Relational Processing Accelerators: From Clouds to Memory Systems

Sudhakar Yalamanchili
School of Electrical and Computer Engineering
Georgia Institute of Technology

Collaborators: M. Gupta, C. Kersey, H. Kim, S. Mukhopadhyay, I. Saeed, S. H. Shon, J. Young, H. Wu, and LogicBlox Inc.

http://www.istc-cc.cmu.edu/

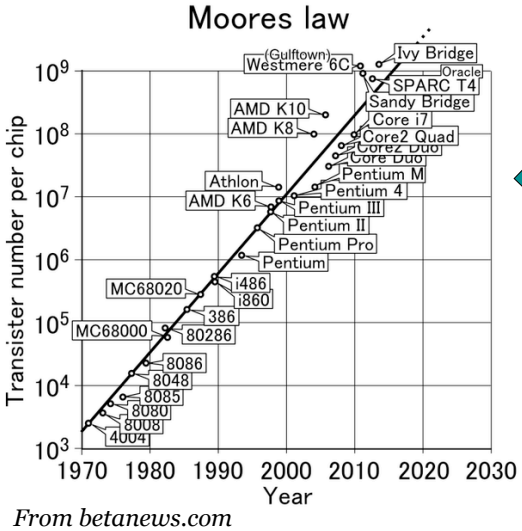
Intel Science & Technology Center for Cloud Computing

Overview

- Drivers
- High Performance Relational Computing
- Benchmark Repository
- Near Memory Processing

How are Technology and Applications Reshaping Systems?

Moore's Law and Dennard Scaling



From betanews.com

Goal: Sustain Performance Scaling

- Performance scaled with number of transistors
- Dennard scaling*: power scaled with feature size

*R. Dennard, et al., "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE Journal of Solid State Circuits*, vol. SC-9, no. 5, pp. 256-268, Oct. 1974.

Post-Dennard Performance Scaling

$$Perf\left(\frac{ops}{s}\right) = Power(W) \times Efficiency\left(\frac{ops}{joule}\right)$$

W. J. Dally, Keynote IITC 2012

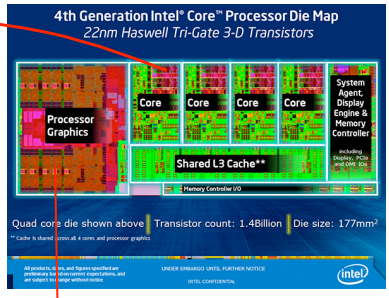
Memory Cost + Operator_cost + Data_movement_cost

Specialization → heterogeneity and asymmetry

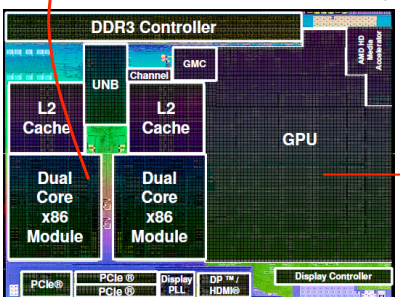
*S. Borkar and A. Chien, "The Future of Microprocessors, *CACM*, May 2011

Heterogeneous Architectures

Multithreaded Cores and Vector Units



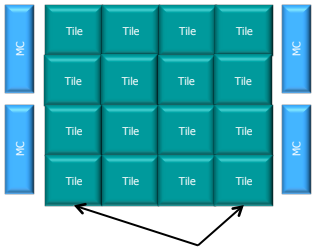
AMD Trinity



General Purpose Graphics Processing Unit (GPGPU) → Bulk Synchronous Parallel Model

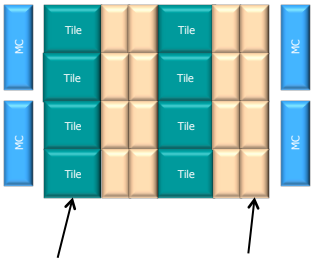
Asymmetric vs. Heterogeneous

Performance Asymmetry



Multiple voltage and frequency islands

Functional Asymmetry



Big Core (out-of-order)

Little Core (in-order)

