
Accelerator-Rich Architectures — From Single-chip to Datacenters

Jason Cong

Chancellor's Professor, UCLA

Director, Center for Domain-Specific Computing

cong@cs.ucla.edu

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

1

Why Customized Computing

2

Why Accelerators?

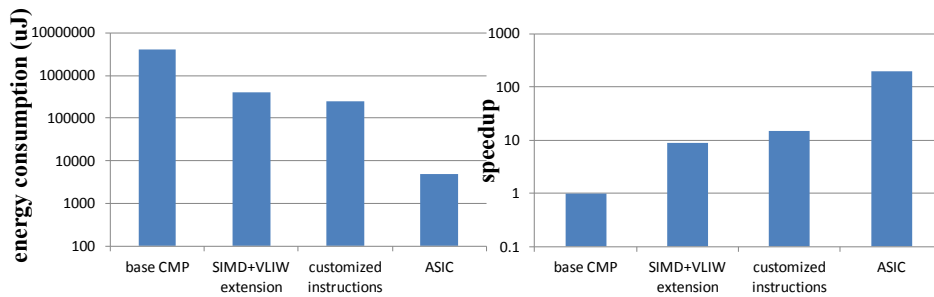
AES 128bit key 128bit data	Throughput	Power	Figure of Merit (Gb/s/W)
0.18mm CMOS	3.84 Gbits/sec	350 mW	11 (1/1)
FPGA [1]	1.32 Gbit/sec	490 mW	2.7 (1/4)
ASM StrongARM [2]	31 Mbit/sec	240 mW	0.13 (1/85)
ASM Pentium III [3]	648 Mbits/sec	41.4 W	0.015 (1/800)
C Emb. Sparc [4]	133 Kbits/sec	120 mW	0.0011 (1/10,000)
Java [5] Emb. Sparc	450 bits/sec	120 mW	0.0000037 (1/3,000,000)

[1] Amphion CS5230 on Virtex2 + Xilinx Virtex2 Power Estimator
 [2] Dag Arne Osvik: 544 cycles AES – ECB on StrongArm SA-1110
 [3] Helger Lipmaa PIII assembly handcoded + Intel Pentium III (1.13 GHz) Datasheet
 [4] gcc, 1 mW/MHz @ 120 Mhz Sparc – assumes 0.25 u CMOS
 [5] Java on KVM (Sun J2ME, non-JIT) on 1 mW/MHz @ 120 MHz Sparc – assumes 0.25 u CMOS

Source: P Schaumont and I Verbauwhede, "Domain specific codesign for embedded security," IEEE Computer 36(4), 2003

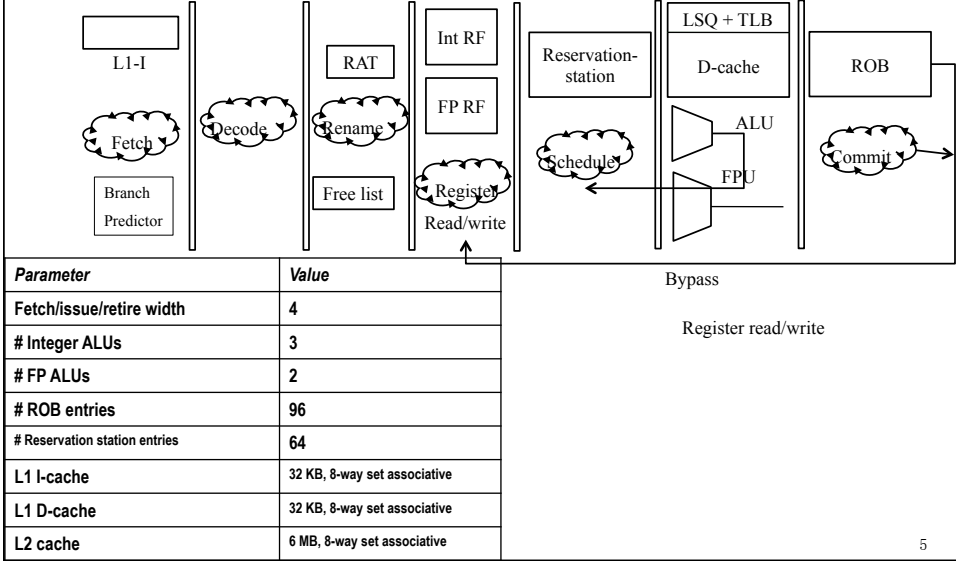
Why Accelerators?

- ◆ Case study of H.264 [Hameed et al., ISCA'2010]
 - Optimization using SIMD + VLIW → 10x energy efficiency
 - Customized instruction fusion → 1.6x energy efficiency
- ◆ Still 50x away from ASICs – Not enough!

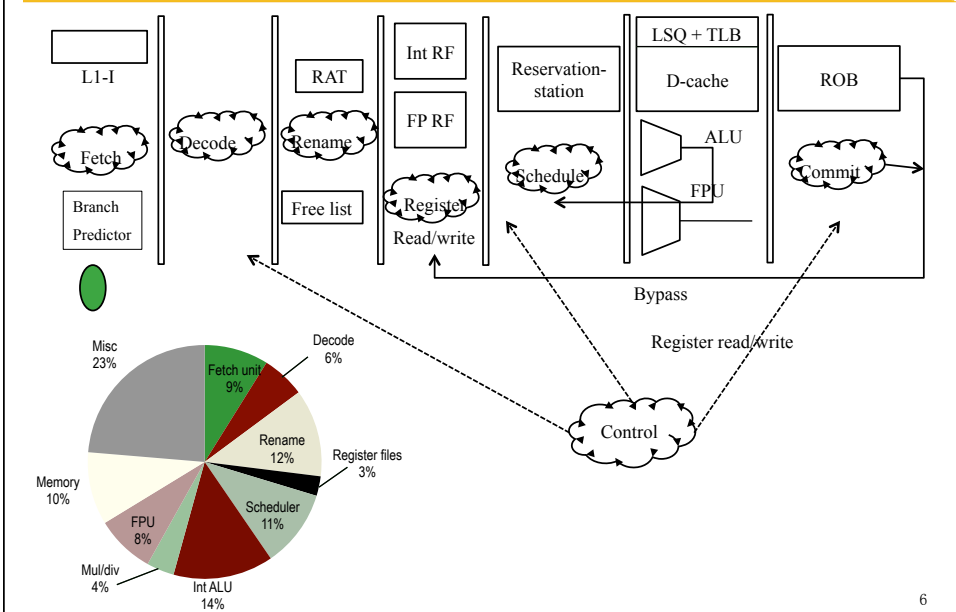


Understanding Energy Inefficiency of Processors [DAC'2014]

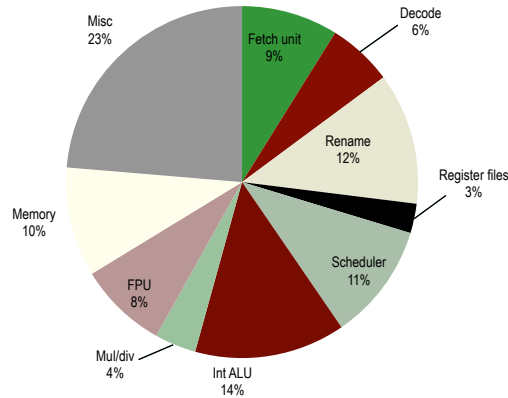
Typical Superscalar OoO Pipeline



Energy Breakdown of Pipeline Components



Removing 'Non-Computing' Portions of the Pipeline



■ Remaining

■ ~ 10% + 26% = 36%

7

Energy Comparison of Processor ALUs and Dedicated Units

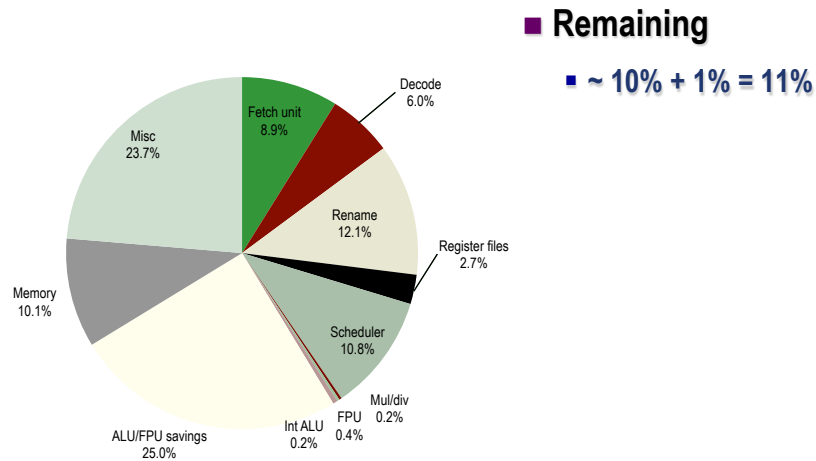
Operation	Processor ALU	45 nm TSMC library
32-bit add	0.122 nJ@2 GHz	0.002 nJ @ 1 GHz
32-bit multiply	0.120 nJ@2 GHz	0.007 nJ @ 1 GHz
Single precision FP operation	0.150 nJ @ 2GHz	0.008 nJ @ 500 MHz

■ Why are processor units so expensive?

