PENGUIN™
SOLUTIONS

Using SR-IOV and DPDK with PCIE-8130

Application Note

P/N: 6806800V47C

January 2023

PENGUIN™
EDGE

Legal Disclaimer*

SMART Embedded Computing, Inc. (SMART EC), dba Penguin Solutions™, assumes no responsibility for errors or omissions in these materials. **These materials are provided "AS IS" without warranty of any kind, either expressed or implied, including but not limited to, the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.** SMART EC further does not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within these materials. SMART EC shall not be liable for any special, indirect, incidental, or consequential damages, including without limitation, lost revenues or lost profits, which may result from the use of these materials. SMART EC may make changes to these materials, or to the products described therein, at any time without notice. SMART EC makes no commitment to update the information contained within these materials.

Electronic versions of this material may be read online, downloaded for personal use, or referenced in another document as a URL to a SMART EC website. The text itself may not be published commercially in print or electronic form, edited, translated, or otherwise altered without the permission of SMART EC.

It is possible that this publication may contain reference to or information about SMART EC products, programming, or services that are not available in your country. Such references or information must not be construed to mean that SMART EC intends to announce such SMART EC products, programming, or services in your country.

Limited and Restricted Rights Legend

If the documentation contained herein is supplied, directly or indirectly, to the U.S. Government, the following notice shall apply unless otherwise agreed to in writing by SMART EC.

Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data clause at DFARS 252.227-7013 (Nov. 1995) and of the Rights in Noncommercial Computer Software and Documentation clause at DFARS 252.227-7014 (Jun. 1995).

SMART Embedded Computing, Inc., dba Penguin Solutions

2900 S. Diablo Way, Suite 190

Tempe, Arizona 85282

USA

*For full legal terms and conditions, visit *https://www.penguinsolutions.com/edge/legal/*

Table of Contents

Table of Contents

About this Manual

Overview of Contents

This document provides information to use Single Root I/O Virtualization (SR-IOV) and the Data Plane Development Kit (DPDK) with the PCIE-8130 PCI Express card to accelerate packet processing workloads.

Abbreviations

This document uses the following abbreviations:

Abbreviation	Definition
CLI	Command Line Interface
DPDK	Data Plane Development Kit
DSP	Digital Signal Processor
IP	Internet Protocol
LAN	Local Area Network
NIC	Network Interface Controller
PCIe/PCIE	PCI Express
PCI	Peripheral Component Interconnect
PF	Physical Function
SR-IOV	Single Root I/O Virtualization
SW	Software
VLAN	Universal Serial Bus
VF	Virtual Function
VM	Virtual Machine

Conventions

The following table describes the conventions used throughout this manual.

Notation	Description
0x00000000	Typical notation for hexadecimal numbers (digits are 0 through F), for example used for addresses and offsets
0b0000	Same for binary numbers (digits are 0 and 1)
bold	Used to emphasize a word
`Screen`	Used for on-screen output and code related elements or commands. `Sample of Programming used in a table (9pt)`
`Courier + Bold`	Used to characterize user input and to separate it from system output
Reference	Used for references and for table and figure descriptions
File > Exit	Notation for selecting a submenu
<text>	Notation for variables and keys
[text]	Notation for software buttons to click on the screen and parameter description
...	Repeated item for example node 1, node 2, ..., node 12
. . .	Omission of information from example/command that is not necessary at the time
..	Ranges, for example: 0..4 means one of the integers 0,1,2,3, and 4 (used in registers)
|	Logical OR
	Indicates a hazardous situation which, if not avoided, could result in death or serious injury
	Indicates a hazardous situation which, if not avoided, may result in minor or moderate injury

Notation	Description
	Indicates a property damage message
	Indicates a hot surface that could result in moderate or serious injury
	Indicates an electrical situation that could result in moderate injury or death
	Indicates that when working in an ESD environment care should be taken to use proper ESD practices
	No danger encountered, pay attention to important information

Summary of Changes

This manual has been revised and replaces all prior editions.

Part Number	Publication Date	Description
6806800V47C	January 2023	Update Section 1.2.4 and 1.3.
6806800V47B	August 2022	Rebrand to Penguin Solutions.
6806800V47A	April 2021	Initial version

About this Manual

Using SR-IOV and DPDK with PCIE-8130

1.1 Overview

The PCIE-8130 card supports Single Root I/O Virtualization (SR-IOV). To use SR-IOV, Virtual Functions (VFs) from the PCIE-8130 Network Interface Controller (NIC) need to be created on the host system.

