nano-tera.ch
swiss scientific initiative in health / security / environment systems

YINS RTD 2013

FNSNF

Accelerating Bioinformatics with Distributed Reconfigurable Systems

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

vlsc

DCSL

SAIL

Motivation

- Bioinformatics -- Big data, big compute
  - Whole genome sequencing processing takes ~hours per person!
  - > 300GB raw data per sequencing!
- Approach -- Accelerate, parallelize
  - Spread computation across many nodes
  - Utilize FPGA-based acceleration
  - Create a **unified programming framework** and **data management strategy**

`...ATCG...` (100-150BP)

X 100,000,000

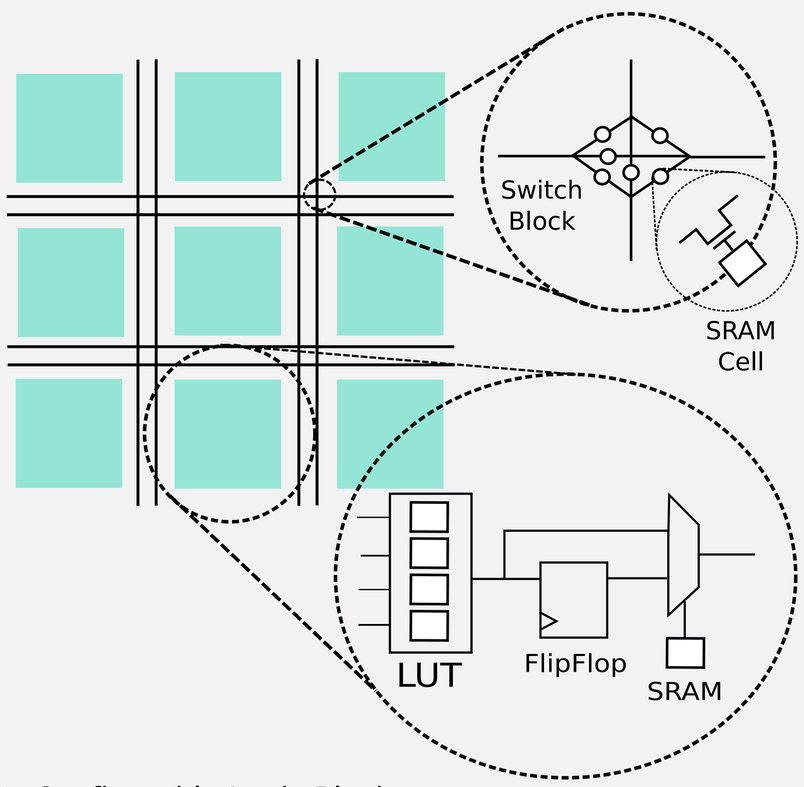
300GB - 1TB!

Context: 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

```
GGGCGGGGCCGGGAGGGGCGGGGCCGC
::  :::::::::::::::::::: : :
GG-CGGGGCCGGGAGGGGCGGGGGCCC
```

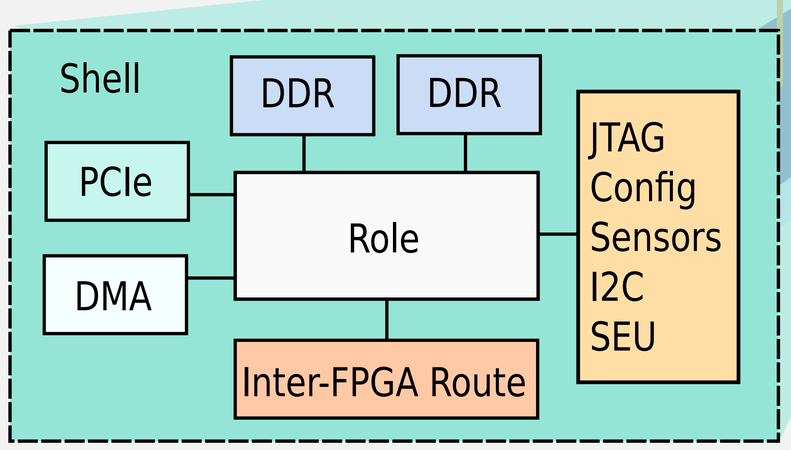
Field-Programmable Gate Arrays (FPGA)

