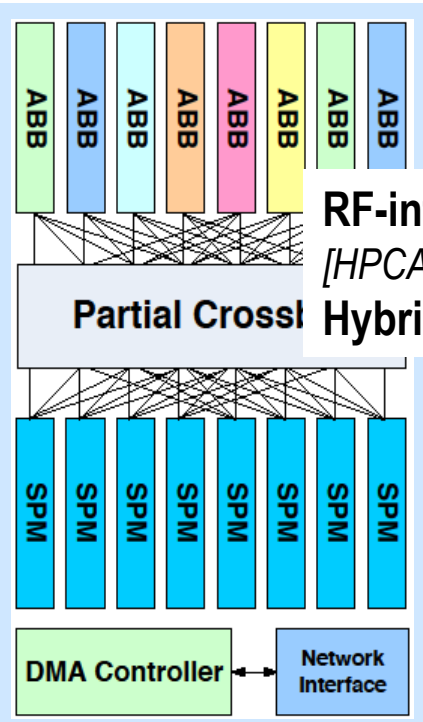
Hybrid L2 Cache with STTRAM + SRAM *[DATE 12]*

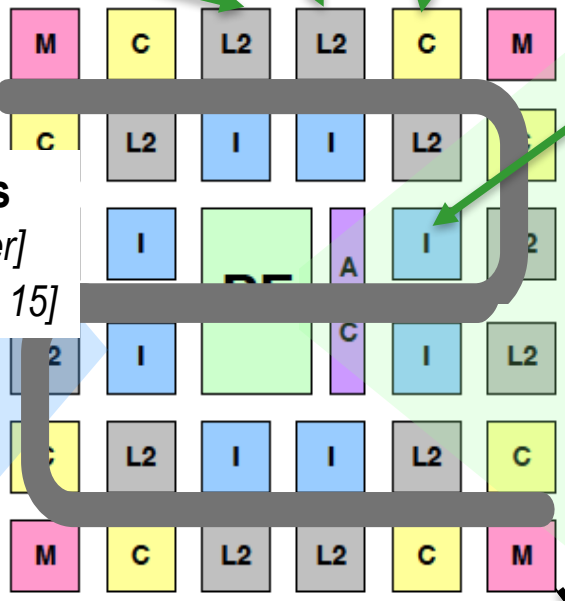
Buffer in NUCA *[ISLPED 12]*

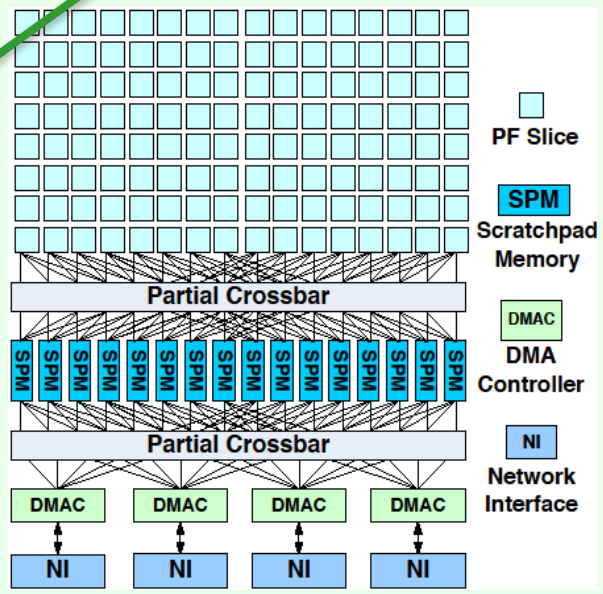
Adaptive L1 Cache + SPM/Buffer *[ISLPED 11]*

Address Translation *[HPCA 17]*

RF-interconnects *[HPCA 08 best paper]*
Hybrid NoC *[DAC 15]*



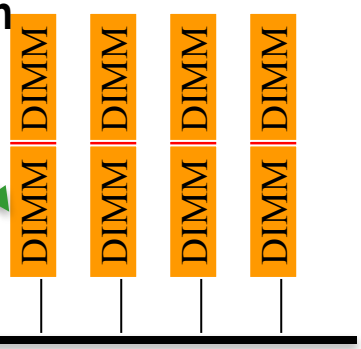




ARC, CHARM, CAMEL
[DAC 12, ISLPED 12, DAC 14]

M Mem Controller
L2 L2 Bank
I ABB Island
PF Programmable Fabric Block
C Core
ABC Accelerator Block Composer

Accelerator in Memory



[JESTCS 12]

RF-interconnects improves DRAM BW

Now the full-system ARA simulator PARADE *[ICCAD 15]* is open source

Levels of Customization

- **Single-chip level**
  - Require new processor designs, e.g. using fixed-function or composable accelerators
- **Server node level**
  - Host CPU + FPGA via PCI-e or QPI connections
- **Data center level**
  - Clusters of heterogeneous computing nodes

