# Workload Management: More Than Just Job Scheduling

James Patton Jones, Bill Nitzberg, and Bob Henderson

Veridian PBS Products Department

*jjones@pbspro.com*


TM

*October 2001*

VERIDIAN

# Outline

- Why Scheduling?
- Workload Management
  - Introduction
  - Overview
  - Components
- Implementation
  - Issues and Considerations
  - Real world answers
- About PBS
- Conclusions & Questions

# The Need

- Enterprises own lots of computing power
  - Environment is distributed and heterogeneous
  - 100's of users, 1000's of jobs
- Demands exceed supply
  - But some resources are under utilized
- Allocation and scheduling is ad-hoc
  - Whoever gets there first or "owns" the system
- Enterprise-wide priorities are impossible to dictate
  - Little data can be gathered on actual usage

# Scheduling

- What is needed is a tool to help collect and execute the jobs (tasks)
- To address this problem, cluster administrators look for a “job scheduler”
- Simplicity is important, so many administrators pick UNIX “cron” or “at” or some other simple scheduling system
- Most quickly realize this is insufficient, and switch to a sophisticated job scheduler
- However, what is needed is MORE than just a job scheduler...

# Workload Management 101

- Workload Management systems provide:
  - Queueing
  - Scheduling
  - Monitoring
  - Process Management
  - Accounting
- Most successful clusters need all of these capabilities
- Overview of each follows.

# Workload Management, cont.

- **Queueing**: the process of collecting together "work" to be executed on a set of resources
- User specifies the tasks to be performed in a a "batch job" (usually a shell script)
- Batch Job contains all info needed to run job:
  - Resource directives / requirements
  - Task(s) to be executed
- Once submitted to workload management system, jobs are held in a "queue" until the "right time" to run the job (see below).

# Workload Management, cont.

- Real life queues: lines at banks or grocery stores
  - sometimes you get lucky, no wait
  - usually you have to stand around for a few minutes
  - some days (like payday) the lines can be quite long

- Same applies to computers
  - the length of time that jobs wait depends on the current demand on the available resources (among other factors)

# Workload Management, cont.

- Queues are also part of the "policy enforcement" aspect of workload management
- System managers are able to:
  - limit access to specific queues
  - control destinations of work submitted to queues
- Real cluster example:
  - creating a queue available only under certain conditions, such as a queue for short jobs only, provides a shorter wait for quick tasks.

# Workload Management, cont.

- **Scheduling**: the process of choosing the best job to run

- But "best" can be a difficult goal, as it is dependent upon:
  - usage policy set by local management
  - the available workload
  - the type and availability of cluster resources
  - the type of applications being run on the cluster

VERIDIAN

PBS Products

# Workload Management, cont.

- Scheduling has two primary activities:
  - policy enforcement
  - resource optimization
- Policy tells how the cluster resources are to be used (e.g. priorities, traffic-control, capability vs. high-throughput)
- Resource optimization addresses specific goals of the cluster (e.g. pack jobs efficiently, or exploit under-used resources)

# Workload Management, cont.

- The difficult part of scheduling then is balancing policy enforcement with resource optimization in order to pick the best job to run.

- Workload management systems give the cluster administrator various means of choosing and tuning job scheduling .

# Workload Management, cont.

- **Monitoring**: collecting node status and resource consumption information
- Before scheduling can be done, availability of the physical hardware is needed. (Can't run a job on a node that's down.)
- If the node is up, then data from that node is needed (e.g. CPU load, memory usage, etc.).

# Workload Management, cont.

- This collected status and resource information is then used for:
  - scheduling decisions
  - cluster status requests
  - historical analysis
  - policy enforcement

# Workload Management, cont.

- **Process Management**: the starting, stopping, and cleaning up after jobs run on cluster nodes

- Starting jobs includes setting limits as specified by the workload management system, local policy, or user requests

# Workload Management, cont.

- **Accounting**: process of collecting resource usage data for the batch jobs run on the cluster.
- Such data includes:
  - job owner information
  - resources requested by job
  - total amount of resources consumed by job
- These data are usually include some of the information collected by the monitoring sub-system, and new information collected specifically for workload accounting.

# Implementation

- In order to implement a successful cluster workload management configuration, you need to first understand the goals and intended usage of the cluster:
  - Why are you building the cluster?
  - Are there particular constraints on the cluster?
  - Who is permitted to use the cluster?
  - Are there time-of-day restrictions on cluster access?
  - Are all jobs equal, or will some have higher priority?

# Implementation Questions

- We are often asked how to configure cluster workload management systems. Below are a few popular questions, and general answers:

- *How many queues do I need?*
  - Two: one for normal use, and one for dedicated-time
- *What scheduling algorithm do I need?*
  - The combination of “fairshare” and “preemptive” algorithms will provide a very capable scheduling system.

