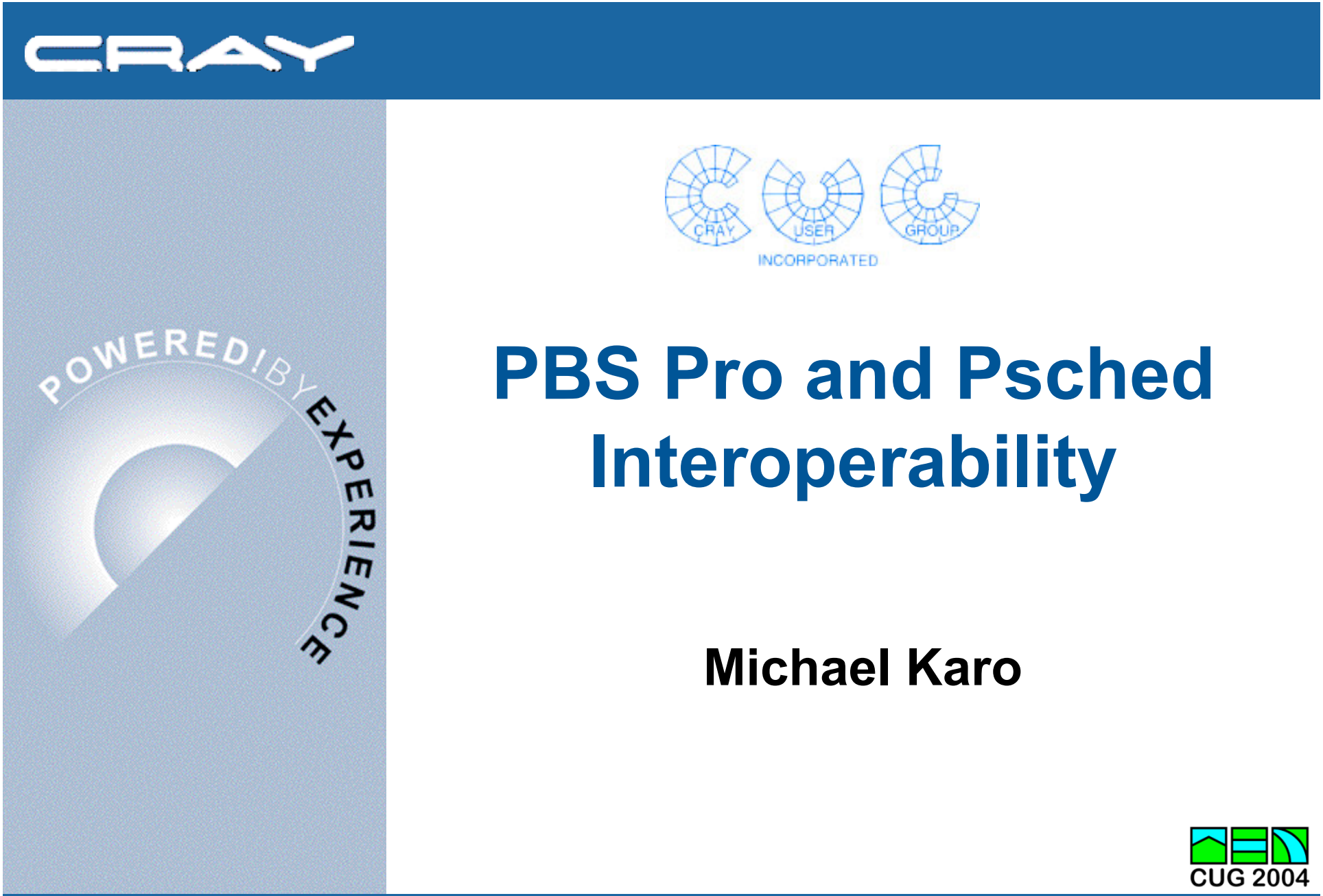


PBS Pro and Psched Interoperability

Michael Karo

CUG 2004

About the Tutorial

- Agenda
  - PBS/Psched Interoperability (~40 minutes)
  - Psched Explained (~50 minutes)
- Prerequisites
  - Familiarity with PBS Pro and Psched components and features for UNICOS/mp (CUG 2003)
- Format
  - Questions welcome during presentations
  - Open discussion afterwards

