SINGLE-ROOT INPUT/OUTPUT VIRTUALIZATION (SR-IOV) WITH LINUX CONTAINERS

09.05 / 10:39 digit.in

Read the full story

Executive Summary Virtualization is a technology offering an abstraction layer between software and hardware. It can emulate a hardware platform in order to provide the operating system...

Executive Summary

Virtualization is a technology offering an abstraction layer between software and hardware. It can emulate a hardware platform in order to provide the operating system abstractions of various resources. It is also a means to easily manage hardware resources and to run services in the cloud. One of the main disadvantages of the traditional hypervisor-based virtualization is, however, that it introduces a large overhead with respect to memory and CPU1 usage. Linux* Containers (abbreviated to LXC) is a lightweight alternative to the traditional approach, providing separation of namespaces and file systems while running all processes on a single kernel. Combined with Single-Root Input/Output Virtualization (abbreviated to SR-IOV) it can provide a solution that allows the containers to appear in the network as separate compute-nodes with exclusive MAC2 addresses, while sharing one link and physical network adapter.

This paper is a result of a joint CERN openlab-Intel research activity with the aim to investigate whether Linux Containers can be used together with SR-IOV in conjunction and complementary to the existing virtualization infrastructure in the CERN Data Centre. This solution could be potentially applied to the storage nodes, which are principally used for Input/Output operations, while keeping the CPU mostly idle. CERN could benefit from LXC/SR-IOV by running both CPU-intensive and storage jobs in containers on one storage node, with full separation of both applications and control over consumed resources.

We describe a successful setup of Linux Containers on top of an Intel® 10-Gigabit Ethernet adapters with SR-IOV support together with results from synthetic network and storage benchmarks.

Authors: Pawel Szostek (CERN openlab) and Marco Righini (Intel)

1 Central Processing Unit
2 Media Access Control

Abstract

This document investigates the possibility to configure SR-IOV with LXC in the CERN environment with the most recent flavour of the operating system used at CERN (CERN CentOS* 7). We present all the steps necessary for an operational configuration. We also present results from the basic network and storage benchmarks run in this configuration. Our results prove that the technology stack is working. However, minor performance drops when running intensive I/O operations were detected. The future work involves running performance and reliability tests with the target technology stack.

1. Introduction

What are Linux Containers?

Linux* Containers, abbreviated to LXC, is a feature implemented in the Linux kernel, allowing processes to be separated inside their own namespaces and file systems, while running on the same kernel. As shown by various studies (Xavier, et al., 2013; Kjallman, et al., 2015), its performance overhead is smaller than for hypervisor-based, which comes for the cost of lower flexibility and manageability. LXC can be bundled together with cgroups, providing a way to control usage of hardware resources inside the containers, which results in even better separation of the applications running inside the different containers.

LXC differs from Docker in that it provides lightweight namespace separation capabilities while removing traditional VM overhead while Docker is a single application virtualization engine which runs on the top of the containers. It differs significantly from LXC in the way that it is meant to be running a single application, enclosed together with its dependencies in a single image. By default Docker disables storage persistence, making the images stateless over executions.

What is SR-IOV?

The Single-Root I/O3 Virtualization (SR-IOV) is a PCI Special Interest Group (PCI-SIG) specification for I/O virtualization. It provides a standard mechanism for devices to advertise their ability to be simultaneously shared among multiple virtual machines. It also allows for the partitioning of a PCI function into many virtual interfaces for the purpose of sharing the resources of a PCI Express* (PCIe*) device in a virtual environment.

Depending on the physical device and the support on the kernel side, a different number of Virtual Functions can be made available inside a system. For this paper we used Intel X520 Cards providing up to 64 Virtual Functions per port. Each of them can be delegated to a container, which will own it during its lifetime. As a result, the device will appear in the network as an independent node with completely isolated traffic.

