BoF: The IO-500 and the Virtual Institute of I/O

George Markomanolis
John Bent, Julian M. Kunkel, Jay Lofstead
BoF Agenda

1. **BoF intro + The Virtual Institute for IO** (5 min) – Julian Kunkel
2. **What’s new with IO-500** (8 min) – George Markomanolis
3. **Community lightning talks** (5 min each)
   a. **In-node storage and memory-like I/O** — Adrian Jackson (EPCC)
   b. **Demonstrating GPUDirect Storage using the IO500** — CJ Newburn/Sven Oehme
4. **Analysis of the IO-500 data** (12 min) – John Bent
5. **Award ceremony** (5 min) — George Markomanolis, John Bent, Julian Kunkel, Jay Lofstead
6. **Roadmap for the IO-500** (5 min) – Julian Kunkel
7. **Voice of the community & Open Discussion** (15 min) – Jay Lofstead
What's new with IO-500

George S. Markomanolis,
The IO-500 and the Virtual Institute of I/O
Denver, Colorado, SC’19
19 November 2019
Outline

• Benchmark improvements
  – MDTest Shift
  – Fix validation with IOR

• Versioning of IO500 via Tagging
  – How IO500 identifies proper versions of IOR/Mdtest

• Webpage: Info-Creator Drop-Down

• Student Cluster Competition 19
MDtest Shift Added; IOR Shift Improved

- It was intentional to implement it since the beginning of IO-500
- Each process handles data that were accessed from another process
- It doesn’t allow local caching
- It can hurt the performance significant
- In most cases, it is more realistic
- About IOR, improved shift how ranks mapped to nodes and IOR detects its mapping, no need to specify in IOR the mapping
Versioning

• IOR/MDTest
  – We were using a tagging version lately
  – Now we use a HASH again inside prepare.sh

• IO500
  – We have a tag of the io-500-dev branch for sc19, …
    • Tag always points to the latest version
    • Changelog shows details of changes
      – SC19-v1 and in Git commit message is the name as well
  – Trying to keep the instructions the same, at least per each list
  – Keep the versions of each IO500 submission through tagging
  – Needs to be improved though
Info-Creator Drop-Down

- https://www.vi4io.org/io500-info-creator/

**Metadata server information**

- Number of nodes
- Number of storage devices in each metadata server
- Type of the storage media in metadata servers
- Volatile memory capacity
- Storage interface used by the servers
- Network interconnect on the servers
- File system software version on the servers
- Operating system software version on the servers
- Overhead of resilience in %

**Data server information**

- Number of data server nodes
- Number of storage devices
- Type of the storage media
- Volatile memory capacity
- Storage interface
- Network interconnect
- File system software version
- Operating system software version
- Overhead of resilience in %
- Whatever
- Comment

Please try to use this web page to help us analyze the data.
Supercomputing Student Cluster Competition 2019

• IO-500 is part of the Supercomputing Cluster Competition 2019 for extra credits!
• New stonewall rule only for the competitions (30 seconds)
• Drop cache option for single node submissions
• We show that IO is important and should be considered part of such competitions
• New list for such competitions will be announced this week
• For vendors: If IO-500 becomes part of SCC, maybe you would like to provide hardware for a team
The new IO-500 list and analysis
Reminder about Computing the Scores

- **IOR easy**
  - Write and read

- **IOR hard**
  - Write and read

- **Mdtest easy**
  - Create, stat, delete

- **Mdtest hard**
  - Create, read, stat, delete

- **Namespace search**
  - Find across all produced files

New metrics used for today’s analysis:

\[
\text{mdt\_consume} = \text{geo\_mean}(\text{mdt\_easy\_stat, mdt\_easy\_delete, mdt\_hard\_read, mdt\_hard\_stat, mdt\_hard\_delete})
\]

\[
\text{mdt\_produce} = \text{geo\_mean}(\text{mdtest\_easy\_create, mdttest\_hard\_create})
\]

\[
\text{mdt\_ratio} = \frac{\text{mdt\_consume}}{\text{mdt\_produce}}
\]
Three new file systems in SC19!
- DAOS
- GekkoFS
- YRCloudFile
Top Bandwidth by List

- SC17
- ISC18
- SC18
- ISC19
- SC19
All Bandwidths by Date
All Bandwidths by Date (log-scale)
Top Metadata by List

- SC17
- ISC18
- SC18
- ISC19
- SC19

Y-axis: Count (0-6000)
X-axis: Years (SC17 to SC19)
All Metadata by Date
All Metadata by Date (log-scale)
Top Score by List
All Scores by Date (log-scale)
Top Scoring Systems by List

**Ten Node**

- Top Bandwidth
- Top Metadata
- Top Overall

**All Systems**

- Top Bandwidth
- Top Metadata
- Top Overall

**Ratios**

- Top Bandwidth
- Top Metadata
- Top Overall

- SC17
- ISC18
- SC18
- ISC19
- SC19
What to do with our lists? New Lists or Merge?

