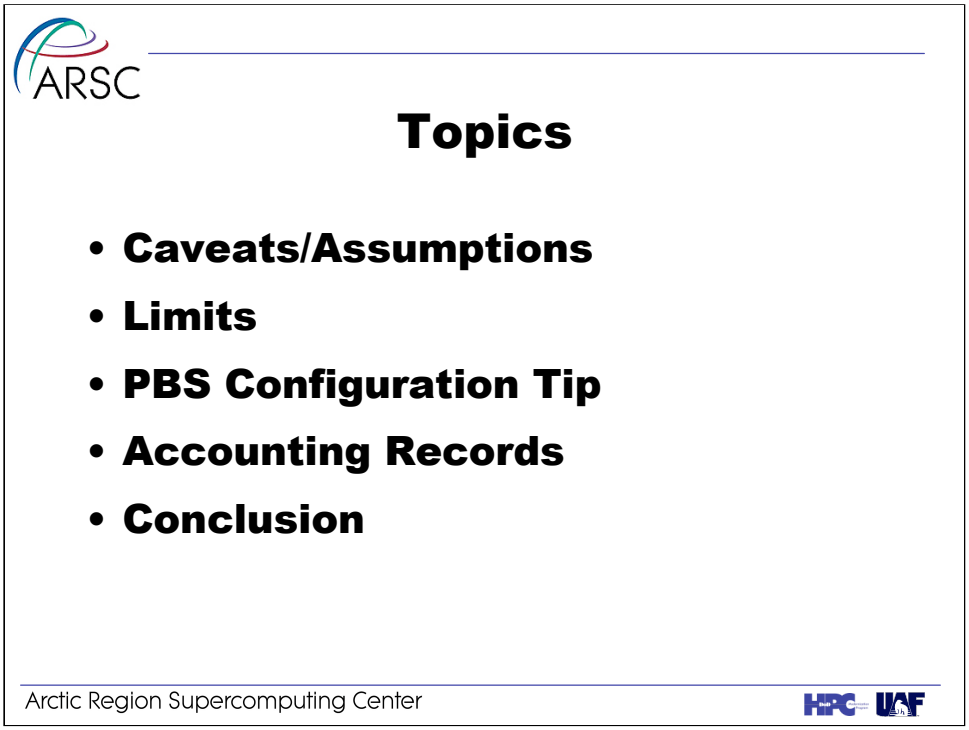
An Initial Foray Into Configuring Resources and Reporting Utilization on the Cray X1

