FreeBSD for High Density Servers

Intel Avoton based 5,000+ cores
NEC Micro Modular Server DX1000

13 Jun, 2015

Daichi GOTO daichi@freebsd.org
at BSD Consulting, Inc. (Tokyo)
Profile

- Daichi GOTO, born in 1980
- ONGS Inc. CEO since 2002 / BSD Consulting, Inc. (Tokyo) CTO since 2012
- FreeBSD committer since 2002 daichi@ / FreeBSD Journal board member since 2013
- Skills: Design and develop of enterprise system, IT-news-writing, book-and-article-writing, etc
- https://jp.linkedin.com/in/daichigoto
Community work - FreeBSD 勉強会
Intention
OPERATING SYSTEM FAMILY / LINUX

http://www.top500.org/statistics/details/osfam/1
TOP500 / Unix

OPERATING SYSTEM FAMILY / UNIX

http://www.top500.org/statistics/details/osfam/3
TOP500 / BSD based

OPERATING SYSTEM FAMILY / BSD BASED

http://www.top500.org/statistics/details/osfam/5
TOP500 / Mac OS X

OPERATING SYSTEM FAMILY / MAC OS

http://www.top500.org/statistics/details/osfam/8
TOP500 / FreeBSD

http://www.top500.org/statistics/details/osfam/1
some reasons of the defeat

- FreeBSD has been mostly running not on HPC systems but on consumer PCs or low-price rack mount servers in the first stage
- Linux vendors (Red Hat, SUSE…) supported Linux to run on HPC systems
- Hardware vendors released Linux version device drivers rather than FreeBSD
- In fact, from the start, FreeBSD was in the wrong circumstances for HPC
Information sharing

• I have been verifying any FreeBSD behaviors with latest NEC’s new rack mount servers continuously.

• NEC Micro Modular Server DX1000 is one of those machines. It is extreme density, exceptional energy efficient and outstanding manageability rack mount server.

• And it is too expensive for someone to buy. So information sharing about FreeBSD is worth while.
Product
NEC Micro Modular Server DX1000

http://jpn.nec.com/slpf/product/cpdc/
DX1000

- up to 46 single-processor server modules in a 2U enclosures
- Atom C2000 series 8-core processor, 4 DIMM slots, 1 SSD slot
- Operation in a 40 degree Celsius environment which minimizes cooling cost
- 80 PLUS platinum certified power supply
- All modules are hot-swappable and easy to replace
DX1000 Modules

- DX1000 is consisted by 5 types of modules (6 if including Power Unit)
- Network Switch Module
- CMM Module
- Server Module
- HDD Module
- Fan Module
DX1000

- Spec: 46 CPU Modules per a enclosure
- Real: 38 Server Modules per a enclosure because of its power unit limitation
- $38 \times 16 = 608$ Server Modules per a rack
- $10 \times 608 = 6,080$ bhyve hosts per a rack
- $100 \times 608 = 60,800$ jail hosts per a rack
Network Switch Module
CMM Module
Server Module
HDD Module
Fan Module
Power Unit
Front panel
Module relationship diagram
<table>
<thead>
<tr>
<th><strong>Server Module</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Form factor</td>
</tr>
<tr>
<td>Number of Processors</td>
</tr>
<tr>
<td>Processors</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Memory Type</td>
</tr>
<tr>
<td>Memory Slots</td>
</tr>
<tr>
<td>Maximum memory</td>
</tr>
<tr>
<td>Storage type</td>
</tr>
<tr>
<td>Maximum internal storage</td>
</tr>
<tr>
<td>Expansion slots</td>
</tr>
<tr>
<td>Network</td>
</tr>
<tr>
<td>Systems management</td>
</tr>
<tr>
<td>Operating systems and virtualization software</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>HDD Module</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Form factor</td>
</tr>
<tr>
<td>Number of HDDs</td>
</tr>
<tr>
<td>Storage type</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Module Enclosure</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Form factor / height</td>
</tr>
<tr>
<td>Server module slots</td>
</tr>
<tr>
<td>Network interconnects</td>
</tr>
<tr>
<td>Redundant cooling fan</td>
</tr>
<tr>
<td>Power supplies</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Redundant power supply</td>
</tr>
<tr>
<td>Temperature and humidity conditions (non-condensing)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Dimensions (W x D x H) and maximum weight</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Mount to a rack

