



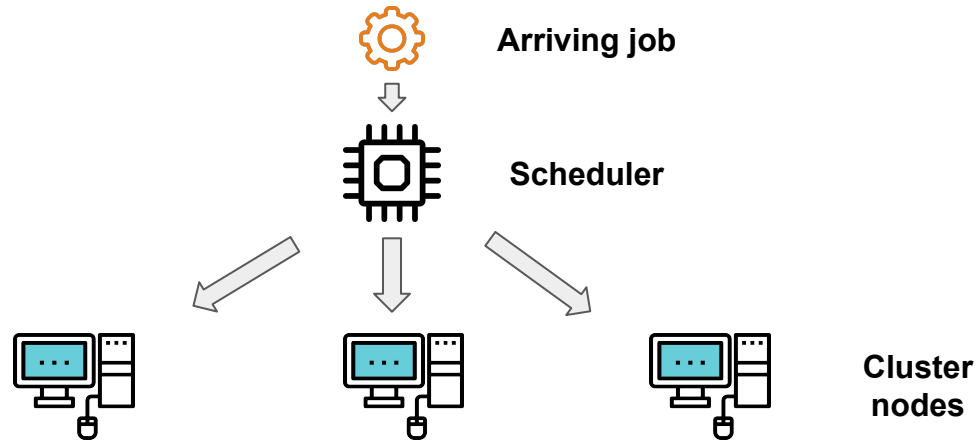
Telemetry data for machine learning based scheduling

Telemetry data for machine learning based scheduling

- Data generated in distributed systems and computing clusters:
 - Memory and CPU usage
 - Network parameters (latency, distance, etc)
 - Affinities between jobs and nodes

Telemetry data for machine learning based scheduling

- Scheduling involves deciding where to allocate arriving jobs in the cluster
- Use of telemetry data to optimize scheduling and other tasks, such as cluster workload estimation



Telemetry data for machine learning based scheduling

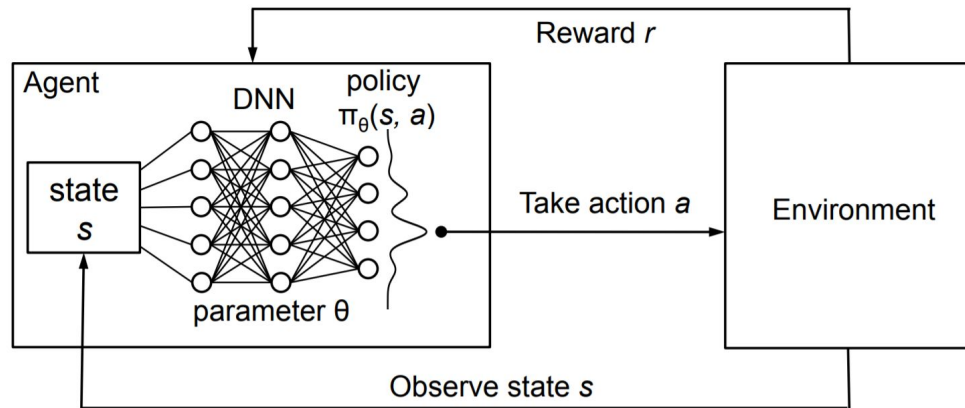
- Use of Machine Learning to optimize scheduling
 - Deep Reinforcement Learning (DRL)

Research questions

- Is it beneficial to use telemetry data as an input to a Machine Learning model to perform scheduling decisions?
- Can telemetry data be used to predict the future workload in a node and to avoid system failures?

DRL scheduling - Methodology

- Why Reinforcement Learning?



DRL scheduling - Methodology

- Policy gradient algorithm

$$\theta = \theta + \alpha \sum_t^{\infty} \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) * (v_t - b_t)$$

- Experiment in simulated cluster
 - Distance based scheduling

Objective: Average job duration

$$\bar{d} = \frac{1}{N} \sum_{j=0}^N d_j$$

where \bar{d} is the average job duration, N is the total number of jobs in the episode and d_j is the duration of the job j .

Reward

$$R_t = \sum_{j \in J_s} -1$$

where R_t is the reward at the current time-step and J_s is the set is jobs that are currently in the system.