Liam Forbes & Jan Julian

Knoxville, TN

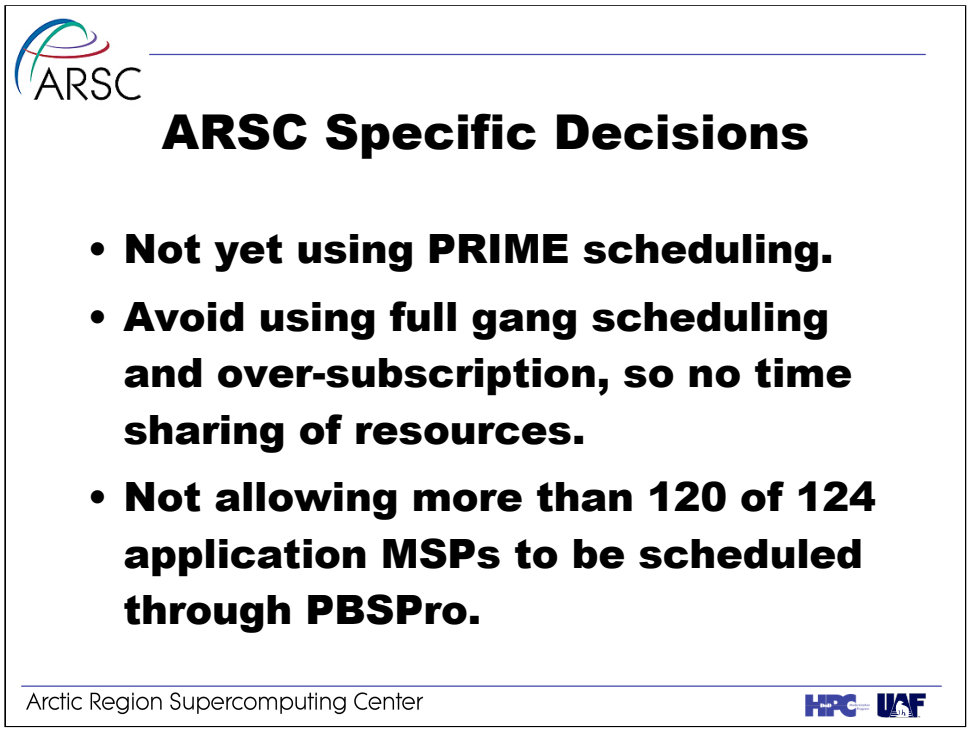
May 17-21


ARSC
Topics
• Caveats/Assumptions
• Limits
• PBS Configuration Tip
• Accounting Records
• Conclusion
Arctic Region Supercomputing Center
HPC UAF

