Consistent Device Naming on HP ProLiant Gen8 servers

Red Hat Enterprise Linux 6.1 and later

Table of contents

Introduction................................................................................................................................................................... 2
Red Hat Enterprise Linux 6.1 and biosdevname......................................................................................................... 4
SMBIOS solution on HP ProLiant Gen8 servers........................................................................................................... 6
ACPI DSM solution for Constant Device naming.......................................................................................................... 7
The biosdevname utility and SR-IOV........................................................................................................................... 8
The biosdevname utility on HP ProLiant G7 servers................................................................................................... 8
Upgrading from conventional to consistent device naming..................................................................................... 10
Reverting back to the conventional naming scheme................................................................................................ 10
Cautionary notes ........................................................................................................................................................ 11
For more information................................................................................................................................................. 12
Documentation feedback........................................................................................................................................... 12
Abstract

Red Hat Linux customers expect network device naming to be deterministic and persistent on their HP server platforms. Persistent device naming ensures that the ordering of network devices remains the same across reboots. Deterministic naming of devices ensures that the network device naming under the OS matches the way the ports are labelled on the server chassis. Ideally, network devices under Linux should be ordered in such a way that the first device label belongs to an embedded network controller (if present), even when an add-on network adapter is inserted into the system.

This white paper outlines the HP support industry-standard solution that ensures deterministic and persistent device naming on systems with Linux 2.6.x kernels. HP implements this solution in the platform firmware on HP ProLiant Gen8 and G7 servers. This white paper describes how to use this solution with systems running Red Hat Enterprise Linux v6.1 or later.

Introduction

A LAN on motherboard (LOM) is a network controller embedded in a server motherboard. On HP ProLiant Gen8 servers, the LOM ports are typically labeled bottom to top, right to left, on the back of the server chassis, as shown in Figure 1.

Figure 1. Quad-port LOM on an HP ProLiant DL360p Gen 8 server: Ports are labeled right to left (1 to 4 in this example)

Customers expect network devices to be enumerated in such a way that eth0 is assigned to the LOM. The assignment of Ethernet names to network devices is not always consistent. This is usually noticeable when adding a stand-up NIC to an HP ProLiant server. For instance, consider an HP ProLiant ML350p Gen8 server that has been configured with a quad-port 1GbE LOM. Figure 2 displays the Ethernet labels eth0 to eth3, corresponding to the four network ports on the LOM and their respective MAC addresses (9c:8e:4c:05:e2 through 9c:8e:4c:05:e5).
When a dual-port NIC card is inserted into the PCI Express slot 2 of the server, `eth0` no longer corresponds to the LOM. Instead, as shown in Figure 3, `eth0` is assigned to one of the ports on the add-on NIC card.

Figure 3. After adding a dual-port NIC card, the `eth0` Ethernet device is no longer associated with the LOM

This unexpected behavior is a result of changes to the default PCI device enumeration algorithm, introduced with Linux 2.6.x kernels. The default PCI device enumeration algorithm uses a depth-first algorithm instead of the breadth-first algorithm. This change causes PCI devices to be enumerated in a different order under the Linux kernel than reported by the Power-On Self-Test (POST) or the ROM-Based Setup Utility (RBSU). This change also affects other PCI devices such as storage controllers. For more information about this problem and its resolution, see:


To resolve this issue, HP ProLiant Gen8 servers support the industry-standard PCI Firmware Specification 3.1 ACPI DSM method and the SMBIOS-based solution. This solution enumerates the LOM devices in a specific predictable order, labels them appropriately, and allows operating systems, such as Red Hat Enterprise Linux 6.1 or later, to take advantage of it. See Table 1.
Table 1. New device-naming conventions by HP ProLiant Gen8 servers

<table>
<thead>
<tr>
<th>Device type</th>
<th>Naming convention</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-board (embedded) NICs</td>
<td>em[1-n] (such as em1, em2, and so forth)</td>
</tr>
<tr>
<td>Cards in PCI slots, port s1…n</td>
<td>p[slot_number]p[port_number] (such as p2p1)</td>
</tr>
<tr>
<td>SR-IOV devices</td>
<td>Add a suffix of s[vf], from 0…n, depending on the number of Virtual Functions exposed on each port (such as p2p1_0,p2p1_1)</td>
</tr>
</tbody>
</table>

Red Hat Enterprise Linux 6.1 and biosdevname

Red Hat Enterprise Linux 6.1 introduces support for biosdevname, the udev helper utility that suggests new names based on the location of the network adapters on the system, as provided by the system firmware. System firmware communicates this information to the OS through standards such as SMBIOS and ACPI.