- Thus far, it seems like the new rules did not affect people’s ability to improve scores
  - Suggests that perhaps we can just merge the new results into the old lists
  - But first let’s consider a bit more carefully

- Let’s zoom in on mdt_produce, mdt_consume, and mdt_ratio
  - The new rule was designed out of fear that historical mdt_consume rates artificially inflated by cache

Reminder:

```
mdt_consume = geo_mean(mdt_easy_stat, mdt_easy_delete, mdt_hard_read, mdt_hard_stat, mdt_hard_delete)

mdt_produce = geo_mean(mdtest_easy_create, mdtest_hard_create)

mdt_ratio = mdt_consume / mdt_produce
```
Top mdt_produce and mdt_consume by List

![Bar Chart]

- max(mdt_produce)
- max(mdt_consume)

- SC17
- ISC18
- SC18
- ISC19
- SC19
All mdt_produce by Date
All mdt_consume by Date

mdt_consume by date

0  2000  4000  6000  8000
All mdt_produce and mdt_consume by Date
Top mdt_ratio by List

- SC17
- ISC18
- SC18
- ISC19
- SC19
Something definitely changed in ISC19. Ratio is probably not expected to increase.
Apples:Apples Comparison of Shift Effects

- Five people submitted apples to apples results
  - Ken Carlile, Andreas Dilger, Sven Oehme, Matt Rásó-Barnett, Ruibo Wang
  - Unfortunately four were for Lustre systems, Ken’s was Vast however

- Reminder that shift was to avoid client side cache
  - IOR phases and mdtest produce phases and find should not be degraded
  - mdtest stat and read phases could be degraded since client side cache could help these
  - mdtest delete phase might be a bit less likely to be degraded since server must be involved
Five people submitted apples to apples results
- Ken Carlile, Andreas Dilger, Sven Oehme, Matt Rásó-Barnett, Ruibo Wang

Reminder that shift was added to avoid client side cache
- IOR phases and mdtest produce phases and find should not be degraded
- mdtest stat and read phases could be degraded since client side cache could help these
- mdtest delete phase might be a bit less likely to be degraded since server must be involved
- If client-side caching had been helping, we would expect to see this in mdt_ratio
Apples: Apples Results Show Effect of Change

But this was only two file systems! Ken was Vast; others were Lustre.
8 File Systems Had Both Historical and SC19 Results

---

**Average mct_ratio**

- **Pre-SC19 Submissions (i.e. without shift)**
- **SC19 Submissions (i.e. with shift)**

<table>
<thead>
<tr>
<th>File System</th>
<th>Pre-SC19</th>
<th>SC19</th>
</tr>
</thead>
<tbody>
<tr>
<td>BeeGFS</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>CephFS</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>GPFS</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>IME</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Lustre</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Qumulo</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Vast</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>WekaIO</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

Drops from blue to red bars suggest that client side caching was benefiting old results.
Four Lists Will be Maintained Going Forward

1. Full List
   a. No historical results
   b. All submissions to SC19 and beyond

2. Ranked List
   a. No historical results
   b. Multiple submissions to SC19 and beyond for “system/institution/file system” collapsed into top

3. Ten Node Ranked List
   a. No historical results
   b. Only 10 node submissions to SC19 and beyond will be included
   c. Multiple submissions for “system/institution/file system” collapsed into top submission

4. Historical List
   a. All submissions both historical and new are included

Thanks to community members Ken Carlile, Andreas Dilger, Glenn Lockwood, Sven Oehme, Matt Rásó-Barnett, and Ruibo Wang for offering valuable opinions and data to help with this key decision!
Awards

10^500
Six SC19 Awards Will Be Now Given

1. Ten-node
   a. Bandwidth
   b. Metadata
   c. Overall

2. All Systems
   a. Bandwidth
   b. Metadata
   c. Overall

Note that due to the decision about making a new list, only SC19 submissions can compete.

Even though historical bandwidths are fully compatible and were not affected by the mdtest-shift, the committee decided that the move to a new list should be complete. This will minimize any confusion as well as reduce the likelihood that incompatible results are ever inadvertently compared.
# 10 Node Challenge SC19 ONLY

This is the official list from Supercomputing 2019 for the 10 Node Challenge. The list shows the best result for a given combination of system/institution/filesystem qualifying for the 10 Node Challenge.