Data Parallel/Vector

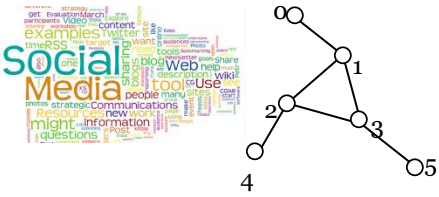
- *Single Instruction Set Architectures (ISA)*
- *Commodity software stacks*

Accelerated Systems

Financial Computing with FPGAs

From www.altera.com

GPU Accelerated Data Analytics



Micron's Automata Processor

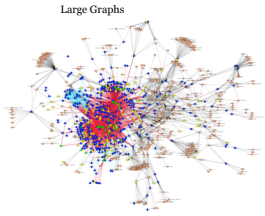
From www.micron.com

Microsoft Research: Web Search with FPGA Accelerators

From theguardian.com

A Data Rich World

Large Graphs



Mixed Modalities and levels of parallelism

Irregular, Unstructured Computations and Data

Pharma

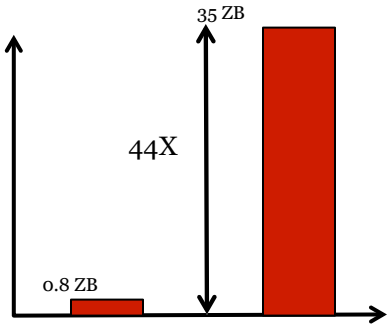
conventioninsider.com

Images from math.nist.gov, blog.thefuturescompany.com, michaelzdinler.blogspot.com

Trend analysis

The Digital Universe

"Between 2009 and 2020, the information in the Digital Universe will grow by a factor of **44**; the number of "files" in it to be managed will grow by a factor of **67**, and storage capacity will grow by a factor of **30**."

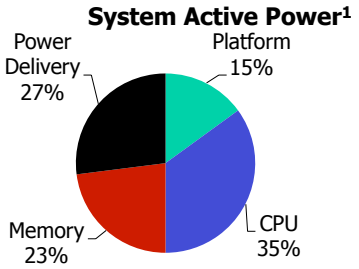


Data is consuming more space ($$)

J. Gantz, **"A Digital Universe Decade – Are You Ready?" ver. 4-26-2010**

Memory Power Costs – System Scale

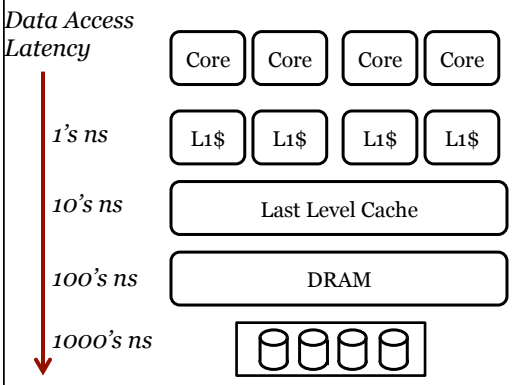
- Memory is consuming an increasingly large percentage of the processing node power
- Consider memory (power) costs for a full scale data center



Data is consuming more energy!

[1]H. David, et.al., "RAPL: Memory Power Estimation and Capping", *ISLPED 2010*.

Shift in the Balance Point



Operation	Energy (pJ)
64-bit integer operation	1
64-bit floating-point operation	20
256 bit on-die SRAM access	50
256 bit bus transfer (short)	26
256 bit bus transfer (1/2 die)	256
Off-die link (efficient)	500
256 bit bus transfer(across die)	1,000
DRAM read/write (512 bits)	16,000
HDD read/write	$O(10^6)$

28nm CMOS, DDR3

Courtesy Greg Astfalk, HP

- Relative costs of operations and memory accesses
  - Time and energy costs have shifted to data movement

Post-Dennard Performance Scaling

$$Perf\left(\frac{ops}{s}\right) = Power(W) \times Efficiency\left(\frac{ops}{joule}\right)$$

