Rockchip Developer Guide PCIe Performance

文件标识：RK-KF-YF-460

发布版本：V1.0.0

日期：2022-05-23

文件密级：□绝密 □秘密 □内部资料 ■公开

免责声明

本文档按“现状”提供，瑞芯微电子股份有限公司（“本公司”，下同）不对本文档的任何陈述、信息和内容的准确性、可靠性、完整性、适销性、特定目的性和非侵权性提供任何明示或暗示的声明或保证。本文档仅作为使用指导的参考。

由于产品版本升级或其他原因，本文档将可能在未经任何通知的情况下，不定期进行更新或修改。

商标声明

“Rockchip”、“瑞芯微”、“瑞芯”均为本公司的注册商标，归本公司所有。

本文档可能提及的其他所有注册商标或商标，由其各自拥有者所有。

版权所有 © 2022 瑞芯微电子股份有限公司

超越合理使用范畴，非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

瑞芯微电子股份有限公司

Rockchip Electronics Co., Ltd.

地址： 福建省福州市铜盘路软件园A区18号

网址： www.rock-chips.com

客户服务电话： +86-4007-700-590

客户服务传真： +86-591-83951833

客户服务邮箱： fae@rock-chips.com

前言

概述

本文介绍 PCIe 各项主要性能测试方案及 RK 测试结果。

产品版本

芯片名称	内核版本
所有支持 PCIe 的芯片	Linux 4.4 及以上

读者对象

本文档（本指南）主要适用于以下工程师:

技术支持工程师

软件开发工程师

修订记录

版本号	作者	修改日期	修改说明
V1.0.0	林鼎强	2022-05-23	初始版本

目录

1. 背景

1.1 简介

本文介绍 RK PCIe 应用相关的性能指标及其测试方法，测试目标为通过简单、准确、可控的应用接口来获取对应的性能指标。

其中 “背景” 章节主要介绍测试过程中涉及到的部分知识背景，“传输速率简介” 章节为 PCIe 理论计算数据，其余章节为具体性能指标及其测试方法。

1.2 PCIe sysfs

Linux PCIe 支持 sysfs 接口，主要提供以下资源：

```
/sys/devices/pci0000:17
|-- 0000:17:00.0
|   |-- class
|   |-- config
|   |-- device
|   |-- enable
|   |-- irq
|   |-- local_cpus
|   |-- remove
|   |-- resource
|   |-- resource0
|   |-- resource1
|   |-- resource2
|   |-- resource2_wc        # 64bits-pref mem resource 并支持写缓冲，对于 PCIe mem
to dev 操作有优化效果。
|   |-- revision
|   |-- rom
|   |-- subsystem_device
|   |-- subsystem_vendor
|   `-- vendor
`-- ...
```

详细参考：

2. 传输速率简介

2.1 PCIe 带宽

PCIe Generation	编码方案	传输速率	x1 Lane	x2 Lane	x4 Lane	x8 Lane
1.0	8b/10b	2.5GT/s	250MB/s	500MB/s	1GB/s	2GB/s
2.0	8b/10b	5GT/s	500MB/s	1GB/s	2GB/s	4GB/s
3.0	128b/130b	8GT/s	984.6MB/s	1.969GB/s	3.938GB/s	7.877GB/s
4.0	128b/130b	16GT/s	1.969GB/s	3.938GB/s	7.877GB/s	15.754GB/s

传输速率

传输速率为每秒传输量GT/s，而不是每秒位数Gbps，因为传输量包括不提供额外吞吐量的开销位； 比如PCIe 1.x和PCIe 2.x使用8b / 10b编码方案，导致占用了20% （2/10）的原始信道带宽。