- Operate follow a user guide document
- However the DX1000 is lighter than any other similar products, it’s heavy. Please do careful.
Installation
Network Switch Module configuration

- Login into 1st Network Switch Module through the serial console port on left side on front panel

- baut rate: 115,200 bps

- ID: admin

- PASSWD: admin
Left side port
# cu -l /dev/cuaU0 -s 115200
Connected

Wind River Linux 5.0.1.10
console

localhost login: admin
Password:
Last login: Tue Jan 20 20:28:53 UTC 2015 from 192.168.36.10 on pts/0

Connecting ... 
Checking ONS status... It may take few minutes...
Checking application states ...
Checking table states ...
Checking platform information ...
Getting user information ...
Authenticating ...

Switch >
```
Switch >enable
Switch #show system

<table>
<thead>
<tr>
<th>System Name</th>
<th>ONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Description</td>
<td>Open Network Software</td>
</tr>
<tr>
<td>Ethernet Switch Type</td>
<td>Fulcrum Switch</td>
</tr>
<tr>
<td>Name</td>
<td>ONS CoreSwitch</td>
</tr>
<tr>
<td>Model</td>
<td>ONS</td>
</tr>
<tr>
<td>Platform</td>
<td>Mercury</td>
</tr>
<tr>
<td>Chip Version</td>
<td>Board:01</td>
</tr>
<tr>
<td>Chip Subtype</td>
<td>fm6000</td>
</tr>
<tr>
<td>API Version</td>
<td>FocalPoint 3.3.5_00268148 + mercury-20131213</td>
</tr>
<tr>
<td>Software Version</td>
<td>1.2.0.1425-2</td>
</tr>
<tr>
<td>CPU</td>
<td>x86_64</td>
</tr>
<tr>
<td>CPU Architecture</td>
<td>x86_64</td>
</tr>
<tr>
<td>OS</td>
<td>Linux</td>
</tr>
<tr>
<td>OS Version</td>
<td>3.4.43-WR5.0.1.10_standard</td>
</tr>
<tr>
<td>Serial Number</td>
<td>N/A</td>
</tr>
<tr>
<td>IP Address</td>
<td>10.1.1.1</td>
</tr>
<tr>
<td>Mask</td>
<td>255.255.255.0</td>
</tr>
<tr>
<td>Gateway</td>
<td>N/A</td>
</tr>
<tr>
<td>MAC Address</td>
<td>74:D4:35:E9:E2:62</td>
</tr>
<tr>
<td>Default VLAN</td>
<td>1</td>
</tr>
<tr>
<td>Current Partition</td>
<td>/dev/sda3</td>
</tr>
</tbody>
</table>
```

Switch #
calc MAC address

- 74:D4:35:E9:E2:61 MAC of ONS
- 74:D4:35:E9:E2:60 MAC of CMM Module
$ arp -a

dullmdaler.ongs.co.jp (202.216.246.94) at 00:0d:b9:2c:6c:62 on vr2 permanent [ethernet]
zenosblead.ongs.co.jp (202.216.246.89) at 00:0b:a2:8c:84:de on vr2 expires in 1180 seconds [ethernet]
natial.ongs.co.jp (202.216.246.90) at 00:0d:b9:2b:d2:38 on vr2 expires in 1072 seconds [ethernet]
hepitas.ongs.net (202.216.246.91) at 00:0d:b9:32:9c:7c on vr2 expires in 1085 seconds [ethernet]
? (192.168.1.40) at 78:31:c1:d5:6e:fc on vr0 expires in 1178 seconds [ethernet]
? (192.168.1.106) at 64:4b:f0:00:13:4c on vr0 expires in 963 seconds [ethernet]
? (192.168.1.10) at 00:0d:0b:80:3e:18 on vr0 expires in 1182 seconds [ethernet]
? (192.168.1.1) at 00:0d:b9:2c:6c:60 on vr0 permanent [ethernet]
? (192.168.1.29) at 74:D4:35:E9:E2:60 on vr0 expires in 1198 seconds [ethernet]
? (192.168.1.34) at 04:db:56:0d:cb:7c on vr0 expires in 1170 seconds [ethernet]
? (192.168.1.101) at e0:69:95:f5:42:84 on vr0 expires in 748 seconds [ethernet]

