An Examination of the Issues in Implementing the PBS Scheduler on an SGI Onyx2/Origin 2000

Sandra Bittner

CUG Conference

May 26, 1999

Argonne National Laboratory

ANL Requirements

- Supports user logins
- Backfill resources with batch jobs
- Batch mechanism
- Advance reservations/Dedicated modes
- Integrate w/graphics pipes
- Generate accounting information
- Balance system activity, user logins, and batch activity

Options Examined

- NQE
- LSF
- PBS
- Miser
- Maui
- Fair Share/Fair Share II

Limited view of the world

- Many schedulers expect to be the only resource manager and the only controlling process on the system
- Dynamic resource allocations not supported
- Sharing excluded using all resources an option
- High learning curve for each package
- Graphics pipes simply not a covered resource

Graphics Pipes

- Development
- Hang & Reboot System
  - Need an individual hardware reset
  - offline, reset, online
- Serious disruptions to batch and other users
- Not a covered item at this time

NQE

- Advantage:
  - Already owned
- Disadvantages:
  - Product End-Of-Life
    - No further development
  - Corrupt binaries
- Conclusion: No

LSF

- Advantage:
  - Works w/IBM SP
  - In house/close colleagues knowledge of operation
- Disadvantages:
  - Needs help on the Origin platform
    - still have to wait for the ABI to be developed
  - High Cost
    - Could buy entire new system for the same $
- Conclusion: No

Miser

- Advantages
  - Already owned
- Disadvantages
  - Limited documentation
  - Particularly buggy, unreliable
- Conclusion: Not a standalone solution

Maui

- Advantages
  - In house knowledge
- Disadvantage
  - Not available at this time for IRIX Systems
  - Programming underway to make it available for IRIX
- Conclusion: Not a possibility

Fair Share/Fair Share II

- Advantages
  - More useful for jobs that do not require large resources
- Disadvantages
  - Large Jobs tend to starve, assume divisible work loads
  - Requires manual control to allow large job processing
  - Costs additional $
- Conclusions: No

PBS

- Advantages
  - Highly flexible and customizable
  - Documentation available
  - Knowledgeable people available
  - No Cost and comes with the source code
- Disadvantages
  - Another Package to Learn
  - Extensive support would cost money
  - Must manipulate to start meet our requirements
- Conclusion: Yes, best path

Iterative Approach

- Login and use
- Weightlessness
- Steps:
  - 1. Development required for resource balancing
  - 2. Full Resource access – Requires OS support
  - 3. Graphics Pipes Support

Missing Hooks/ABI

- Limited kernel structures exist to obtain information
- Operating system hooks are still needed to obtain useful data
  - ABI/Hooks under development
  - Applaud open ABI for scheduling information
- Thursday 10am “IRIX Resource Management Plans and Status”

PBS Resources Available

- PBS Administrator's Guide
- PBS External Reference Specification
- PBS Internal Design Specification
- PBS Manual Pages
- PBS Source
- FAQ/Mailing List/Support Group
- Users Guide due in summer

Installation of PBS

- TCL challenges
  - Required installation of TCL, TK, & TCLX to be installed in the same directory
- Installation modifications
  - Modified install scripts to make PBS self contained
  - /etc/services
    - Regular Port
    - Test Ports

Selecting C scheduler over TCL

- **TCL required program/development of own scheduler**
- **Routines to accomplish this do exist**
- **Preferred not to write a scheduler**

Establishing Base Bounds

- resources default, max, min
- *resources_available
- Required for bounding PBS otherwise the bound is infinity or the entire system

Establishing Artificial Bounds

- **Active Bounding**
  - **Feed PBS preset or computed values**
  - **Use presets or computed values to determine job scheduling**
  - **Subtract active or in use values from set resources**
  - **Float window**

Flexibility

- Priorities
- Advance reservations
- Dedicated mode reserves entire system
  - will require search & destroy scripts to drain resources
- Future: Smaller slices of system for dedicated modes

Batch coexisting with interactive

- 2 levels of batch offered
  - historical batch
  - interactive batch
- User Logins

Accounting

- Extended Accounting
- PBS tracks resources consumed on a per job basis
- Comparisons are possible

Future

- Resource Management Improvements on the Horizon
- Influencing System Resource
- Code to be added to future PBS releases
- Linux Clusters? PBS already there too

Conclusions

- Selected PBS
  - Offered the most cost effective, flexible system that can be readily adjusted to meet changing and challenging needs
  - The developers are open to improvements
  - Continuous Improvement
- Most important the product performs as advertised

Thank you

- Robert L. Henderson, MRJ Technology Solutions
- Bhroam A. Mann, NASA Ames Research Center
- Jens Petersohn, NASA Ames Research Center
- Joe Boyd, SGI
- Daryl Coulthart, SGI
- Dan Higgins, SGI
- Argonne National Laboratory
  - Dinesh Kumar Kaushik
  - MCS Staff
  - MCS Systems Group
- Cray User's Group, CUG

Website

- http://pbs.mrj.com
- http://science.nas.nasa.gov/Software/PBS