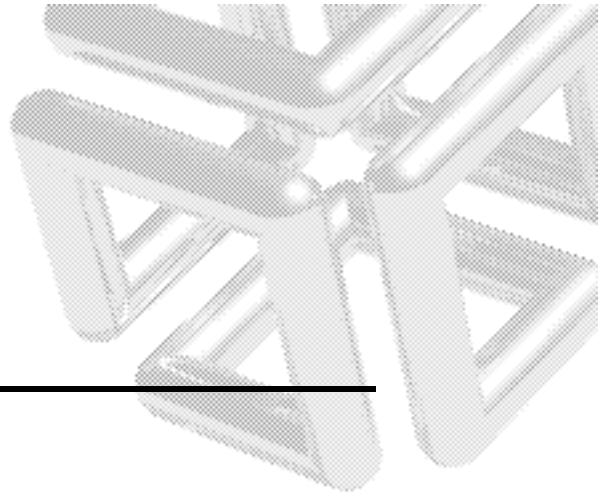
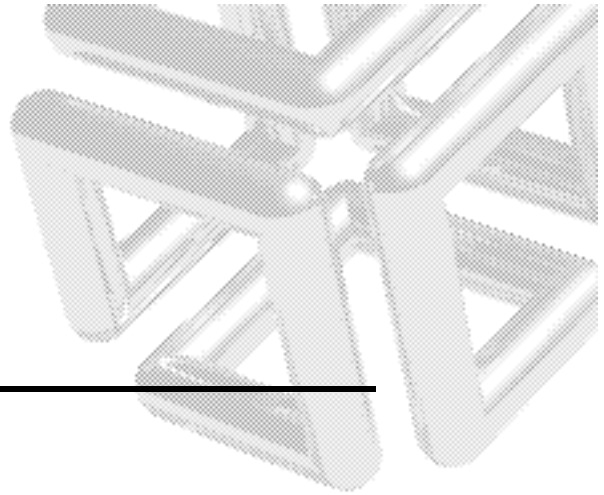


CUG 98 T3E Scheduling Tutorial

Jim Grindle

June 15, 1998

SiliconGraphics Computer Systems



Special Thanks

I want to thank

Jay Blakeborough

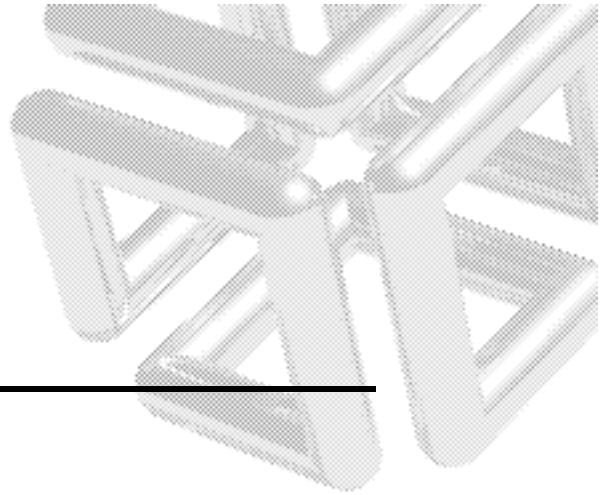
and

Stephan Gipp

for their invaluable assistance in preparing this tutorial

SiliconGraphics Computer Systems

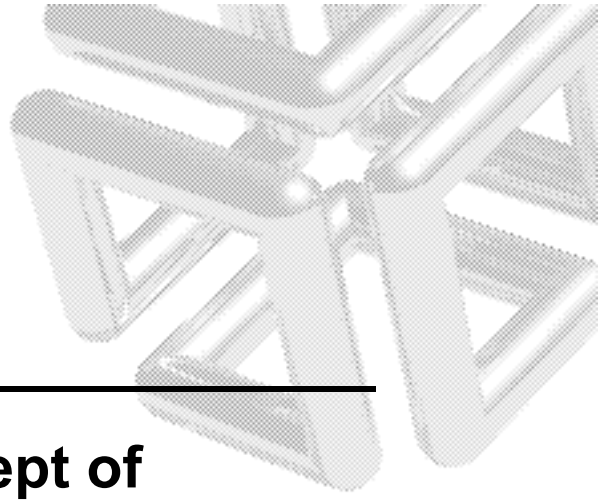
Topics



- **A Bit of History**
- **What You Can Do Today**
- **Possibilities for the Future**