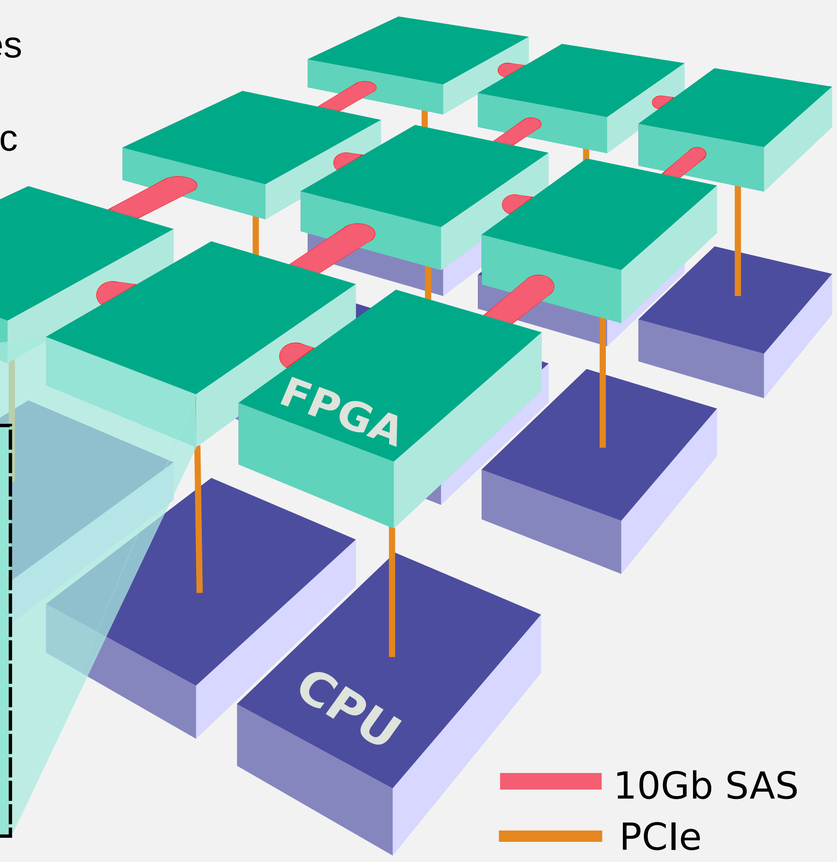


- Configurable Logic Blocks
- Programmable Routing Fabric
- Hard Blocks -- Memories, DSPs, I/O controls

Platform: Microsoft Catapult

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





Catapult bing

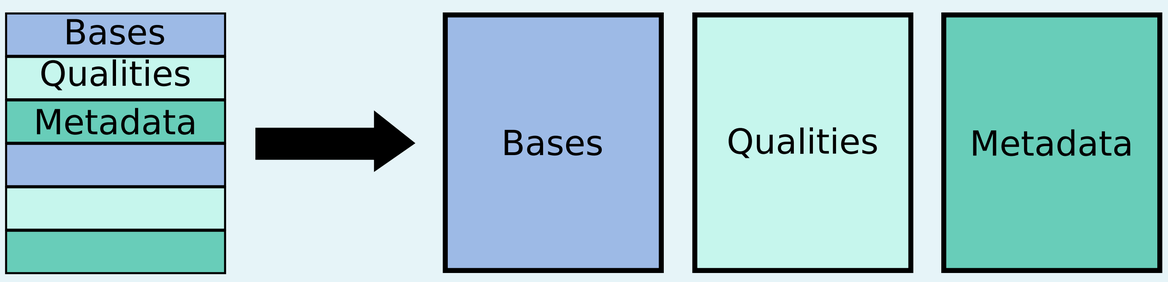
- Seven stages across 8 FPGAs

Ranking Acceleration

- 2X system performance increase!

Challenge - Data Management

- High FPGA streaming thoughput
- 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 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