State

- Input to DRL scheduler
 - Memory available in nodes
 - Memory request of waiting jobs in buffer
 - Affinity preference of waiting jobs in buffer

	Node 1 Memory available	Node 2 Memory available	Job 1 Memory request	Job 1 Affinity preference	Job 2 Memory request	Job 2 Affinity preference
Categorical data	500	1000	750	1	250	2
One-hot encoded data	001000	000100	0001	010	0100	001

Scheduler actions

Action	Description
0	Do not schedule any job
1	job 1: schedule to node 1 job 2: remain in buffer
2	job 1: remain in buffer job 2: schedule to node 1
3	job 1: schedule to node 1 job 2: schedule to node 1
4	job 1: schedule to node 2 job 2: remain in buffer
5	job 1: remain in buffer job 2: schedule to node 2
6	job 1: schedule to node 2 job 2: schedule to node 2
7	job 1: schedule to node 1 job 2: schedule to node 2
8	job 1: schedule to node 2 job 2: schedule to node 1

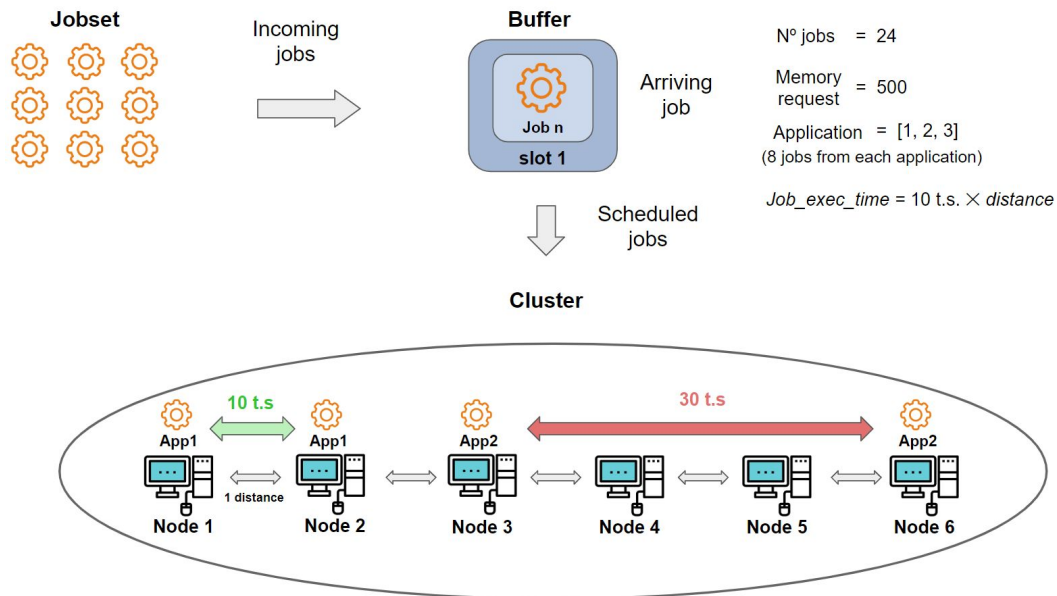
Objective: Average job duration

$$\bar{d} = \frac{1}{N} \sum_{j=0}^N d_j$$

where \bar{d} is the average job duration, N is the total number of jobs in the episode and d_j is the duration of the job j .

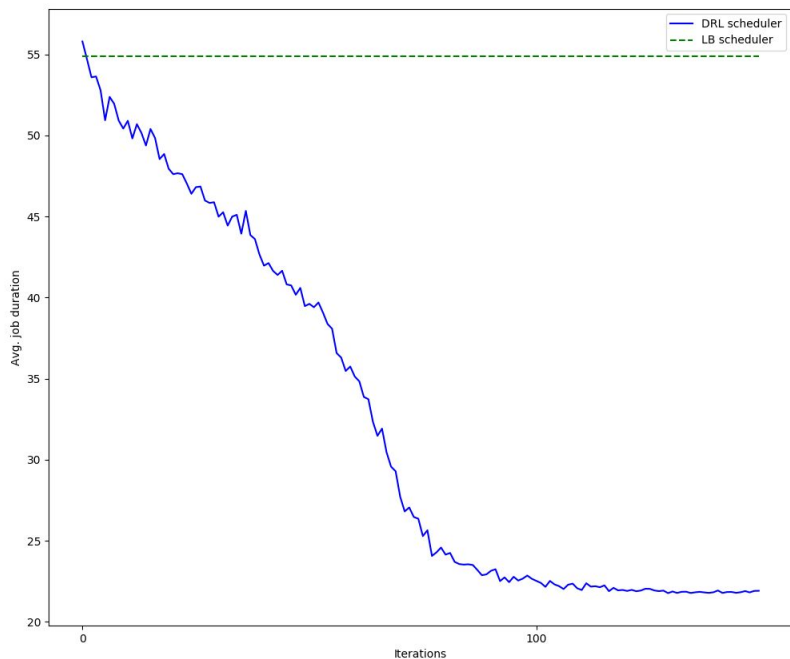
DRL scheduling - Distance based scheduling