$ ping 192.168.1.29
PING 192.168.1.29 (192.168.1.29): 56 data bytes
64 bytes from 192.168.1.29: icmp_seq=0 ttl=64 time=0.698 ms
64 bytes from 192.168.1.29: icmp_seq=1 ttl=64 time=0.648 ms
64 bytes from 192.168.1.29: icmp_seq=2 ttl=64 time=0.548 ms
^C

--- 192.168.1.29 ping statistics ---
3 packets transmitted, 3 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 0.548/0.631/0.698/0.062 ms

$
$ ./mng_niclist -I 192.168.1.29 -C all

<table>
<thead>
<tr>
<th>CPU Board</th>
<th>ManagementLAN MAC IP</th>
<th>DataLAN-1 MAC</th>
<th>DataLAN-2 MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Board23</td>
<td>74:D4:35:83:78:16 0.0.0.0</td>
<td>74:D4:35:83:78:14 74:D4:35:83:78:15</td>
<td></td>
</tr>
<tr>
<td>CPU Board24</td>
<td>74:D4:35:83:75:60 0.0.0.0</td>
<td>74:D4:35:83:75:5E 74:D4:35:83:75:5F</td>
<td></td>
</tr>
<tr>
<td>CPU Board27</td>
<td>74:D4:35:83:75:75 0.0.0.0</td>
<td>74:D4:35:83:75:73 74:D4:35:83:75:74</td>
<td></td>
</tr>
<tr>
<td>CPU Board29</td>
<td>74:D4:35:83:75:3C 0.0.0.0</td>
<td>74:D4:35:83:75:3A 74:D4:35:83:75:3B</td>
<td></td>
</tr>
<tr>
<td>CPU Board30</td>
<td>74:D4:35:83:79:45 0.0.0.0</td>
<td>74:D4:35:83:79:43 74:D4:35:83:79:44</td>
<td></td>
</tr>
<tr>
<td>CPU Board31</td>
<td>74:D4:35:83:78:52 0.0.0.0</td>
<td>74:D4:35:83:78:50 74:D4:35:83:78:51</td>
<td></td>
</tr>
<tr>
<td>CPU Board32</td>
<td>74:D4:35:83:78:55 0.0.0.0</td>
<td>74:D4:35:83:78:53 74:D4:35:83:78:54</td>
<td></td>
</tr>
</tbody>
</table>

$
calc MAC address

- 74:D4:35:83:78:16 obtained by command
- 74:D4:35:83:78:5F MAC of NIC2
- 74:D4:35:83:78:5E MAC of NIC1
$ cat /usr/local/etc/dhcpd.conf
option domain-name "bsdconsulting.co.jp";
option domain-name-servers 192.168.1.1;
default-lease-time 600;
max-lease-time 7200;

# DHCP settings
subnet 192.168.1.0 netmask 255.255.255.0 {
  range 192.168.1.20 192.168.1.60;
  option routers 192.168.1.1;
  option broadcast-address 192.168.1.255;
  option subnet-mask 255.255.255.0 ;
}