SiliconGraphics
Computer Systems

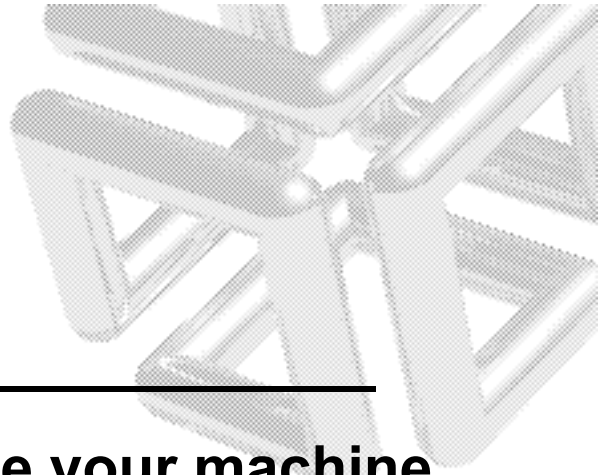
A Bit of History



- **Started with a more specialized concept of scheduling**
- **Has become more general over time**
- **Still have some items we see the need to do**

SiliconGraphics Computer Systems

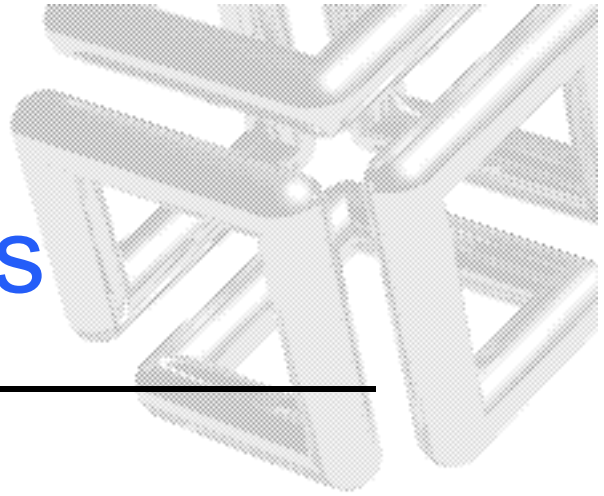
What you can do Today



- **Think about how you want to schedule your machine**
  - Processor usage
  - Memory usage
  - Priority
  - Overall ‘best’ Throughput
- **You can reasonably schedule for 1, sometimes 2, of the 4**

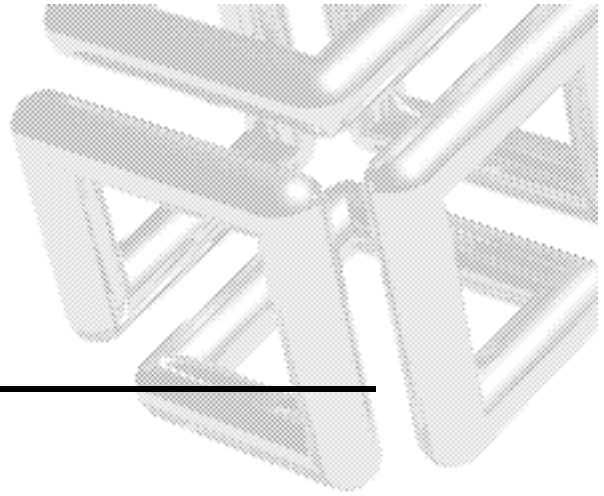
SiliconGraphics Computer Systems

Overview of Components



- **GRM**
- **Political Scheduler (psched)**
  - Gang Scheduler (GS)
  - Load Balancer (LB)
  - Multi-Layered User-Fair Scheduling Environment (MUSE)

