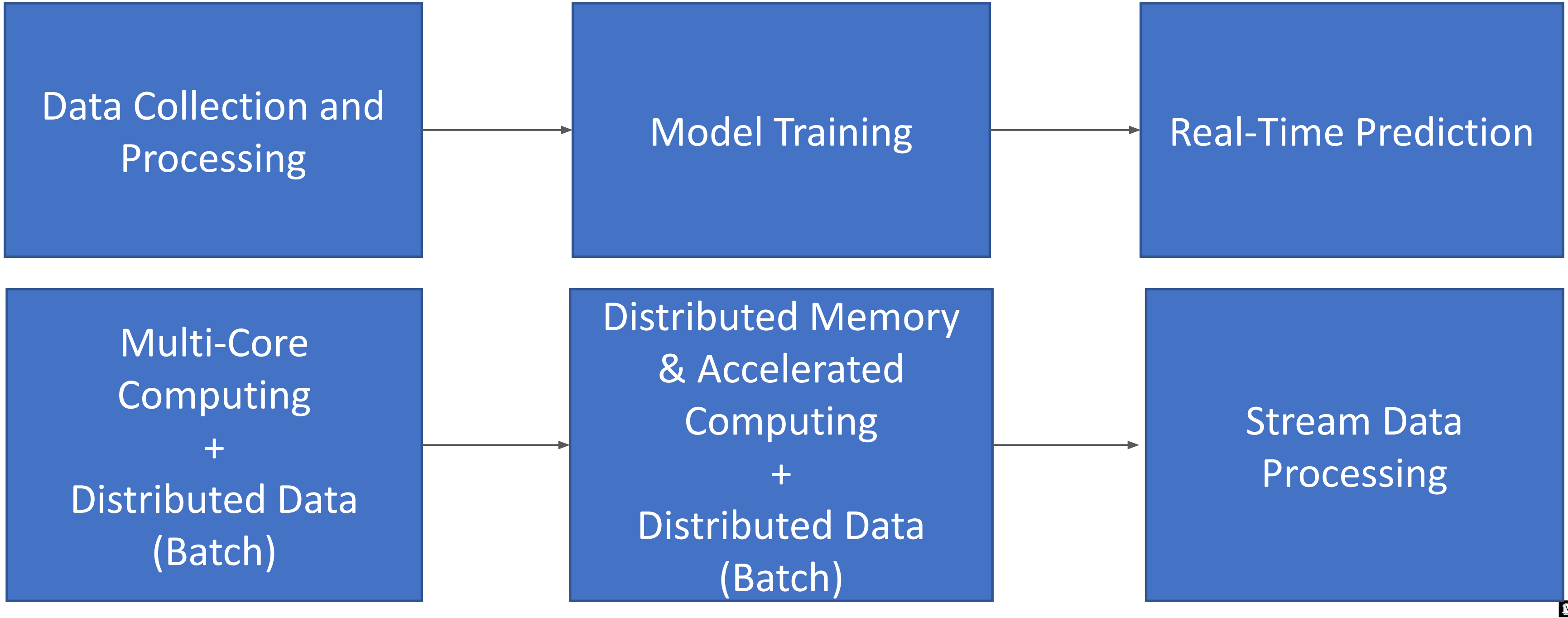
Harvard John A. Paulson
School of Engineering
and Applied Sciences

Parallelization of stock prediction

Group 11: Kevin Hare, Junkai Ong, Sivananda Rajananda

Project Phases



Data Collection and Processing

Model Training

Real-Time Prediction

Multi-Core Computing
+
Distributed Data (Batch)

Distributed Memory & Accelerated Computing
+
Distributed Data (Batch)

Stream Data Processing

Data collection & processing

$$O(S * D * N)$$

where:

- S = # of securities
- D = # of days of trading data
- N = Sequence per day (roughly 325/day)

Data collection & processing

Overheads:

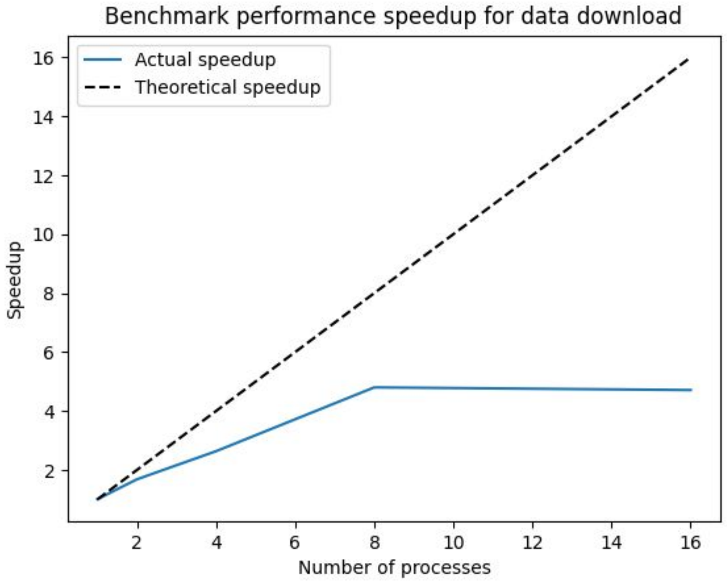
- Communication with API & download of data
- Conversion to sequence is security-specific

Mitigation strategy:

- Data download is embarrassingly parallel
- Python *multiprocessing* module (bound by the cores available)

Data collection & processing

- Theoretical speedup:
  - *v*, number of cores
- Est. serial processing time: 40 min.
- Est. parallel processing time: 9 min.



