CS 744: GAVEL

Shivaram Venkataraman

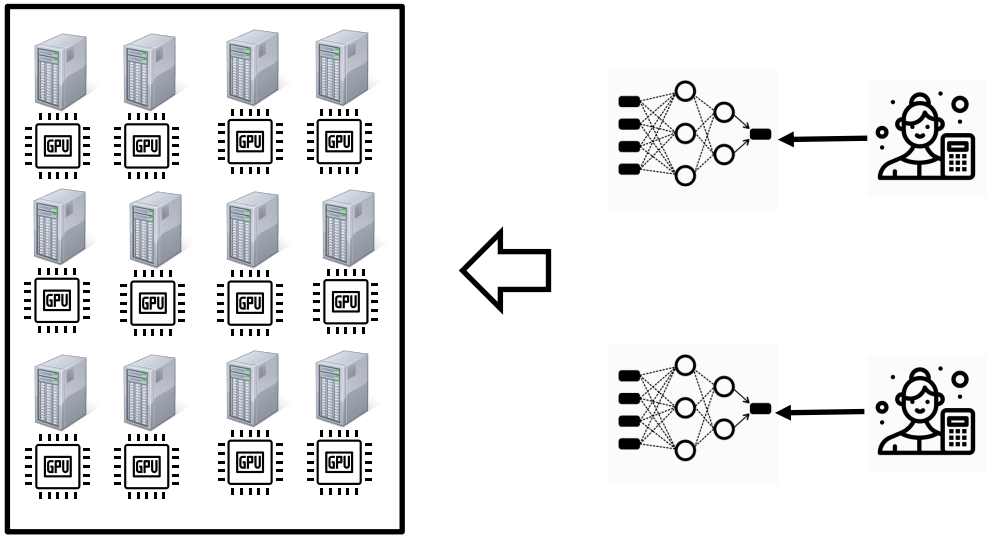
Spring 2024

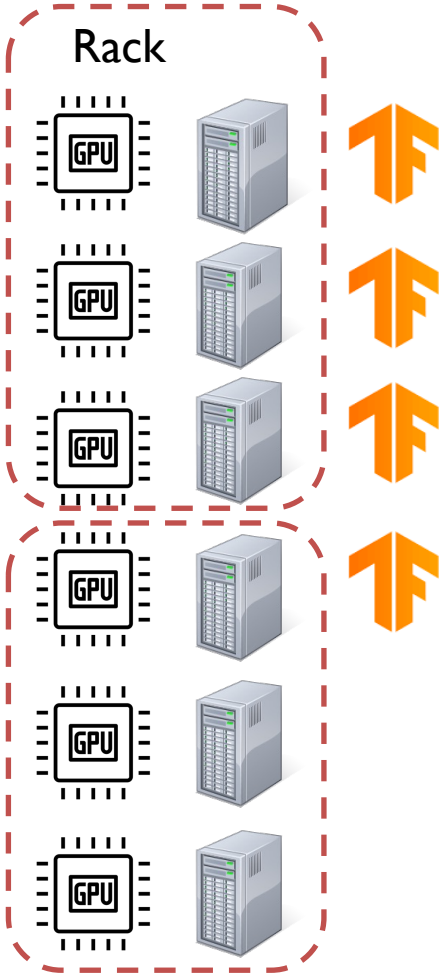
ADMINISTRIVIA

- Course project assignments
  - Emails will go out end of this week (March 1)
  - Introductions due March 8th

- Midterm Exam
  - In class on March 14th
  - Includes everything from beginning to the end of scheduling (including INFaaS)

MACHINE LEARNING: TRAINING





WORKLOAD CHARACTERISTICS

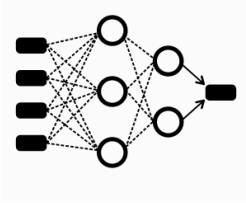
Long running tasks

Gang scheduling

Heterogeneity?