- ALU can perform multiple operations
 - Add/sub/bitwise XOR/OR/AND
- 64-bit ALU
- Dynamic/domino logic used to run at high frequency
 - Higher power dissipation

8

Energy for Custom ASICs



9

Additional Energy Efficiency from Custom ASICs

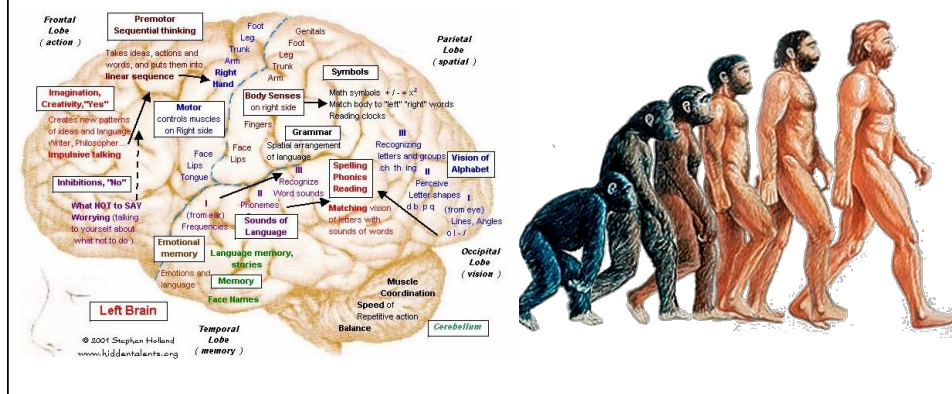
- ◆ Bitwidth customization
- ◆ More efficient memory operations
 - Energy/mem-op could be less than processor
 - Custom memory architecture
 - Better utilization of on-chip data
- ◆ Predictable communication patterns
- ◆ All these lead to 10-100X energy efficiency over the processors
- ◆ Problem: too costly and too much time to build an ASIC

So, What Shall We Do with Processors? Our Proposal – Accelerator-Rich Architectures

- ◆ A customizable heterogeneous platform (CHP)
 - With a sea of dedicated and composable accelerators
 - Most computations are carried on accelerators – not on processors!
- ◆ A fundamental departure from von Neumann architecture
- ◆ Why now?
 - Previous architectures are device/transistor limited
 - Von Neumann architecture allows maximum device reuse
 - One pipeline serves all functions, fully utilized
- ◆ Future architectures
 - Plenty of transistors, but power/energy limited (dark silicon)
 - Customization and specialization for maximum energy efficiency
- ◆ A story of specialization

Lessons from Nature: Human Brain and Advance of Civilization

- ◆ High power efficiency (20W) of human brain comes from specialization
 - Different region responsible for different functions
- ◆ Remarkable advancement of civilization also from specialization
 - More advanced societies have higher degree of specialization



UCLA Newsroom

Home [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 awarded a \$10 million grant by the National Science Foundation's Expeditions in Computing program to develop high-performance, energy efficient, customizable computing that could revolutionize the way computers are used in health care and other important applications.

In particular, UCLA Engineering researchers will demonstrate how the new technology, known as domain-specific computing, could transform the role of medical imaging and hemodynamic simulation, providing more cost-effective and convenient solutions for preventive, diagnostic and therapeutic procedures and dramatically improving health care quality, efficiency and patient outcomes.

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

In an effort to meet ever-increasing computing needs in various fields, the computing industry has entered an "era of parallelization," in which tens of thousands of computer servers are connected in warehouse-scale data centers, said Jason Cong, the Chancellor's Professor of Computer Science and director of the new UCLA Center for Domain-Specific Computing (CDSC), which will oversee the research. But these parallel, general-purpose computing systems still face serious challenges in terms of performance, energy, space and cost.

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

The goal of the new UCLA center, Cong said, is to look beyond parallelization and focus on domain-specific customization to bring significant power-performance efficiency improvement to important application domains.

All Stories
 All Stories
 Featured News
 News Releases
 Advisories
 Images
 Multimedia

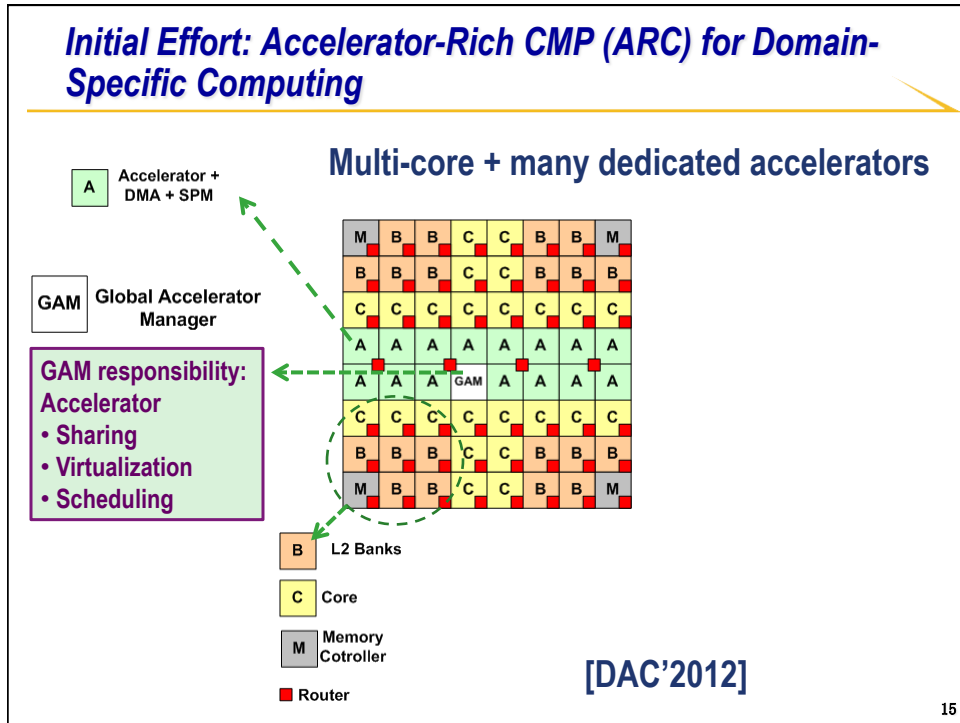
Research
 Health Sciences
 Arts & Humanities
 Student Affairs
 Academics & Faculty
 Campus News
 Media Contacts

Images
Video
Blogs

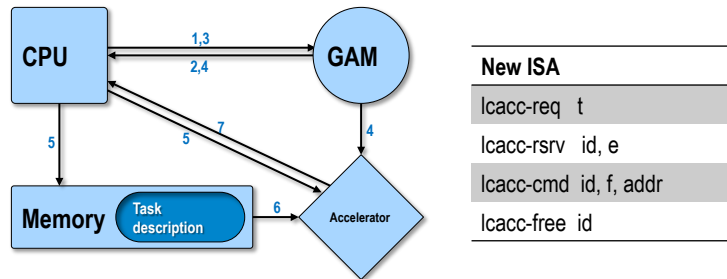
For the Media
 Contacts
 News releases
 Advisories
 About UCLA

Accelerator-Rich Architectures
-- Single-Chip Level Designs

Initial Effort: Accelerator-Rich CMP (ARC) for Domain-Specific Computing



ISA Extension and Overall Workflow in ARC



1. The core requests for a given type of accelerator (lcacc-req).
2. The GAM responds with a "list + waiting time" or NACK
3. The core reserves (lcacc-rsv) and waits.
4. The GAM ACK the reservation and send the core ID to accelerator
5. The core shares a task description with the accelerator through memory and starts it (lcacc-cmd).
6. The accelerator reads the task description, and begins working
7. When the accelerator finishes its current task it notifies the core.
8. The core then sends a message to the GAM freeing the accelerator (lcacc-free).

Medical Image Processing Pipeline

reconstruction

Medical images exhibit sparsity, and can be sampled at a rate \ll classical Shannon - Nyquist theory :

$$\min_u \sum_{\text{sampled points}} \|ARu - S\|^2 + \lambda \sum_{\forall \text{voxels}} \|grad(u)\|$$

compressive sensing

denoising

$$\forall \text{voxel} : u(i) = \sqrt{\left(\sum_{\text{voxel} \in \text{volume}} w_{ij} f(j)^2 \right) - 2\sigma^2}, w_{ij} = \frac{1}{Z(i)} e^{-\frac{\sqrt{\sum_{k=1}^5 |v_k - z_k|^2}}{h}}$$

total variational algorithm

registration

$$v = \frac{\partial u}{\partial t} + v \cdot \nabla u$$

$$\mu \Delta v + (\mu + \eta) \nabla(\nabla \cdot v) = -[T(x-u) - R(x)] \nabla T(x-u)$$

fluid registration

segmentation

$$\frac{\partial \phi}{\partial t} = |\nabla \phi| \left[F(\text{data}, \phi) + \lambda \text{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right]$$

$$\text{surface}(t) = \{ \text{voxels } x : \phi(x,t) = 0 \}$$

level set methods

analysis

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v = -\nabla p + \nu \Delta v + f(x,t)$$

$$\frac{\partial v_i}{\partial t} + \sum_{j=1}^3 v_j \frac{\partial v_i}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \nu \sum_{j=1}^3 v_j \frac{\partial^2 v_i}{\partial x_j^2} + f_i(x,t)$$

Navier-Stokes equations

17

ARC Evaluation on Medical Imaging Benchmarks

		GPU * (NVIDIA Tesla M2075)	FPGA (Xilinx V6)	Monolithic Accelerators
Deblur	Performance	2.4X	3.4X	7.8X
	Energy	0.3X	3.2X	32X
Denoise	Performance	16.6X	1.6X	3.5X
	Energy	1.4X	1.2X	13X
Segmentation	Performance	73X	16X	16X
	Energy	6.1X	3.6X	53X
Registration	Performance	3.9X	6.7X	15X
	Energy	0.4X	3.2X	60X
Average	Performance	24X	6.9X	10X
	Energy	2X	2.8X	39.8X

* NOTE: GPU power values were full-system measurements obtained using the Kill-A-Watt device, making them relatively inflated compared to other McPAT-generated values
 Results relative to Quad Core Intel Xeon (E5405 @ 2 GHz)
Accelerators are synthesized in 32nm technology

- But workload coverage is narrow
- Dedicated accelerators contain replicated structures, such DMA engines and SPMs

18

Problems with ARC

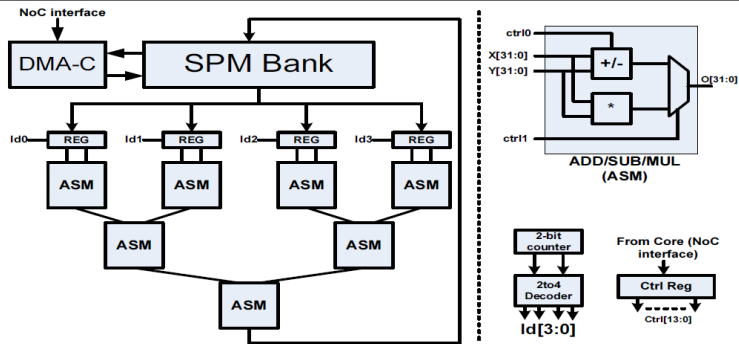
- ◆ **Dedicated accelerators are inflexible**
 - An LCA may be useless for new algorithms or new domains
 - Often under-utilized
 - LCAs contain many replicated structures
 - Things like fp-ALUs, DMA engines, SPM
 - Unused when the accelerator is unused
- ◆ **We want flexibility and better resource utilization**
 - Solution: use composable accelerators

19

Possibility of Accelerator Composition – Use of Accelerator Building Blocks (ABBs)

ABBs	Denoise	Deblur	Registration	Segmentation
Float Reciprocal (FInv)	✓	✓		✓
Float Square-Root (FSqrt)	✓	✓	✓	✓
Float Polynomial-16 (Poly16)	✓	✓	✓	✓
Float Divide (FDiv)	✓	✓	✓	✓