The Big Picture

The CERN Data Centre lies at the heart of the Worldwide LHC Computing Grid. It is a place where data from the LHC experiments are gathered, stored and processed. It hosts almost 110,000 processor cores and 11,000 servers that run round-the-clock to ensure uninterrupted services. All the nodes hosted at the CERN Data Centre can be divided into several categories, depending on their use case. In this paper we focus on the storage front-end nodes, whose aim is to collect data coming from the experiments and serve this data on demand to the computing nodes. They are characterized by low CPU usage and intensive use of their persistent storage resources. The aim of the investigation described in this Openlab paper is to look at the possibility of using Linux Containers to provide additional computing resources from the storage servers without the need to sacrifice storage reliability. A diagram of the target setup is shown in Figure 1.

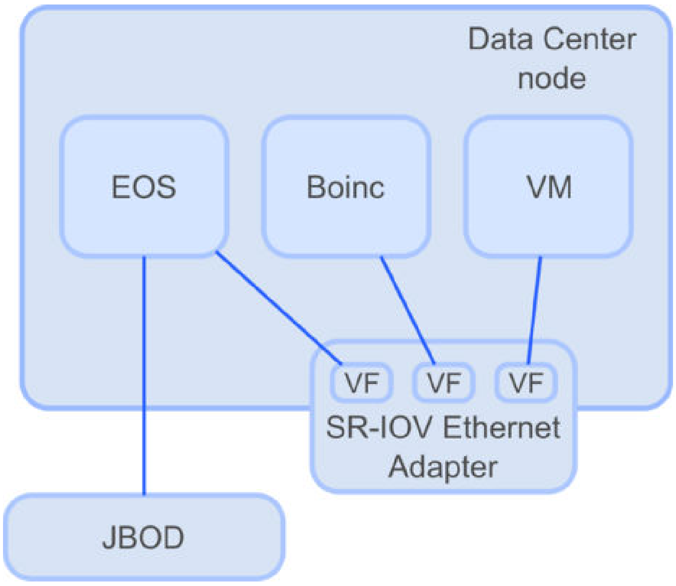


Figure 1: Example of a storage node configuration.

EOS is a disk storage system used at CERN for large physics data – it is I/O intensive with moderate need for computing resources.

Boinc is a distributed application for data processing – CPU intensive with negligible I/O. In the target configuration EOS and Boinc would run on a single storage node. Such a setup would allow better utilizing CPUs, which are currently underutilized, while keeping the applications separated from each other and providing them with the necessary hardware resources. To this end, SR-IOV enables us to create many virtual network interfaces to be used inside the containers independently, while cgroups provides the capability for resource separation.

In this paper we describe a setup of LXC with SR-IOV tested with a handful of easy-to-run synthetic benchmarks. We do not investigate the target configuration as depicted in Figure 1, since such a setup is complicated, time-consuming and error-prone. Instead, we focus on running synthetic benchmarks with the aim of stressing I/O-related hardware resources.

3 Input/Output (operations)

2. Hardware and software setup

The node tested is equipped with a JBOD array of 24 Hitachi HUA72302 A800 drives and with an Intel® 82599ES 10-Gigabit Ethernet Adapter featuring SR-IOV support. In addition, there is one data and one management 1-Gigabit Ethernet port on the system board.
The machine is booted with pre-production CentOS* 7 with the 3.10.0-123.13.2.el7.x86_64 kernel and CERN-specific add-ons. Standard packages are installed, unless explicitly mentioned.

BIOS configuration

SR-IOV requires the following BIOS settings to work correctly: VT enabled, SR-IOV enabled and VT-D disabled.

3. SR-IOV configuration in the OS

In this section we describe in detail all the steps needed to obtain a working setup of SR-IOV on the test system. In the remainder of the text the host machine will be called p01001534852033, while the containers will be called centos1 and centos2.

