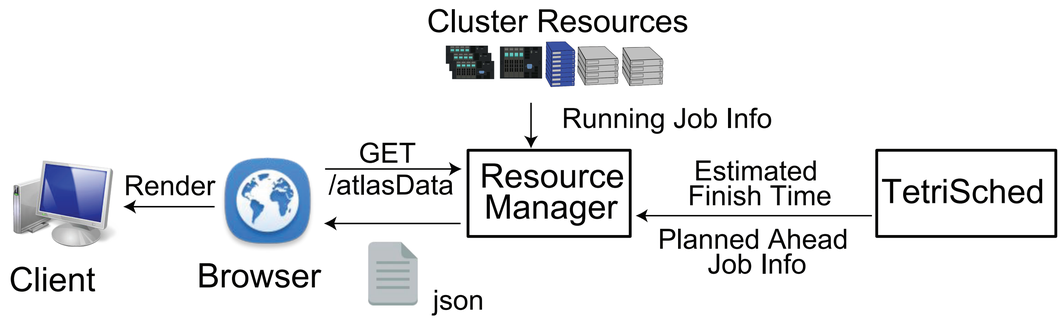
TetriScope: Visualizing Scheduling with Adaptive Plan-ahead

Jun Woo Park, Alexey Tumanov, Xiaolin Zang, Michael A. Kozuch (Intel Labs), Bill Courtright, Greg Ganger

Motivation

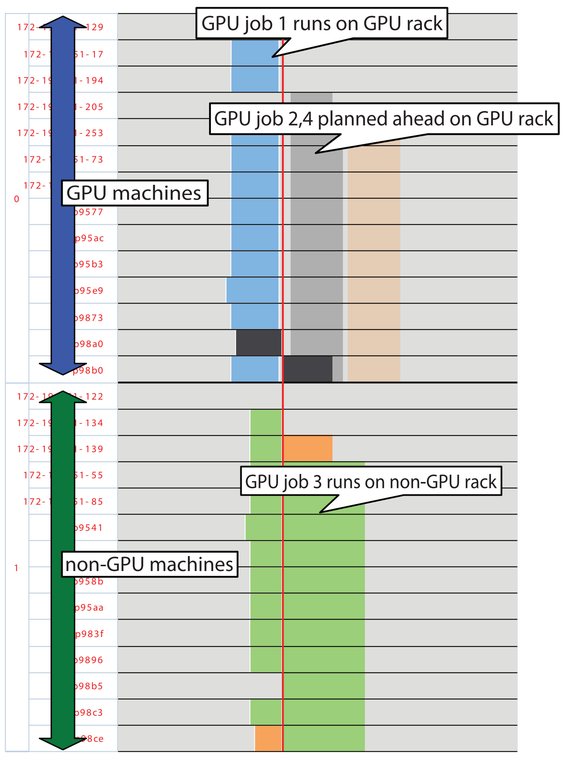
- **Need to understand/debug complex scheduling decisions**
  - **Effect of Job Properties**
  - **Scheduler performance**
- **Scheduler state is difficult to analyze at scale**
- **TetriScope: a visualization tool that presents YARN scheduler state visually so that users can grasp the cluster state and analyze scheduler behavior**

