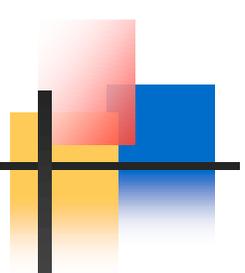


OSCAR

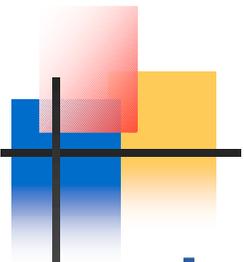
Workload Management



Jeremy Enos

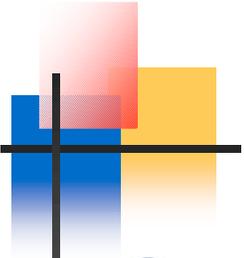
OSCAR Annual Meeting

January 10-11, 2002



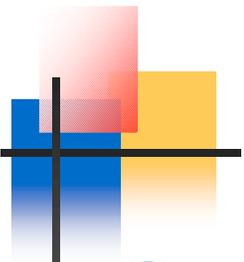
Topics

- Current Batch System – OpenPBS
- How it Works, Job Flow
- OpenPBS Pros/Cons
- Schedulers
- Enhancement Options
- Future Considerations
- Future Plans for OSCAR



OpenPBS

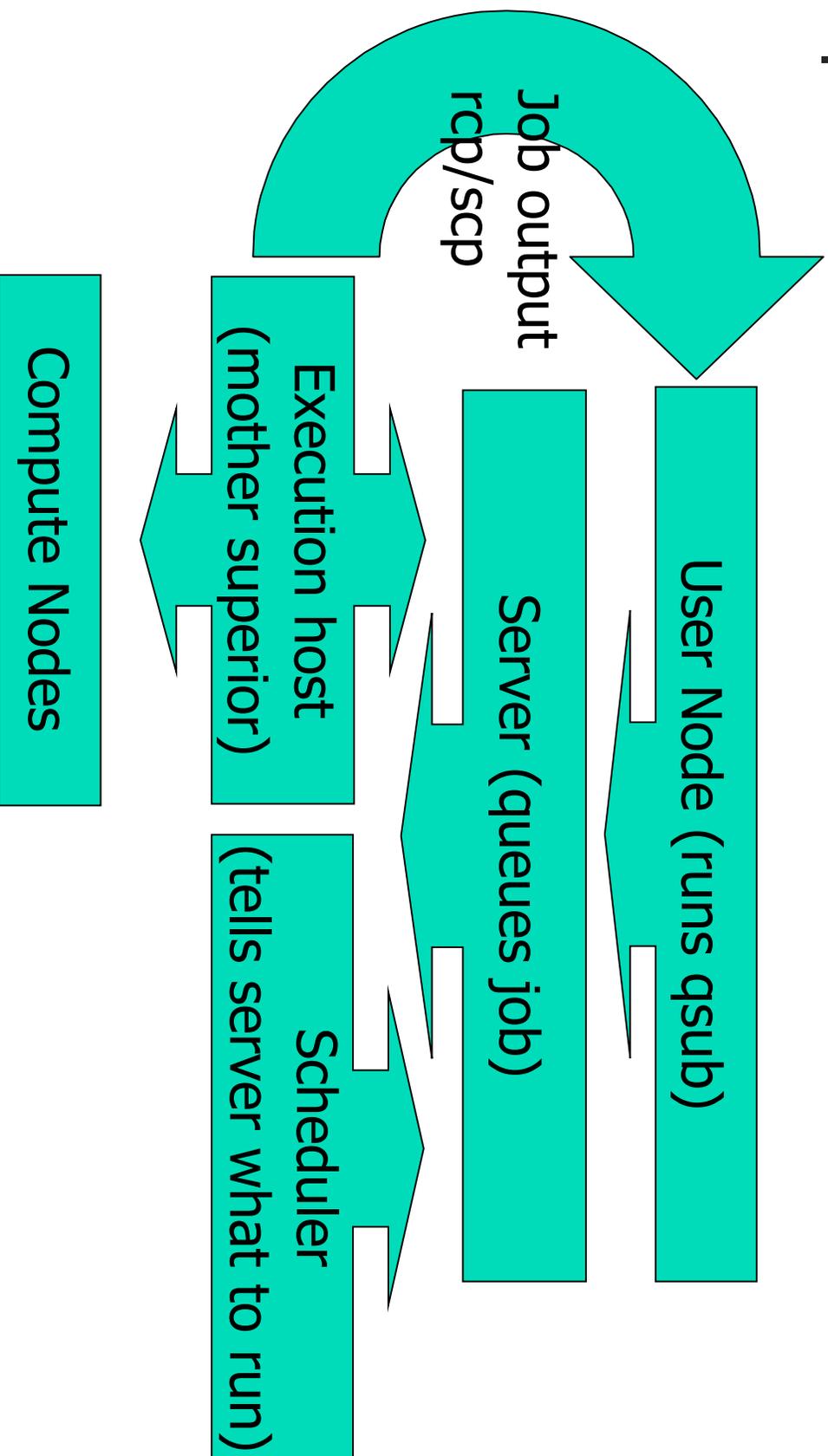
- PBS = Portable Batch System
- Components
 - Server – single instance
 - Scheduler – single instance
 - Mom – runs on compute nodes
 - Client commands – run anywhere
 - qsub
 - qstat
 - qdel
 - xpbsmon



