CRAY X1 System Administration UNICOS/mp 2.4 Update

Peggy Gazzola

Software Product Support

Cray, Inc.

UNICOS/mp 2.4 Update

- Released 22 March, 2004
- UNICOS/mp 2.4 Release Overview (S-2336-24)
- Emphasis for this OS release:
  - Improved system performance, resiliency
  - Additional application placement, scheduling capabilities
  - Accounting changes to better distinguish between SSP and MSP cpu time
  - Ability to run multiple programs as single apteam (MPMD, Multiple Program, Multiple Data)
  - And…

UNICOS/mp 2.4 Update

- **General Availability for UNICOS/mp 2.3 Limited Availability Features**
  - **Path-managed disk driver**
  - **Fibre Channel IP Bonding driver**
  - **ADIC StorNext File System client**

UNICOS/mp 2.4 Update

- Support follows UNICOS/mk model
- Updates (fix packages) potentially released weekly
- Field Notices for critical problems
- Supported via updates until next major release is available
- Limited critical fix support after final update

CUG 2004

UNICOS/mp 2.4 Update

- **Application Placement and Scheduling Enhancements**
  - **psched configuration changes**
  - **Formerly, placement decisions based strictly on processor requirements**
  - **As of UNICOS/mp 2.4, can also take application Resident Set Size (RSS) into consideration**

UNICOS/mp 2.4 Update

- **New psched configuration parameters**
  - **/Global/UseMemoryLimit ** 2.4 only**
    - **Default value 0 (disabled)**
  - **/Global/DefaultMemoryMsp**
    - **Default value 1/4 application node memory**
  - **/Global/DefaultMemorySsp**
    - **Default value 1/16 application node memory**

UNICOS/mp 2.4 Update

- **Implementation of RSS for psched placement consideration**
  - **User determines memory requirements of application (using new acctcom(1) –L option)**
  - **User adds –m size or –c memoryuse=xxx to application aprun command line**
  - **Administrator enables RSS memory limit in psched**
    - **Set /Global/UseMemoryLimit non-zero in /etc/psched.conf**
    - **OR psmgr –c 'set /Global/UseMemoryLimit 1'**

***** NOTE applications must be linked with UNICOS/mp 2.4 libc***

CUG 2004

UNICOS/mp 2.4 Update

- **Node numbering change**
  - **For future UNICOS/mp systems (X1E)**
  - **2 nodes per module**
  - **Node #s 0x000, 0x002, 0x004, etc.**
  - **Various commands modified at UNICOS/mp 2.4 to support new numbering scheme:**
    - **snflv**
    - **psview**
    - **apstat**
  - **aprun –l specify module vs. node**

UNICOS/mp 2.4 Update

- **Sample output (node vs. module)**

```
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
```

UNICOS/mp 2.4 Update

- **Sample output (node vs. module)**

```
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____
```

UNICOS/mp 2.4 Update

- **Sample output (node vs. module)**

```
x1% setenv _XPG 0
x1% ps -el
 F S   UID     PID    PPID  C PRI NI  P        SZ:RSS    WCHAN TTY      TIME CMD
 0 S     0       1       0  0  20 20  *        569:38    ktwait  ?      0:04  init
...
 4 R     0     688       1  0  20 20  3e2      560:40         -  ?      0:07  pbs_sched
...
 a0 R  207    1985    1984  0  20 20  329 2307331:49667      -  pts/2  48:45  wrf.exe.t
 a0 R  207    2027    1985  0  20 20  340 2307075:49154      -  pts/2  49:06  wrf.exe.t
 a0 R  207    2028    1985  0  20 20  369 2307075:49154      -  pts/2  49:07  wrf.exe.t
 a0 R  207    2029    1985  0  20 20  381 2307075:49154      -  pts/2  49:06  wrf.exe.t
 a0 R  207    2030    1985  0  20 20  3ae 2307075:48898      -  pts/2  49:07  wrf.exe.t
 a0 R  207    2031    1985  0  20 20  3c1  576771:29698      -  pts/2  49:15  wrf.exe.t
 a0 R  207    2032    1985  0  20 20  321 2307331:49667      -  pts/2  49:09  wrf.exe.t
 a0 R  207    2033    1985  0  20 20  32c 2307331:49667      -  pts/2  49:09  wrf.exe.t
...
  0 R  224    9837    8851  0  20 20  3ec      541:21        -  pts/10 0:00  ps
```