Scheduling Domains



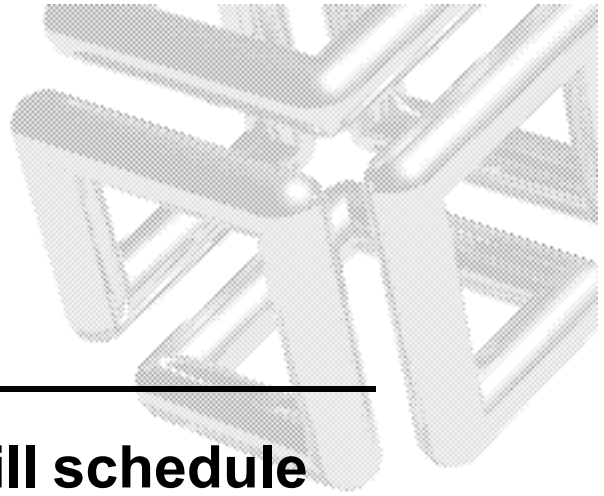
- **Command**
- **Application**

- **Batch**
- **Interactive**

- **Other combinations - site configurable (labels)**
- **scheduling domains must match GRM idea of domains**
- **GS, LB, MUSE, psched can be configured differently for the different domains**

SiliconGraphics Computer Systems

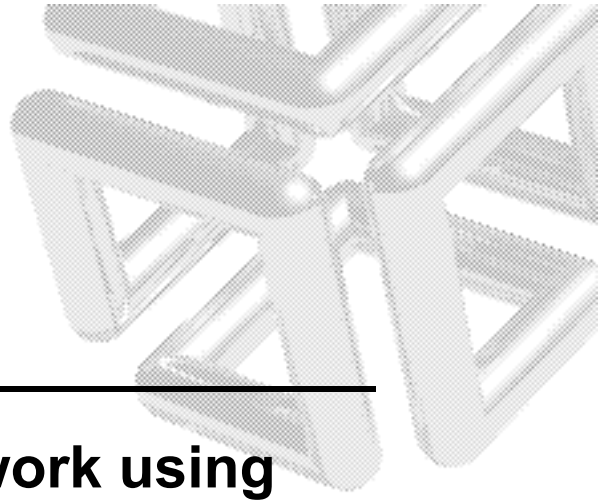
Scheduling Processors



- **Using default settings, UNICOS/mk will schedule processors**
- **This can have the effect of large apps languishing in queue for long periods if there are small apps to fill small spaces**
- **This can also leave lots of free memory if the apps are small in terms of size in memory**
- **Started apps will run to completion with no swapping and no migration**

