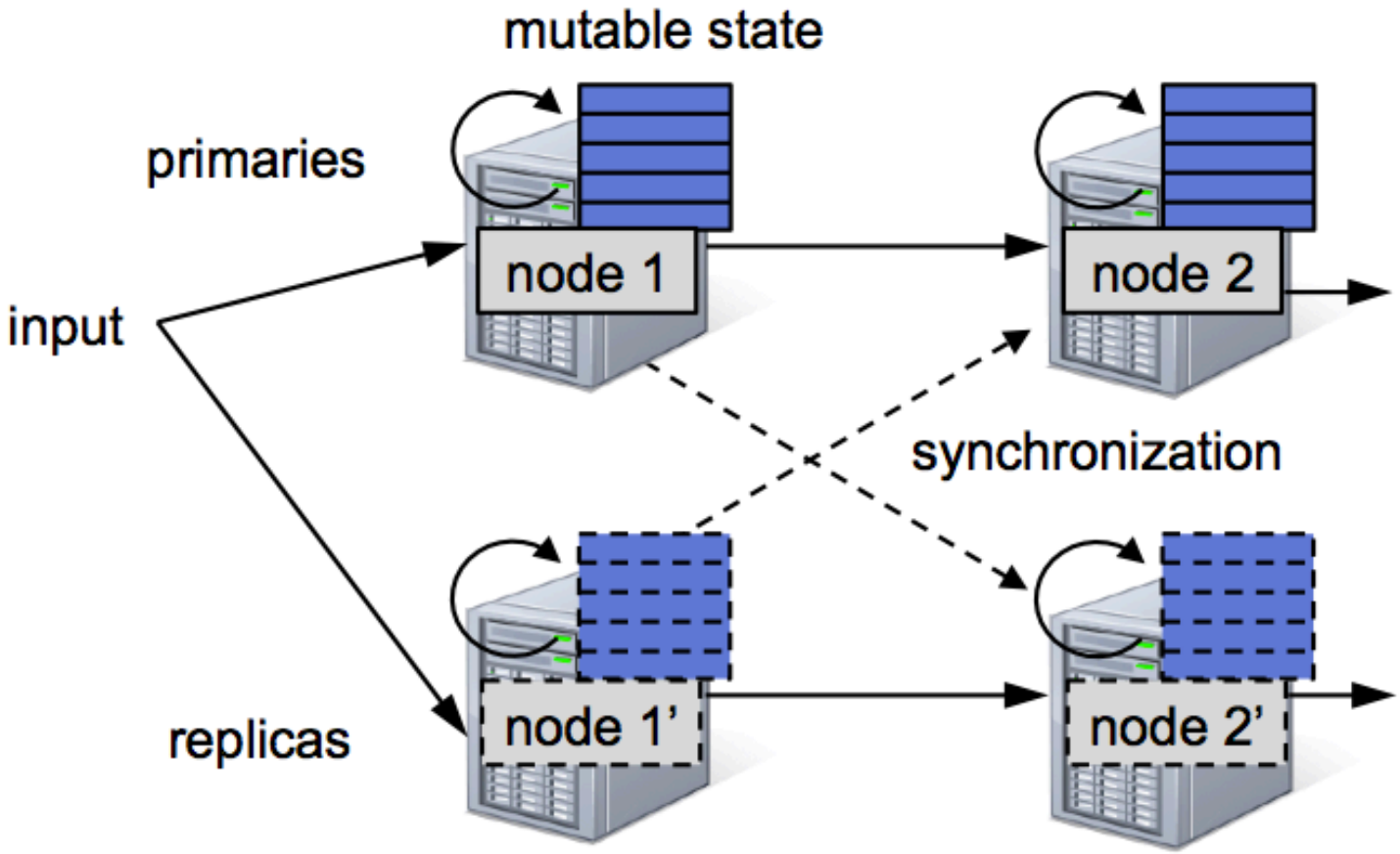
Discretized Streams: Fault-Tolerant Streaming Computation at Scale

