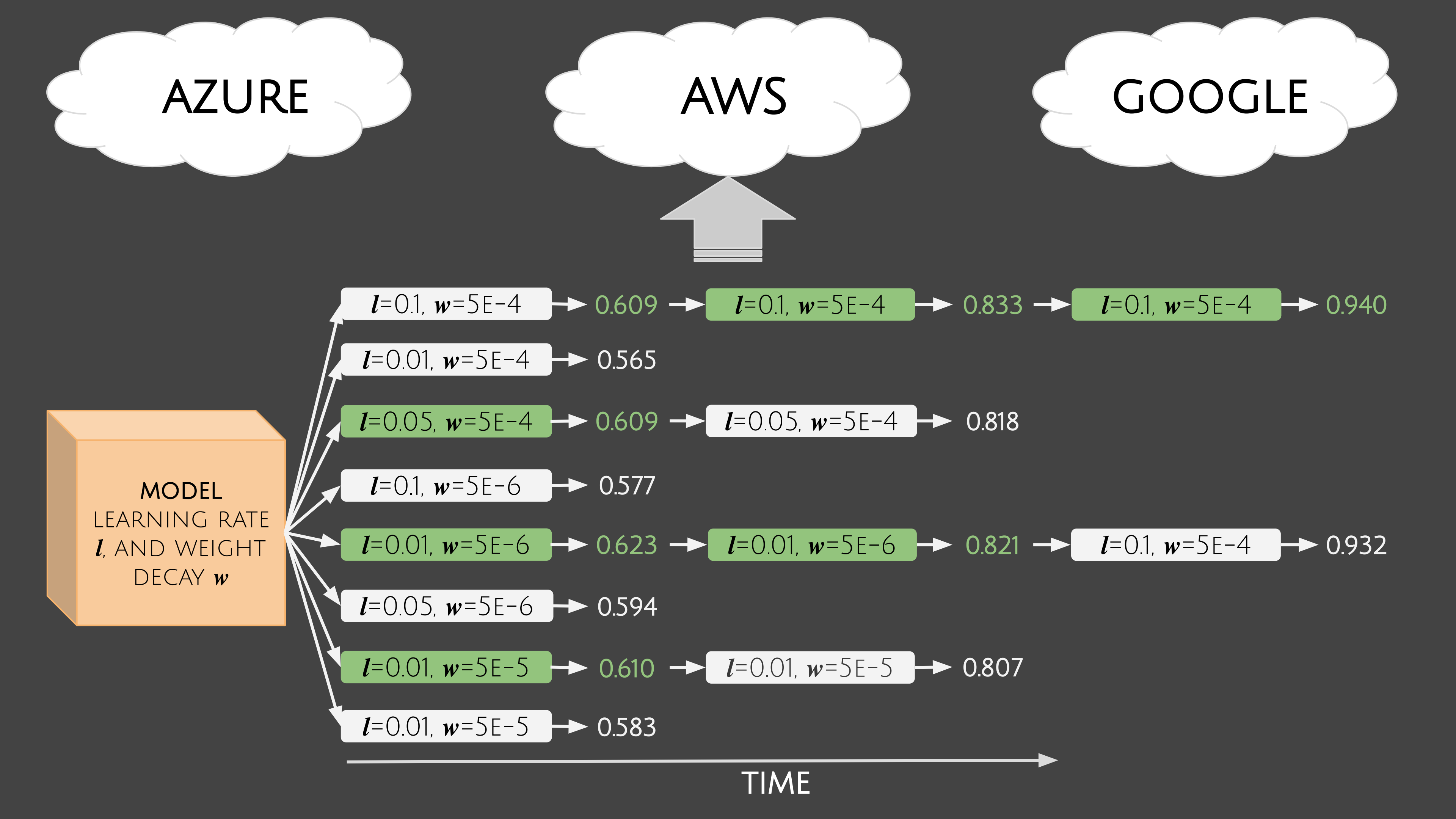
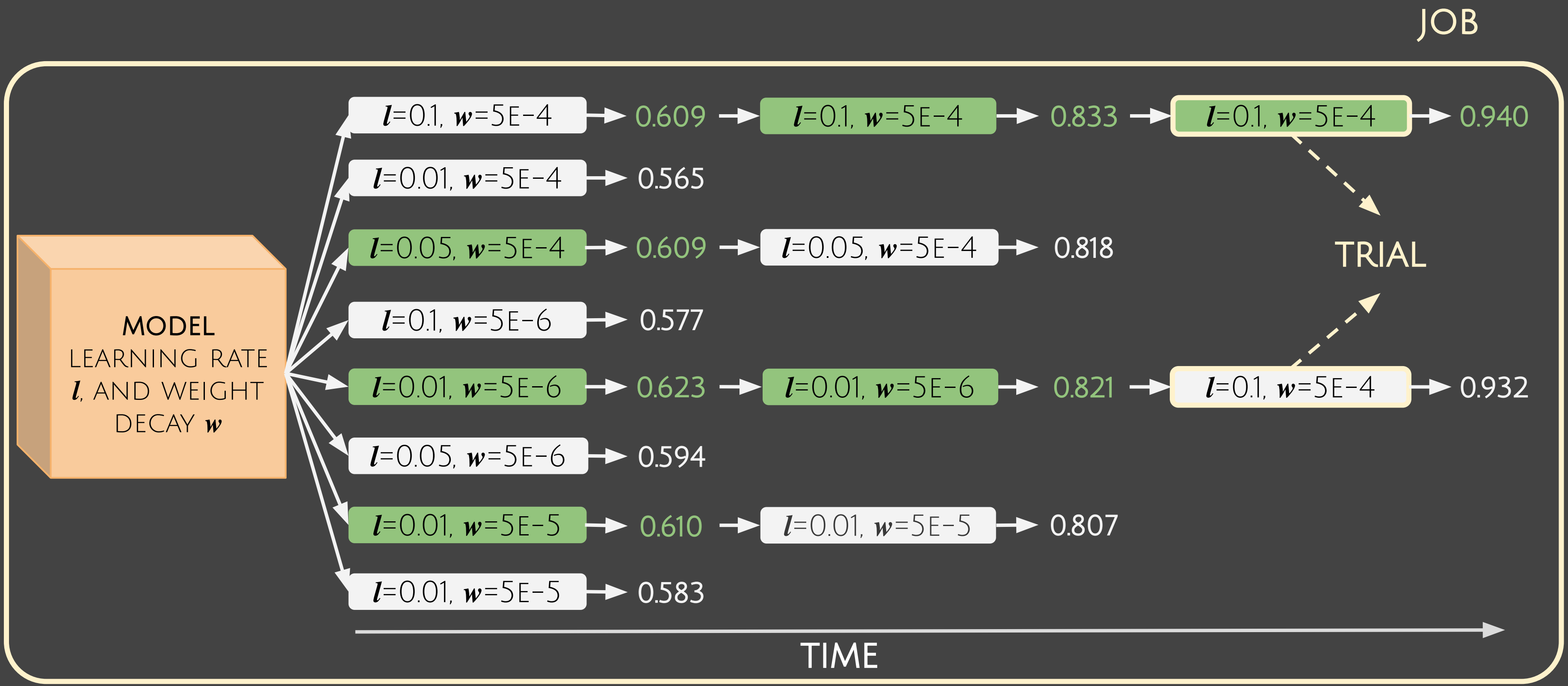
Rubberband

