A Study of Multiple Platforms’ I/O Workload

HUONG LUU, WILLIAM GROPP, MARIANNE WINSLETT
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
We use lightweight I/O profiler to collect applications’ I/O characteristics.

Darshan is enabled by default on Mira, Edison and Blue Waters.

<table>
<thead>
<tr>
<th></th>
<th>Mira</th>
<th>Edison</th>
<th>Blue Waters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>BG/Q</td>
<td>Cray XC30</td>
<td>Cray XE6/XK7</td>
</tr>
<tr>
<td>Peak Flops</td>
<td>10 PF</td>
<td>2.57 PF</td>
<td>13.34 PF</td>
</tr>
<tr>
<td>Cores</td>
<td>768K</td>
<td>130K</td>
<td>792K+59K smx</td>
</tr>
<tr>
<td>Total Storage</td>
<td>24 PB</td>
<td>7.56 PB</td>
<td>26.4 PB</td>
</tr>
<tr>
<td>Peak I/O Throughput</td>
<td>240 GB/s</td>
<td>168 GB/s</td>
<td>963 GB/s</td>
</tr>
<tr>
<td>File System</td>
<td>GPFS</td>
<td>Lustre</td>
<td>Lustre</td>
</tr>
<tr>
<td>Time period</td>
<td>Jan – Sep’15</td>
<td>Jan – Sep’14</td>
<td>Jan – Sep’15</td>
</tr>
<tr>
<td>Number of jobs</td>
<td>152K</td>
<td>703K</td>
<td>570K</td>
</tr>
<tr>
<td>Number of apps</td>
<td>455</td>
<td>1183</td>
<td>1657</td>
</tr>
</tbody>
</table>

Table 1: Target platforms’ characteristics
These platforms have similarly distributed I/O workload.

- **Mira**
  - 44% apps with max bytes < 1 GB

- **Edison**
  - 60% apps with max bytes < 1 GB
  - 75% apps with max throughput < 1 GB/s

- **Blue Waters**
  - 55% apps with max bytes < 1 GB
  - 60% apps with max throughput < 1 GB/s
Job Size Distribution per I/O Throughput

Number of processes

Amount of data read/written

Jobs Count

- 1 - 10
- 11 - 100
- 101 - 1k
- 1k1 - 10k
- 10k1 -100k
Choices of I/O interface

- Text I/O
- POSIX I/O
- MPI-IO

Mira: 70% POSIX I/O, 30% MPI-IO
Edison: 90% POSIX I/O, 10% MPI-IO
Blue Waters: 80% POSIX I/O, 20% MPI-IO
Choices of I/O interface

![Graph showing I/O interface choices for Mira, Edison, and Blue Waters]

<table>
<thead>
<tr>
<th>I/O Throughput</th>
<th>Mira</th>
<th>Edison</th>
<th>Blue Waters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1TB/s</td>
<td>MPI</td>
<td>MPI</td>
<td>MPI</td>
</tr>
<tr>
<td>1GB/s</td>
<td>POSIX only</td>
<td>POSIX only</td>
<td>POSIX only</td>
</tr>
<tr>
<td>1MB/s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1KB/s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B/s</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- Yellow: Text I/O
- Red: POSIX I/O
- Green: MPI-I/O
Open Questions

1. What would you most like to see in these graphs that we don’t currently show?
   • Additional analysis
   • Collecting additional data to improve the interpretation of the data

2. Is there a difference in I/O performance between Lustre and GPFS based systems?
   • Does this reflect the performance capabilities and differences of Lustre and GPFS, the I/O approach of the applications workload, or is just a co-incidence?
Thank you very much for your attention!