- Goal
 - Minimize job duration, but how?
 - By learning from interaction and experience the distances between the nodes
- Implementation

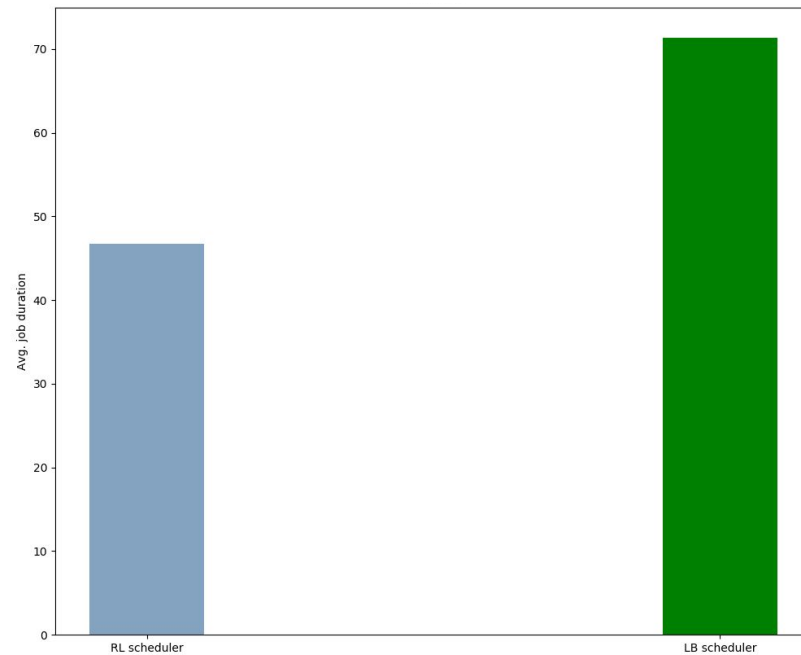


DRL scheduling - Distance based scheduling

- Results