Created by Phonlaphat Thongsriphong
from Noun Project

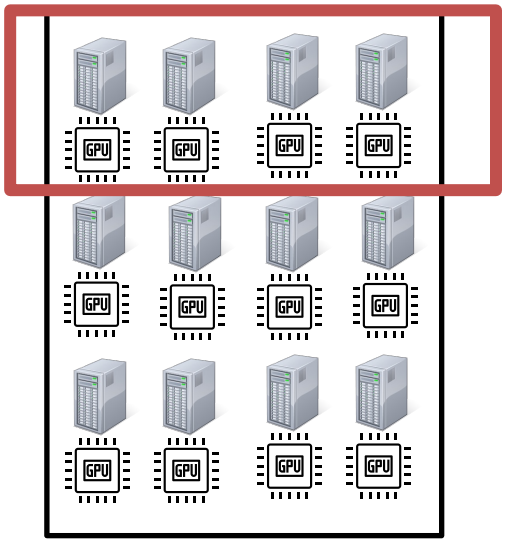
DL SCHEDULER INTERFACE

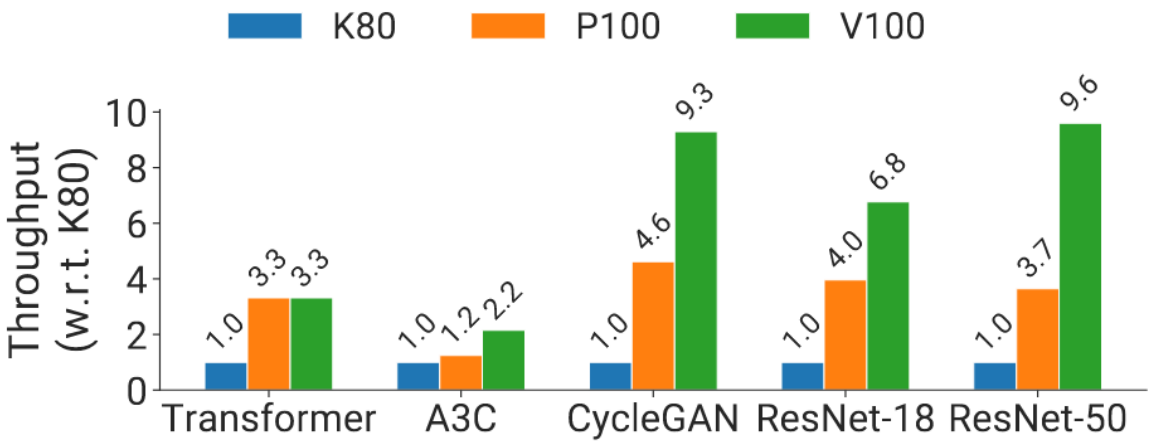


Run job Resnet18
With BatchSize = 64
on Num GPUs = 4

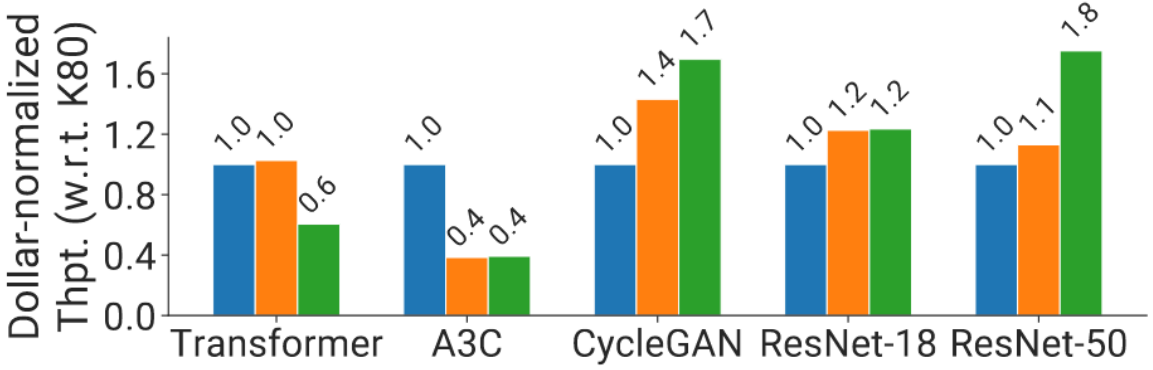
Goals:

- Maximize throughput
- Fairness
- Minimize JCT
- …





(a) Throughput.



(b) Dollar-normalized.