# Implementation Questions

- *Which workload management system do I need?*
  - Depends on what your requirements are. You should define your requirements, and then pick the system that meets your requirements.

- The most popular workload management system for cluster configurations (and supercomputing sites) is the Portable Batch System, PBS.

VERIDIAN

PBS Products

# History of PBS

- Originally developed for NASA by MRJ Technology Solutions, as existing systems did not meet NASA requirements
- Between 1995 and 1998, NASA distributed PBS to about 70 sites in the U.S. under a restricted-access beta-software program.
- In 1998 MRJ took over development, distribution and support.
- Now over 4000 registered PBS sites around the world.
- In late 1999 MRJ was acquired by Veridian, and in April 2000 the PBS team became the PBS Products Dept.

# Two versions of PBS:

- OpenPBS: www.OpenPBS.org
  - *de facto* standard scheduler for Linux Clusters
  - open source, source code distribution
  - based on last NASA release of PBS
  - active user community (> 1800 active sites)
  - supported and maintained by Veridian
- PBS Pro: www.PBSpro.com
  - new enhanced, "hardened", professional version
    - binary distribution, new docs, lots of new features
  - primary engineering focus of PBS team
  - open technology
    - advantages of open source, in a commercial package
    - binary, "shrink-wrapped" package
    - source code availability

VERIDIAN

PBS Products

# PBS Pro 5.x Provides

- Support for both best-effort and advanced reservation scheduling (and a dozen other popular scheduling algorithms)
- Computational Grid support
- Enhanced fault tolerance and reliability
- Support for IBM SP PSSP 3.2.x
- Support for SGI Origin “cpusets”
- Job Suspend/Resume on all systems
- Checkpoint/Restart for Unicos and IRIX
- Optimized inter-daemon communication

# PBS Pro 5.1 -- Summer 2001

- Full SMP cluster support
- Increased fault tolerance
- Increased integration with specific systems
- Simplified, more flexible installation for binary release
- Preemptive Scheduling
- Enhanced advance reservation features
  - ACL control over reservations
- Ability to tie specific nodes to a queue
- Simplified user authentication for sites with common user name space
- New node specification syntax (now consistent across all architectures, offers user greater control)
- Support for OpenMP jobs

# Currently Supported Systems

## Workstations/Servers

- Sun SPARC w/ Solaris 2.3-2.8
- DEC ALPHA w/ Digital Unix 4.x, Tru64
- HP 9000 w/ HP-UX 9.x, 10.x, 11.x
- IBM RS/6000 w/ AIX 3.2, 4.1-4.3
- SGI systems w/ IRIX 5.x, 6.1-6.5.x
- Intel & Alpha systems: FreeBSD, NetBSD, Redhat Linux 5.x, 6.x, SGI Linux

## Parallel Supercomputers

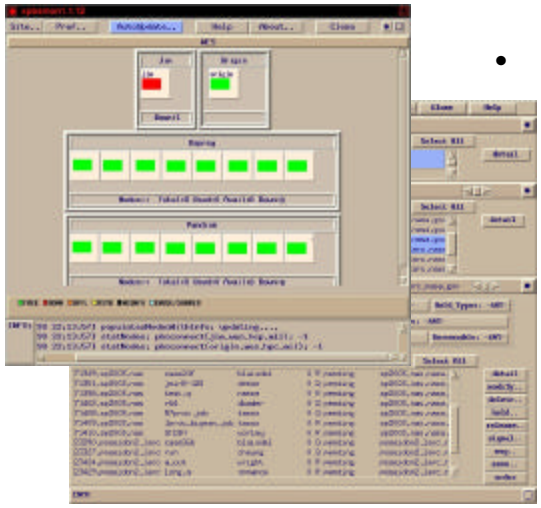
- Cray T3D w/ UNICOSMK
- Cray T3E w/ UNICOS/mk2
- SGI O2000/O3000 w/ IRIX 6.4, 6.5.x
- IBM SP-series w/ AIX 3.2, 4.1-4.2 with (PSSP 2.1) and AIX 4.3 (PSSP 2.3)

## Vector Supercomputers

- Cray SV1 w/ UNICOS 10
- Cray C90 w/ UNICOS 8, 9, 10
- Cray J90 w/ UNICOS 8, 9, 10
- Fujitsu VPP300 w/ UXP/v

# PBS - The Portable Batch System

## Flexible workload management and job scheduler



- Unified interface to all computing resources
  - All major UNIXs supported, heterogeneous environment, SMPs & clusters, parallel jobs (MPI)
  - Single interface handles both interactive and batch processing
  - GUI tools for user and administrator
  - POSIX batch standard
  - Source code included

- Fully configurable scheduler module -- any site policy
  - fair share, load balancing, priorities, back-filling, meta-scheduling
- Sophisticated fault tolerance, accounting, security (ACLs), automatic file staging
- Professional services: commercial support & training

VERIDIAN

PBS Products

www.pbspro.com