GT/s： Giga transation per second （千兆传输/秒），即每一秒内传输的次数

Gbps： Giga Bits Per Second （千兆位/秒）。GT/s 与Gbps 之间不存在成比例的换算关系

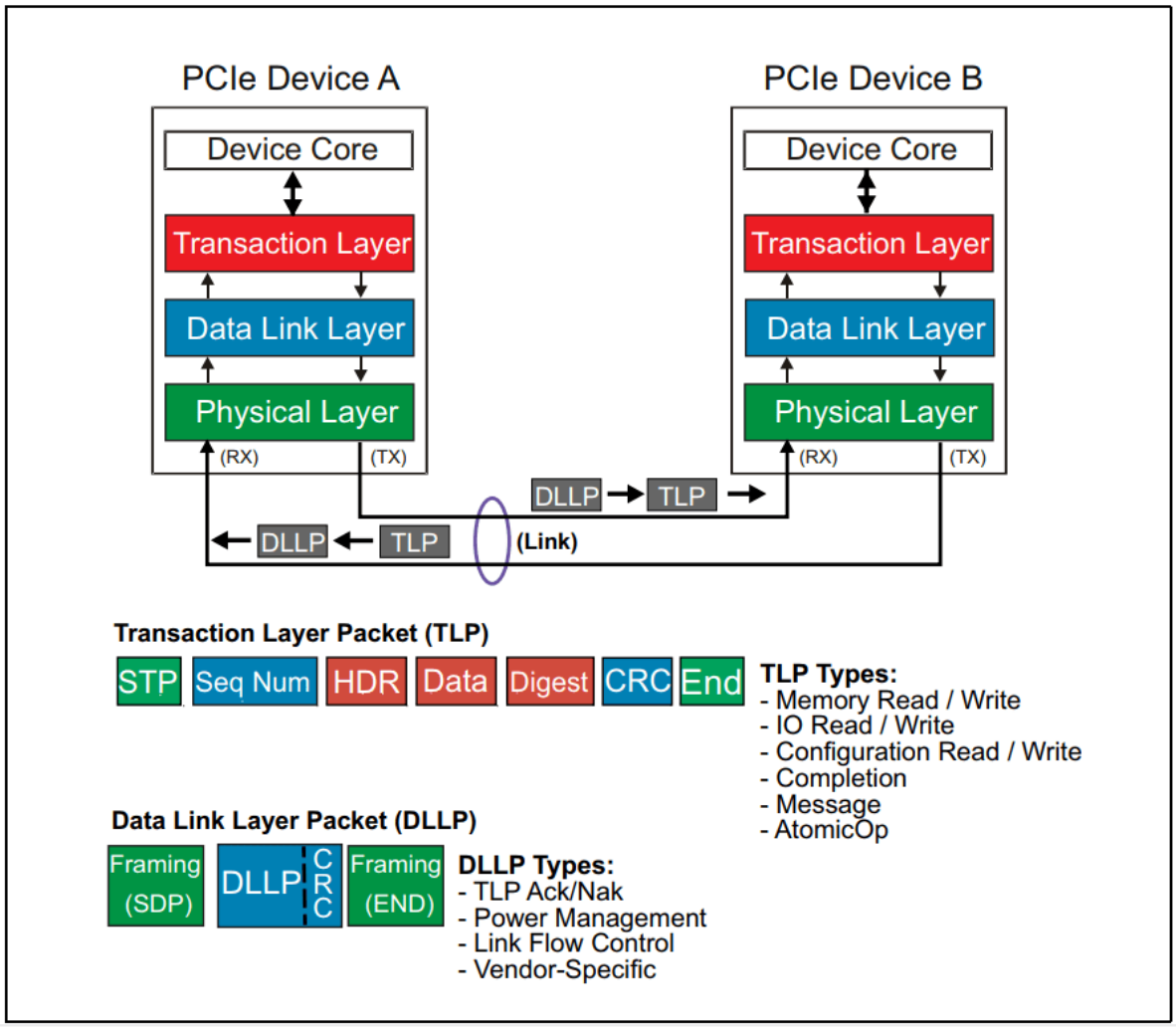
PCIe 可用带宽

吞吐量 = 传输速率 * 编码方案

例如：PCIe 2.0 协议的每一条 Lane 支持 5*8 / 10 = 4 Gbps = 500 MB/s 的速率，PCIe 2.0 x8的通道为例，x8 的可用带宽为 4*8 = 32 Gbps = 4 GB/s。

2.2 PCIe TLP 损耗

由于数据经过 Transaction Layer 和 Data Link Layer 后会添加打包信息，例如数据包的包头、包尾（ LCRC 和 ECRC 等）、数据链路层添加的包编号等等，所以 Data Payload 的真实速率相较于带宽会更低。



28 Byte TLP overhead 是冗余最大的传输包，包含 the header (16 bytes), the optional ECRC (4 bytes), the LCRC (4 bytes), the sequence number (2 bytes) and the framing Symbols STP and END (2 bytes).

