# Heterogeneous Architectures for CHIPS

**Jason Cong**

*Distinguished Professor and Chancellor's Professor, UCLA*

*Director, Center for Domain-Specific Computing*

cong@cs.ucla.edu

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

# Why Heterogeneous Architecture?

| 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*

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
Faculty experts



# 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.

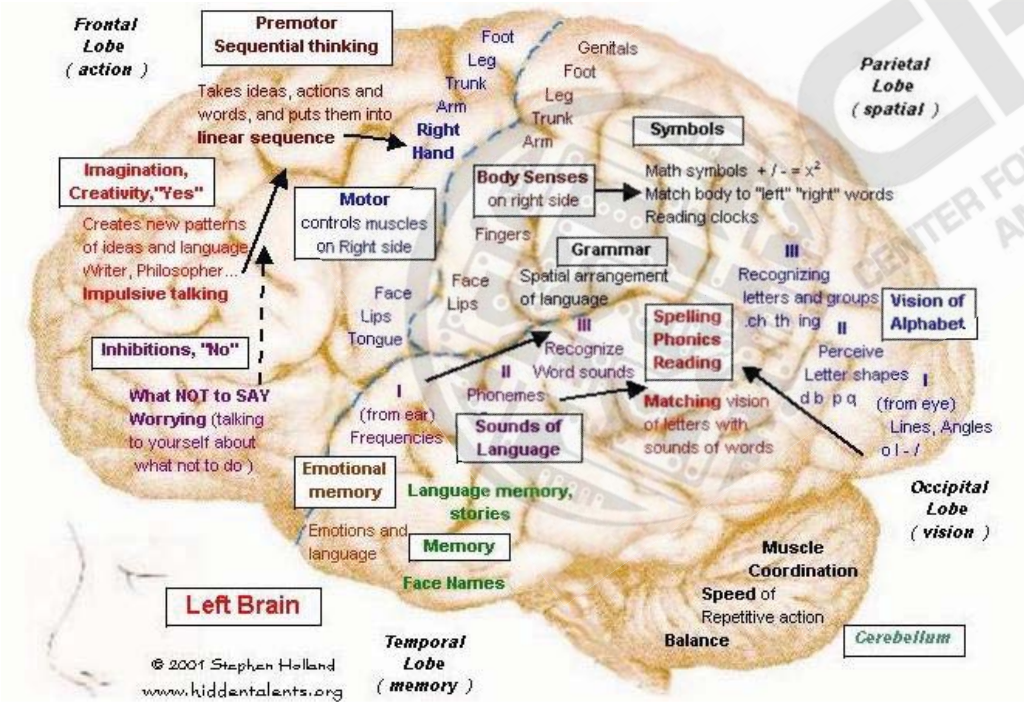
The goal of the new UCLA center, Cong said, is to look beyond parallelization and focus on domain-specific customization to bring significant power-performance efficiency improvement to important application domains.

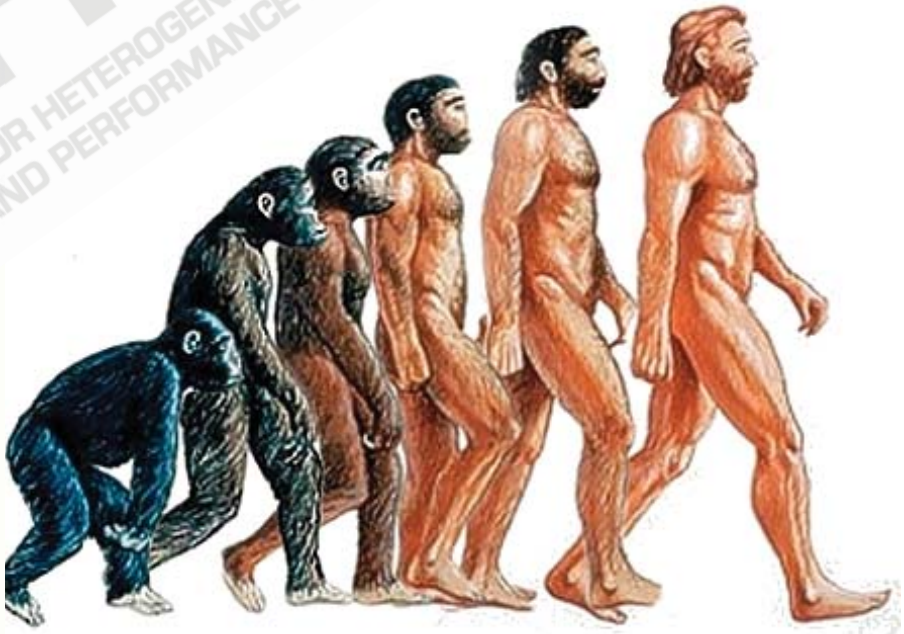
# *Our Proposal – Customized Computing with Accelerator-Rich Architectures*