Example:
Poly-8 ABB

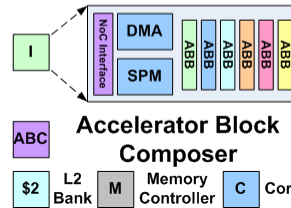


20

Second Effort: Accelerator Composition [ISLPED'12]

- ◆ **ABB**
 - Accelerator building blocks (ABB)
 - Primitive components that can be composed into accelerators
- ◆ **ABB islands**
 - Multiple ABBs
 - Shared DMA controller, SPM and NoC interface
- ◆ **ABC**
 - Accelerator Block Composer (ABC)
 - Runtime composition of virtual accelerators from ABBs
 - Arbitrate requests from cores
- ◆ **Other components**
 - Cores
 - L2 Banks
 - Memory controllers
- ◆ **Accelerator composition: Static mapping + dynamic allocation**

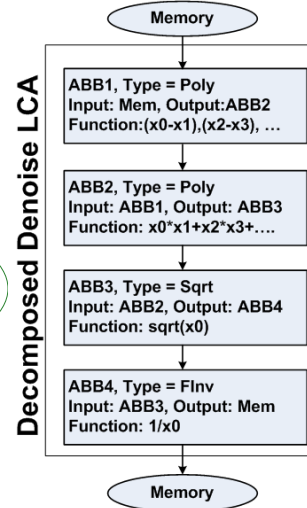
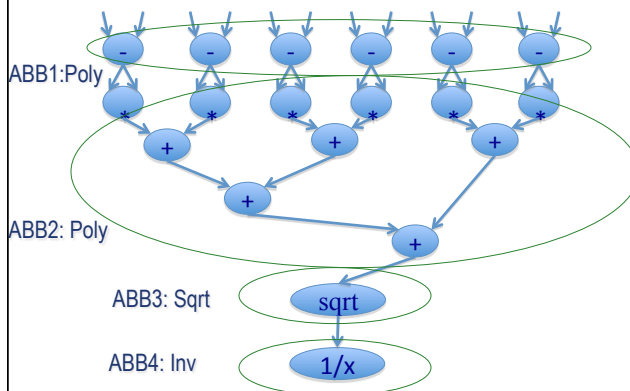
M	\$2	\$2	C	C	\$2	\$2	M
C	C	C	C	C	C	C	C
\$2	I	\$2	I	\$2	I	\$2	I
\$2	I	\$2	I	\$2	I	\$2	I
\$2	I	\$2	ABC	\$2	I	\$2	I
\$2	I	\$2	I	\$2	I	\$2	I
C	C	C	C	C	C	C	C
M	\$2	\$2	C	C	\$2	\$2	M



21

Static Mapping/Decomposition into ABBs

$$1 / \sqrt{\sum_{i=0}^6 (Xi - Y)^2}$$

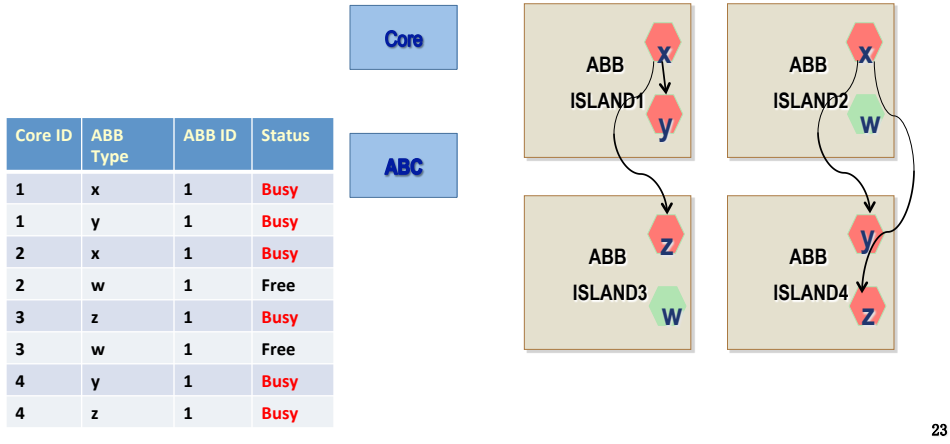


22

Accelerator Composition Process at Runtime

Accelerator cloning

- Repeat to generate more LCAs if ABBs are available



23

Composable Accelerator Results

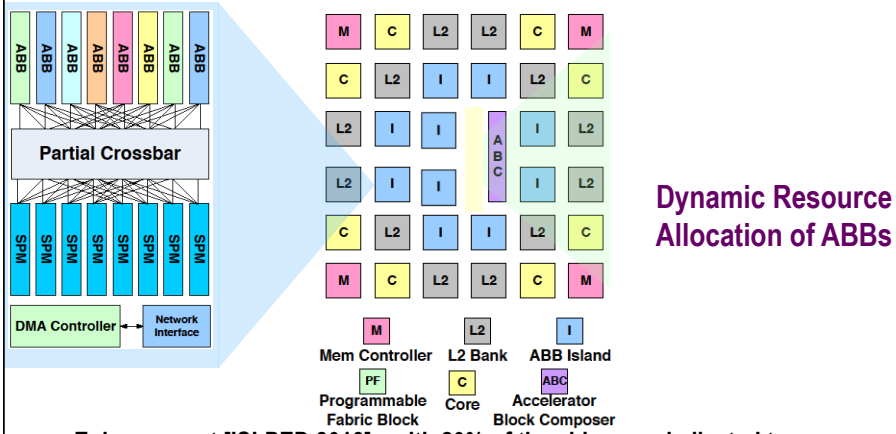
		GPU * (NVIDIA Tesla M2075)	FPGA (Xilinx V6)	Monolithic Accelerators	Composable Accelerators
Deblur	Performance	2.4X	3.4X	7.8X	21X
	Energy	0.3X	3.2X	32X	55X
Denoise	Performance	16.6X	1.6X	3.5X	11X
	Energy	1.4X	1.2X	13X	29X
Segmentation	Performance	73X	16X	16X	77X
	Energy	6.1X	3.6X	53X	186X
Registration	Performance	3.9X	6.7X	15X	58X
	Energy	0.4X	3.2X	60X	144X
Average	Performance	24X	6.9X	10X	42X
	Energy	2X	2.8X	39.8X	103X

* NOTE: GPU power values were full-system measurements obtained using the Kill-A-Watt device, making them relatively inflated compared to other McPAT-generated values

Results relative to Quad Core Intel Xeon (E5405 @ 2 GHz)
Accelerators are synthesized in 32nm technology

24

Latest Effort: Composable Accelerators with Programmable Fabrics [ISLPED'2013]



- ◆ **Enhancement [ISLPED 2013]:** with 20% of the chip area dedicated to programmable fabric, we can achieve more:
 - **Flexibility:** An average 8.2x (up to 146x) speedup in other domains, such as commercial, vision and navigation
 - **Longevity:** 22x speedup on a new application within the medical imaging domain

25

New Research Opportunities for Architecture-Rich Architecture

- ◆ **Memory support**
- ◆ **Communication support**
- ◆ **Prototyping and validation**
- ◆ **Software support**

26

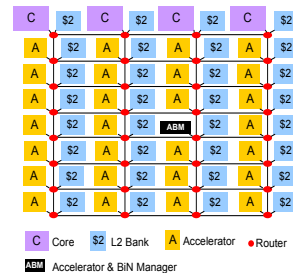
Buffer Management for Accelerator-Rich Architectures

- ◆ Accelerator uses buffers or scratchpad-memory (SPM)
 - With much higher throughput than processors
- ◆ Providing a private buffer for each accelerator is very inefficient.
 - Large private buffers: occupy a considerable amount of chip area
 - Small private buffers: less effective for reducing off-chip bandwidth
- ◆ Not all accelerators are powered-on at the same time
- ◆ Opportunities: efficient schemes for buffer sharing among cores and accelerators
 - Accelerator store (AS): shared buffer among accelerators [Lyonsy et al. TACO'12]
 - Buffer in Cache (BiC)[Fajardo et al. DAC'11]
 - Adaptive hybrid cache: AH-cache [ISLPED'11]
 - Share buffers with banked L2 cache: BiN -- A Buffer-in-NUCA Scheme for Accelerator-Rich CMPs [ISLPED'12]

27

BiN: Buffer-in-NUCA [ISLPED'2012]

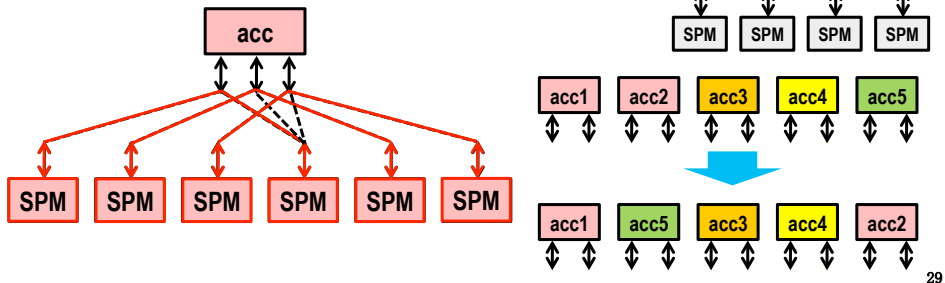
- ◆ **ABM: a centralized resource manager**
 - Accelerator allocation/management +
 - Buffer allocation/management
 - Towards optimal on-chip storage utilization
- ◆ **Contributions of BiN:**
 - Dynamic interval-based global buffer allocation:
 - Deal with the buffer resource contention
 - Flexible paged buffer allocation:
 - Avoid buffer resource fragmentation
- ◆ **Results**
 - 32% and 35% performance improvement over AS and BiC
 - 12% and 29% energy improvement over AS and BiC



28

Needs of Optimized Interconnects between Accelerators and SPMs

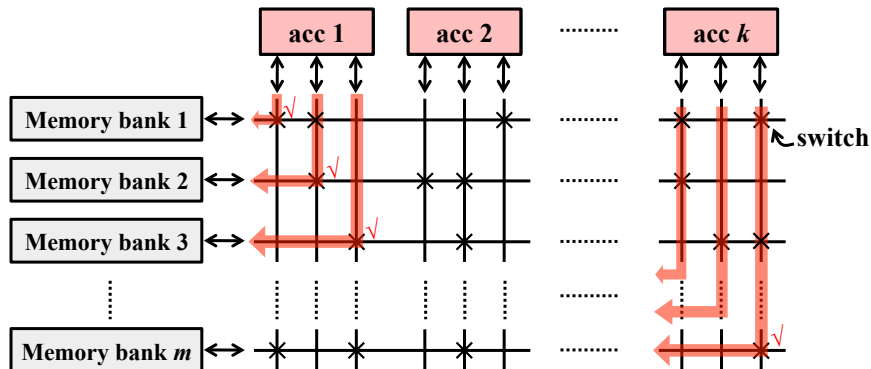
- ◆ **Conventional interconnects**
 - either insufficient bandwidth (bus or NoC) or resource-consuming (full crossbar)
- ◆ **Opportunity for optimization**
 - Concurrent memory accesses of a single accelerator
 - Sparsity of accelerator execution
 - Application specific patterns of accelerator execution



