Scale-Out Acceleration for Machine Learning

Jongse Park
Hardik Sharma
Divya Mahajan
Joon Kyung Kim
Preston Olds
Hadi Esmaeilzadeh

Alternative Computing Technologies (ACT) Lab
Georgia Institute of Technology
University of California, San Diego
Machine learning is growing exponentially!

ML model complexity
Training data size
Two Disjoint Solutions

Distributed computing

Accelerators
Scale-Out Acceleration for ML
Research Challenges

How to distribute?   How to customize?   How to coordinate?

We need a full stack!
CoSMIC

Computing Stack for ML Acceleration In the Cloud

<table>
<thead>
<tr>
<th>Layer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming layer</td>
<td>High-level mathematical language</td>
</tr>
<tr>
<td>Compilation layer</td>
<td>Accelerator operation scheduling and data mapping</td>
</tr>
<tr>
<td>System layer</td>
<td>System software orchestrating distributed accelerators</td>
</tr>
<tr>
<td>Architecture layer</td>
<td>Multi-threaded template architecture</td>
</tr>
<tr>
<td>Circuit layer</td>
<td>Constructing RTL Verilog</td>
</tr>
</tbody>
</table>
Come and learn about CoSMIC phenomenon for ML!

Session 4-A
At 2:30PM
Ballroom ABC