Sonexion 2000 Power On-Off Procedures

Contents

About *Sonexion 2000 Power On-Off Procedures*

Sonexion 2000 Power On-Off Procedures provides step-by-step instructions for powering on or off a Sonexion 2000 system running the Sonexion 1.5 or 2.0 operating system.

Scope and Audience

The procedures presented in this manual are to be carried about by site administrators and technicians where Sonexion systems are installed, employed by either Cray Inc., or the customer organization.

Typographic Conventions

`Monospace`	A `Monospace` font indicates program code, reserved words or library functions, screen output, file names, path names, and other software constructs
`Monospaced Bold`	A **`bold monospace`** font indicates commands that must be entered on a command line.
`Oblique` or *`Italics`*	An *`oblique`* or *`italics`* font indicates user-supplied values for options in the syntax definitions
Proporational Bold	A **proportional bold** font indicates a user interface control, window name, or graphical user interface button or control.
`Alt-Ctrl-f`	`Monospaced` hypenated text typically indicates a keyboard combination

Record of Revision

Publication Number	Date	Description
HR5-6130-0	December 2014	Original Printing, release 1.5
HR5-6130-A	April 2015	Release 2.0

Power On Sonexion 2000

Prerequisites

- **System access requirements**:

  Root user access is required to perform this procedure on a Sonexion system. If you do not have root user access, contact Cray Support.

- **Service Interruption Level**:

  This procedure requires taking the Lustre file system offline.

- **Required Tools and Equipment**:
  - ESD strap, boots, garment or other approved protection
  - Console with monitor and keyboard (or PC with a serial port configured for 115.2 Kbps, 8 data bits, no parity and one stop bit)

About this task

Use this procedure to power on a Sonexion 2000 system using the CSCLI interface.

The MMU component can use either a 2U24 EBOD or 5U84 EBOD, which hosts the MDT (storage target for the MGMT, MGS, and MDS nodes). This document uses the terms *2U24 EBOD* and *5U84 EBOD* to refer to the two types of storage enclosures. The term *2U Quad Server* refers to the MMU component. The term "5U84" refers to the SSU and ESU components.

Procedure

1. At the back of the rack, confirm that power is off to the 5U84s. Each 5U84 has two power supply units (PSUs) with power switches, located below the fan modules.

2. Verify that the power switches on MMU 2U24 EBOD enclosure are in the ON position. The 2U24 EBOD has two power cooling modules (PCMs) with power switches, located next to the EBOD I/O modules.

3. Confirm that the rack PDU power ON/OFF switches are in the OFF position.

4. Power on the Sonexion 2000 rack, noting the following:
   - If necessary, plug the PDU cords into power receptacles.
   - If using a Raritan PDU, place each PDU power ON/OFF switch in the ON position.

     Raritan PDUs have three lines, with one power switch for each line. Each line must be powered on. Some Raritan PDUs also have two line inputs and a new 60A PDU will only have one line input.
   - ServerTech PDUs do not have power ON/OFF switches.

5. If your cluster contains Expansion Storage Units (ESUs), power on the ESUs (5U84) as follows:

   a. Place the power switches in the ON position. In an ESU, the power switches are on the PSU.

b. Wait 30 seconds for the drives to spin up.

6. Verify that the 2U24 (or 5U84) EBOD is powered and LEDs indicate that the drives are spinning.

Make certain the primary and secondary MGMT nodes (logical nodes 00 and 01) are connected to the public network, as shown in the following figures.

Figure 1. Cabling the MGMT Nodes (eth1)