UNICOS/mp 2.4 Update

- **Multiple Program, Multiple Data Feature**
  - **Allows multiple, independent programs to be run as a single ApTeam**
  - **Cray Programming Environment Release 5.2**
  - **Application-to-application communications supported with MPI, SHMEM, or CAF**
  - **All apps must be same mode (MSP or SSP)**
  - **Use aprun or mpirun with ':' separating apps:**
    - **aprun –n 4 prog0 : -n 2 prog1 : -n 4 prog2**
  - **Documented in Cray C and C++ Reference Manual (S-2179-52) and Cray Fortran Compiler Commands and Directives Reference Manual (S-3901-52)**

CUG 2004

UNICOS/mp 2.4 Update

- **Dynamic large page tuning**
  - **Former large page tunables removed**
    - **os_percent_nodemem_xx_pages**
    - **ap_percent_nodemem_xx_pages**
  - **Replaced with new text vs. other page ratio parameters**
    - **app_text_page_weight**
    - **app_other_page_weight**
  - **When application is placed, kernel tries to form required large pages on app node based on text vs. other page weight ratio**
  - **Default ratio 1:10**

UNICOS/mp 2.4 Update

- **Path-managed Disk Driver**
  - **Replaces dksc driver**
  - **Manages multiple active paths to a single LUN**
  - **Improved failover capabilities**
  - **pm(8) command for monitoring/managing paths**
  - **parts(8) still used for partitioning LUNs**

***** Cray-X1 systems must be converted to pmd driver before upgrading to UNICOS/mp 2.5***

UNICOS/mp 2.4 Update

- **Disk device naming conventions**
  - **dksc driver → device named for i/o path**

    dks<C>d<T>l<L>s<P>

    C → controller #

    T → disk target #

    L → LUN #

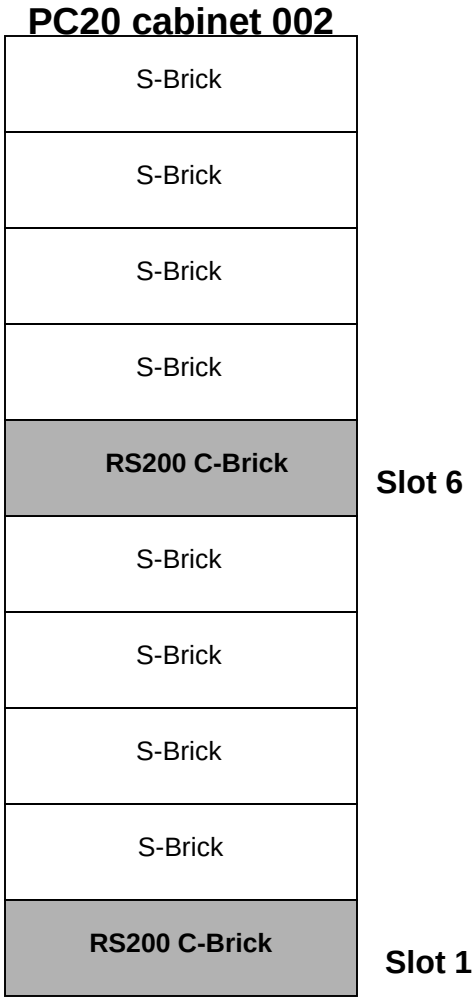
    P → partition (slice) #

    Example: dks2d2l1s0

    controller 2 target 2 LUN 1 slice 0

UNICOS/mp 2.4 Update

- **Disk device layout**



Device named for physical location

- PC20 chassis number

- PC20 slot number for C-Brick

- LUN number (generated by csm)

- slice number (generated by parts(8))

