Accelerating Bioinformatics with Distributed Reconfigurable Systems

Stuart Byma, Sam Whitlock, Edouard Bugnion, Christos Kozyrakis, James Larus

Motivation

- Bioinformatics -- Big data, big compute
  - Whole genome sequencing processing takes ~hours per person!
  - > 300GB raw data per person!

- Approach -- Accelerate, parallelize
  - Spread computation across many nodes
  - Utilize FPGA-based acceleration
  - Create a unified programming framework and data management strategy

Bioinformatics

- Analysis and interpretation of biological data

  Example: Sequence Alignment
  - Millions of small "reads" produced by DNA sequencing machines
  - Alignment reconstructs genome from reads using reference

Field-Programmable Gate Arrays (FPGA)

- Configurable Logic Blocks
- Programmable Routing Fabric
- Hard Blocks -- Memories, DSPs, I/O controls

Microsoft Catapult

- Reconfigurable fabric for datacenter services
- One FPGA / server -- parallel compute fabric
- Less than 10% additional power

Catapult: Bing Ranking Acceleration

- Seven stages across 8 FPGAs
- 2X system performance increase!

Challenge - Data Management

- High FPGA streaming throughput
- Legacy formats inefficient for streaming
  - Plain-text records
  - Row storage format
- New dense column storage format
  - Binary format for bases
  - Streaming Compression
  - Made for parallel storage I/O

Challenge - Programming

- Need a high-level interface for easy programming
- Use dataflow
  - Kernels distributed across devices
  - Intelligently balance for max throughput / low latency

Spec

- Read data
- Global Alignment
- Local Alignment
- Output

Implementation

- Distribute
- Parallelize
- Optimize