Cray X1 System Administration -- UNICOS/mp 2.4 Update

Peggy Gazzola
Cray, Inc.

Abstract:
This session will cover general Cray X1 administration. Discussion topics will include system configuration, system monitoring, and problem diagnosis.

I presented a very general UNICOS/mp system administration talk last year for CUG. This year, another request was made for a similar session. Now that there are more Cray X1 customers and more field experience with the UNICOS/mp Operating System, I decided to focus this talk on more specific UNICOS/mp topics, in particular some of the new features introduced with the 2.4 Release of the UNICOS/mp Operating System.

Some of the changes introduced in UNICOS/mp 2.4, for example some operating system changes focusing on improved system performance, are inherent in the OS. No action is required by the administrator or end-user to access these enhancements. A handful of the features do require explicit steps by the system administrator and/or the end-user. These are the items highlighted here.

The 2.4 Release of the UNICOS/mp Operating System was initially made available on 22 March, 2004. The UNICOS/mp 2.4 Release Overview (Cray publication S-2336-24) describes the release content, and is a good general overview of the changes included with this major release of the Cray X1 OS.

The key UNICOS/mp 2.4 features covered in this presentation are:

- Additional application placement and scheduling capabilities
- Node numbering changes
- Ability to run multiple programs as single apteam (MPMD)
- Dynamic large page tuning
- Path-managed disk driver**
- Fibre Channel IP Bonding driver**
- ADIC ™ StorNext File System client**

** these last 3 items were Limited Availability features with UNICOS/mp 2.3. They are generally available with the UNICOS/mp 2.4 release.

The UNICOS/mp support model is similar to the model Cray used for the UNICOS/mk software on the Cray T3E systems. Major releases are issued 1-3 times per year, with the frequency decreasing as the operating system matures. Important fixes are collected and integrated into the released leg of the operating system every week, for exposure on in-house systems. If the exposure period proves successful, an update release will be shipped consisting of a new UNICOS/mp kernel and/or any modified commands. The update releases occur potentially weekly, but less frequently if there are no important fixes to ship or if the update shows stability problems.

UNICOS/mp 2.4 Feature Highlights

1) Application Placement and Scheduling Enhancements

The UNICOS/mp Placement Scheduler (Psched daemon) is responsible for assigning resources to applications on the Cray X1 system. The Cray X1 system normally consists of one OS/Support flavored node, used for the bulk of the operating system and for the command load on the machine. The remainder of the nodes are configured as Application nodes.

In UNICOS/mp 2.3 and earlier releases, Psched assigned application resources based strictly on processor count. In a standard configuration, with no

processor oversubscription, Psched could place an application load of up to 4 MSPs or 16 SSPs (or some combination thereof) on a single node. The memory use of an application was not taken into consideration. This could cause resource issues on the system. Consider, for example, eight single-MSP applications on the system. Four of those applications require 6 GBytes of memory each; the remaining four require 1 GByte of memory each. The application nodes on the system have 16 GBytes of memory. If the four 6 GByte applications happen to be placed on the same node, and the four 1 GByte applications are all placed on another node, the system will have to begin swapping on that first node. If Psched's placement algorithm could take into consideration the memory requirements of each of these applications, they could be split across those same two application nodes with no swapping overhead.

As of UNICOS/mp 2.4, application memory requirements can be taken into consideration when Psched places an application. Psched now supports an application Resident Set Size (RSS) memory limit. The user may specify a per PE RSS limit on the aprun(1) command line via the *-m size* or the *-c memoryuse* argument. Psched will consider the RSS memory requirements for the application when selecting a target node for placement. For applications with no explicit RSS specification, a default value is assigned.

***** NOTE ***** Applications must be relinked with the UNICOS/mp 2.4 libc library before RSS memory limit use is enabled in Psched. Applications linked with a pre-2.4 libc may abort in startup if RSS memory limit use is enabled.