所以以 Gen3 PCIe 带宽为例：

PCIe Generation	编码方案	传输速率	x1 Lane	x2 Lane	x4 Lane	x8 Lane
3.0	128b/130b	8GT/s	984.6MB/s	1.969GB/s	3.938GB/s	7.877GB/s

计入 TLP overhead 后：

PCIe Generation Gen3	x1 Lane	x2 Lane	x4 Lane	x8 Lane
PCIe 带宽（未计入 TLP overhead）	984.6MB/s	1.969GB/s	3.938GB/s	7.877GB/s
256B payload 扣除 TLP 损耗（256/284）	887.5MB/s	1.775GB/s	3.550GB/s	7.1002GB/s
128B payload 扣除 TLP 损耗（128/156）	807.8MB/s	1.616GB/s	3.232GB/s	6.464GB/s

2.3 PCIe 应用场景传输速率

如果再考虑用于 Ack/Nak 和 Flow Control 等的 DLLP 和用于链路训练和 Skip 的 Order Sets 等不定因素对速率的影响，实际真实数据传输速率会更低。

具体应用场景下以 eDMA 传输为最优解。

3. 控制器 - DMA Interface 吞吐率 - 芯片互联测试

3.1 理论速率

DMA 传输速率理论上接近 "PCIe TLP 损耗" 计入后的速率，理论值参考 “PCIe TLP 损耗” 章节说明。

3.2 测试方法

DMA 测试基于 “芯片互联测试” 测试，详细内容参考 《Rockchip_Developer_Guide_PCIe_CN.pdf》 文中 “芯片互联功能”，以 EP 写数据测试来测量 DMA 速率，其特点为：

- local EP memory to remote RC memory

特殊配置：

1. CPU 和 DDR 设定为高性能模式
2. 由于默认 mps 默认配置为 128B，可以修改 mps 为 256B 以降低 TLP 损耗，详细参考附录 “PCIe MPS 补丁参考”，如果测试使用 RK 设备互联，则 RC/EP 都应该添加补丁

3.3 测试结果

3.3.1 RK3568

测试设备

RC：RK3568-EVB1-DDR4-V10 PCIe 3.0 2 lanes ARM：1.99G DDR：1.5G

- 测试命令：./data/test-pcie-ep-new 500 2048 0 1006632960

EP：RK3568-IOTEST-DDR3-V10 PCIe 3.0 2 lanes ARM: 1.99G DDR：1.0G

- 测试命令（读 1000 loops x 2047MB 数据为例）：./data/test-pcie-ep-new 500 2048 0 1006632960 & ./data/test-pcie 1 1000 2047 0 0

测试结果

DMA 大小	256KB	512KB	1024KB	2048KB	4096KB
传输速率	430MB/s	850MB/s	1.3GB/s	1.5GB/s	1.5GB/s
传输时间（ms）	600	580	680	1320	2820

3.3.2 RK3588

RC：RK3588-EVB1-LP4X-V10 PCIe 3.0 4 lanes 大核 2.4G 小核 1.8G

- 测试命令：./data/test-pcie-ep-new 500 2048 0 1006632960

EP：RK3588-EVB4-LP4X-V10 PCIe 3.0 4 lanes 大核 2.4G 小核 1.8G

- 测试命令（读 1000 loops x 2047MB 数据为例）：./data/test-pcie-ep-new 500 2048 0 1006632960 & ./data/test-pcie 1 1000 2047 0 0

DMA 大小	256KB	512KB	1024KB	2048KB	4096KB
传输速率	428MB/s	875MB/s	1.58GB/s	3.07GB/s	3.34GB/s
传输时间/ms	584	688	590	635	1167

4. 控制器 - Bar Interface 吞吐率

4.1 简介

通过 Linux 提供的 PCIe sysfs 的 pci_mmap_resource_wc 接口做读写测试。

4.2 测试 APP

APP 参考附录 “resource_mmap_test.c” 源码，要求确认源码内以下参数：

- 选择 Write combine mmap Resource，例如：resource2_wc

APP 编译方式参考，以 RK3588 Android 为例：

```
/home1/ldq/rk-linux/prebuilts/gcc/linux-x86/aarch64/gcc-arm-10.2-2020.11-x86_64-aarch64-none-linux-gnu/bin/aarch64-none-linux-gnu-gcc -mcpu=cortex-a76 -O3 -static -DROPT -o resource_mmap_test resource_mmap_test.c
```