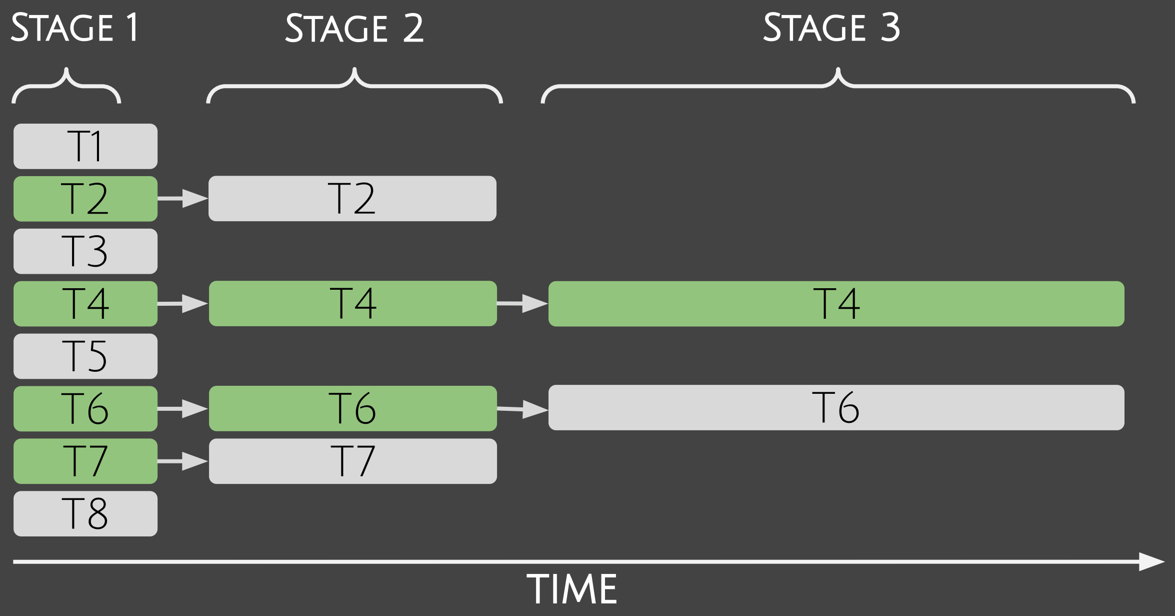
Ujval Misra*, Richard Liaw*, **Lisa Dunlap**, Romil Bhardwaj, Kirthevasan Kandasamy, Joseph E. Gonzalez, Ion Stoica, Alexey Tumanov


AZURE
AWS
GOOGLE
MODEL
LEARNING RATE
l, AND WEIGHT
DECAY w
l=0.1, w=5E-4
0.609
l=0.1, w=5E-4
0.833
l=0.1, w=5E-4
0.940
l=0.01, w=5E-4
0.565
l=0.05, w=5E-4
0.609
l=0.05, w=5E-4
0.818
l=0.1, w=5E-6
0.577
l=0.01, w=5E-6
0.623
l=0.01, w=5E-6
0.821
l=0.1, w=5E-4
0.932
l=0.05, w=5E-6
0.594
l=0.01, w=5E-5
0.610
l=0.01, w=5E-5
0.807
l=0.01, w=5E-5
0.583
TIME

Distributed, Parallel Hyperparameter Tuning

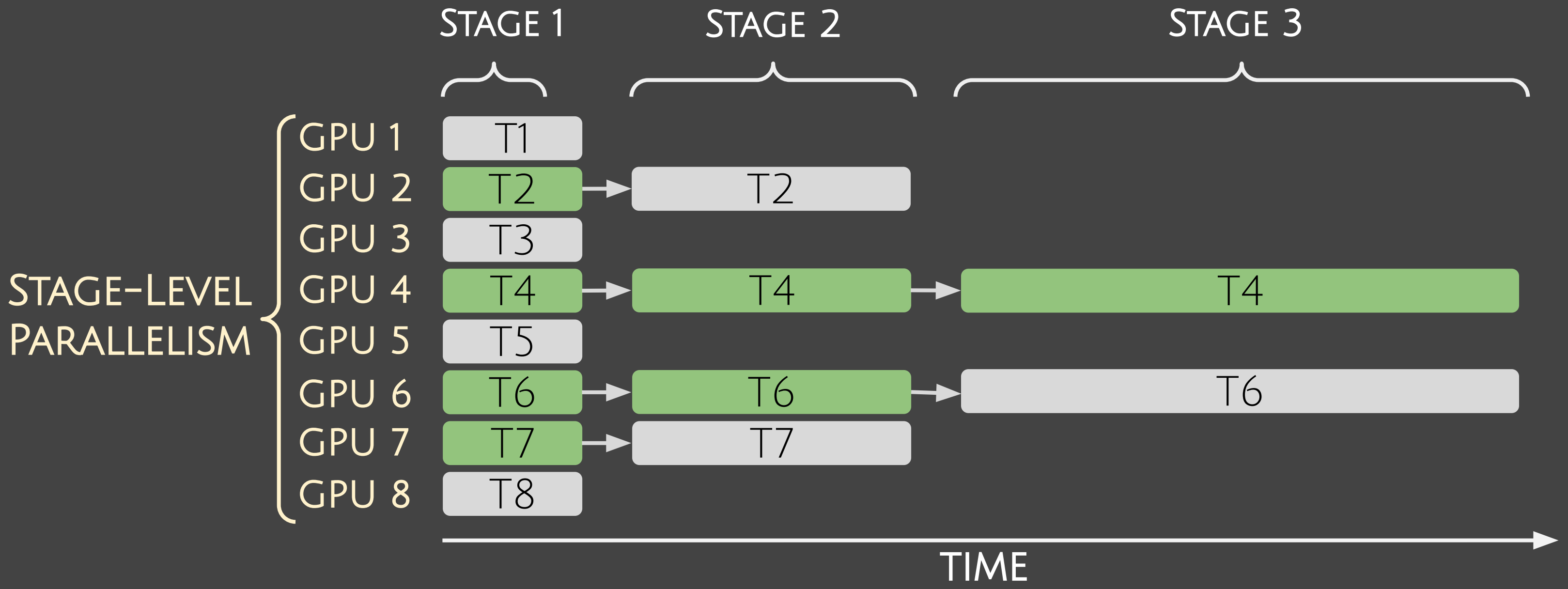


Distributed, Parallel Hyperparameter Tuning



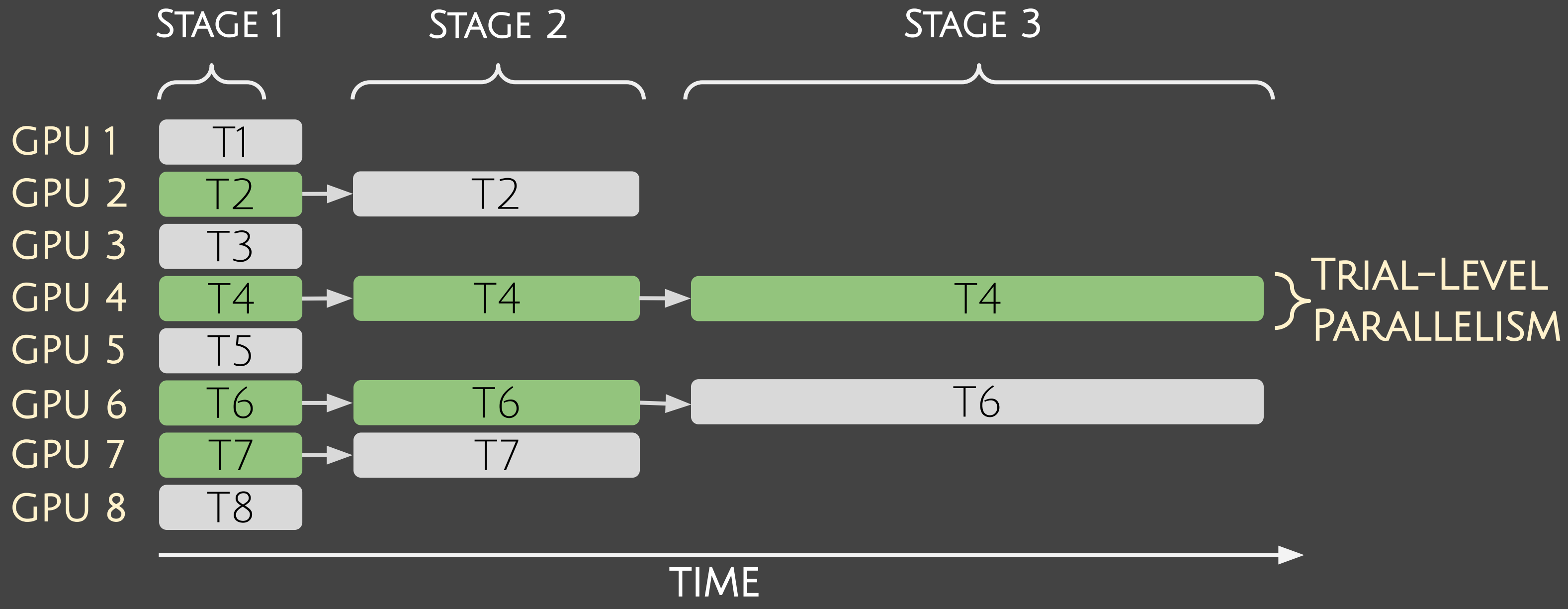
Successive Halving Algorithm (SHA)