There are three new Psched configuration parameters associated with this feature:

/Global/UseMemoryLimit
/Global/DefaultMemoryMsp
/Global/DefaultMemorySsp

The UseMemoryLimit parameter is disabled by default; set this value non-zero in the Psched configuration file (/etc/psched.conf), or via the psmgr(8) command to enable use of RSS memory limits in Psched. This parameter is temporary, for UNICOS/mp 2.4 only. As of UNICOS/mp 2.5, the Psched RSS memory limit capability will always be enabled.

The DefaultMemoryMsp and DefaultMemorySsp values are assigned to applications that do not explicitly specify RSS memory limits on the aprun command line. The default value for DefaultMemoryMsp is 1/4 of the smallest application node memory size on the system; the default value for DefaultMemorySsp is 1/16 of the smallest application node memory size.

Users may determine their application RSS memory requirements using a new acctcom(1) option, the -L option. When acctcom is run with the –A and -L options (-A for application records), the output includes a column headed "APP PE MAX MEM" which displays the application's per PE maximum memory use in MBytes.

2) Node Numbering Scheme Changed

A future UNICOS/mp platform, the Cray X1E, will have two nodes per physical processor module, rather than one node per module as in the Cray X1. In preparation for this change, some system utilities were changed in UNICOS/mp 2.4. On systems with single-node modules, the nodes are now numbered differently, using only even numbers. For example, an 8-node Cray X1 system now consists of nodes 0x000, 0x002, 0x004, 0x006, 0x008, 0x00a, 0x00c, 0x00e. The former familiar node numbers are now classified as module numbers. Commands that display node or processor numbers have been modified to support this distinction.

For example, the snflv(1) command, which is used to set node flavors, now displays the new node numbers, as shown here on a 4-node Cray X1 system:

```
x1% snflv
Node  Node
Start Count Type Resource Flavor(s)
===== ===== ==== ==================
0x000    3   X1  Application
0x006    1   X1  Support OS

x1% snflv -v
Module  Node Type   Pages Mbytes SSPs MSPs Resource Flavor(s)
====== ===== ==== ======= ====== ==== ==== ==================
 0x000 0x000  X1   522992  32687   16    4 Application
 0x001 0x002  X1   522992  32687   16    4 Application
 0x002 0x004  X1   522996  32687   16    4 Application
 0x003 0x006  X1   519632  32477   16    4 Support OS
```

The psview command also displays the new node numbers in the standard display (no arguments). Some psview displays have been modified to list both nodes and modules; in some cases, the word "Module" has replaced the word "Node" in the output.

```
x1% psview
Posted list is empty

Launched applications: 4
Age          Apid   User|Uid     w:d:N   Mode Node  Num Command    Note
  1:19     268016      user0     1:4:0   NbFS 0x000   1 l1.exe     -
  0:58     301026      user0     1:1:0   NbFS 0x004   1 l1.exe     -
  0:16     387403      user0     1:2:0   NbFS 0x002   1 l1.exe     -
  0:00     425080      user1     1:1:0   NiFS 0x004   1 equake     -
x1% psview -m
                 000 001 002
Command    Apid  000 002 004
======== ======= === === === ===
l703.exe  268016 __4_____________
l502.exe  301026 __________1____
l502.exe  387403 ______2________
equake    425080 __________1____
```

The ps command includes a processor number in the ps -l output if the environment variable _XPG is set to 0. The displayed processor numbers in UNICOS/mp 2.3 and earlier for a 4-node Cray X1 system would range from 0x00 - 0x3f. In UNICOS/mp 2.4, the processor numbers are 0x00-0x0f, 0x20-0x2f, 0x40-0x4f, 0x60-0x6f.

3) Multiple Program, Multiple Data (MPMD)

UNICOS/mp 2.4 supports the Multiple Program, Multiple Data feature, the capability to run multiple programs as part of a single application team. This allows independent programs to communicate using a shared memory programming model. The applications must all be of the same mode (MSP or SSP), and must use MPI, SHMEM, or CAF for communication.