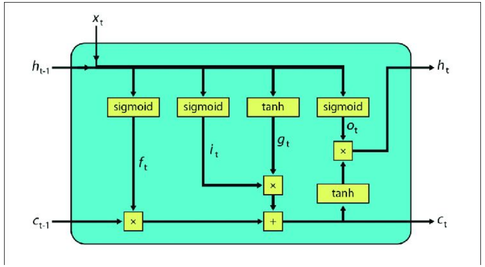
Model Training

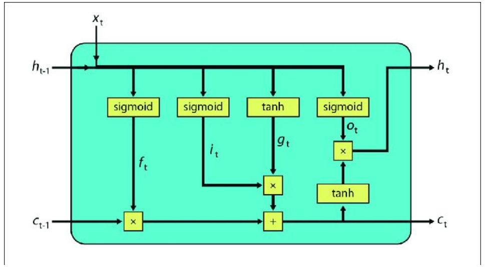
- The success of our LSTM training hinges on the efficiency of processing massive amounts of computation
  - matrix multiplication

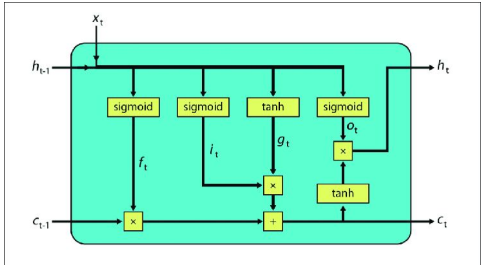
- The computational parallelism in such a graph can be characterized by two main parameters:
  - **the graph’s work W**, which corresponds to the total number of vertices
  - **graph’s depth D**, which is the number of vertices on any longest path

Model Training - Theoretical Speed-Up

- Assuming one operation per processor per unit time:
  - the execution time of such a DAG on p processors is bounded by: max{W/p, D} ≤ Tp ≤ O(W/p + D)
  - LSTM being a sequential model is limited mostly by D







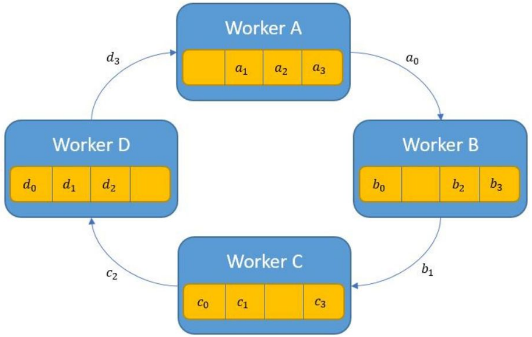
Model Training - Implementation

- Parallel computation within batch
  - Keras LSTM cell, implemented with CuDNN
  - NVIDIA M60 GPU
  - Est. 6x speedup (NVIDIA, 2021)

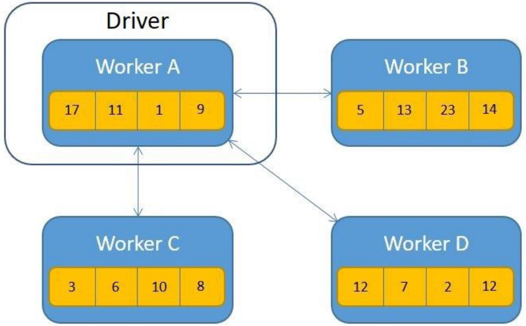
- Parallelize training between batches
  - Horovod
  - Theoretical speedup: p, for p processes
  - Limited by communication overhead

Ring-AllReduce vs Parameter Server

- Bandwidth-optimal message-passing algorithm
- Phases: (1) share-reduce, and (2) share
- Theoretical reduction in complexity (Ring-AllReduce) = p

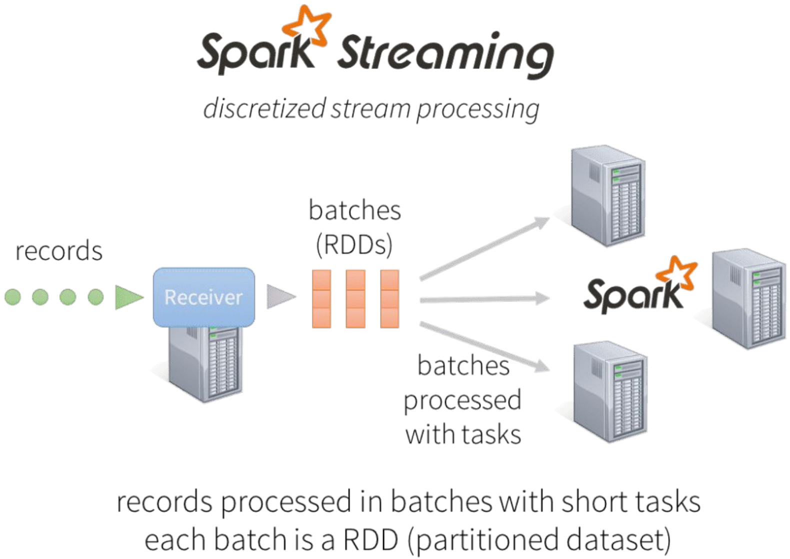


Ring-AllReduce Model



Parameter Server Model

Real-Time Prediction


Spark Streaming
discretized stream processing
records
Receiver
batches (RDDs)
Spark
batches processed with tasks
records processed in batches with short tasks
each batch is a RDD (partitioned dataset)

- Parallelization of real-time processing & prediction
- Scalability to process and predict every minute

Source: https://databricks.com/blog/2015/07/30/diving-into-apache-spark-streamings-execution-model.html

Infrastructure (AWS)

Data Processing:

- t2.2xlarge instances

Model Training

- g3.4xlarge instances (NVIDIA Tesla M60 GPU)
  - Horovod
  - OpenMPI

Prediction:

- g3.4xlarge instances

Storage:

- S3 Standard