W. J. Dally, Keynote IITC 2012

Memory Cost + Operator_cost + Data_movement_cost

Specialization → heterogeneity and asymmetry

Three operands x 64 bits/operand

$$Energy^{*} = \#\,bits \times dist - mm \times energy - bit - mm$$

*S. Borkar and A. Chien, "The Future of Microprocessors, *CACM*, May 2011

Systems Should Be Designed to be Optimized for Data Usage

Overview

- Drivers
- High Performance Relational Computing
  - The software stack
  - Relational algebra and arithmetic kernels
- Benchmark Repository
- Near Memory Processing

Relational Queries and Data Analytics

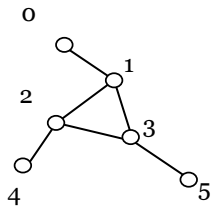
- The Opportunity
  - Significant potential data parallelism
  - High memory bandwidth and compute bandwidth of accelerators [1]
- The Problem
  - Need to process 1-50 TBs of data[2]
  - Fine grained computation
  - 15–90% of the total time spent in data movement [1] (for accelerators)

[1] B. He, M. Lu, K. Yang, R. Fang, N. K. Govindaraju, Q. Luo, and P. V. Sander. Relational query co-processing on graphics processors. In *TODS*, 2009.

[2] Independent Oracle Users Group. A New Dimension to Data Warehousing: 2011 *IOUG Data Warehousing Survey*.

High Performance Relational Computing

- Fraud detection
  - Looking for patterns in data sets

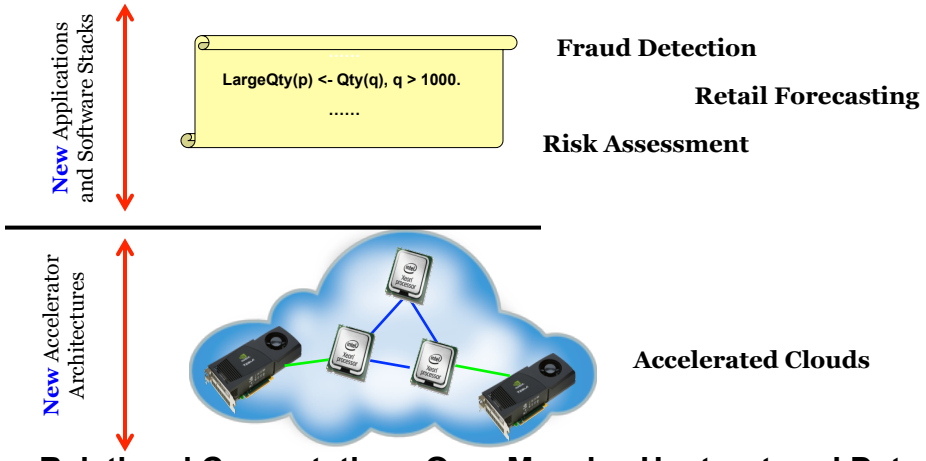


- Retail Forecasting
  - Processing billions of sales events each week

Prod#	A	C	D	E

- Risk Assessment
  - Insurance quotations for automobile and health

Red Fox: HPRC on Accelerator Clouds



Relational Computations Over Massive Unstructured Data

Goal: Sustain 10X – 100X throughput over multicore using GPU Accelerators

Red Fox: Goal and Strategy

GOAL
Build a compilation chain to bridge the semantic gap between ***Relational Queries*** and ***GPU*** execution models
10x-100X speedup for relational queries over multicore

Strategy

1. Optimized Design of Relational Algebra (RA) Operators
   1. Fast GPU RA primitive implementations (PPoPP2013)
   2. Multi-predicate Join Algorithm (ADMS2014)
2. Data Movement Optimizations (MICRO2012)
3. Query level compilation and optimizations (CGO2014)
4. Out of core computation

Source Language: LogiQL

- LogiQL: A declarative programming language based on Datalog
- Find more in http://www.logicblox.com/technology.html
- Examples

```
number(n)->int32 (n) .
number(0).
// other number facts elided for brevity
next(n,m)->int32(n), int32(m).
next(0,1).
// other next facts elided for brevity

even(n)-> int32(n).
even(0).
even(n)<-number(n),next(m,n),odd(m).

odd (n)->int32(n).
odd (n)<-next(m,n),even(m).
```