First of all, we ensure that the Ethernet adapters are visible in the system. Since these are PCIe devices, we use the lspci command:

[root@p01001534852033 tmp]# lspci | grep 10-Gig 04:00.0 Ethernet controller: Intel Corporation 82599ES 10-Gigabit SFI/SFP+ Network Connection (rev 01) 04:00.1 Ethernet controller: Intel Corporation 82599ES 10-Gigabit SFI/SFP+ Network Connection (rev 01)
Kernel drivers for the card employed are publicly available on the Intel website. Once the drivers are downloaded, we unpack, build and install both ixgbe and ixgbevf kernel modules:

```
tar xvzf ixgbe-*.tar.gz
cd ixgbe-*/src
make install
cd ../..
tar xvzf ixgbevf-*.tar.gz
cd ixgbevf-*/src
make install
```

Afterwards, we configure the system to load the kernel modules automatically at boot-time:
echo "ixgbevf\nixgbe" >> /etc/modules-load.d/modules.conf
In addition, the ixgbe driver requires virtual memory pages of large size (also called huge memory pages) to be mounted:

mkdir -p /mnt/huge chmod 777 /mnt/huge echo "huge /mnt/huge" hugetlbfs defaults 0 0 echo "vm.nr_hugepagesz=2M" >> /etc/sysctl.conf
The driver requires the following parameters to be passed at boot to the kernel:

intel_iommu=off default_hugepagesz=2M hugepagesz=2M ixgbe.max_vfs=8,8.
The last parameter specifies the number of virtual interfaces available for every physical interface. In this particular case we use 8 virtual interfaces for every physical one.

To this end, we add a custom entry to grub.cfg. Since CentOS7 uses grub2, this has to be done indirectly by adding an entry to /etc/grub.d/40_custom. The entry below is a clone of another entry with the extra parameters added. For further details please consult /boot/grub2/grub.cfg:

```
cat << EOF > /etc/grub.d/40_custom
menuentry 'CentOS Linux (3.10.0-123.9.3.el7.x86_64) 7 Intel setup' --class centos -
-class gnu-linux --class gnu --class os --unrestricted $menuentry_id_option
'gnulinux-3.10.0-123.el7.x86_64-advanced-3d81a16b-9af7-4ba0-ae53-
ef16fb54f864' {
    load_video
    set gfxpayload=keep
    insmod gzio
    insmod part_msdos
    insmod ext2
    set root='hd0,msdos1'
    if [ x$feature_platform_search_hint = xy ]; then
      search --no-floppy --fs-uuid --set=root --hint-bios=hd0,msdos1 --hint-
efi=hd0,msdos1 --hint-baremetal=ahci0,msdos1 --hint='hd0,msdos1'  0b7ebc67-
```

f22d-4760-abd9-7503761e827f
else
search --no-floppy --fs-uuid --set=root 0b7ebc67-f22d-4760-abd9-7503761e827f
fi
linux16 /vmlinuz-3.10.0-123.8.1.el7.x86_64 root=/dev/mapper/cc_p01001534852033-root ro rd.lvm.lv=cc_p01001534852033/swap crashkernel=auto vconsole.font=latarcyrheb-sun16 rd.lvm.lv=cc_p01001534852033/root vconsole.keymap=us rhgb quiet LANG=en_US.UTF-8 intel_iommu=off default_hugepagesz=2M hugepagesz=2M ixgbe.max_vfs=8,8
initrd16 /initramfs-3.10.0-123.8.3.el7.x86_64.img
}
EOF

Next, we recreate the GRUB2 configuration file to allow it include the latest changes:
grub2-mkconfig --output=/boot/grub2/grub.cfg
Finally we reboot the system to check that the SR-IOV-related configuration is correct and reproducible with every boot.

Once the system is rebooted with the new parameters, we test whether virtual functions are visible in the system. If they are not listed by lspci, it might mean that the drivers were not loaded or loaded incorrectly (in this case, consult lsmod and dmesg).