29

Partial Crossbar Design for High Routability [ICCAD'2013]

- ◆ **Definition of routability**
 - Given a randomly selected workload of c accelerators, the probability that their ports can be routed to different memory banks via separate data paths
- ◆ **Optimization goal**
 - Place as few switches in the partial crossbar as possible while keeping high routability



30

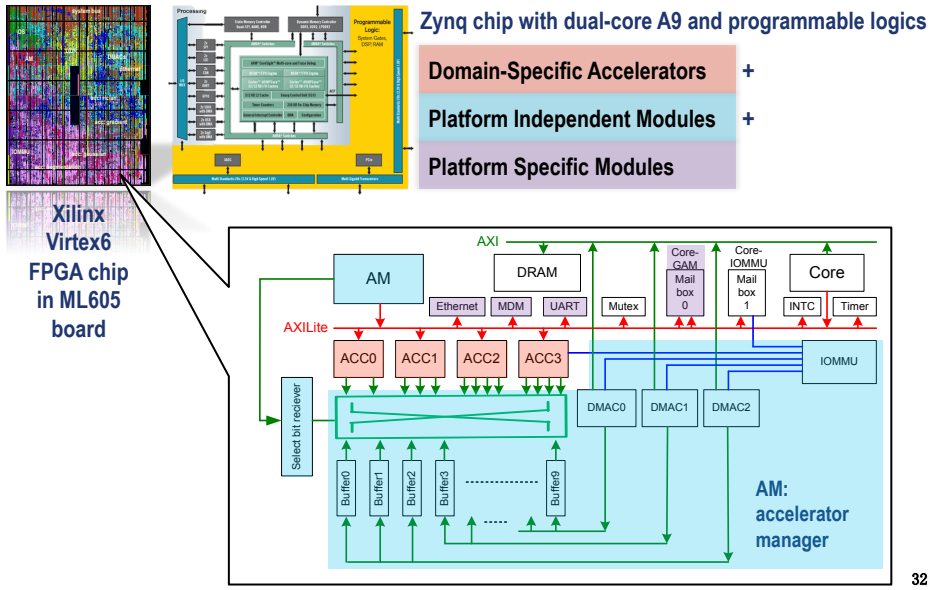
Comparison with Conventional Bus

- ◆ Use Xilinx AXI4 bus IP
- ◆ Memory sharing among accelerators vs private memories → huge area savings
- ◆ Our customized crossbar vs conventional bus → both area savings and performance improvement
 - Conventional bus performs arbitration at every memory access, and optimized for general-purpose access patterns → extra logics and delay spent on arbitrators

	Memory usage	Interconnect cost in # of LUTs	Accelerator subtask runtime
Private memories	3328KB (177%)	0	10.3us
Shared memories via AXI buses	768KB (41%)	50043 (33%)	117us
Shared memories via our crossbar	768KB (41%)	3169 (2%)	10.3us

31

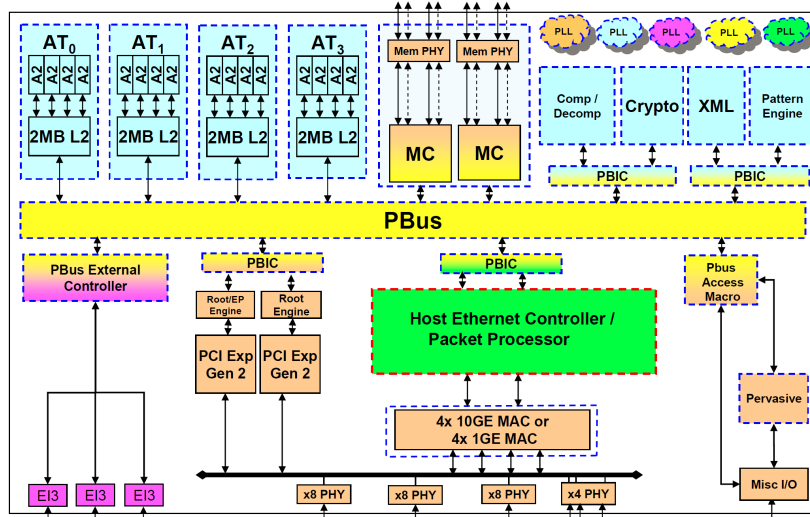
Prototyping of Accelerator-Rich Platform in FPGA [ICCD'2013]



32

IBM Wirespeed Processor [ISSCC'2010]

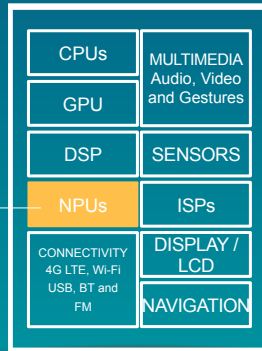
- ◆ 16 A2 cores (each 4-way SMT) sharing a number of accelerators (45nm SOI)



33

Qualcomm Neural Processing Units (NPU)

A new class of processors mimicking human perception and cognition (Oct. 2013)



Massively parallel,
reprogrammable

Comprehensive
tools

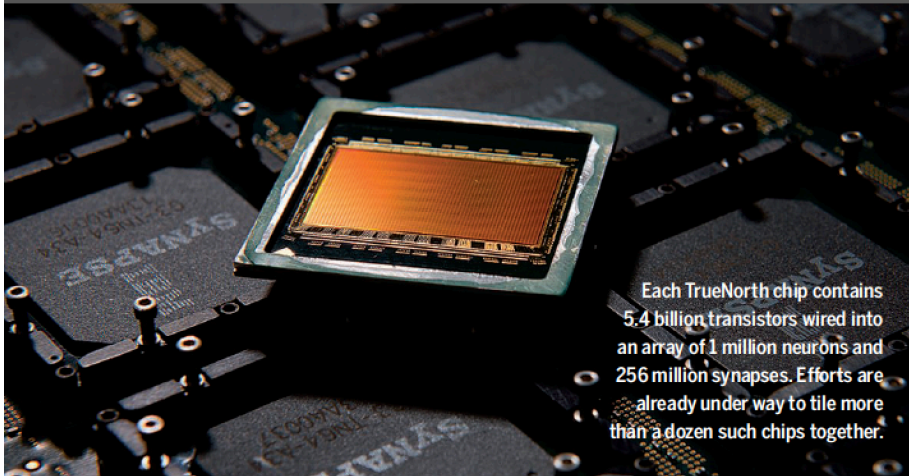
Human-like
functions

Targeting 1M neurons and 1B synapses

Source: Casimir Wierzynski, Qualcomm Research

34

IBM TrueNorth Chip [Science, 2014]



Each TrueNorth chip contains 5.4 billion transistors wired into an array of 1 million neurons and 256 million synapses. Efforts are already under way to tile more than a dozen such chips together.

- Consumes 20 milliwatts
- Both Qualcomm NPU and IBM TrueNorth are ideal candidates as accelerators

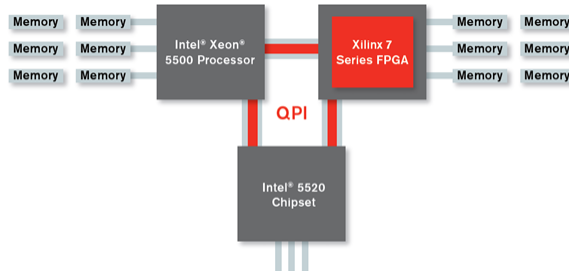
35

Accelerator-Rich Architectures

-- Server Node Level Platforms

36

Server Node-level Interconnects



	PCI-E Gen3	FSB	QPI
Vendors / Products	Xilinx Virtex Eval Bd Altera Stratix Bd Nallatech PCIE Bd Alphadata PCIE Bd Convey HC-2	Nallatech FSB Convey HC-1	Xilinx Virtex-7 QPI IP Altera Stratix5 QPI IP
Bandwidth	8~16 GB/s	5.8 ~ 8 GB/s	25.6 GB/s

37

Server Node-level Interconnects

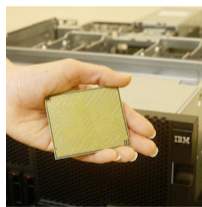
Press releases

Xilinx Demonstrates Industry's First QPI 1.1 Interface with FPGAs at Intel Developer Forum

April 23, 2014

What Power8 and OpenPOWER Might Mean for HPC

Timothy Prickett Morgan



IBM is making a big play in hybrid computing, seeking to marry its POWER8 processors with various kinds of accelerators and high-speed networking and opening up its chip and system software through the OpenPOWER Foundation. The OpenPOWER Foundation is a consortium of...

IBM is working with FPGA makers Xilinx running over the CAPI interface, so this the Impact2014 event. IBM and Xilinx will be accelerated by FPGAs and show order of magnitude lower latency. A Mo machines accelerated by Altera FPGAs adapter and switch maker Mellanox Technologies using Remote Direct Memory Access (RDMA) boosted throughput and cut latencies.