Distributed, Parallel Hyperparameter Tuning

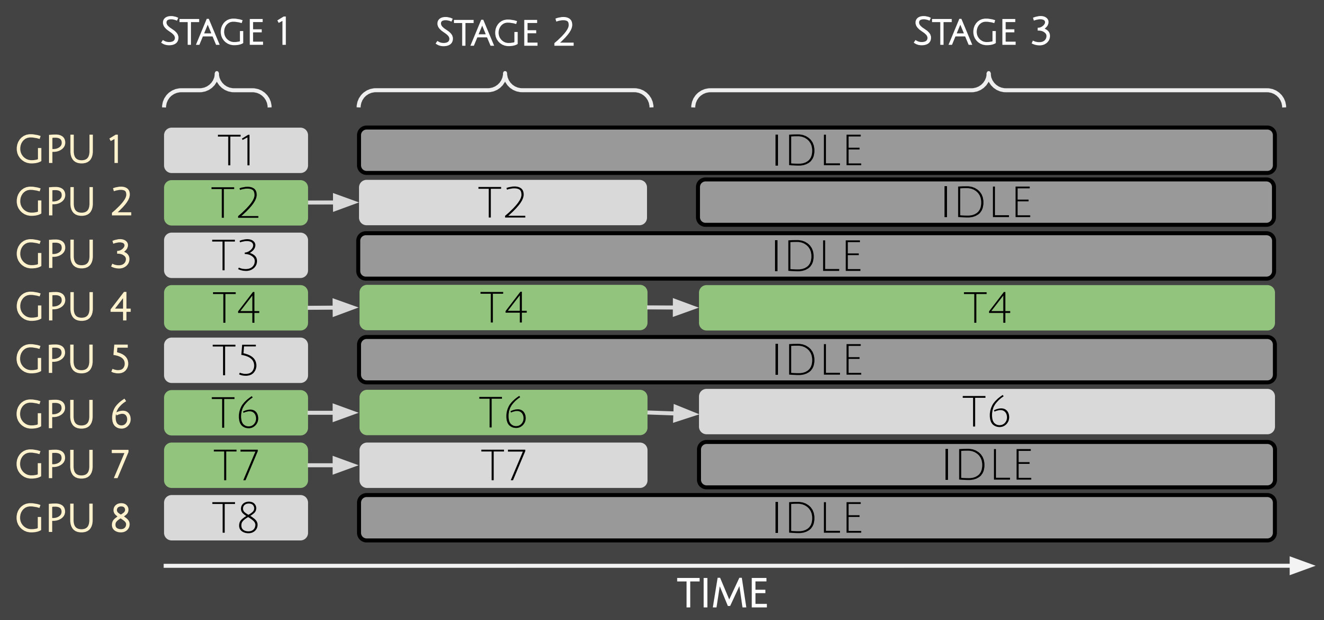


Successive Halving Algorithm (SHA)

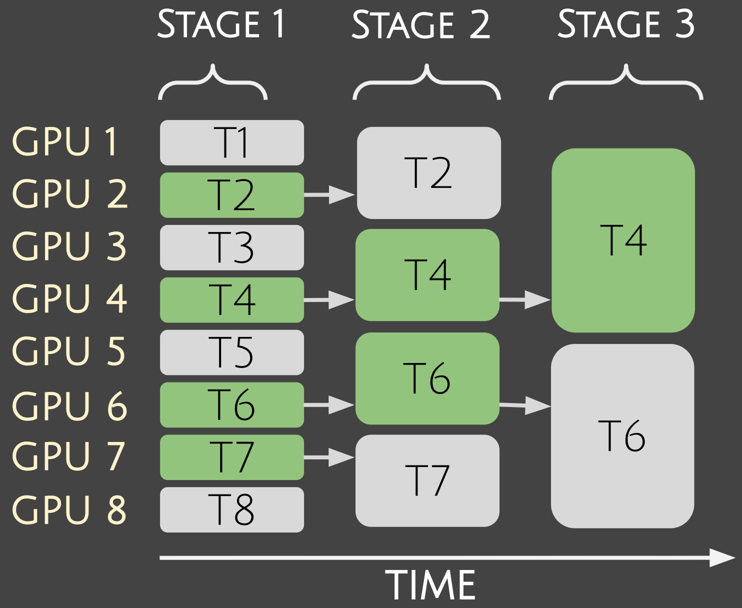
Distributed, Parallel Hyperparameter Tuning



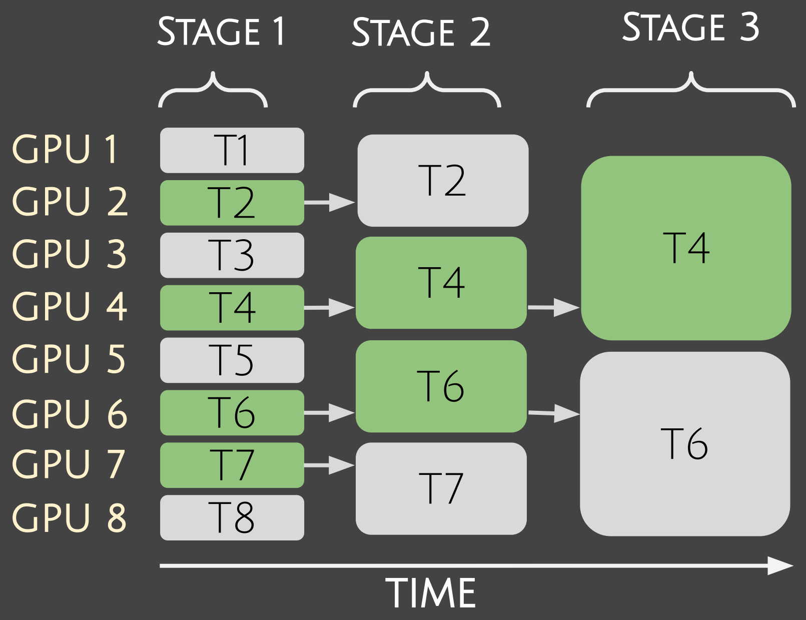
Distributed, Parallel Hyperparameter Tuning



