CS 744: GAVEL

Shivaram Venkataraman
Fall 2022
ADMINISTRIVIA

- Course project assignments
  - Emails will go out end of this week (Oct 14th)
  - Introductions due Oct 25th

- Midterm Exam
  - In class on Oct 27th
  - Includes everything from beginning to the end of ML (including Nexus)
MACHINE LEARNING: TRAINING
Run job Resnet18
With BatchSize = 64
on Num GPUs = 4

Goals:
Maximize throughput
Fairness
Minimize JCT
...
WORKLOAD CHARACTERISTICS

- Long running tasks
- Gang scheduling
- Heterogeneity?
MOTIVATION: HETEROGENEITY
ADDITIONAL GOALS

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

- Placement sensitivity/Co-location
GAVEL: SYSTEM DESIGN

PyTorch

TensorFlow

Training jobs written in existing frameworks

Throughput Estimator

Throughput measurements from runs fed back into throughput estimator

If measurements provided by user

User objective

Policy

Throughput tensor

Allocation

Scheduling Mechanism

Per-round placement

... V100

... P100
SCHEDULING POLICY: OPTIMIZATION PROBLEM

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

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

\[
0 \leq X_{m,j} \leq 1 \quad \forall (m, j) \quad (1)
\]

\[
\sum_j X_{m,j} \leq 1 \quad \forall m \quad (2)
\]

\[
\sum_m X_{m,j} \cdot \text{scale\_factor}_m \leq \text{num\_workers}_j \quad \forall j \quad (3)
\]

\[
X^{\text{example}} = \begin{pmatrix}
V100 & P100 & K80 \\
0.6 & 0.4 & 0.0 \\
0.2 & 0.6 & 0.2 \\
0.2 & 0.0 & 0.8
\end{pmatrix}
\]

job 0

job 1

job 2
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

\[ T = \begin{pmatrix} V_{100} & K_{80} \\ 40.0 & 10.0 \\ 12.0 & 4.0 \\ 100.0 & 50.0 \end{pmatrix} \]

job 0

job 1

job 2

\[ X^{\text{hom.}} = \begin{pmatrix} \end{pmatrix} \]

\[ X^{\text{het.}} = \begin{pmatrix} V_{100} & K_{80} \\ 0.45 & 0.0 \\ 0.45 & 0.09 \\ 0.09 & 0.91 \end{pmatrix} \]

job 0

job 1

job 2
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., \(w_4 = 1\) and \(w_5 = 0\))

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

\[ X_{\text{example}} = \begin{pmatrix} 0.6 & 0.4 & 0.0 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.0 & 0.8 \end{pmatrix} \]

\text{job 0}

\text{job 1}

\text{job 2}

\begin{array}{c|c|c|c}
\text{V100} & \text{P100} & \text{K80} & \text{rounds\_received}_n \\
\hline
3 & 1 & 0 & \text{job 0} \\
1 & 3 & 0 & \text{job 1} \\
0 & 0 & 4 & \text{job 2} \\
\end{array}

\begin{array}{c|c|c|c}
\text{V100} & \text{P100} & \text{K80} & \text{priorities}_n \\
\hline
0.2 & 0.4 & 0 & \text{job 0} \\
0.2 & 0.2 & \infty & \text{job 1} \\
\infty & 0 & 0.2 & \text{job 2} \\
\end{array}

\begin{array}{c|c|c|c}
\text{V100} & \text{P100} & \text{K80} & \text{rounds\_received}_{n+1} \\
\hline
3 & 2 & 0 & \text{job 0} \\
1 & 3 & 1 & \text{job 1} \\
1 & 0 & 4 & \text{job 2} \\
\end{array}

Jobs placed on resources where they have high priority (marked in red)
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/Y5J5NrD4ZwoKGjw76
What are some similarities or differences between Mesos/DRF and DL schedulers like Gavel?
Next Class: Nexus
Course Project Introductions!
Midterm after that!