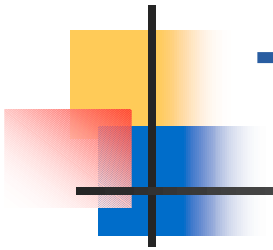
OSCAR

Workload Management

Jeremy Enos

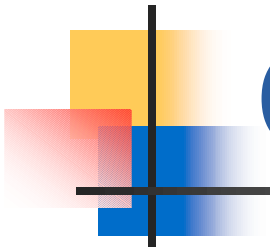
OSCAR Annual Meeting

January 10-11, 2002



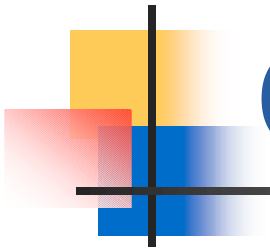
Topics

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



OpenPBS

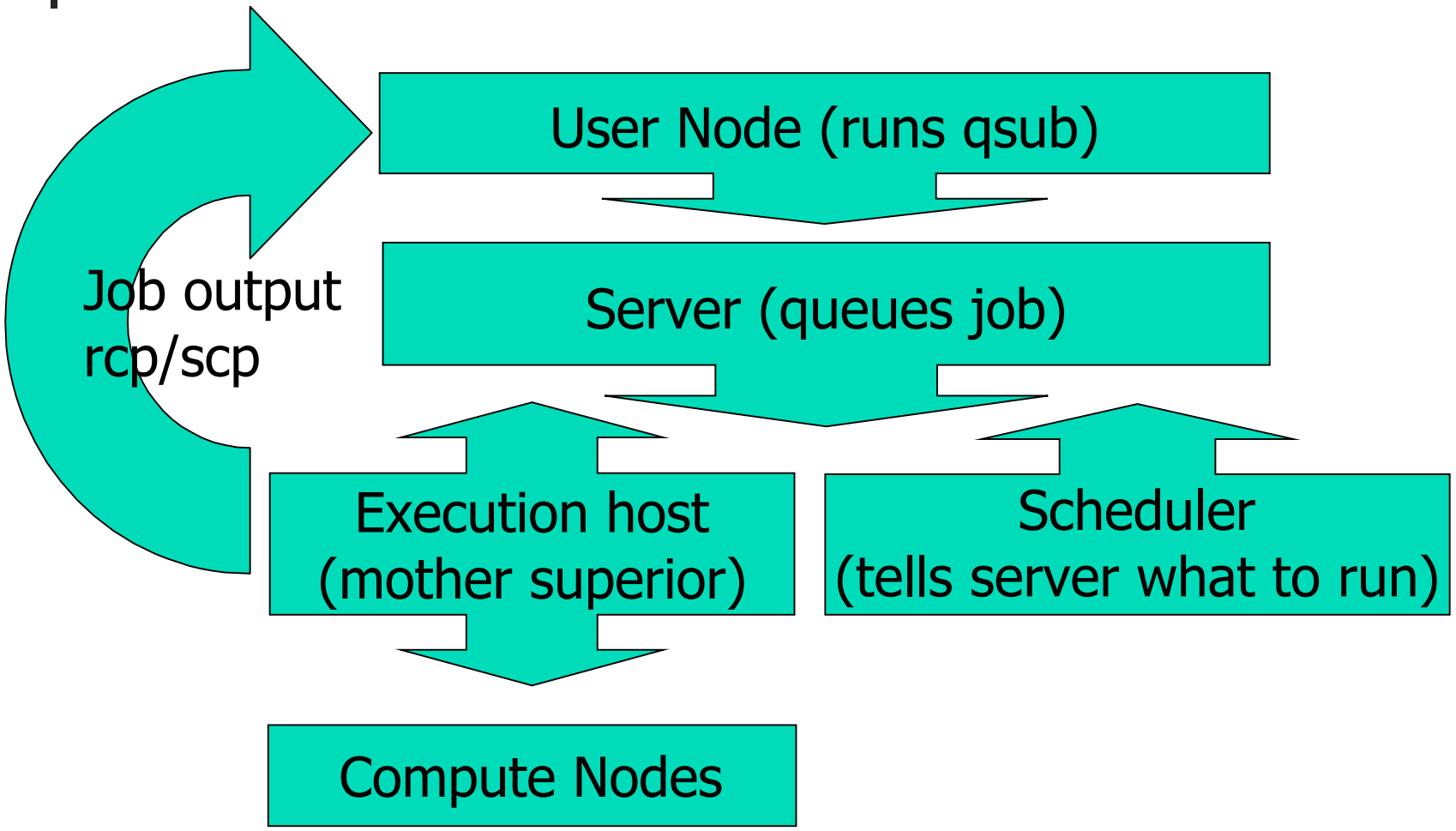
- PBS = Portable Batch System
- Components
  - Server – single instance
  - Scheduler – single instance
  - Mom – runs on compute nodes
  - Client commands – run anywhere
    - qsub
    - qstat
    - qdel
    - xpbsmon



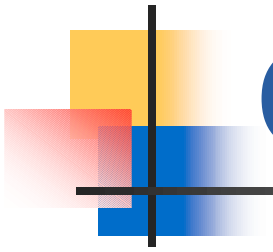
OpenPBS - How it Works

- 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)

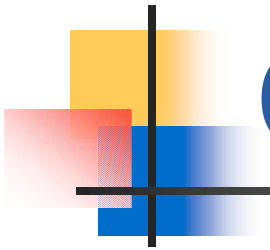


OpenPBS - Schedulers

- 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