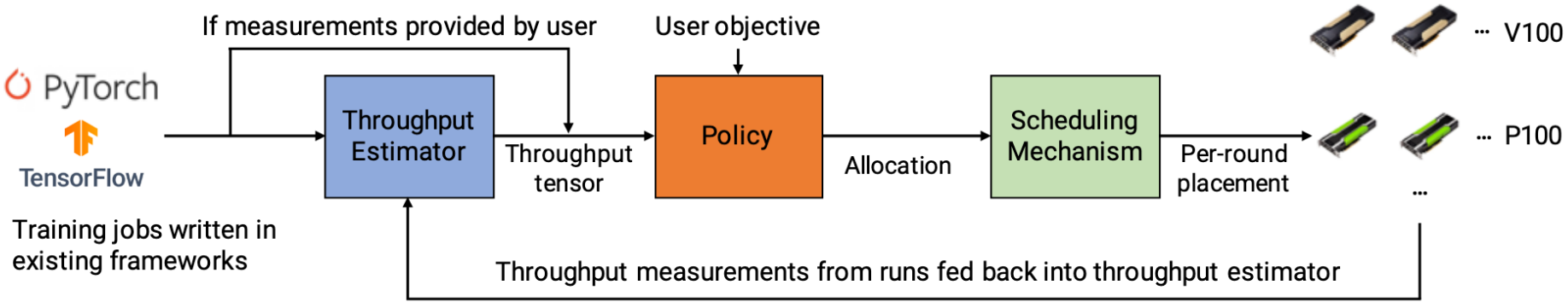
MOTIVATION: HETEROGENEITY

ADDITIONAL GOALS

- Support a wide range of objectives
  - Minimize makespan
  - Average JCT
  - Fairness (Sharing incentive)
  - …

- Placement sensitivity/Co-location

GAVEL: SYSTEM DESIGN



SCHEDULING POLICY: OPTIMIZATION PROBLEM

$$\text{Maximize}_X \sum_{m \in \text{jobs}} \text{throughput}(m, X)$$

$$\text{throughput}(m, X) = \sum_{j \in \text{accelerator types}} T_{mj} \cdot X_{mj}$$

$$0 \leq X_{mj} \leq 1 \qquad \forall (m, j) \quad (1)$$

$$\sum_j X_{mj} \leq 1 \qquad \forall m \quad (2)$$

$$\sum_m X_{mj} \cdot \text{scale\_factor}_m \leq \text{num\_workers}_j \qquad \forall j \quad (3)$$

$$X^{\text{example}} = \begin{array}{c} \begin{matrix} V100 & P100 & K80 \end{matrix} \\ \begin{pmatrix} 0.6 & 0.4 & 0.0 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.0 & 0.8 \end{pmatrix} \end{array} \begin{array}{l} \text{job 0} \\ \text{job 1} \\ \text{job 2} \end{array}$$

POLICY: MAX-MIN FAIRNESS

Classic: Weighted max-min fairness based on accelerator hours consumed

$$\text{Maximize}_X \min_m \frac{1}{w_m} X_m$$

Gavel: Use weighted normalized effective throughputs

$$\text{Maximize}_X \min_m \frac{1}{w_m} \frac{\text{throughput}(m, X)}{\text{throughput}(m, X_m^{\text{equal}})}$$

$$\text{throughput}(m, X) = \sum_{j \in \text{accelerator types}} T_{mj} \cdot X_{mj}$$

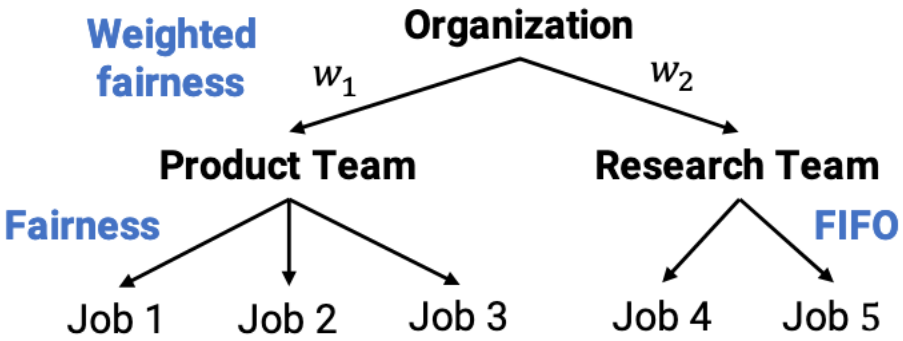
EXAMPLE

$$T = \begin{array}{c} \begin{matrix} V100 & K80 \end{matrix} \\ \begin{pmatrix} 40.0 & 10.0 \\ 12.0 & 4.0 \\ 100.0 & 50.0 \end{pmatrix} \end{array} \begin{array}{l} \\ \text{job 0} \\ \text{job 1} \\ \text{job 2} \end{array}$$

