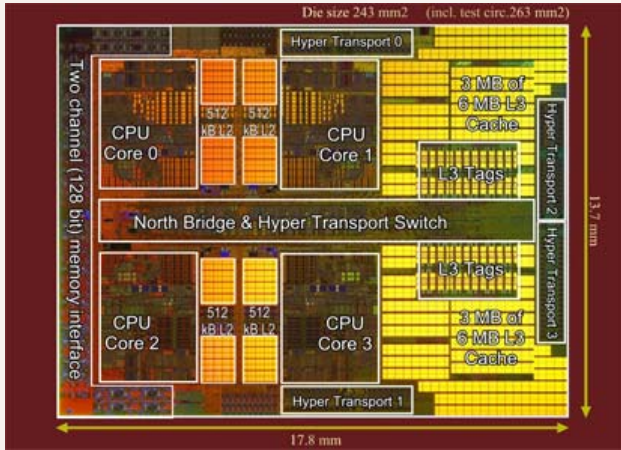# Programming Models for Accelerator Clouds
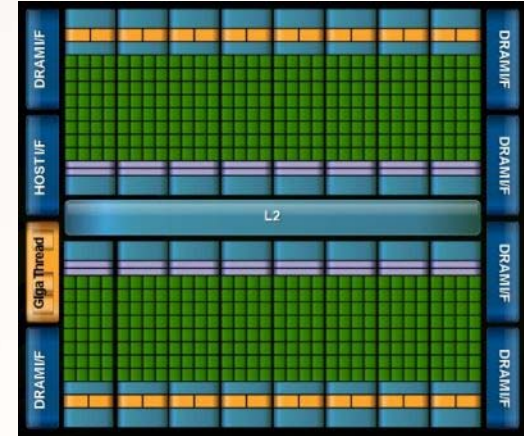
Nate Clark, Sudha Yalamanchili, Greg Diamos,
Andrew Kerr, Haicheng Wu, Kirak Hong,
and many people at LogicBlox

cercs

Georgia Tech | College of Computing
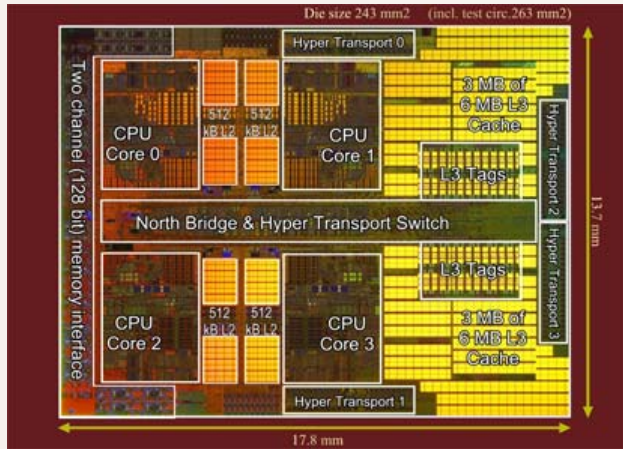
What / Why Accelerator Clouds?

AMD Shanghai

NVIDIA Fermi

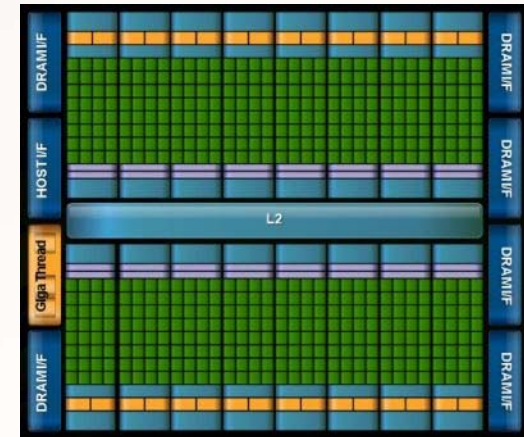- Power wall → heterogeneity
- Compute / $

Problem: Programming

MIMD Coherent Shared Memory

SIMD Data Parallel

AMD Shanghai

NVIDIA Fermi

- Productivity
- Performance/system portability

A PARALLEL CHANNEL COMPUTING MACHINE

Lecture by
J. P. Eckert, Jr.
Electronic Control Company

... Again I wish to reiterate the point that all the arguments for parallel operation are only valid provided one applies them to the steps which the built in or wired in programming of the machine operates. Any steps which are programmed by the operator, who sets up the machine, should be set up only in a serial fashion. It has been shown over and over again that any departure from this procedure results in a system which is much too complicated to use.
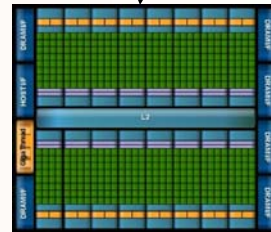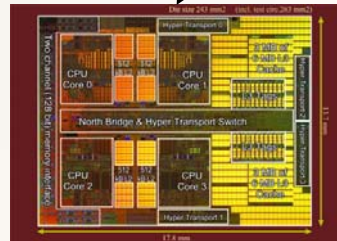
Portability Solution

- Java, Fortran, MapReduce, etc.
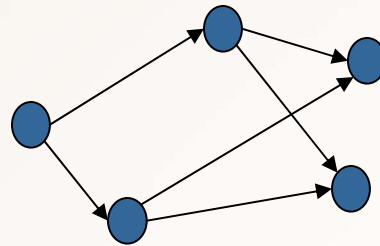
-Side effect free kernels
-Explicit dependencies
-Enables aggressive optimization