Altera Demonstrates Industry's First QPI 1.1 FPGA Home Agent for Enhanced Server Capabilities

Demo Features Stratix V FPGA Configured to Extend Co-processing for Intel Processors

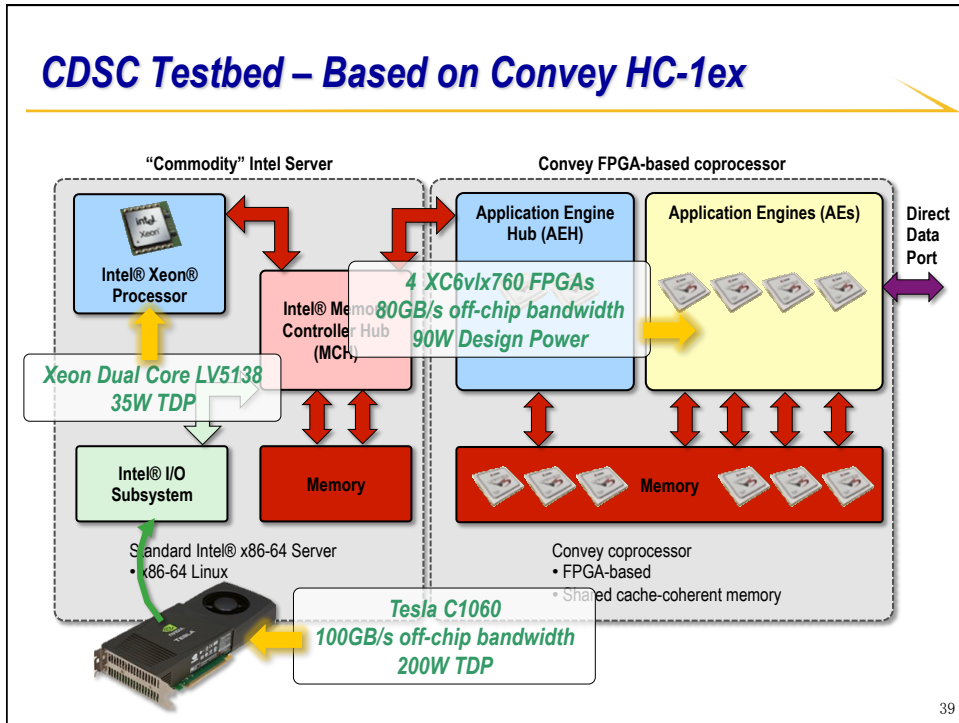
San Jose, Calif., and IDF Beijing, April 10, 2013—Altera Corporation (NASDAQ: ALTR) today announced the industry's first demonstration of an FPGA Home Agent enabled by the Intel QuickPath Interconnect (QPI) protocol 1.1. Connecting to Intel's Sandy Bridge XEON processors, the demonstration leverages an Altera® Stratix® V FPGA configured as the Home Agent, and it supports both the Caching Agent and Home Agent in a Pactron Vigor Development Platform. This solution is ideal for designers of low-latency signal-processing, packet processing and embedded applications, such as high-frequency trading and big data that need higher computation performance-per-watt than traditional CPU configurations can deliver. Altera is demonstrating its QPI 1.1 intellectual property (IP) solution to support both the Caching Agent and Home Agent in a Pactron Vigor Development Platform at the Intel Developers Forum (IDF) Beijing, April 10-11, in Altera's booth #E120.

QPI is the only way to coherently connect to an Intel server processor. The Altera Stratix V FPGA transceiver has been qualified to support the Intel QPI electrical specification at 8 Gbps. Developers of low-latency, high-bandwidth systems looking to extend the flexible shared memory model that Intel uses for x86 programming can now efficiently integrate a Stratix V FPGA into their systems. The Home Agent demo addresses 32 GB of memory on the motherboard connected to the socket with support for two channels connecting to four 8 GB RDIMMs.

"Our QPI 1.1 solution provides developers of data centers and high-performance computing applications a platform to significantly increase their compute performance while reducing system cost and power," said David Gamba, director of the compute and storage product line at Altera. "FPGAs deliver a highly effective, efficient way to speed the processing of large data sets through parallel processing and accelerated data transfers."

38

CDSC Testbed – Based on Convey HC-1ex



39

Server Node-level Acceleration Example : 3D CT-Reconstruction Algorithm (EMTV)

◆ Optimization techniques:

- Data reuse
- Prefetching
- Pipelining
- Fixed-point optimization
- Parallelization



Platform	Exec Time	Speedup
CPU	54.6 hours	1
GPU	110 mins	29.8X
FPGA	38.8 mins	84.5X

→ Sinogram reuse
 → Sinogram + voxel reuse

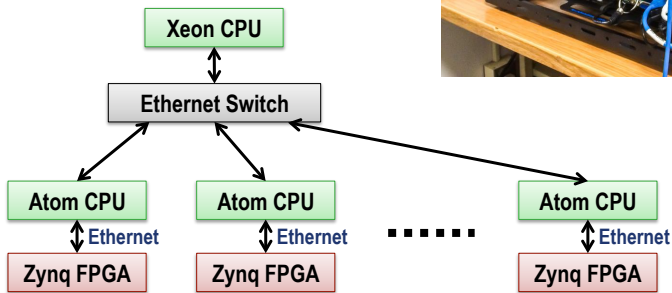
40

Accelerator-Rich Architectures -- Datacenter Level Integrations

41

FPGA “FARM” at UCLA – A Small Multi-FPGA Rack

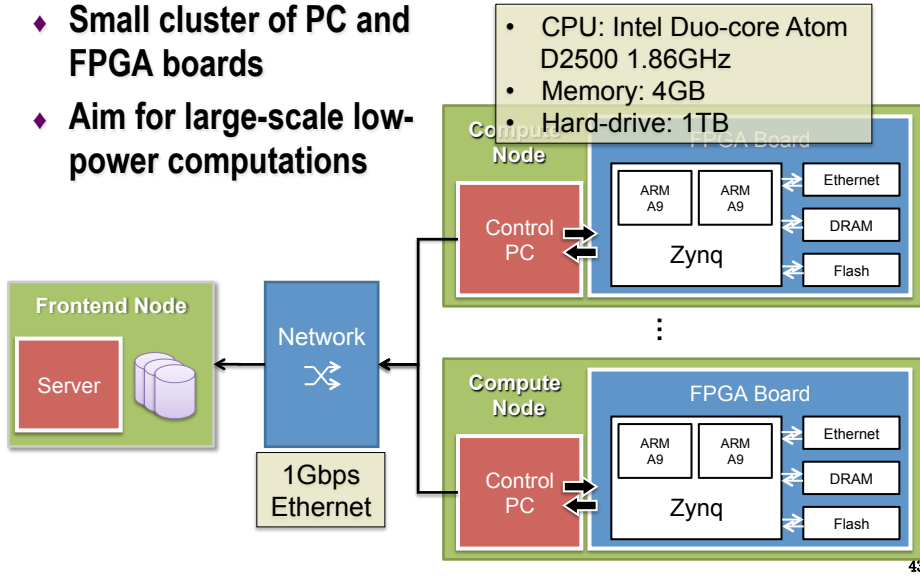
- Deployed in 2013
- Used for research and teaching
 - CS133 (60+ students)
 - CS259 (18 students)



42

FPGA “Farm” at UCLA – A Multi-FPGA platform

- ◆ Small cluster of PC and FPGA boards
- ◆ Aim for large-scale low-power computations



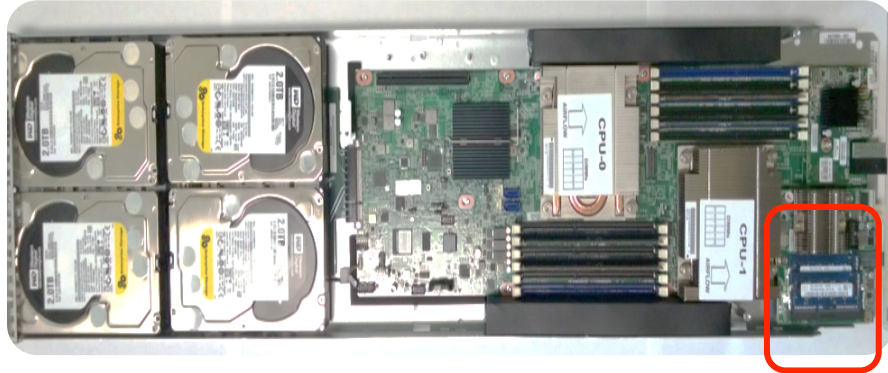
Datacenter Level Integration at Microsoft

- Microsoft Open Compute Server
- 1U, 1/2 wide servers
- Enough space & power for 1/2 height, 1/2 length PCIe card
- Squeeze in a single FPGA
- Won't fit (or power) GPU



A. Putnam, "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services", ISCA'2014

Microsoft Open Compute Server



- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs, 2 SSDs
- 10 Gb Ethernet
- No cable attachments to server

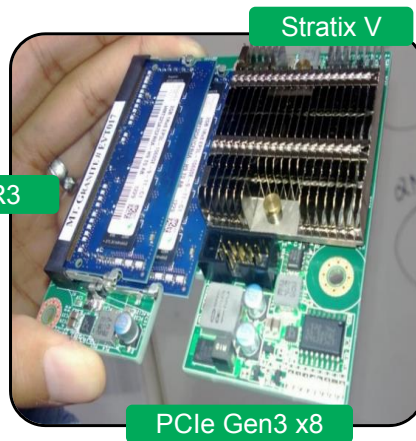
Air flow

A. Putnam, "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services", ISCA'2014

45

Catapult FPGA Accelerator Card

- Altera Stratix V D5
- 172,600 ALMs, 2,014 M20Ks, 1,590 DSPs
- PCIe Gen 3 x8
- 8GB DDR3-1333
- Powered by PCIe slot
- Torus Network



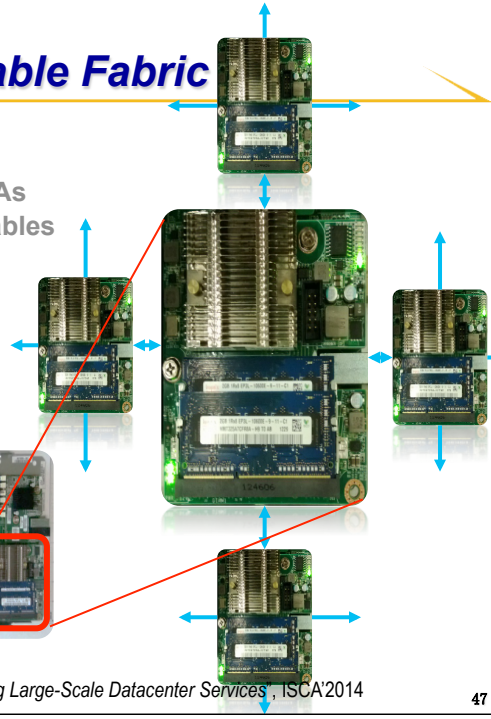
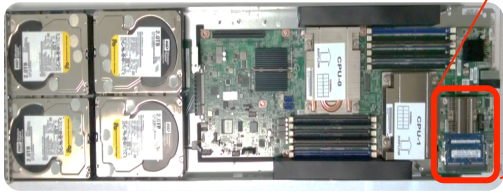
A. Putnam, "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services", ISCA'2014

46

Scalable Reconfigurable Fabric

- 1 FPGA board per Server
- 48 Servers per 1/2 Rack
- 6x8 Torus Network among FPGAs
 - 20 Gb over SAS SFF-8088 cables

