Cray X1 System Administration and Configuration

Peggy Gazzola, Cray Software Product Support

The Cray X1 system consists of one or more Cray X1 mainframe chassis, along with an array of other hardware components and systems. This paper touches on a brief subset of the duties required to administer and configure a Cray X1 system, beginning with an overview of the system components.

An administrator of the Cray X1 system must have an understanding of the complete system configuration. This includes one or more mainframe chassis, liquid-cooled (LC) or air-cooled (AC), one or more I/O chassis (IOC), one or more I/O drawers (IOD), a set of RAID disk subsystems, one or more Cray Network Subsystems (CNS), a Cray Programming Environment Server (CPES), the System Control Facility, and a Cray Workstation (CWS) as the central point of operation and configuration of the system.

The Cray X1 hardware configuration is managed on the CWS by the x1config utility. The configuration is maintained in a text file, named /opt/craycfg/cray.cfg by default. The configuration file describes the physical components of the system, including node modules, router modules, i/o chassis, and IODs. Every configuration file must define one system component, either an LC or AC system. Within the system component, chassis components are defined. An LC system consists of two brick pair components, each of which contains eight node slots, four router slots, and two Brick Cooling Unit (BCU) slots. AC systems consists of four node slots, two router slots, and one BCU slot.

The full system configuration can be dumped using the x1config "-d all:DUMPFILE" option. This will place the complete configuration in the file "DUMPFILE", listing every definable component in the system, including each chip on each node module. It is not necessary for the configuration file to include every physical component. A standard set of subcomponents will be associated with each defined component as appropriate. For example, the configuration file must contain nodemodule specifications for nodes with attached System Port channels (SPCs), but the individual processors, memory chips, E chips are not normally included.

In case of a component failure, for example a single processor on a node module (SSP), that component may be marked down in the configuration file by explicitly listing the component, and setting its state to 'Disable'.

The RAID subsystem configuration is managed from the CWS by a variety of commands, called the Cray Storage Management (csm) utilities. There is also a graphical user interface, SMgui, which may be used to monitor the RAID subsystem.

The System Control Facility consists of L0 and L1 processors which reside on node modules, router modules, BCUs, and IODs. The L0 processors control and monitor the power and cooling subsystem. The L1 processors provide the operational interface to the system, for Master Clear and initialization of chips, boundary scan, and mainframe hardware error reporting.

The CWS is the focal point for system operation and monitoring. The bootsys(8) command on the CWS performs all necessary mainframe initialization steps to bring up the operating system. Similar to earlier Cray systems (PVP, T3E), the bootsys(8) command invokes a machine specific command, bootx1(8) in this case, to load the operating system and bring the mainframe to single-user mode.

System dumps are also performed from the CWS, using the dumpsys(8) command. Initially, dumpsys(8) calls the hwerrdump(8) command to capture hardware state information from the mainframe. Next, like the dumpsys operation on the T3E, the Cray X1 dumpsys(8) will dump one node (typically an application node) to the CWS, then invoke the mboot(8) command to perform a maintenance-mode boot on the dumped node. The remaining nodes are dumped by default to a date-stamped subdirectory of the /dumps directory (file system) on the mainframe.

The CWS contains a central repository, /opt/craylog, for various system component log files, including l0, l1, cpes, ops, cns. Pertinent error, warning, and information messages are forwarded to the CWS by these components and placed in the corresponding log file. The CWS also houses the x1ms and x1wacs commands, normally used by Cray engineers to monitor the system.

The Cray Programming Environment Server (CPES) is a Sun V480 system running Solaris. This system is used for compiling and linking user codes targetted to run on the Cray X1. All users requiring access to the Programming Environment must have accounts setup on the CPES to match their accounts on the Cray X1 (same uids). The interface to the Programming Environment is "invisible" to users. A user logged into the Cray X1 mainframe issues a 'cc' command, for example, and that 'cc' triggers the execution of the C compiler on the CPES. Data files are shared via NFS. All directories from which users on the Cray X1 may issue any compile commands must be mounted on the CPES, under a designated mount point such as /x1. Relative paths under /x1 on

the CPES must match the full paths on the mainframe. The CPES /x1/opt/ctl file system is exported to the Cray X1, mounted as /opt/ctl.

So what about the mainframe? The Cray X1 mainframe runs the UNICOS/mp operating system. UNICOS/mp is based on IRIX, but includes a number of extensions and modifications for support of the Cray X1 system. In particular, enhancements have been made for support of the multi-streaming processor (MSP) available on the Cray X1. Unlike IRIX, UNICOS/mp has limited device driver support: fibre channel device drives for disk and network are provided.

UNICOS/mp includes the concept of node flavors, similar to the T3E defined PE types. The Cray X1 supports OS, Support, and Application flavored nodes. A typical single-chassis LC configuration consists of one combined OS/Support node (for running the bulk of the operating system, and all user commands), and 15 Application nodes. Users launch work onto the application nodes via the aprun(1) or mpirun(1) commands. Another carryover from UNICOS/mk is the Political Scheduling daemon, psched. The psched daemon on UNICOS/mp combines the functions of the UNICOS/mk Global Resource Manager (GRM) and the psched daemon. UNICOS/mp psched supports the load balancer and gang scheduler functions, provided to manage resources in the application nodes. The configuration for psched is similar to the UNICOS/mk format, via the /etc/psched.conf file as well as via the psmgr(8) command.