Recursive Definition

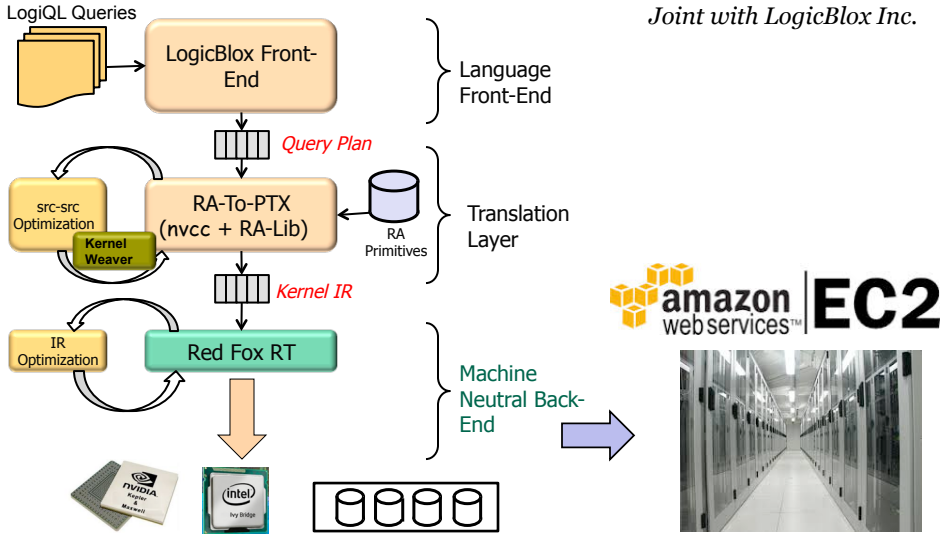
LogicBlox

Domain Specific Compilation

Haicheng Wu

Joint with LogicBlox Inc.

```
tr(x,y,z) ← E(x,y),E(y,z),E(x,z),x<y<z.
```



Red Fox TPC-H (SF=1) Comparison with Multicore CPU

Based on NVIDIA Geforce GTX Titan and LogicBlox v4.0 RT

- **On average (geo mean)**
- **GPU w/ PCIe : Parallel CPU = 11x**
- **GPU w/o PCIe : Parallel CPU = 15x**

Find latest performance and query plans in
http://gpuocelot.gatech.edu/projects/red-fox-a-compilation-environment-for-data-warehousing/

22

Multi-Predicate Join Algorithm

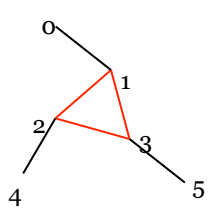
```
tr(x,y,z) ← E(x,y),E(y,z),E(x,z),x<y<z.
```

- Goal: Implementation of Leapfrog Triejoin (LFTJ) on GPU
  - A worst-case optimal multi-predicate join algorithm
  - Details (e.g., complexity analysis) in T. L. Veldhuizen, *ICDT 2014*
- Benefits
  - Smaller memory footprint and data movement
  - No data reorganization (e.g. sorting or rebuilding hash table) after changing join key
- Approach
  - CPU version
  - CPU-Friendly GPU version
  - Customized GPU version

Example: Graphs as Relations

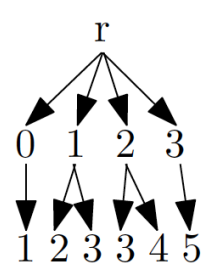
- Finding cliques
  - triangle(x,y,z)<-E(x,y),E(y,z),E(x,z), x<y<z.
  - 4cl(x,y,z,w)<-E(x,y),E(x,z),E(x,w),E(y,z),E(y,w),E(z,w), x<y<z<w.