[root@p01001534852033 tmp]# lspci | grep -i virt
00:11.0 PCI bridge: Intel Corporation C600/X79 series chipset PCI Express Virtual Root Port (rev 06)
04:10.0 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:10.1 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:10.2 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:10.3 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:10.4 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:10.5 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:10.6 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:10.7 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:11.0 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:11.1 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:11.2 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:11.3 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:11.4 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:11.5 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
04:11.6 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual

```
Function (rev 01)
04:11.7 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual
Function (rev 01)
```

We double check that huge memory pages are mounted:

```
[root@p01001534852033 tmp]# mount | grep huge
cgroup on /sys/fs/cgroup/hugetlb type cgroup
(rw,nosuid,nodev,noexec,relatime,hugetlb)
hugetlbfs on /dev/hugepages type hugetlbfs (rw,relatime)
huge on /mnt/huge type hugetlbfs (rw,relatime)
```

Also, all the virtual interfaces should be reported by ifconfig:

```
[root@p01001534852033 ~]# ifconfig | grep enp4s | cut -c 1-53
enp4s16: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s17: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s0f0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s0f1: flags=4099<UP,BROADCAST,MULTICAST>  mtu 150
enp4s16f2: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s16f4: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s16f6: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s17f2: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s17f4: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
enp4s17f6: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>
```

If this is not the case, try reloading ixgbe and ixgbevf. However, if you use SSH to connect to the machine, the connection will break. This can be avoided by running.

```
nohup 'rmmod ixgbe && rmmod ixgbevf && insmod ixgbe && insmod ixgbevf'.
```

Note that ixgbevf has to be loaded after ixgbe, otherwise the virtual functions will not be reported by the system.

When working on the solution we stumbled upon a bug in the CentOS7 LXC template that results in a very slow container boot4. The bug can be fixed by applying a patch on the template file5.

Once the above-mentioned configuration is in place, we create a new CentOS container from scratch:
lxc-create -t centos -n centos1

Before we can start the container, we need to change its settings so that it does not use the bridged network adapter, but a virtual function instead. A sample configuration is shown below:

```
[root@p01001534852033 tmp]# cat /var/lib/lxc/centos1/config
lxc.autodev = 1
lxc.network.type = phys # was veth
lxc.network.flags = up
lxc.network.link = enp4s16f1  # was virbr0
lxc.rootfs = /var/lib/lxc/centos1/rootfs
lxc.include = /usr/share/lxc/config/centos.common.conf
lxc.arch = x86_64
lxc.utsname = centos1.cern.ch
lxc.autodev = 1
```

Each time the ixgbevf kernel module is loaded, the virtual functions are assigned new MAC addresses. To prevent any consequences (like misrouting the packets through the CERN network), we use Python script below to reset the addresses. We are aware that this is a suboptimal solution, but should work as a temporary workaround.

```
#!/usr/bin/env python
import subprocess
intf_mac_list0 = [("enp4s16","d6:d6:d6:54:f9:f8"),
("enp4s16f2", "ee:2b:1d:e6:f7:aa"),
("enp4s16f4", "1e:99:1e:f0:bb:64"),
("enp4s16f6", "6e:82:0c:ff:f8:47"),
("enp4s17",   "da:61:59:96:ab:1d"),
("enp4s17f2", "fe:66:3f:87:74:12"),
("enp4s17f4", "72:7b:1a:0c:bf:fc"),
("enp4s17f6", "46:02:c9:b1:3a:a7")]
intf_mac_list1 = [("enp4s16f1", "7a:e7:81:9c:43:38"),
("enp4s16f3",  "ca:fa:89:58:95:92"),
("enp4s16f5",  "ba:7d:2e:d3:1b:96"),
("enp4s16f7",  "3e:72:0d:07:2b:e9"),
("enp4s17f1",  "52:7b:24:90:11:01"),
("enp4s17f3",  "fe:14:4f:45:6d:49"),
("enp4s17f5",  "86:1e:f4:cc:fa:bb"),
("enp4s17f7",  "0a:93:df:2e:62:2b")]
for idx, tup in enumerate(intf_mac_list0):
  intf, mac = tup
  cmd = "ip link set %s vf %d mac %s" % ("enp4s0f0", idx, mac)
  print(cmd)
  subprocess.Popen(cmd, shell=True, stdin=subprocess.PIPE,
stdout=subprocess.PIPE)
for idx, tup in enumerate(intf_mac_list1):
  intf, mac = tup
  cmd = "ip link set %s vf %d mac %s" % ("enp4s0f1", idx, mac)
  print(cmd)
  subprocess.Popen(cmd, shell=True, stdin=subprocess.PIPE,
stdout=subprocess.PIPE)
```