User limits under UNICOS/mp are managed by a limits database, created by the administrator using the limit_mkdb(8) command. There are four unique limit scopes defined, batch/command limits (BC), interactive command (IC), batch/application (BA), and interactive application (IA). Limits are explicitly defined for each scope. For each limit, two types may be defined: INITIAL and MAXIMUM. The INITIAL limit type sets the current, default limit. The MAXIMUM limit type is the upper bound allowed for a user. Available limits include core file size, memory size, cpu time, etc. The limits are described in the limit_mkdb(8)manual page.

One of the more significant differences an administrator will encounter with UNICOS/mp on the Cray X1 system vs. UNICOS or UNICOS/mk on earlier Cray systems is in the area of disk configuration. On a UNICOS or a UNICOS/mk system, the administrator had to manually (or with the assistance of the installation tool) configure the physical disks. On the Cray X1 system, UNICOS/mp uses a hardware discovery feature at boot time to identify the physical disks attached to the system. The device nodes are automatically created at boot time based on the discovery.

As previously noted, the RAID device configuration is managed from the CWS. This configuration is normally completed at the factory, prior to machine shipment. The csm utilities currently support six distinct RAID configuration options for a single storage brick. The default configuration uses two of these options, one targeted for the system disk, containing root and other standard utility file systems, and a second targeted for large data bandwidth file systems. The RAID devices are partitioned into logical units, or LUNs. The csmadd(8) command on the CWS is used to configure the RAID subsystem.

The CWS-resident Cray X1 configuration file, /opt/craycfg/cray.cfg by default, describes the iopath information, the connections from the mainframe to the RAID controllers. The I/O chassis are listed, with associated IODs (an IOD is described as an "iomodule" in the configuration file). The IODs each contain an A board and a B board; each board supporting two channel adapters, each channel adapter supporting two PCI-X slots. Dual-ported Fibre Channel host bus adapter cards are used in the PCI-X slots, and provide the connections to the RAID controllers. The dual ports are referenced as functions in the configuration file, 'func0' and 'func1'. The IOD configuration section in the cray.cfg file lists each board, channel adapter, PCI-X slot, and function associated with disk (or network) devices.

When the system is booted, the configured LUNs are discovered, and disk device nodes in /dev/rdsk (character special) and /dev/dsk (block special) are created for the discovered LUNs. The administrator can then use the parts(8) command to configure slices on the LUNs. The parts(8) command, created for UNICOS/mp, performs some of the functions of the IRIX fx(8) command. It reads and writes disk partition information for a particular LUN into the volume header on the device. A system administrator uses the parts(8) command to create slices on the disks, in preparation for creating file systems. The parts(8) command requires an input file which specifies the desired slice configuration for a device. The input file includes partition number, partition type, and partition size. The partition type can be xfs, log, swap, or raw (swap and raw are equivalent). The default partition type is xfs, used for all file systems. The log type is used to designate partitions for use as external logs for XFS file systems. The partition size parameter can be specified as a number of units (blocks, kilobytes, megabytes, gigabytes), or as a percentage of the total disk volume capacity. A set of standard parts command files are provided with the UNICOS/mp operating system, under the /etc/parts directory. The files cover options for configuring LUNs with a single slice, two slices each comprising 50% of the LUN, four slices each using 25% of the LUN, eight slices -- four

small log slices and four remaining each using 25% of the space, and finally a standard system drive file, with three XFS slices and one raw (swap) slice.

The assigned disk device names for the device entries in /dev/dsk (and /dev/rdsk) are of the form dksWdXlYsZ. The W from dksW represents the disk number assigned in the cray.cfg configuration file, in the iomodule definition section. The X from dX is the Fibre Channel loop-ID of the host-port on the RAID controller (each controller has 4 host ports, so this value will be in the range 0-3). The Y from lY is the associated LUN number. Note that the lY portion of the name is omitted for LUN 0 devices. For example, dks2d0s4 references slice 4 of LUN 0 on disk 2 in the configuration file. And finally, the Z from sZ is the slice number as configured via the parts(8) command.

Like IRIX, UNICOS/mp supports XLV, the logical volume disk driver. XLV allows you to combine physical disk slices into larger, logical units. The xlv_make(8) command is used to create the logical volume objects. XLV is required, for example, to configure an XFS file system with an external log device. XLV is also used to create, for example, striped logical volumes, to allow for greater bandwidth to the disks.

This is not a complete review of Cray X1 System Administration and Configuration, nor is it intended to be so. It is only a brief glimpse, highlighting a few of the key points that may be new to administrators familiar with Cray systems running UNICOS and UNICOS/mk. A more thorough presentation of the topic is available in a variety of Cray X1 Software Publications, as well as through the Cray Technical Training Department.