Caveats/Assumptions

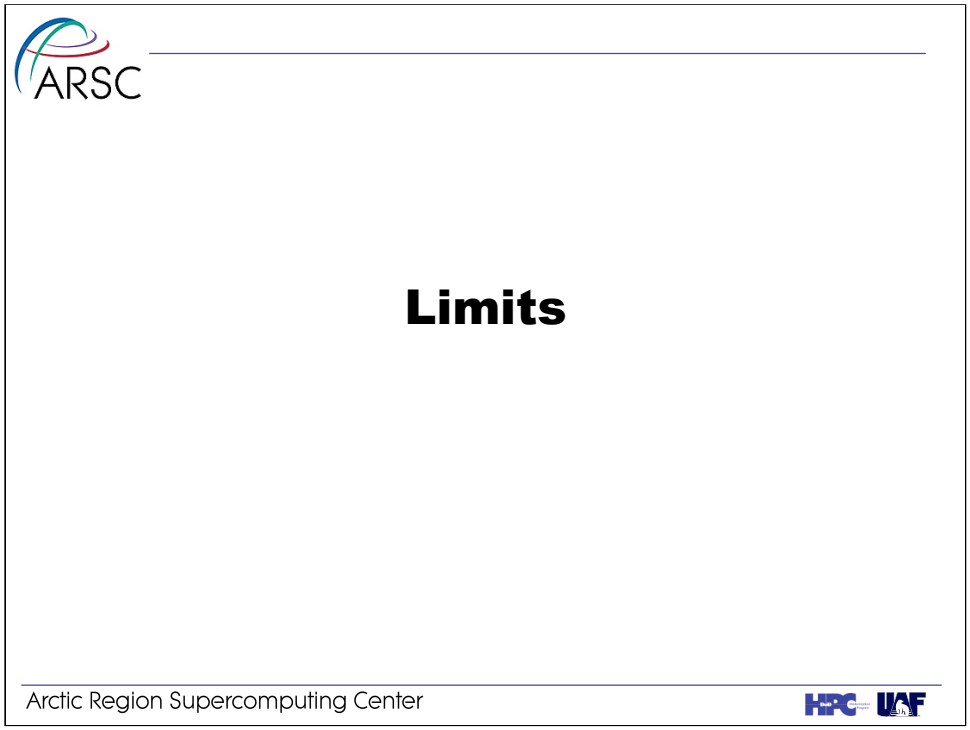
Software Levels

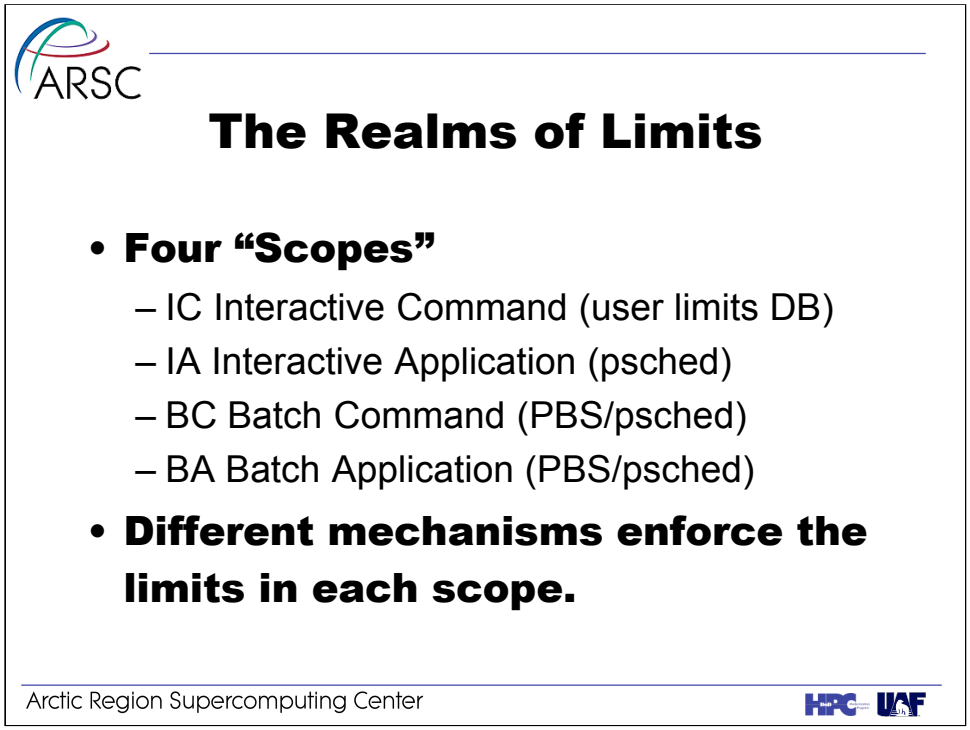
- **Unicos/mp 2.4 (kernel 2.4.17)**
- **PBSpro 5.3.4c**
- **Programming Environment 5.2**



Resources are dedicated to a job until it completes.

By limiting the resources_available.mppe, we guarantee PEs are available for interactive jobs (mostly debugging), even if the system is “full”.


ARSC
Limits
Arctic Region Supercomputing Center
HPC
UAF



The first mechanism employed to manage resources, is limits.

PBS documentation refers to the command scopes as "support".

Interactive == launched from the command line.

Batch == launched from a PBS job.

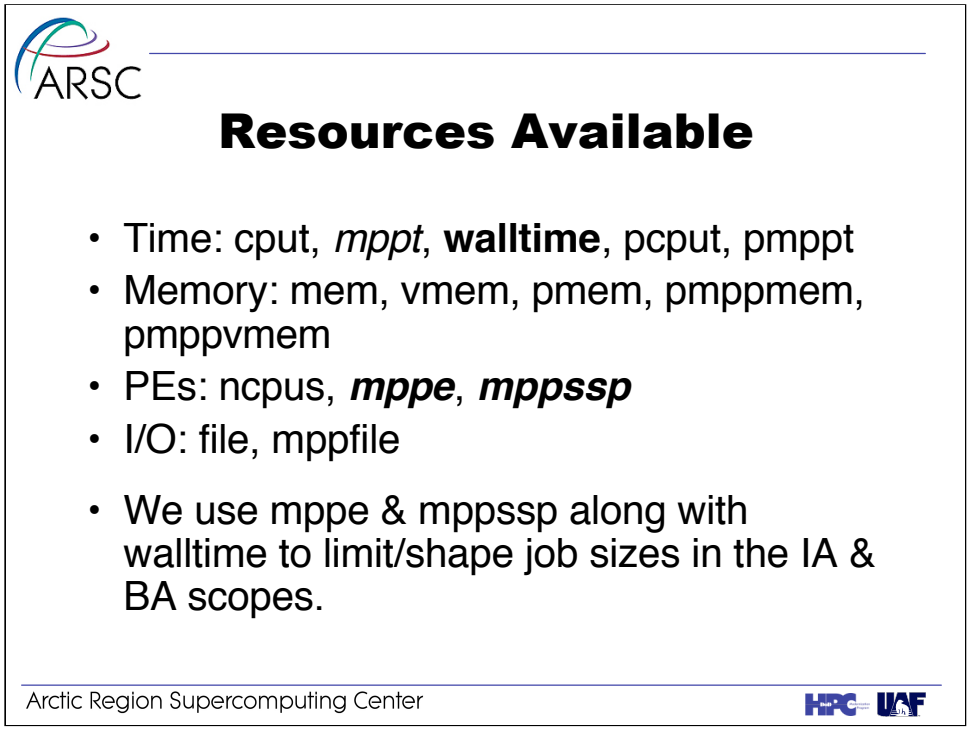
Command == process executed on support node.

Application == process execute on application nodes.

Interactive Command is generally user login sessions. We use the ULDB to manage the support node resources.

Batch Application is the bulk of the work performed on ARSC's system ,so we use the PBS queue configuration to enforce limits on the number of PEs and the length of execution time in walltime.