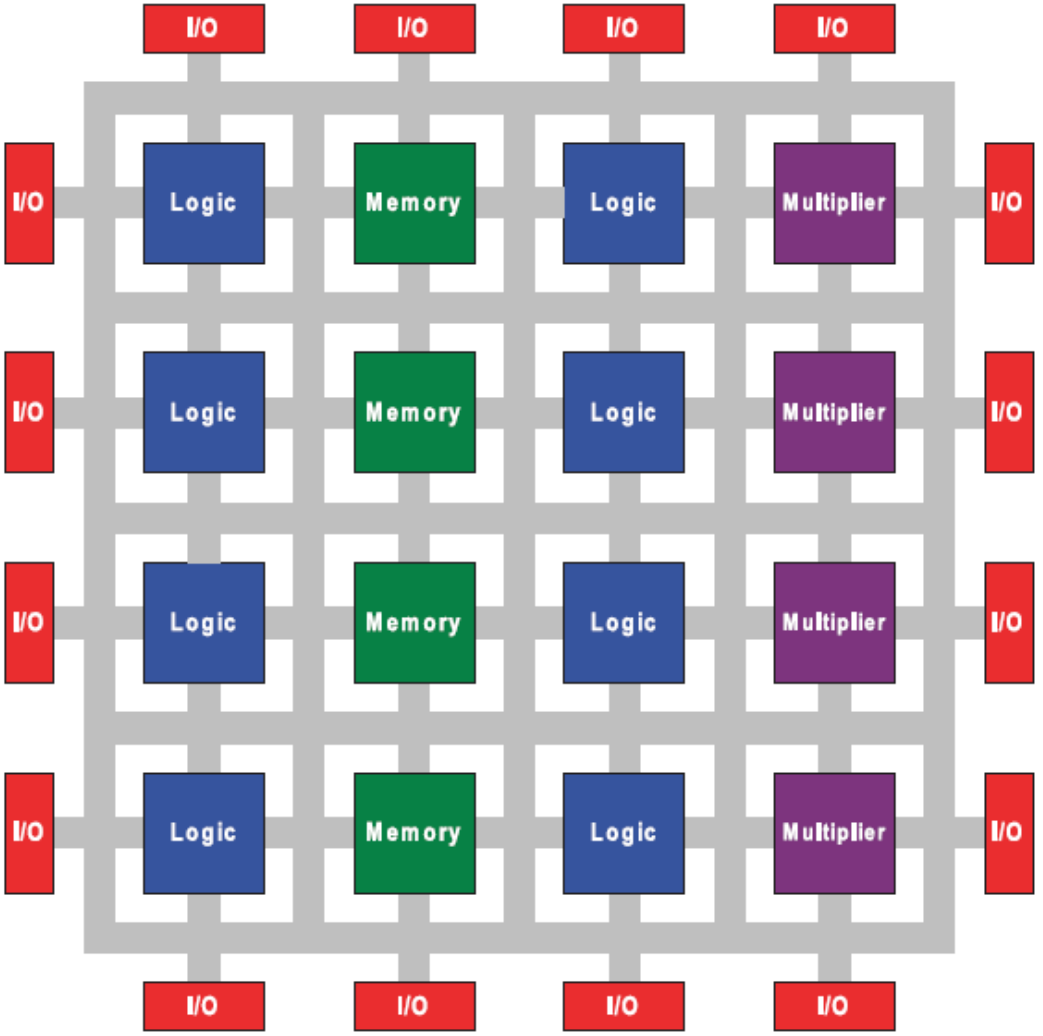
Use of FPGAs for Accelerator Implementation

Field Programmable Gate Arrays (FPGA)

- Reconfigurable hardware to accelerate specific computations
- Mature compute platforms integrated with CPU

FPGA benefits

- Low-power, energy efficient (5~30W)
- Customized high performance
  - Smith-waterman [FCCM'15]: 26x over 24-thread CPU
  - CT Recon [FPGA'14]: 4x over GPU

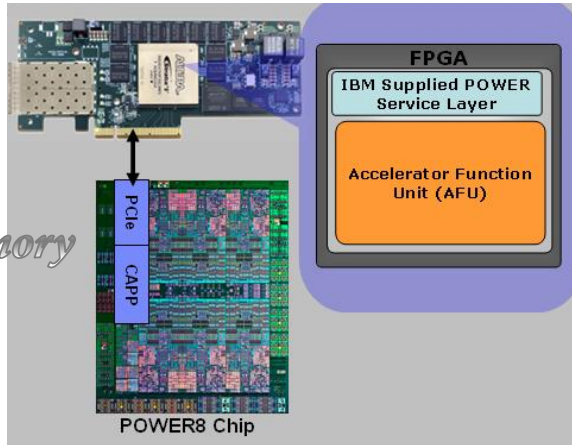


Source: I. Kuon, R. Tessier, J. Rose. FPGA Architecture: Survey and Challenges. 2008.

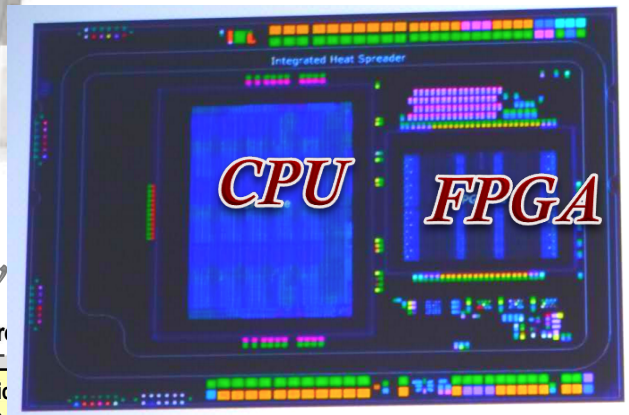
Modern CPU-FPGA Platforms

Alpha Data, 2014
PCIe-based, Separate Memory (Mainstream)

