Data-Intensive Document Clustering on Graphics Processing Unit (GPU) Clusters

Yongpeng Zhang, Frank Mueller, Xiaohui Cui, Thomas Potok

Jay Thakkar

Dept. of Computer & Information Sciences
University of Delaware
Motivation

- Increasing number of uncategorized documents.
- Optimize an existing TF-IDF algorithm.
Contributions

- CUDA enabled GPU clusters.
- Speed-ups for entire application.
- Design highly parallelized methods to build hash tables.
TF-IDF?

- Term Frequency.
- Inverse Document Frequency.
- Used to compare documents.
Flocking-based Algorithm

- Document Clustering.

- Similar – all three; Non-similar – only separation.

- Three set of rules:
  - separation
  - alignment
  - cohesion
• Message Passing Interface

• Used to exchange data between nodes.

• Host memory – GPU memory.
- Double buffering the document raw data in the GPU.

- Overlapping hash table memory copy in the current batch with the stream pre-processing of the next batch.
Hash Tables

• Saved as linked list.

• Bucket size can be sufficiently large because of limited number of unique terms.

• Key (unsigned long)

• identity (unsigned int) – to prevent hash collisions.
With Atomic Operations

• Document stream is evenly distributed to a set of threads.

• A small number of buckets \( k \) is chosen for sorting.

• Atomic increment is performed for each new term.

• Each linked list per bucketed is accessed in mutual exclusion.
Without Atomic Operations

- Document stream is split in $M$ packets.
- Write protection is guaranteed by giving each thread a separate hash table.
- No write conflicts due to sorting by key values.
- Sub tables are merged to create the final document hash table.
Without Atomic Operation

- Domain is news article with no more than 10,000 words per document.
- TF-IDF calculation cannot start until all documents have been processed.

- High payload for communication between clusters.

- Term Frequency – Inverse Corpus Frequency (TF-ICF) has been proposed.
TF-IDF experiments.
- Clustering experiments.

Table 1

<table>
<thead>
<tr>
<th>Experiment platform</th>
<th>16 GPUs (NCSU)</th>
<th>16 CPUs (NCSU)</th>
<th>3 GPUs (ORNL)</th>
<th>3 CPUs (ORNL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>16</td>
<td>16</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>CPU cores</td>
<td>AMD Athlon Dual</td>
<td>AMD Athlon Dual</td>
<td>Intel Quad Q6700</td>
<td>Intel Quad Q6700</td>
</tr>
<tr>
<td>CPU frequency</td>
<td>2.0 GHz</td>
<td>2.0 GHz</td>
<td>2.67 GHz</td>
<td>2.67 GHz</td>
</tr>
<tr>
<td>System memory</td>
<td>1 GB</td>
<td>1 GB</td>
<td>4 GB</td>
<td>4 GB</td>
</tr>
<tr>
<td>GPU</td>
<td>16 GTX 280 s</td>
<td>Disabled</td>
<td>3 Tesla C1060</td>
<td>Disabled</td>
</tr>
<tr>
<td>GPU memory</td>
<td>1 GB</td>
<td>N/A</td>
<td>4 GB</td>
<td>N/A</td>
</tr>
<tr>
<td>Network</td>
<td>1 Gbps</td>
<td>1 Gbps</td>
<td>1 Gbps</td>
<td>1 Gbps</td>
</tr>
</tbody>
</table>
Fig. 9. Per-module performance: CPU baseline versus CUDA.
Fig. 10. Per-module contribution to overall execution time.
Fig. 11. Execution time with different corpus size.
Results

- Detection – detecting neighbor documents.
- Similarity – document similarity calculation.

Fig. 13. Speedups for similarity and detection kernels.
Results

Fig. 14. GTX 280 GPUs.

Fig. 15. Tesla C1060 GPUs.
### Table 2

Fraction of communication in GPU and CPU clusters (GPU/CPU) [in %].

<table>
<thead>
<tr>
<th>Docs (k)</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>800</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 nodes</td>
<td>74/9</td>
<td>67/8</td>
<td>64/5</td>
<td>58/3</td>
<td>52/1.5</td>
<td>49/0.9</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>8 nodes</td>
<td>67/12</td>
<td>71/11</td>
<td>65/8</td>
<td>68/6</td>
<td>62/3.5</td>
<td>56/2</td>
<td>52/1.2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12 nodes</td>
<td>67/17</td>
<td>69/12</td>
<td>68/10</td>
<td>71/8</td>
<td>68/6</td>
<td>63/3</td>
<td>57/1.4</td>
<td>54/1.2</td>
<td>NA</td>
</tr>
<tr>
<td>16 nodes</td>
<td>63/18</td>
<td>63/13</td>
<td>71/12</td>
<td>69/9</td>
<td>65/7</td>
<td>66/4.2</td>
<td>59/1.9</td>
<td>60/1.5</td>
<td>55/1.1</td>
</tr>
</tbody>
</table>