- **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

# 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

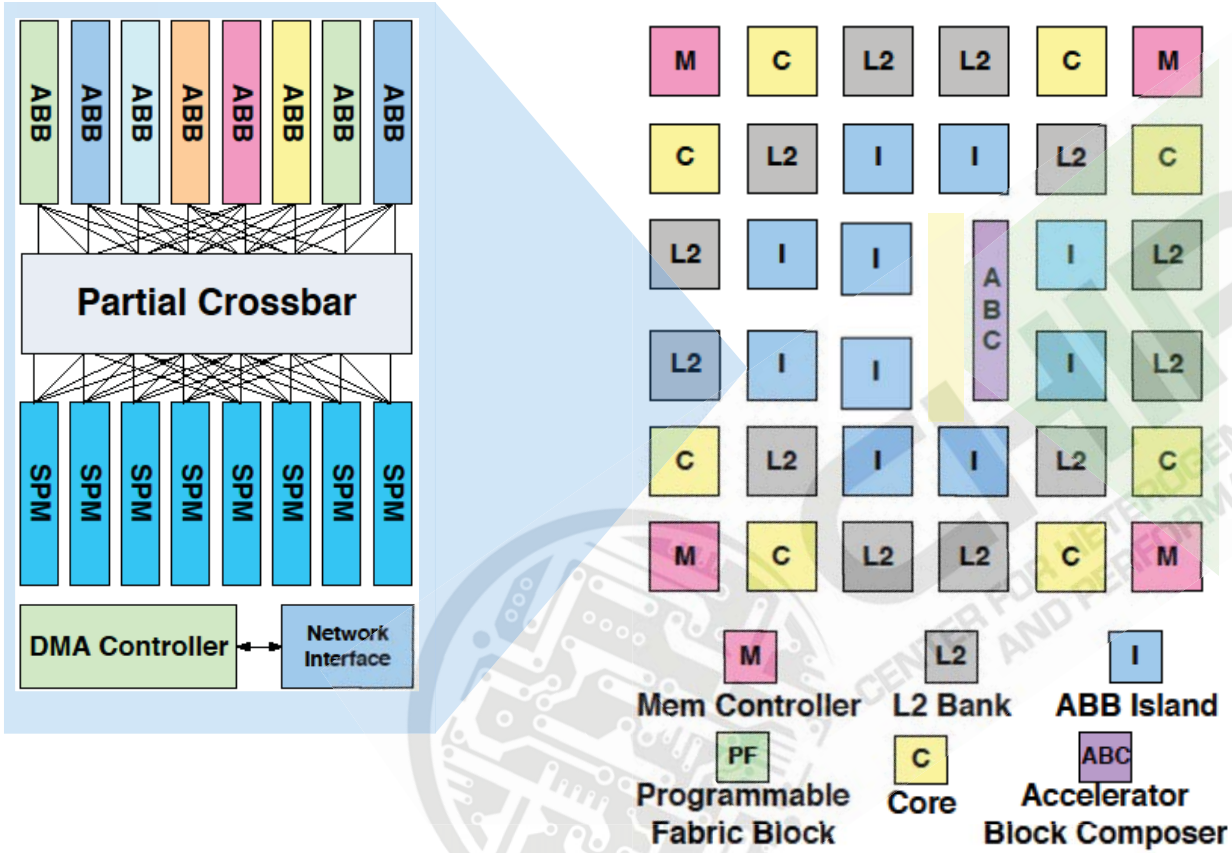


# *Current Research at CDSC:*
## *-- Three Levels of Customization*

- **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

# Composable Accelerators with Programmable Fabrics [ISLPED'2013]




**Dynamic Resource Allocation of ABBs**

- **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

# *Levels of Customization*

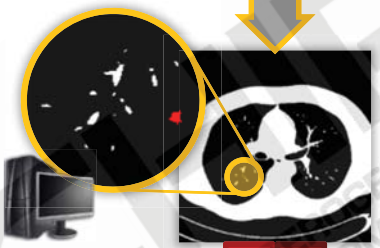
- **Single-chip level**
  - Require new processor designs, e.g. using composable accelerators [ISLPED' 12, DAC'14]
- **Server node level**
  - Host CPU + FPGA via PCI-e or QPI connections
- **Data center level**
  - Clusters of heterogeneous computing nodes

# A Success Application: Low-Dose Adaptive CT Scan

1. Initial Scan

(low-dose CT scan)

2. Image Reconstruction



3. Automated Detection

*If no nodules are seen on the low-dose CT study, the patient can go home without any further imaging required.*

5. Clinical Interpretation

4. Adaptive Diagnostic Scan

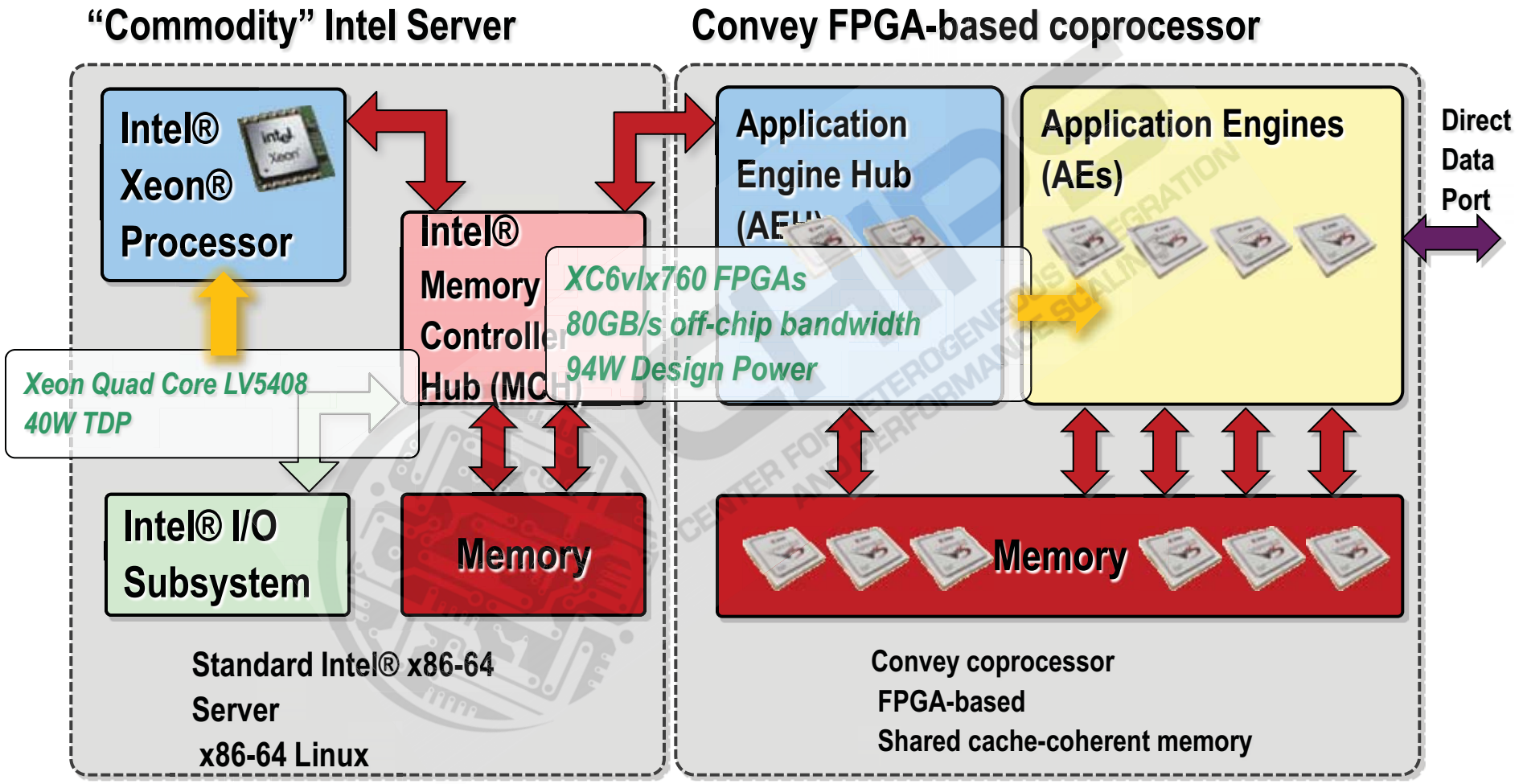
If nodules are detected, a diagnostic follow-up scan is performed during the same visit.