Multi-predicate Join



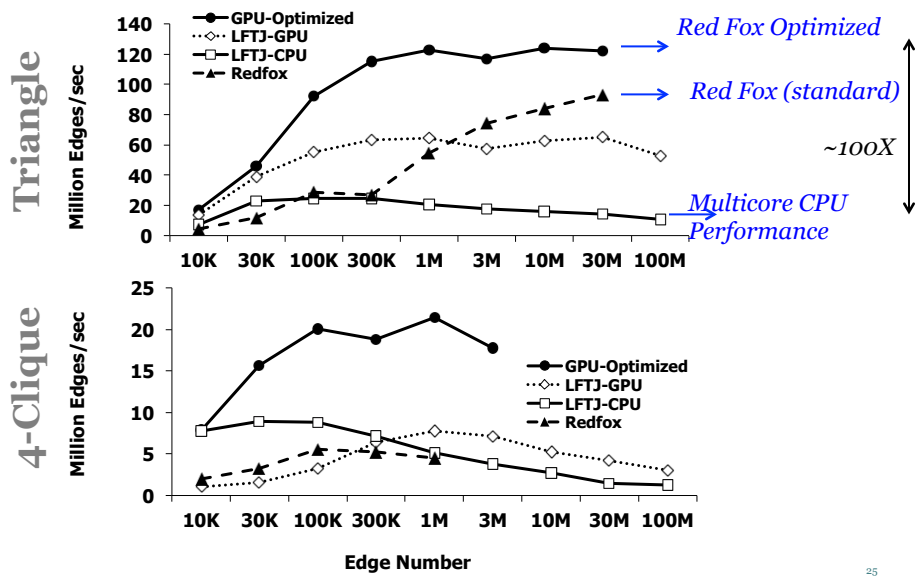
Edge:

From	To
0	1
1	2
1	3
2	3
2	4
3	5



H. Wu, D. Zinn, M. Aref, and S. Yalamanchili, "Multipredicate Join Algorithms for Accelerating Relational Graph Processing on GPUs," *Proceedings of ADMS*, September 2014

Performance

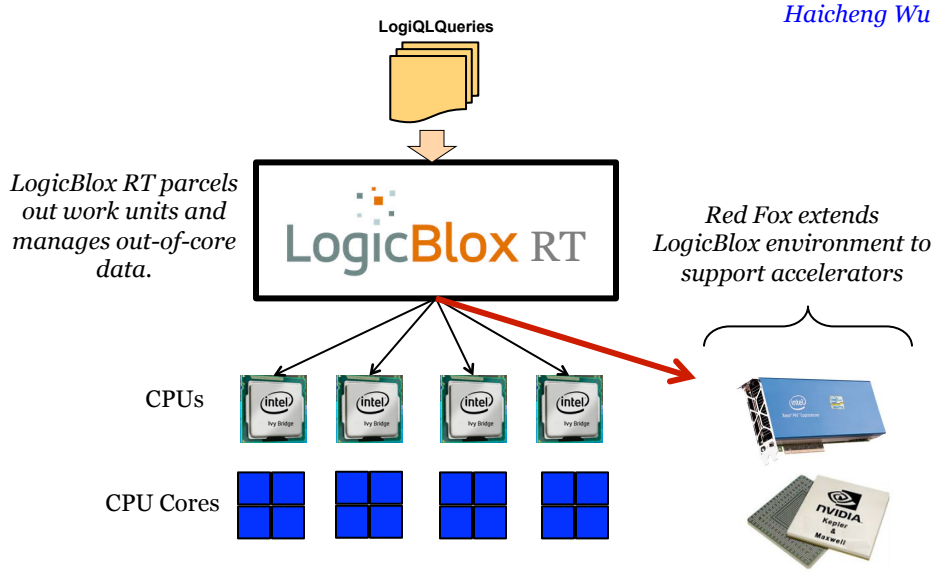


25

Algorithmic Innovation can Deliver Performance but Data Movement?