1.2 Process to Create Virtual Functions

1.2.1 List the PCIE-8130 Cards in the System

To create VFs for a PCIE-8130 card in the system, it is necessary to find the PCI bus:dev.fn for the card. To do that, run pcie8130-listdev. In the example below, the system has a half-length (HL) and a full-length (FL) card:

```
# pcie8130-listdev
PCIE-8130-HL-6#0
   CPLD0: 37:00.0 ens2np0
PCIE-8130-FL-12#1
   CPLD1: 86:00.0 ens5np0
```

The output above shows the PCI bus:dev.fn in RED for each PCIe card in the system. \

1.2.2 Create Virtual Functions for PCIE-8130 Card

Using the PCI information obtained above, it is very simple to create VFs for a PCIE-8130 card. The maximum number of VFs supported by a PCIE-8130 card is 16.

For example, to create one VF on the HL card and two VFs on the FL card:

```
# echo 1 > /sys/bus/pci/devices/0000\:37\:00.0/sriov_numvfs
# echo 2 > /sys/bus/pci/devices/0000\:86\:00.0/sriov_numvfs
```

At this point you have one VF on the HL card and two VFs on the FL card and you can see that in the output of lspci:

```
# lspci | grep Broadcom
37:00.0 Ethernet controller: Broadcom Inc. and subsidiaries
BCM57412 NetXtreme-E 10Gb RDMA Ethernet Controller (rev 01)
37:02.0 Ethernet controller: Broadcom Inc. and subsidiaries
NetXtreme-E Ethernet Virtual Function
```

```
86:00.0 Ethernet controller: Broadcom Inc. and subsidiaries
BCM57412 NetXtreme-E 10Gb RDMA Ethernet Controller (rev 01)
86:02.0 Ethernet controller: Broadcom Inc. and subsidiaries
NetXtreme-E Ethernet Virtual Function
86:02.1 Ethernet controller: Broadcom Inc. and subsidiaries
NetXtreme-E Ethernet Virtual Function
```

You can also see Ethernet interfaces created for the VFs on the host. For the HL card the Physical Function (PF) interface is ens2np0. To find the VF interface:

```
# ifconfig -a | grep ens2
ens2np0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
ens2v0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
```

For the FL card the PF interface is ens5np0. To find the VF interfaces:

```
# ifconfig -a | grep ens5
ens5np0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
ens5v0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
ens5v1: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
```

The DSPs from a PCIE-8130 card can be segregated using VLANS. In the example above there were two VFs created for the FL card and it can be segregated in two groups of six DSPs, i.e., DSP0 - DSP5 and DSP6 - DSP11.

To assign a VLAN to a VF run the command: ip link set PHYS_INTF_NAME vf VF_NUM vlan VLAN_NUM.

```
# ip link set ens5np0 vf 0 vlan 15
# ip link set ens5np0 vf 1 vlan 25
```

We have assigned VLAN 15 to ens5v0 and VLAN 25 to ens5v1. The PCIE-8130 Marvell software also needs to be updated to assign DSP0 - DSP5 to VLAN 15 and DSP6 - DSP11 to VLAN 25 via the Marvell SW CLI. The mapping of DSP to switch port is as follows:

Table 1-1 DSP to Switch Port Map

DSP Number	Switch Port
0	0/26
1	0/4
2	0/24
3	0/12
4	0/27
5	0/8
6	0/20

Table 1-1 DSP to Switch Port Map (continued)

DSP Number	Switch Port
7	0/16
8	0/0
9	0/1
10	0/2
11	0/3

The host interface is connected to switch port 0/25.