Center for Domain Specific Computing

UCLA Health RADIOLOGY

CFAR

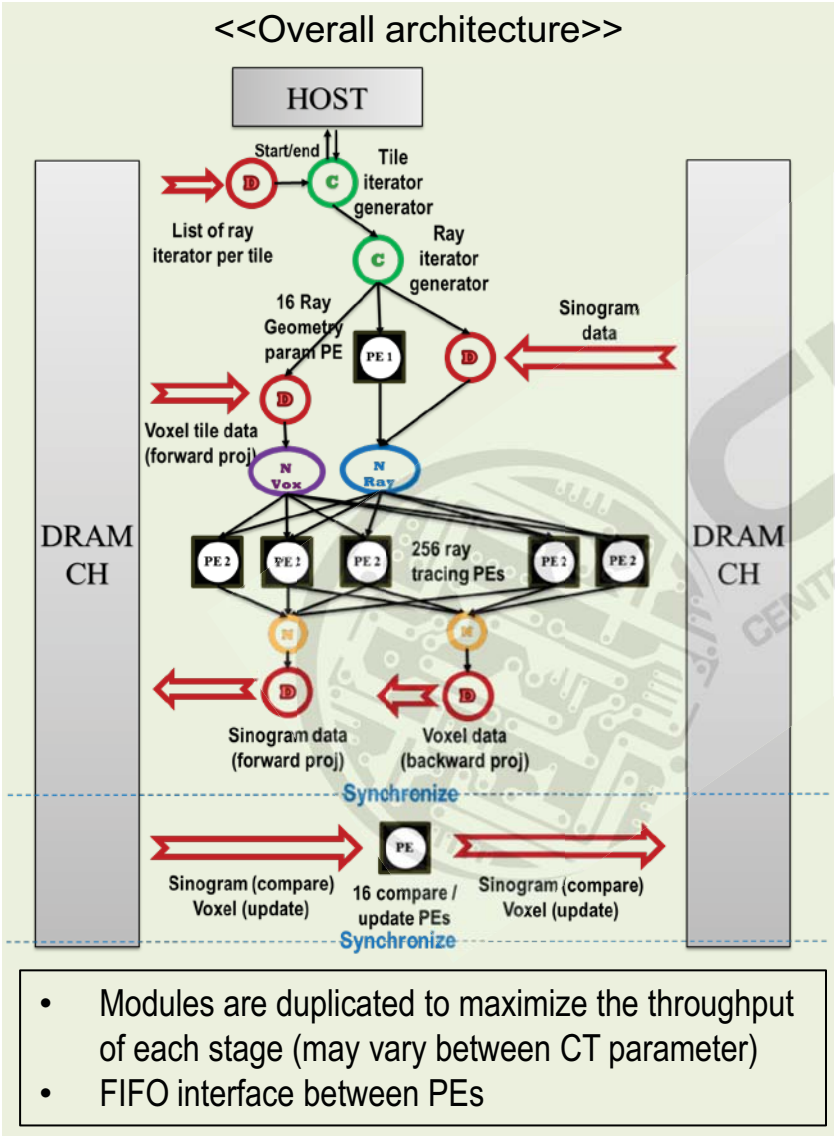
# Example of CDSC Heterogeneous Computing Server



# Accelerator Architecture / Performance [FCCM 2014]

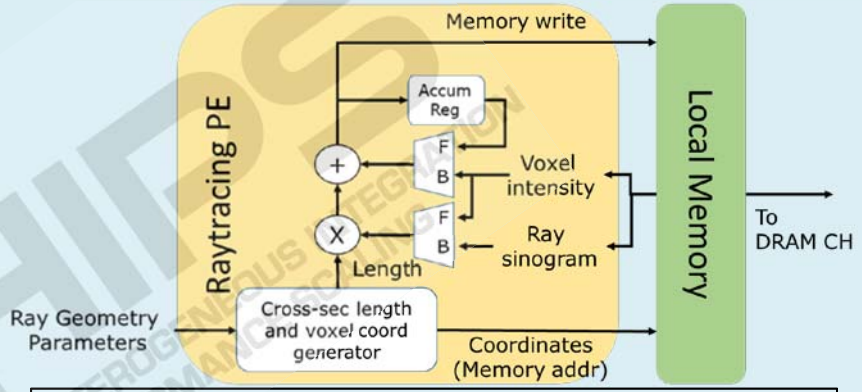
## Throughput optimization using PE array architecture