CAPI, 2015
PCIe-based, Shared Memory

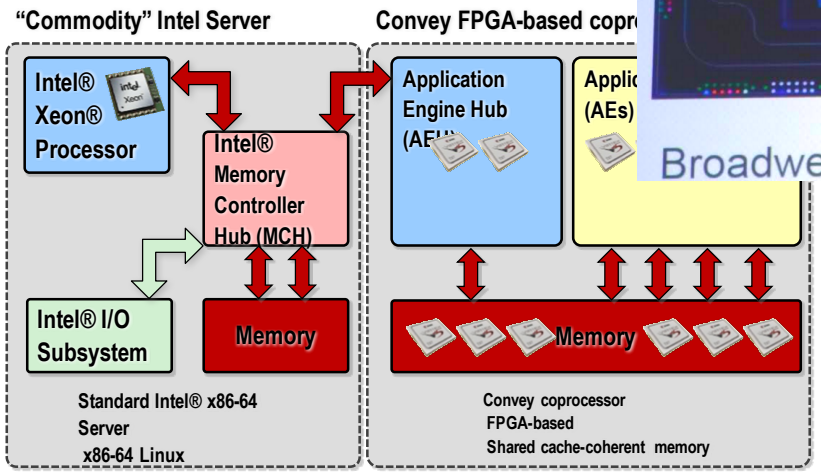


HARP 2, 2016
Multi-Chip Package, Shared Memory

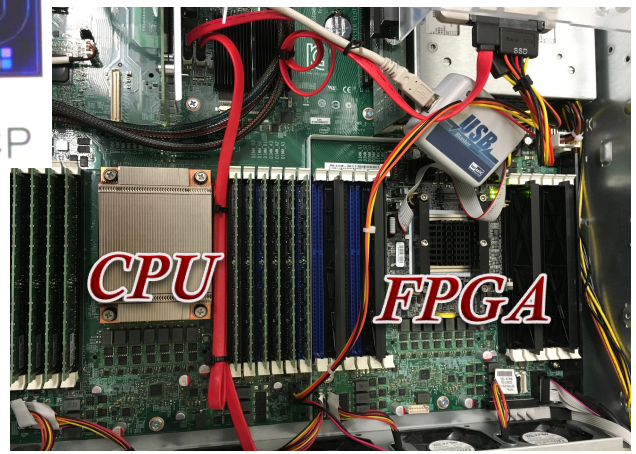


Broadwell + Arria 10 GX MCP

Convey, 2010
FSB-based, Shared Memory



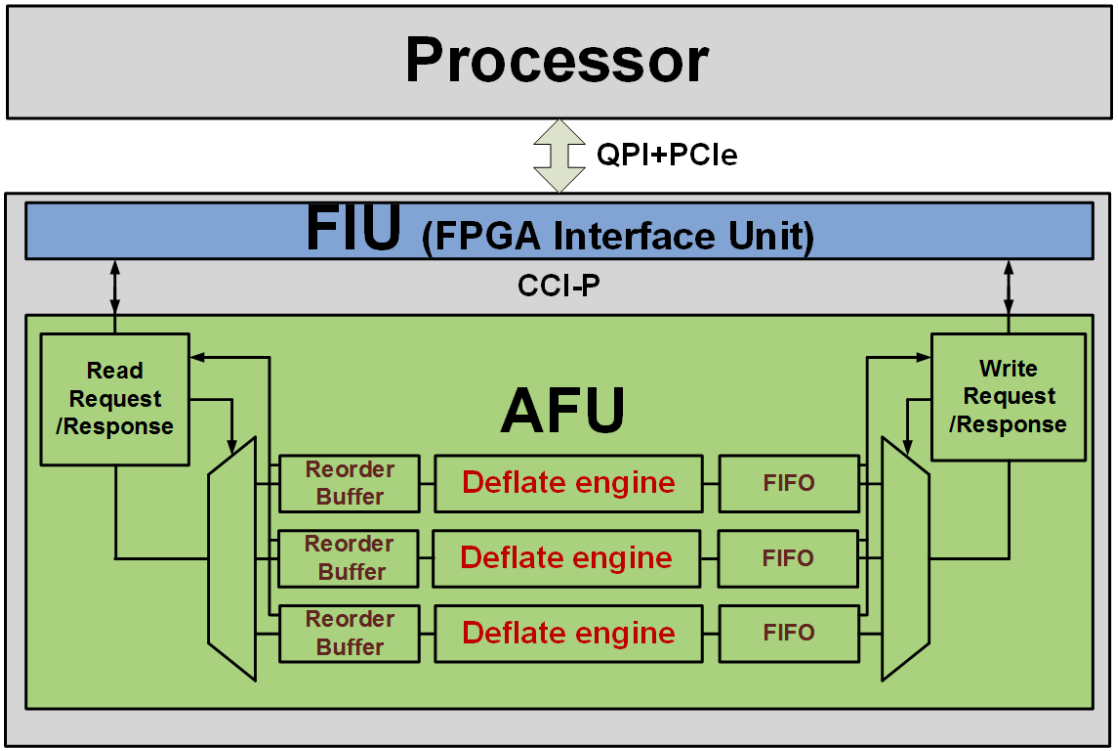
HARP, 2015
QPI-based, Shared Memory



Example: Acceleration of Lossless Data Compression on HARP-2

- **Scalable FPGA-based parallel architecture**
  - Multi-engine Deflate compressor which can be easily scaled
  - Fully pipelined in each engine
  - Valuable in multi-thread environment applications

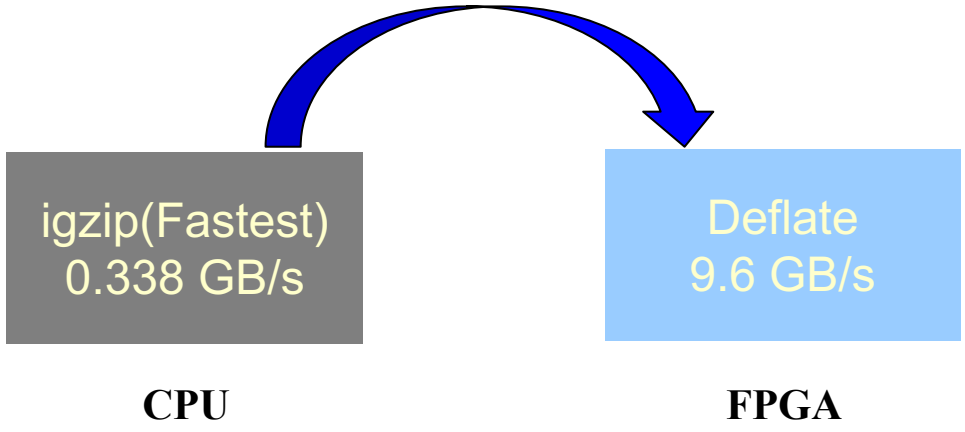
Integration on HARPv2



Acceleration of Lossless Data Compression

- **Throughput**
  - Kernel throughput: 9.6 GB/s @ 200 MHz
  - End-to-End throughput: >9 GB/s
  - Best published result

~ 28x speedup !!!



Kernel Throughput
9.6 GB/s
End-to-End Throughput on HARP
3.9 GB/s
End-to-End Throughput on HARPv2
9.3 GB/s

- **Compression ratio**
  - Average 1.95x on Calgary Corpus benchmarks

Levels of Customization

- **Single-chip level**
  - Require new processor designs, e.g. using fixed-function or composable accelerators
- **Server node level**
  - Host CPU + FPGA via PCI-e or QPI connections
- **Data center level**
  - Clusters of heterogeneous computing nodes

