Three Pieces of the MapReduce Workload Management Puzzle

Abhishek Verma*, Ludmila Cherkasova*, Vijay S. Kumar*, Roy H. Campbell*
*{verma7, rhc}@illinois.edu University of Illinois at Urbana-Champaign, #lucy.cherkasova, vijay.s.kumar}@hp.com HP Labs, Palo Alto

Job Profiles compactly summarize performance metrics of different job stages collected from logs

- Automatic Resource Inference and Allocation (ARIA) with novel performance models:
  - Can predict job completion time as a function of resources
  - Given a deadline, compute minimum resources to allocate to the job, so that it finishes within deadline

Three Pieces of the Puzzle

1. **Job Ordering**
   - Which order should the jobs be allocated to the chosen job?
2. **Tailoring amount of resources**
   - How many slots should be allocated to the job?
3. **Allocating spare resources**
   - How to allocate the spare resources in the system and de-allocate them in case of a new urgent job?

Job Scheduling using Different Mechanisms

1. **Earliest Deadline First**
   - Allocate all the resources in the system to the job with the earliest deadline
2. **Min-EDF**
   - Compute the minimum resources to allocate to the job with the earliest deadline
3. **Min-EDF-WC**
   - Allocate any spare resources among running jobs
   - When new job arrives, compute if enough slots will be released in the future to satisfy the job
   - If not, cancel spare tasks of the currently running jobs

Evaluation Setup and Workloads

- **Testbed Setup**
  - 66 HP DL145 machines: 2 masters + 64 slaves
  - Four 2.39 GHz cores, 8 GB RAM, 2 x 160 GB hard disks
  - Two racks, Gigabit Ethernet

- **Workloads**
  - Real testbed trace of 1000 jobs with combinations of: Wordcount, Sort, Bayesian classification, TF-IDF, WikiTrends, Twitter on 3 different datasets
  - Synthetic Facebook trace: generated using LogNormal distribution fit to 6 months of jobs

Simulator and Metrics

- Replay traces using the simulator SimMR
  - Discrete event simulator replays job traces at task-level
  - Accuracy > 95%
  - Can replay two weeks workload in 2 seconds

- Comparison metrics
  1. % of missed jobs
  2. Average job completion time
  3. Number of tasks released in the future to satisfy the job

Future Work

- All three mechanisms are required for deadline-based workload management
- Incorporate these mechanisms in existing schedulers
- Scale smaller datasets to simulate larger ones
- Dynamic resource adjustment
  - Compare expected behavior against observed behavior and adjust
  - Deal with stragglers, input data skew

The simulation results with the Facebook workload are similar and reflect the same conclusions.