Distributed, Parallel Hyperparameter Tuning

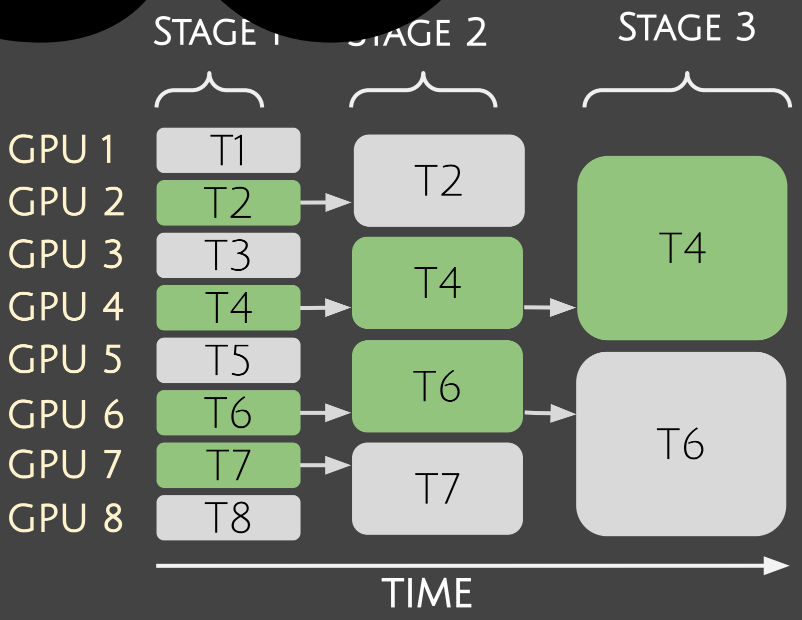


Distributed, Parallel Hyperparameter Tuning



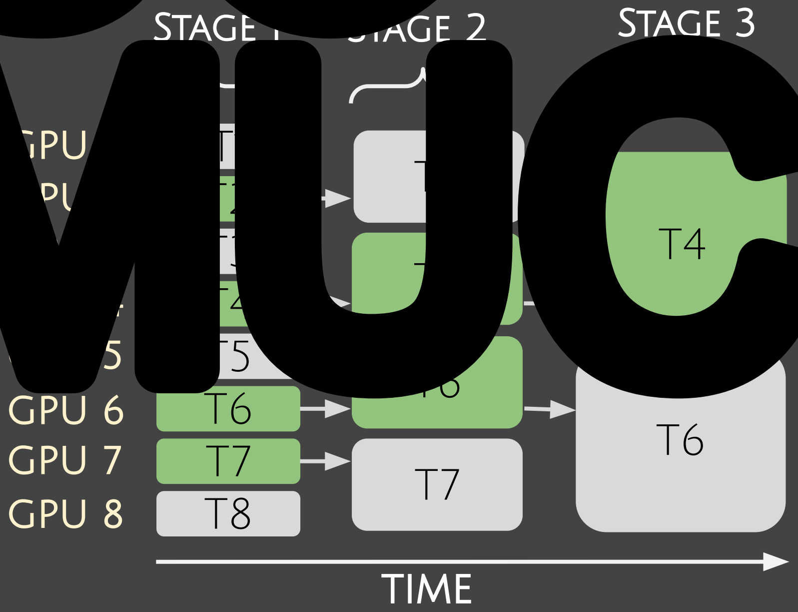
SO

…TRI…UTED, …RALLEL HYPERPARAMETER TUNING

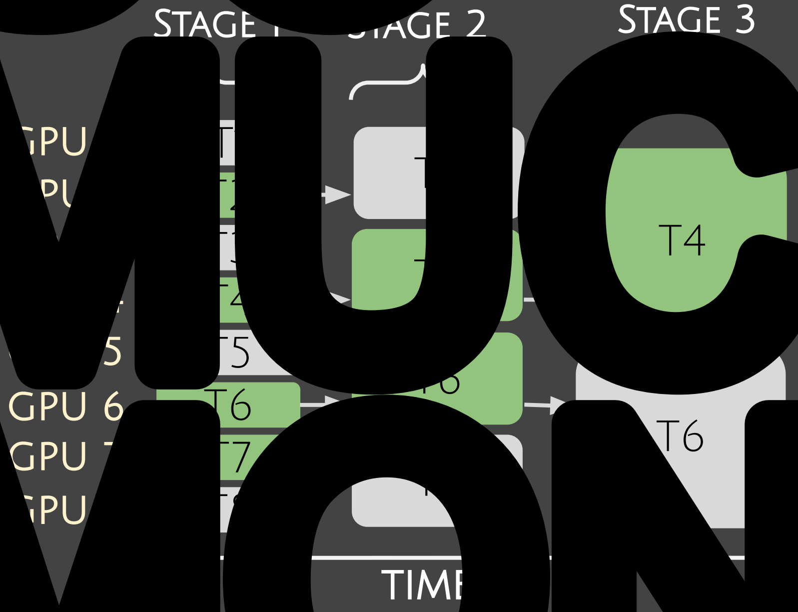


SO MUCH

Distributed, Parallel Hyperparameter Tuning

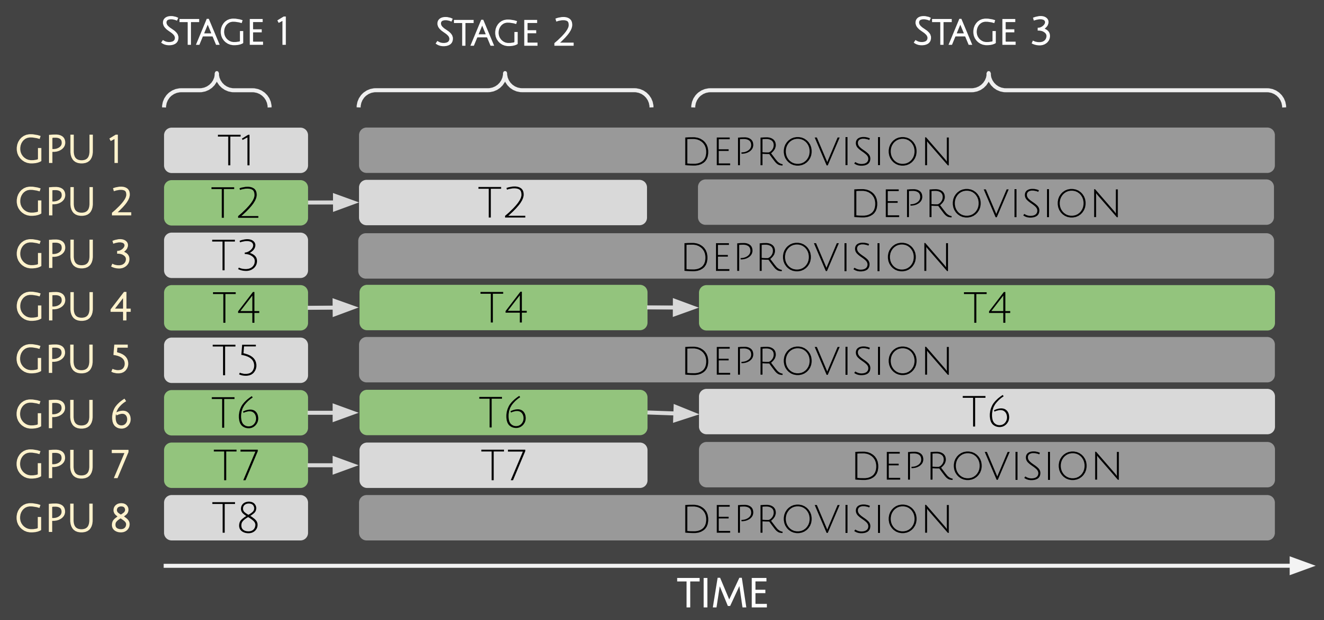


…TED, …RALLEL HYPERPARAMETER TUNING

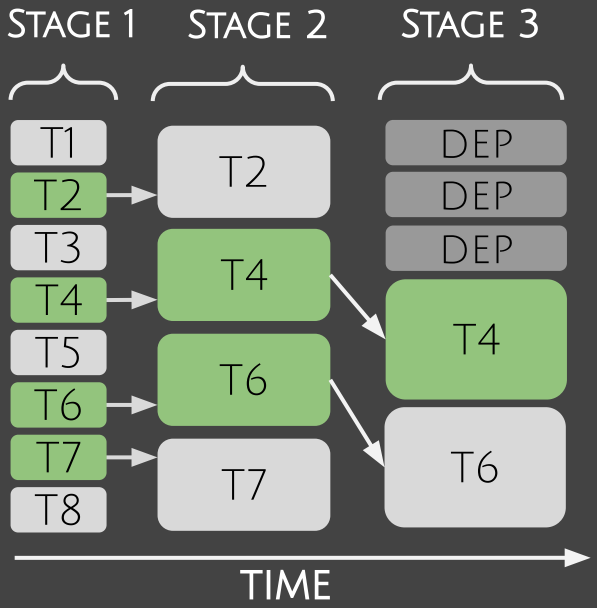


SO MUCH MONEY

Distributed, Parallel Hyperparameter Tuning



Given a time constraint, Minimize the cost of executing a hyperparameter tuning job.