Lianghong Xu

CMU 15-799 ligthning talk

Matei Zaharia etc., SOSP'13

Continuous Processing Model

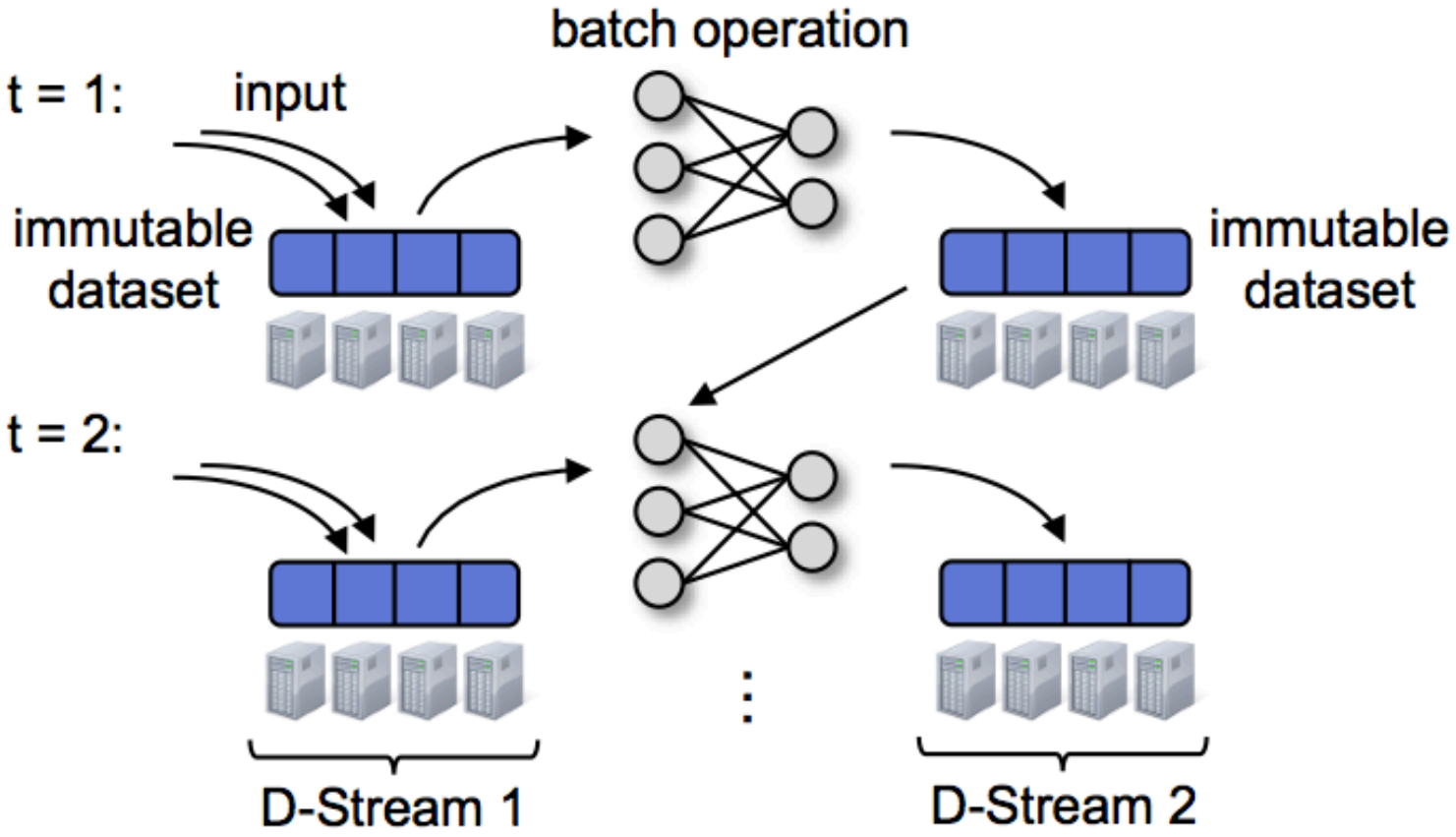


Problem: hot standby or long recovery times

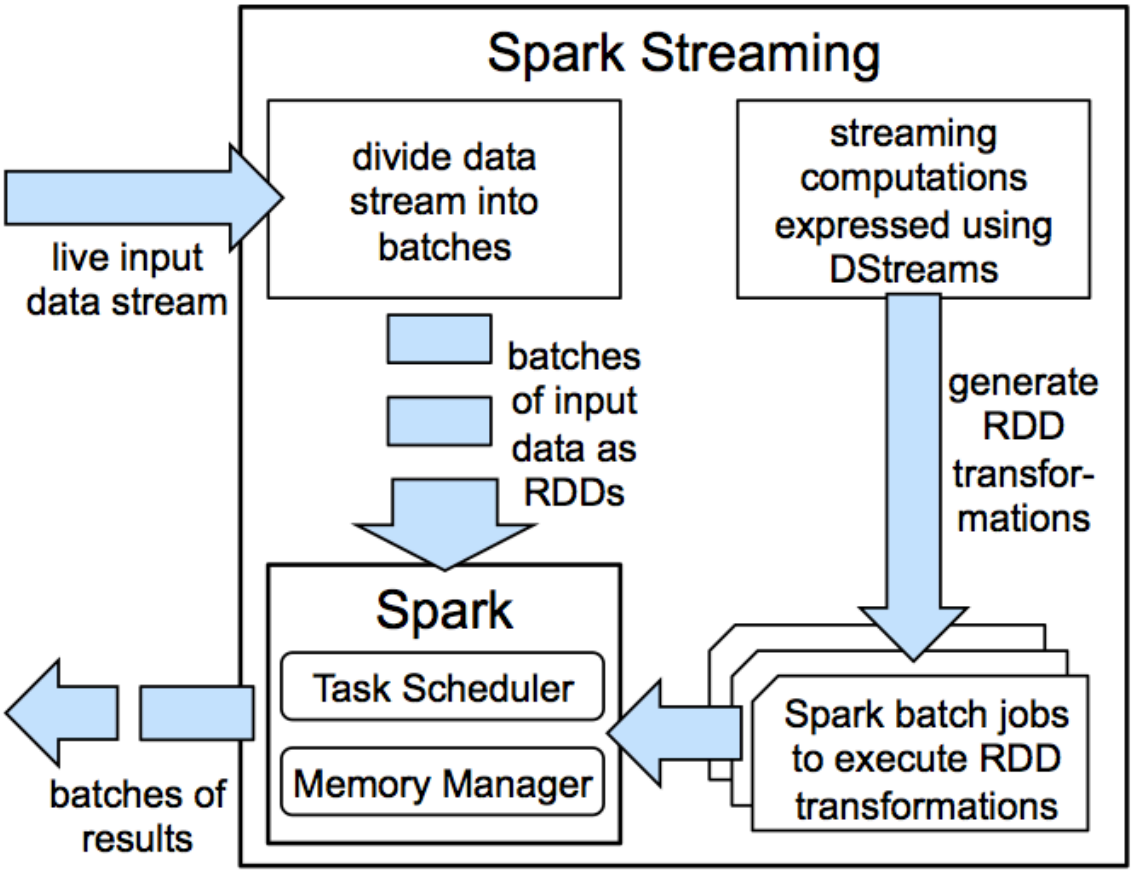
Discretized Streams (D-Streams)

- Same execution model as batch processing
- A series of *short*, *stateless*, *deterministic* batch computations
  - No long-lived operators
- Fast parallel recovery
  - RDD lineage graphs, asynchronous checkpointing
- Handle stragglers
  - Fine-grained speculative execution
- Linear scaling to 100 nodes

D-Stream Processing Model



Spark Streaming Overview



Performance

- 64M records/s for Grep (100 nodes)
- 25M records/s for TopKCount and others
- Comparable to commercial products on per-node performance
  - But linearly scales to 100 nodes
- Significantly outperforms S4 and Storm

Open source at http://spark-project.org