Finally, we can run the container and connect to its terminal:

```
[root@p01001534852033 ~]# lxc-start –d -n centos1
[root@p01001534852033 ~]# lxc-attach -n centos1
[root@centos1 ~]#
```

https://lists.linuxcontainers.org/pipermail/lxc-users/2014-July/007443.html
https://github.com/tukiyo/lxc/commit/dbf45a526bf5a2f0503737b0b68d244e9389a2a7

4. Tests

Network Bandwidth Tests

We conducted very simple tests to prove that the network bandwidth obtained inside and outside of the container is at the same level. For this purpose, we set up two machines with 10GbE cards connected to the same 10GbE Ethernet switch. We used the iperf3, which benchmarks client-server data transmission. Our client is the machine hosting the containers, while the server is another machine (cd1001534-0004132575) connected directly to the same switch. First, we ran the benchmark both in and outside container and measured the bandwidth. Then we repeated the tests in reverse mode, i.e. with traffic going from the server to the client. The results show that the link is symmetrical and using LXC does not influence network traffic in this configuration.

```
[root@centos2 ~]# iperf3 -c cd1001534-0004132575 -O 3 -t 20
Connecting to host cd1001534-0004132575, port 5201
...
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bandwidth       Retr
[  4]   0.00-20.00  sec  21.9 GBytes  9.41 Gbits/sec    0             sender
[  4]   0.00-20.00  sec  22.0 GBytes  9.43 Gbits/sec                  receiver
[root@p010001534074188 ~]# iperf3 -c cd1001534-0004132575 -O 3 -t 20
...
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bandwidth       Retr
[  4]   0.00-20.00  sec  21.9 GBytes  9.41 Gbits/sec  0          ...
```

Tags: #Linux #Software

0 Comments Myinforms.com Login

Recommend Share Sort by Newest

Start the discussion…

Be the first to comment.

Subscribe Add Disqus to your site Add Disqus Add Privacy

Last news

Samsung mass produces 14-Nanometer processor

09.06 / 06:57 itp.net

Exynos 7570 is the first Exynos processor to fully integrate a Cat 4 LTE 2CA...

Exciting Images From Juno Show Different Faces Of Jupiter

09.06 / 06:55 hngn.com

After five years, Juno spacecraft got into orbit around Jupiter and captured rare and amazing images of the gas...

Razer announces the world's first gaming keyboards with Mecha-Membrane technology

09.06 / 06:52 neowin.net

Razer has launched the Ornata line of keyboards featuring Mecha-Membrane technology, This new technology combines the best features of mechanical and rubber dome keyboards for comfort and

Intel is acquiring Movidius, maker of compact computer vision chips: think VR, AR, drones, robots - CNET

09.06 / 06:47 cnet.com

Intel's push into AR, VR, and robotics continues with hardware that could potentially work on untethered headsets and...

Take a digital dive into the grisly sunken remains of Henry VIII's flagship - CNET

09.06 / 06:47 cnet.com

Artefacts from Henry VIII's sunken flagship, the Mary Rose, have been scanned in detailed 3D and posted online for the world to...

Capgemini in North America is selected by Wolters Kluwer for a multi-year contract extension for Finance and Accounting

09.06 / 06:40 capgemini.com

Paris – Capgemini, a global leader in consulting, technology and outsourcing services, has secured a five-year contract extension from Wolters Kluwer, a global leader in information services and solutions

Chinook Dredging Plans on the Table

09.06 / 06:39 dredgingtoday.com

The U.S. Army Corps of Engineers and the Washington Department of Ecology have received an application from Port of Chinook, WA to perform maintenance dredging operations in Port of Chinook Marina.

What you need to know about the Pennovation Center

09.06 / 06:37 thedp.com

The Pennovation Center is a 58,000 square foot business incubator and laboratory located on Grays Ferry Avenue by the Schuylkill...