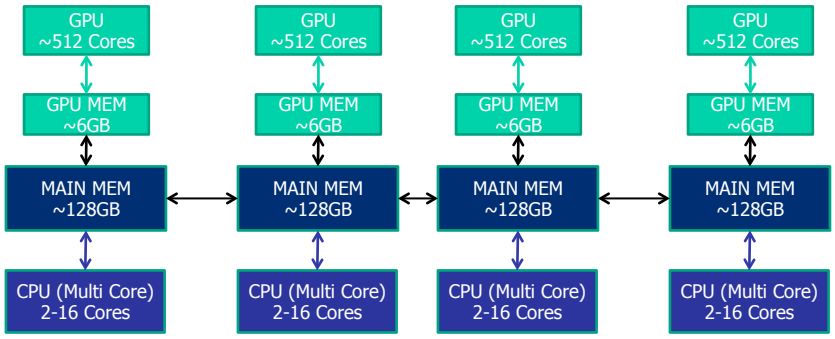
Getting to Out-of-Core Data Sets

Haicheng Wu



27

Memory Aggregation



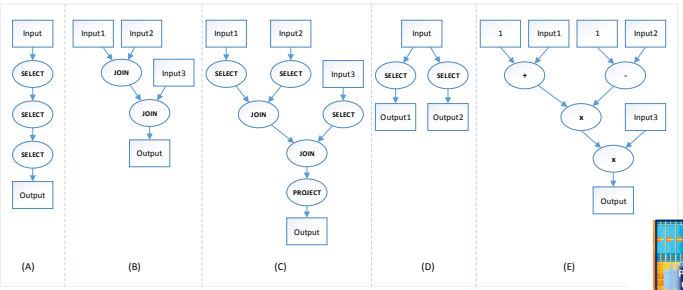
- Cluster-based memory aggregation
- Hardware support for global non-coherent, physical address space system
- Change the ratio of host-memory : GPU-memory
- Oncilla runtime (HPCC2012, Cluster2013)

Overview

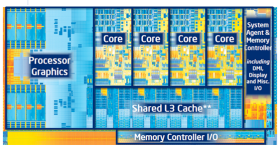
- Drivers
- High Performance Relational Computing
- Benchmark Repository
- Near Memory Processing

Common Patterns

The frequently occurring patterns of operators in the TPC-H benchmark suite



Multicore x86 CPUs, Gen, and Phi



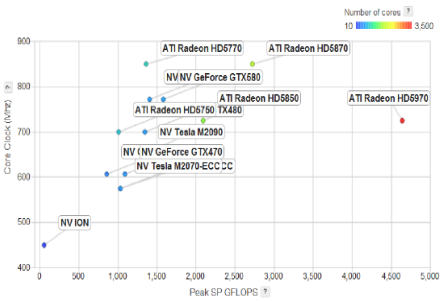
Relational Algebra Operators

- PROJECT
- SELECT
- INNER JOIN
- CROSS PRODUCT
- SET Family (SET INTERSECTION, SET UNION, SET DIFFERENCE)
- REDUCE
- REDUCE BY KEY
- UNIQUE
- SORT

The Scalable Heterogeneous Computing Benchmark Suite

Courtesy: Oak Ridge National Laboratories *Jeffery Young*

- Early focus on scientific computing workloads
- Kernels implemented in CUDA, OpenCL MIC port was developed in collaboration with Intel
  - System and stability tests
  - Multi-accelerator & cluster scale support



Max FLOPS Benchmark from SHOC

- Our current efforts → adding Red Fox kernels, TPC-H microbenchmarks, and TPC-H queries

-Danalis, et al, The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite, GPGPU '10
-https://github.com/vetter/shocOur

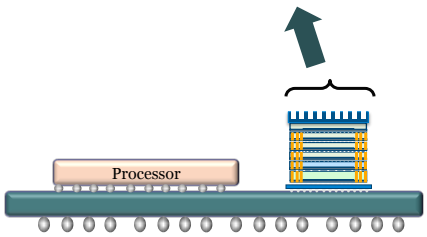
Overview

- Drivers
- High Performance Relational Computing
- Benchmark Repository
- Near Memory Processing

Near Memory Data Intensive Computing

Kim (CS), Mukhopadhyay (ECE), Yalamanchili (ECE)
Collaborative Discussions with Intel Labs (N. Carter)

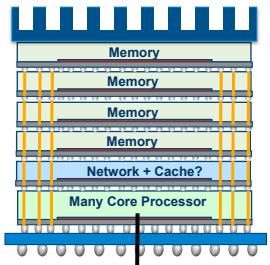
- Move Analytics Primitives (RA) into the memory system
  - Data movement optimization
  - Data locality optimizations
- Explore novel programming models and abstractions
- Explore novel compute and memory architectures
  - Memory consistency and coherency models
  - Integrated thermal and power management



Heterogeneous Architecture Research Prototype (HARP)