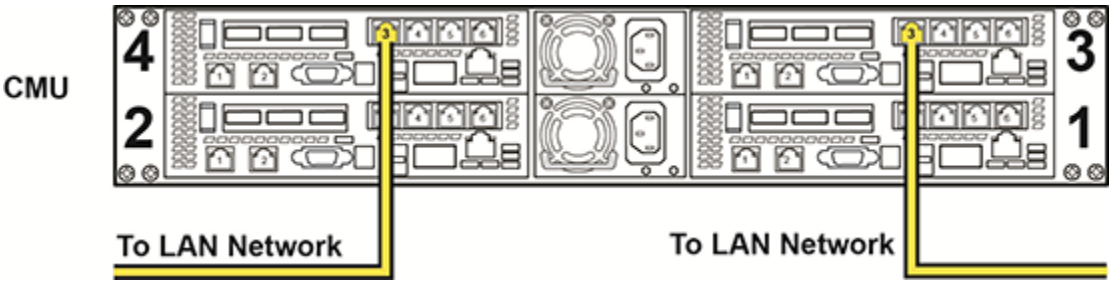
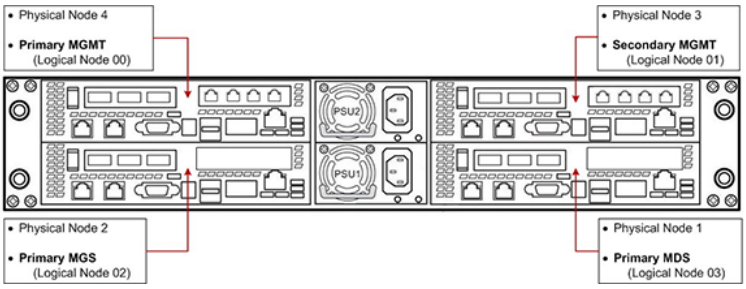


Figure 2. Logical and Physical nodes



7. Both management servers are normally configured to automatically power on. If this is not the case for you system, power on manually using the front panel switch. Verify the servers are on via the power indicator LED.

8. Optionally you may connect a keyboard and monitor to the primary MGMT node and log into the console using the admin account credentials. Otherwise log in over the public LAN. The systems may take 10 minutes to become available.

9. Check that the shared storage targets are available for the management nodes:

```
[admin@n000]$ pdsh -g mgmt cat /proc/mdstat
```

For example:

```
[admin@snx11000n000 ~]$ pdsh -g mgmt cat /proc/mdstat | dshbak -c
----------------
snx11000n000
----------------
Personalities : [raid1] [raid6] [raid5] [raid4] [raid10]
md64 : active raid10 sda[0] sdc[3] sdw[2] sdl[1]
      1152343680 blocks super 1.2 64K chunks 2 near-copies [4/4] [UUUU]
      bitmap: 2/9 pages [8KB], 65536KB chunk
md127 : active raid1 sdy[0] sdz[1]
      439548848 blocks super 1.0 [2/2] [UU]
unused devices: <none>
----------------
```

```
snx11000n001
---------------
Personalities : [raid1] [raid6] [raid5] [raid4] [raid10]
md67 : active raid1 sdi[0] sdt[1]
      576171875 blocks super 1.2 [2/2] [UU]
      bitmap: 0/5 pages [0KB], 65536KB chunk
md127 : active raid1 sdy[0] sdz[1]
      439548848 blocks super 1.0 [2/2] [UU]
unused devices: <none>
```

10. Check HA status once the node is completely up and HA configuration has been established:

```
[admin@n000]$ sudo crm_mon -1r
```