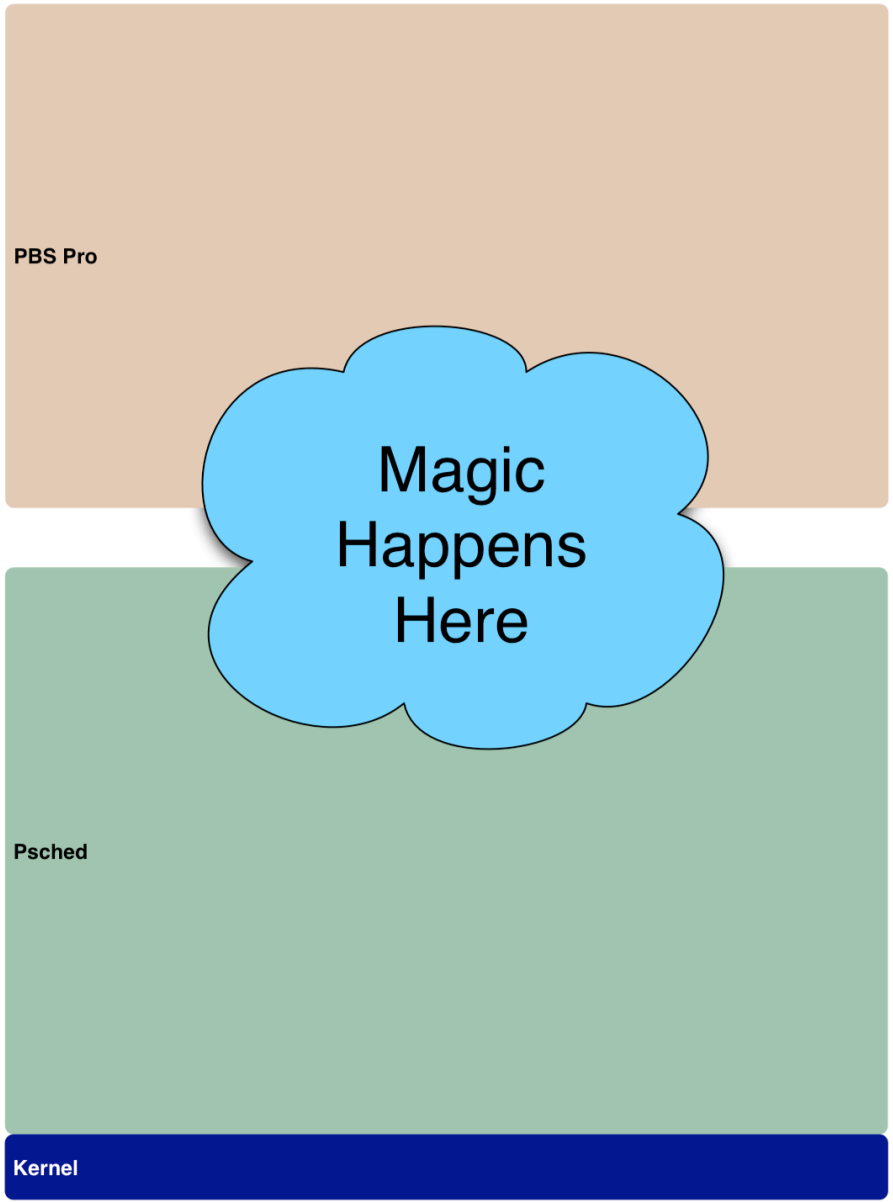
Problem Statement

- **Provide a means for Psched and PBS Pro to cooperatively schedule and manage applications on UNICOS/mp systems.**