Chad Kersey
Meghana Gupta

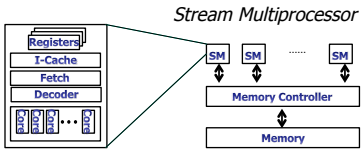
- Parametric, C++ processor generator environment
- Harmonica v2 in Altera FPGAs
- Assembler, emulator, and linker
- **OpenCL programming environment and Compiler** (in progress)



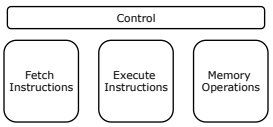
- RISC ISA
- C++ generator flow
- gcc compilation support
- Basic multicore/ multithreaded support
- Testing with cycle level simulators (in progress)

Platform Architecture

Stream Multiprocessor

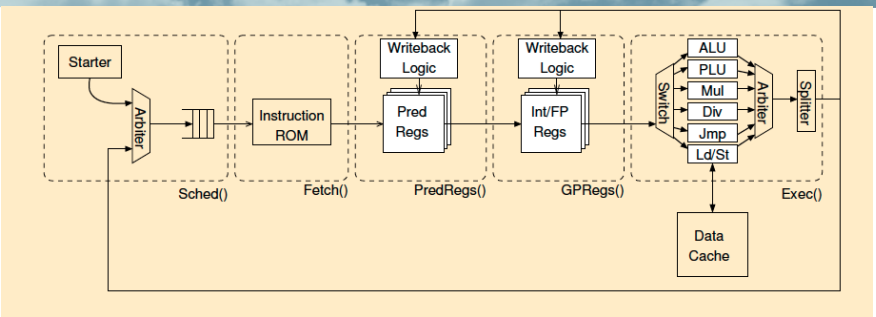


HARP Core



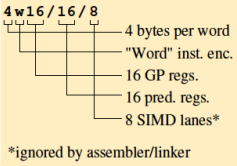
RISC Core

Heterogeneous Architecture Research Prototype (HARP)

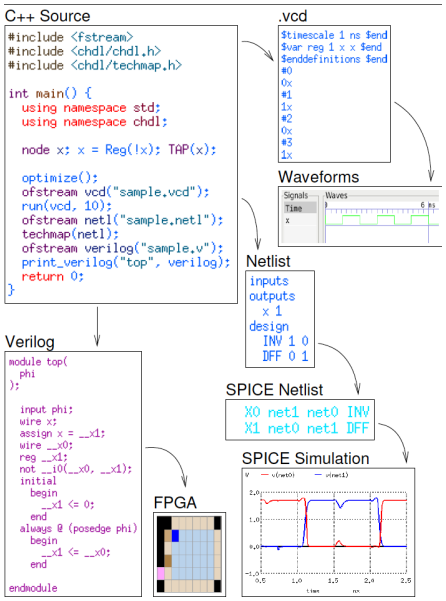


- Customizable, multithreaded, SIMD soft core
  - Generated from an architecture specification
- Supported by a generated HARP Tool assembler/ linker/emulator
- Small - ~1500 lines of C++

```
/* Return in r0 dot product of vectors
   of real values pointed to by r0 and
   r1, length in r2 */
dotprod: ldi %r3, #0;
dploop:  ld %r4, %r0, #0;
         ld %r6, %r1, #0;
         subi %r2, %r2, #1;
         addi %r0, %r0, __WORD;
         addi %r1, %r1, __WORD;
         rtop @p0, %r2;
         fmul %r4, %r4, %r6;
         fadd %r3, %r3, %r4;
  @p0 ? jmpi dploop;
         ori %r0, %r3, #0;
         jmpr %r5;
```