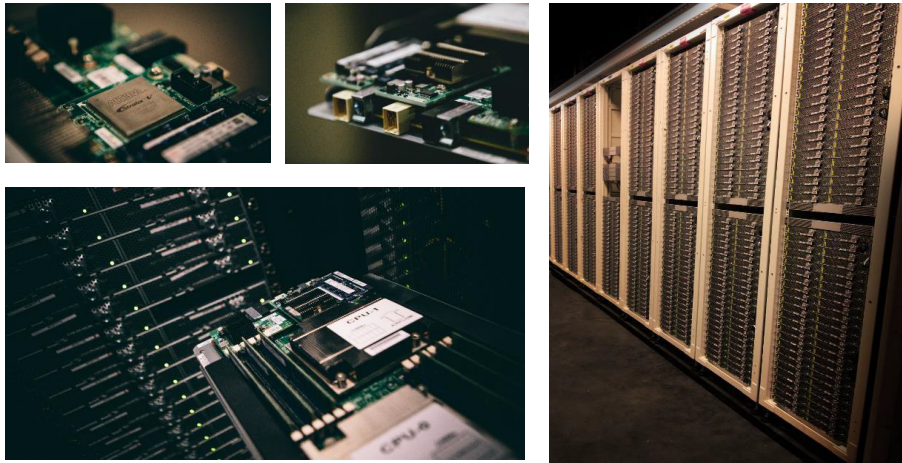
Data Center Server (1U, 1/2 width)



A. Putnam, "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services", ISCA'2014

47

1632 Server Pilot Deployed in a Production Datacenter

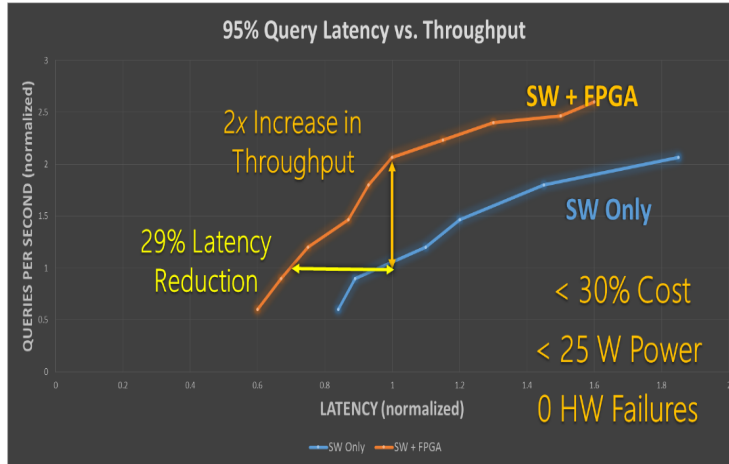


A. Putnam, "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services", ISCA'2014

48

Accelerating Large-Scale Services – Bing Search

1,632 Servers with FPGAs Running Bing Page Ranking Service (~30,000 lines of C++)

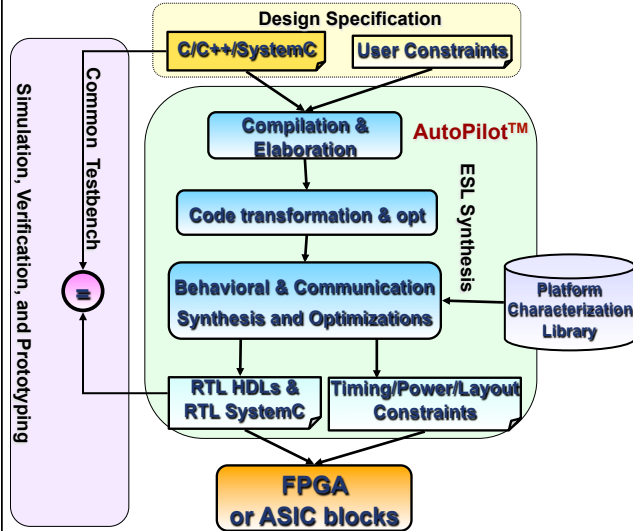


A. Putnam, "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services", ISCA'2014

How to Program Such "Beasts"?

-- Datacenter Level Solutions

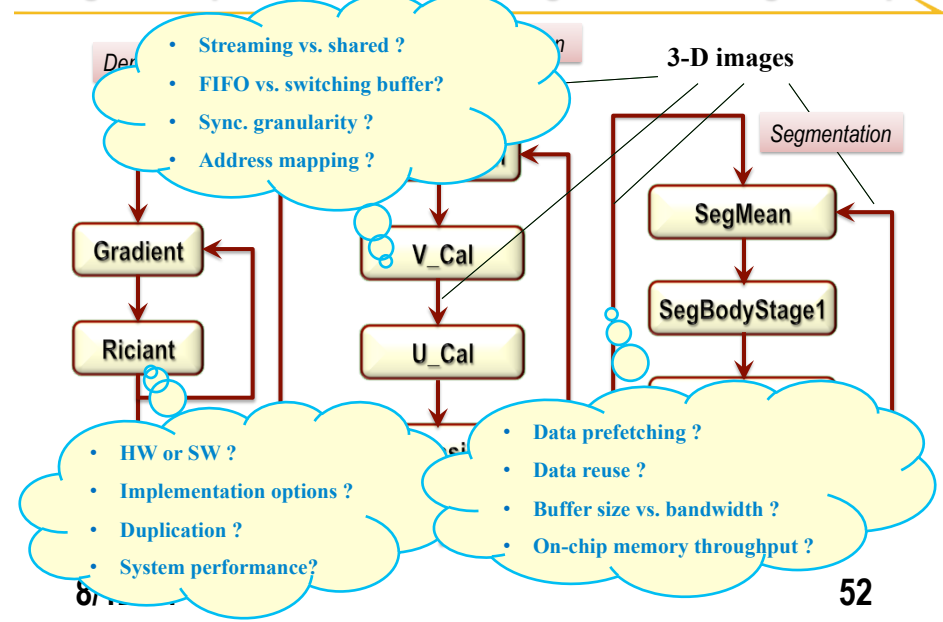
1. C/C++ Based Synthesis for Dedicated Accelerators xPilot (UCLA) -> AutoPilot (AutoESL) -> Vivado HLS (Xilinx)



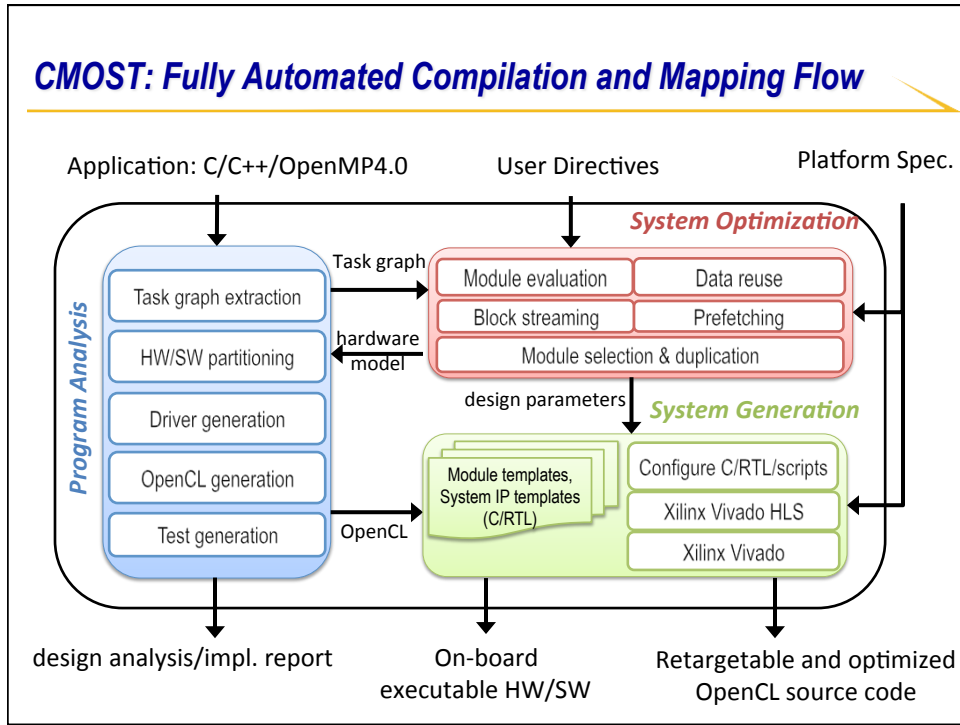
- ◆ Platform-based C to RTL synthesis
 - ◆ Synthesize pure ANSI-C and C++, GCC-compatible compilation flow
 - ◆ Full support of IEEE-754 floating point data types & operations
 - ◆ Efficiently handle bit-accurate fixed-point arithmetic
 - ◆ SDC-based scheduling
 - ◆ Automatic memory partitioning
 - ◆ ...
- QoR matches or exceeds manual RTL for many designs

Developed by AutoESL, acquired by Xilinx in Jan. 2011

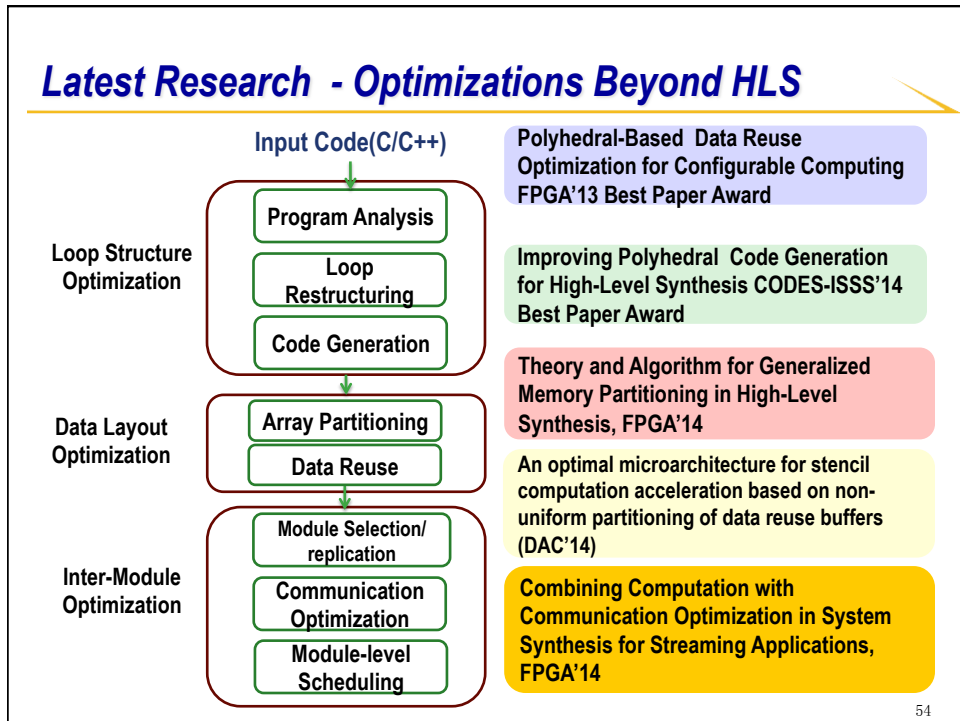
Design Complexity: Medical Image Processing Example