OpenPBS - How it Works

- User submits job with “qsub”
- Execution host (mom) must launch all other processes
 - mpirun
 - ssh/rsh/dsh
 - pbsdsh
- Output
 - spooled on execution host (or in user’s home dir)
 - moved back to user node (rcp/scp)

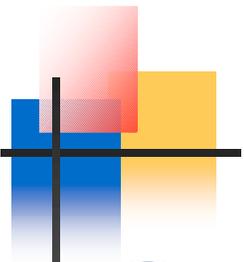
OpenPBS – Job Flow



OpenPBS – Monitor (xpbsmon)

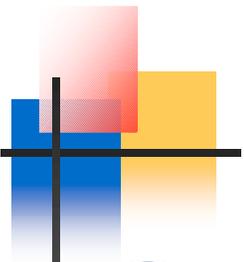
The screenshot shows the OpenPBS Monitor interface for the local site 'ntscl169.ncsa.uiuc.edu'. The window title is 'xpbsmon2.2'. The interface includes a menu bar with 'Site...', 'Pref...', 'Autoupdate...', 'Help', 'About...', and 'Close'. Below the menu bar is a grid of node status information. The grid has 10 columns and 10 rows. The first column contains node IDs (e.g., c0009, c0010, c0011, c0012, c0013, c0014, c0015, c0016, c0018, c0020, c0021, c0022). The second column contains node IDs (e.g., ntscl09, ntscl09). The third column contains node IDs (e.g., c0001, c0002, c0003, c0004, c0005, c0006, c0007, c0008). The fourth column contains node IDs (e.g., ntscl08, ntscl08). The fifth column contains node IDs (e.g., ntscl07, ntscl07). The sixth column contains node IDs (e.g., ntscl06, ntscl06). The seventh column contains node IDs (e.g., ntscl05, ntscl05). The eighth column contains node IDs (e.g., ntscl04, ntscl04). The ninth column contains node IDs (e.g., ntscl03, ntscl03). The tenth column contains node IDs (e.g., ntscl03, ntscl03). The status of each node is indicated by a colored square: green for 'FREE', red for 'DOWN', blue for 'OFFL', yellow for 'RSVD', black for 'NOINFO', and brown for 'INUSE/SHARED'. The 'INFO:' field at the bottom left shows 'done'.