## 10 Node Challenge SC19 ONLY

<table>
<thead>
<tr>
<th>#</th>
<th>list id</th>
<th>institution</th>
<th>system</th>
<th>storage vendor</th>
<th>filesystem type</th>
<th>client nodes</th>
<th>client total procs</th>
<th>bw</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>10</td>
<td>310</td>
<td>123.89</td>
</tr>
<tr>
<td>2</td>
<td>sc19</td>
<td>National Supercomputing Centre, Singapore</td>
<td>Aspire 1</td>
<td>DDN</td>
<td>IME</td>
<td>10</td>
<td>160</td>
<td>101.75</td>
</tr>
<tr>
<td>3</td>
<td>sc19</td>
<td>NVIDIA</td>
<td>DGX-2H SuperPOD</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>400</td>
<td>86.97</td>
</tr>
<tr>
<td>4</td>
<td>sc19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>10</td>
<td>2610</td>
<td>56.22</td>
</tr>
<tr>
<td>5</td>
<td>sc19</td>
<td>State Key Laboratory of High-end Server &amp; Storage Technology (HSS)</td>
<td>TStor3000</td>
<td>INSPUR</td>
<td>BeeGFS</td>
<td>10</td>
<td>300</td>
<td>41.14</td>
</tr>
<tr>
<td>6</td>
<td>sc19</td>
<td>National Supercomputing Center in Changsha</td>
<td>Tianhe-2E</td>
<td>National University of Defense Technology</td>
<td>Lustre</td>
<td>10</td>
<td>160</td>
<td>35.06</td>
</tr>
<tr>
<td>7</td>
<td>sc19</td>
<td>CSIRO</td>
<td>bracewell</td>
<td>Dell/ThinkParQ</td>
<td>beegfs</td>
<td>10</td>
<td>160</td>
<td>33.77</td>
</tr>
<tr>
<td>8</td>
<td>sc19</td>
<td>Janelia Research Campus, HHMI</td>
<td>weka</td>
<td>WekaIO</td>
<td>wekaio</td>
<td>18</td>
<td>1368</td>
<td>26.22</td>
</tr>
<tr>
<td>9</td>
<td>sc19</td>
<td>University of Cambridge</td>
<td>Data Accelerator</td>
<td>Dell EMC</td>
<td>Lustre</td>
<td>10</td>
<td>320</td>
<td>21.73</td>
</tr>
<tr>
<td>10</td>
<td>sc19</td>
<td>EPCC</td>
<td>NEXTGenIO Prototype</td>
<td>BSC (NEXTGenIO) &amp; JGU (Ada-FS)</td>
<td>Adhoc Filesystem</td>
<td>10</td>
<td>240</td>
<td>21.47</td>
</tr>
</tbody>
</table>

## Sorted by BW

Sorted by BW
Certificate
IO-500 Performance Certification

This Certificate is awarded to:

Intel
#1 in the 10 Node BW Score

IO500  Nov 2019
IO-500 Steering Board

http://io500.org/list/19-11/
### 10 Node Challenge SC19 ONLY

This is the official list from Supercomputing 2019 for the 10 Node Challenge. The list shows the best result for a given combination of system/institution/filesystem qualifying for the 10 Node Challenge.

<table>
<thead>
<tr>
<th>#</th>
<th>list id</th>
<th>institution</th>
<th>system</th>
<th>storage vendor</th>
<th>filesystem type</th>
<th>client nodes</th>
<th>client total procs</th>
<th>data</th>
<th>bw</th>
<th>md</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>10</td>
<td>310</td>
<td>zip</td>
<td>123.89</td>
<td>2152.46</td>
</tr>
<tr>
<td>2</td>
<td>sc19</td>
<td>EPCC</td>
<td>NEXGenIO Prototype</td>
<td>BSC (NEXGenIO) &amp; JGU (Ada-FS)</td>
<td>Adhoc Filesystem</td>
<td>10</td>
<td>240</td>
<td>zip</td>
<td>21.47</td>
<td>728.68</td>
</tr>
<tr>
<td>3</td>
<td>sc19</td>
<td>NVIDIA</td>
<td>DGX-2H SuperPOD</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>400</td>
<td>zip</td>
<td>86.97</td>
<td>715.76</td>
</tr>
<tr>
<td>4</td>
<td>sc19</td>
<td>iFLYTEK</td>
<td>iFLYTEK</td>
<td>Yanrong</td>
<td>YRCloudFile</td>
<td>10</td>
<td>200</td>
<td>zip</td>
<td>13.55</td>
<td>455.18</td>
</tr>
<tr>
<td>5</td>
<td>sc19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td></td>
<td>10</td>
<td>2610</td>
<td>zip</td>
<td>56.22</td>
<td>435.76</td>
</tr>
<tr>
<td>6</td>
<td>sc19</td>
<td>DDN</td>
<td>AI400</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>240</td>
<td>zip</td>
<td>19.65</td>
<td>207.63</td>
</tr>
<tr>
<td>7</td>
<td>sc19</td>
<td>University of Cambridge</td>
<td>Data Accelerator</td>
<td>Dell EMC</td>
<td>Lustre</td>
<td>10</td>
<td>320</td>
<td>zip</td>
<td>21.73</td>
<td>167.09</td>
</tr>
<tr>
<td>8</td>
<td>sc19</td>
<td>State Key Laboratory of High-end Server &amp; Storage Technology (HSS)</td>
<td>TStor3000</td>
<td>INSPUR</td>
<td>BeeGFS</td>
<td>10</td>
<td>300</td>
<td>zip</td>
<td>41.14</td>
<td>165.71</td>
</tr>
<tr>
<td>9</td>
<td>sc19</td>
<td>CSIRO</td>
<td>bracewell</td>
<td>Dell/ThinkParQ</td>
<td>beegfs</td>
<td>10</td>
<td>160</td>
<td>zip</td>
<td>33.77</td>
<td>132.15</td>
</tr>
<tr>
<td>10</td>
<td>sc19</td>
<td>Janelia Research Campus, HHMI</td>
<td>weka</td>
<td>WekaIO</td>
<td>wekaio</td>
<td>18</td>
<td>1388</td>
<td>zip</td>
<td>26.22</td>
<td>90.62</td>
</tr>
</tbody>
</table>