SiliconGraphics Computer Systems

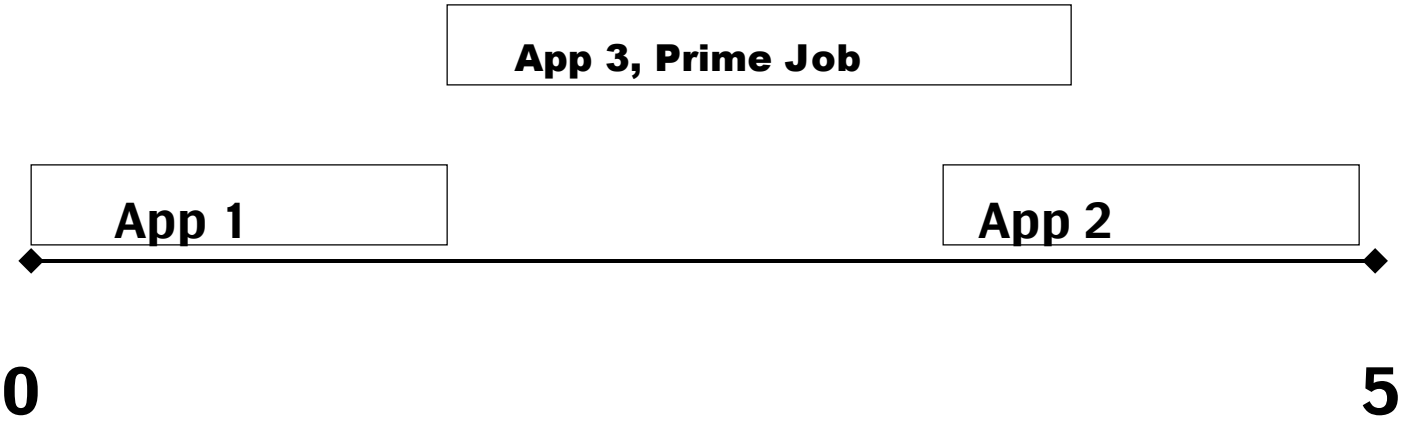
Scheduling by priority



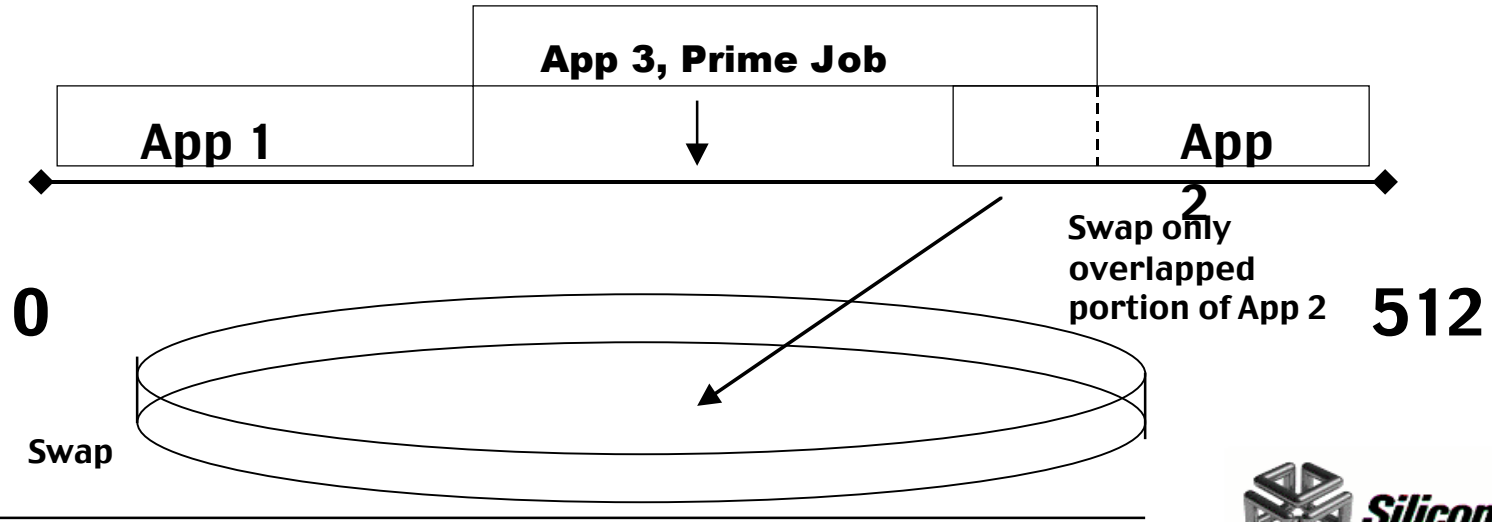
- **You can schedule very high priority work using psched and the use of the ‘prime job’ feature.**
- **Prime job will force all applications in a job into a set of processors if there isn’t an obvious hole for it.**
- **Use of prime job will cause swapping in many cases**
- **Prime jobs compete with one another if they overlap on any PEs.**

SiliconGraphics Computer Systems

Example of use of Prime Job



Example of use of Prime Job



SiliconGraphics Computer Systems

Use of Gang Scheduling

- **Helps where you have long and short running applications**
- **Allows preemption for important work**
- **Avoid swapping (use user limits and NQE limits)**
- **app_max 2> - need latest UNICOS/mk 2.0.3 and corresponding psyched**