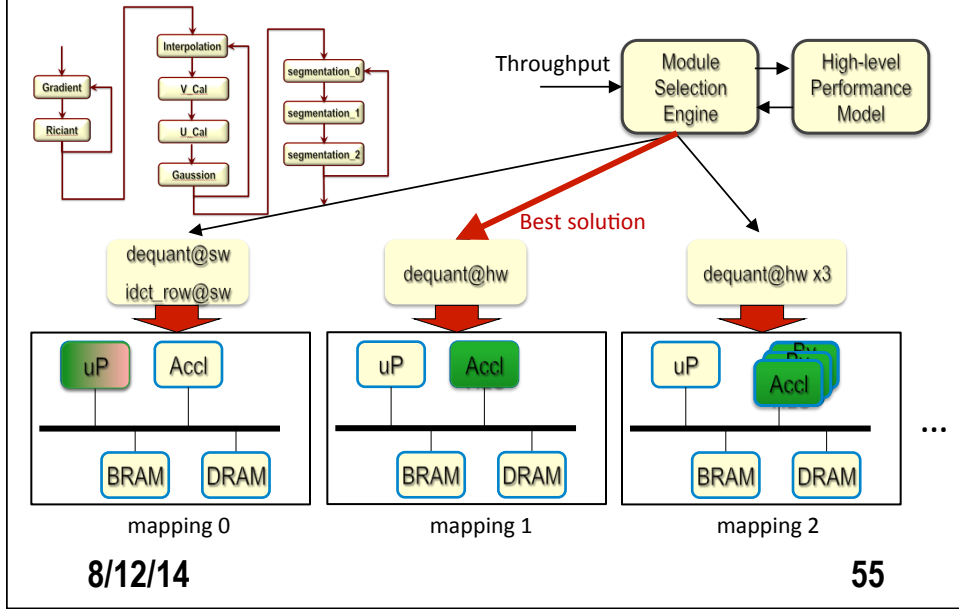
CMOST: Fully Automated Compilation and Mapping Flow



Latest Research - Optimizations Beyond HLS

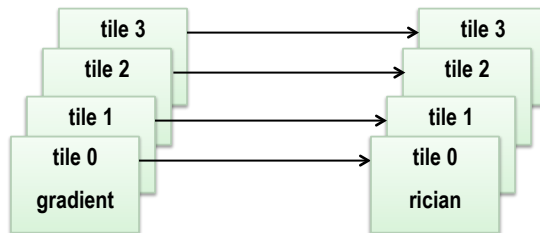


Example: Throughput-Driven Task Scheduling and Mapping [FPGA'2014]



Motivation

Tile size: 32x32
Image: 64x64, 4 tiles



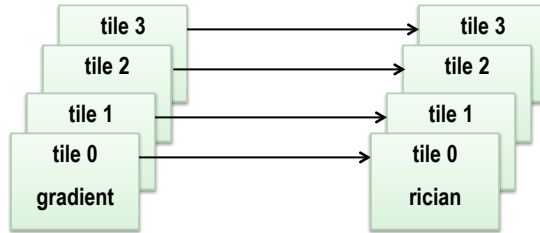
■ Which implementation to use for each module?

■ Memory partitioned v.s. Memory non-partitioned

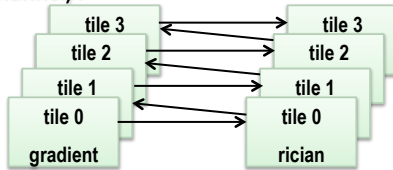
	BRAM	DSP	FF	LUT
non-partitioned gradient	128	21	2511	2125
partitioned gradient	176	56	7147	7262
partitioned rician	128	22	4692	3991
non-partitioned rician	176	88	14475	15537

Motivation

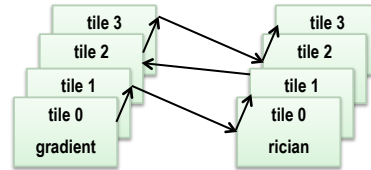
Tile size: 32x32
Image: 64x64, 4 tiles



- How many number of replicas?
- Scheduling and Communication cost (number of tiles in the communication channel)?



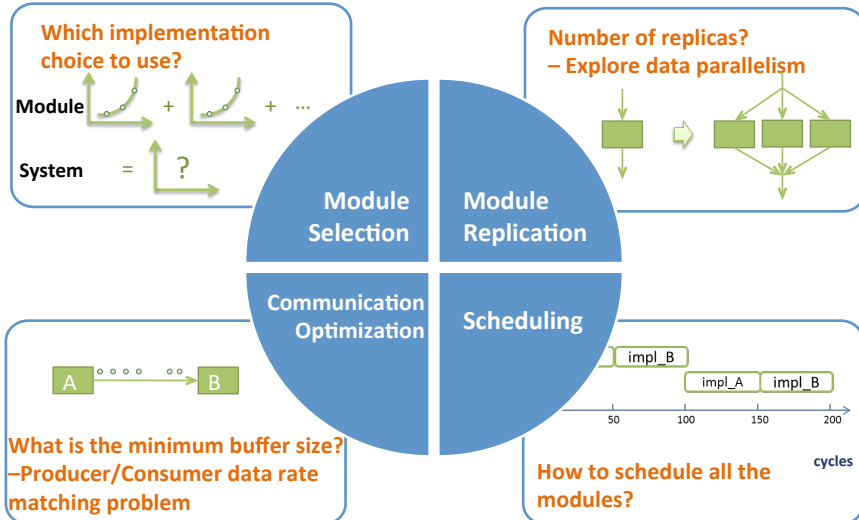
scheduling 0 → 1 tile



scheduling 0 → 2 tiles

57

A Rich Design Space: System-Level Synthesis with HLS for Streaming Applications

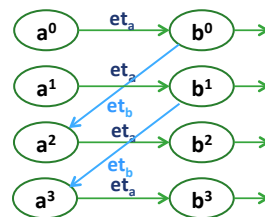


58

Formulation (1/2)

Derive a scheduling graph

- Associate each node with a time variable, denoting the starting time of the node
- Scheduling graph: delineates all the scheduling constraints
 - Module latency, Module replication, System throughput requirement, Buffer constraints



Module latency constraints
 et_a : execution time of task a
 et_b : execution time of task b

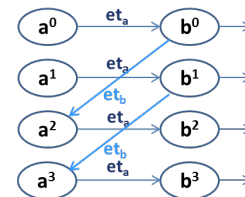
Buffer Constraints
 If buffer size between a and b is 2,
 then add edges: $b^0 \rightarrow a^2$ $b^1 \rightarrow a^3$

59

Formulation (2/2)

Associate each node with a scheduling variable

- $t(b^0) - t(a^0) \geq et_a$
- $t(a^2) - t(b^0) \geq et_b$
- ...
- Scheduling variables are integer variables



Schedulability checking problem is a System of Difference Constraints (SDC) problem

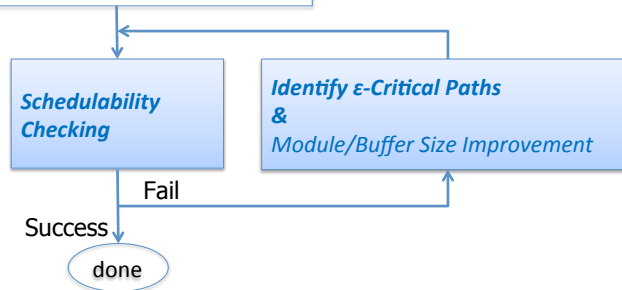
- It can be solved optimally *in polynomial time* by linear programming relaxation
- And the solution is guaranteed to be integers

60

Streaming Synthesis (ST-Syn)

- **Formulation – Schedulability checking**
 - System of Difference Constraint Problem
- **Exploration – Identify critical path, module/buffer improvement**
 - Find ϵ -critical paths in the scheduling graph
 - Minimum cut problem
- **All can be solved by linear programming relaxation**

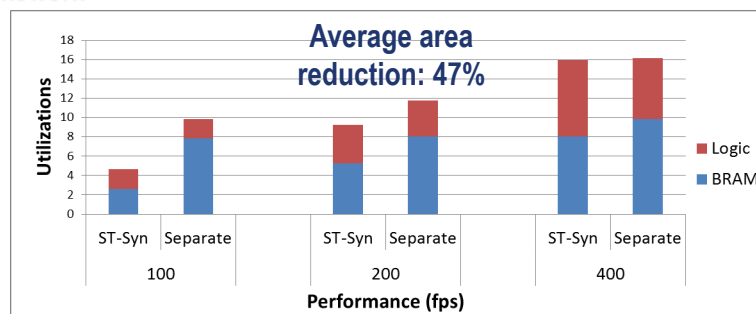
Start from the impl with the smallest logic, minimum buffer size



61

Experiments on Example Denoise

- **Our methodology: ST-Syn**
 - computation & communication co-optimization
- **Separate:**
 - separate computation opt. + communication opt.
- **→ Communication and computation should be considered in a unified framework**



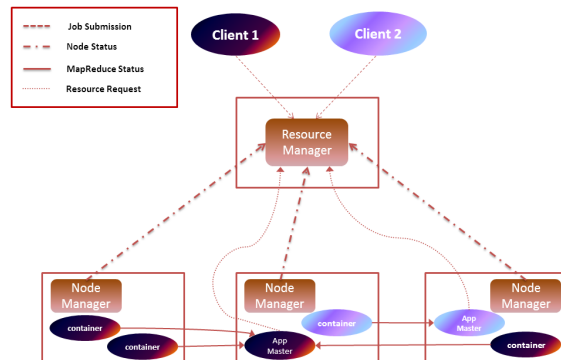
62

What about Runtime Management? -- Datacenter Level Integration

63

Example of Data Center Level Computing Model: MapReduce (or Hadoop)