We do not execute anything in the Batch Command scope; our support node is busy enough already.



Based upon experience with the T3E, our first choice was to try limiting based on mppt and number of PEs. However, mppt is the one resource not yet available on the X1. Instead we use walltime to limit the life of a job, and number of PEs to limit the size of the job. We use the ULDB for IA limits and PBS queue configuration for BA limits.

mppe - Specifies the maximum number of MSP processing elements used by all processes running on application nodes in a job.

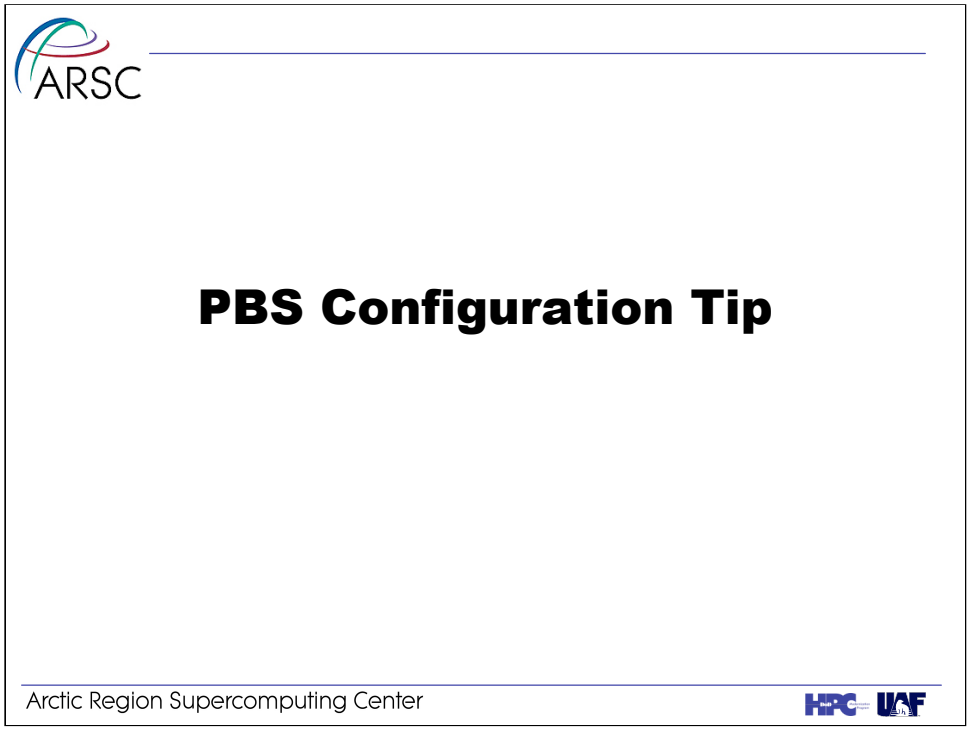
mppssp - Specifies the maximum number of SSP processing elements used by all processes running on application nodes in a job.

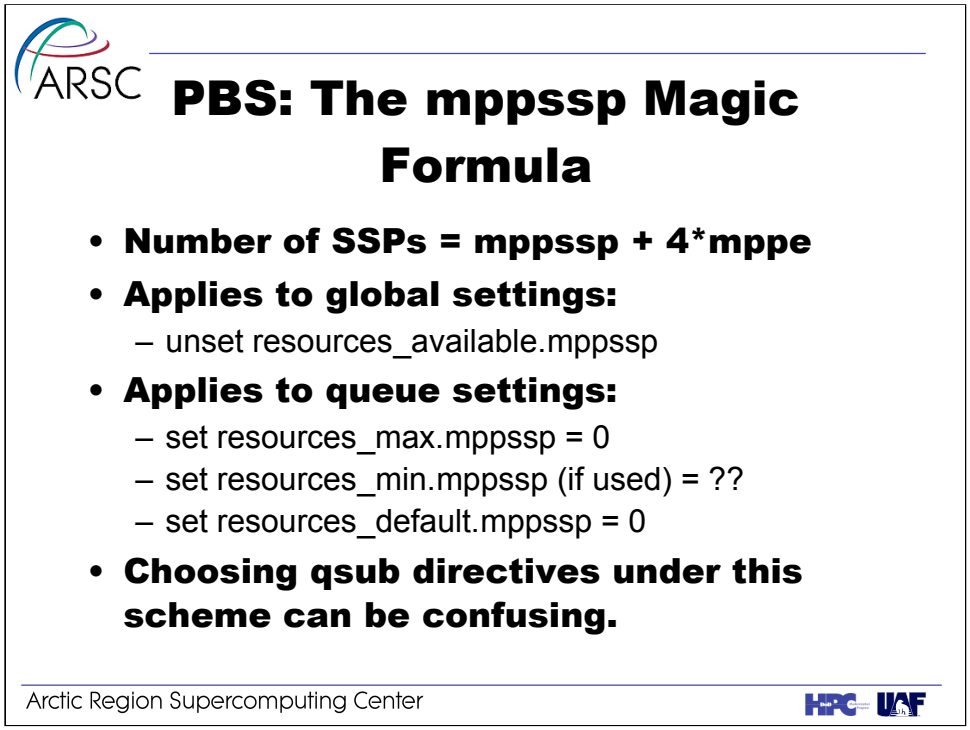
pmppfile - Specifies the maximum size of any single file that may be created by a process running on application nodes in a job.

