**Task Fusion: Improving Utilization of Multi-user Clusters**

**Motivation**
- Users Submitting Jobs
- Shared Cluster (Hadoop, etc)
- Still Waiting
- Cluster Over-Utilized!

**Solutions?**
- Better Utilization
  - Optimize Single Jobs (traditional approach)
  - Fuse Jobs Together (for data locality)

**Insight:** Load data once, run multiple analyses

**Can we automatically merge related tasks from different users?**

---

### Task Fusion

**Naive Approach**
- Maps from different tasks might output same keys, sending to wrong reducer

**Proof of concept implementation in Boa - http://boa.cs.iastate.edu/**

**When can I use task fusion?**
- Task fusion currently has the following assumptions:
  1. No side-effects.
  2. No shared state.
  3. No dependency conflicts.

**In the future...**
- Relax these assumptions:
  - Automated program transformations
  - Separate class spaces (a la OSGi)

---

### Table of Results

<table>
<thead>
<tr>
<th>Task Size</th>
<th># of Tasks</th>
<th>Times (in minutes)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>21</td>
<td>8.1m</td>
<td>0.8m</td>
</tr>
<tr>
<td>Medium</td>
<td>22</td>
<td>2.3h</td>
<td>1.8h</td>
</tr>
<tr>
<td>Large</td>
<td>18</td>
<td>4.6h</td>
<td>3.9h</td>
</tr>
<tr>
<td>Mixed</td>
<td>9</td>
<td>1.3h</td>
<td>0.9h</td>
</tr>
</tbody>
</table>

**Does task fusion increase throughput?**

**Does task fusion decrease user wait times?**

---

**Robert Dyer**

[rdyer@iastate.edu]