LEARNED
QUERY OPTIMIZATION

ML/RL

Query Optimizer

Query

Latency

Database

Results
### Issues with Prior Approaches and Benefits of BAO

<table>
<thead>
<tr>
<th>PRIOR APPROACHES</th>
<th>BAO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Long training time</td>
<td>1. Short training time</td>
</tr>
<tr>
<td>2. Inability to adjust to database and workload changes</td>
<td>2. Robustness to schema, data, and</td>
</tr>
<tr>
<td></td>
<td>workload changes</td>
</tr>
<tr>
<td>3. Tail catastrophe</td>
<td>3. Improved tail latency</td>
</tr>
<tr>
<td>4. Black-box decisions</td>
<td>4. Interpretability and easier debugging</td>
</tr>
<tr>
<td>5. Integration cost</td>
<td>5. Low integration cost</td>
</tr>
</tbody>
</table>
THE BIG IDEA

Query latency (seconds)

<table>
<thead>
<tr>
<th>JOB Query</th>
<th>PostgreSQL</th>
<th>PostgreSQL (no loop join)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16b</td>
<td>60.2s</td>
<td>21.2s</td>
</tr>
<tr>
<td>24b</td>
<td>0.4s</td>
<td>19.7s</td>
</tr>
</tbody>
</table>
BAO, THE BANDIT OPTIMIZER
SELECTING QUERY PLAN

Thompson Sampling
THE PREDICTIVE MODEL
THE PREDICTIVE MODEL (VECTORIZING QUERY PLAN TREES)
THE PREDICTIVE MODEL
(TREE CONVOLUTIONAL NEURAL NETWORKS)
THE PREDICTIVE MODEL (PREDICTING THE PERFORMANCE)
BAO, THE BANDIT OPTIMIZER

Parser → Query Optimizer

Hint set 1 → TCNN → Reward Predictions

Hint set 2 → Execution Engine

Hint set 3

Training → Experience

User provided
Query plan
External component
Bao
EXPERIMENTS
<table>
<thead>
<tr>
<th>Size</th>
<th>Queries</th>
<th>WL</th>
<th>Data</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDb</td>
<td>7.2 GB</td>
<td>5000</td>
<td>Dynamic</td>
<td>Static</td>
</tr>
<tr>
<td>Stack</td>
<td>100 GB</td>
<td>5000</td>
<td>Dynamic</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Corp</td>
<td>1 TB</td>
<td>2000</td>
<td>Dynamic</td>
<td>Static</td>
</tr>
</tbody>
</table>
METHODS

Data

Train  Test

Train  Test

Train  Test

Train  Test
IS BAO PRACTICAL?

1. Cost and performance on the cloud
2. Cost and performance on varied hardware
3. Tail latency analysis
4. Training time and convergence
5. Query regression analysis
6. Query optimization analysis
7. Prior learned optimizers
I. COST AND PERFORMANCE ON THE CLOUD
2. COST AND PERFORMANCE ON VARIED HARDWARE
3. TAIL LATENCY ANALYSIS
4. TRAINING TIME AND CONVERGENCE
5. QUERY REGRESSION ANALYSIS
6. QUERY OPTIMIZATION TIME

![Workload Time Chart]

- **Optimization**
- **Execution**
7. PRIOR LEARNED OPTIMIZERS
01 Do we need Neural Networks?

02 How accurate is BaO’s predictive model?

03 How long does training on GPU take?
I. DO WE NEED NEURAL NETWORKS?
2. HOW ACCURATE IS BAO’S PREDICTIVE MODEL?
3. HOW LONG DOES TRAINING ON GPU TAKES?