Facebook joins forces with Samaritans support service to prevent suicides

09.06 / 06:37 tech.firstpost.com

Ahead of the World Suicide Prevention Day which falls on September 10, social networking giant Facebook has joined hands with support service Samaritans to launch a series of suicide prevention tools. The

Oculus issues warning against using the Samsung Galaxy Note 7 with the Gear VR headset

09.06 / 06:37 tech.firstpost.com

Oculus has put up a warning on the Gear VR page requesting users to get their Note 7 smartphones exchanged before using it with the Gear VR. The post Oculus issues warning against using the Samsung Galaxy

G20 Summit Roundup

09.06 / 06:35 soylentnews.org

Obama Cancels Meeting with Philippine President Duterte takyon writes:President Obama has cancelled a planned meeting with Filipino President Rodrigo Duterte after Duterte described Obama as a "son of

Archos 133 Oxygen Tablet Launched in IFA 2016 with a Huge 13.3-Inch Display – Check here

09.06 / 06:35 fitnhit.com

Archos 133 Oxygen Tablet: Archos a French multinational electronics company has Launched a New tablet to its Oxygen series at IFA 2016 with a huge 13.3-inch Display – the 133 Oxygen tablet with a Price tag

Twin Peaks meets X-Files in new Xbox One video game, Virginia

09.06 / 06:33 winbeta.org

Mystery/thriller video game, Virginia, has just been confirmed for release on Microsoft's Xbox One and is due out on September 22nd. The game is set within the first-person perspective and explores a story

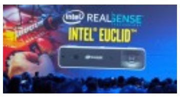
IFA 2016: Modular Desktop PCs Still In Vogue

09.06 / 06:28 techtree.com

At the IFA 2016 in Germany, the technology flowing about, has been noteworthy, as it seems there is something for everybody, even those who like PCs. Though most PCs aren't on the shopping list of most

Intel buys Movidius to build the future of computer vision

09.06 / 06:28 engadget.com



Intel is making it extra-clear that computer vision hardware will play a big role in its beyond-the-PC strategy. The computing behemoth has just acquired Movidius, a specialist in AI and

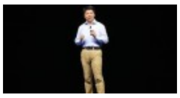
Sillpa leverages the capabilities of the Azure cloud to build applications that help ordinary people solve everyday prob

TN 09.06 / 06:23 blogs.technet.com

Sillpa is an award-winning, Ecuadorian software startup specializing in application development. The company is dedicated to developing a wide range of technological solutions that help solve every day

Huawei rotating CEO Guo Ping hails new era of ICT ecosystem alliances as 'An Age of Heroes'

09.06 / 06:19 vanillaplus.com



On the third day of HUAWEI CONNECT 2016 in Shanghai, Huawei rotating CEO Guo Ping, Intel CEO Brian Krzanich, and Huawei Products & Solutions president Ryan Ding took centre stage. Guo Ping's keynote

Microsoft Blocking Some Third-Party Windows Browsers as Part of Family Settings

09.06 / 06:12 news.softpedia.com

A change that Microsoft has recently made to its family filters on Windows devices might be

blocking some third-party browsers, according to one report, with the software giant recommending users to turn to Edge, the new default in

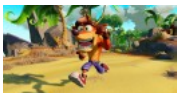
Destroy All Humans PS2 on PS4 Release Coming? Rated For PS4 In Europe

09.06 / 06:10 thebitbag.com

It seems like THQ Nordic is finally going to do something with the Destroy All Humans license. The open-world alien game looks… The post Destroy All Humans PS2 on PS4 Release Coming? Rated For PS4 In

Crash Bandicoot PS4 Remaster: Grand Reveal At PlayStation Meeting?

09.06 / 06:10 thebitbag.com



Gamers that grew up in the 1990's are excited for the Crash Bandicoot PS4 re-release. Ever since the remastered collection was announced… The post Crash Bandicoot PS4 Remaster: Grand Reveal At PlayStation