# PXE boot settings
# kernel looking path (look at /etc/inetd.conf too)
next-server 192.168.1.1;
filename "pxeboot";
# nfs server path
option root-path "192.168.1.10:/home/pxefreebsd";
}

host dx1000mmc1 { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.80; }
host dx1000s23bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.81; }
host dx1000s24bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.82; }
host dx1000s25bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.83; }
host dx1000s26bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.84; }
host dx1000s27bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.85; }
host dx1000s28bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.86; }
host dx1000s29bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.87; }
host dx1000s30bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.88; }
host dx1000s31bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.89; }
host dx1000s32bmc { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.90; }
host dx1000s23nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.91; }
host dx1000s24nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.92; }
host dx1000s25nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.93; }
host dx1000s26nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.94; }
host dx1000s27nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.95; }
host dx1000s28nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.96; }
host dx1000s29nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.97; }
host dx1000s30nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.98; }
host dx1000s31nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.99; }
host dx1000s32nic { hardware ethernet 74:D4:35:E9:E2:60; fixed-address 192.168.1.100; }
$
PXE Boot

- 1) Host obtains an IP address and tftp information from DHCP server
- 2) loading pxeboot kernel from tftp server
- 3) loading installer from NFS server
- write /etc/rc.conf, /etc/inetd.conf, /etc/exports, /usr/local/etc/dhcpd.conf for DHCPd, tftp and NFS
DHCP configuration

$ pkg install isc-dhcp43-server

/etc/rc.conf
  dhcpd_enable="YES"

/usr/local/etc/dhcpd.conf
  look at 2 pages back
tftp configuration

/etc/rc.conf
inetd_enable="YES"

/etc/inetd.conf
tftp dgram udp wait root /usr/libexec/tftpd \
tftpd -l -s /tftpboot/amd64/10.1

$ cp /boot/pxeboot /tftpboot/amd64/10.1/
$ ls -l /tftpboot/amd64/10.1/
total 260
-r--r--r-- 1 daichi daichi 264192 Mar 9 14:49 pxeboot
$
NFS configuration

/etc/rc.conf
rpcbind_enable="YES"
rpc_statd_enable="YES"
rpc_lockd_enable="YES"
nfs_server_enable="YES"
nfs_server_flags="-u -t -n 4"
nfsd_enable="YES"
mountd_enable="YES"

/etc/exports
V4: / -sec=sys -network 192.168.1.0 -mask 255.255.255.0
/home/pxefreebsd -ro -alldirs -maproot=root
$ mdconfig -a -t vnode
FreeBSD-10.1-RELEASE-amd64-disc1.iso
md0
$ mount_cd9660 /dev/md0 /mnt
$ tar zvxf /mnt/usr/freebsd-dist/base.txz -C /home/pxefreebsd/
...
$ ls /home/pxefreebsd/
COPYRIGHT    README.TXT    docbook.css    proc    usr
ERRATA.HTM    RELNOTES.HTM    etc    rescue    var
ERRATA.TXT    RELNOTES.TXT    lib    root
HARDWARE.HTM    bin    libexec    sbin
HARDWARE.TXT    boot    media    sys
README.HTM    dev    mnt    tmp
$ vi /home/pxefreebsd/boot/loader.conf /etc/ttys
$ cat /home/pxefreebsd/boot/loader.conf
boot_serial="YES"
console="comconsole"
comconsole_port="1000"
comconsole_speed="115200"
$ grep ^ttyu2 /home/pxefreebsd/etc/ttys
ttyu2  "/usr/libexec/getty std.115200" xterm  on secure
Installation
Power on the Server Module

$ ipmitool -I lanplus \
-U Administrator -P Administrator \
-H 192.168.1.81 power on

$ ipmitool -I lanplus \
-U Administrator -P Administrator \
-H 192.168.1.81 sol activate
IPMI Base Board Management Controller was detected.