The algorithm for the Linux biosdevname udev helper utility is simple and works as follows:

- If the system BIOS exposes the new PCI Firmware Specification 3.1 ACPI DSM method, the algorithm obtains and uses the interface label and index
- Else, if system BIOS exposes an index and label in SMBIOS 2.6 (or above) types 9 and 41, the algorithm exposes the index value
- Else, the algorithm falls back to using the legacy PCI IRQ Routing Table to determine the slots devices are located in, sorting the PCI device list in breadth-first order and assigning index values

The biosdevname utility is installed by default with Red Hat Enterprise Linux installations (except minimum installs). To enable the biosdevname functionality, the user must specify the boot parameter biosdevname=1 during the installation, as shown in Figure 4.

Figure 4. Enabling biosdevname during install

On Gen8 servers, use of biosdevname results in a change in nomenclature for network devices. For example, LOMs will no longer be called eth0 and eth1. Instead, they will be named em1 and em2, as shown in Figure 5. The em string stands for “embedded”. Unlike the eth names that start from 0, the em devices are ordered starting from 1. This matches the chassis labeling of ports, which also starts at 1 (see Figure 1).
The name `p2p1` denotes port 1 of the add-on NIC card in slot 2. Similarly, `p2p2` denotes port 2 of the card in slot 2. Comparing the MAC address information in Figure 5 with that of Figure 3, you can see that the LOM devices are consistently enumerated ahead of the add-on ports when `biosdevname` is in effect. Table 2 compares the traditional nomenclature with the new naming convention for LOM and add-on network devices.

Table 2. Comparing old and new methods for naming network devices

<table>
<thead>
<tr>
<th>NIC Type</th>
<th>MAC address</th>
<th>Traditional method</th>
<th>biosdevname method</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOM</td>
<td>9c:8e:99:4c:05:e2</td>
<td>eth2</td>
<td>em1</td>
</tr>
<tr>
<td>LOM</td>
<td>9c:8e:99:4c:05:e3</td>
<td>eth3</td>
<td>em2</td>
</tr>
<tr>
<td>LOM</td>
<td>9c:8e:99:4c:05:e4</td>
<td>eth4</td>
<td>em3</td>
</tr>
<tr>
<td>LOM</td>
<td>9c:8e:99:4c:05:e5</td>
<td>eth5</td>
<td>em4</td>
</tr>
<tr>
<td>ADD ON</td>
<td>A0:36:9f:00:41:a4</td>
<td>eth0</td>
<td>p2p1</td>
</tr>
<tr>
<td>ADD ON</td>
<td>A0:36:9f:00:41:a5</td>
<td>eth1</td>
<td>p2p2</td>
</tr>
</tbody>
</table>

After completing the installation of Red Hat Enterprise Linux with `biosdevname` support, the new naming scheme for network interfaces will be as shown in the following `ls` output example:

```
root@localhost ~]# ls /sys/class/net/
em1 em2 em3 em4 lo p2p1 p2p2
```

On Gen 8 HP ProLiant servers, a new label and index files are generated in the `/sys` file system.

**Note**

No label or index file will be found in the `/sys` file system. Instance name and label are exported to the `/sys` file system only on servers that support the Type 41 record (namely, HP ProLiant Gen8 servers).
As discussed at the beginning of this section, LOM devices are enumerated based on the information provided by the SMBIOS records or by the ACPI DSM, if one or both of these are supported. If neither is supported, device enumeration is based on the PCI IRQ Routing Table. The next two sections discuss the SMBIOS-based enumeration solution and the ACPI DSM solution, respectively.

**SMBIOS solution on HP ProLiant Gen8 servers**

The `biosdevname` utility’s ability to support new names for network devices requires system BIOS support for SMBIOS Version 2.6 Type 41 and Type 9 records. The type 9 record in the SMBIOS table that is included in the platform firmware on HP ProLiant servers can help you determine the bus address of a given PCI Express slot in the system, as shown in Example 1.