UNICOS/mp 2.4 Update

- **csm configuration example**

```
cws$ csmreport –O 2 6 layout
02d06  - RS200 starting at Chassis:  2   Slot:  6      Profile(02d06.prf)
+=============================================================+
|  02d08s SB201[20]          RG12C w/  2 LUNs                 |
|+------+------+------+------+------+------+------+------+------+------+------+------+-----+-----+|
|| 02d08_B1_L01  583.0GB                                | 02d08_B3_L03  291.5GB  ||
|+------+------+------+------+------+------+------+------+------+------+------+------+-----+-----+|
+=============================================================+
|  02d07s SB201[ 0]          RG12C w/  2 LUNs                 |
|+------+------+------+------+------+------+------+------+------+------+-----+------+------+-----+|
|| 02d07_A0_L00  583.0GB                                | 02d07_A2_L02  291.5GB  ||
|+------+------+------+------+------+------+------+-----+------+------+-----+------+------+------+|
+=============================================================+
|  02d06s CB200              N5884-530869-001                 |
|+------------------------------------------------------------------------------------------ --------+|
|| CTLR-A Optimal    10.0.117.252   A/1: Tid:  0 Alpa:0xEF                                          ||
||                   05.30.05.01    A/2: Tid:  2 Alpa:0xE4                                          ||
|+-----------------------------------------------------------------------------------------------------+|
|| CTLR-B Optimal    10.0.117.253   B/1: Tid:  1 Alpa:0xE8                                          ||
||                   05.30.05.01    B/2: Tid:  3 Alpa:0xE2                                          ||
|+-----------------------------------------------------------------------------------------------------+|
+=============================================================+
```

UNICOS/mp 2.4 Update

- **Disk device naming conventions**
  - **pmd driver → device named for physical location**

    pm<C>d<S>L<U>s<P>

    C → PC20 chassis #

    S → PC20 C-Brick slot #

    U → LUN #

    P → partition (slice) #

    Example: pm2d6L1s0

    chassis 2 slot 6 LUN 1 slice 0

UNICOS/mp 2.4 Update

- **Disk device naming conventions**
  - **pmd driver → underlying path identifiers**

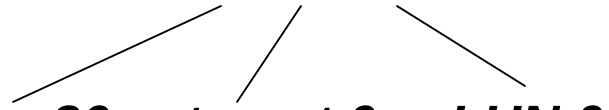
    fc<C>d<T>L<U>s

    C → host controller # (Disk identifier from cray.cfg)

    T → disk target # on the fibre channel

    U → LUN #

Example: **fc20d0L0**



UNICOS/mp 2.4 Update

- **pm(8) command**
  - **Monitor disk device paths**
  - **Perform manual path switches (failover)**
  - **View path information**
  - **Interactive status display (pm –r)**

UNICOS/mp 2.4 Update

x1# *pm pm4d1*

pm4d1L0

path	port	state	read blks	write blks	errs	MB/Sec
fc20d0L0	pri	active	307042174	633261550	0	0.00
fc22d2L0	pri	active	307862636	632467160	0	0.00
fc21d1L0	alt	standby	0	0	0	0.00
fc23d3L0	alt	standby	0	0	0	0.00

pm4d1L1

path	port	state	read blks	write blks	errs	MB/Sec
fc21d1L1	pri	active	309941777	634968704	0	0.00
fc23d3L1	pri	active	309681137	635141952	0	0.00
fc20d0L1	alt	standby	0	0	0	0.00
fc22d2L1	alt	standby	0	0	0	0.00

UNICOS/mp 2.4 Update

- **pmd error recovery**
  - **Normally two primary paths active, two alternate paths standby**
  - **In case of error on a path, the other active path is used**
  - **Consecutive failures on a given path result in that path being temporarily suspended**
  - **Persistent failures on primary paths trigger path switch to alternate paths**