4.3 测试结果

4.3.1 RK3588

RC：RK3588-EVB1-LP4X-V10 PCIe 3.0 4 lanes 大核 2.4G 小核 1.8G

- 传输命令：./data/resource_mmap_test

EP：RK3588-EVB4-LP4X-V10 PCIe 3.0 4 lanes 大核 2.4G 小核 1.8G

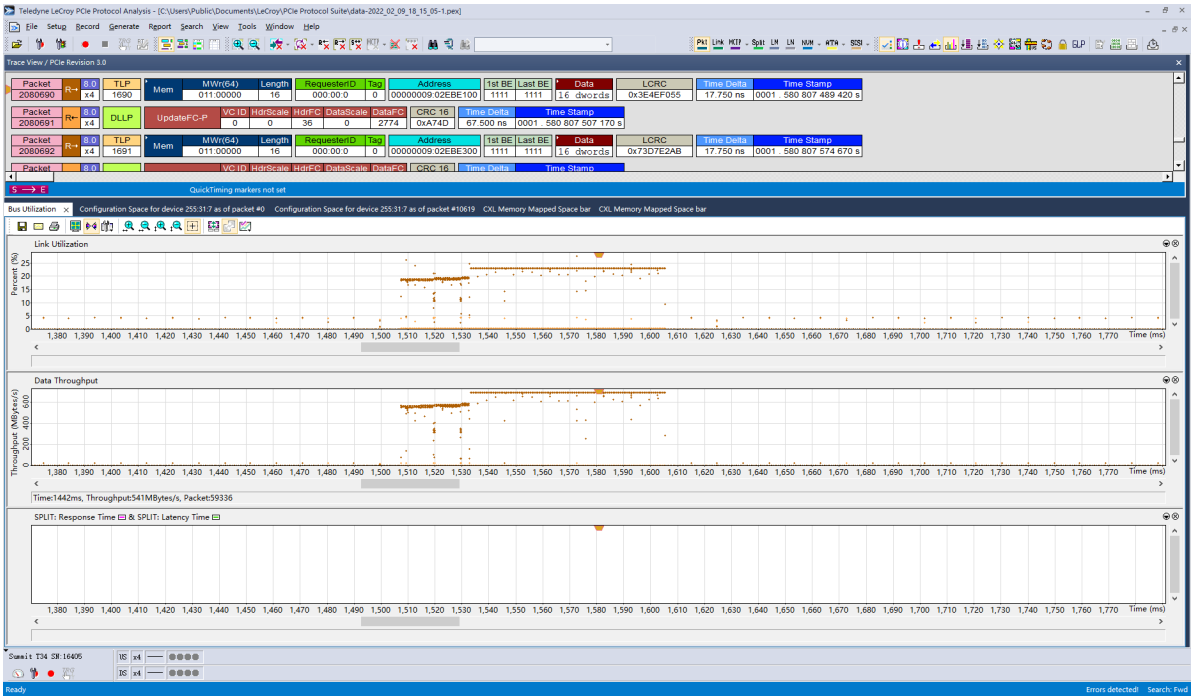
传输速率(MB/s)	写	读
64bits-pref (resource0_wc..N_wc)	680	24.9
64bits-pref (resource0..N)	13.3	16.3
32bits-np (resource0..N)	13.3	not support