Stage	Epochs	Trials	GPUs/Trial
1	0-4	8	1
2	5-12	4	2
3	13-28	2	3

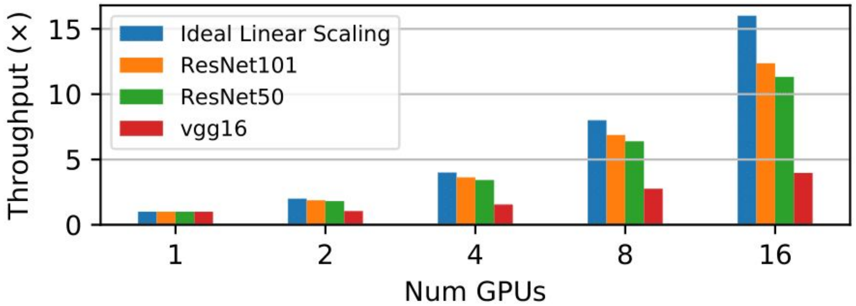
Allocation plan

Challenges

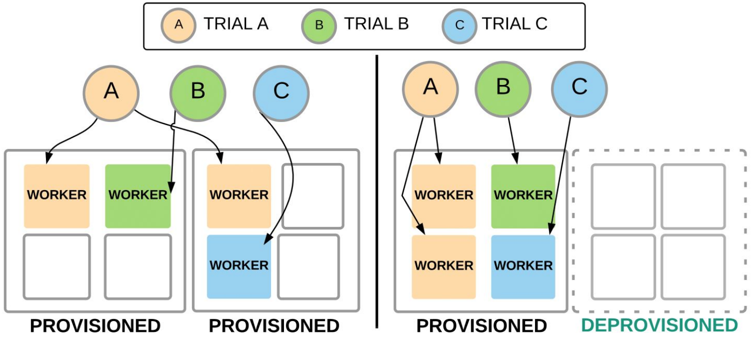
How can we model the job completion time and cost of the given allocation plan?

How can we generate a low cost allocation plan that completes on time?

How can we schedule said allocation to optimize worker co-location + cluster utilization?

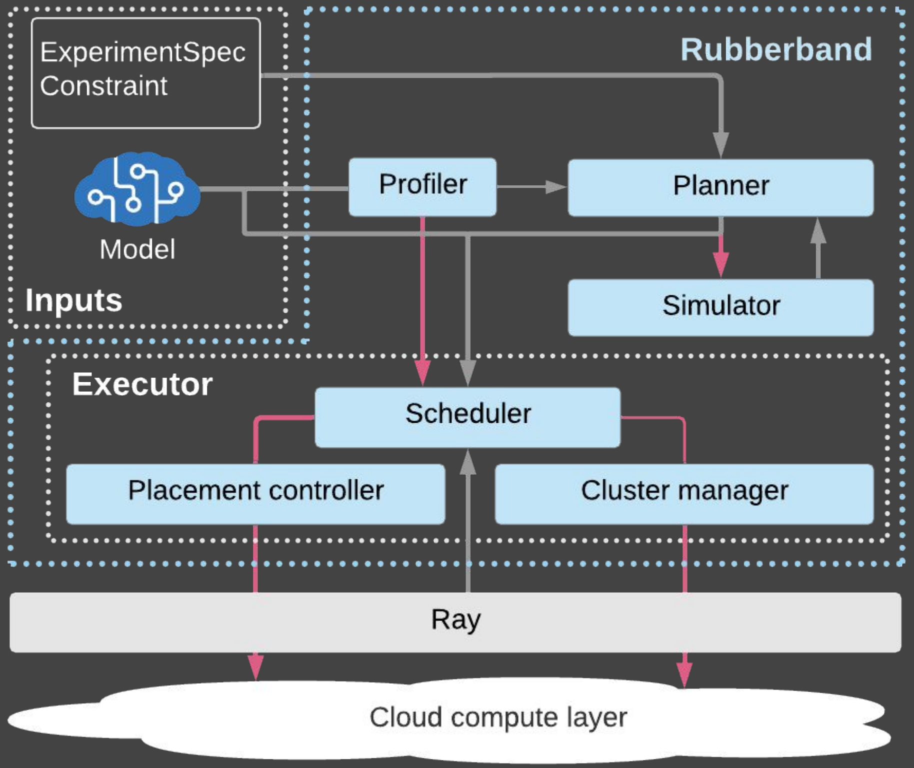


Model Scaling



Worker Co-location

Rubberband



ExperimentSpec
Constraint
Model
Inputs
Rubberband
Profiler
Planner
Simulator
Executor
Scheduler
Placement controller
Cluster manager
Ray
Cloud compute layer

Challenges	Rubberband
How can we model the job completion time and cost of the given allocation plan?	Cost/performance model via profiling DL model training latency and provisioning overheads
How can we generate a low cost allocation plan that completes on time?	DAG-based execution model which finds feasible and cost-efficient resource allocations
How can we schedule said allocation to optimize worker co-location + cluster utilization?	Full-stack system for placement, scheduling, and scaling

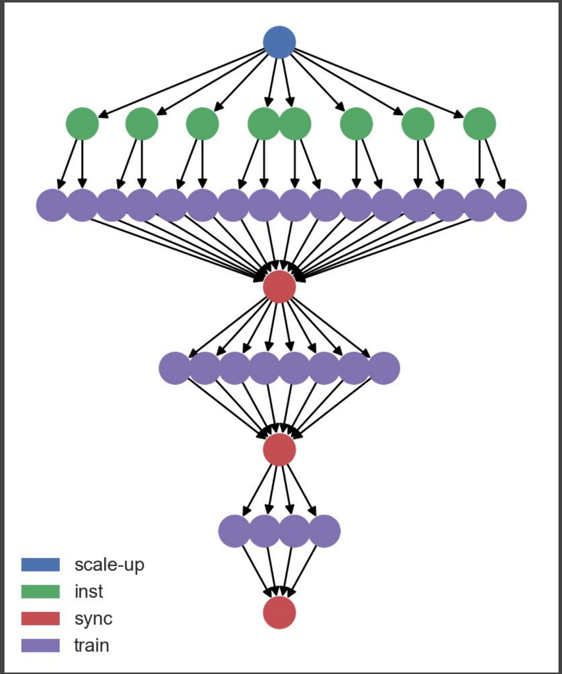
Issue 1: Modeling job completion time and cost

Performance modeling

- training latency
- provider queuing delay
- instance initialization latency

Cost modeling

- compute price
- billing granularity
- data price



Issue 2: finding a low-cost allocation plan

Issue 2: finding a low-cost allocation plan

Step 1: generate candidates

Issue 2: finding a low-cost allocation plan

Step 1: generate candidates

Step 2: use simulator to predict job completion time

Issue 2: finding a low-cost allocation plan

Step 1: generate candidates

Step 2: use simulator to predict job completion time

Step 3: greedily select best candidate

Issue 2: finding a low-cost allocation plan

Step 1: generate candidates

Step 2: use simulator to predict job completion time

Step 3: greedily select best candidate

Maximize ***cost-marginal benefit:***

$$M = \frac{\text{Cost of current best plan - Cost of proposed plan}}{\text{JCT of proposed plan- JCT of current best plan}}$$