The user can run an MPMD program by specifying multiple applications on the aprun command line, using ':' delimiters, for example:

x1% aprun -n 8 prog0 : -n 32 \ prog1 : -n 16 prog2

This feature is supported in conjunction with the Cray Programming Environment Release 5.2.

4) Dynamic Large Page Tuning

In prior releases of the UNICOS/mp OS, a variety of large page parameters were available for tuning via the systune(8) command. The parameters specified high water percentage values for each supported page size, for OS and for Application flavored nodes. When large pages were required on a given node, these values were used by the kernel as guidelines when coalescing large pages. These tunables have been replaced by a new pair of parameters used for dynamic tuning of large pages. The new parameters:

app_text_page_weight
app_other_page_weight

determine a page-weight ratio of text pages to data/other pages. The default values are 1 and 10 respectively. When an application is placed via aprun(1), the OS starts forming large pages based on the page sizes specified on the aprun(1) command line, rather than waiting until the large pages are needed by the application. The page-weight ratio is used when forming large pages for the application.

5) Path-managed Disk Driver

The UNICOS/mp path-managed disk driver (pmd) was a limited availability feature in the UNICOS/mp 2.3 release. The pmd driver becomes a general availability feature in UNICOS/mp 2.4. This new driver replaces the former dksc disk driver, which was inherited from SGI Irix ™. The dksc disk driver had limited dynamic path management and failover capabilities, and these capabilities were only supported for devices configured using the XLV volume manager. The Cray X1 system uses RAID devices supporting multiple host connections; the path-managed disk driver offers better control over multiple paths to each device, and better failover capabilities in the event of a failure of a disk path component.

The disk device naming convention has changed with the pmd driver. The dksc disk devices were named for the path from the host to the device. The pmd disk devices are named for the physical location of the device, incorporating the disk chassis number and the RAID Controller Brick (C-Brick) slot number within the chassis, along with the LUN number and the partition number. Multiple active paths to each disk are supported. The active path(s) to the disk can change without the disk device name changing.

Cray X1 systems shipped prior to the release of UNICOS/mp 2.4 included four host connections to each RAID subsystem, with two of these connections commented out in the system configuration file on the CWS. With the pmd driver, these connections can be brought into the configuration resulting in four paths to each disk device.

The pm(8) command was introduced with the pmd driver. pm(8) is used to monitor and manage the disk device paths. The default pm(8) display shows each disk device (LUN), and the four underlying paths associated with each disk, along with the state of each path. Each disk device is configured with two primary paths and two alternate paths. The primary paths are used as the active paths, unless a path switch occurs (manually, or due to failure of both primary paths). The pmd driver uses a round-robin algorithm to send i/o requests across both active paths. The path state is stored in non-volatile RAM on the RAID controller, so it is maintained across Cray X1 system reboots (a probing procedure

is invoked at OS boot time to identify all LUN paths and states). The conversion process for moving from the dksc to the pmd disk driver is documented in the UNICOS/mp Disk and Filesystems Reference Manual (S-2377-24), in Appendix A. The process involves the following set of steps:

1) designate a root file system as the pmd conversion root device
2) add 'pmd_enable' to NVRAM file on CWS for the designated root device
3) boot Cray X1 to single-user mode
4) run 'pm -v conversion' to display dksc vs. pmd disk device names
5) modify /etc/fstab file to reference new pmd device names
6) modify NVRAM file on CWS to reference new pmd name for root device
7) reboot UNICOS/mp using updated NVRAM, pmd root file system

**** NOTE **** As of the next UNICOS/mp release, the dksc driver will no longer be supported. All Cray X1 systems must go through the dksc to pmd conversion process under UNICOS/mp 2.4.