Sorted by md
Certificate
IO-500 Performance Certification

This Certificate is awarded to:

**Intel**

#1 in the 10 Node MD Score

IO 500

Nov 2019

IO-500 Steering Board

http://io500.org/list/19-11/
# 10 node challenge - Winner

This is the official list from Supercomputing 2019 for the 10 Node Challenge. The list shows the best result for a given combination of system/institution/filesystem qualifying for the 10 Node Challenge.

<table>
<thead>
<tr>
<th>#</th>
<th>list id</th>
<th>institution</th>
<th>system</th>
<th>storage vendor</th>
<th>filesystem type</th>
<th>client nodes</th>
<th>client total procs</th>
<th>data</th>
<th>score</th>
<th>bw</th>
<th>md</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>10</td>
<td>310</td>
<td>zip</td>
<td>516.41</td>
<td>123.89</td>
<td>2152.46</td>
</tr>
<tr>
<td>2</td>
<td>sc19</td>
<td>NVIDIA</td>
<td>DGX-2H SuperPOD</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>400</td>
<td>zip</td>
<td>249.50</td>
<td>86.97</td>
<td>715.76</td>
</tr>
<tr>
<td>3</td>
<td>sc19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>10</td>
<td>2610</td>
<td>zip</td>
<td>156.51</td>
<td>56.22</td>
<td>435.76</td>
</tr>
<tr>
<td>4</td>
<td>sc19</td>
<td>EPCC</td>
<td>NEXTGenIO Prototype</td>
<td>BSC (NEXTGenIO) &amp; JGU (Ada-FS)</td>
<td>Adhoc Filesystem</td>
<td>10</td>
<td>240</td>
<td>zip</td>
<td>125.08</td>
<td>21.47</td>
<td>728.88</td>
</tr>
<tr>
<td>5</td>
<td>sc19</td>
<td>State Key Laboratory of High-end Server &amp; Storage Technology (HSS)</td>
<td>TStor3000</td>
<td>INSPUR</td>
<td>BeeGFS</td>
<td>10</td>
<td>300</td>
<td>zip</td>
<td>82.57</td>
<td>41.14</td>
<td>165.71</td>
</tr>
<tr>
<td>6</td>
<td>sc19</td>
<td>ifLYTEK</td>
<td>iFLYTEK</td>
<td>Yanrong</td>
<td>YRCloudFile</td>
<td>10</td>
<td>200</td>
<td>zip</td>
<td>78.54</td>
<td>13.55</td>
<td>455.18</td>
</tr>
<tr>
<td>7</td>
<td>sc19</td>
<td>CSIRO</td>
<td>bracewell</td>
<td>Dell/ThinkParQ</td>
<td>beegfs</td>
<td>10</td>
<td>160</td>
<td>zip</td>
<td>66.80</td>
<td>33.77</td>
<td>132.15</td>
</tr>
<tr>
<td>8</td>
<td>sc19</td>
<td>DDN</td>
<td>A400</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>240</td>
<td>zip</td>
<td>63.88</td>
<td>19.65</td>
<td>207.63</td>
</tr>
<tr>
<td>9</td>
<td>sc19</td>
<td>University of Cambridge</td>
<td>Data Accelerator</td>
<td>Dell EMC</td>
<td>Lustre</td>
<td>10</td>
<td>320</td>
<td>zip</td>
<td>60.25</td>
<td>21.73</td>
<td>167.09</td>
</tr>
<tr>
<td>10</td>
<td>sc19</td>
<td>National Supercomputing Center in Changsha</td>
<td>Tianhe-2E</td>
<td>National University of Defense Technology</td>
<td>Lustre</td>
<td>10</td>
<td>160</td>
<td>zip</td>
<td>52.58</td>
<td>35.06</td>
<td>78.86</td>
</tr>
</tbody>
</table>

Sorted by score.
Certificate
IO-500 Performance Certification

This Certificate is awarded to:

Intel
#1 in the 10 Node Challenge

IO 500
Nov 2019

10-500 Steering Board

http://io500.org/list/19-11/
# Full List

This is the full list from Supercomputing 2019. The list shows all submissions.