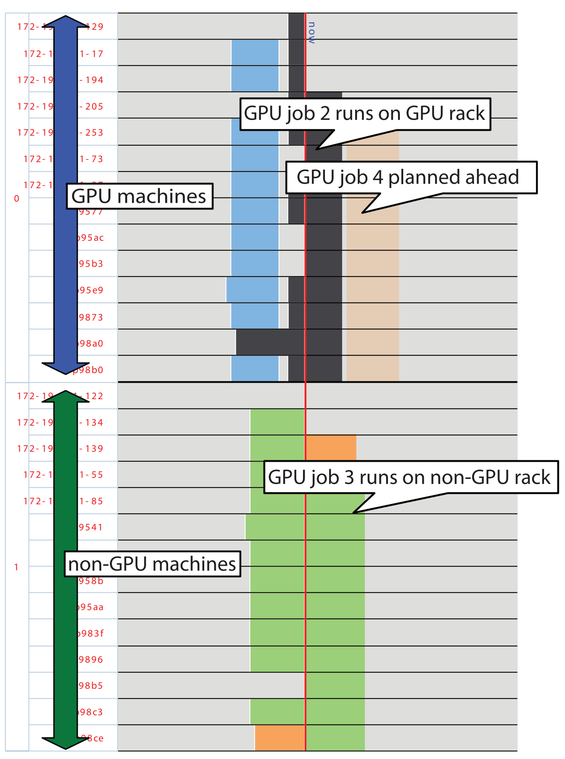
TetriScope System Architecture



Demo Scenarios

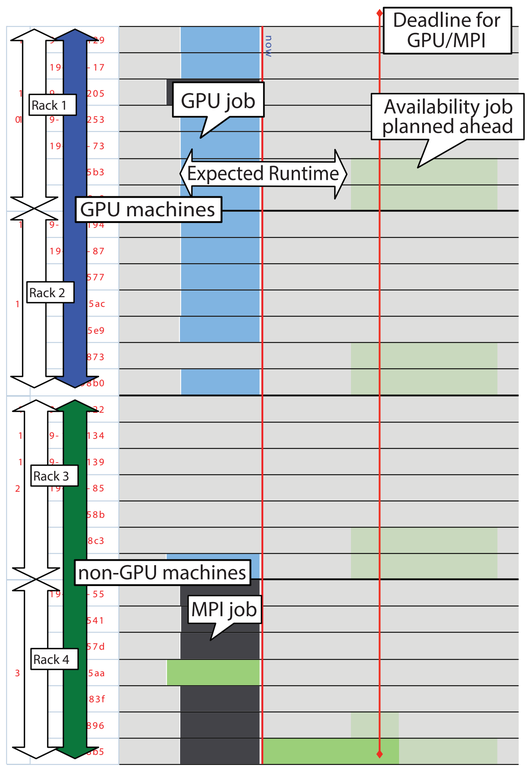
PLAN AHEAD

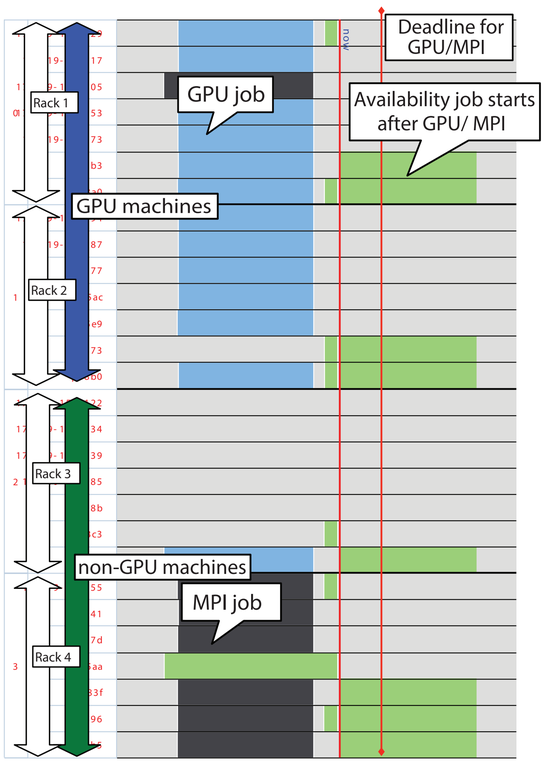




- **GPU job 1 schedules on GPU machine**
- **GPU job 2 runs on non-GPU rack since it is faster than waiting for GPU rack**
- **GPU job 3 and 4 is planned ahead in GPU rack**

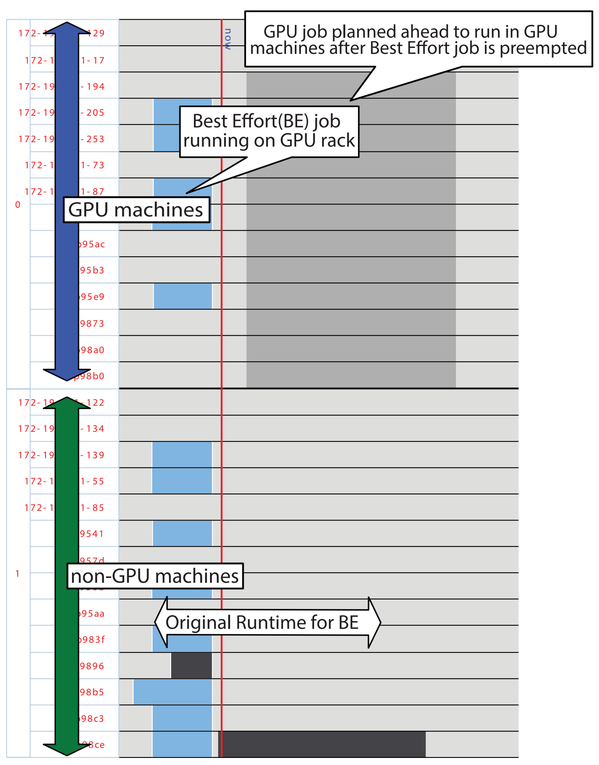
GLOBAL SCHEDULING

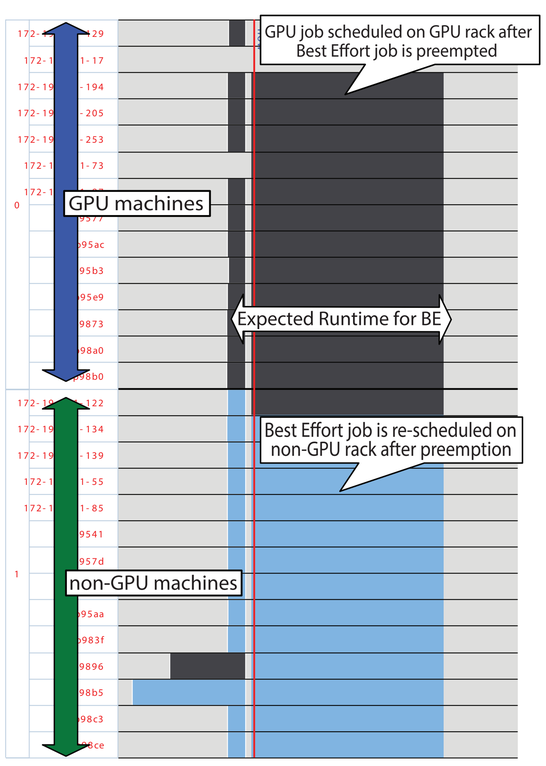




- **Availabilty job arrives with loose deadline**
- **GPU/MPI arrives with tighter deadline**
- **Only way to meet deadline: run GPU and MPI job before Availability Job**

PREEMPTION





- **Best Effort job already occupying GPU rack when GPU job arrives**
- **Best Effort job is preempted to let GPU job run on GPU rack**
- **Best Effort job is scheduled on non-GPU rack after preemption**

Carnegie Mellon University

Intel Science & Technology Center for Cloud Computing

NSERC CRSNG

Parallel Data Laboratory

PDLSVD16