Ming-hao Hu, Chang-jian Wang, Yu-xing Peng, 2017. Meeting deadlines for approximation processing in MapReduce environments. *Frontiers of Information Technology & Electronic Engineering*, **18**(11): 1754-1772.
http://dx.doi.org/10.1631/FITEE.1601056

Meeting deadlines for approximation processing in MapReduce environments

Key words: MapReduce; Approximation jobs; Deadline; Task scheduling; Straggler mitigation

Corresponding author: Ming-hao HU
E-mail: minghao_hu@yeah.net

ORCID: http://orcid.org/0000-0003-2986-4139

Motivation

1. It is crucial to satisfy deadline requirements for MapReduce jobs in today's production environments so as to provide timely results for big data analytics.

2. Current solutions fail to satisfy the deadline requirement and maximize the volumes of processed data simultaneously, by pre-allocating appropriate resources or pre-sampling a subset of the entire dataset.

Main idea

A deadline-oriented task scheduler for MapReduce jobs that

1. makes dynamic scheduling decisions based on an approach-revise algorithm;

2. efficiently handles task failures and data skew.

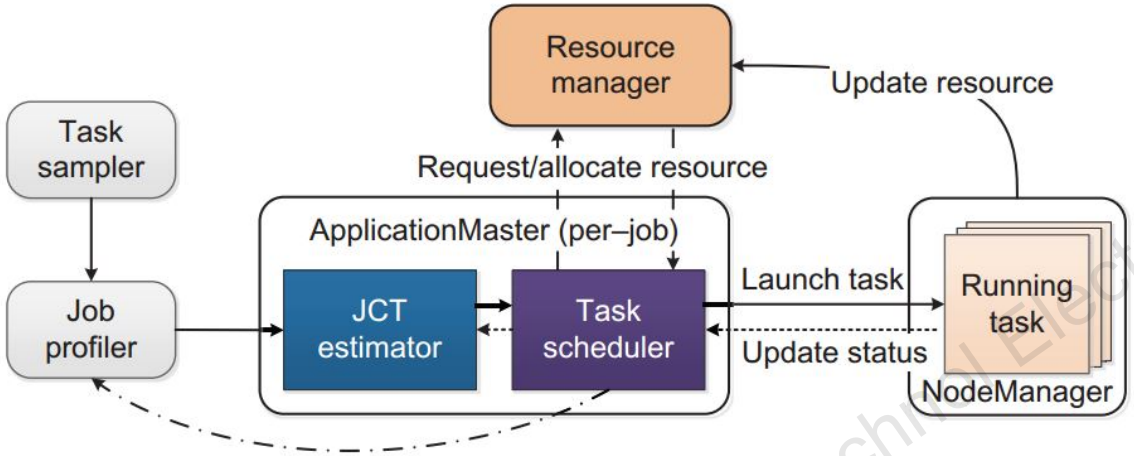
The results:

1. effectively meet the deadline;

2. process near-maximum volumes of data;

3. successfully work even with tight deadlines and limited resources.

Method



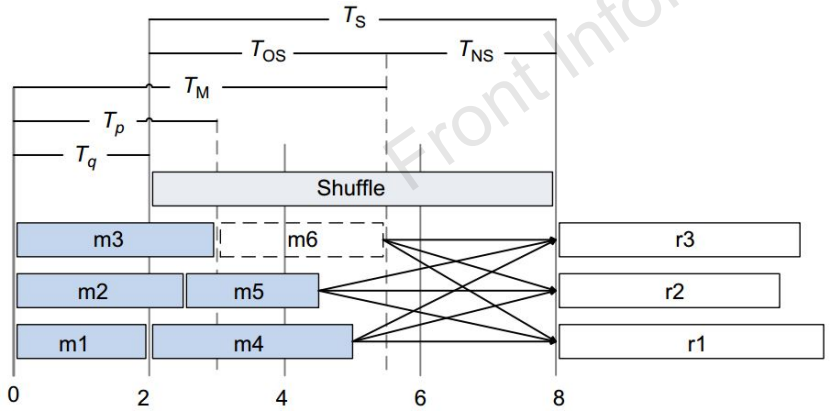
A architecture overview of our proposed scheduler