Requirements/Constraints

- Work within the bounds of current product architectures
- Maintain platform independence in PBS Pro
- Support MSP/SSP resource types in PBS Pro
- Support both batch and interactive workloads
- Support multiple applications per job
- Provide checkpoint/restart support
- Manage per-process and aggregate job limits for distinct limit domains
- Maintain resource usage and accounting data

Initial Design



Pros:

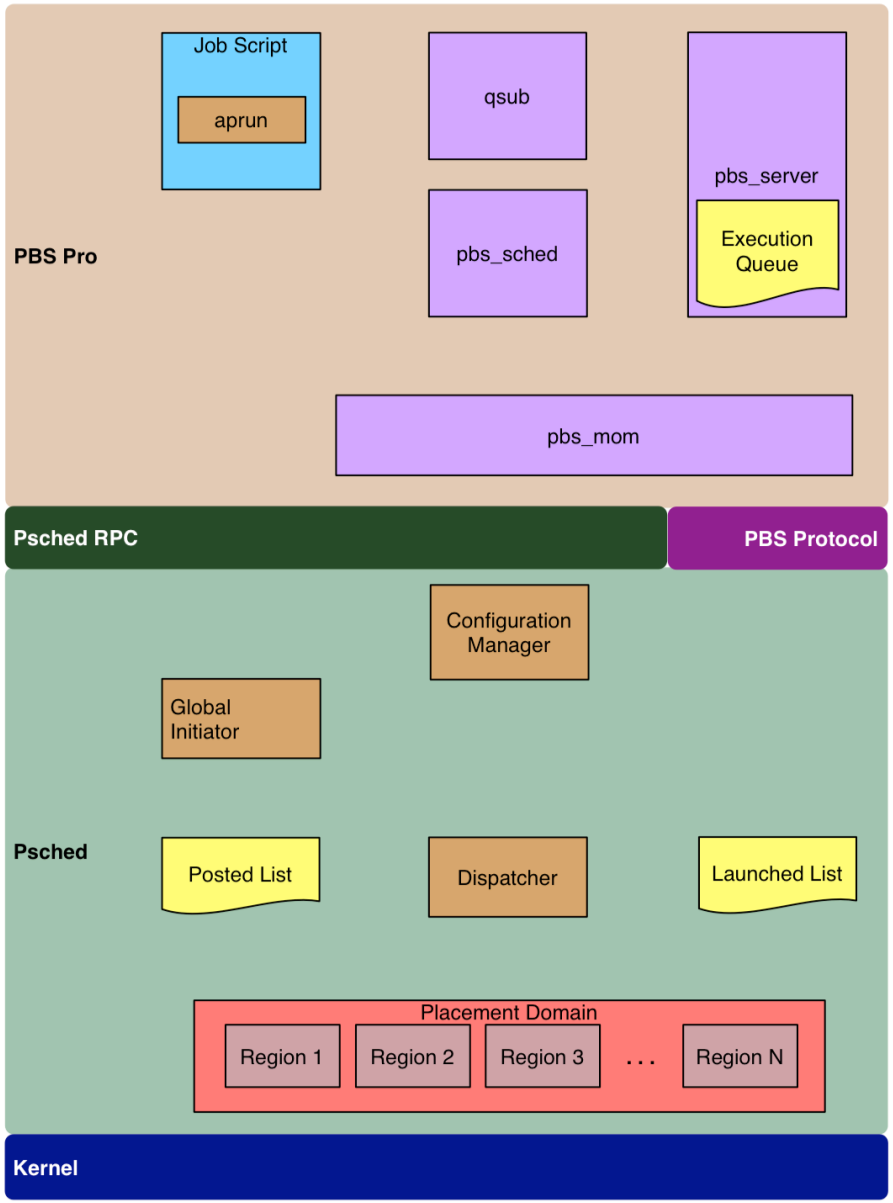
- **Conceptually simple**
- **Intuitive**
- **Few moving parts**