**Example 1:** A typical type 9 record in SMBIOS table

```bash
root@localhost ~]# dmidecode -t 9
Handle 0x0902, DMI type 9, 17 bytes
  System Slot Information
    Designation: PCI-E Slot 2
    Type: x8 PCI Express 3
    Current Usage: In Use
    Length: Long
    Characteristics:
      - 3.3 V is provided
      - PME signal is supported
    Bus Address: 0000:07:00.0
```

Example 1 shows that the PCI Express slot 2 belongs to bus address 7, device 0, and function 0. The In Use state indicates that an add-on card has been inserted in the slot.

The type 41 record in the SMBIOS table indicates the ordering of embedded devices on a server platform, as shown in Example 2. Such devices could include network, storage, and video controllers.

**Example 2:** Snippet of type 41 record entries in the SMBIOS table

```bash
root@localhost ~]# dmidecode -t 41
Handle 0x2901, DMI type 41, 11 bytes
  Onboard Device
    Reference Designation: NIC Port 1
    Type: Ethernet
    Status: Enabled
    Type Instance: 1
    Bus Address: 0000:03:00.0

Handle 0x2902, DMI type 41, 11 bytes
  Onboard Device
    Reference Designation: NIC Port 2
    Type: Ethernet
    Status: Enabled
    Type Instance: 2
    Bus Address: 0000:03:00.1
```

Example 2 shows that LOM ports located in the specific bus addresses are ordered as 1 and 2, as indicated by the Type Instance field.

The string used in the Reference Designation field in the type 41 SMBIOS record (see Example 2) is also exposed as the label in `/sys`. For instance, the label for the device `em1` is shown in the following example:

```bash
root@localhost ~]# cat /sys/class/net/em1/device/label
NIC Port 1
```

Together, the type 9 and type 41 SMBIOS records provide a mechanism for the OS to order the LOM devices and the add-on NIC devices consistently and to label them suitably. This not only enables you to distinguish LOMs from add-on controllers but also ensures that the first Ethernet device is always assigned to the LOM. However, for the functionality to exist, the `biosdevname` utility must be operational.
ACPI DSM solution for Constant Device naming

As noted in the preceding section, the SMBIOS Type 41 record provides Type Instance and Reference Designation as "NIC Port n"; the SMBIOS Type 41 record does not reveal whether the embedded NIC is an ordinary LOM or an Adaptive LOM (ALOM)/Blade Adaptive LOM (BLOM). It does not inform the end user about the type of motherboard NIC present in the server.

This shortcoming is addressed by ACPI DSM (Advanced Configuration and Power Interface and Device Specific Method), which were added to the HP ProLiant Gen8 BIOS. This solution provides and exports to the /sys file system the firmware instance number and string name of PCI devices, as defined by PCI Firmware Specification Revision 3.1 section 4.6.7 (DSM for Naming a PCI or PCI Express Device under Operating Systems).

Information provided by ACPI DSM is given higher priority than that given by SMBIOS Type 41 and Type 9 records. In absence of DSM, firmware will fall back to the SMBIOS-based approach.

With ACPI DSM, the entity ‘index’ is changed to ‘acpi_index’. The semantics of the “device type instance” provided by SMBIOS and the firmware “instance number” provided by ACPI DSM differ. With the ACPI DSM feature enabled, the label and index are indicated as shown in the following example:

```bash
root@localhost ~]#cat /sys/class/net/em1/device/label
Embedded FlexibleLOM Port 1
```

```bash
root@localhost ~]#cat /sys/class/net/em1/device/acpi_index
1
```

With the ACPI DSM feature disabled, firmware resorts to the SMBIOS-based approach and the label and index are indicated as shown in the following example:

```bash
root@localhost ~]#cat /sys/class/net/em1/device/label
NIC Port 1
```

```bash
root@localhost ~]#cat /sys/class/net/em1/device/index
1
```

Support for DSM is optional. This can be enabled or disabled through RBSU. When supported, the platform BIOS provides support for DSM, and any ACPI-aware OS can take advantage of DSM and use its device-specific information for deterministic display in the OS device manager or equivalent user interfaces. DSM returns an ACPI index, which is the instance number and a label (string name) assigned to the network device by the system firmware.