pmppt - Specifies the maximum amount of CPU time used by a single process running on application nodes in a job.

pmppmem - Specifies the maximum resident memory segment size used by a single process running on application nodes in a job.

pmppvmem - Specifies the maximum amount of virtual memory used by a single process running on application nodes in a job.


ARSC
PBS Configuration Tip
Arctic Region Supercomputing Center
HPC
UAF

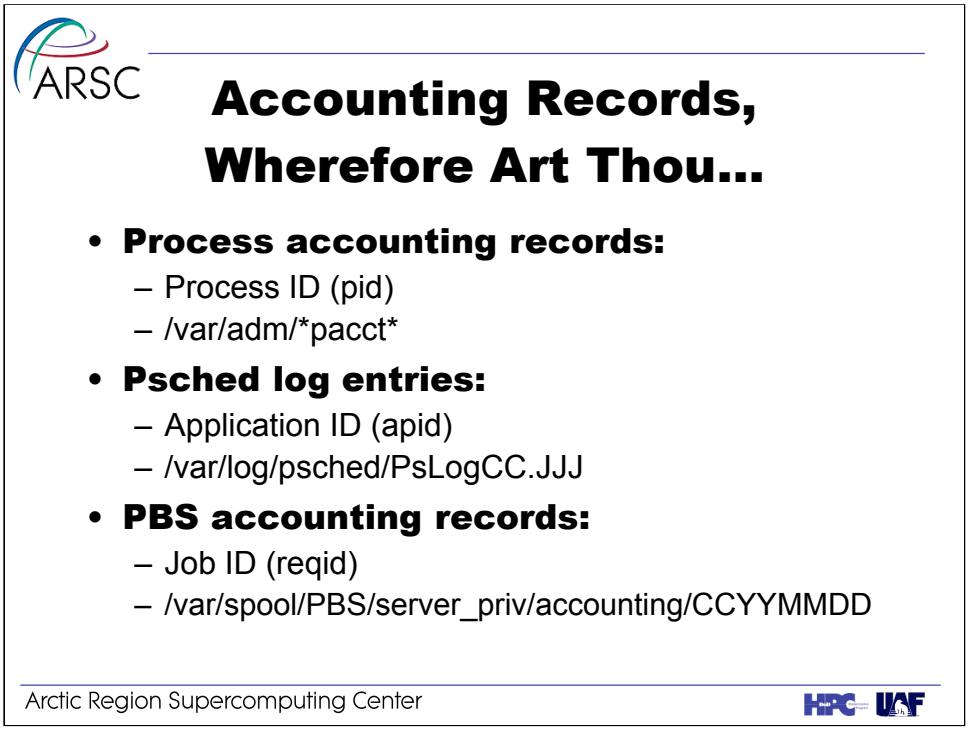


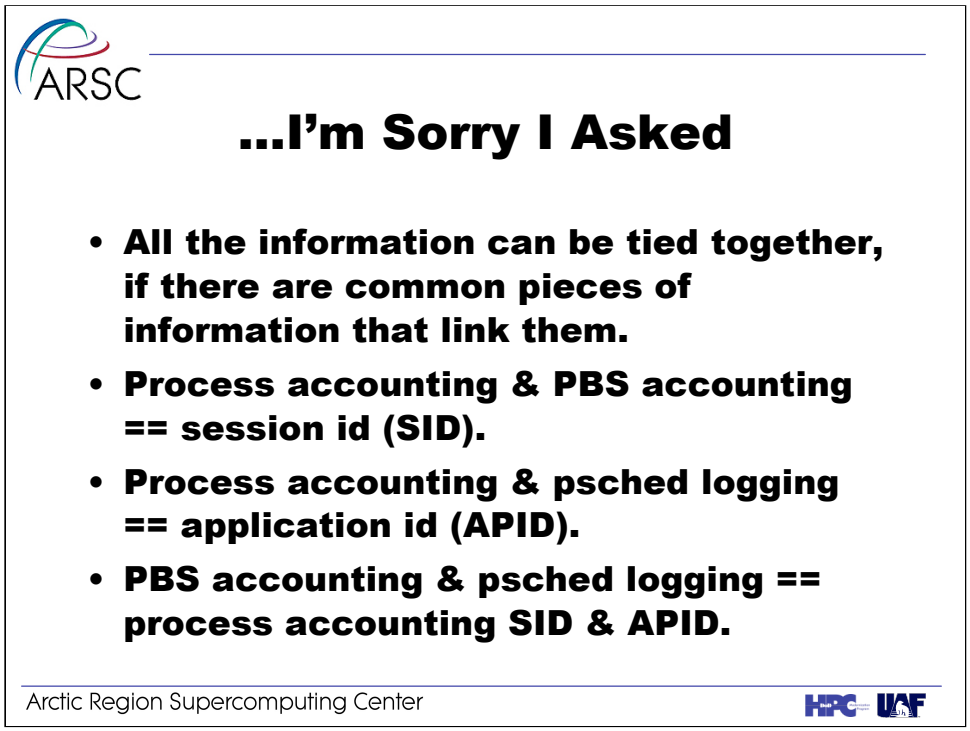
When limiting on the number of PEs, we use PBS server and queue configuration. However, it's important to use the right PE limit. We use the same queues for both MSP and SSP jobs. Thus we have to be wary of the interaction between mppe and mppssp.