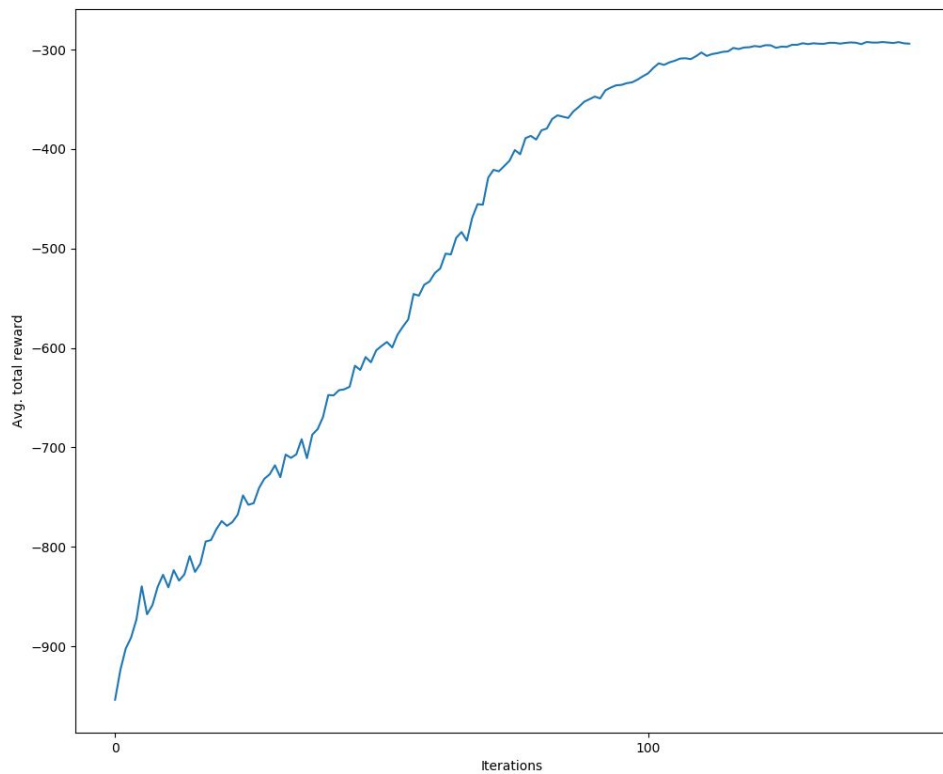


Average job duration evolution over training iterations



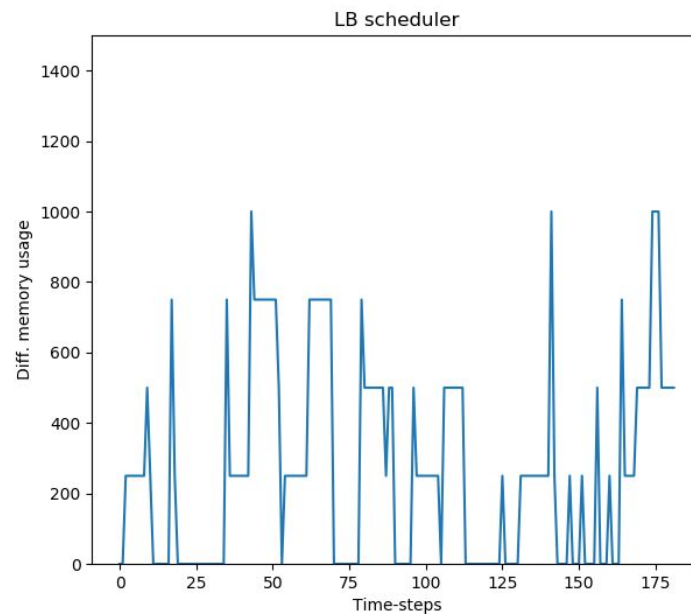
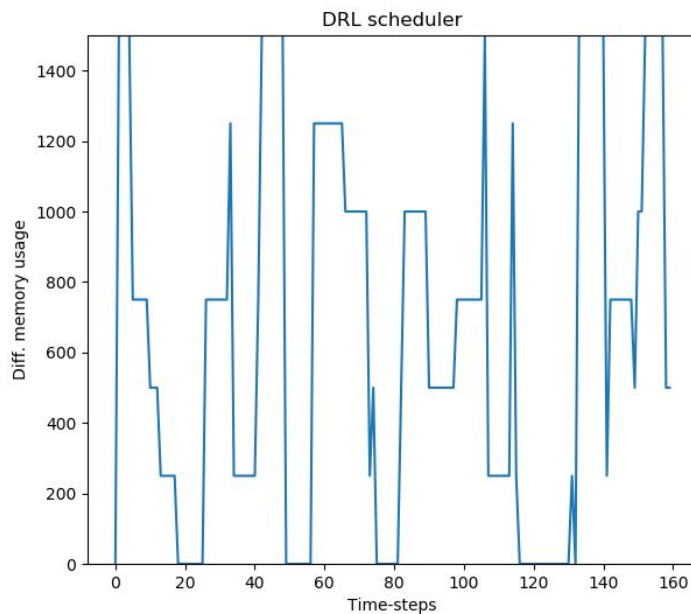
Average job duration of DRL and LB schedulers on test jobset