UNICOS/mp 2.4 Update

- **pmd error pathswitch example:**

```
x1# pm pm4d4L0
pm4d4L0
          path   port     state   read blks   write blks  errs   MB/Sec
          ------ ------   ------  ----------- ----------- -----  ----------
      fc10d0L0   pri     active    6448561    20610032      0       0.00
      fc12d2L0   pri     active    6459440    19914049      0       0.00
      fc11d1L0   alt   standby          0           0      0       0.00
      fc13d3L0   alt   standby          0           0      0       0.00

x1# pm -p pm4d4L0                    # perform path switch on this LUN
...
```

CUG 2004

UNICOS/mp 2.4 Update

- **pmd error pathswitch example:**

on console:

May 15 11:01:13 5A:x1 unix: CPU 0VN2S11 (0x6b): NOTICE: pmd_pathswitch: initiating controller path switch of pm4d4L0

May 15 11:01:14 5A:x1 unix: CPU 0VN2S11 (0x6b): NOTICE: pmd_pathswitch: successful controller path switch of pm4d4L0

x1# pm pm4d4L0

pm4d4L0

path	port	state	read blks	write blks	errs	MB/Sec
fc10d0L0	pri	standby	6448561	20610420	0	0.00
fc12d2L0	pri	standby	6459440	19914404	0	0.00
fc11d1L0	alt	active	0	6840	0	0.00
fc13d3L0	alt	active	0	6832	0	0.00

UNICOS/mp 2.4 Update

- **dksc to pmd conversion steps**
  - **Add pmd_enable to NVRAM file on CWS**
  - **Boot system to single-user mode, run pm to display the pmd to dksc device node mapping:**

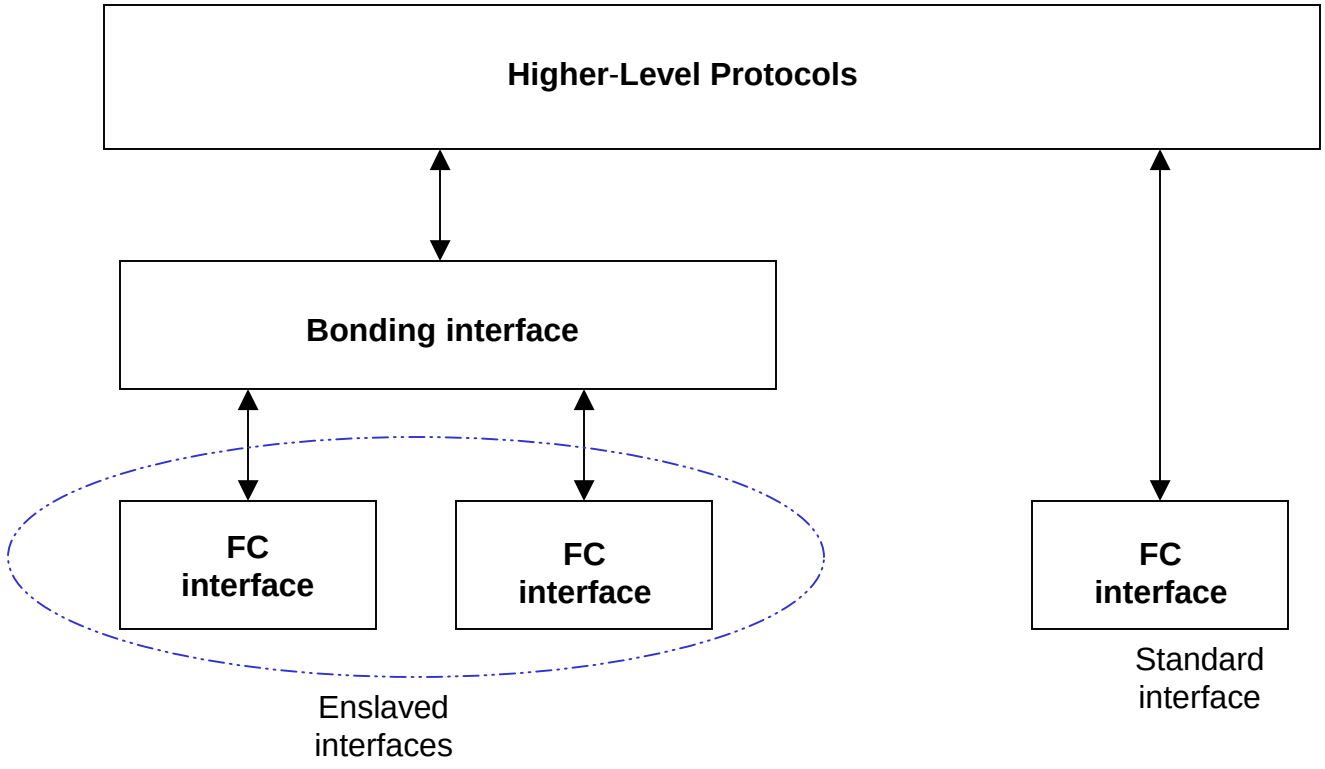
    # pm –v conversion
  - **Save fstab (fstab.dks), modify new fstab to include pmd device names**
  - **Modify NVRAM file on CWS to refer to pmd device name for root**
  - **Reboot UNICOS/mp**