Cons:

- **Lack of "magic" support in Programming Environment tools**

CUG 2004

Functional Components

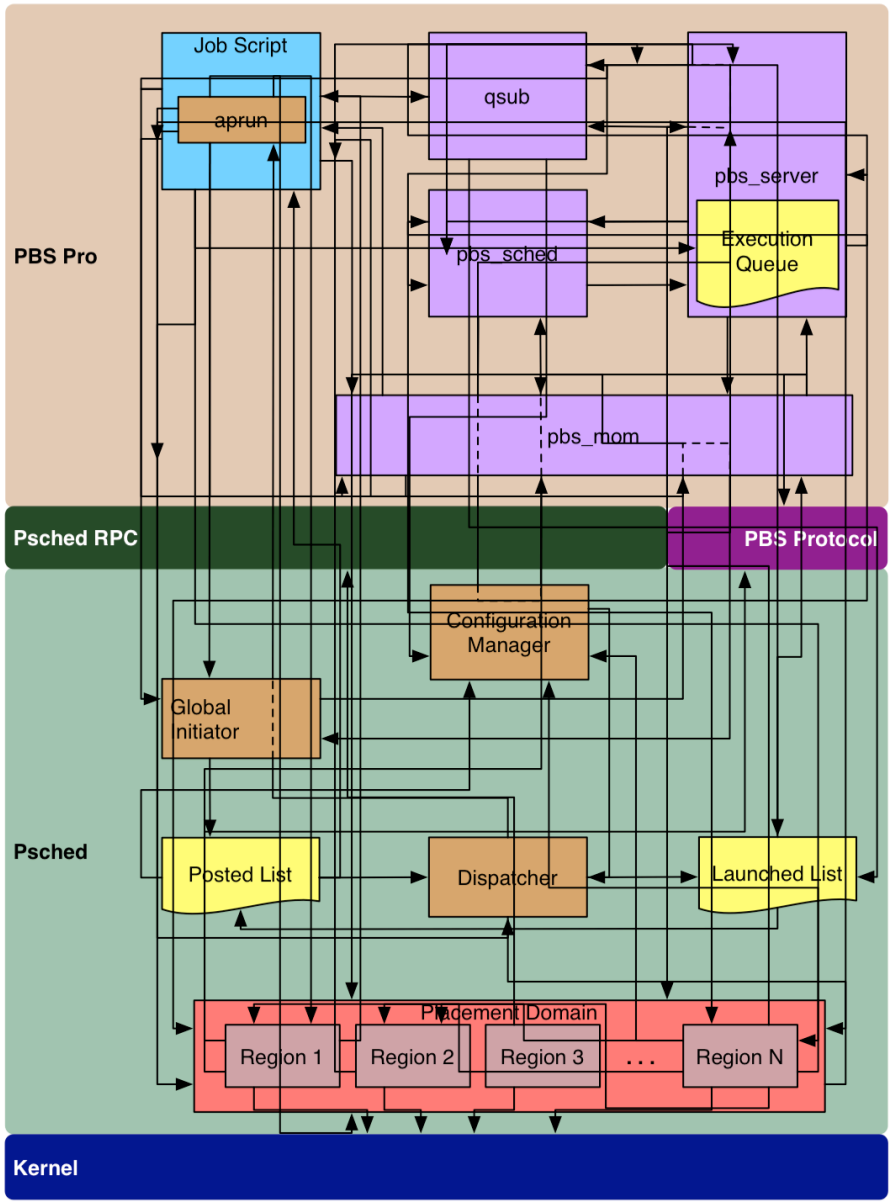


Concerns:

- Several moving parts
- Platform dependencies
- Multiple communication protocols
- Data types and representation
- Preventing version dependencies

CUG 2004

Communication Paths



Pros:

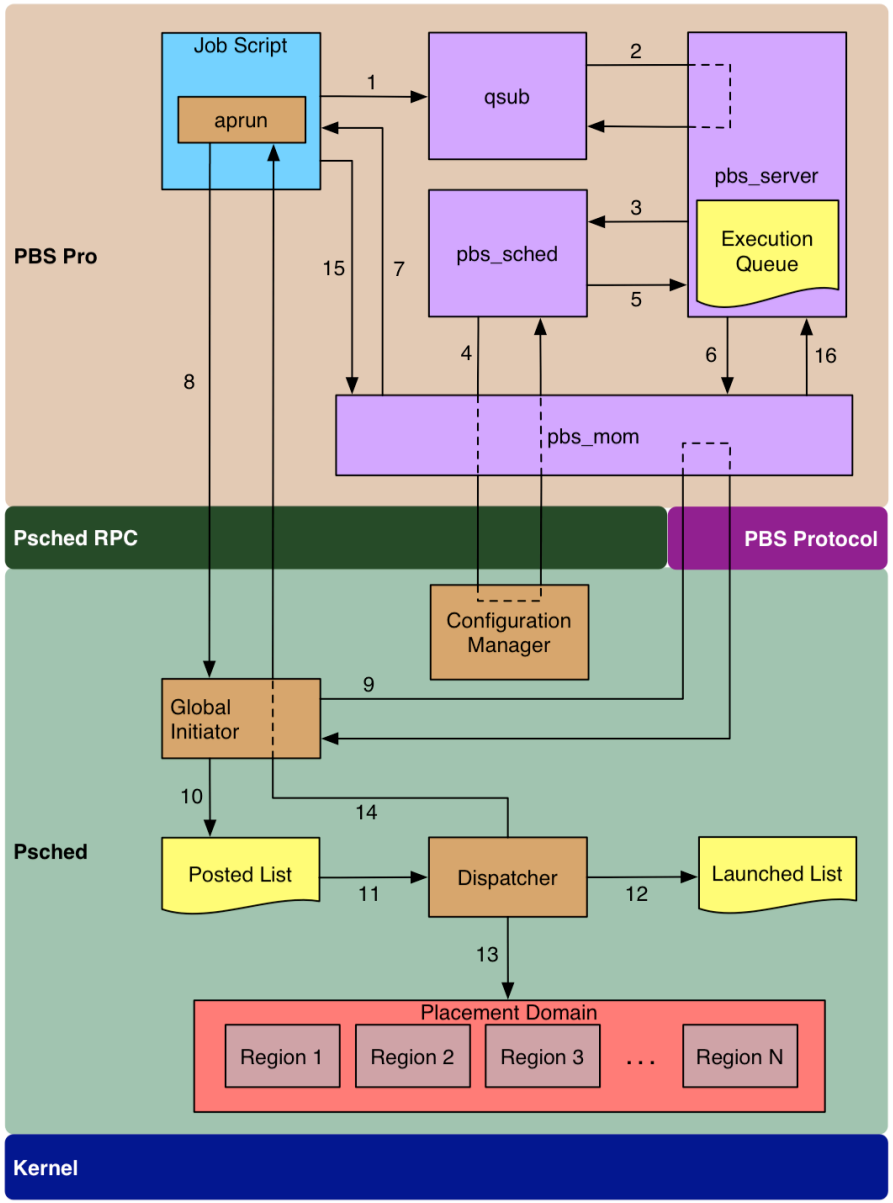
- Very robust
- Somewhat hypnotic

Cons:

- Too complex
- Reminiscent of a hardware circuit diagram

CUG 2004

Component Interactions



Action Index:

1. Job script created and passed to qsub.
2. Client contacts server, job is queued, response returned.
3. Current job and node status passed to scheduler.
4. Psched placement data proxied via MOM. (psched_fit)
5. Run request returned from scheduler to server.
6. Job sent to MOM for execution.
7. Job session initialized, job begins execution.
8. Call to aprun in job script, Psched contacted.
9. Job status request sent to MOM, response received. (UseQueueLimits)
10. Job queued on Psched posted list.
11. Psched dispatcher examines requests on posted list.
12. Resources are allocated and job moves to launched list.
13. Physical placement occurs.
14. Signal returned to aprun, application begins execution.
15. Script completes, SIGCHLD caught by MOM.
16. Obituary returned to server, job purged.