For example:

```
[admin@snx11000n000 ~]$ sudo crm_mon -1r
Last updated: Thu Aug 7 01:30:36 2014
Last change: Wed Aug 6 23:58:18 2014 via crm_resource on snx11000n001
Stack: Heartbeat
Current DC: snx11000n001 (0828104e-8d91-44ad-892a-13dbd1fd7c6c) - partition
with quorum
Version: 1.1.6.1-6.el6-0c7312c689715e096b716419e2ebc12b57962052
2 Nodes configured, unknown expected votes
53 Resources configured.
============
Online: [ snx11000n000 snx11000n001 ]
Full list of resources:
snx11000n000-1-ipmi-stonith (stonith:external/ipmi): Started snx11000n000
snx11000n001-1-ipmi-stonith (stonith:external/ipmi): Started snx11000n001
snx11000n000-2-ipmi-stonith (stonith:external/ipmi): Started snx11000n000
snx11000n001-2-ipmi-stonith (stonith:external/ipmi): Started snx11000n001
Clone Set: cln-diskmonitor [diskmonitor]
Started: [ snx11000n000 snx11000n001 ]
Clone Set: cln-last-stonith [last-stonith]
Started: [ snx11000n000 snx11000n001 ]
Clone Set: cln-kdump-stonith [kdump-stonith]
Started: [ snx11000n000 snx11000n001 ]
prm-httpd (lsb:httpd): Started snx11000n000
prm-mysql (lsb:mysqld): Started snx11000n000
prm-nfslock (lsb:nfslock): Started snx11000n001
prm-bebundd (lsb:bebundd): Started snx11000n000
Clone Set: cln-cerebrod [prm-cerebrod]
Started: [ snx11000n001 snx11000n000 ]
prm-conman (lsb:conman): Started snx11000n000
prm-dhcpd (lsb:dhcpd): Started snx11000n001
Clone Set: cln-syslogng [prm-syslogng]
Started: [ snx11000n001 snx11000n000 ]
Clone Set: cln-dnsmasq [prm-dnsmasq]
Started: [ snx11000n001 snx11000n000 ]
prm-nodes-monitor (lsb:nodes-monitor): Started snx11000n000
Clone Set: cln-ses_mon [prm-ses_monitor]
Started: [ snx11000n001 snx11000n000 ]
Clone Set: cln-nsca_passive_checks [prm-nsca_passive_checks]
Started: [ snx11000n001 snx11000n000 ]
Resource Group: grp-icinga
prm-icinga (lsb:icinga): Started snx11000n000
prm-nsca (lsb:nsca): Started snx11000n000
prm-npcd (lsb:npcd): Started snx11000n000
```

```
prm-repo-local (ocf::heartbeat:Filesystem): Started snx11000n001
prm-repo-remote (ocf::heartbeat:Filesystem): Started snx11000n000
prm-db2puppet (ocf::heartbeat:oneshot): Started snx11000n000
Clone Set: cln-puppet [prm-puppet]
Started: [ snx11000n001 snx11000n000 ]
prm-nfsd (ocf::heartbeat:nfsserver): Started snx11000n001
prm-vip-eth0-mgmt (ocf::heartbeat:IPaddr2): Started snx11000n000
prm-vip-eth0-ipmi-sec (ocf::heartbeat:IPaddr2): Started snx11000n000
prm-vip-eth0-nfs (ocf::heartbeat:IPaddr2): Started snx11000n001
Resource Group: snx11000n000_md64-group
snx11000n000_md64-raid (ocf::heartbeat:XYRAID): Started snx11000n000
snx11000n000_md64-fsys (ocf::heartbeat:XYMNTR): Started snx11000n000
snx11000n000_md64-stop (ocf::heartbeat:XYSTOP): Started snx11000n000
Resource Group: snx11000n000_md67-group
snx11000n000_md67-raid (ocf::heartbeat:XYRAID): Started snx11000n001
snx11000n000_md67-fsys (ocf::heartbeat:XYMNTR): Started snx11000n001
snx11000n000_md67-stop (ocf::heartbeat:XYSTOP): Started snx11000n001
prm-ctdb (lsb:ctdb): Started snx11000n001
Resource Group: grp-plex
prm-rabbitmq (lsb:rabbitmq-server): Started snx11000n000
prm-plex (lsb:plex): Started snx11000n000
Resource Group: grp-dcs
prm-zabbix-server (lsb:zabbix-server): Started snx11000n000
prm-zabbix-listener (lsb:agent-listener): Started snx11000n000
baton (ocf::heartbeat:baton): Started snx11000n001
snx11000n000_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n000
snx11000n001_mdadm_conf_regenerate (ocf::heartbeat:mdadm_conf_regenerate):
Started snx11000n001
```

Correct output indicates that all resources have started and are balanced between two nodes. If not, use one of the following, and re-check `crm_mon` output after 5 minutes.

In cases when all resources started on a single node (for example, all resources have started on node 00 and did not failback to node 01), you may need to execute the failback operation:

```
[admin@n000]$ cscli failback –n secondary_MGMT_node
```

Or the opposite may be required. For example, if the secondary management node has assumed all of the resources at boot time, run the following. (This is likely the case if you see the message `cscli: Please, run cscli on active management node`.)

```
[admin@n001]$ cscli failback –n primary_MGMT_node
```

11. Power on the MGS and MDS nodes:

```
[admin@n000]$ cscli power_manage -n mgs_node,mds_node --power-on
```

Example:

```
[root@snx11000n000 ~]$ cscli power_manage -n snx11000n[002-003] --power-on
power_manage: processing snx11000n002 ...
power_manage: processing snx11000n003 ...
power_manage: Operation performed successfully.
```

12. Physically power on the 5U84 SSUs. On each SSU, place the power switch in the ON position.

After switching on an SSU you must wait at least 50 seconds before performing Step 13.