UNICOS/mp 2.4 Update

- **Bonded Fibre Channel Driver**
  - **Channel bonding – multiple Fibre Channel interfaces configured as single logical interface**
  - **Resiliency feature (ride through link failure between Cray X1 and CNS)**
  - **Requires two Fibre Channel connections between Cray X1 and CNS**
  - **Available with UNICOS/mp 2.4, CNS 1.2**

UNICOS/mp 2.4 Update

- **Bonded Fibre Channel Driver**



UNICOS/mp 2.4 Update

- **Bonding interfaces: bfc0, bfc1, …**
- **Managed via new bfc(8) command**
- **Slave interface selection, 2 algorithms:**
  - **Active backup**

    single slave interface (first one configured) is used for all traffic; if first interface fails, traffic is routed to next slave interface
  - **Round-robin ** not supported in UNICOS/mp**

    packets are routed round-robin across available interfaces

UNICOS/mp 2.4 Update

- **bfc(8) command**
  - **Create a bonding interface**
  - **Remove a bonding interface**
  - **Assign FC interfaces to bonding interface (enslavement)**
  - **Remove FC interfaces from bonding interface (emancipation)**
  - **Select algorithm used for distributing outbound packets**

UNICOS/mp 2.4 Update

- **Bonded FC configuration on UNICOS/mp**
  - **Example startup script /etc/init.d/bond.local**
    - **Identify slaves for each bonding interface**
    - **Startup commands, e.g.:**
      - **bfc attach 0**
      - **bfc enslave bfc0 qfa0 qfa2**
    - **Shutdown commands:**
      - **bfc free bfc0 all**
      - **bfc detach bfc0**
  - **Modify /etc/config/netif.options**
    - **Include new bonding interface:**

      if<n>name=bfc0

      If<n>addr=<name or IP address>
    - **Remove slave FC interfaces (qfa0, qfa2)**

CUG 2004

UNICOS/mp 2.4 Update

- **Bonded FC configuration on UNICOS/mp**
  - **Link startup/shutdown script:**
    - **# ln –s /etc/init.d/bond.local /etc/rc2.d/S29network**
    - **# ln –s /etc/init.d/bond.local /etc/rc2.d/K41network**
      - ***** startup must precede S30network,***
      - ***shutdown must follow K40network***
  - **Create /etc/config/ifconfig-<n>.options**
    - **e.g. 'netmask 0xfffffffc**

UNICOS/mp 2.4 Update

- **Bonded FC configuration on CNS**
  - **Run cns_gen_config(8) command to configure**
    - **Answer 'Y' to "Configure interface bond0?"**
      - **Specify IP address and netmask**
      - **lpfn0 and lpfn1 configured as slaves to bond0**
  - **Run cns_config(8) to load new configuration**
  - **Reboot CNS to activate bond0 interface**