<<Overall architecture>>




- Modules are duplicated to maximize the throughput of each stage (may vary between CT parameter)
- FIFO interface between PEs

<<Raytracing PE>>




- Tightly coupled local memory provides 1 voxel/s throughput
- Forward / Backward projection combined to one PE

<<Performance>>

| Compute Node | Release Year | Forward Proj | Backward Proj | Other | Total |
|---|---|---|---|---|---|
| CPU - 1thrd (Xeon 5138) | 2008 | 1253 | 2025 | 2 | 3280 (1.0X) |
| CPU - 8thrds (Xeon E5-2420) | 2012 | 88 | 154 | 2 | 243 (13.5X) |
| GPU - 3072 thrds (Tesla C1060) | 2008 | 19.8 | 90.6 | - | 110 (29.8X) |
| GPU - 9216 thrds (Tesla M2070 3X) | 2010 | 9.5 | 23.5 | 0.4 | 33.5 (97.9X) |
| FPGA - 64PEs[5] (Vir6 LX760 4X) | 2009 | 28.3 | 93.3 | 3.1 | 125 (26.2X) |
| Proposed (Vir6 LX760 4X) | 2009 | 3.1 (404X) | 3.0 (675X) | 1.0 - | 7.2 (455X) |

# *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
  - How about programming at data center level?

# CDSC FPGA-Enabled Cluster

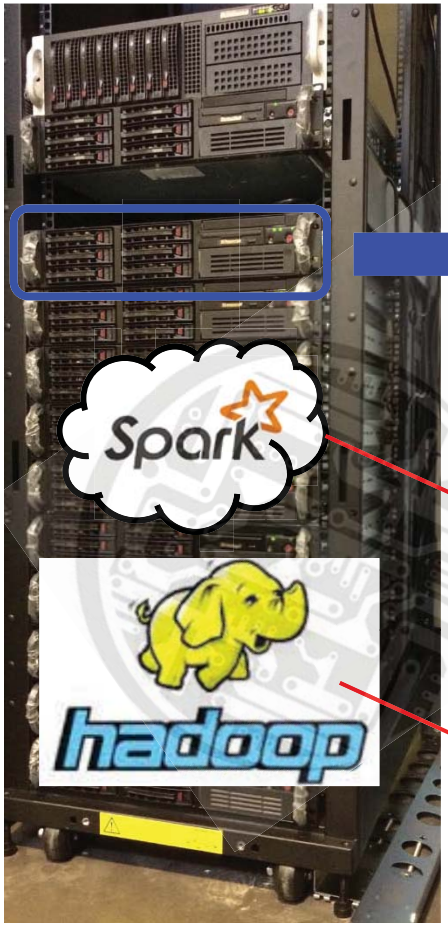
- A 24-node cluster with FPGA-based accelerators
  - Run on top of Spark and Hadoop (HDFS)

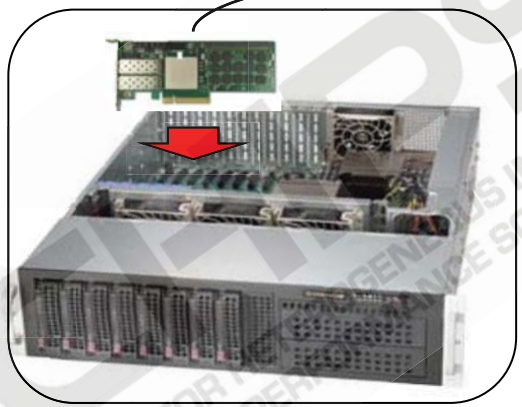
1 master / driver

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

# Example Application: Personalized Cancer Treatment

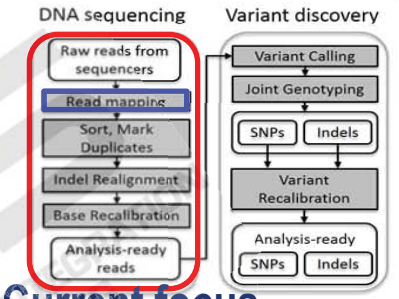
A large genome collection of healthy population and cancer patients