**Figure 6** shows an RBSU screen example displaying default support enabled for embedded LOMs. To modify these defaults, the user has to navigate through **F9 → Advance Options → Advance System ROM Options → Consistent Device Naming (CDN)**.
The biosdevname utility and SR-IOV

Single Root I/O Virtualization (SR-IOV) is an I/O virtualization technology defined in the PCI Sig SR-IOV Specification. On platforms that support SR-IOV, each physical interface can be split into multiple virtual instances or functions. Each of these virtual instances or functions can then be assigned to virtual machines.

More information about SR-IOV and implementing SR-IOV for Red Hat Enterprise Linux on HP ProLiant servers can be found at:


The interface names for virtual functions exposed by SR-IOV adapters (such as the Intel 82599 network adapter) are as follows (assuming that one virtual function is enabled per physical function):

```
root@localhost ~]# ls /sys/class/net/
em1 em2 em3 em4 lo p2p1 p2p1_0 p2p2_1 p2p2
```

Table 3 lists the naming scheme for physical and virtual functions.

<table>
<thead>
<tr>
<th>Physical function</th>
<th>p2p1</th>
<th>p{slot_number}[port_number]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual functions</td>
<td>p2p1_0</td>
<td>p{slot_number}[port_number][vf_instance]</td>
</tr>
</tbody>
</table>

The biosdevname utility on HP ProLiant G7 servers

HP ProLiant G7 and earlier servers do not support the type 41 SMBIOS record. Support for Type 41 has been made available only for Gen8 and future HP ProLiant servers. As discussed earlier in this paper, if ACPI DSM is not supported, biosdevname resorts to the SMBIOS method. If SMBIOS is not supported, biosdevname then falls back to the legacy PCI IRQ Routing Table.

If the Type 41 record is missing in SMBIOS, biosdevname (0.3.11 or later) falls back to using the correct PCIe slot number for retrieving all the details. To determine the “In Use” NIC card and the slot positions, biosdevname collects the correct slot information from SMBIOS Type 209 and Type 221 records.

1 Older versions of biosdevname use the PIRQ table, which is inconsistent on many BIOS versions.
Installing Red Hat Enterprise Linux 6.4 on the HP ProLiant DL580 G7 (a server that is in the Linux white list for default sorting of PCI devices), if "biosdevname=1" is set as the boot parameter, network interfaces are renamed from ethx to emy, and add-on cards are renamed from ethx to pxpy.

The naming of the network interfaces are as follows:

```bash
root@localhost ~]# ls /sys/class/net/
em1  em2  em3  em4  lo  p7p1  p7p2
```

```
[root@localhost ~]# dmidecode -t 41
# dmidecode 2.11
SMBIOS 2.7 present.
```

```
[root@localhost ~]# dmidecode -t 9
# dmidecode 2.11
SMBIOS 2.7 present.
```

Handle 0x0907, DMI type 9, 17 bytes
System Slot Information
- Designation: PCI-E Slot 7
- Type: X4 PCI Express 2 x8
- Current Usage: In Use
- Length: Long
- ID: 7
- Characteristics:
  - 3.3 V is provided
  - PME signal is supported
- Bus Address: 0000:47:00.0

```
[root@localhost ~]# dmidecode -t 209
# dmidecode 2.11
SMBIOS 2.7 present.
```

Handle 0xD100, DMI type 209, 36 bytes
HP BIOS NIC PCI and MAC Information
- NIC 1: PCI device 04:00.0, MAC address 78:E7:D1:58:ED:8C
- NIC 2: PCI device 04:00.1, MAC address 78:E7:D1:58:ED:8D
- NIC 3: PCI device 04:00.2, MAC address 78:E7:D1:58:ED:8E
- NIC 4: PCI device 04:00.3, MAC address 78:E7:D1:58:ED:8F

```
[root@localhost ~]# dmidecode -t 221
# dmidecode 2.11
SMBIOS 2.7 present.
```

Users can check the entry logs logged in dmesg:

```
udev: starting version 147
udev: starting version 147
udev: starting version 147
udev: renamed network interface eth0 to em1
udev: renamed network interface eth3 to em4
udev: renamed network interface eth1 to em2
udev: renamed network interface eth2 to em3
udev: renamed network interface eth0 to p7p1
udev: renamed network interface eth1 to p7p2
```

**Note**
No label or index file will be found in the /sys file system. The instance name and label are exported to the /sys file system only on servers that support the Type 41 record (namely, HP ProLiant Gen8 servers).
Upgrading from conventional to consistent device naming

For upgrades from Red Hat Enterprise Linux 6.0 to 6.1 or later, the older device naming scheme using eth_n is preserved. If you want to use the new nomenclature outlined in this document, HP recommends a fresh installation.