SiliconGraphics Computer Systems

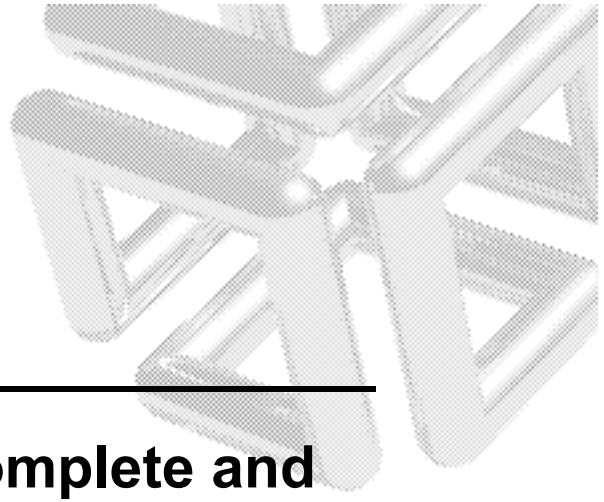
System data

- **Context switching - 2-3 milliseconds per PE for small numbers of PEs. Have measured ~300 milliseconds for the context switch of two 512 PE job.**
- **Migrate ~170 Mbytes/sec between PEs**
- **Migrate performance - by size of "chunk" (show 2 PE example)**
- **Swap - highly dependent upon devices and configuration. We are able to drive swap at ~device rates (ignoring contention).**

Interactions/Tidbits

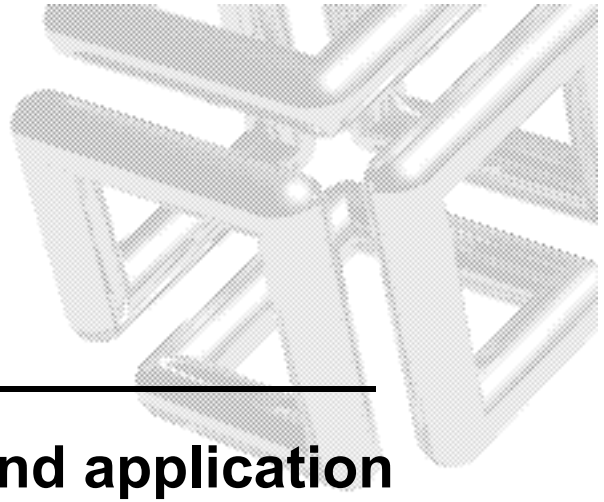
- **LB is separate from GS - can get benefit from running load balancer with app_max of 1 and no GS.**
- **Use of app_max 2**
  - might swap
  - prime job(?) [requires GS/RM] - set abs_app_max
- **LB and GS slice times can be short (increases psched overhead). Lower limit of 5 seconds, increase in 5 second increments.**
- **LB - look-ahead of one**

Interactions/Tidbits



- **Psched designed to operate with incomplete and stale data. Psview(1) output is "old" from last daemon request.**
- **Now log reason for migrate request (since 2.0.3)**

Reconfiguring on fly



- **Can reconfigure between command and application Pes on the fly(limited use, may work, don't have a lot of experience)**
  - stop psched
  - reconfigure GRM
  - change psched configuration to match
  - restart psched

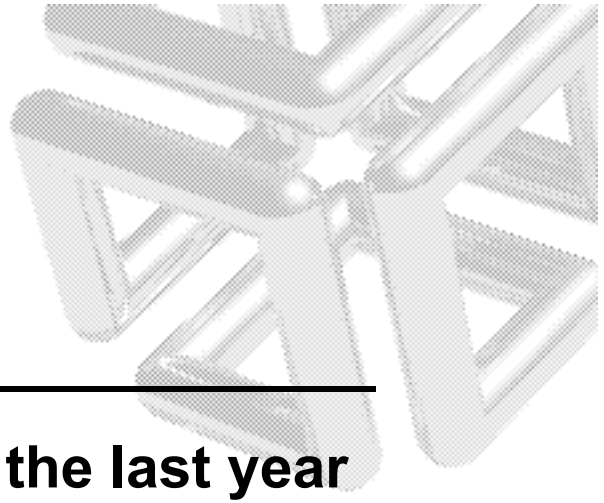
SiliconGraphics Computer Systems

Top areas for improvement

- **GRM queue - prioritization, starvation**
- **single PE applications**
- **Optimization**
- **Others**

SiliconGraphics Computer Systems

Questions and Answers



- **Some common questions asked over the last year and the current answers.**

SiliconGraphics Computer Systems

GRM Questions/Answers

- **How can psched daemon be configure to handle PE attributes?**
  - Psched currently does not schedule by PE attributes. If your GRM configuration includes attributes or labels such as ACID, GID or UID lists you may need to configure multiple psched scheduling domains that correspond to your PE attributes. See section 4.10.2 in the UNICOS/mk Resource Administration manual.