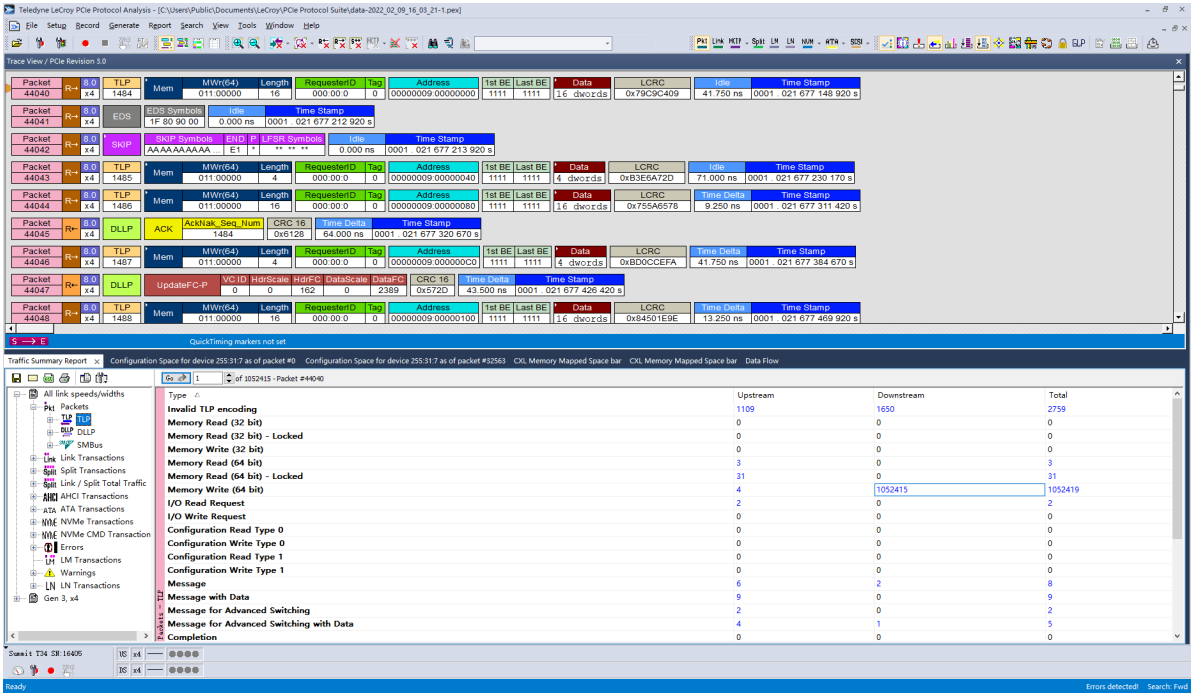
备注:

- freq 定频到最高，写速率提升不明显，为 682.761MB/s，读无提升
- 如果使用 resource2（非 write combine）：写 209.358MB/s、读 15.717MB/s

4.4 测试结果分析

写数据:

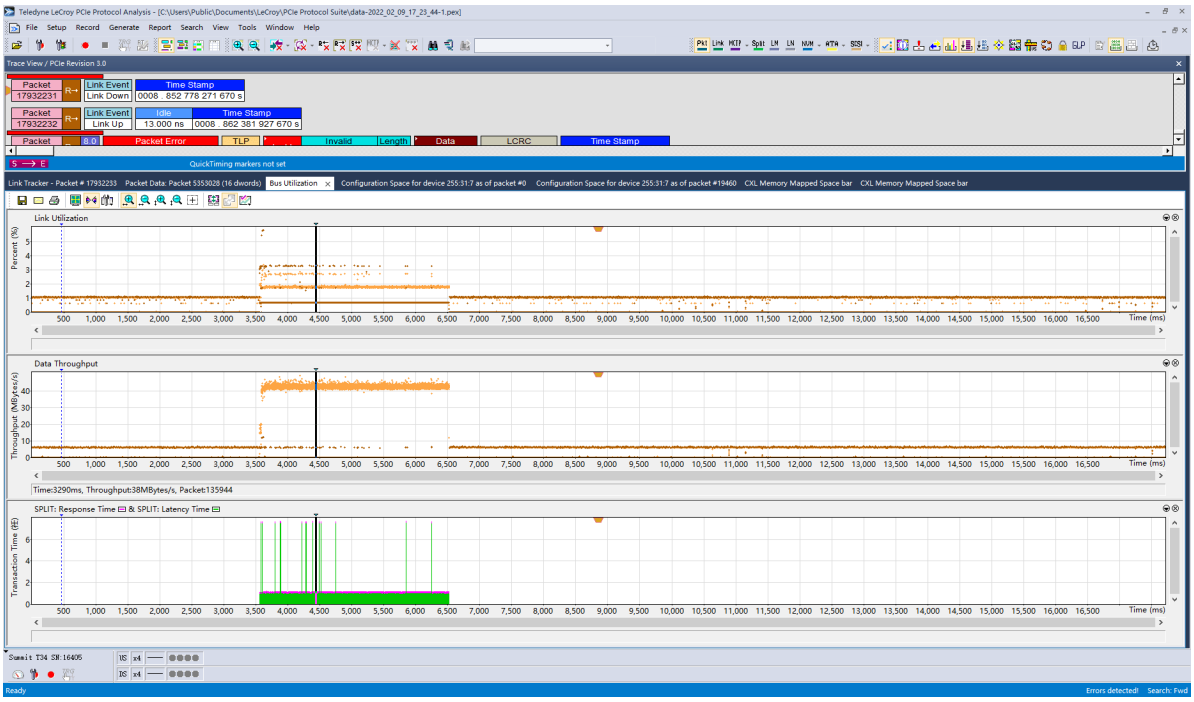


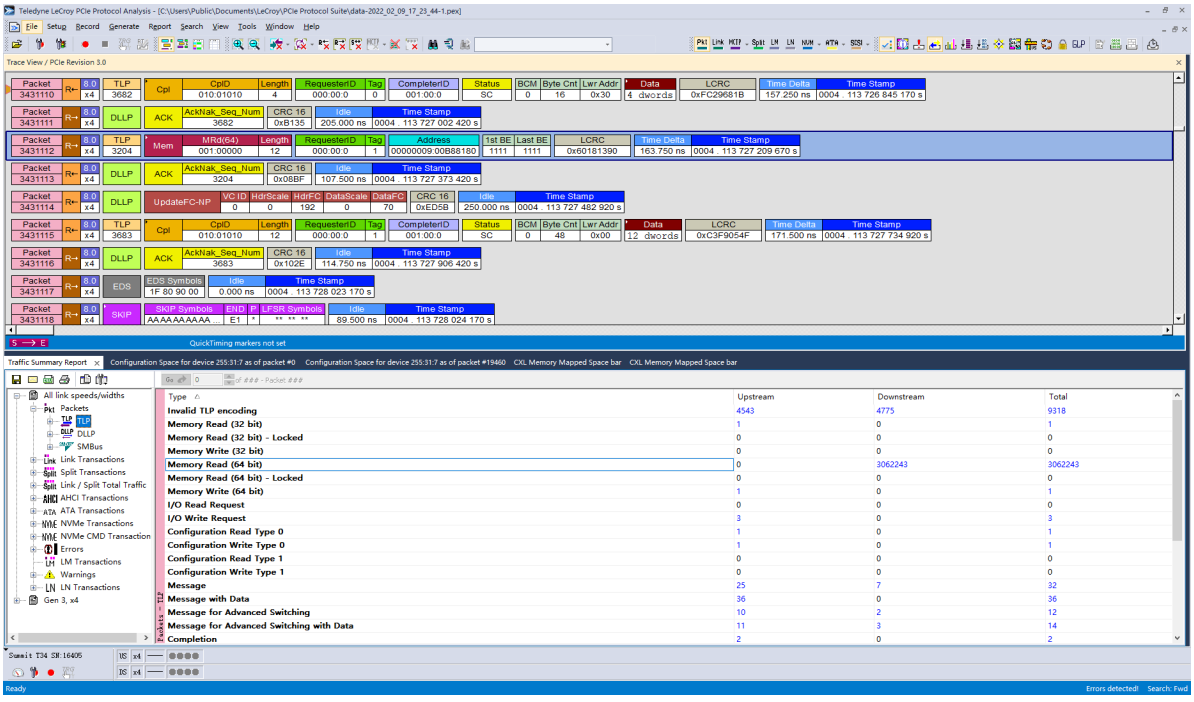


说明：

- 吞吐率大部分在 600MB/s
- Meory Write（64 bit）tlp count 1052415，平均 64B per tlp

读数据：





说明：

- 吞吐率大部分在 40MB/s
- Meory Read（64 bit）tlp count 3062243，平均 22B per tlp

4.5 性能优化方向

4.5.1 多线程测试

多个 resource_mmap_test.c 同时运行，读写不同 Bar，通过逻辑分析仪确认，总线利用率没有提高。

对于实际带宽没有显著提升（不超过 10%）,带宽利用率也没有提升，但是确实看到 TLP 不同 Bar 空间交替的情况。

5. 外设 - NVME 吞吐率

5.1 测试方法

测试方法详细参考《Rockchip_Developer_Guide_NVME.pdf》中 “性能评估” 章节，该文档目前为内部文档，如有需要请联系内部工程师获取。

