Optimizing Resource Management for Machine Learning Workloads in High-Performance Clusters

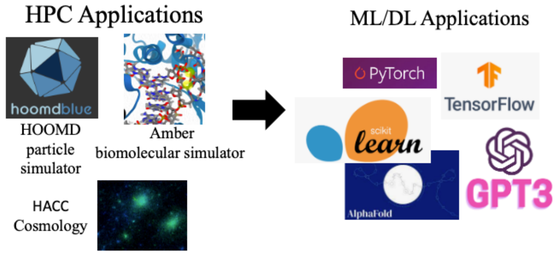
Di Zhang, UNC Charlotte
Dong Dai, DIRLAB

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Introduction

About High-Performance Clusters…

- HPC systems are traditionally optimized for long-term, resource-heavy scientific simulations.
- An increasing integration of DL applications presents new challenges due to their different characteristics:
  1. Heterogeneous resource use, leveraging both CPUs and GPUs.
  2. High cancellation rate due to feedback-driven exploration.
- The rising popularity of DL in HPC is significantly impacting job scheduling.
- It's crucial for the HPC community to understand and adapt to these changes to maintain system performance and efficiency.



Objectives

Goals

1. Compare and contrast the characteristics of traditional HPC jobs and DL jobs to gain a comprehensive understanding of their similarities and differences.
2. Identify novel opportunities in resource management to effectively cater to the demands of these emerging workloads.

Collected Data

Four Job Traces

Mira (ANL)

Blue Waters (NCSA)

Microsoft

Philly (Microsoft)

商汤 sensetime

Helios (Sensetime)

How does Job Trace look like?

Job

- Submit Time
- Run Time
- Wall Time
- Requested Nodes
- User ID
- Project ID
- …

Job Trace

Job 1

Job 2

Job 3

…

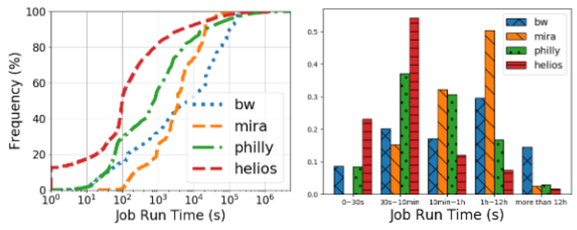
Characteristic Aspects

Analyzing Key Attributes

1. Job Geometries Characterization
   - Job Run Time
   - Job Arrival Patterns
   - System Utilization and Resource Occupation
   - Job Waiting Time
2. Job Failure Characterization
   - Job Failures Distribution
   - Correlation between Job Failure and Job Geometries.
3. Users' Behaviors Characterization
   - Users' Repeated Behaviors
   - Users' Submission Behaviors.
   - Correlation between Per-User Job Run Time and Job Statuses.

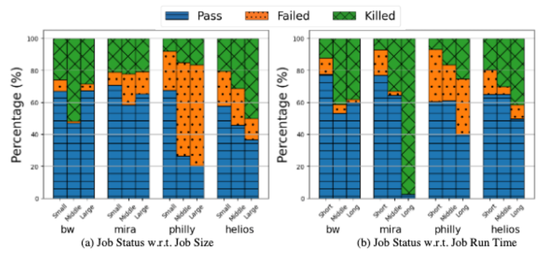
Results

7 Takeaways:



(a) CDF of Run Time (b) Details of Run Time Distribution

(Job Run Time) Takeaway 1: DL jobs tend to be shorter than traditional HPC jobs



(a) Job Status w.r.t. Job Size (b) Job Status w.r.t. Job Run Time

(Job Failure) Takeaway 2: the larger and longer jobs present higher failure rates across the systems

…

(b) Mira

(User Behavior) Takeaway 7: the elapsed time of users' jobs, can be used to predict job runtimes and be utilized further for better scheduling efficiency

Research Opportunities:

Traces	Metrics	Baseline	Adaptive	Improved
Blue Waters	*wait*	7513.02	7520.38	<1%
	bsld	39.02	38.99	<1%
	util	0.7164	0.7165	<1%
	violation	1258.35	1200.81	**5%**
Mira	*wait*	34199.90	36210.37	-6%
	bsld	37.81	39.40	-4%
	util	0.8792	0.8805	<1%
	violation	670.31	344.89	**49%**

Benefit on leveraging user behaviors in relaxed backfilling in HPC scheduling

Conclusions

Conclusions

- A cross-system analysis was conducted on four real-world clusters, including classic HPC, classic DL, and hybrid setups, to understand the impact of emerging DL workloads on HPC scheduling.
- Seven key takeaways were identified, revealing notable differences and similarities among the clusters.
- These insights will guide the design of more efficient job schedulers for future HPC systems.
- Two use case studies were introduced - job run time prediction and adaptive relaxed backfilling - to enhance existing job scheduling.

Future Work

All data processing logic and simulator will be publicly accessible and online analysis services will be provided, helping researchers design more efficient HPC schedulers in the future.

References

[1] Zhang, Di, et al. "RLScheduler: an automated HPC batch job scheduler using reinforcement learning." *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2020.

[2] Zhang, Di, Dong Dai, and Bing Xie. "SchedInspector: A Batch Job Scheduling Inspector Using Reinforcement Learning." *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing*. 2022.