GRM Questions/Answers

- **Is it possible to stop and start multi-PE application allocation?**
  - As of the UNICOS/mk 2.0.2 release, the grmgr(8) command is capable of starting and stopping the allocation of multi-PE applications. If the grmgr stop_allocation option is run, GRM is prevented from launching multi-PE applications. Single-PE applications (commands) are not affected. The grmgr start_allocation option can then be used to reinstate the launching of multi-PE applications.

GRM Questions/Answers

- **Is it possible to configure a PE to run both single-PE applications and multi-PE applications?**
  - As of the UNICOS/mk 2.0.2 release, adminstrators are not able to configure a PE that runs both single-PE applications (commands) and multi-PE appliations. This change was made to protect the psched gang scheduling feature and to help reduce signal 34 problems. See grmgr(8) man page.

GRM Questions/Answers

- **When the load balancing and gang scheduling features of the political scheduler are running, GRM often places applications on PEs that already have active jobs, instead of placing them on PEs that have no jobs. Can GRM be configured differently to change this?**
  - Yes; if a site is running psched and has configured ap_max to be greater than 1 on application PEs, GRM may prefer placing applications on PEs that already have work instead of placing them on idle PEs. To correct this problem, configure the *close_max*, *contiguity*, and *label_match* attributes to hav a value of 0. GRM then places new application on PEs with fewest apps.

GRM Questions/Answers

- **How can GRM be configured to stop uneven PE utilization when serial (command) PEs have different memory sizes?**
  - In environments where there are mixed-size serial PEs, and access to the larger PEs is not controlled with labels or authorization lists, you should configure the *memory _fit* attribute to have a multiplier value of 0. This configuration allows GRM to balance the number of assignments on each PE; without this configuration, you may encounter uneven PE utilization on your machine

GRM Questions/Answers

- **Why do we sometimes encounter a number of applications waiting to start up because of Route BESU when there are enough PEs available? I realize they are waiting for a barrier, but what would prevent them from getting one?**
  - Working this issue - There are many reasons why GRM can fail to allocate a barrier tree. All of these reasons are related to partial planes, renumbered PEs, or places where PE number is disjoint (from the physical proximity viewpoint) due to the numbering algorithm.

GRM Questions/Answers

- **How can I override all GRM configuration limits, including labeled PE restrictions, when running online diagnostics?**
  - A privileged user can override the following GRM configuration attributes by using the mpprun -l lpe command:
    - PE labels
    - Service lists
    - UID lists
    - GID lists
    - ACID lists

psched Questions/Answers

- **Why do migrations fail with psched?**
  - Psched attempts migration is based on information that it has obtained from the system. While analyzing this data, conditions can change (process exit, launch, I/O, et.) which affect the success of the migration that psched decides to attempt. The most common reasons for migration failure are that the application no longer exists (EINVAL) or that the application cannot be currently frozen(EAGAIN/EBUSY).

psched Questions/Answers

- **Where are migration failures reported?**
  - Migration failures are reported in the psched log.

psched Questions/Answers

- **What are the strategies, goals, and implementation rules for the psched application load balancer?**
  - Perform only one migration per load balancing cycle.
  - Make recently migrated applications ineligible for migration
  - Attempt to ensure that load balancing actions do not involve swapping.
  - Assume that PEs in load balancing domains are configure similarly.

psched Questions/Answers

- **What is the best way to stop the psched(8) daemon?**
  - Psched should be considered a critical daemon. If you need to stop psched, never use the *kill -9* command because it does not allow psched to perform necessary cleanup tasks or to reset kernel information before it terminates, Instead, use the sdaemon command to start and stop psched so that the daemon can shut down properly.

