Implications of Public Cloud Resource Heterogeneity for Inference Serving

Jashwant Raj Gunasekaran, Cyon Subhra Mishra, Prashanth Thinakaran, Mahmut Kandemir, Chita Das

Sixth International Workshop on Serverless Computing (WoSC6)

Dec 8, 2020
## Executive Summary

<table>
<thead>
<tr>
<th>Tenants</th>
<th>Providers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster Response Times</td>
<td>Multiple Service Offerings</td>
</tr>
<tr>
<td>SLO violations, Variable Cost</td>
<td>Model Selection Resource Selection</td>
</tr>
<tr>
<td>Low Cost, SLO-aware</td>
<td>Self Managed Automated Framework</td>
</tr>
</tbody>
</table>
Model Serving Hosted on Cloud

How to optimize both model selection and resource selection?

Resources for Applications
Model Serving in Public Cloud
Model Serving Requirements

Image Classification using pretrained keras models

Latency (ms)

Accuracy (%)

M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11

Model Serving Challenges?

Response Latency

Provisioning Latency

Model Latency

Cost

Model Serving Challenges?

Top1-Accuracy

Latency

Accuracy (%)

Latency (ms)

M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11

70.00
75.00
80.00
85.00
90.00

0
100
200
300
400
Model Selection

ISO Latency

ISO Accuracy

What about resource selection?
Analyzing Prior Works

Over-provisioned VMs
SLO violations

Cost of Different Policies

VM-Cost
#VMS
Mixed

Resource Offerings from Cloud

Azure ML
Amazon SageMaker

Baremetal
Serverless Functions

Normalized VMs

Normalized Cost

SLO Violations

Berkley
WITS
Twitter
Wiki

PennState College of Engineering
Challenges with Serverless

 Arrival rate variability

 Request Rate

 Wiki WITS berkley Twitter

 Serverless Function Configuration

 Model Type

 exec_time(ms) Memory(GB) Cost($)

 Wiki Twitter

 Cost is 1.5x higher for 0.2x lower latency
What we need?

• How to make the users oblivious of model selection from the extensive pool of models?

• How to right-size VMs and appropriately configure the serverless functions?

• What is the right degree to combine serverless functions along with VMs for dynamic load?
Proposed Solutions

• Feedback-driven learning based model selection.

• Load-Based Procurement Policies

• Provisioning latency and SLO aware resource selection

• Dynamic serverless configurations.
Implementation and Evaluation

- Mxnet Framework.
- AWS resources.
- Pretrained ML models on imagenet dataset.

<table>
<thead>
<tr>
<th>Query Type</th>
<th>Memory Required (GB)</th>
<th>Memory Allocated (GB)</th>
<th>Average Execution (ms)</th>
<th>Requests per vCPU for VMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffenet</td>
<td>1.024</td>
<td>3.072</td>
<td>300</td>
<td>4</td>
</tr>
<tr>
<td>Googlenet</td>
<td>0.456</td>
<td>2.048</td>
<td>450</td>
<td>3</td>
</tr>
<tr>
<td>Squeezenet</td>
<td>0.154</td>
<td>2.048</td>
<td>130</td>
<td>6</td>
</tr>
<tr>
<td>Resnet-18</td>
<td>0.304</td>
<td>3.072</td>
<td>320</td>
<td>3</td>
</tr>
<tr>
<td>Resnet-200</td>
<td>1.024</td>
<td>3.072</td>
<td>956</td>
<td>1</td>
</tr>
<tr>
<td>Resnext-50</td>
<td>0.645</td>
<td>3.072</td>
<td>560</td>
<td>2</td>
</tr>
</tbody>
</table>
Initial Results

60% less SLO Violations.
10% reduction in deployment costs