13. If the system has Additional DNE Unit (ADU) nodes, physically power them on. On each ADU, place the power switches in the “ON” position.

    The ADU is a 2U24; therefore, there are two PSUs to power on.

14. Power on the OSS nodes and, if present, the ADU nodes:

    ```
    [root@n000]# cscli power_manage -n oss_adu_nodes --power-on
    ```

    For example:

    ```
    [root@snx11000n000 ~]# cscli power_manage -n snx11000n[004-007] --power-on
    power_manage: processing snx11000n004 ...
    power_manage: processing snx11000n005 ...
    power_manage: processing snx11000n006 ...
    power_manage: processing snx11000n007 ...
    power_manage: Operation performed successfully.
    ```

    Cray advises to have no more than 60 nodes powering on at the same time. If there are more than 60 nodes, repeat this step as necessary (powering on 60 nodes each time) until all nodes are powered on.

15. Check the status of the nodes:

    ```
    [admin@n000]$ pdsh -a date
    ```

    The correct output includes the date for each host in the cluster. For example:

    ```
    [admin@snx11000n000 ~]$ pdsh -a date
    snx11000n000: Thu Aug 7 01:29:28 PDT 2014
    snx11000n003: Thu Aug 7 01:29:28 PDT 2014
    snx11000n002: Thu Aug 7 01:29:28 PDT 2014
    snx11000n001: Thu Aug 7 01:29:28 PDT 2014
    snx11000n007: Thu Aug 7 01:29:28 PDT 2014
    snx11000n006: Thu Aug 7 01:29:28 PDT 2014
    snx11000n004: Thu Aug 7 01:29:28 PDT 2014
    snx11000n005: Thu Aug 7 01:29:28 PDT 2014
    ```

    This completes the power on procedure. Mount the cluster as required.

Power Off Sonexion 2000

Prerequisites

- **System access requirements**:

  Root user access is required to perform this procedure on a Sonexion system. If you do not have root user access, contact Cray Support.

- **Service Interruption Level**:

  This procedure requires taking the Lustre file system offline.

- **Required Tools and Equipment**:
  - ESD strap, boots, garment or other approved protection
  - Console with monitor and keyboard (or PC with a serial port configured for 115.2 Kbps, 8 data bits, no parity and one stop bit)

About this task

Use this procedure to power off the Sonexion 2000 system using the CSCLI interface. The procedure stops Lustre on all Sonexion 2000 nodes.

The MMU component can use either a 2U24 EBOD or 5U84 EBOD, which hosts the MDT (storage target for the MGMT, MGS, and MDS nodes). This document uses the terms *2U24 EBOD* and *5U84 EBOD* to refer to the two types of storage enclosures. The term *2U Quad Server* refers to the MMU component. The term "5U84" refers to the SSU and ESU components.

> **IMPORTANT:** Unmount the Lustre file system from all clients before starting the power off procedure. Failure to do so may cause clients to hang.

Procedure

1. Log in to the primary MGMT node via SSH:

   ```
   [Client]$ ssh -l admin primary_MGMT_node
   ```

2. Change to root user:

   ```
   [admin@n000]$ sudo su -
   ```

3. Stop the Lustre file system:

   ```
   [root@n000]# cscli unmount -f filesystem_name
   ```

4. Verify that resources have been stopped by running the following on an even-numbered node:

```
[root@n000]# ssh nodename crm_mon -r1 | grep fsys
```

This is sample output showing the nodes are stopped:

```
[MGMT0]# ssh snx11000n006 crm_mon -r1 | grep fsys
snx11000n006_md0-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n006_md1-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n006_md2-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n006_md3-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n006_md4-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n006_md5-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n006_md6-fsys (ocf::heartbeat:XYMNTR): Stopped
snx11000n006_md7-fsys (ocf::heartbeat:XYMNTR): Stopped
```

5. Log in to the MGS node via SSH:

```
[root@n000]# ssh MGS_node
```

6. To check the MGS/MDS nodes to determine whether Resource Group **md65-group** is stopped, use the `crm_mon` utility to monitor the status of the MGS and MDS nodes and check that their resources have stopped:

```
[MGS]# crm_mon -1r | grep fsys
```