Data-Center Level Customization: Example: CDSC FPGA-Accelerated Cluster

Scale-up: FPGA acceleration inside each node

- A 24-node cluster with FPGA-based accelerators

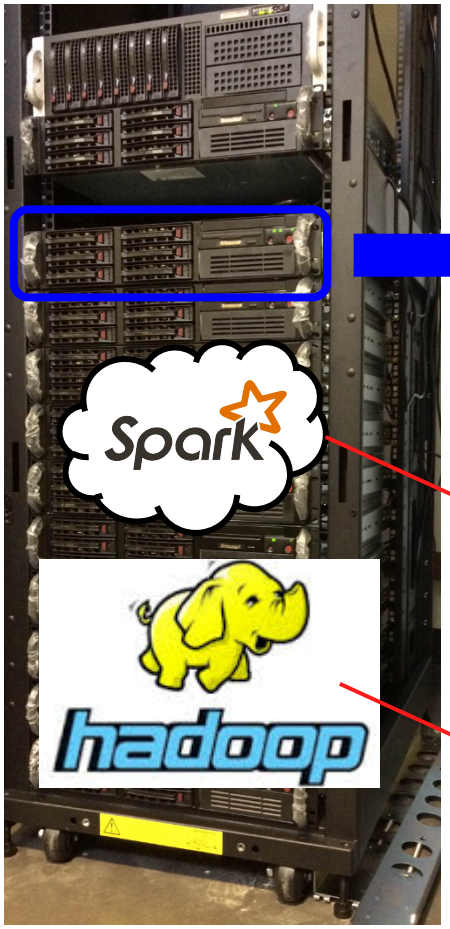
Scale-out: an in-memory cluster

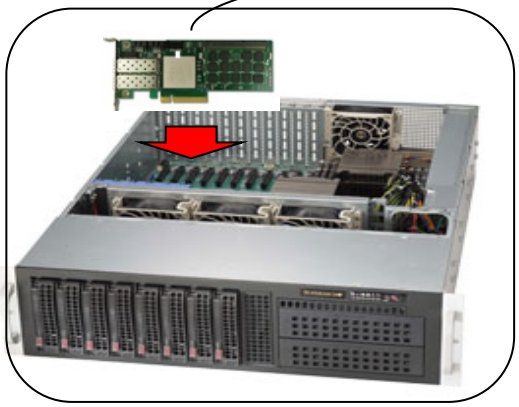
1 master / driver

1 10GbE switch

22 workers

1 file server





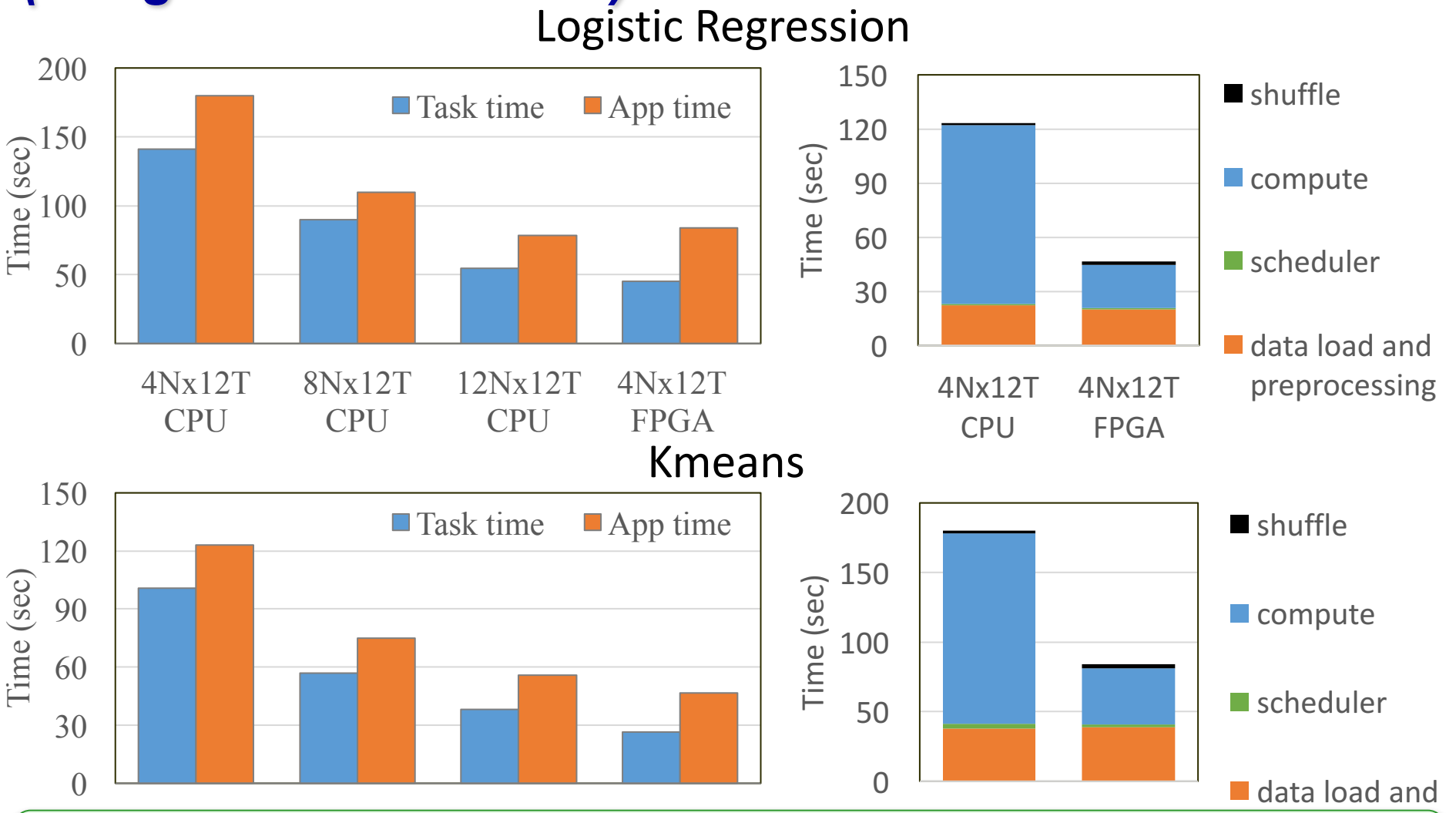
Alpha Data board:
1. Virtex 7 FPGA
2. 16GB on-board RAM

Each node:
1. Two Xeon processors
2. One FPGA PCIe card (Alpha Data)
3. 64 GB RAM
4. 10GBE NIC

Spark:
- Computation framework
- In-memory MapReduce system

HDFS:
- Distributed storage framework

Overall Performance with Accelerators (Integrated with Blaze)