6) Fibre Channel IP Bonding Driver

Another feature which was offered with limited availability in the UNICOS/mp 2.3 release is the Fibre Channel IP Bonding driver. This driver is also generally available in the UNICOS/mp 2.4 release. The Fibre Channel IP Bonding driver provides a channel bonding feature for fibre channel network interfaces -- multiple interfaces can be configured as "slaves" of a single logical interface, providing a network failover capability. The Cray X1 system supports channel bonding for two fibre channel interfaces between the Cray Network Subsystem (CNS) and the Cray X1.

UNICOS/mp currently supports the "active backup" mode for bonded interfaces. In this mode, one interface path is used for all network traffic for a given bonded interface. If that path fails, traffic is routed to the alternate (backup) slave interface. The new bfc(8) command is used to manage and monitor the bonded Fibre Channel interfaces.

Creating a bonded interface involves configuration steps on both the Cray X1 system and on the CNS. The bonding interfaces on UNICOS/mp are named bfc0, bfc1, and so on. They are created via network startup scripts invoked at system boot time. A sample startup script, /etc/init.d/bond.local, is included with the UNICOS/mp 2.4 release. This script contains the necessary commands to create the bfc interfaces, and assign the underlying slaves to a given bfc interface. The configuration steps are described in the UNICOS/mp Networking Facilities Administration guide (publication S-2341-24) and the Cray Network Subsystem (CNS) Software Installation and Administration guide (S-2366-13).

7) StorNext File System (SNFS) Client and Related Fibre Channel Support

The final limited availability feature that was initially offered with UNICOS/mp 2.3, and is now generally available with UNICOS/mp 2.4, is the Cray X1 client for the Advanced Digital Information Corporation (ADIC ™) StorNext File System. This feature allows the Cray X1 system to support a Storage Area Network (SAN) file system, with multiple hosts sharing access to the same data.

The StorNext client communicates across a private network with a StorNext metadata server (MDS). The MDS controls client access to user data on the SAN. SNFS file systems are configured and created on the MDS, and mounted on available client systems.

StorNext is enabled on UNICOS/mp systems using the chkconfig utility,

and is started and stopped by the /etc/init.d/cvfs script. StorNext file systems are mounted using the mount(8) command with the '*-t cvfs*' argument to specify the StorNext file system type.

In order to support the SNFS client, Fibre Channel fabric support has also been included in UNICOS/mp 2.4. In a SAN environment, a Fibre Channel switch is used to provide multiple host connections to the same disk storage. At system boot time, a fabric probe is initiated to identify all available paths to fabric-attached disk LUNs. The pmd driver limits the number of paths per LUN to the number of host ports on the device, eliminating the confusion of potentially dozens of paths per LUN on a system with multiple host connections to a Fibre Channel switch.

The StorNext client software is currently built into the UNICOS/mp kernel. SNFS commands are installed in /usr/cvfs/bin on the Cray X1 client. Administrator commands include cvadmin(1M), which is used to monitor status and view attributes of SNFS file systems, and cvlabel(1M), which displays available disk devices on the system. User commands include cvcp(1), a StorNext copy command that utilizes SNFS I/O strategies for better performance, and cvmkfile(1), which can preallocate and align a user data file for better I/O performance on a StorNext file system.

This completes the summary of a subset of the features introduced with the 2.4 release of the UNICOS/mp operating system. The UNICOS/mp Release Overview, Cray publication S-2336-24, contains additional information on the content of this release. All Cray, Inc. publications supporting the UNICOS/mp 2.4 release are available at the Cray public website, http://www.cray.com, under the Training & Support section.

About the author: Peggy Gazzola is an OS Product Support Specialist working in the Product Support Group at Cray, Inc. She has been with Cray for 20 years, primarily in the area of OS support, and specializing in file systems. Address/phone/email:
1340 Mendota Heights Road
Mendota Heights, MN 55120
651-605-8966 peggy@cray.com