Issue 2: finding a low-cost allocation plan

Step 1: generate candidates

Step 2: use simulator to predict job completion time

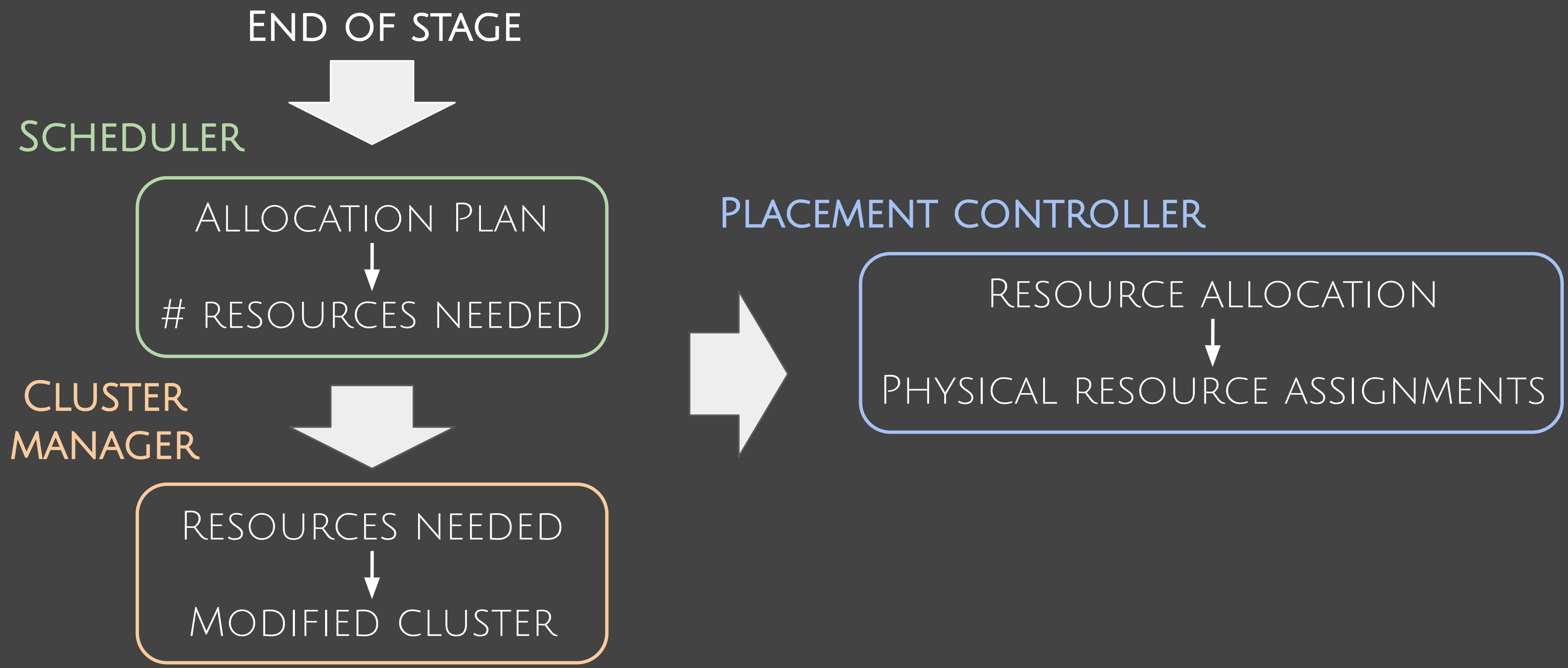
Step 3: greedily select best candidate

Step 4: iterate with new best candidate

Maximize ***cost-marginal benefit:***

$$M = \frac{\text{Cost of current best plan - Cost of proposed plan}}{\text{JCT of proposed plan- JCT of current best plan}}$$

Issue 3: effectively execute allocation plan



Issue 3: effectively execute allocation plan

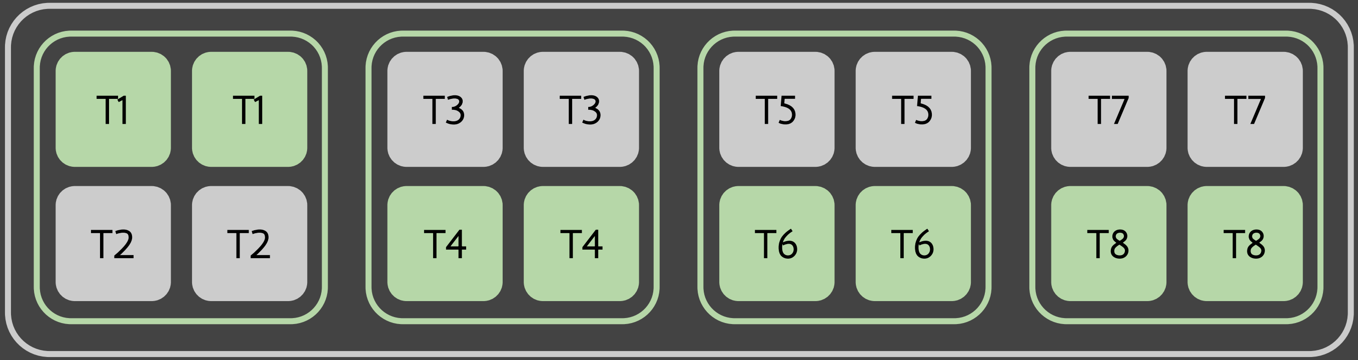
Stage	Epochs	Trials	GPUs/trial	Cluster size
1	0-4	8	2	4
2	5-12	4	3	3