Also, it's important to set resources_default.mppe to 0 or jobs that include no directives will be assigned values that either overcharge a job if fewer PEs are used, or not enough resources will be assigned causing the job to be killed when it executes on too many PEs. By setting both mppe & mppssp to 0, a uses is required to specify resources.

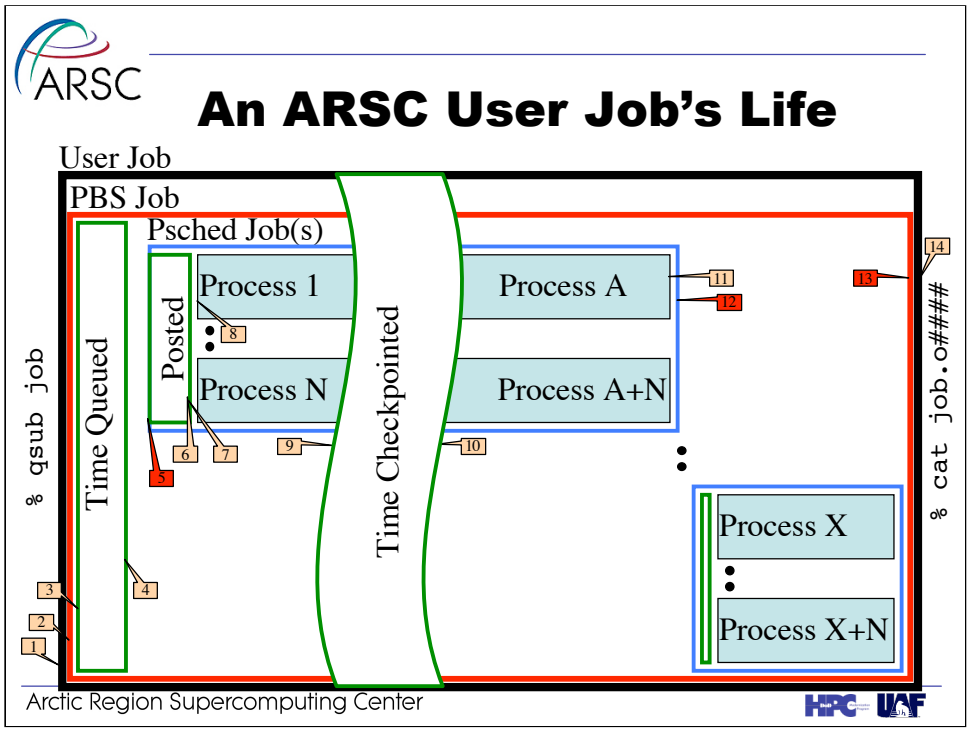
Mike Karo discussed this "problem" during the PBSpro/psched tutorial. The above is not the same as his "magic formula", but is functionally equivalent. I claim the above is the way users experience the problem versus his approach which is how PBS/psched experience the problem.

