Old scheduler

- Arbitrary selection of nodes
Topology-aware scheduler

- All jobs get rectangular prisms of nodes
- Minimizes interference between adjacent jobs
- [https://bluewaters.ncsa.illinois.edu/topology-aware-scheduling](https://bluewaters.ncsa.illinois.edu/topology-aware-scheduling)
Benefits

• Run-time consistency is better
• Performance approaching dedicated-system performance all the time
  • Large jobs benefit more, we saw 20% improvement
• Your allocation will go further
• You do not have to change your job script
Considerations

- Job must fit available shapes, not merely available idle nodes
- Taking advantage of backfill is different
System details

- Torus is 24x24x24 Gemini network routers
- XK block is 15x6x24 Gemini routers
- Each Gemini has 2 nodes
- Every 2 Geminis (4 nodes) in Y is fast on-board connection
- Beyond immediate neighbors, Y is slower
- Z direction is a single cabinet: high bandwidth
- X direction is medium-bandwidth
- Scheduler takes network characteristics into account
- Service nodes distributed throughout torus
Wrap-around communication

- Communication is shortest-path on the torus
- For small prisms all communication is internal
- Beyond $\frac{1}{2}$ the torus dimension (12 Geminis), some communication will wrap around
- “Gap job”: a job running next to a large job, which may have external traffic passing through
Geometry

- Scheduler models communication network performance to select best shape from available prisms
- Two multi-node jobs will not share Gemini
- Rectangular prisms are generally long in Z, short in Y
  - Long Z: maximize use of better bandwidth
  - Short Y: minimize use of poorer bandwidth
- Prism dimensions usually shorter than 12 Geminis in any dimension to avoid wrap-around
Default Geometry

• Rectangular prisms will likely not match requested number of compute nodes
  • Prism may contain service nodes
  • Prism may be bigger than requested nodes
• Scheduler only considers prisms that contain up to 15% more compute nodes than requested
• Jobs currently only charged for nodes requested, not geometry allocations
## Scheduler shape table

<table>
<thead>
<tr>
<th>Size</th>
<th>Shapes</th>
</tr>
</thead>
<tbody>
<tr>
<td>252</td>
<td>6x3x7 7x3x6</td>
</tr>
<tr>
<td>256</td>
<td>8x2x8 4x4x8</td>
</tr>
<tr>
<td>288</td>
<td>6x3x8 8x3x6 3x6x8 9x2x8 8x2x9</td>
</tr>
<tr>
<td>294</td>
<td>7x3x7</td>
</tr>
<tr>
<td>320</td>
<td>5x4x8 10x2x8 8x2x10</td>
</tr>
</tbody>
</table>

- 256-node jobs may not fit in a 8x2x8 prism
  - Service nodes
  - Down nodes
- Scheduler also considers 288 and 294-node class (15% over)
Command-line tools

• Command changes
  • qstat, showq, showres: as before
  • checkjob: shows topology information
  • showbf: includes available prism sizes
  • Total available nodes (e.g. from showq) not representative of what jobs will run, since only prisms of nodes are available for scheduling
New job communication attributes

- May improve job turn-around time by making job easier to schedule
- *Please make sure settings are appropriate: abuse will be policed*
- Contact us for assistance in determining the communication patterns in your code
New job communication attributes

• #PBS -f flags=commtolerant
  • Allow job to be scheduled next to a large job (i.e. as a “gap job”)
  • External communication may route through job
  • Intended for low-communication jobs
New job communication attribute

- #PBS -f flags=commlocal
  - Promises that the job will not generate much traffic outside immediate neighboring nodes
  - If scheduler gives job a wide prism (one that wraps), not much traffic will be generated
  - “Gap jobs” scheduled next to it should experience limited route-through impact
- #PBS -f flags=commtolerant:commlocal
  - Apply both flags
  - Good for collections of single-node jobs
Specifying geometry

- `#PBS -l geometry=XxYxZ`
- `#PBS -l geometry=8x4x12/12x4x8/8x6x8`
- Capable of also specifying origin: not recommended
- References Gemini coordinates, not nodes
- Allow 5-10% extra nodes (service nodes, down nodes)
- Too many idle nodes will cause job to be held (30% threshold)
When to use geometry

- Latency sensitivity: shorter paths more important than bandwidth
- Take advantage of a specific backfill opportunity
- Problem fits a particular geometry
- Maximize rank reordering benefit
Using Backfill

- `showbf` shows available geometries
- Job may not be automatically considered for some shapes
- `-l geometry` can override automatic shapes
- Debundle jobs to improve backfill opportunities
## Backfill example

```
showbf -p nid11293 -f xe  
<table>
<thead>
<tr>
<th>Partition</th>
<th>Tasks</th>
<th>Nodes</th>
<th>Duration</th>
<th>StartOffset</th>
<th>StartDate</th>
<th>Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>nid11293</td>
<td>3584</td>
<td>224</td>
<td>23:44:33</td>
<td>00:00:00</td>
<td>22:50:05_10/29</td>
<td>4x4x7</td>
</tr>
<tr>
<td>nid11293</td>
<td>1920</td>
<td>120</td>
<td>1:15:44:33</td>
<td>00:00:00</td>
<td>22:50:05_10/29</td>
<td>5x2x6</td>
</tr>
<tr>
<td>nid11293</td>
<td>1600</td>
<td>100</td>
<td>INFINITY</td>
<td>00:00:00</td>
<td>22:50:05_10/29</td>
<td>5x2x5</td>
</tr>
</tbody>
</table>
```

- 192-node job not automatically considered for 4x4x7 geometry
- **Use** `qsub -l geometry=4x4x7 myjob.pbs`
- `showbf` is a snapshot of last scheduler iteration
- Not guaranteed to fill block, due to other work, other recently-completed job
Using Rank Reordering

- **MPICH_RANK_REORDER_METHOD**
- Cray perftools and `grid_order`
- Topaware: in-house tool for enhanced rank reordering. Contact us for more information
- See Topology Considerations documentation [https://bluewaters.ncsa.illinois.edu/topology-considerations](https://bluewaters.ncsa.illinois.edu/topology-considerations)
Miscellaneous

• Debug and High reserved nodes initially unavailable (priority boost still in effect)
• Low-queue pre-emption initially disabled

Availability

• Going into production in early January
• We will wait for planned or unplanned outage
• Look for email notice