psched Questions/Answers

- **Can the mpprun -f command option be used with psched(8)?**
  - The *mpprun(1) -f* command and psched gang scheduling can coexist. This affects launch only, after launch the application competes with other applications in the system.

psched Questions/Answers

- **How can I display the execution status of parallel applications (in memory, swapped, swap-in, swap-out, etc.) and display relative priorities, even for swapped jobs?**
  - As of the UNICOS/mk 2.0.2 release, the *psview(8)* command shows swapping status with the -g option; however status is somewhat delayed because it is only updated at each gang scheduling cycle.

psched Questions/Answers

- **What is the difference between the Global Resource Manager (GRM) and psched?**
  - GRM is part of the UNICOS/mk kernel, while psched is a daemon. GRM does no perform CPU scheduling and has no concept of global scheduling in terms of laod balancing the system. It simply assigns PEs at launch time, assigns Global Segment Registers, and allocates barrier trees. GRM is a required system component; it does not need psched psched needs information from GRM to do its job. Configurations must be compatible between GRM and psched.

psched Questions/Answers

- **Is it recommended to run multiple parallel jobs in the same PEs, even if they oversubscribe memory?**
  - We do NOT recommend extreme oversubscription of the CRAY T3E system. On this architecture, throughput is increased by running jobs sequentially, rather than suffering the overhead of context switching between jobs. Can use prime job to allow a job to run to completion while allowing other multiple parallel jobs to share PEs.

psched Questions/Answers

- **How do better nice values influence CPU cycles?**
  - Nice values have no impact if MUSE is enabled. If MUSE is not enabled, nice values are used to calculate the length of a gang scheduling time slice. Subsequent changes of nice values take effect on the next time slice.

psched Questions/Answers

- **Do programs with better nice values preempt programs with worse nice values?**
  - In the tradition of the UNIX nice value, psched does not preempt applications based on their nice value. Although the current job is not preempted by programs with better nice values, the high-priority program is chosen to run at the next time slice. The program's priority is directly impacted by its nice value.

psched Questions/Answers

- **Is there an administrator utility that controls the affect of nice values to a reasonable degree?**
  - Yes administrators can use the renice(1) command.

psched Questions/Answers

- **What is the impact of nice values with regard to MUSE and gang scheduling?**
  - Time-slice calculations for application are based on either MUSE or nice values

psched Questions/Answers

- **Are the accounting records accurate and reproducible?**
  - The accounting records are accurate; that is, they correctly reflect the resources utilized by the process or application. Howver, they may not be exactly reproducible.

psched Questions/Answers

- **Is there any logging of scheduling actions, in order to assist with tuning and debugging?**
  - Yes, psched logs all decision except those concerning gang scheduling. Because there are so many decisions, this logging information is conditioned by a DEBUG mode setting (0 is not active; 1 is full debugging):
  - psmgr -c set/ObjMgr/Debug 1

psched Questions/Answers

- **What happens if psched domains overlap?**
  - Multiple domains are not bounds checked by psched. This means domains that overlap can cause scheduling problems; therefore, you should not overlap domans. Administrators can use the *psview(8)* command to check the configured PE ranges of scheduling domains.

psched Questions/Answers

- **Is it OK to oversubscribe memory with psched?**
  - We do NOT recommend extreme over subscription of the CRAY T3E system. If PE memory is oversubscribed, psched may trigger local swapping when switching from one gang to another. Psched schedules PEs in such a way to maximize CPU utilization, but this may not be the most efficient memory utilization. Swapping is local to each PE's memory manager. Swapping may also take up a lot of the machine time if large applications are involved.

psched Questions/Answers

- **What is the best migration configuration?**
  - In the worst case, applications migrate at a speed of about 170 Mbytes/sec. Once an application is migrated, it should not be chosen for migration in the near future. All applications are exempted from migration for the amount of time specified by the *MigrationDelay* configuration value. The size of typical applications should be use to determine how much time the system wil need for migration. The *MigrationDelay* value should be set to balance migration overhead with reduced PE utilization caused by fragmentation. If your goal is to not have jobs interrupted, set the value to a multiple of the gang scheduling time slice . If your goal is full machine utilization, set the value less than the gang scheduling time slice.