Following is an example `crm_mon` output showing the MGS and MDS nodes in a partial stopped state.

```
[MGS]# crm_mon -1r
sonexion11000n003_md66-fsys (ocf::heartbeat:XYMNTR): Stopped
sonexion11000n003_md65-fsys (ocf::heartbeat:XYMNTR): Started
```

If the node is not stopped, issue the `stop_xyraid` command:

```
[MGS]# stop_xyraid nodename_md65-group
```

This is sample `crm_mon` output showing the MGS and MDS nodes in a stopped state.

```
[MGS]# crm_mon -1r
sonexion11000n003_md66-fsys (ocf::heartbeat:XYMNTR): Stopped
sonexion11000n003_md65-fsys (ocf::heartbeat:XYMNTR): Stopped
```

7. Power off the diskless nodes:

```
[root@n000]# pdsh –f 100 –g diskless poweroff
```

8. Check the power-off status of the diskless nodes:

```
[root@n000]# pm –q
```

Repeat this step until all non-MGMT nodes have been powered down. Example:

```
[root@snx11000n000 ~]# pm -q
on: snx11000n[000-001]
```

```
off: snx11000n[002-011]
unknown:
```

9. From the primary MGMT node, power off the MGMT nodes:

```
[root@n000]# pdsh -g mgmt poweroff
```

10. Once all of the nodes are shut down, physically power off the enclosures by turning off the power switches (at the back of the rack).

    For the SSUs, ESUs, and 5U84 EBOD (if used), the power switches are on the PSU.

    The 2U24 EBOD power switches should remain switched on.

Automated System Power Up

Prerequisites

- The Sonexion system must be ready to be powered-up. Specifically, the PDUs must be powered on and all nodes must have their power switches in the on-position.
- The user must have a client system connected to the same LAN segment as **eth1** on the management nodes. That client system must have been installed with a Wake-on-LAN packet generator. For Linux systems, this functionality is provided by the etherwake package. For Windows systems, there are several Wake-on-LAN generators. A Windows client can be found at *http://sourceforge.net/projects/aquilawol/*.
- The client network switch must be configured to allow the ingress of Wake-on-LAN packets.
- Wake-on-LAN must be enabled in the management node BIOS and on the externally-facing network card on that node. The latter setting can be checked using the `ethtool` command:

  ```
  [root@n000]# ethtool eth1 | grep 'Wake'
  ```

  Output looks as follows:

  ```
  Supports Wake-on:pumbgWake-on:g
  ```

  In the command output, a 'd' indicates that Wake-on-LAN is disabled for Wake-on-LAN packets, while a 'g' indicates it is enabled. If it is disabled, boot into the BIOS and change it.
- The Sonexion system must be in daily mode.
- The operator must know the MGMT node MAC addresses. To find these MAC addresses run the following command from any available node:

  ```
  pdsh -g mgmt cat /sys/class/net/$(sed -nre '/mgmtLoginNetworkIfName:
  /s/.*\s//p' /etc/puppet/data/CSSM/cfg.yaml)/address
  ```

  Example:

  ```
  [root@snx11000n000 ~]# pdsh -g mgmt cat /sys/class/net/$(sed -nre '/
  mgmtLoginNetworkIfName:/s/.*\s//p' /etc/puppet/data/CSSM/cfg.yaml)/address
  snx11000n001: 00:1e:67:3f:07:44
  snx11000n000: 00:1e:67:3f:07:9c
  ```

About this task

Use this procedure to power on a Sonexion system using the automated system power up functionality.

Procedure

1. If you are using a Linux client, use the WoL generator to send a Wake-on-LAN packet to one of the management nodes:

   ```
   [Client]# ether-wake MGMT_node_MAC_address -i ifname-use_interface
   ```

   For example:

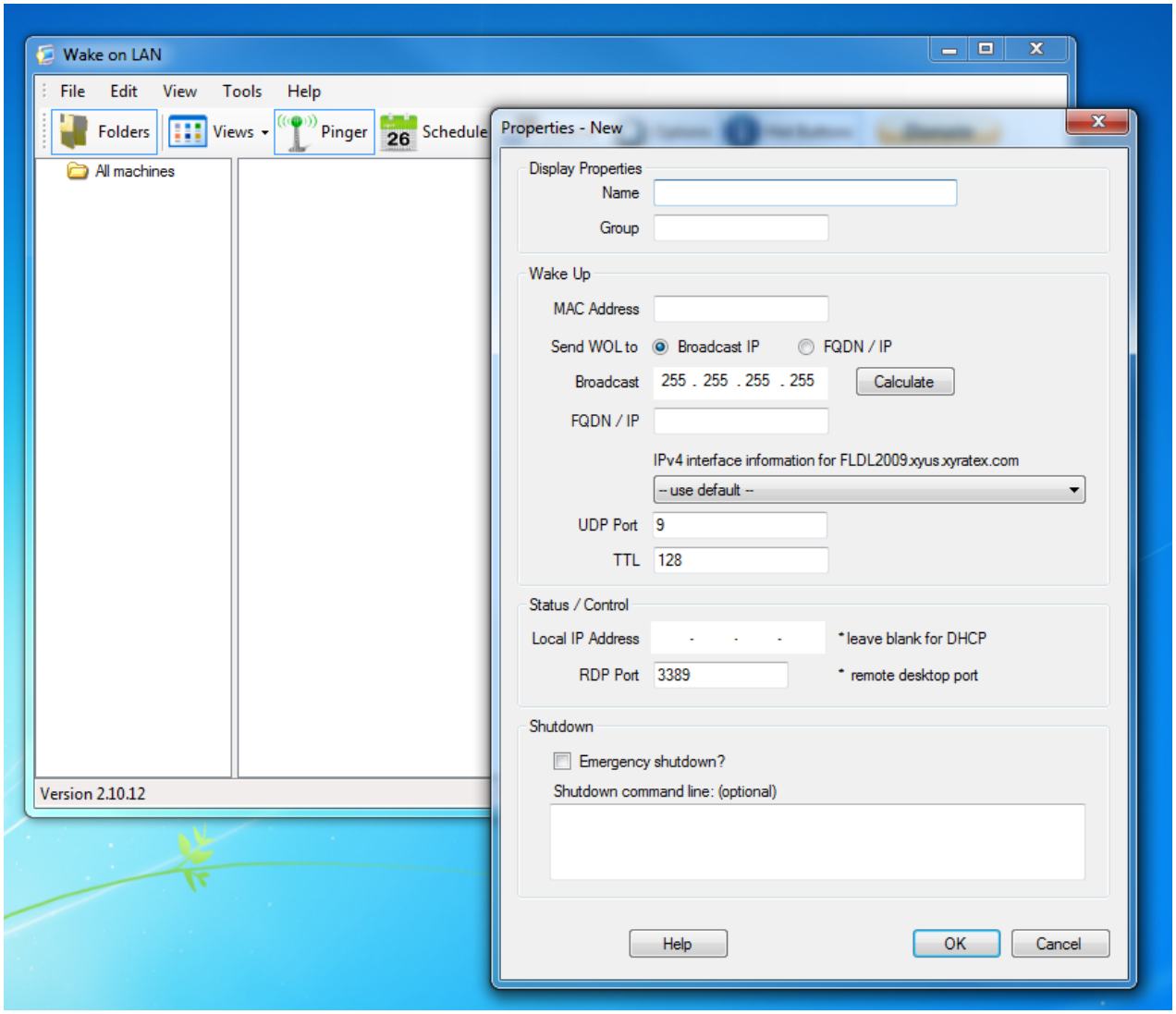
   ```
   [Client~]# ether-wake 00:1e:67:9f:75:72 -i eth1
   ```

2. If you are using a Windows client, use the Wake-on-LAN tool referenced in the introductory section. Then:

   a. Click file > **New Host**. The **Properties - New** window opens.

   b. In the **Properties - New** window (see the following figure) enter the settings for the target MGMT node.