| Node ID |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| c0009 | ntscl09 | c0001 | ntscl08 | c0002 | ntscl07 | c0003 | ntscl06 | c0004 | ntscl05 |
| c0010 | ntscl09 | c0002 | ntscl08 | c0003 | ntscl07 | c0004 | ntscl06 | c0005 | ntscl04 |
| c0011 | ntscl09 | c0003 | ntscl08 | c0004 | ntscl07 | c0005 | ntscl06 | c0006 | ntscl04 |
| c0012 | ntscl09 | c0004 | ntscl08 | c0005 | ntscl07 | c0006 | ntscl06 | c0007 | ntscl04 |
| c0013 | ntscl09 | c0005 | ntscl08 | c0006 | ntscl07 | c0007 | ntscl06 | c0008 | ntscl04 |
| c0014 | ntscl09 | c0006 | ntscl08 | c0007 | ntscl07 | c0008 | ntscl06 | | ntscl04 |
| c0015 | ntscl09 | c0007 | ntscl08 | | ntscl07 | | ntscl06 | | ntscl04 |
| c0016 | ntscl09 | c0008 | ntscl08 | | ntscl07 | | ntscl06 | | ntscl04 |
| c0018 | ntscl09 | | ntscl08 | | ntscl07 | | ntscl06 | | ntscl04 |
| c0020 | ntscl09 | | ntscl08 | | ntscl07 | | ntscl06 | | ntscl04 |
| c0021 | ntscl09 | | ntscl08 | | ntscl07 | | ntscl06 | | ntscl04 |
| c0022 | ntscl09 | | ntscl08 | | ntscl07 | | ntscl06 | | ntscl04 |



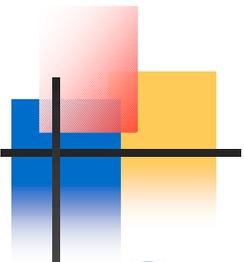
OpenPBS - Schedulers

- Stock Scheduler
 - Pluggable
 - Basic, FIFO
- Maui
 - Plugs into PBS
 - Sophisticated algorithms
 - Reservations
 - Open Source
 - Supported
 - Redistributable



OpenPBS – in OSCAR2

1. List of available machines
2. Select PBS for queuing system
 1. Select one node for server
 2. Select one node for scheduler
 1. Select scheduler
 3. Select nodes for compute nodes
 4. Select configuration scheme
 - staggered mom
 - process launcher (mpirun, dsh, pbsdsh, etc)



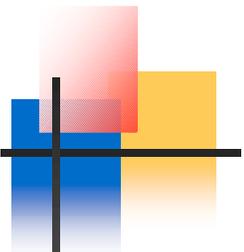
OpenPBS – On the Scale

Pros

- Open Source
- Large user base
- Portable
- Best option available
- Modular scheduler

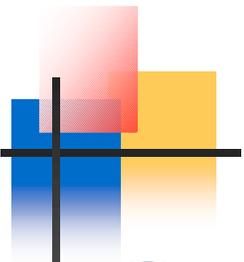
Cons

- License issues
- 1 year+ devel lag
- Scalability limitations
 - number of hosts
 - number of jobs
 - monitor (xpbsmon)
- Steep learning curve
- Node failure intolerance
- Not developed on Linux



OpenPBS – Enhancement Options

- qsub wrapper scripts/java apps
 - easier for users
 - allows for more control of bad user input
- 3rd party tools, wrappers, monitors
- Scalability source patches
- “Staggered moms” model
 - large cluster scaling
- Maui Silver model
 - “Cluster of clusters” diminishes scaling requirements
 - never attempted yet



Future Considerations for OSCAR

- Replace OpenPBS (with what? when?)
 - large clusters are still using PBS
- Negotiate better licensing with Veridian
- Continue incorporating enhancements
 - test Maui Silver, staggered mom, etc.
 - 3rd party extras, monitoring package