CUG 2004

Psched Related Limits

The following table enumerates the application domain limits supported in PBS Pro version 5.3.5c:

mppe*	Application MSP limit (normalized)
mppssp*	Application SSP limit (additive)
mppfile†	Application file size limit
pmppt†	Application CPU time limit
pmppmem†	Application resident memory size limit
pmppvmem†	Application virtual memory size limit

*** Aggregate job limit enforced by PBS Pro**
† Per-process limit enforced by UNICOS/mp

CUG 2004

MSP/SSP Limit Enforcement

- **mppe = Normalized MSPs**
  - **SSP usage gets "packed" into mppe calculation**
  - **Example: `qsub -l mppe=4 -l mppssp=0`**
  - **$mppe_{used} = MSP_{used} + ((SSP_{used} + 3) / 4)$**
- **mppssp = Additional SSPs**
  - **SSP usage is additive**
  - **Example: `qsub -l mppe=0 -l mppssp=16`**
  - **No math involved, no MSPs allowed!**

MSP/SSP Limit Caveats

- **Leaving mppssp unspecified**
  - **Example: `qsub -l mppe=4`**
  - **Limited to normalized value of MSP/SSP usage**
- **Leaving mppe unspecified**
  - **Example: `qsub -l mppssp=16`**
  - **Limited to 16 SSPs, but *unlimited* MSPs!**
- **Leaving both mppe and mppssp unspecified**
  - **Unlimited SSPs and MSPs**
- **Specifying both mppe and mppssp**
  - **Example: `qsub -l mppe=4 -l mppssp=16`**
  - **Limited to 4 MSPs *plus* 16 SSPs, *or* 32 SSPs!**
  - **Other combinations are possible**

CUG 2004

User Expectations

- **Most users…**
  - **run one application per job**
  - **want to specify either mppe or mppssp resource limits, but not both**
- **Wouldn't it be nice if there was a way for administrators to automatically set the mppe and mppssp resource limits if unspecified?**
- **There is! And here's how…**
  - `Qmgr: set queue foo resources_default.mppe = 0`
  - `Qmgr: set queue bar resources_min.mppssp = 4`

CUG 2004

Psched Configuration

- The /Global/UseQueueLimits Psched parameter
  - Enables/disables communication with PBS Pro
  - Specifies limit domain for interactive jobs

Psched Configuration

- With /Global/UseQueueLimits disabled…
  - Psched determines interactive/batch domain based on presence of tty
  - Psched initializes application with appropriate ULDB limits
- With /Global/UseQueueLimits enabled…
  - Psched sends session ID of aprun to PBS Pro
    - PBS protocol select/status request to local pbs_mom
  - PBS Pro returns resource limits of job
    - PBS protocol encapsulating ASCII string
  - Psched determines interactive/batch domain based on PBS Pro job information or tty presence
  - Psched initializes application with the lesser of ULDB and job limits

CUG 2004

Psched Configuration

- /Global/UseQueueLimits settings
  - A value of 0 or unspecified means:
    - No communication with PBS Pro
    - Presence of tty determines IA/BA limit domain
  - A value of 1 means:
    - Perform PBS Pro query
    - IA limits used for interactive jobs
  - A value of 2 means:
    - Perform PBS Pro query
    - BA limits used for interactive jobs
- Setting the parameter
  - `/usr/sbin/psmgr -c "set /Global/UseQueueLimits 1"`
  - Add "`new int /Global/UseQueueLimits 1`" to /etc/psched.conf