To enter the Marvell SW CLI telnet to the Marvell SW IP address with port 12345:

```
# telnet 192.168.11.2 12345
Console# configure
Console(config)# interface vlan device 0 vid 15
Console(config-if)# exit
Console(config)# interface vlan device 0 vid 25
Console(config-if)# exit
Console(config)# interface range ethernet 0/26,4,24,12,27,8
Console(config-if)# switchport allowed vlan add 15 untagged
Console(config-if)# switchport allowed vlan remove 1
Console(config-if)# switchport customer vlan 15
Console(config-if)# exit
Console(config)# interface range ethernet 0/0-3,16,20
Console(config-if)# switchport allowed vlan add 25 untagged
Console(config-if)# switchport allowed vlan remove 1
Console(config-if)# switchport customer vlan 25
Console(config-if)# exit
Console(config)# interface ethernet 0/25
Console(config-if)# switchport allowed vlan add 15 tagged
Console(config-if)# switchport allowed vlan add 25 tagged
Console(config-if)# exit
Console(config)# exit
Console# CLIexit
```

NOTE: The steps listed above are not persistent and need to be executed every time the host boots.

1.2.3 Create Virtual Machines

Create Virtual Machines (VMs) using virt-manager. The default Network Address Translation (NAT) networking is just fine for the VM. Pass the PCI device from the PCIE-8130 to the VM using pcie-passthrough. Make sure to add a dchp server and tftp server to the VM so the DSPs are able to boot from the VM if desired.

1.2.4 Configure the VM to Boot the DSPs

Configuring the VM to boot the DSPs is a manual process as the utilities provided to manage the PCIE-8130 cards work on bare metal only.

1. Create the directory `/var/lib/tftpboot/pcie8130`
2. Create the directory `/opt/bladeservices/etc`
3. Copy the file `/var/lib/tftpboot/pcie8130/hdt_img_t121.bin` from the host to the same directory on the VM.
4. Copy the file `/var/lib/tftpboot/autoboot.hdt` from the host to the same location on the VM
5. Edit the file `/var/lib/tftpboot/autoboot.hdt` with the command to load the Vocallo image for example: `dsprun -s octmgw_app.imgx`. Make sure this line is not commented. This command is executed by the DSP to load the Vocallo image.
6. Copy the file `/var/lib/tftpboot/octmgw_app.imgx` from the host to the same location on the VM
7. Copy the file `/opt/bladeservices/etc/pdevXX-dhcpd.conf` from the host to the same directory on the VM
8. Edit `/etc/dhcp/dhcpd.conf` to add a line:
   `include "/opt/bladeservices/etc/pdevXX-dhcpd.conf"`
9. Edit `/opt/bladeservices/etc/pdevXX-dhcpd.conf` on the VM to have the correct IP address information and interface name
10. Copy the file `/etc/systemd/system/tftp.service` from the host to the same directory on the VM
11. Reload the system daemon on the VM: `systemctl daemon-reload`
12. Restart dhcpd on the VM: `systemctl restart dhcpd`
13. Configure firewalld and selinux if they are running on the VM
14. On the host comment out the line:
    `include "/opt/bladeservices/etc/pdevXX-dhcpd.conf";` from the system dhcpd.conf
15. Restart dhcpd on the host: `systemctl restart dhcpd`

16. On the host, power the DSPs assigned to VMs down and up, so they can boot from the VM

NOTE: In the file name pdev**XX**-dhcpd.conf **XX** represents the PCI BUS for the PCIE-8130 card passed to the VM.

1.3 DPDK and the PCIE-8130

To use DPDK with the PCIE-8130 card it is necessary to have a VM for the PCIE-8130. The VM runs an unmodified version of DPDK. In this example DPDK 19.11.6 is used. Download DPDK 19.11.6 from dpdk.org on the VM.

1. Compile DPDK

```
# mkdir /root/DPDK
# cd /root/DPDK
# tar -Jxf /root/dpdk-19.11.6.tar.xz
# cd dpdk-stable-19.11.6
# make config T=x86_64-native-linuxapp-gcc
# make install T=x86_64-native-linuxapp-gcc
# cd examples/kni
# make RTE_SDK=/root/DPDK/dpdk-stable-19.11.6 RTE_TARGET=x86_64-native-linuxapp-gcc
# cd ../..
```

2. Load DPDK kernel modules

```
# mkdir -p /mnt/huge
# echo 1024 >  /sys/devices/system/node/node0/hugepages/hugepages-2048kB/nr_hugepages
# mount -t hugetlbfs nodev /mnt/huge
# modprobe uio
# insmod x86_64-native-linuxapp-gcc/kmod/igb_uio.ko
# insmod x86_64-native-linuxapp-gcc/kmod/rte_kni.ko carrier=on
```

