**Large-Scale Reconfigurable Systems with DataFlow**

**Stuart Byma, James Larus**

### Motivation
- Emerging datacenter architectures
  - MS Catapult
  - CPUs, reconfigurable HW (FPGAs)
  - Unified "flat" architecture
- Can program CPUs, single FPGAs, but:
  - How to link these cohesively?
  - Move data across (many) devices?
  - Reason about system as a whole?
- We need a unified programming framework

### System-Level Dataflow
- Coarse-grain compute kernels
  - Different device types, bits and bitstreams
- User-defined data / types
  - Cross-device compatible
- Data flow graph = system "program"
  - Specified with high-level API or language

![Diagram of system-level dataflow](image)

```
in_q = resources.fifo_queue()
out_q = resources.fifo_queue()
enqueue = ops.enqueue(in_q, [in_files])

with device("/machine:0"):
dqd = ops.dequeue(in_q)
r0 = ops.someCompute(dqd)

with device("/machine:1"):
dqd = ops.dequeue(in_q)
r1 = ops.someFPGAcompute(dqd)

with device("/machine:2"):
enq0 = ops.enqueue(out_q, r0)
enq1 = ops.enqueue(out_q, r1)
out = ops.dequeue(out_q)

system.run(enq, enq0, enq1)
result = system.run(out)
```

### Why Dataflow?
- Dataflow makes things more explicit:
  - Concurrent execution, multi-device
  - Queuing
    - Separate system into subgraphs
    - Easily reason about queuing behavior
  - Specification of shared resources
    - Network? An implicit queue
  - Total system architecture
    - Contained in several files/graphs
  - Other benefits:
    - Modularization, reusability
    - Simple deploy and run

### Application: Bioinformatics
- Genome Sequence Alignment
- Challenge: Big data, big compute

![Diagram of genome sequence alignment](image)

1. File formats and data access
   - Legacy formats row-based ASCII
   - New format: column store, parallel access, binary format
   - Chunked for efficient disk/network access

2. Distributed dataflow execution

**Next Steps**
- Sequence alignment goal
  - WGS alignment in < 1 minute
  - From disk or network storage
- Bioinformatics continued
  - Downstream analysis pipelines
- System Dataflow implementation based on TensorFlow
  - Deployable on MS Catapult datacenter