Harmony Runtime

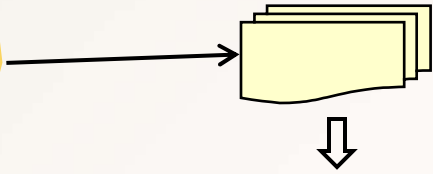**Georgia Tech** | College of Computing

Productivity: LogicBlox's Datalog

- Declarative language (not imperative)
  - Think high-powered Excel

- Many examples of business uses
  - Risk analysis
 - Business analytics

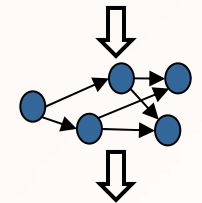- (Usually) maps well to kernel IR

Current Status

Application Development

Legend
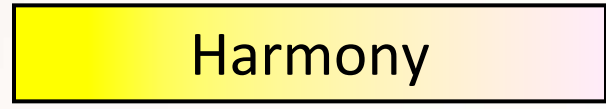- **Open Source available** (yellow)
- **In progress** (orange)
- **Commodity SW/GT** (purple)

Datalog Front-End — Translate to Harmony Kernel IR

Kernel IR

Harmony — **Harmony**: Run-Time

Ocelot: Code Generation and Dynamic Optimization Environment

Ocelot/LLVM | NVIDIA JIT

OS/VM

Georgia Tech | College of Computing

Risk Analysis Application

GPU interaction overhead dominates

Harmony distributes the work across 4 CPU cores, 2 slow GPU boards, and 1 fast GPU board

Measured execution times

82 min
20 min
6 min
3 min

Legend:
- AMD Phenom 9550 (C)
- 9800 GTX (CUDA)
- 280 GTX (CUDA)
- Complete System (Harmony)
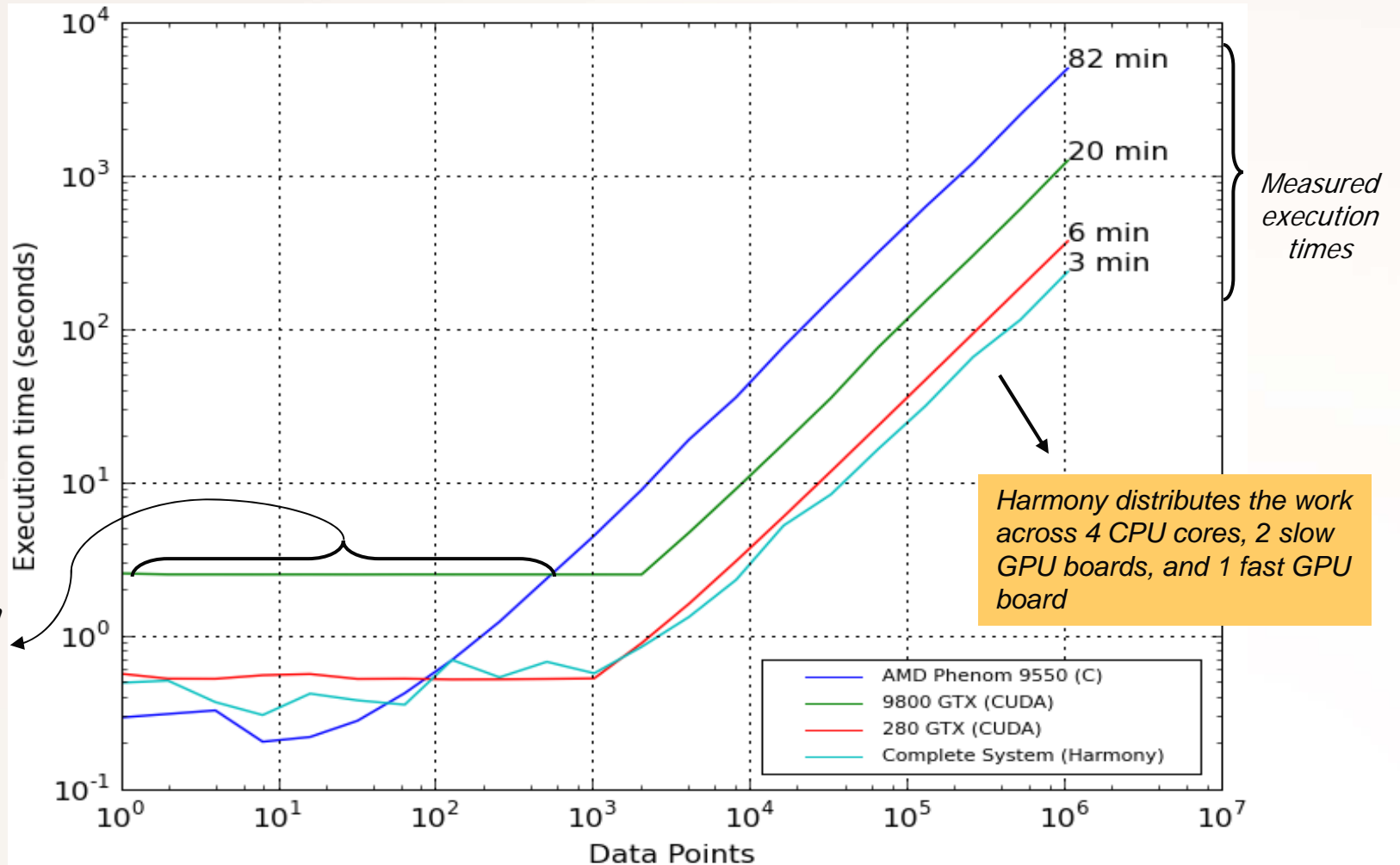
Conclusion

- Preliminary results look good

- Interaction benefits both LB and GT

- Open research
  - Scaling, scaling, scaling
  - Further optimizations
  - Other programming paradigms?