If a fresh installation is not suitable, follow these steps to upgrade from the conventional naming scheme to the Consistent Device Naming scheme:

1. Back up the data and OS configurations as per instructions in the Red Hat documentation.
2. Upgrade to the destination OS (refer to the Red Hat documentation), providing biosdevname=1 as a boot parameter.
3. If your previous installation used the traditional naming scheme, after upgrading the OS, the system continues to use the traditional naming (or same) scheme. To get the mapping from the current eth_n names to the new emx or pxpy names, issue the biosdevname –d command.

For more information, refer to the following Linux man page:
http://linux.die.net/man/1/biosdevname

4. List all the configuration files in the system using the eth naming scheme. Replace each occurrence of ethx with the corresponding emx or pxpy name, based on the mapping obtained in step 3. You can update the system’s network configuration manually or by using the system-config-network-cmd utility:
   A. **Updating the network configuration manually**
      Assume the updated system has the /etc/sysconfig/network-scripts/ifcfg-eth0 network script. By issuing biosdevname -d, you would get p2p1 as the corresponding BIOS name for eth0. Rename ifcfg-eth0 to ifcfg-p2p1.
      Edit ifcfg-p2p1 and change ”DEVICE=eth0” to ”DEVICE=p2p1”. Do the same for all network scripts.
   B. **Updating the network configuration using system-config-network-cmd utility**
      i. Save (export) the network configuration of the system to the file /tmp/network-config by issuing the following command as root:

         system-config-network-cmd -e > /tmp/network-config

      ii. Replace each occurrence of the ethx name in /tmp/network-config with the corresponding emx or pxpy name as obtained in step 3.

      iii. Restore or import the modified network configuration file by issuing the following command as root:

         system-config-network-cmd -i -c -f /tmp/network-config

         **Note:** The effect of each specified option is as follows:
         - -i option Imports the data
         - -c option Clears the existing configuration prior to importing the data
         - -f option Specifies the file to import

      iv. Verify in /etc/sysconfig/network-scripts/ ifcfg-* that the new names are used for the file names and for the DEVICE variable in each of the configuration files.

5. Update any custom scripts, iptables rules, and service configuration files that might include network interface names.
7. Reboot the server and make sure all the configuration files, iptables, and other related files reflect the new naming scheme.

Reverting back to the conventional naming scheme

In case you install the operating system with biosdevname=1 and would like to revert back to the old ethx naming conventions, follow the below steps:

2. Remove (or comment) the HWADDR lines from all /etc/sysconfig/network-scripts/ifcfg-* files.
3. Rename the ifcfg-e* or ifcfg-p* files to use the old naming convention (for example, change ifcfg-p2p1 to ifcfg-eth0). The new names will be in effect after reboot.
4. Update any custom scripts, iptables rules, and service configuration files that might include network interface names.
5. Reboot the system.

**Cautionary notes**

- The `biosdevname` feature is an option enabled during OS installation only; enabling or disabling the feature after installation is possible but not recommended because the feature will not work as expected. To modify or resort to the previously-used naming scheme, adhere to the steps delineated in the preceding sections.
- Label and index files are not present in the `/sys` file system on HP ProLiant G7 and earlier servers. The attribute is created only if the firmware has given a name to the PCI device. Due to the absence of Type 41 records on HP Proliant G7 servers, string name and instance number are not exported to `/sys` file system.
- Label and index files are available only for embedded NICs on HP ProLiant Gen8 servers using ACPI/SMBIOS records. These files are not exposed for add-on NIC cards.
For more information

HP ProLiant Gen8 Agentless Management
hp.com/go/proliantgen8

Linux documentation
hp.com/go/linux-docs
Select HP Linux Server Management Software

SMBIOS Specification document
dmtf.org/standards/smbios

ACPI 5.0 Specification document
acpi.info/spec.htm

PCI Specification document
pcisig.com/members/downloads/pcfw_r3_1_13Dec10.pdf

Documentation feedback

HP is committed to providing documentation that meets your needs. To help us improve the documentation, send any errors, suggestions, or comments to Documentation Feedback (docsfeedback@hp.com). When submitting your feedback, include the document title and part number, version number, or the URL.