- Results



Average total reward evolution over training iterations

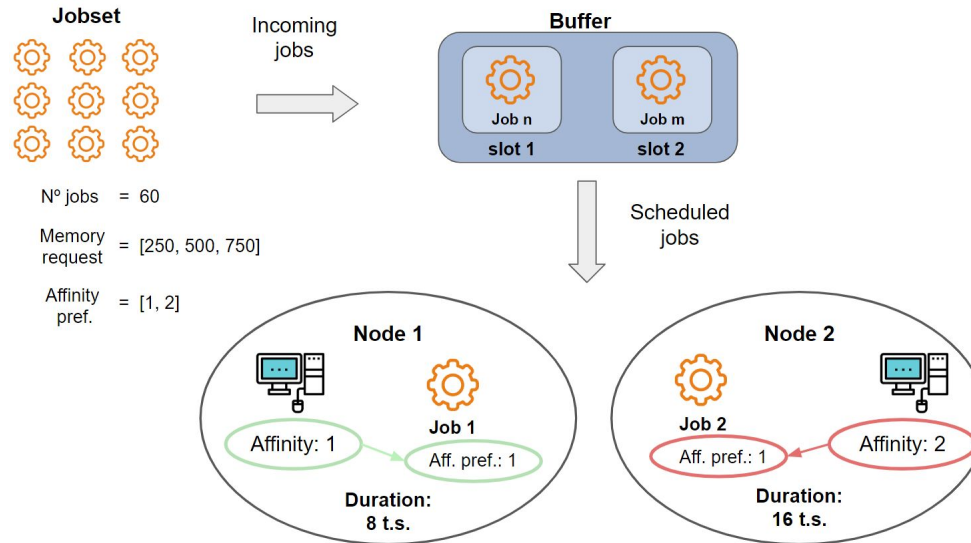
- Results



Difference of memory usage between nodes for DRL and LB schedulers

DRL scheduling - Affinity based scheduling

- Goal
 - Minimize job duration, but how?
 - Is DRL scheduling capable of learning the affinities of the nodes to reduce job duration?
- Implementation



State

- Input to DRL scheduler
 - App of Jobs in Nodes
 - App of buffer jobs
 - App of three next jobs

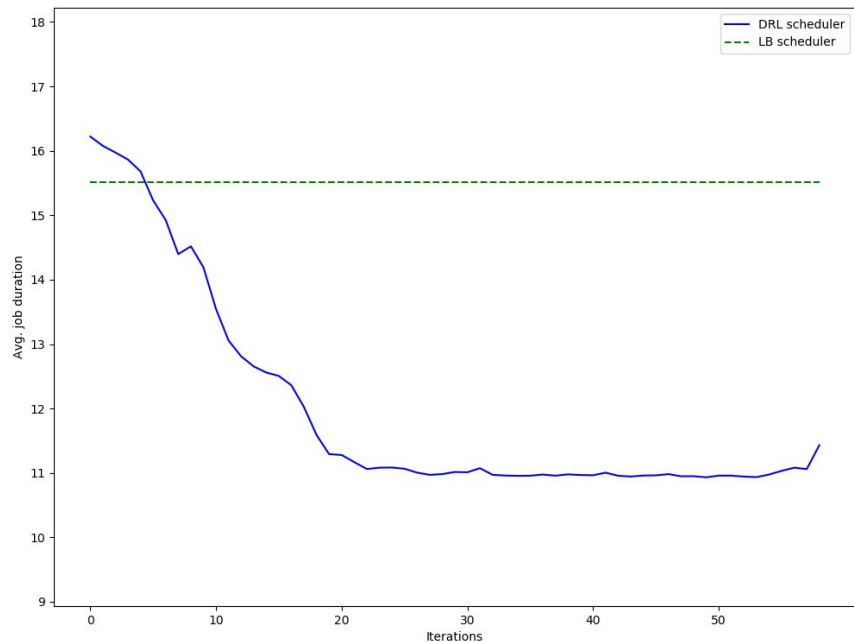
	Node 1 App	...	Node 6 App	Buffer App	Job + 1 App	Job + 2 App	Job + 3 App
Categorical data	1	...	0	2	3	1	0
One-hot encoded data	0010	...	0001	0100	1000	0010	0001