1. The approaching stage

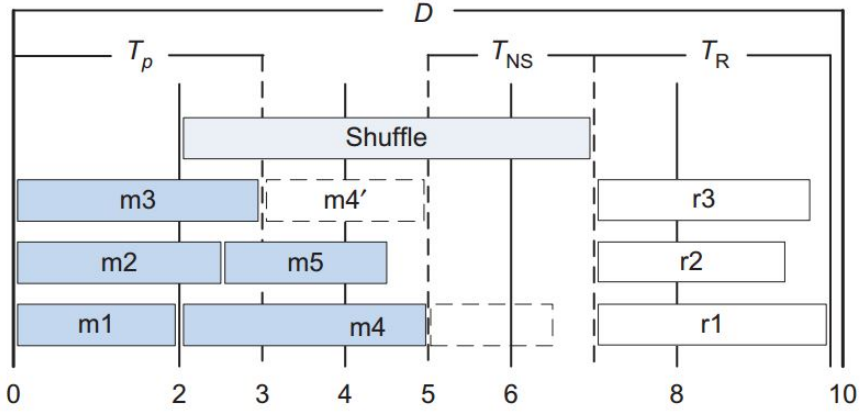
- Iterative estimation
- Dynamical scheduling

2. The revising stage

- Speculative duplication
- Task size optimization



Job completion time estimating



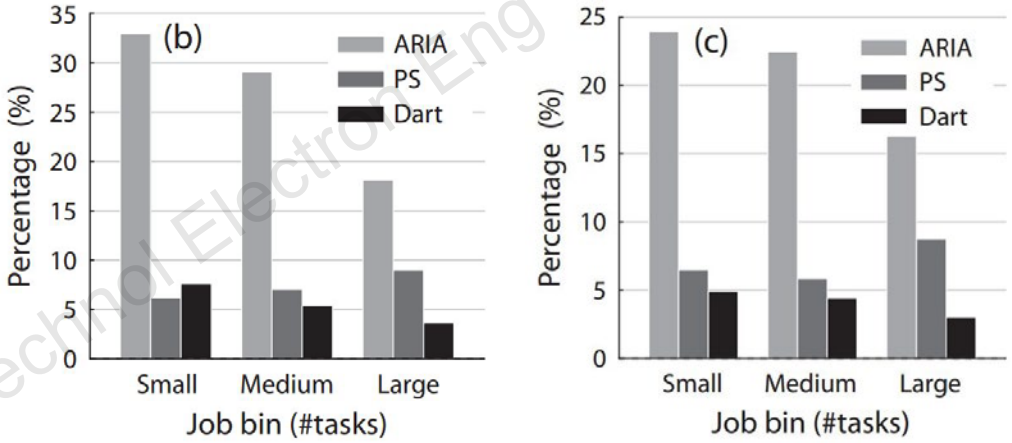
Straggler mitigation

Major Results

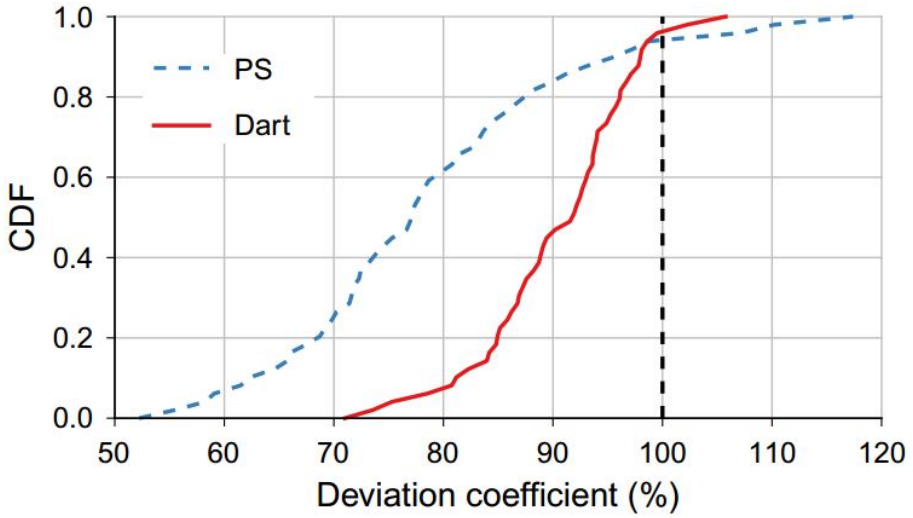
1. Deadlines are met despite of different job's sizes even when

1.1 tight deadlines are given

1.2 limited resources are given



2. Volumes of processed data are maximized while satisfying the deadline requirement



Conclusions

1. We investigate the dilemma in MapReduce environments including the inability to meet deadlines, unable to maximize processed data and the straggler problem.

2. We proposed a deadline-oriented task scheduling approach that consists of an approaching stage and a revising stage, and offers solutions to task failures and data skew problem.

3. Extensive experiments are conducted to demonstrate that our approach can not only effectively meet the deadline but also process near-maximum volumes of data even with tight deadlines and limited resources.