<table>
<thead>
<tr>
<th>#</th>
<th>list id</th>
<th>institution</th>
<th>system</th>
<th>storage vendor</th>
<th>filesystem type</th>
<th>client nodes</th>
<th>client total proc</th>
<th>data</th>
<th>bw (GiB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sc19</td>
<td>National Supercomputing Center in Changsha</td>
<td>Tianhe-2E</td>
<td>National University of Defense Technology</td>
<td>Lustre</td>
<td>480</td>
<td>5280</td>
<td>zip</td>
<td>209.43</td>
</tr>
<tr>
<td>2</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>26</td>
<td>728</td>
<td>zip</td>
<td>183.36</td>
</tr>
<tr>
<td>3</td>
<td>sc19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>345</td>
<td>8625</td>
<td>zip</td>
<td>174.74</td>
</tr>
<tr>
<td>4</td>
<td>sc19</td>
<td>University of Cambridge</td>
<td>Data Accelerator</td>
<td>Dell EMC</td>
<td>Lustre</td>
<td>128</td>
<td>2048</td>
<td>zip</td>
<td>131.25</td>
</tr>
<tr>
<td>5</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>10</td>
<td>310</td>
<td>zip</td>
<td>123.88</td>
</tr>
<tr>
<td>6</td>
<td>sc19</td>
<td>National Supercomputing Centre, Singapore</td>
<td>Aspire 1</td>
<td>DDN</td>
<td>IME</td>
<td>10</td>
<td>160</td>
<td>zip</td>
<td>101.76</td>
</tr>
<tr>
<td>7</td>
<td>sc19</td>
<td>NVIDIA</td>
<td>DGX-2H SuperPOD</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>400</td>
<td>zip</td>
<td>85.97</td>
</tr>
<tr>
<td>8</td>
<td>sc19</td>
<td>CEA</td>
<td>Tera-1000</td>
<td>DDN</td>
<td>Lustre</td>
<td>128</td>
<td>4096</td>
<td>zip</td>
<td>81.01</td>
</tr>
<tr>
<td>9</td>
<td>sc19</td>
<td>CSIRO</td>
<td>bracewell scratch2</td>
<td>Dell/ThinkParQ</td>
<td>beegfs</td>
<td>26</td>
<td>260</td>
<td>zip</td>
<td>69.10</td>
</tr>
<tr>
<td>10</td>
<td>sc19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>10</td>
<td>2610</td>
<td>zip</td>
<td>56.22</td>
</tr>
</tbody>
</table>
Certificate
IO-500 Performance Certification

This Certificate is awarded to:
National Supercomputing Center in Changsha
#1 in the IO-500 BW Score

IO 500 Nov 2019
IO-500 Steering Board

http://io500.org/list/19-11/
### Full List - Metadata Winner

This is the full list from Supercomputing 2019. The list shows all submissions.

<table>
<thead>
<tr>
<th>#</th>
<th>list id</th>
<th>institution</th>
<th>system</th>
<th>storage vendor</th>
<th>filesystem type</th>
<th>client nodes</th>
<th>client total procs</th>
<th>data</th>
<th>bw</th>
<th>md</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GIB/s</td>
<td>kIOP/s</td>
</tr>
<tr>
<td>1</td>
<td>sc19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>345</td>
<td>8625</td>
<td>zip</td>
<td>174.74</td>
<td>5045.33</td>
</tr>
<tr>
<td>2</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>26</td>
<td>728</td>
<td>zip</td>
<td>183.36</td>
<td>4753.79</td>
</tr>
<tr>
<td>3</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>10</td>
<td>310</td>
<td>zip</td>
<td>123.89</td>
<td>2152.46</td>
</tr>
<tr>
<td>4</td>
<td>sc19</td>
<td>National Supercomputing Center in Changsha</td>
<td>Tianhe-2E</td>
<td>National University of Defense Technology</td>
<td>Lustre</td>
<td>480</td>
<td>5280</td>
<td>zip</td>
<td>209.43</td>
<td>982.78</td>
</tr>
<tr>
<td>5</td>
<td>sc19</td>
<td>EPCC</td>
<td>NEXTGenIO Prototype</td>
<td>BSC (NEXTGenIO) &amp; JGU (Ada-FS)</td>
<td>Adhoc Filesystem</td>
<td>10</td>
<td>240</td>
<td>zip</td>
<td>21.47</td>
<td>728.68</td>
</tr>
<tr>
<td>6</td>
<td>sc19</td>
<td>NVIDIA</td>
<td>DGX-2H SuperPOD</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>400</td>
<td>zip</td>
<td>86.97</td>
<td>715.76</td>
</tr>
<tr>
<td>7</td>
<td>sc19</td>
<td>CEA</td>
<td>Tera-1000</td>
<td>DDN</td>
<td>Lustre</td>
<td>128</td>
<td>4096</td>
<td>zip</td>
<td>81.01</td>
<td>545.74</td>
</tr>
<tr>
<td>8</td>
<td>sc19</td>
<td>iFLYTEK</td>
<td>iFLYTEK</td>
<td>Yanrong</td>
<td>YRCloudFile</td>
<td>10</td>
<td>200</td>
<td>zip</td>
<td>13.55</td>
<td>455.18</td>
</tr>
<tr>
<td>9</td>
<td>sc19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>10</td>
<td>2610</td>
<td>zip</td>
<td>56.22</td>
<td>435.76</td>
</tr>
<tr>
<td>10</td>
<td>sc19</td>
<td>University of Cambridge</td>
<td>Data Accelerator</td>
<td>Dell EMC</td>
<td>Lustre</td>
<td>128</td>
<td>2048</td>
<td>zip</td>
<td>131.25</td>
<td>401.13</td>
</tr>
</tbody>
</table>
Certificate
IO-500 Performance Certification

This Certificate is awarded to:

WekalO
#1 in the IO-500 MD Score

IO 500
Nov 2019
IO-500 Steering Board

http://io500.org/list/19-11/
### Full List - Winner

This is the full list from Supercomputing 2019. The list shows all submissions.

