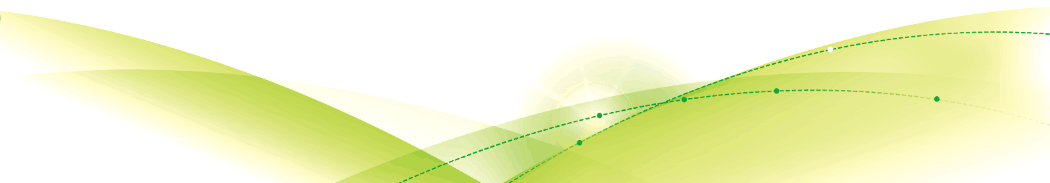


[编号ODCC-0500X-201X]

无损网络测试规范

V1.0



开放数据中心委员会 2017-08-22 发布

目 录

[1. 测试环境 4](#_Toc514751692)

[1.1. 测试网络拓扑 4](#_Toc514751693)

[1.2. 配合设备 4](#_Toc514751694)

[1.3. 测试工具 5](#_Toc514751695)

[1.4. 业务场景 5](#_Toc514751696)

[2. 测试项目 6](#_Toc514751697)

[2.1. 交换机裸性能：吞吐量、时延、丢包率测试 6](#_Toc514751698)

[2.2. 无损网络HPC业务性能测试 8](#_Toc514751699)

[2.2.1. 无损网络&InfiniBand网络测试 8](#_Toc514751700)

[2.2.2. 与传统Ethernet网络对比测试 11](#_Toc514751701)

[2.3. 无损网络DPDK性能测试 15](#_Toc514751702)

[2.3.1. DPDK无丢包转发性能测试 15](#_Toc514751703)

[2.4. 无损网络虚拟化场景性能测试 16](#_Toc514751704)

[2.4.1. 虚拟化SR-IOV场景下的带宽和延迟性能测试 16](#_Toc514751705)

[2.4.2. 虚拟化SR-IOV场景下的DPDK吞吐性能和不丢包测试 19](#_Toc514751706)

[2.5. 无损网络传统存储业务性能测试 20](#_Toc514751707)

[2.5.1. iSCSI over TCP/RDMA随机读写性能对比测试(FIO) 20](#_Toc514751708)

[2.5.2. iSCSI over TCP/RDMA OLTP模型读写性能对比测试(VDBENCH) 22](#_Toc514751709)

[2.5.3. iSCSI over TCP/RDMA OLAP模型读写性能对比测试(VDBENCH) 24](#_Toc514751710)

[2.6. 无损网络NVMF（NVMe over Fabric）业务性能测试 26](#_Toc514751711)

[2.6.1. 无损网络&InfiniBand网络测试 27](#_Toc514751712)

[2.6.2. 与传统Ethernet网络对比测试 31](#_Toc514751713)

[2.7. 无损网络分布式数据处理业务性能测试 35](#_Toc514751714)

[2.7.1. 高性能Spark HiBench排序性能测试 35](#_Toc514751715)

**前 言**

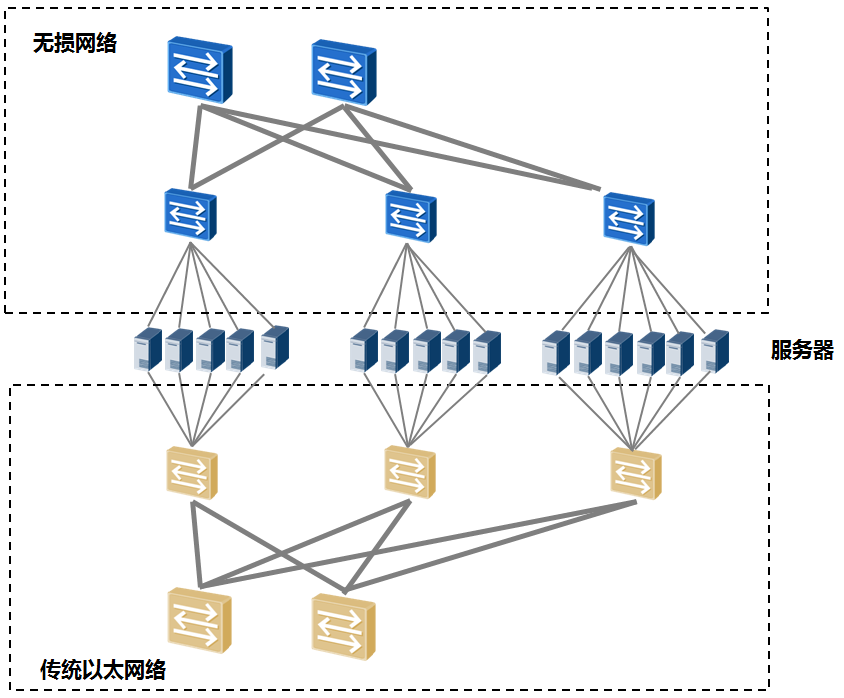
本标准规定了数据中心的无损网络测试规范。本标准是数据中心xxx产品商用部署所需要遵从的技术文件。