1 server with FPGA offers the same throughput of 3 servers

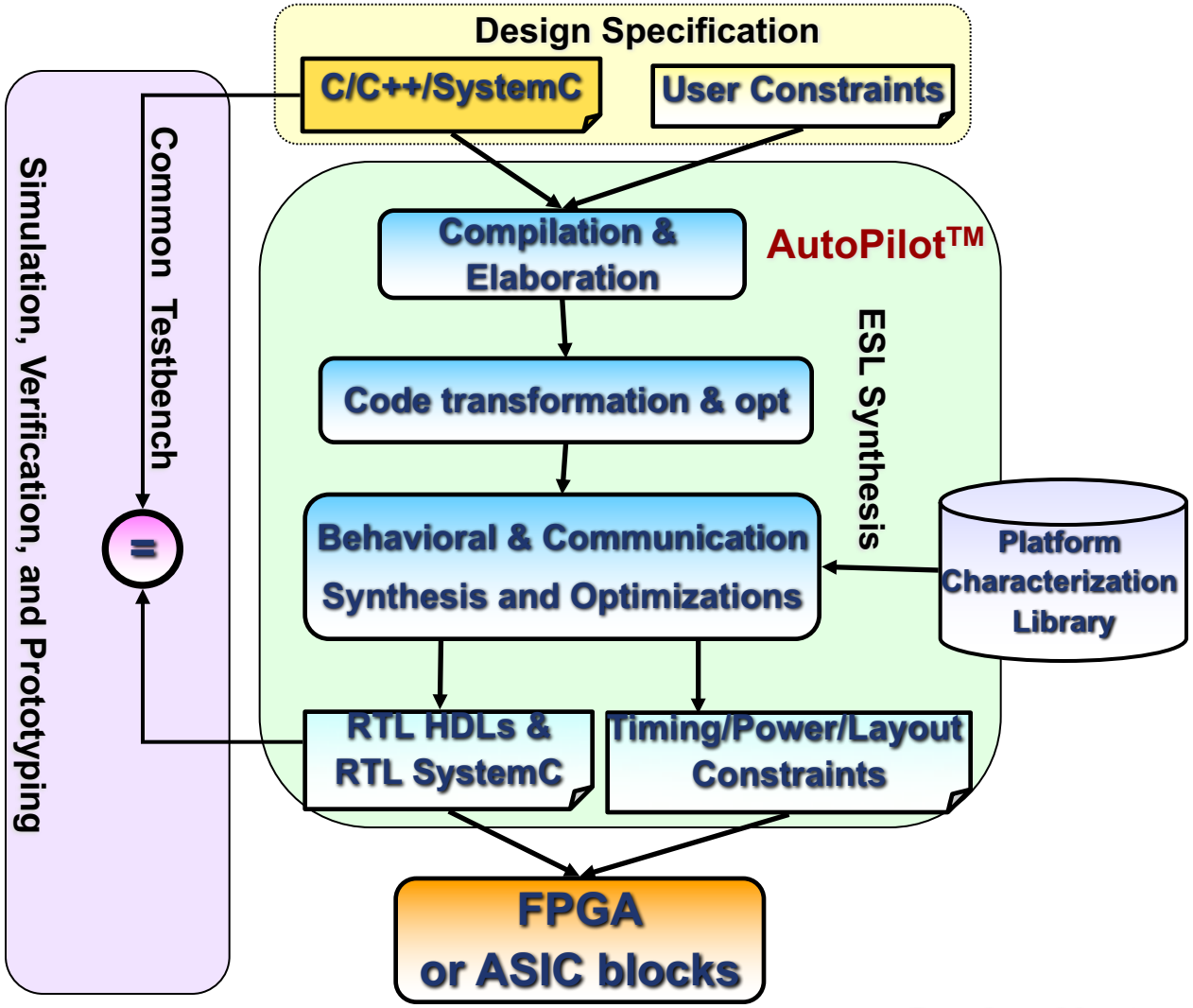
FPGA-Based Customized Computing is Taking Off

- FPGA is gaining popularity as a compute device
  - Used by many industry giants
  - First public cloud adoption (AWS F1) in Feb. 2017
  - Intel prediction: 30% datacenter nodes with FPGA by 2020

HOW TO DESIGN AND DEPLOY ACCELERATORS

C/C++ Based Synthesis for Accelerator Design

xPilot (UCLA 2006) -> AutoPilot (AutoESL) -> Vivado HLS (Xilinx 2011-)