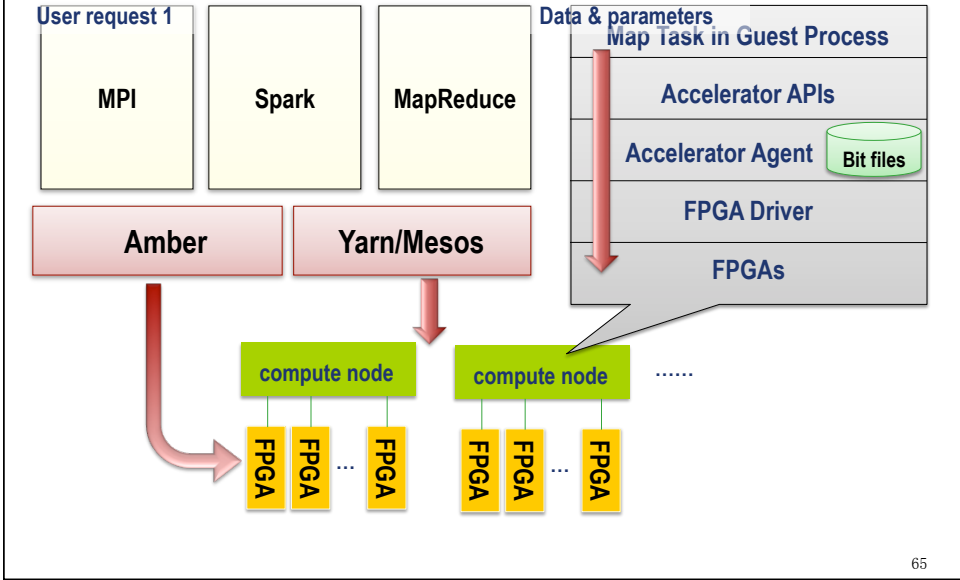
- A popular programming model for data centers
- Automatically compose and launch tasks in distributed compute nodes
- Limitation: unaware of accelerators



source: <http://www.bigo.co.in/2013/06/hadoop-yarn-next-generation-mapreduce.html>

64

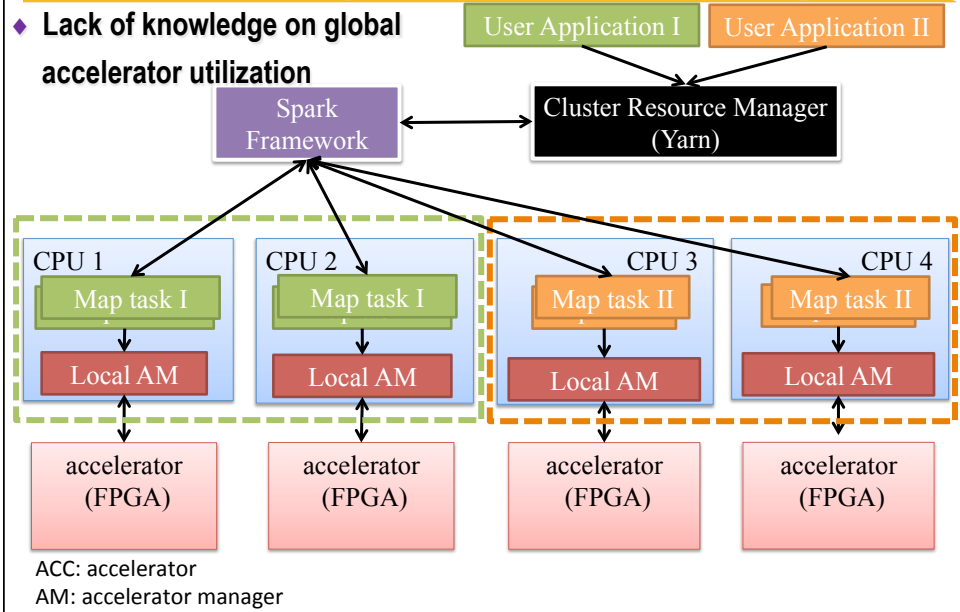
Runtime Framework with Global Accelerator Manager (Amber)

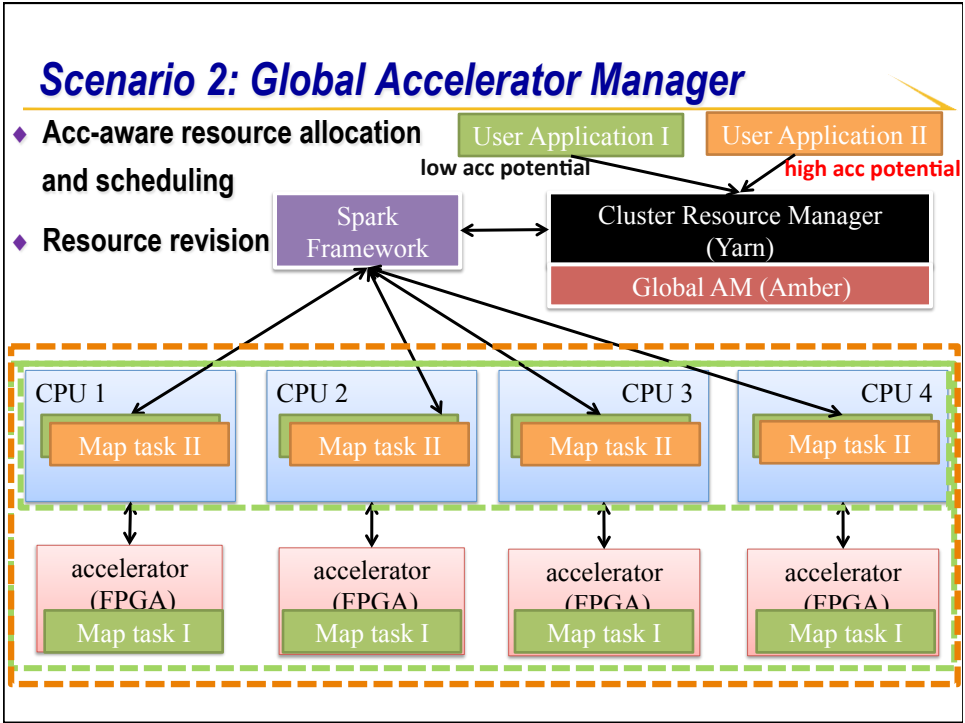


65

Solution with Only Local Accelerator Manager

- ◆ Lack of knowledge on global accelerator utilization





Initial Experimental Result

- ◆ **Cluster setting**
 - 4 CPU nodes (Xeon), each connected to one FPGA node (ML605)
- ◆ **User application**
 - Application I: Logistic Regression (LR) – 2x FPGA speedup
 - Application II: Neural Network (NN) Training – 9x FPGA speedup

	LR first	NN first	LR, NN simul.
Local AM	6.14s	0.62s	1.23s
Global AM	0.85s	0.62s	0.62s
Speedup	7.22x	--	2x
Energy saving	10.2x	--	1.45x

- With local AM, the first application will occupy all the accelerator resources
- With global AM (resource revision), more acc/FPGA resources will be allocated to applications with higher acceleration potential (NN)

A New Phase of CDSC -- Innovation Transition

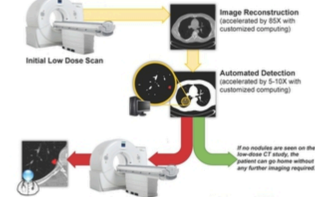


National Science Foundation
Directorate for Computer & Information Science & Engineering (CISE)

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

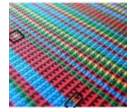


Real-time adaptive low-dose CT-scan enabled by customized computing.
Credit and Larger Version

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.



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](#)

Overall Goals and Research Vectors


Director: Jason Cong, UCLA, cong@cs.ucla.edu



Center for Domain Specific Computing

- ◆ **Overall Goal**
 - Accelerator-rich architectures (ARA) for Big Data applications in healthcare
 - Demonstrate 10-100X improvement in performance/energy medical imaging and personalized cancer treatment
- ◆ **Research Vectors (RVs)**
 - RV1. Accelerator-rich Architecture (ARA)
 - RV2. Compilation and Runtime Support for ARAs
 - RV3. Personalized Cancer Treatment (Application Domain 1)
 - RV4. Medical Imaging (Application Domain 2)
 - RV5. ARA Prototyping and Deployment

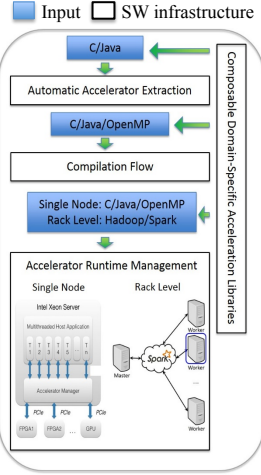
Customized accelerators






Supercomputer in a rack

Software

■ Input □ SW infrastructure



Composable Domain-Specific Acceleration Libraries

71

Concluding Remarks

- ◆ **New era of computing**
 - Future computing platforms will have a sea-of-accelerators
 - With efficient support for customization and specialization
- ◆ **Accelerators at all levels**
 - Chip-level
 - Server node level
 - Data center level
- ◆ **Customizable and composable accelerators offer the right trade-off between flexibility and efficiency**
- ◆ **Software is the key**
 - Programming models
 - OpenMP 4.0, OpenCL, Hadoop/MapReduce + C/C++, ...
 - Compilation support
 - Runtime management

72

Acknowledgements – CDSC and C-FAR

- ◆ Center for Domain-Specific Computing (CDSC) under the NSF Expeditions in Computing Program and C-FAR Center under the STARnet Program
- ◆ CDSC faculty:



Aberle
(UCLA)



Baraniuk
(Rice)



Bui
(UCLA)



Chang
(UCLA)



Cheng
(UCSB)



Cong (Director)
(UCLA)



Palsberg
(UCLA)



Potkonjak
(UCLA)



Reinman
(UCLA)



Sadayappan
(Ohio-State)



Sarkar
(Associate Dir)
(Rice)



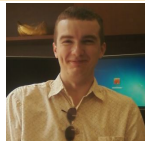
Vese
(UCLA)

73

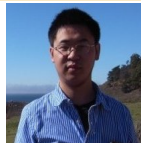
Postdoc and Graduate Students in Collaboration



Mohammad Ali Ghodrat
(UCLA)



Michael Gill
(UCLA)



Yi Zou
(UCLA)



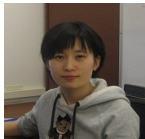
Beayna Grigorian
(UCLA)



Chunyue Li
(UCLA)



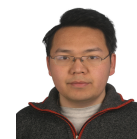
Prof. Deming Chen
(UIUC/ADSC)



Hui Huang
(UCLA)



Muhuan Huang
(UCLA)



Dr. Peng Li
(UCLA)



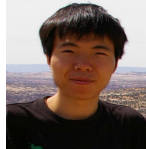
Prof. Louis-Noël Pouchet
(UCLA)



Bo Yuan
(UCLA)



Yuxin Wang
(PKU)



Bingjun Xiao
(UCLA)



Dr. Peng Zhang
(UCLA)



Wei Zuo
(UIUC)

74