Big data, compute-intensive

Customized accelerators


Supercomputer in a rack

*Genomic* analysis pipeline

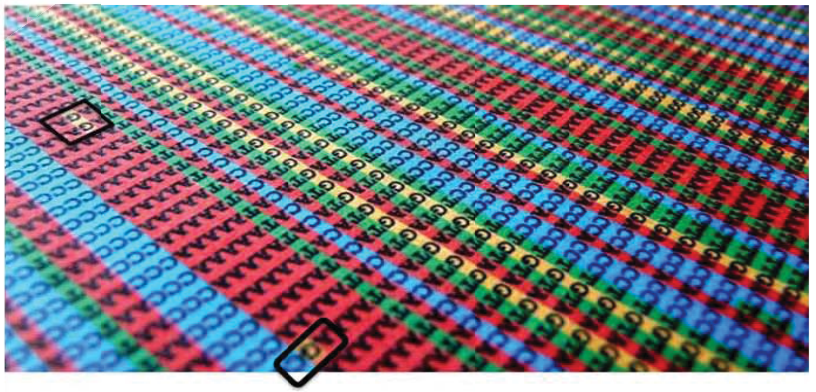



**Current focus**

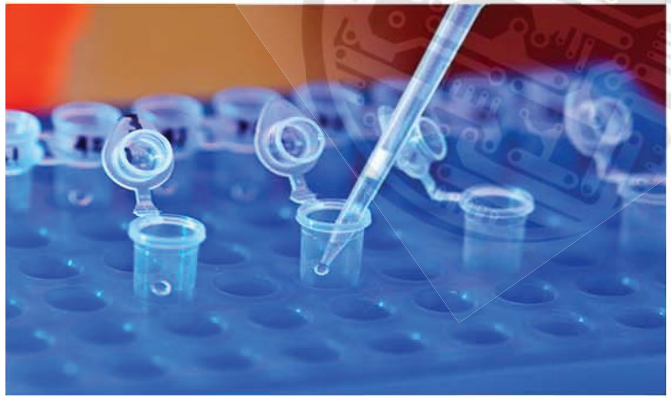
Automatic Accelerator Extraction

Accelerator Runtime Management

Gene mutations discovered in codon reading frames

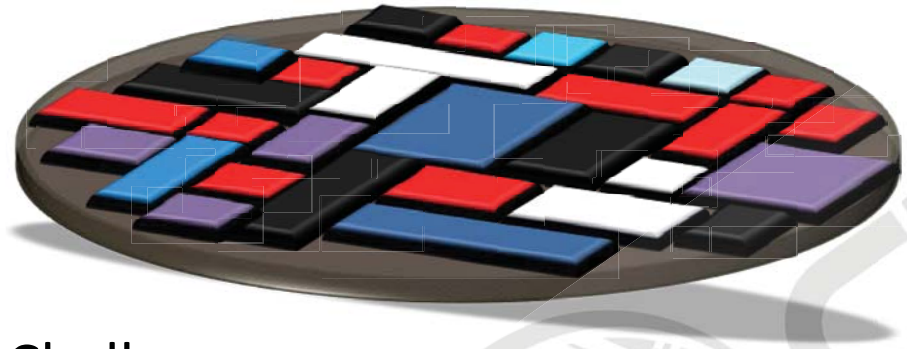


Enable effective searches of precision medicine for cancer treatment



# *New Opportunity*
# *-- Customization at CHIPS Level*

# Mega SoIFs by re-integration on an Interconnect Fabric



The ”right” interconnect fabric

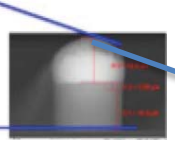
- Mechanically robust (flat, stiff, tough…)
- Capable of fine wiring, fine pitch interconnects
- Thermally conductive
- Can have active and passive built-in components

**Silicon Fits the Bill**

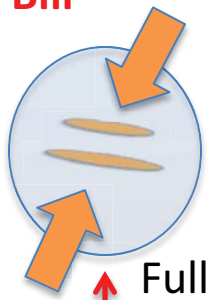
Challenges:

- Warpage
- Topography
- Assembly /Thru’put
- Thermal

>100 μm pitch
Mass reflow




100μm >pitch > 40 μm
Mass re-flow+TCB



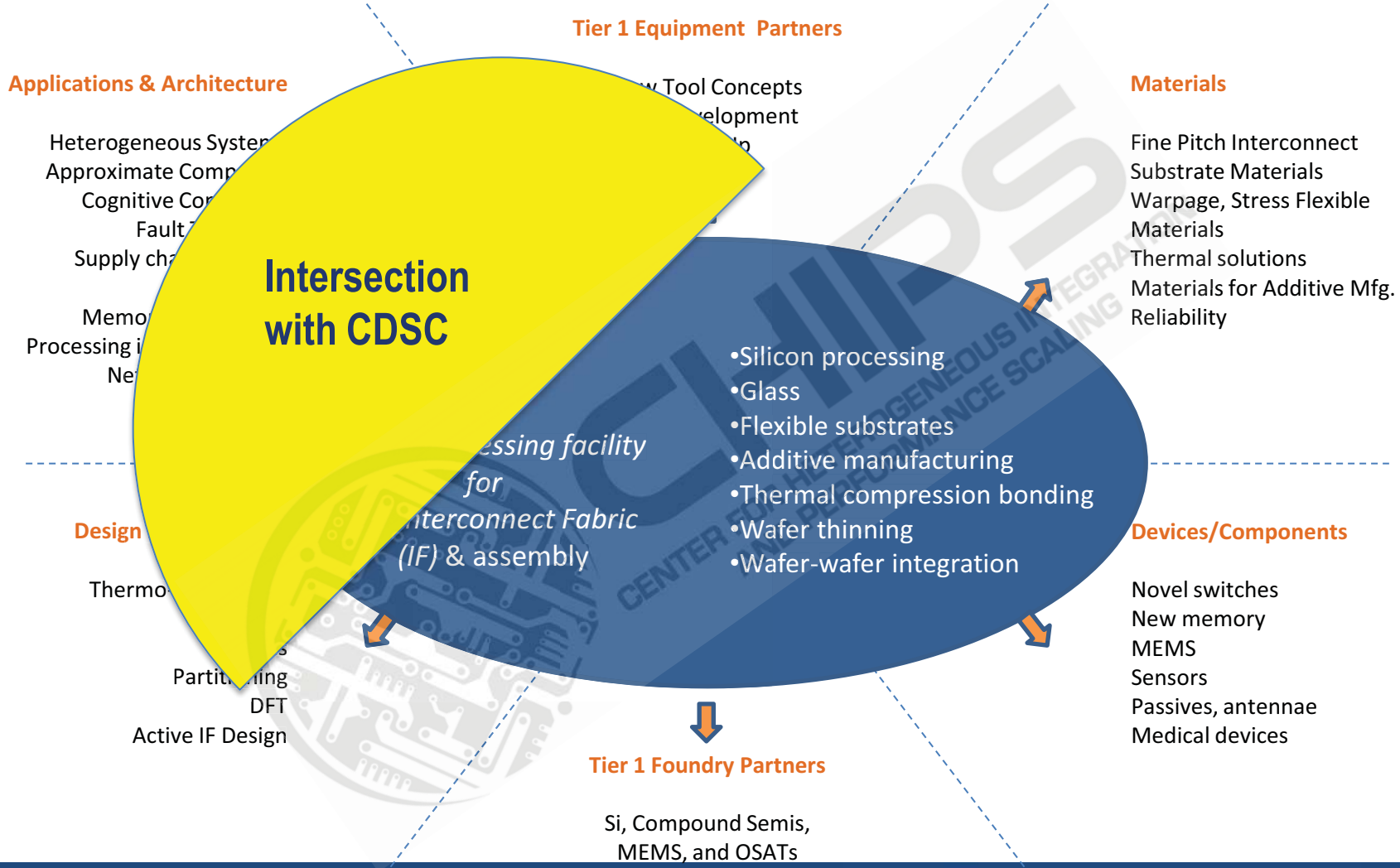
2-10μm

Full contact – TCB

Proximate

- Inductive
- Capacitive

# CHIPS Framework



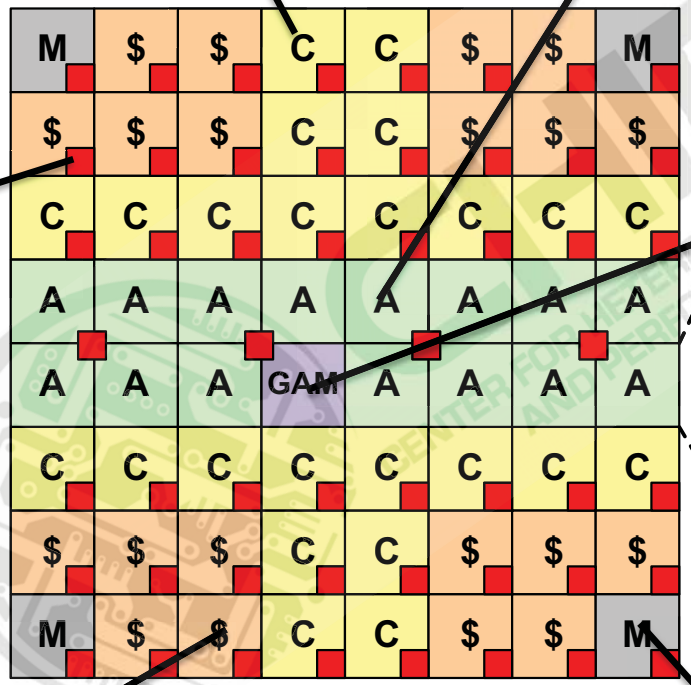
# Design Tool Support is Critical

- Modeling
- Simulation
- Synthesis
- Verification
- …

# Example: PARADE -- Platform for ARA Design & Exploration [ICCAD'15]

Extended m5 (McPAT) for X86 CPU, with OS

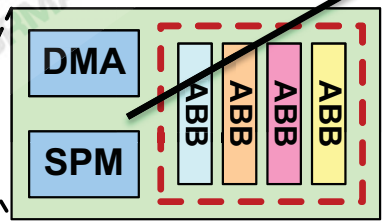
Auto-generated accelerators based on HLS (AutoPilot)