Accounting Records


ARSC
Accounting Records,
Wherefore Art Thou...
• Process accounting records:
– Process ID (pid)
– /var/adm/*pacct*
• Psched log entries:
– Application ID (apid)
– /var/log/psched/PsLogCC.JJJ
• PBS accounting records:
– Job ID (reqid)
– /var/spool/PBS/server_priv/accounting/CCYYMMDD
Arctic Region Supercomputing Center
HPC UAF



The ability to recreate the majority of what was available in CSA exists on the X1, it's just a matter of tying it all together.



1. Qsub executed, work is submitted to PBS.
2. Jobid assigned.
3. Queued record(s) created in PBS accounting file.
4. Job selected for execution, SID assigned, start record created in PBS accounting file.
5. Application posted for execution, apid assigned, post record created in psched log file.
6. Application placed on node(s). Placed record created in psched log file.
7. Application launched. Launched record created in psched log file. Processes forked, pids assigned, placed on PEs, and execution begins.
8. Checkpoint initiated. Delete record created in PBS accounting file. Checkpoint record created in psched log file. Process records written into pacct file.
9. Restart initiated. Restart record created in PBS accounting file and psched log file. New processes forked and execution begins again.
10. Process(es) complete. Process records written into pacct file.
11. Application completes. Delete record written into psched logging file.
12. Job completes. Exit record written into PBS accounting file.
13. User reviews output.

It is possible for a PBS job to execute multiple applications, as long as limits are obeyed. We have no experience with this situation.

Interactive Accounting Records

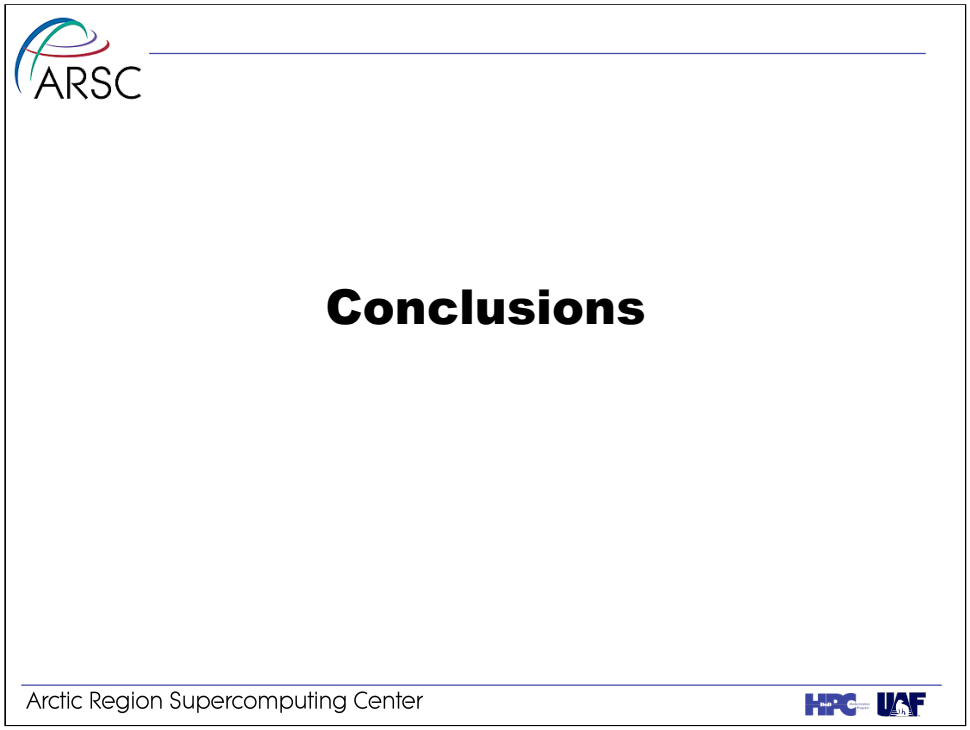
- **The psched log file contains just enough in the "posted" and "deleted" records.**