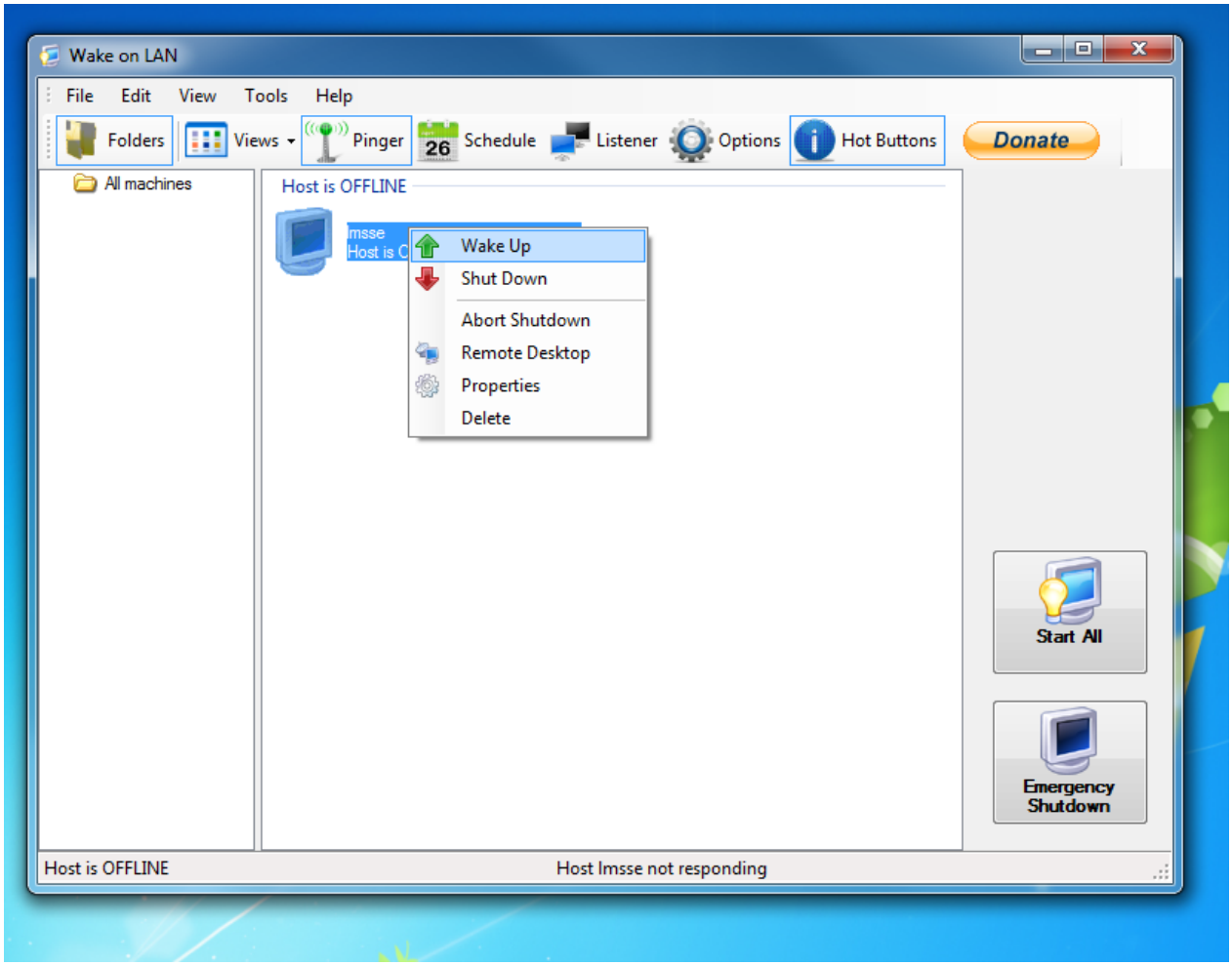
   Figure 3. Wake-on-Lan from an MS Windows Client



c. Click **OK**.

d. Right-click on the target machine and select **Wake Up** (see the following figure).

Figure 4. Select Wake Up



The next three steps are used to monitor the system power up. It is also possible to monitor the system power up from the Node Control tab of the CSSM GUI.

3. Wait several minutes for the management node to power up, and then log into the primary MGMT node via SSH:

```
[Client]$ ssh -l admin primary_MGMT_node
```

4. Verify that the other nodes booted successfully and are responding to SSH:

```
[MGMT]$ pdsh -g all date
```

5. Start the Lustre file system:

```
[MGMT]$ cscli mount -f fs_name
```

6. Verify that Lustre mounted successfully:

```
[MGMT]$ cscli fs_info
```