<table>
<thead>
<tr>
<th>#</th>
<th>list id</th>
<th>institution</th>
<th>system</th>
<th>storage vendor</th>
<th>filesystem type</th>
<th>client nodes</th>
<th>client total proc</th>
<th>data</th>
<th>score</th>
<th>bw</th>
<th>md</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SC19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>345</td>
<td>8625</td>
<td>zip</td>
<td>938.95</td>
<td>174.74</td>
<td>5045.33</td>
</tr>
<tr>
<td>2</td>
<td>SC19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>26</td>
<td>728</td>
<td>zip</td>
<td>933.64</td>
<td>183.36</td>
<td>4753.79</td>
</tr>
<tr>
<td>3</td>
<td>SC19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>10</td>
<td>310</td>
<td>zip</td>
<td>516.41</td>
<td>123.89</td>
<td>2152.46</td>
</tr>
<tr>
<td>4</td>
<td>SC19</td>
<td>National Supercomputing</td>
<td>Tianhe-2E</td>
<td>National University of Defense Technology</td>
<td>Lustre</td>
<td>450</td>
<td>5280</td>
<td>zip</td>
<td>453.68</td>
<td>209.43</td>
<td>952.78</td>
</tr>
<tr>
<td>5</td>
<td>SC19</td>
<td>NVIDIA</td>
<td>DGX-2H SuperPOD</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>400</td>
<td>zip</td>
<td>249.50</td>
<td>86.97</td>
<td>715.76</td>
</tr>
<tr>
<td>6</td>
<td>SC19</td>
<td>University of Cambridge</td>
<td>Data Accelerator</td>
<td>Dell EMC</td>
<td>Lustre</td>
<td>128</td>
<td>2048</td>
<td>zip</td>
<td>229.45</td>
<td>131.25</td>
<td>401.13</td>
</tr>
<tr>
<td>7</td>
<td>SC19</td>
<td>CEA</td>
<td>Tera-1000</td>
<td>DDN</td>
<td>Lustre</td>
<td>128</td>
<td>4096</td>
<td>zip</td>
<td>210.26</td>
<td>81.01</td>
<td>545.74</td>
</tr>
<tr>
<td>8</td>
<td>SC19</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO</td>
<td>WekaIO Matrix</td>
<td>10</td>
<td>2610</td>
<td>zip</td>
<td>156.51</td>
<td>56.22</td>
<td>435.76</td>
</tr>
<tr>
<td>9</td>
<td>SC19</td>
<td>EPCC</td>
<td>NEXTGenI0 Prototype</td>
<td>BSC (NEXTGenI0) &amp; JGU (Ada-FS)</td>
<td>Ad hoc Filesystem</td>
<td>10</td>
<td>240</td>
<td>zip</td>
<td>125.08</td>
<td>21.47</td>
<td>728.68</td>
</tr>
<tr>
<td>10</td>
<td>SC19</td>
<td>CSIRO</td>
<td>bracewell scratch2</td>
<td>Dell/ThinkParQ</td>
<td>beegfs</td>
<td>26</td>
<td>260</td>
<td>zip</td>
<td>94.86</td>
<td>69.10</td>
<td>130.23</td>
</tr>
</tbody>
</table>
### Full List - Winner

This is the full list from Supercomputing 2019. The list shows all submissions.