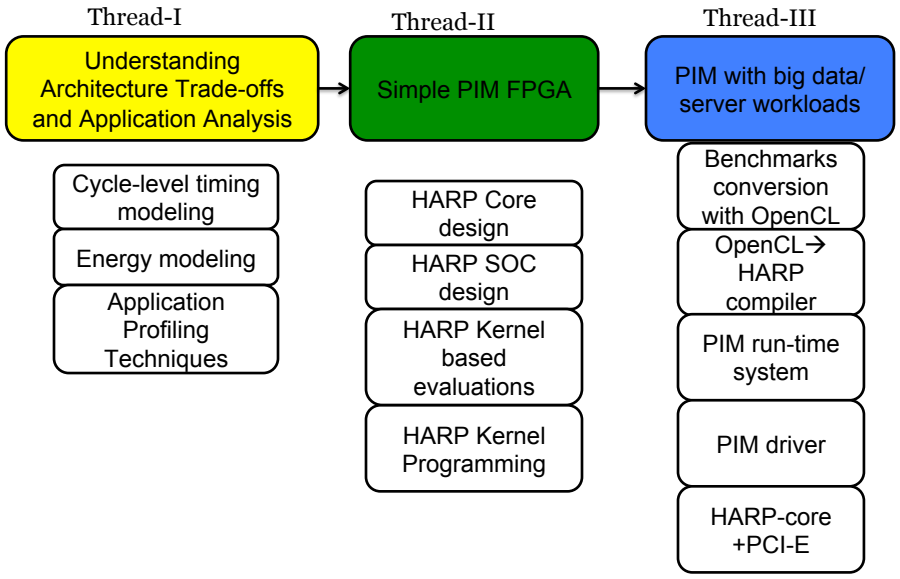


CHDL Tool Chain

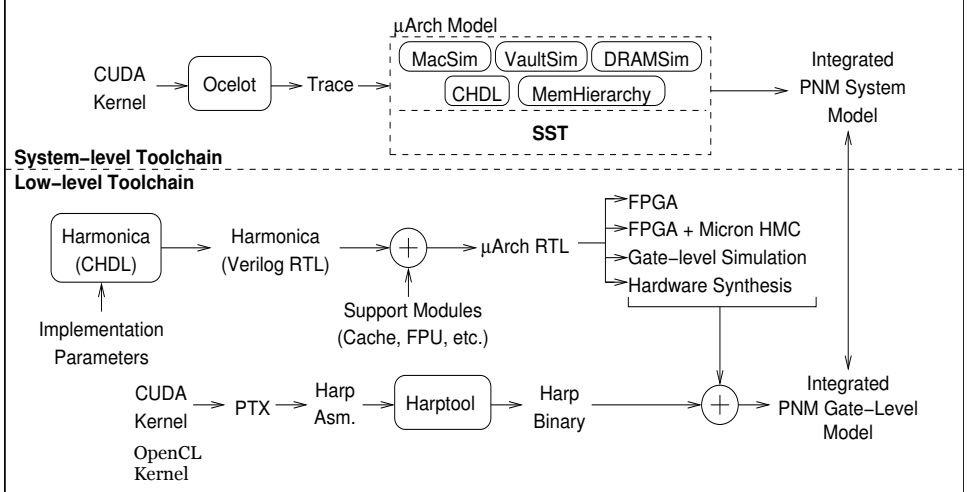


- Vertically integrated CAD environment
- Strong emphasis on code reusability
- Same model for generating both gate level and system level simulations
- CAD tool interfaces
- v2 running in Altera FPGAs

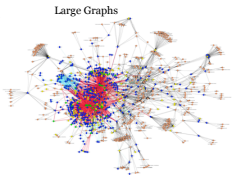
Project Elements



Modeling Flow



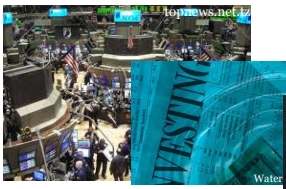
System Model: Summary

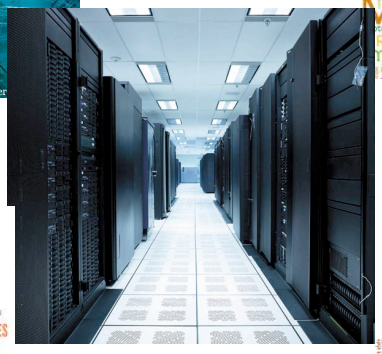


Programming Models	*Domain Specific Languages*
Data Movement Optimizations	*Compiler and Run-Time Support*
System Abstractions e.g. GAS, Virtual DIMMs, etc	*Cluster Wide Hardware Consolidation*

Hardware Customization

The Future is Acceleration





Thank You