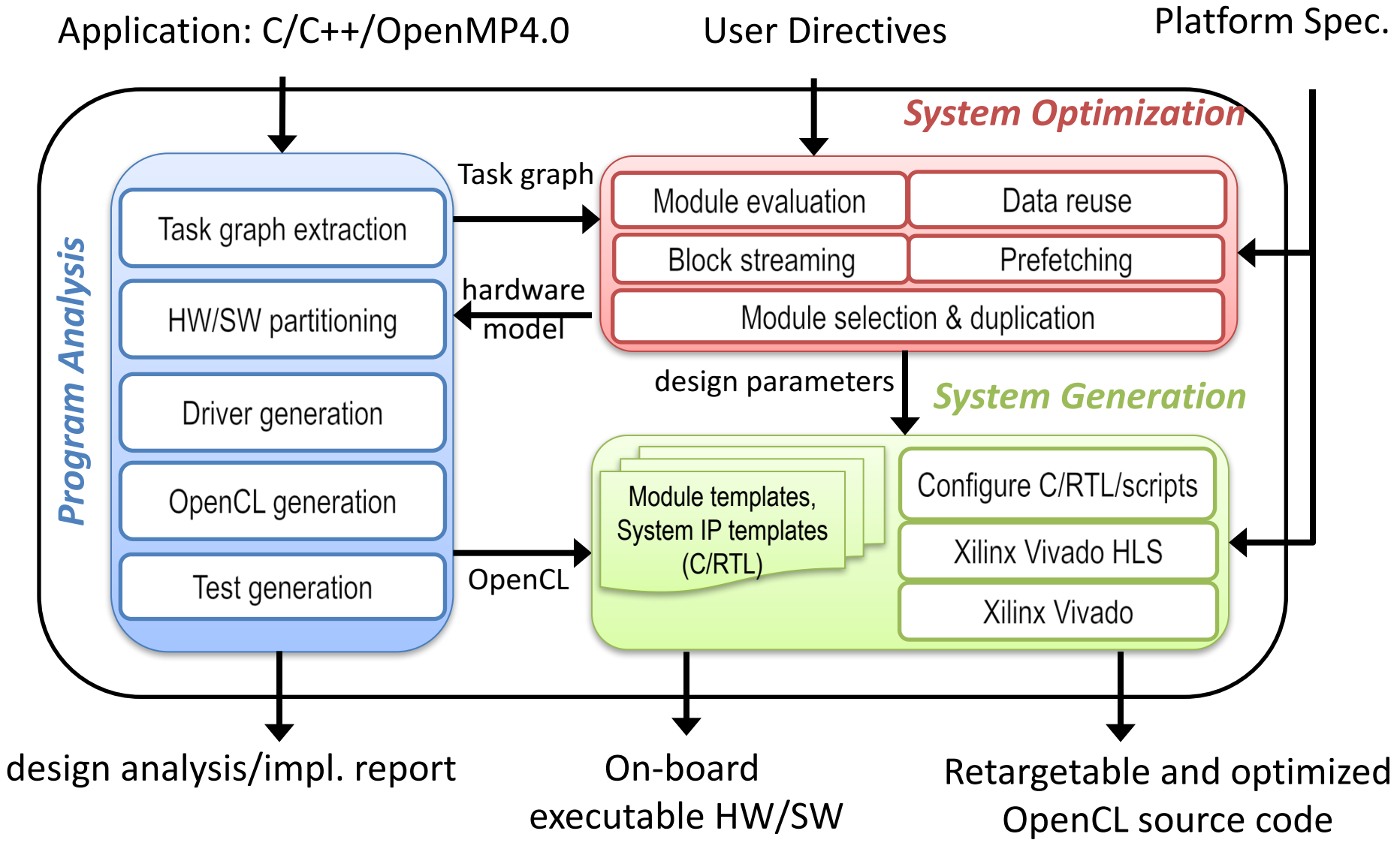


- Platform-based C to RTL synthesis
- Synthesize pure ANSI-C and C++, GCC-compatible compilation flow
- Full support of IEEE-754 floating point data types & operations
- Efficiently handle bit-accurate fixed-point arithmetic
- SDC-based scheduling
- Automatic memory partitioning
- …

QoR matches or exceeds manual RTL for many designs

Developed by AutoESL, acquired by Xilinx in Jan. 2011

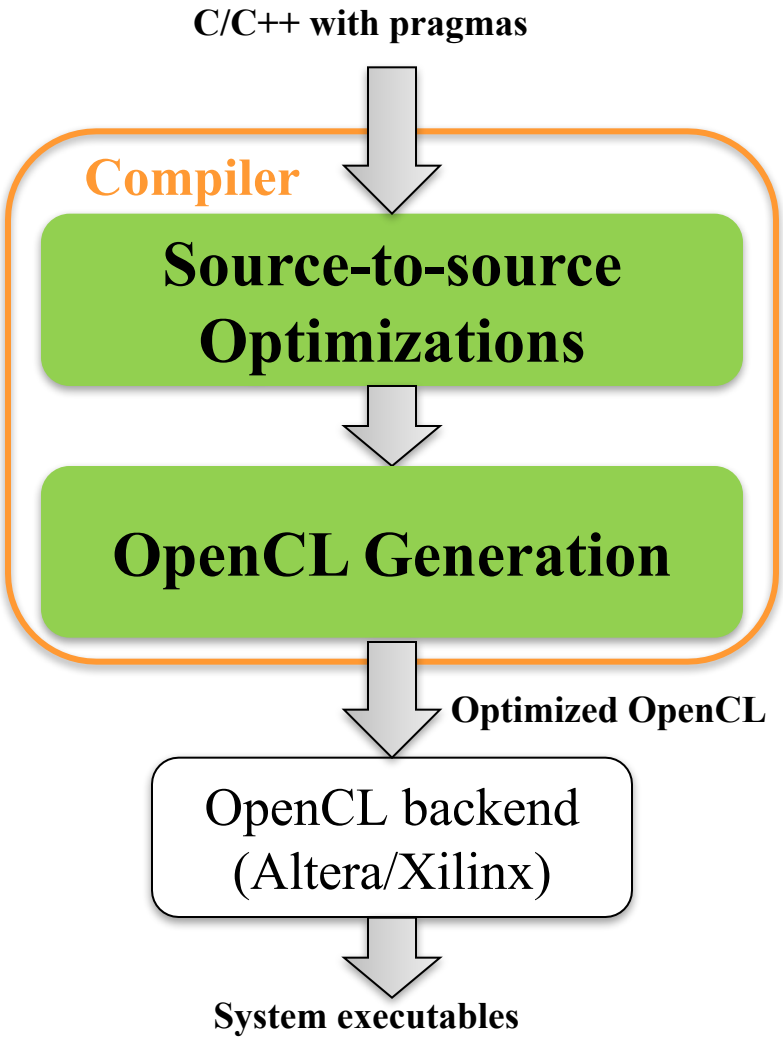
CMOST: Fully Automated Compilation and Mapping Flow [DAC 2015]



Further Advance in Programming FPGAs in High-Level Languages

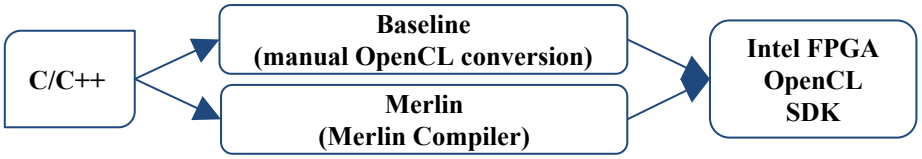
Merlin Compiler from Falcon Computing Solutions: http://www.falcon-computing.com

- ✓ C-based design flow
- ✓ OpenMP-like high-level programming model
- ✓ Automatic optimizations for productivity and QoR
- ✓ Same input for multi-vendors and multi-platforms



Merlin 2017.2 Preliminary Results

FPGA platform: Intel Arria 10 DevKit
Software configuration: AOCL 16.1



Case	Baseline (quick Opt)	Merlin	Manual Opt OCL
aes	7600	2.4	1
gemm	0.2	0.2	0.2
viterbi	14.5	0.16	0.2
NW	180	5.3	3.9
bfs-queue	9.7	0.4	NA
kmp	6700	920	111
spmv-el	10	1.5	1.5

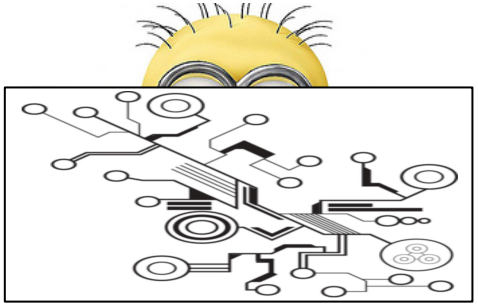
Execution time: ms

Merlin speedup over Baseline: 32.6x (1x – 3167x), excluding 'aes' and 'gemm'

What about Accelerator Deployment?