Scheduler actions

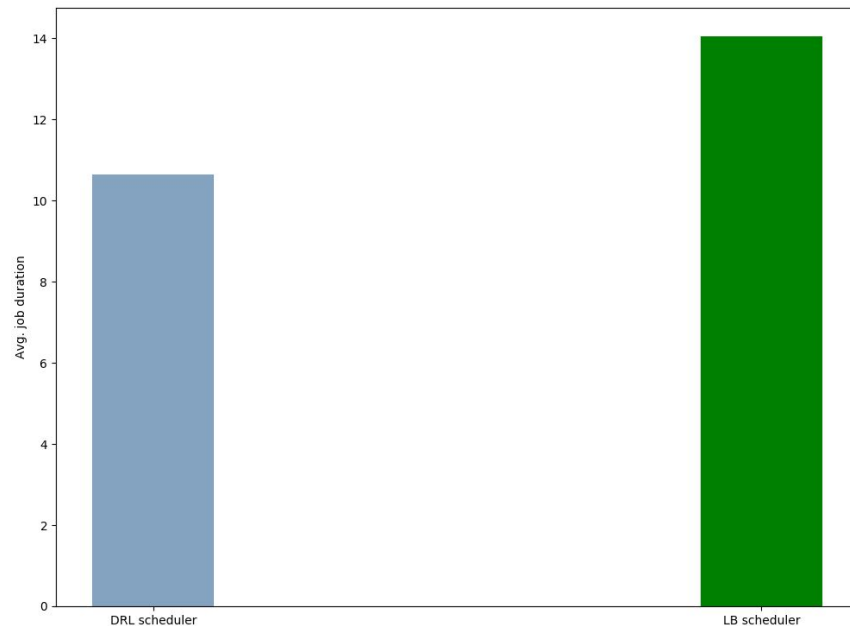
Action	Description
0	Do not schedule the job
1	Schedule the job to Node 1
2	Schedule the job to Node 2
3	Schedule the job to Node 3
4	Schedule the job to Node 4
5	Schedule the job to Node 5
6	Schedule the job to Node 6

DRL scheduling - Affinity based scheduling

- Results

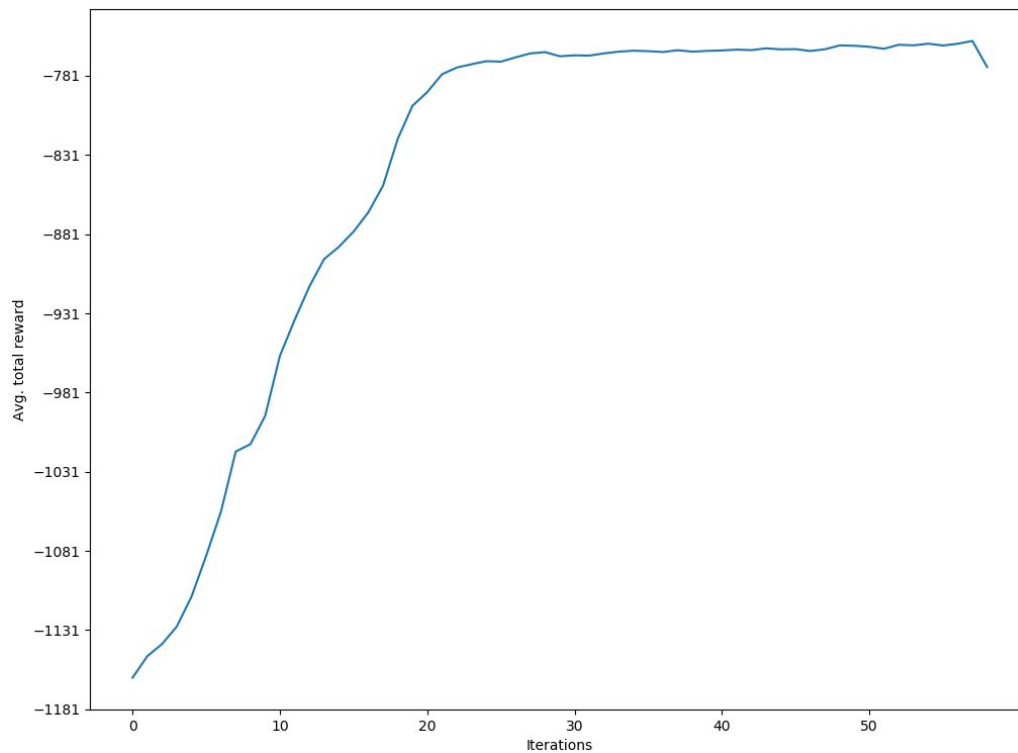


Average job duration evolution over training iterations



Average job duration of DRL and LB schedulers on test jobset

- Results



Average total reward evolution over training iterations