CUG 2004

PBS Pro Configuration

- The $batch_domain_limits MOM parameter
  - Controls how per-process job limits are initialized
  - All jobs use BC domain if true
  - Interactive jobs use IC domain if false
- The psched_fit PBS Pro scheduler parameter
  - Enables/disables communication with Psched
  - Addresses conflicts between “pure interactive” and PBS Pro application requests
  - Restricts Psched posted list entries
  - Favors “pure interactive” applications

PBS Pro Configuration

- With psched_fit disabled, the PBS scheduler…
  - Checks to determine whether user/queue run limits have been reached
  - Checks each specified resource and runs a job if: $(resources_{requested} + resources_{assigned}) <= resources_{available}$
- With psched_fit enabled, the PBS scheduler…
  - Performs its normal checks
  - Collects Psched node data and posted/launched lists (proxied via pbs_mom)
  - Determines whether sufficient mppe and mppssp resources exist based on Psched node data
  - Checks for a backlog of requests on the Psched posted list that may prevent placement

PBS Pro Configuration

- **Setting the psched_fit parameter**
  - **Add "`psched_fit: true`" to $PBS_HOME/sched_priv/sched_config**
- **Setting the $batch_domain_limits parameter**
  - **Add "`$batch_domain_limits true`" to $PBS_HOME/mom_priv/config**

Per-process Limit Domains

Method	Mode	BDL	UQL	Session	App
login	Background	N/A	N/A	N/A	BA
login	Foreground	N/A	N/A	N/A	IA
qsub	Batch	False	0,1,2	BC	BA
qsub	Batch	True	0,1,2	BC	BA
qsub	Interactive	False	0,1	IC	IA
qsub	Interactive	False	2	IC	BA
qsub	Interactive	True	0,1	BC	IA
qsub	Interactive	True	2	BC	BA

BDL = PBS Pro MOM $batch_domain_limits setting
UQL = Psched /Global/UseQueueLimits setting
IC, BC = Interactive/Batch Command Domain
IA, BA = Interactive/Batch Application Domain

CUG 2004

Tips

- Utilizing job dependency in PBS Pro will…
  - help to maximize system utilization by providing more accurate resource specifications
  - reduce the risk of “pure interactive” applications interfering with PBS Pro jobs
- Reducing IA domain ULDB limits and setting /Global/UseQueueLimits to 2 will encourage use of PBS Pro
- Setting resources_default.mppe and resources_default.mppssp will allow users to specify a single job resource limit

CUG 2004

Additional Resources

- **PBS Pro Release Overview, Installation Guide, and Administration Addendum for Cray Systems (Publication S-2345-535c)**
- **PRIME yourself! (Please Read Instructional Materials Extensively)**

Next!

Everything you ever wanted to know about Psched… and more!

Just In Case

- And you thought we were done. :-)
- The following slides were cut due to time constraints, but may be used for discussion

Other Resource Limits

- ncpus
  - Aggregate job load average
  - Based on SSPs
  - Useful in accounting/auditing
- mppt
  - Not supported in PBS Pro for UNICOS/mp
  - Use of walltime resource is equivalent when Psched oversubscription is disabled

Prime Jobs

- Supported in Psched, not in PBS Pro
- PBS Pro supports “express queues”
- Preemption mechanisms
  - Checkpoint/Restart
    - CPR overhead
    - $restart_background true
  - Suspend/Resume
    - Suspended applications still allotted time slice
  - Kill/Restart
    - Application progress lost
- Apply ACL to express queue
- Manual intervention (qrun, psmgr -f)

CUG 2004

Oversubscription

- In Psched…
  - Processor/memory oversubscription factors
  - Gang scheduling provides time slicing at configurable intervals
  - Virtual processor counts reported to PBS Pro
- In PBS Pro…
  - resources_available.mppe controls availability
  - If availability exceeds virtual processor count, jobs collect in Psched posted list
  - Provides for backfill
  - Caution: jobs on the posted list accrue walltime