End of stage

Issue 3: effectively execute allocation plan

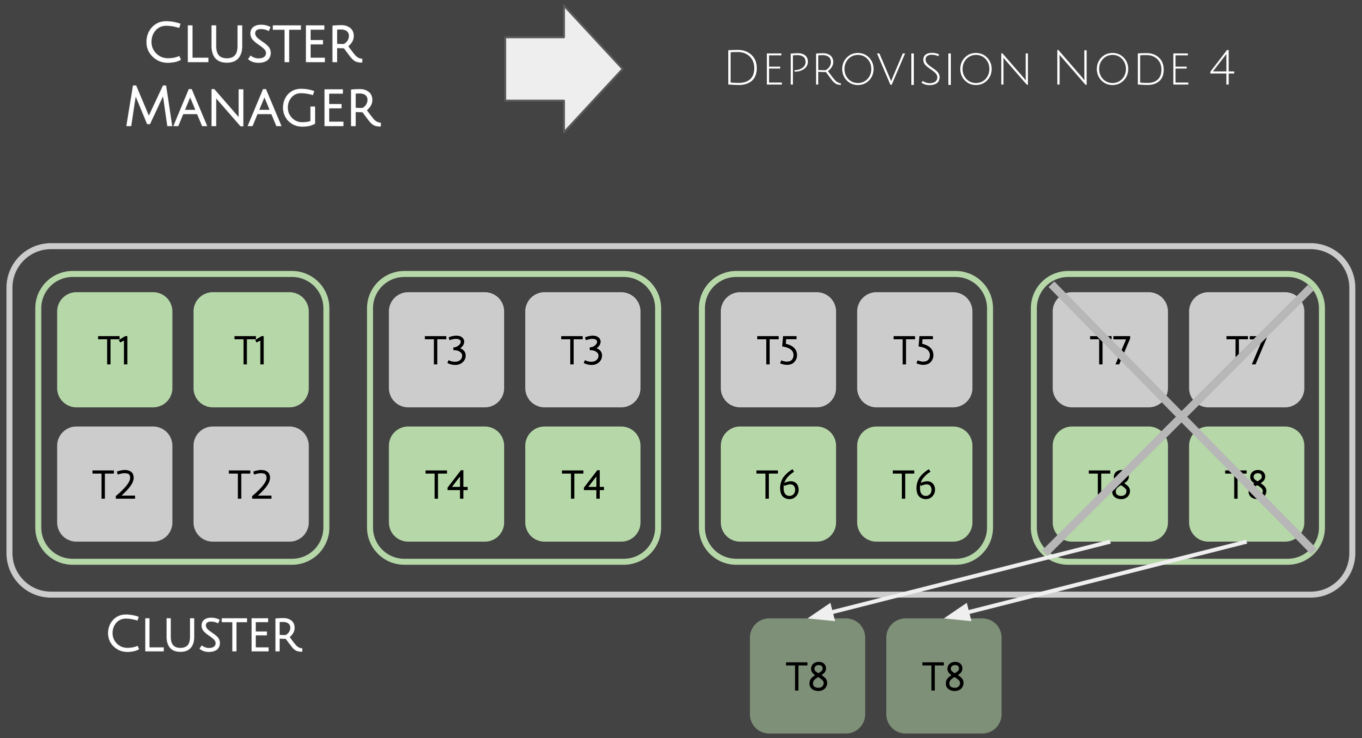
Scheduler

STOP – T2, T3, T5, T7
Continue – T1, T4, T6, T8

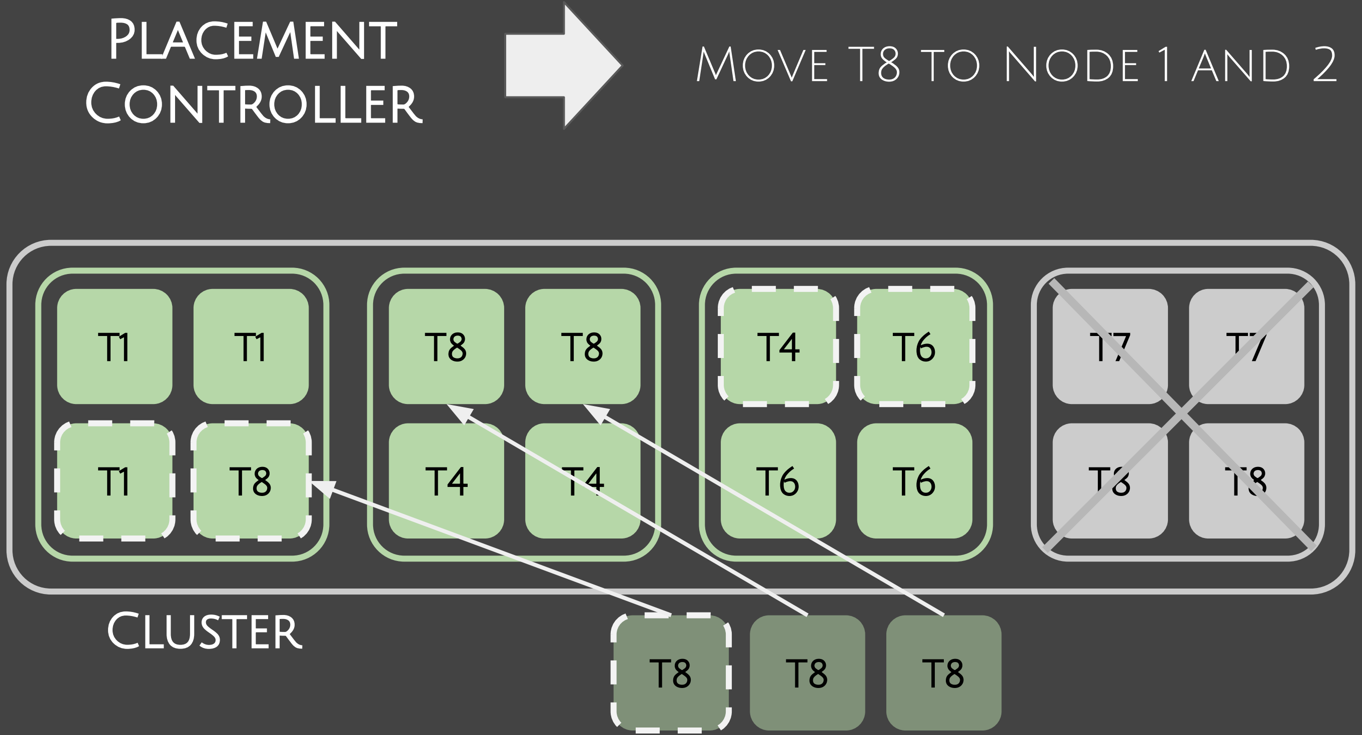


Cluster

Issue 3: effectively execute allocation plan



Issue 3: effectively execute allocation plan

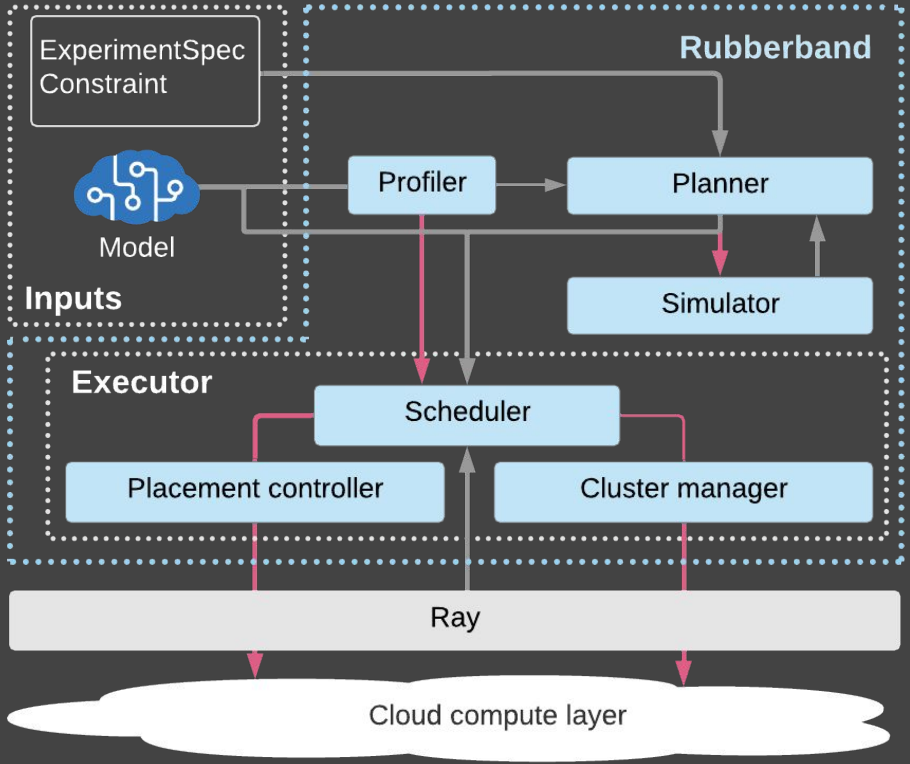


System

1. **Profiler** and **Simulator** model job completion time + cost of potential allocations

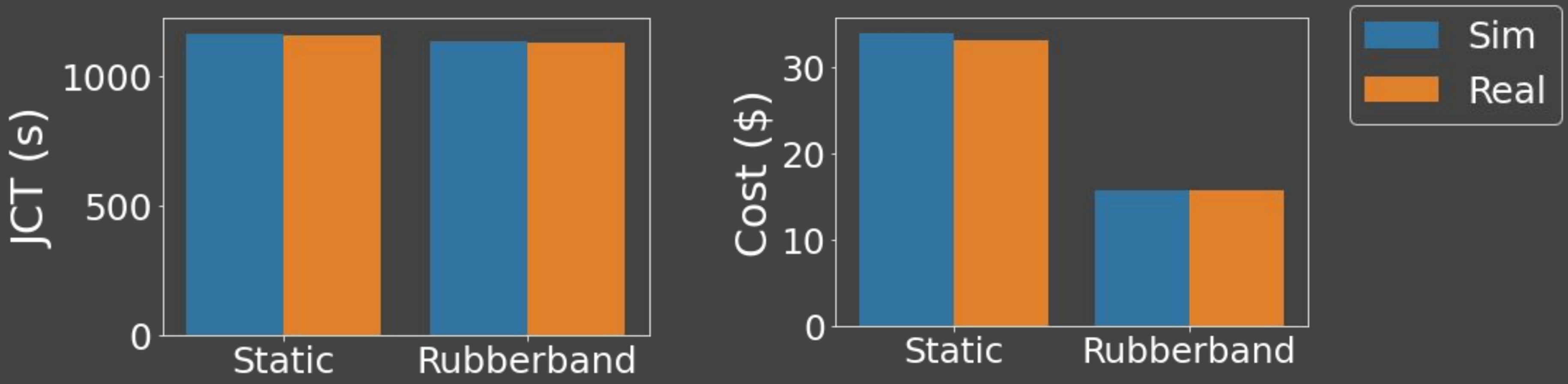
2. **Planner** generates a low cost allocation plan that completes on time

3. **Scheduler, Placement controller** and **Cluster manager** executes the allocation plan such that worker co-location and cluster utilization are maximized

