CS 267, Describe a Parallel Application
Distributed Time Series Analysis with Apache Spark

Francois Belletti
January 28, 2016

Bio:
Passionate about time series analysis and distributed computing, I have graduated with a MS in stochastic calculus specialized in time series analysis for finance from Ecole Polytechnique - France, and with a MS in distributed system from Imperial College London - UK. I started my PhD in EECS so as to blend my two passions and today I work for Amplab on designing and implemented SparkTimeSeries, a scalable time series analysis module for Spark.

Project:
Communication avoidance has been at the core of the speedups in the project so far. The basic linear correlation and model analytics have been implemented for that system and new exciting research is on the way. In particular, overlapping cached streaming feeds offer a compelling opportunity to provide cross-correlation analytics on the fly and to recalibrate new models efficiently as new data feeds into the system. For geophysical system, partitioning of data with respect to space also provides a significant reduction of the need for communication between distributed nodes. Indeed, identifying exactly what part of the data is necessary to compute meaningful local previsions gives a compelling opportunity to reduce latency in computations. Applications of these techniques are many but I am particularly interested in analyzing news feeds, smart cities and most of all, air traffic control optimization for NASA, whose performance could greatly be improved if exceptional meteorological events were better taken into account. Time series is a very pervasive field so if there is an application that drives you, I'll be more than happy to tailor the system to the computational aspects implied by the problems and models you are interested in.