$X^{\text{hom.}}$

$$X^{\text{het.}} = \begin{array}{c} \begin{matrix} V100 & K80 \end{matrix} \\ \begin{pmatrix} 0.45 & 0.0 \\ 0.45 & 0.09 \\ 0.09 & 0.91 \end{pmatrix} \end{array} \begin{array}{l} \\ \text{job 0} \\ \text{job 1} \\ \text{job 2} \end{array}$$

HIERARCHICAL POLICIES



Share physical cluster among sub-organizations
Different policies at levels of hierarchy

Solve an LP problem across the organization
Weights constrained by policy within entity
(e.g., w4 = 1 and w5 = 0)

Use water-filling to remove bottlenecked jobs

MECHANISM: ROUND-BASED SCHEDULING

Schedule in "rounds" – every round is ~6 mins

In every round:

Consider a list of schedulable jobs and X^{opt} (from policy)

Decide which jobs are chosen to run in this round

Track time spent by job m on accelerator type j

Give high priority to jobs which are farthest from X^{opt}

Greedy policy that converges across rounds

MECHANISM: PRIORITIES

$$X^{\text{example}} = \begin{array}{c} \begin{matrix} V100 & P100 & K80 \end{matrix} \\ \begin{pmatrix} 0.6 & 0.4 & 0.0 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.0 & 0.8 \end{pmatrix} \end{array} \begin{matrix} \text{job 0} \\ \text{job 1} \\ \text{job 2} \end{matrix}$$

V100 | P100 | K80

$$\begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{matrix} \text{job 0} \\ \text{job 1} \\ \text{job 2} \end{matrix}$$

rounds_received_n

V100 | P100 | K80

$$\begin{pmatrix} 0.2 & \mathbf{0.4} & 0 \\ 0.2 & 0.2 & \infty \\ \infty & 0 & 0.2 \end{pmatrix} \begin{matrix} \text{job 0} \\ \text{job 1} \\ \text{job 2} \end{matrix}$$

priorities_n

Jobs placed on resources where they have high priority (marked in **red**)

V100 | P100 | K80

$$\begin{pmatrix} 3 & \mathbf{2} & 0 \\ 1 & 3 & \mathbf{1} \\ \mathbf{1} & 0 & 4 \end{pmatrix} \begin{matrix} \text{job 0} \\ \text{job 1} \\ \text{job 2} \end{matrix}$$

$\text{rounds\_received}_{n+1}$

SUMMARY

DL training workloads properties

Clusters with mix of accelerators

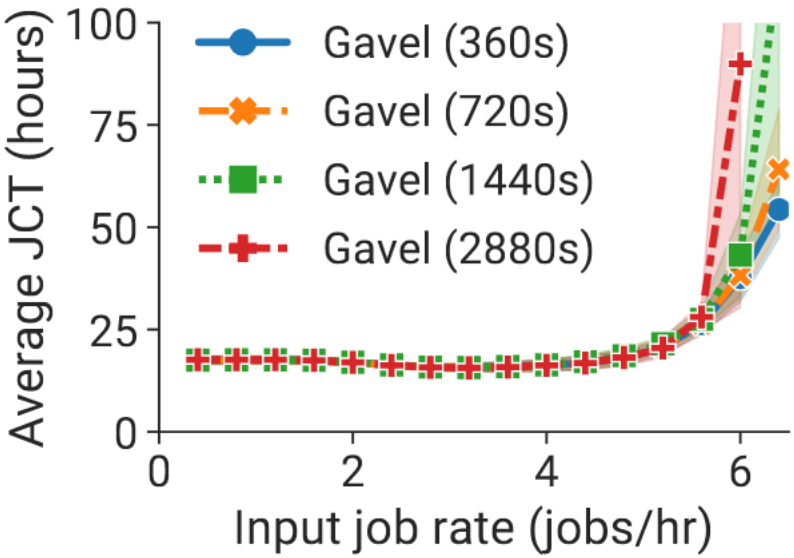
Gavel: Framework to capture many scheduling goals

Mechanism based on round-based assignments

DISCUSSION

https://forms.gle/pYnFErGi54HEHcuj7

What are some similarities or differences between Mesos/DRF and DL schedulers like Gavel?


Average JCT (hours)
100
75
50
25
0
0
2
4
6
Input job rate (jobs/hr)
Gavel (360s)
Gavel (720s)
Gavel (1440s)
Gavel (2880s)

NEXT STEPS

Next Class: INFaaS

Course Project Introductions!