Router

C Core

$ LLC Banks




Dedicated/Composable Accelerators

M Memory Controller

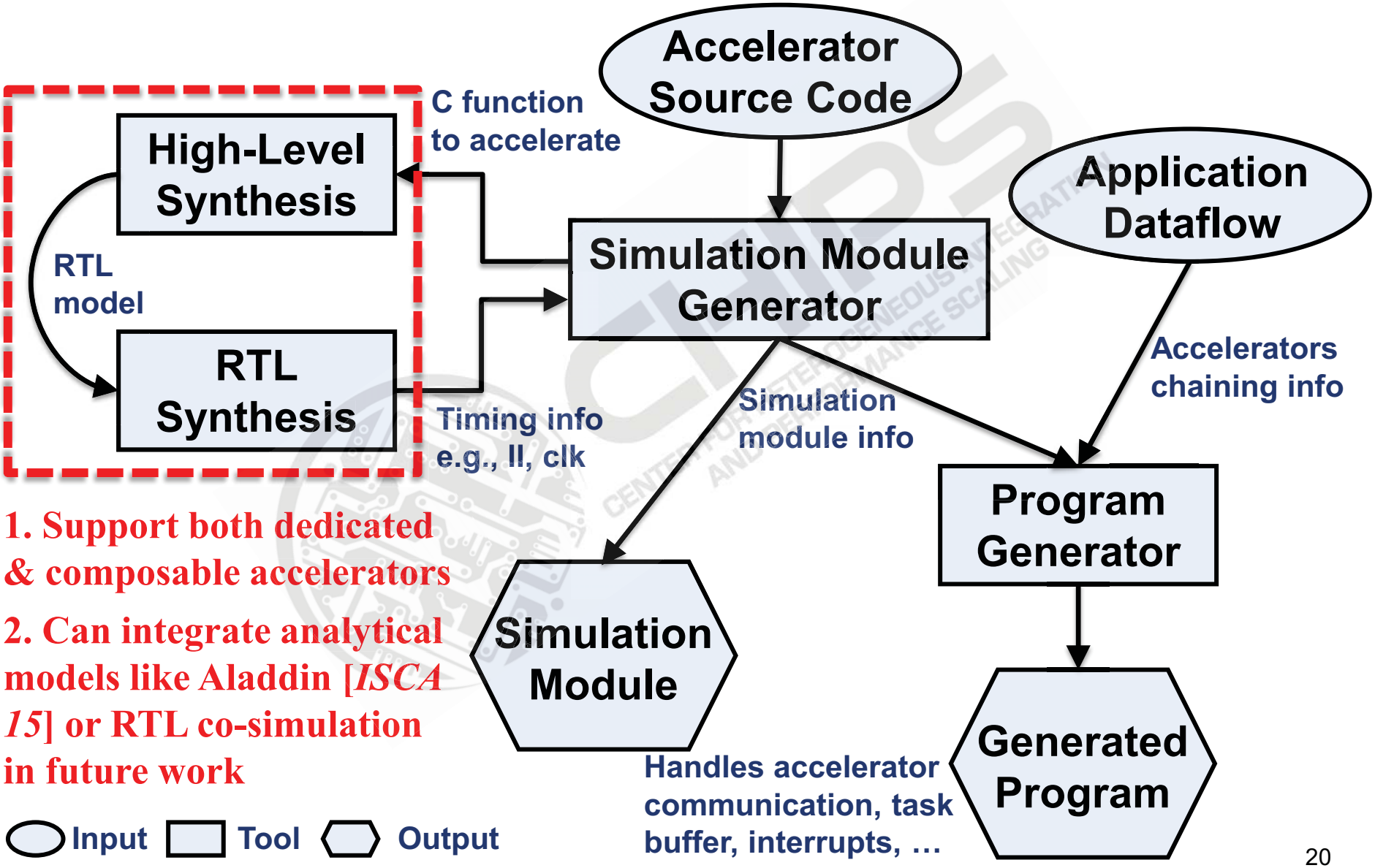
Added SPM, DMA, GAM & TLB model

Extended Garnet (DSENT) for NoC

Extended Ruby (CACTI) for coherent cache hierarchy

gem5 memory model [ISPASS 14]

# HLS-based Automatic Accelerator/App Generation




1. Support both dedicated & composable accelerators

2. Can integrate analytical models like Aladdin [*ISCA 15*] or RTL co-simulation in future work

# Customize Your Own Accelerator (e.g., Denoise)

Denoise core computation:

$$1 / \sqrt{\sum_{i=0}^{5} (x_c - x_i)^2}$$

Auto-generated accelerator for each ABB function

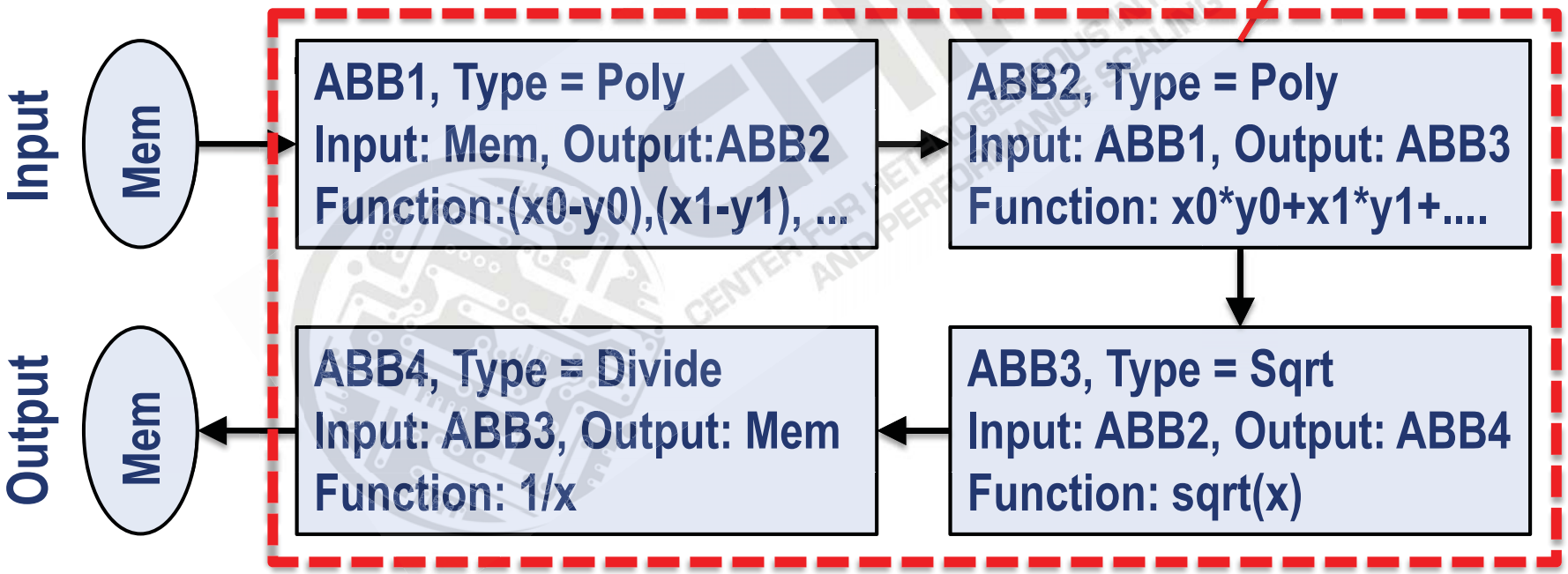
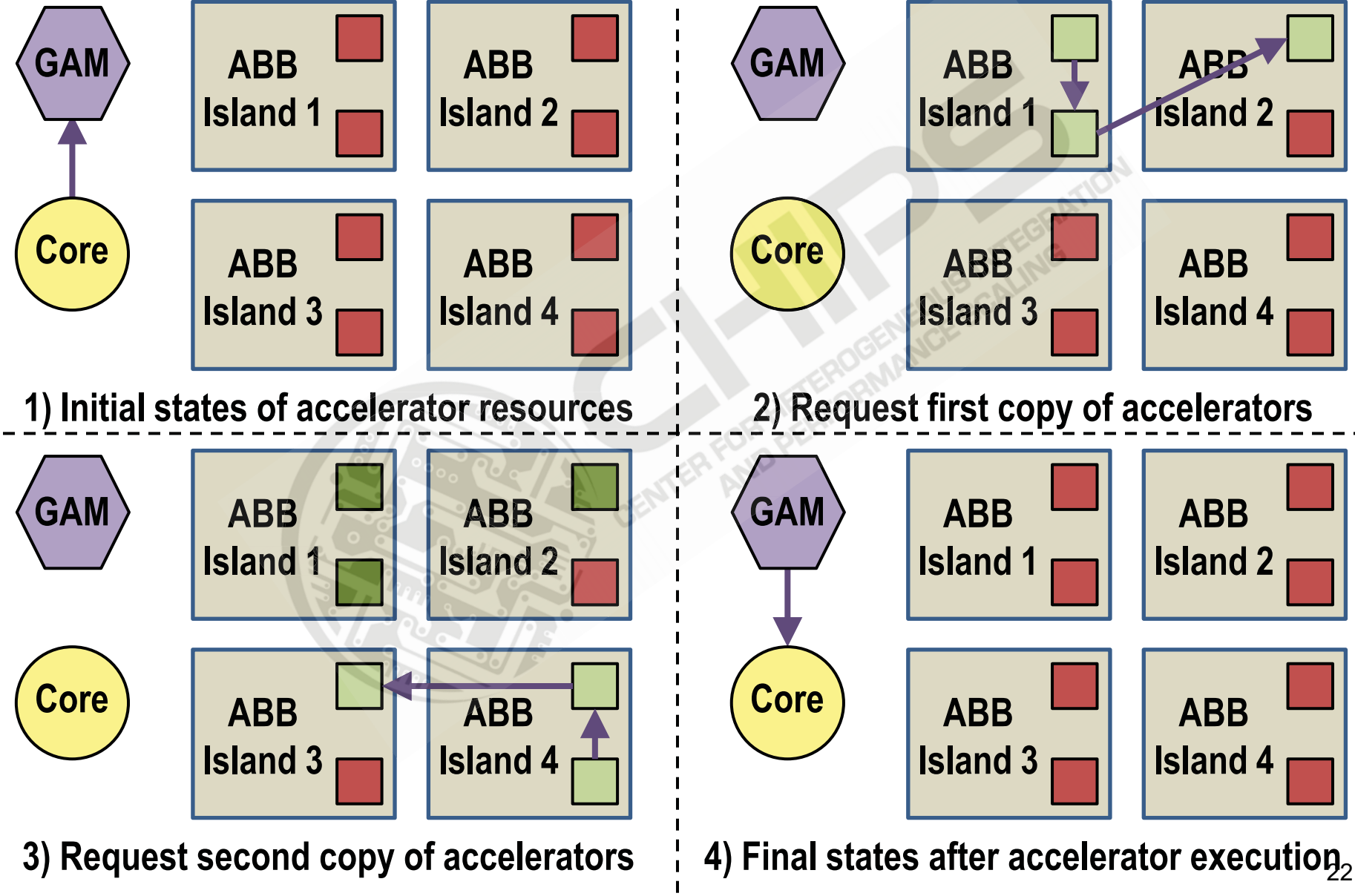



ABB: Accelerator Building Block

Auto-generated application using accelerator chaining data flow

# Global Accelerator Management (GAM)



## *Open Source Release*

***http://vast.cs.ucla.edu/software/parade-ara-simulator***

# Concluding Remarks

- New era of computing
  - Accelerator-centric computing
  - Need efficient support for customization and specialization
- Customization at all levels
  - Chip-level
  - Server node level
  - Data center level
  - New opportunites -- CHIPS level!
- Design automation support is key
  - Modeling
  - Simulation
  - Synthesis …

# 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)

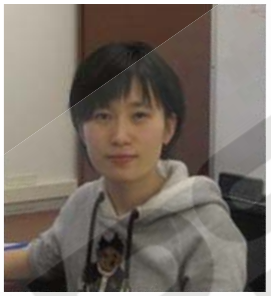

Sarkar (Associate Dir) (Rice)


Vese (UCLA)

# Postdocs, Graduate Students, and Collaborators

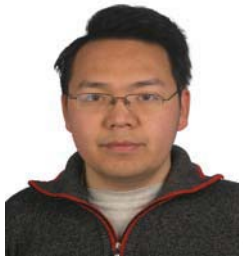
Prof. Deming Chen
(UIUC/ADSC)

Yuting Chen
(UCLA)



Hui Huang
(UCLA)

Muhuan Huang
(UCLA)

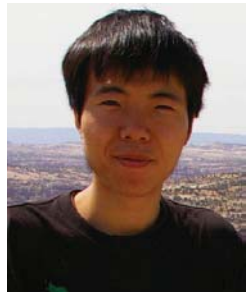


Dr. Peng Li
(UCLA)

Prof. Louis-Noël Pouchet
(UCLA)

Yuxin Wang
(PKU)

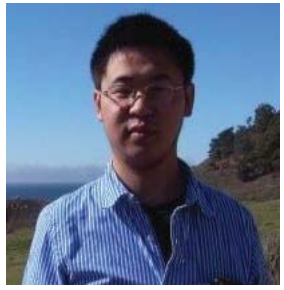
Di Wu
(UCLA)



Bingjun Xiao
(UCLA)

Hao Yu
(UCLA)

Dr. Peng Zhang
(UCLA)



Yi Zou
(UCLA)

Wei Zuo
(UIUC)