<table>
<thead>
<tr>
<th>#</th>
<th>list id</th>
<th>institution</th>
<th>system</th>
<th>storage vendor</th>
<th>filesystem type</th>
<th>client nodes</th>
<th>client total procs</th>
<th>data</th>
<th>score</th>
<th>bw</th>
<th>md</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sc19</td>
<td>WekaI/O</td>
<td>WekaI/O</td>
<td>WekaI/O</td>
<td>WekaI/O Matrix</td>
<td>345</td>
<td>8625</td>
<td>zip</td>
<td>938 95</td>
<td>174.74</td>
<td>5045.33</td>
</tr>
<tr>
<td>2</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>26</td>
<td>728</td>
<td>zip</td>
<td>933.64</td>
<td>183.36</td>
<td>4763.79</td>
</tr>
<tr>
<td>3</td>
<td>sc19</td>
<td>Intel</td>
<td>Wolf</td>
<td>Intel</td>
<td>DAOS</td>
<td>10</td>
<td>310</td>
<td>zip</td>
<td>516.41</td>
<td>123.89</td>
<td>2152.46</td>
</tr>
<tr>
<td>4</td>
<td>sc19</td>
<td>National Supercomputing Center in Changsha</td>
<td>Tianhe-2E</td>
<td>National University of Defense Technology</td>
<td>Lustre</td>
<td>480</td>
<td>5280</td>
<td>zip</td>
<td>453.68</td>
<td>209.43</td>
<td>952.78</td>
</tr>
<tr>
<td>5</td>
<td>sc19</td>
<td>NVIDIA</td>
<td>DGX-2H SuperPOD</td>
<td>DDN</td>
<td>Lustre</td>
<td>10</td>
<td>400</td>
<td>zip</td>
<td>249.50</td>
<td>86.97</td>
<td>715.76</td>
</tr>
<tr>
<td>6</td>
<td>sc19</td>
<td>University of Cambridge</td>
<td>Data Accelerator</td>
<td>Dell EMC</td>
<td>Lustre</td>
<td>128</td>
<td>2048</td>
<td>zip</td>
<td>229.45</td>
<td>131.25</td>
<td>401.13</td>
</tr>
<tr>
<td>7</td>
<td>sc19</td>
<td>CEA</td>
<td>Tera-1000</td>
<td>DDN</td>
<td>Lustre</td>
<td>128</td>
<td>4096</td>
<td>zip</td>
<td>210.26</td>
<td>81.01</td>
<td>545.74</td>
</tr>
<tr>
<td>8</td>
<td>sc19</td>
<td>WekaI/O</td>
<td>WekaI/O</td>
<td>WekaI/O</td>
<td>WekaI/O Matrix</td>
<td>10</td>
<td>2610</td>
<td>zip</td>
<td>156.51</td>
<td>56.22</td>
<td>435.76</td>
</tr>
<tr>
<td>9</td>
<td>sc19</td>
<td>EPCC</td>
<td>NEXTGenI/O Prototype</td>
<td>BSC (NEXTGenI/O) &amp; JGU (Ada-FS)</td>
<td>Adhoc Filesystem</td>
<td>10</td>
<td>240</td>
<td>zip</td>
<td>125.06</td>
<td>21.47</td>
<td>728.68</td>
</tr>
<tr>
<td>10</td>
<td>sc19</td>
<td>CSIRO</td>
<td>bracewell scratch2</td>
<td>Dell/ThinkParQ</td>
<td>beegfs</td>
<td>26</td>
<td>260</td>
<td>zip</td>
<td>94.86</td>
<td>69.10</td>
<td>130.23</td>
</tr>
</tbody>
</table>

Sorted by score

0.57% difference
Certificate
IO-500 Performance Certification

This Certificate is awarded to:

WekalO

#1 in the IO-500

IO 500
Nov 2019

http://io500.org/list/19-11/
## SC19 List of Awarded Systems

<table>
<thead>
<tr>
<th></th>
<th>Metadata</th>
<th>Bandwidth</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10-Node</strong></td>
<td>Intel</td>
<td>DAOS</td>
<td>2152 KIOPS</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>Intel</td>
<td>DAOS</td>
<td>124 GiB/s</td>
</tr>
<tr>
<td>Overall</td>
<td>Intel</td>
<td>DAOS</td>
<td>516 score</td>
</tr>
<tr>
<td><strong>All Systems</strong></td>
<td>WekaIO</td>
<td>Matrix</td>
<td>5045 KIOPS</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>Tianhe</td>
<td>Lustre</td>
<td>209 GiB/s</td>
</tr>
<tr>
<td>Overall</td>
<td>WekaIO</td>
<td>Matrix</td>
<td>939 score</td>
</tr>
</tbody>
</table>
SC19 IO500
Winners and Honorable Mentions

<table>
<thead>
<tr>
<th>10-Node</th>
<th>Metadata</th>
<th>Intel</th>
<th>DAOS</th>
<th>2152</th>
<th>KIOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td>Intel</td>
<td>DAOS</td>
<td>124</td>
<td>GiB/s</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>Intel</td>
<td>DAOS</td>
<td>516</td>
<td>score</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All Systems</th>
<th>Metadata</th>
<th>WekaIO</th>
<th>Matrix</th>
<th>5045</th>
<th>KIOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td>Tianhe</td>
<td>Lustre</td>
<td>209</td>
<td>GiB/s</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>WekaIO</td>
<td>Matrix</td>
<td>939</td>
<td>score</td>
<td></td>
</tr>
</tbody>
</table>

Top Five SC19 10 Node Challenge Systems: IO500 Scores

Top Five SC19 Systems: IO500 Scores
# SC Student Cluster Competition Preliminary results*

*Results received less than 18 hours ago and have not been fully validated by the committee yet.