Application developer

How to program with your accelerators…?



Accelerator designer

How to acquire accelerator resource …?

Cloud or edge service provider

How to install my accelerators…?

Challenges in the Accelerator Deployment

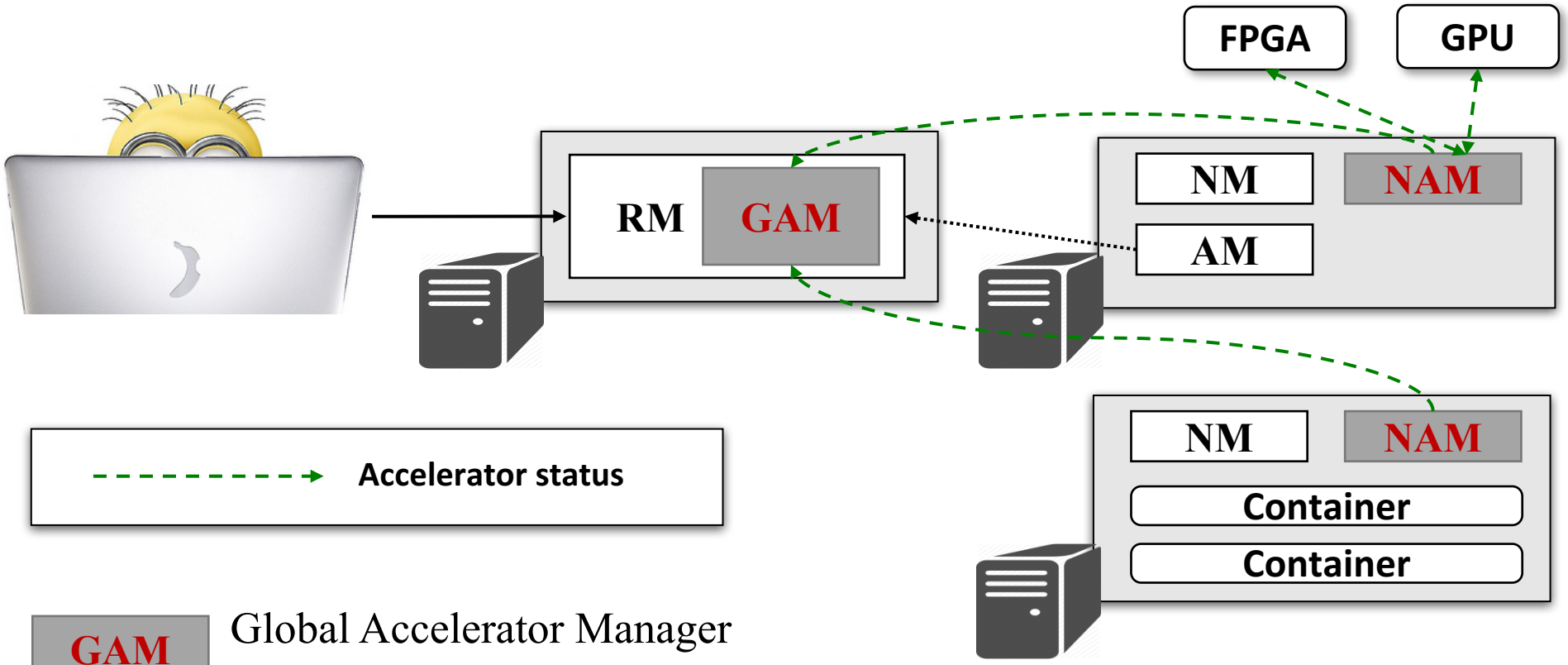
Complex programming

- **High-level language (*C/C++/Java*) for applications (*Spark, AR*) vs. low-level language (*OpenCL*) & HW expertise for accelerators (*FPGA*)**
- **Explicit accelerator sharing by multiple threads and applications**
- **[HotCloud'16] Manual integration of Spark + FPGA: ~900 lines of code with HW expertise, and has to repeat for every integration**

Runtime performance overhead

- **#1: Large JVM/host-to-accelerator data transfer overhead, [HotCloud'16] 1000x slowdown for straightforward integration**
- **#2: Long FPGA (partial) reconfiguration overhead (0.5 - 2 seconds), Naïve FPGA sharing by multi-accelerators may lead to 2x slowdown**

Blaze: Accelerator-as-a-Service [SoCC 2016]