Device ID : 20  
Device Revision : 01  
IPMI Version : 2.0  
Firmware Revision : 01.21  
Self Test Result : 5500  

Press <F2> SETUP
BIOS Information
BIOS Version 5.6.0007
Build Date 08/20/2014

Memory Information
Total Memory 32768 MB (DDR3)

System Language [English]
System Date [Tue 03/10/2015]
System Time [14:02:00]
Access Level Administrator

Choose the system default language
> <: Select Screen
^v: Select Item
Enter: Select
+/-= Change Opt.
F1: General Help
F2: Previous Values
F3: Optimized Defaults
F4: Save & Exit
ESC: Exit

Version 2.16.1242n Copyright (C) 2013 American Megatrends, Inc.
Boot Configuration

Boot Option #1 [P0: TOSHIBA THNSNH1...]
Boot Option #2 [IBA GE Slot 00A0 v1543]

Network Device BBS Priorities
Hard Drive BBS Priorities

Sets the system boot order

> <: Select Screen
^v: Select Item
Enter: Select
+/ -: Change Opt.
F1: General Help
F2: Previous Values
F3: Optimized Defaults
F4: Save & Exit
ESC: Exit
Welcome to FreeBSD

1. Boot Multi User [Enter]
2. Boot Single User
3. Escape to loader prompt
4. Reboot

Options:
5. Kernel: default/kernel (1 of 2)
6. Configure Boot Options...

/boot/kernel/kernel text=0xf8f898 data=0x124a30+0x2055c0 syms=[0x8+0x1405e0+0x8+0x15b077-]
add net ::ffff:0.0.0.0: gateway ::1
add net ::0.0.0.0: gateway ::1
Generating host.conf.
eval: cannot create /etc/host.conf: Read-only file system
eval: cannot create /etc/host.conf: Read-only file system
eval: cannot create /etc/host.conf: Read-only file system
/libexec/resolvconf/libc: cannot create /etc/resolv.conf: No such file or directory
Creating and/or trimming log files.
Starting syslogd.
ELF ldconfig path: /lib /usr/lib /usr/lib/compat
32-bit compatibility ldconfig path: /usr/lib32
Clearing /tmp (X related).
Starting local daemons:
Welcome to FreeBSD!

Please choose the appropriate terminal type for your system.
Common console types are:

ansi Standard ANSI terminal
vt100 VT100 or compatible terminal
xterm xterm terminal emulator (or compatible)
cons25w cons25w terminal

Console type [vt100]: _
\
Welcome to FreeBSD! Would you like to begin an installation or use the live CD?
Conclusion
Conclusion

• NEC Micro Modular Server DX1000 is affordable as Hadoop clusters or to bundle many physical servers into a rack

• Installation into DX1000 Server Modules is a little bit confusion, but no problem

• FreeBSD can run on DX1000 Server Modules
Appendix a.
change QSFP to RJ45
```
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x17 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x18 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x19 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x1a 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x1b 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x1c 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x1d 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x1e 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x1f 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
$ ipmitool -I lanplus -U Administrator -P Administrator -H 192.168.1.80 raw \
 0x30 0x06 0x40 0x20 0x00 0x20 0x30 0xb0 0x26 0x00 0x01 0x01 0x14 0x00 0x00 0xc4
```

$ cu -s 115200 -l /dev/cuaU0
Connected

Wind River Linux 5.0.1.10
    console

localhost login: admin
Password:
Last login: Mon Mar 9 18:14:21 UTC 2015 on console

Connecting ...    
Checking ONS status... It may take few minutes...
Checking application states ... 
Checking table states ... 
Checking platform information ... 
Getting user information ... 
Authenticating ... 

Switch >
Switch >enable
Switch #configure
Switch (config)#interface range xe47,xe50
Switch (config-if-range)#switchport pvid 1
Switch (config-if-range)#no switchport vlan add 4092
Switch (config-if-range)#switchport vlan add 1 untagged
Switch (config-if-range)#exit
Switch (config)#interface range xe1-xe46
Switch (config-if-range)#shutdown
Switch (config-if-range)#no switchport vlan add 4092
Switch (config-if-range)#no shutdown
Switch (config-if-range)#exit
Switch (config)#no spanning-tree
Switch (config)#exit
Switch #save config

Configuration saving is in progress. It may take few minutes.

Switch #exit
Appendix b.
contact list
interested in purchasing

- North America: NEC Corporation of America
  http://www.necam.com/servers
- Europe: NEC Enterprise Solutions
  http://www.nec-enterprise.com/
- APAC: NEC Corporation
  http://www.nec.com/express