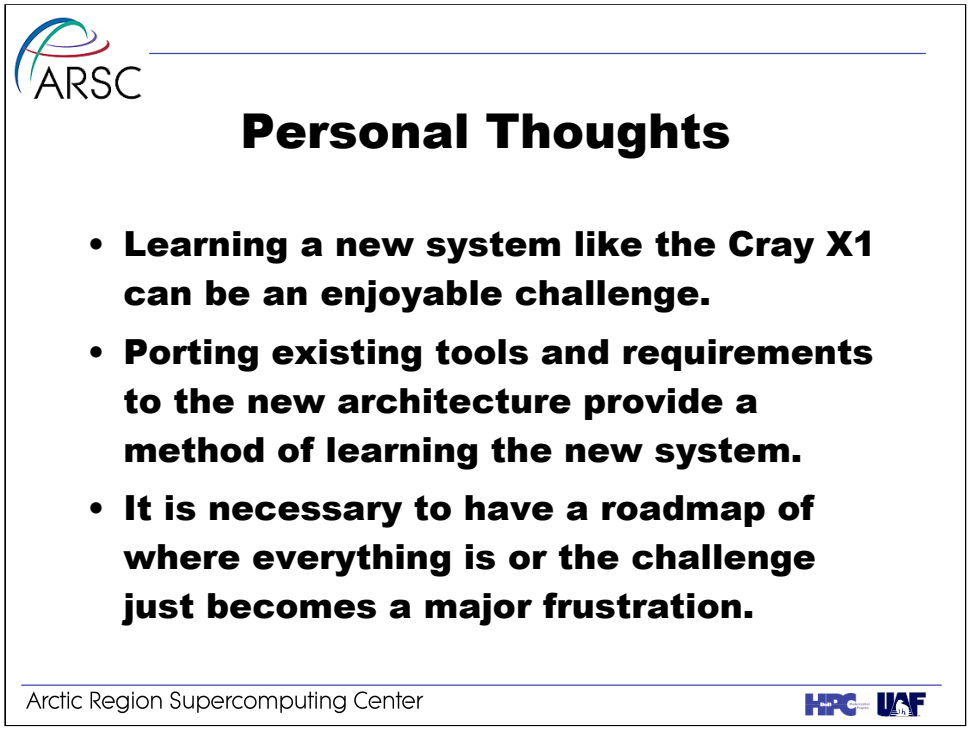
```
18.07:58:43 Posted apid 64702 uid 929
  flags iMNAX w:d:N 4:1:0 time
  UNLIMITED memory UNLIMITED cmd
  ./pxfd
18.07:59:01 Deleted apid 64702 Connect
  time 00:00:00:16 dd:hh:mm:ss
```

Batch Accounting Records

- **The PBS accounting file contains enough information in the "exit" record.**

```
04/29/2004 16:15:04;E;8835.klondike;user=lforbes group=staff
   jobname=pxfd.q queue=Qsmall ctime=1083283836
   qtime=1083283836 etime=1083283950 start=1083283951
   exec_host=klondike/0 Resource_List.mppe=4
   Resource_List.mppssp=0 Resource_List.walltime=00:30:00
   session=298073 end=1083284104 Exit_status=0
   resources_used.cpupercent=77 resources_used.cput=00:13:18
   resources_used.mem=2165440kb resources_used.mppe=4
   resources_used.mppssp=0 resources_used.ncpus=1
   resources_used.vmem=439229376kb
   resources_used.walltime=00:02:50
```


ARSC
Conclusions
Arctic Region Supercomputing Center
HPC
UAF



Much of the information that we had to learn by trial and error, and the earlier PBSpro/psched tutorial really needed /needs to be in Cray documentation. The areas of Limits and the Cray specific modifications to PBSpro need to be better dealt with in the online manuals and the system administration training courses.

Attributions

- **Judith Conrad, Cray Inc.: the mppssp magic formula**
- **John Metzner, Cray Inc.: work on limits & PBS configuration**
- **Derek Bastille & Kurt Carlson, ARSC: direction and inspiration on the accounting**
- **Jan Julian, ARSC: putting up with me**

Thank You For Your Time

- **Questions?**
- **lforbes@arsc.edu**
- **http://www.arsc.edu/**