PBS Pro Workload Management: Preemption & Reservations

Bill Nitzberg
nitzberg@pbspro.com
May 25, 2001

PBS Products Department
2672 Bayshore Pkwy, Suite 810
Mountain View, CA 94043
Tel: (650) 967-4675
Fax: (650) 967-3080

43rd CUG Conference
Indian Wells, California, USA

Outline

- General motivation
- PBS Pro architecture
- Scheduling
- Preemptive Scheduling
- Advance Reservation Scheduling
- Current Status & Futures

The Problem

- 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

VERIDIAN
PBS Products

The Solution: PBS Pro

The Portable Batch System

- Flexible batch queuing & workload management
- Systems software -- runs "everywhere"
  - Users package their work in "containers"
  - Stake-holders set enterprise-wide scheduling and use policy
  - PBS Pro maximizes efficient use of resources while enforcing policy & provides detailed usage data

VERIDIAN
PBS Products

Workload Management

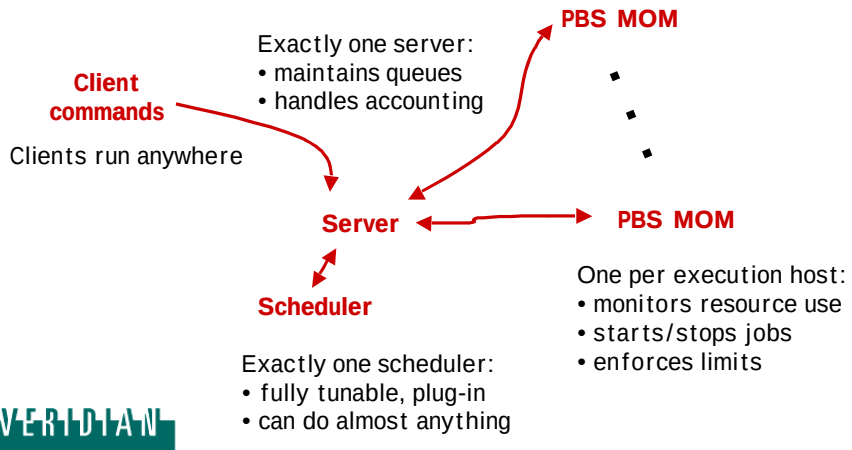
- The Goal

  "Best" "Use" of "Resources"

- Six aspects of workload management
  - Monitoring
  - Queuing
  - Scheduling
  - Starting / stopping jobs
  - Enforcing limits
  - Accounting

VERIDIAN
PBS Products

PBS Pro Architecture



VERIDIAN
PBS Products

Scheduling in General

- Policy enforcement
  - Priorities
  - Traffic control
  - Capability vs. throughput
- Resource optimization
  - Exploit unused resources
  - Pack jobs efficiently

VERIDIAN
PBS Products

High-priority Jobs

- "When it absolutely, positively, has to be computed by …"
- PBS supports two styles of high-priority job
  - Preemptive scheduling
  - Advance reservations

VERIDIAN
PBS Products

Preemptive Scheduling

Run myjob NOW!

- Running jobs are “preempted” to free resources
- Best option when job arrival can’t be predicted

VERIDIAN
PBS Products

Preemption Setup

- Turn on (`sched_config` file)

```
preemptive_sched:   TRUE   ALL
preempt_queue_prio: 150
preempt_suspend:    TRUE
preempt_checkpoint: TRUE
preempt_requeue:    TRUE
```

- Create a queue with high enough priority:

```
Qmgr: set queue BumpQ priority=150
```

VERIDIAN
PBS Products

Using Preemption

- Submit a per-empting, high-priority job

```
qsub -q BumpQ -l ncpus=128 myjob
```

- PBS will (suspend, checkpoint, and requeue) jobs to make 128 cpus available for myjob
- Note: Policy may allow jobs which have been checkpointed or requeued to resume before “myjob” finishes…

VERIDIAN
PBS Products

Advance Reservations

Run myjob at 8:00 am

- In the real-world
  - “I’d like to fly to Chicago for Thanksgiving…”
  - Airline, hotel, rental car, dinner
- PBS can optimize around the reservation, packing in jobs more efficiently
- Enables co-scheduling (especially with people)

VERIDIAN
PBS Products

Using PBS Reservations

Reserve {R} for time T

- Run myjob…when you get around to it…

```
% qsub myjob
```

- Run myjob... at 8am tomorrow

```
% qsub -W reserve_start=05260800 myjob
```

VERIDIAN
PBS Products

Reservation Queues

- Reservation is independent of any one job
  - Many jobs can share a single reservation
  - A simple error (e.g., mispelled file name) can be corrected without going to the end of the queue
- pbs_rsub -R 1400 -E 1600 -l nodes=4
  - Make a reservation (2-4 pm for 4 nodes)
  - returns reservation / queue id, e.g., “R1234”
- pbs_rdel R1234 / pbs_rstat
  - Delete a reservation, query status
- qsub -q R1234 myjob
  - Standard qsub interface to submit jobs to run under a reservation

VERIDIAN
PBS Products

PBS Pro 5.1 -- Early Summer 2001

- Improved IRIX `cpusets` support
- Preemptive Scheduling
- Enhanced advance reservation features
  - ACL control over reservations
- Ability to tie specific nodes to a queue
- Full SMP cluster support
- Increased fault tolerance
- 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
- Hardened to 10,000's of jobs, 1000's of CPUs, 100's of users

VERIDIAN
PBS Products

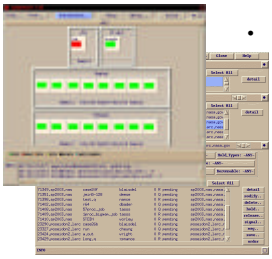
PBS Pro 5.2 and Beyond...

- System-specific capabilities
  - Sun CRE (parallel task spawning via PBS TM API)
  - SGI MPI and job accounting
- Windows 2000 port
- Web-enabled interface
- High-availability option
- Automatic resource detection and configuration
- High-throughput file-staging module
- Accounting, metrics, and reporting tools
- New security model (PKI, Kerberos, DCE)
- Standards & Grids -- **www.GridForum.org**

VERIDIAN
PBS Products

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
  - advance reservations, preemption, 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

About Veridian

Veridian is an information technology solutions company serving government and commercial customers. A private company with annual revenues of $650 million, Veridian operates at more than 50 locations in the US and overseas, and employs nearly 5,000 computer scientists and software development engineers, information security specialists, systems analysts, scientists, engineers and other information technology professionals. The company is known for building strong, long-term relationships with a highly sophisticated customer base.

For more information, visit:

www.veridian.com

VERIDIAN

PBS Products