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Team</th>
<th>Score</th>
<th>BW</th>
<th>MD</th>
<th>Nodes</th>
<th>DropCache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nanyang</td>
<td>28.28</td>
<td>4.139</td>
<td>193.21</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>Purdue</td>
<td>26.0845</td>
<td>9.8</td>
<td>69.4</td>
<td>5</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>Shanghai Tech</td>
<td>20.4175</td>
<td>2.61</td>
<td>159.381</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>ETH Zurich</td>
<td>19.588</td>
<td>3.23</td>
<td>118.575</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>Peking</td>
<td>16.21</td>
<td>2.54</td>
<td>103.2</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>FAU</td>
<td>13.82</td>
<td>2.02</td>
<td>94.4</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>UIUC</td>
<td>12.69</td>
<td>6.5</td>
<td>24.77</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>NTHU</td>
<td>9.4</td>
<td>1.05</td>
<td>84.16</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>Tennessee</td>
<td>8.96</td>
<td>1.85</td>
<td>43.38</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>Wake Forest</td>
<td>5.44</td>
<td>0.98</td>
<td>30.2</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>Warsaw</td>
<td>4.95</td>
<td>1.33</td>
<td>18.38</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>12</td>
<td>Washington</td>
<td>4.39</td>
<td>0.248</td>
<td>77.916</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>Shanghai Jiao Tong</td>
<td>3.279</td>
<td>1.188</td>
<td>9.05</td>
<td>1</td>
<td>yes</td>
</tr>
</tbody>
</table>

Not valid

<table>
<thead>
<tr>
<th>Team</th>
<th>Score</th>
<th>BW</th>
<th>MD</th>
<th>Nodes</th>
<th>DropCache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tsinghua</td>
<td>30.55</td>
<td>3.08</td>
<td>303.121</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>NC State</td>
<td>24.64</td>
<td>2.39</td>
<td>253.351</td>
<td>1</td>
<td>no</td>
</tr>
</tbody>
</table>
Roadmap
Roadmap for the IO-500

- Rewrite the IO-500 into a C-application instead of the script solution
  - Run by using a configuration file and no additional arguments
  - Improved error handling and validation for submitters
  - Produces the same results as the current bash solution

- Using MDTest data validation during mdtest hard read
  - Compares read data with the expectation

- Integration of tools to automatically harvest system configuration

- Rewrite the webpage
  - Move 100% of code into github
  - Mirror at io500.vi4io.org and io500.org
C-Application / Thoughts

- Running the application should be as simple as (e.g. SLURM)

  ```bash
  #!/bin/bash -e
  #SBATCH -p compute2
  #SBATCH --nodes=10
  module load OpenMPI
  mpiexec -np 20 ./io500 final-config.ini
  ```

- Configuration could be INI or JSON files
  - Providing only options that are allowed to tune; options for additional testing in extra section

- The tool provides a dry-run option showing the exact commands it runs
  - e.g. `mpiexec -np 20 ./io500 --dry-run final-config.ini`
  - Dumps the full INI options that are available and their current values
  - Shows predicted execution behavior:
    - I run mdtest with these arguments, then this then that....
    - The result should be valid or will definitely be invalid based on the options provided
[find]
external-program = ./bin/mmfind.sh # wrapper returns similar output

[ior-hard]
API = MPIIO # Like when using ior -O <OPTION>=<VALUE>
hintsFileNm = my-hints.txt

[ior-hard write] # Some options might be valid for specific sections
posix.odirect = 1

[optional]
ior-random = enable

[debug] # the program will warn if anything is invalid
drop-caches = TRUE
stonewall-time = 10
Discussion

\[10^{500}\]
Edit or add functionalities to IO-500

Change Request

The IO-500 aims to be a robust and long-living benchmark. Nevertheless, the community recognizes the need to consider modifications occasional modifications such as including new access patterns, removing deprecated access patterns, or any other modifications deemed necessary by the community. Therefore, we have established a process to add further benchmarks, which works as follows:

1. A member of the community prepares a (up to) 1-page proposal for the new access pattern to include. This should include a motivation, a rough sketch of the access pattern and justification why the pattern is important. This proposal can then be sent to the community mailing list or the steering board. Deadline: 1 month before the next community meeting – at the moment, these are the birds-of-a-feather sessions at ISC or Supercomputing.
2. The steering board will give feedback to the technical quality of the proposal.
3. The member is given the opportunity to present the proposal at the next following community IO-500 meeting.
4. Given there are no technical concerns, the IO-500 benchmark will be modified for the next submission period to allow the execution of a benchmark that represents the pattern as an *optional* benchmarking step. Additionally, the optional field is introduced into subsequent lists and the changes to the benchmark are documented on the webpage.
5. The optional pattern is kept for at least two subsequent IO-500 lists and community meetings. The results and effectiveness of the new pattern are discussed during the community meetings. As a result, it may be removed, remain optional, or may become mandatory.

The committee can be reached at committee@io500.org.

http://io500.org/rules/proposals
Open Floor
Issues about Fair Comparisons

- Non-erasure vs erasure systems
- Production system versus benchmark-only system
- Vendor submission versus customer submission
- GA File system versus research file system
- Cloud vs on-prem
- Ephemeral vs persistent file system
- Storage media