UNICOS/mp 2.4 Update

- **Bonded FC example**

```
x1# netstat -i
Name Mtu   Network          Address          Ipkts  Ierrs   Opkts Oerrs  Coll
bfc0 65280  192.168.192.76  mfeg6-sv2-fc-c  100750      0  102569      0     0
qfa0 65280  none            none            100778      0  102596      0     0
qfa1* 65280 192.168.192.44  mfeg6-sv2-fc-b       0      0       0      0     0
qfa2* 65280 none            none                 0      0       0      0     0
lo0  65535    loopback      localhost          529      0     529      0     0

x1# bfc show bfc0
bfc0:   mode backup (0)
      2 slaves:
         Intf  State    Ipkts  Ierrs   Opkts  Oerrs
         qfa0  Up     103327      0  105190      0
         qfa2  Down        0      0       0      0
         bfc0  Up     103327      0  105190      0
```

UNICOS/mp 2.4 Update

- **Bonded Fibre Channel documentation:**
  - **UNICOS/mp Networking Facilities Administration, SR-2341-24**
  - **Cray Network Subsystem (CNS) Software Installation and Administration, S-2366-13**
  - **FN5211a – “CNS-2 automatic failover and Fibre Channel IP bonding”**

UNICOS/mp 2.4 Update

- **StorNext File System (SNFS) Client**
  - **ADIC Storage Area Network (SAN) file system**
  - **Requires pmd driver**
  - **Currently supported client version: 2.2.1**
  - **Utilizes Fibre Channel fabric support**
  - **Client software built into UNICOS/mp kernel**
  - **StorNext commands, man pages included with UNICOS/mp OS release**

UNICOS/mp 2.4 Update

- **Fibre Channel fabric support**
  - **Connections to fibre channel switch defined in cray.cfg file as Disk xyy (convention: x = switch ordinal #, yy = port # on switch)**
  - **csm used to setup RAID configuration, choose a site unique 'chassis #' for switch accessed by multiple Cray X1 systems**
  - **At OS boot time, fibre channel controllers probed to find all LUNs on each path; 4 paths configured per LUN (# ports on array)**

UNICOS/mp 2.4 Update

- **Fabric-attached device naming conventions**
  - **pmd driver → underlying path identifiers**

fc<SPP>d<T>L<U>s

SPP → host controller # (Disk id from cray.cfg)

S → switch ordinal #

PP → port # on switch

T → disk target #, 129 + switch port #

U → LUN #

Example: fc149d134L5

switch 1, port 49 ***target 134 (port 5)*** ***LUN 5***

UNICOS/mp 2.4 Update

- **fabric-attached storage pm(8) excerpt:**

pm80d1L5

path	port	state	read blks	write blks	errs	MB/Sec
fc149d134L5	pri	active	10192896	26398720	0	0.00
fc149d136L5	pri	active	10192896	26398720	0	0.00
fc148d133L5	alt	standby	0	0	0	0.00
fc148d135L5	alt	standby	0	0	0	0.00

pm80d4L4

path	port	state	read blks	write blks	errs	MB/Sec
fc148d141L4	pri	active	10188800	26399232	0	0.00
fc148d143L4	pri	active	10188800	26399232	0	0.00
fc149d142L4	alt	standby	0	0	0	0.00
fc149d144L4	alt	standby	0	0	0	0.00

UNICOS/mp 2.4 Update

- **SNFS on UNICOS/mp**
  - **Disk devices labeled for SNFS use on Metadata server (MDS)**
  - **File systems configured/created on MDS**
  - **Disk (LUN) configuration on CWS using csm utilities**
  - **Network connection to StorNext private network (for metadata traffic)**
  - **StorNext enabled on UNICOS/mp via chkconfig (‘/sbin/chkconfig cvfs on’)**

UNICOS/mp 2.4 Update

- **SNFS, Cray-X1 client configuration**
  - **/usr/cvfs/config/fsnameservers (contains private network info. for MDS)**
  - **cvlabel(1M) to view SAN disk devices**
  - **/etc/init.d/cvfs script to start/stop SNFS**
  - **‘mount –t cvfs …’ to mount SNFS file systems**
  - **cvadmin(1M) to view active SNFS configuration**

UNICOS/mp 2.4 Update

Peggy Gazzola

Cray, Inc. – Software Product Support

651-605-8966, peggy@cray.com

CUG 2004