GAM Global Accelerator Manager

Accelerator locality aware scheduling

NAM Node Accelerator Manager

Local accelerator service management,
JVM-to-ACC communication optimization

RM: Resource Manager
NM: Node Manager
AM: Application Master

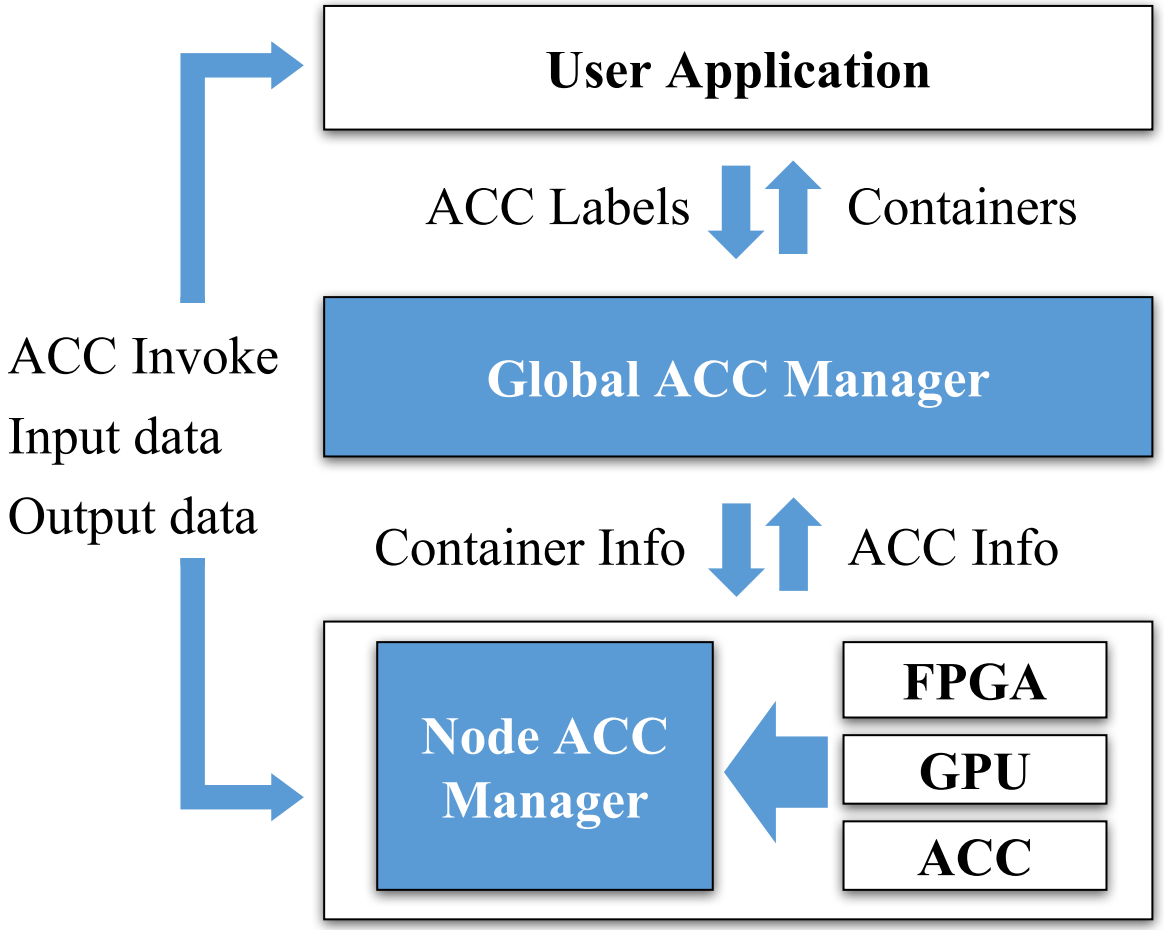
Blaze Deployment Flow Overview

Register Accelerators

- Interface to add accelerator service to corresponding nodes

Request Accelerators

- Use acc_id as label
- GAM allocates corresponding nodes to applications

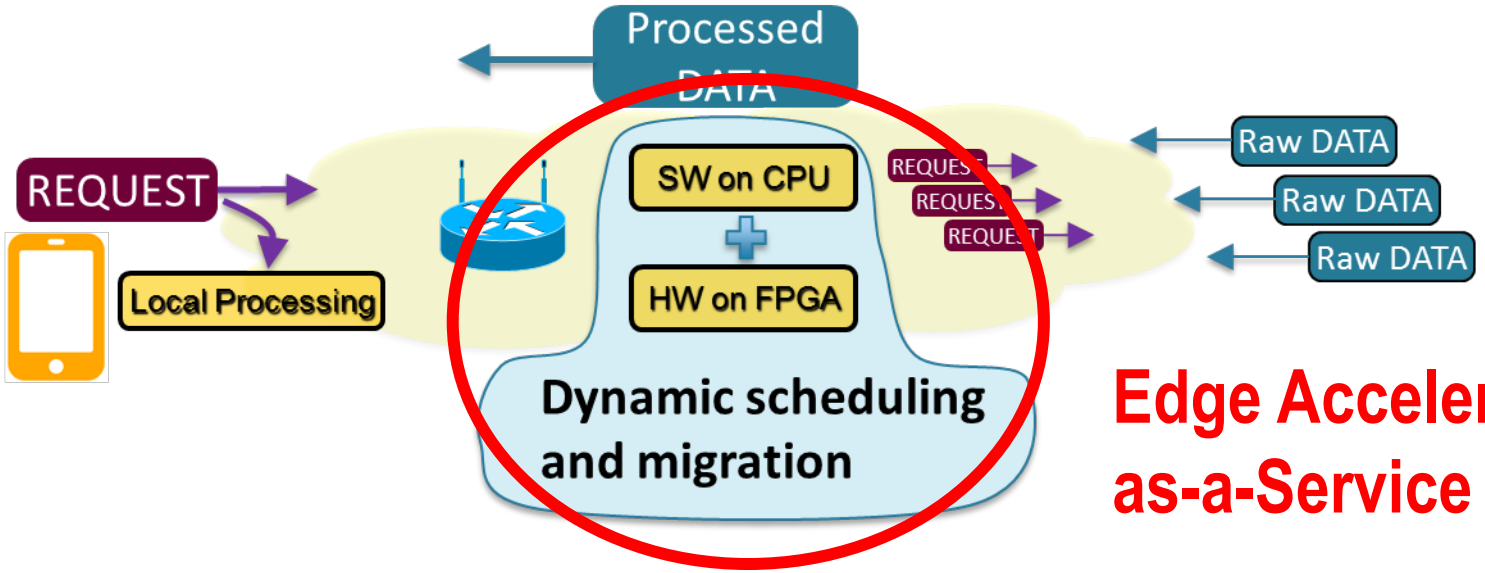


New Research Theme -- Acceleration in Fog

- **Single-chip level**
  - Require new processor designs, e.g. using fixed-function or composable accelerators
- **Server node level**
  - Host CPU + FPGA via PCI-e or QPI connections
- **Data center level**
  - Clusters of heterogeneous computing nodes
- **Fog-level**
  - Acceleration at the edge of wireless network
  - Acceleration as a service (AaaS)

Recap: The Need for In-network Acceleration

- In-network and en-route aggregation
  - IoT streams are processed once generated
  - Deploy and customize NFs for IoT processing
- Location-based aggregation
  - Location as the first landmark for streamlining
- On-demand migration btw compute & comm.



Edge Acceleration-as-a-Service (AaaS)

Research Opportunities of AaaS in NDN

- **NDN for acceleration**
  - Acceleration function F(x)
  - F = bitstream is data: NDN helps
  - x is data: NDN helps to minimize the redundant computation
- **Acceleration for NDN**
  - Name checking – hashing
  - Compression/decompression
  - Encryption/decryption
- **Enable new 5G applications: e.g. AR/VR**
  - CPU is not sufficient to meet latency requirment