3. Unbind 8130 NIC from BCM driver and bind it to DPDK driver

```
# python3 usertools/dpdk-devbind.py -u <PCI BUS FOR PCIE-8130 NIC on VM>
# python3 usertools/dpdk-devbind.py -b igb_uio <PCI BUS FOR PCIE-8130 NIC on VM>
```

4. Start DPDK KNI application

```
# examples/kni/build/kni -c 0xf -n 1 -- -p 0x1 -P --config="(0,1,2,3)"
```

5. On a different window connected to the VM assign an IP address to virtual NIC interface vEth0_0 created by KNI sample application

```
# ifconfig vEth0_0 192.168.10.1/24 up
```

6. Update the dhcpd.conf file on the VM to use the interface name created by the DPDK KNI application and restart dhcpd
7. On the host, power the DSPs for the card used by the VM down and up, so the DSPs can boot from the VM

NOTE: More info on commands need can be found in Section 5.2.4.1 pcie8130octmezz in the PCIE-8130 Installation and Use manual. Refer to *Section A.1, Penguin Solutions Documentation on page 17* for information on how to obtain the manual.

1.4 FAQ/Troubleshooting

Things to look at in your system:

1. `cat /var/lib/tftpboot/autoboot.hdt`

   # The commands in this file will be automatically executed upon hdt boot (pin reset)

   # Edit this file with the command to load the Vocallo image for example:

   `dsprun -s octmgw_app.img`**`x`**

   Make sure this line is not commented. This command is executed by the DSP to load the Vocallo image
2. The file `octmgw_app.imgx` must be present in `/var/lib/tftpboot`

   ```
   # ls /var/lib/tftpboot/octmgw_app.imgx
   /var/lib/tftpboot/octmgw_app.imgx
   ```
3. The file `/var/lib/tftpboot/pcie8130/hdt_img_t121.bin` should be present

   ```
   # ls /var/lib/tftpboot/pcie8130/hdt_img_t121.bin
   ```
4. For each PCIE-8130 card, there should be a file named pdev* in `/opt/bladeservices/etc`

   ```
   # ls /opt/bladeservices/etc/pdev*
   /opt/bladeservices/etc/pdev1a-dhcpd.conf
   /opt/bladeservices/etc/pdev29-dhcpd.conf
   ```
5. For each PCIE-8130 card there should be a `#include` line in `/etc/dhcpd.conf`. In my system I have 2

   ```
   # grep pdev /etc/dhcpd.conf
   include "/opt/bladeservices/etc/pdev29-dhcpd.conf";
   include "/opt/bladeservices/etc/pdev1a-dhcpd.conf";
   ```

6. dhcpd must be running

   ```
   # systemctl status dhcpd
   ```

   dhcpd.service - DHCPv4 Server Daemon

   Loaded: loaded (/etc/systemd/system/dhcpd.service; enabled; vendor preset: disabled)

   Active: active (running) since Wed 2021-05-26 02:23:10 UTC; 1 months 2 days ago

   Docs: man:dhcpd(8)

   man:dhcpd.conf(5)

   Main PID: 2225 (dhcpd)

   Status: "Dispatching packets..."

   Tasks: 1

   CGroup: /system.slice/dhcpd.service

   /usr/sbin/dhcpd -d -f -cf /opt/bladeservices/etc/dhcpd.conf -user dhcpd -...

7. TFTP is started by tftp.socket

   ```
   # systemctl status tftp.socket
   ```

   tftp.socket - Tftp Server Activation Socket

   Loaded: loaded (/usr/lib/systemd/system/tftp.socket; enabled; vendor preset: disabled)

   Active: active (listening) since Wed 2021-05-26 02:22:43 UTC; 1 months 2 days ago

   Listen: [::]:69 (Datagram)

8. firewalld could be blocking tftp.Disable the firewall or configure it to allow the tftp service

Related Documentation

A.1 Penguin Solutions Documentation

Technical documentation can be found by using the Documentation Search at *https://www.penguinsolutions.com/edge/support/* or you can obtain electronic copies of documentation by contacting your local sales representative.

Table A-1 Penguin Solutions Documentation

Document Title	Documentation Number
PCIE-8130 Installation and Use	6806877A01
PCIE-8130 Quick Start Guide	6806877A03
PCIE-8130 Safety Notes	6806877A02
PCIE-8130 Data Sheet	PCIE-8130 DS

PENGUIN™
SOLUTIONS