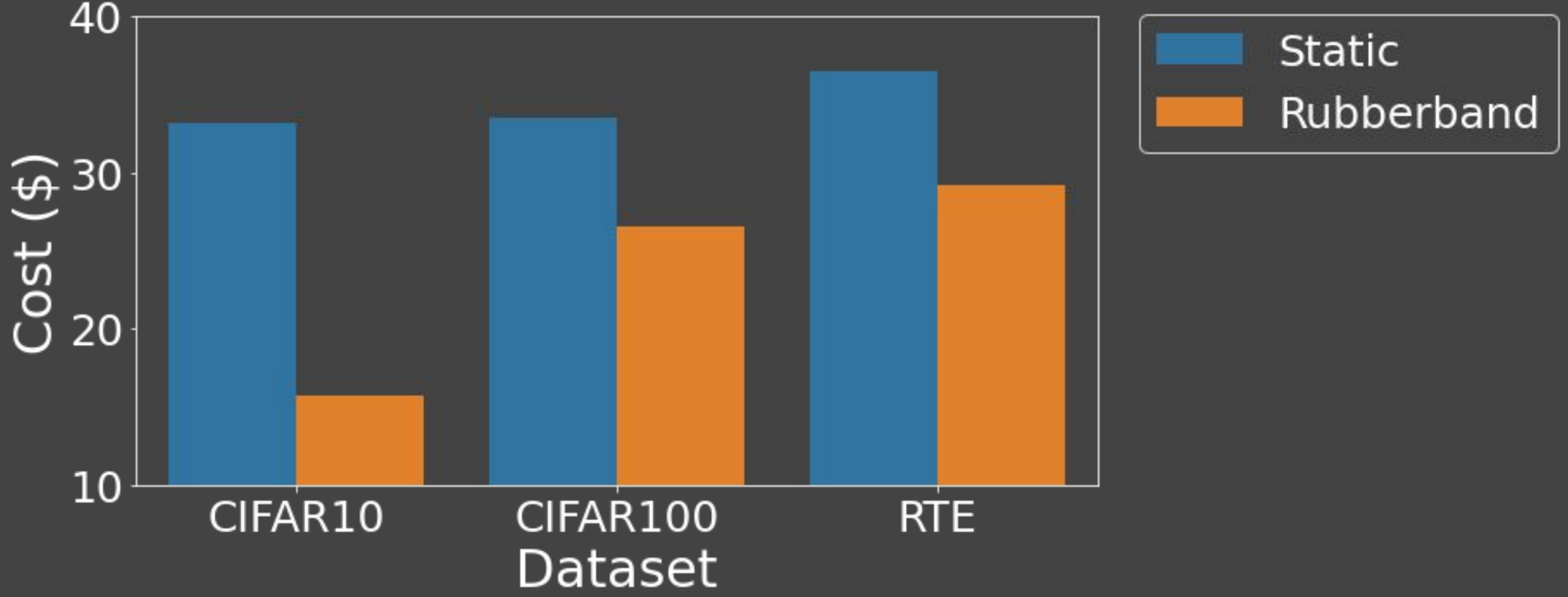


End-to-End Results

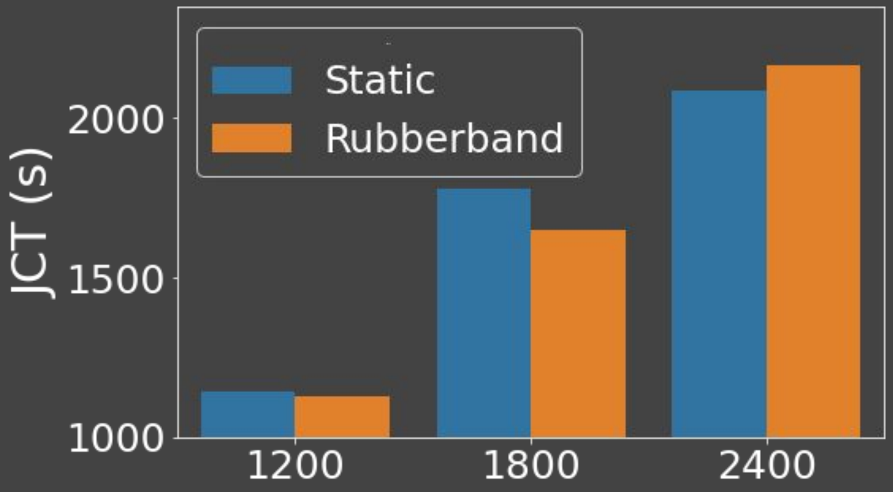
Simulation Quality

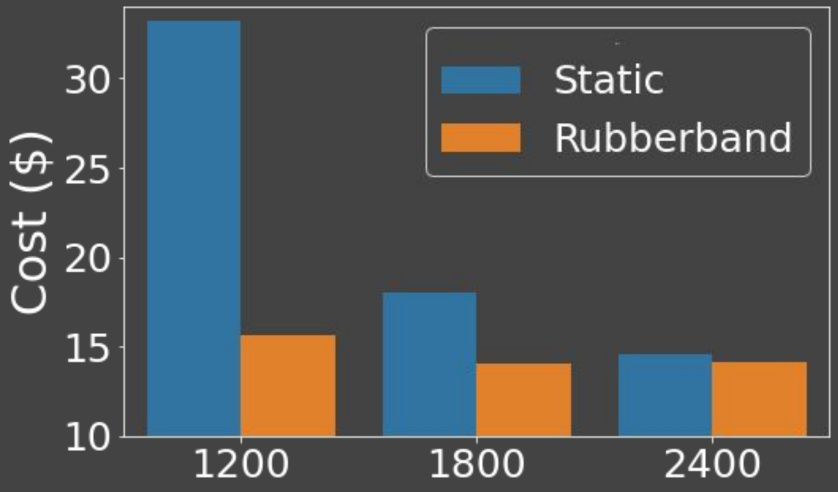


Across Datasets

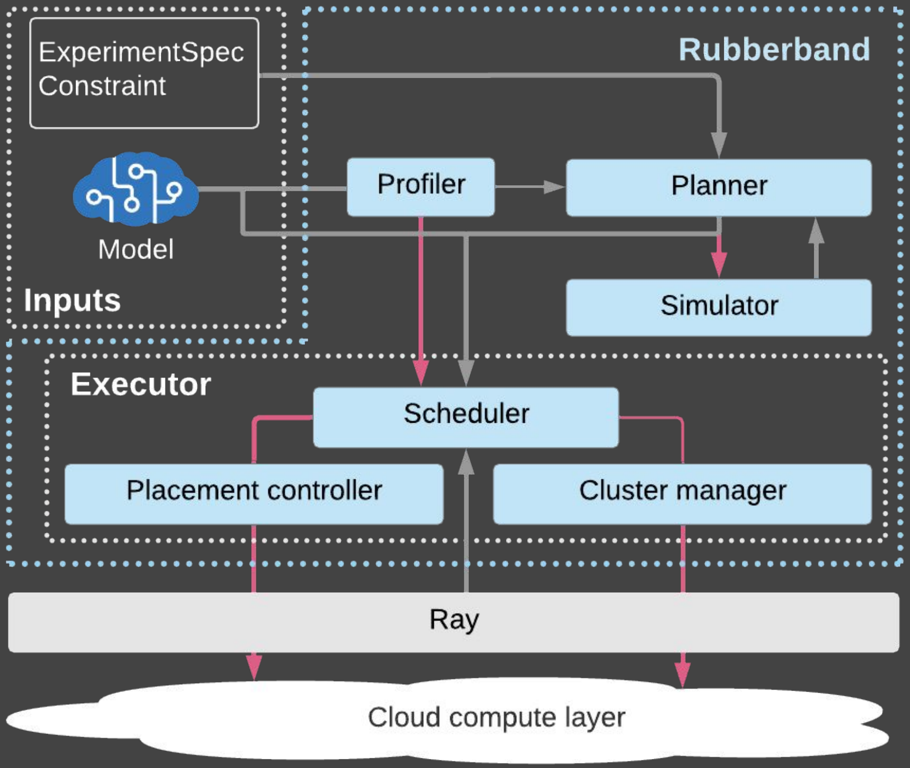


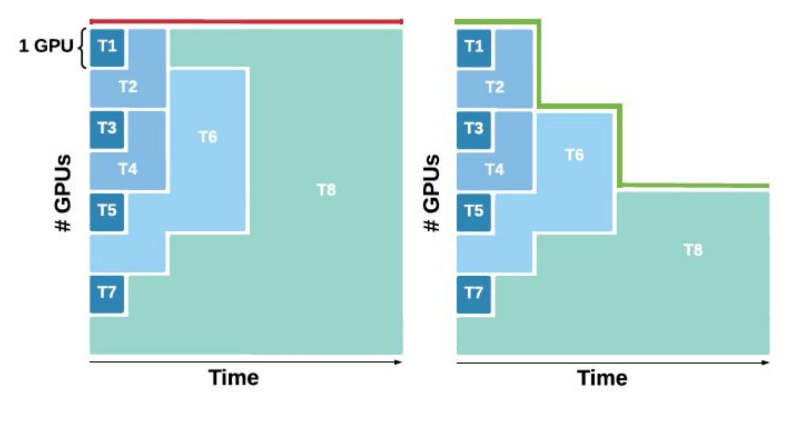
Across Deadlines





Rubberband


ExperimentSpec
Constraint
Model
Inputs
Rubberband
Profiler
Planner
Simulator
Executor
Scheduler
Placement controller
Cluster manager
Ray
Cloud compute layer


1 GPU
T1
T2
T3
T4
T5
T6
T7
T8
GPUs
Time
T1
T2
T3
T4
T5
T6
T7
T8
GPUs
Time

Thank You!