5.2 测试结果

5.2.1 RK3588 Gen3x4

主机简介

- RK3588 EVB1 pcie3x4 主控 Port0/1 x4 RC
- 定频大核 2.4G 小核 1.8G、mq-deadline、f2fs nobarrier、LPDDR4X, 2112MHz

外设

三星980 M.2 NVMe SSD 1TBPCIe Gen3x4

PC CrystalDiskMark 测试结果：

PC crystalmark 8.0.1	**传输速率MB/s**	**IOPS（K）**
SEQ1M Q32T1 READ	3222	\
SEQ1M Q32T1 WRITE	2600	\
RND4K Q32T1 READ	708	\
RND4K Q32T1 WRITE	486	\

RK3588 fio 模拟 crystalmark 8.0.1测试结果

APP 测试名	传输速率MB/s	IOPS（K）
SEQ1M Q32T1 READ	2514.2	\
SEQ1M Q32T1 WRITE	1342.1	\
RND4K Q32T1 READ	556	143
RND4K Q32T1 WRITE	436	112

5.2.2 RK3588 Gen3x2

主机简介

- RK3588 EVB1 pcie3x4 主控 Port0 x2 RC
- 定频大核 2.4G 小核 1.8G、mq-deadline、f2fs nobarrier、LPDDR4X, 2112MHz

外设

PHISON 128G BGA SSD

PC CrystalDiskMark 测试结果：

crystalmark 8.0.1	传输速率MB/s	IOPS（K）
SEQ1M Q32T1 READ	1615	\
RND4K Q32T1 READ	434	\
SEQ1M Q32T1 WRITE	626	\
RND4K Q32T1 WRITE	230	\

RK3588 fio 模拟 crystalmark 8.0.1 测试结果

APP 测试名	传输速率MB/s	IOPS（K）
SEQ1M Q32T1 READ	1572.4	\
SEQ1M Q32T1 WRITE	600	\
RND4K Q32T1 READ	422	101
RND4K Q32T1 WRITE	355	90

5.2.3 RK3588s Gen2x1

主机简介

- RK3588s EVB1 pcie2x1l1主控 Port0 x1 RC
- 定频大核 2.4G 小核 1.8G、mq-deadline、f2fs nobarrier、LPDDR4X, 2112MHz

外设

PHISON 128G BGA SSD。

RK3588 fio 模拟 crystalmark 8.0.1 测试结果

APP 测试名	传输速率MB/s	IOPS（K）
SEQ1M Q32T1 READ	407	\
SEQ1M Q32T1 WRITE	350	\
RND4K Q32T1 READ	400	102
RND4K Q32T1 WRITE	134	32

6. 附录

6.1 PCIe MPS 补丁参考

```
diff --git a/drivers/pci/controller/dwc/pcie-dw-rockchip.c
b/drivers/pci/controller/dwc/pcie-dw-rockchip.c
index bab81e368c98..aabef18883e3 100644
--- a/drivers/pci/controller/dwc/pcie-dw-rockchip.c
+++ b/drivers/pci/controller/dwc/pcie-dw-rockchip.c
@@ -1718,7 +1719,7 @@ static int rk_pcie_really_probe(void *p)
        enum rk_pcie_device_mode mode;
        struct device_node *np = pdev->dev.of_node;
        u32 val = 0;
-       int irq;
+       int irq, reg;

        match = of_match_device(rk_pcie_of_match, dev);
        if (!match) {
@@ -1866,6 +1867,15 @@ static int rk_pcie_really_probe(void *p)
                rk_pcie->is_signal_test = true;
        }

+       /* Set MPS 256B */
+       reg = dw_pcie_find_capability(rk_pcie->pci, PCI_CAP_ID_EXP);
+       val = dw_pcie_readl_dbi(rk_pcie->pci, reg + PCI_EXP_DEVCTL);
+       val &= ~PCI_EXP_DEVCTL_PAYLOAD;
+       val |= (1 << 5);
+       dw_pcie_writel_dbi(rk_pcie->pci, reg + PCI_EXP_DEVCTL, val);
+
        /* Skip waiting for training to pass in system PM routine */
        if (device_property_read_bool(dev, "rockchip,skip-scan-in-resume"))
                rk_pcie->skip_scan_in_resume = true;
```