本规范的起草感谢以下单位（排名不分先后）：

**无损网络测试规范**

# 测试环境

## 测试网络拓扑



图表 1网络拓扑图

## 配合设备

|  |  |  |  |
| --- | --- | --- | --- |
| 配合设备名称 | 数量 | 单位 | 备注 |
| InfiniBand交换机 | 5 | 台 |  |
| 无损网络交换机 | 5 | 台 |  |
| Ethernet交换机 | 5 | 台 |  |
| 服务器/CPU/内存 | 16 | 台 | 测试HPC应用和SPARK，要求配置尽量高 |
| AOC | 若干 | 根 |  |
| NVME | 若干 |  | 如果没有的话就用NULL\_BLK设备 |

图表 2配合设备列表

## 测试工具

|  |  |
| --- | --- |
| 测试工具名称 | 备注 |
| FIO | 存储性能测试软件 |
| OpenFoam | HPC性能测试软件，V1606+ 版本 |
| OFED | 版本3.4-1.0.0.0，背景流产生命令Ib\_send\_bw |
| IMB\OSU | OSU Micro-Benchmarks 5.4.2  Intel(R) MPI Benchmarks 2018 Update 1 |

图表 3测试工具列表

## 业务场景

**无损网络旨在为主流业务所用的数据中心网络提供时延无损、丢包无损、吞吐无损。**当前这些业务对于数据中心网络的性能提出了更高的要求，未来网络需要更高带宽、更低时延、更大吞吐。本测试针对当前主流业务，例如高**性能数据库、大数据处理平台、HPC业务、虚拟化场景**等，开展了数据中心无损网络的各项测试。希望能为高性能、低时延、虚拟化场景下的业务提供更好的数据中心网络。

# 测试项目

## 交换机裸性能：吞吐量、时延、丢包率测试

|  |
| --- |
|  |
| 测试项目：交换机时延测试 |
| 测试目的：以太网络和IB网络交换机的时延测试 |
| 测试配置：分别进行Ethernet和IB组网及配置，测试交换机延时 |
| 网络拓扑：单台交换机组网测试。如下图为例：  Server 1  Server 2 |
| 测试步骤：  1、配置两台服务器，使用双端口网卡。每个网卡的两个端口，端口1连到交换机，端口2直连对端服务器。对交换机做简单配置，保证连通性。  2、测试通过端口1的IB\_write\_latency延时，以及通过端口2的对应延时。  3、记录平均时延结果。 |
| 预期结果：   1. IB交换机延时增加在100ns以内； 2. 以太交换机延时300ns左右。 |
| 备注： |

|  |
| --- |
|  |
| 测试项目：带宽测试 |
| 测试目的：以太网络和IB网络交换机的带宽测试 |
| 测试配置：分别进行Ethernet和IB组网及配置，测试交换机端口带宽 |
| 网络拓扑：单台交换机组网测试。如下图为例：  Server 1  Server 2 |
| 测试步骤：  1、配置两台服务器，使用双端口网卡。每个网卡的两个端口，端口1连到交换机，端口2直连对端服务器。对交换机做简单配置，保证连通性。  2、测试通过端口1的IB\_write\_bandwidth延时，以及通过端口2的对应带宽。  3、记录平均带宽结果。 |
| 预期结果：   1. IB交换机带宽为25Gb线速； 2. 以太交换机带宽接近25Gb线速。 |
| 备注： |

## 无损网络HPC业务性能测试

### 无损网络&InfiniBand网络测试

|  |
| --- |
|  |
| 测试项目：规模对比测试 |
| 测试目的：不同网络规模下无损网络和IB网络的性能测试 |
| 测试配置：分别进行无损和IB无阻塞组网，网络拓扑、端口速率和节点配置均一致。测试不同规模下，两种组网下OpenFoam测试用例motorBike的运行时间。 |
| 网络拓扑：以两级CLOS、无阻塞组网，服务器规模16台    网络拓扑图 |
| 测试步骤： **1**、在每台服务器（Huawei 2288 v3，操作系统redhat7.1）上分别安装OpenFoam（版本V1606+），OFED（版本3.4-1.0.0.0），并配置所有节点间无密码互相登录。  **2**、任选一台服务器作为master（上图中最左边一台），其余服务器作为slave。16台时IB和无损组网配置分别如下。  3、执行前处理；  Master上执行./run1.sh，完成前处理。  4、执行运算；  Master上执行./run2.sh，进行并行计算  5、记录执行时间。 |
| 预期结果：无损网络交换机与IB交换机均无丢包，任务完成时间相当。 |
| 备注： |

|  |
| --- |
|  |
| 测试项目：有背景流测试 |
| 测试目的：有背景流时，无损网络和IB网络的性能测试 |
| 测试配置：分别进行无损和IB无阻塞组网，网络拓扑、端口速率和节点配置均一致。测试增加背景流时，两种组网下OpenFoam测试用例motorBike的运行时间。 |
| 网络拓扑：以两级CLOS、16台服务器组网为例 |
| 测试步骤：  **1**、在每台服务器（Huawei 2288 v3，操作系统redhat7.1）上分别安装OpenFoam（版本V1606+），OFED（版本3.4-1.0.0.0），并配置所有节点间无密码互相登录。  **2**、任选一台服务器作为master（上图中最左边一台），其余服务器作为slave。16台时IB和无损组网分别进行配置。  3、添加背景流；  4、执行前处理；  5、执行运算；  6、记录执行时间； |
| 预期结果：无损网络交换机与IB交换机均无丢包，任务完成时间相当。 |
| 备注： |

### 与传统Ethernet网络对比测试

|  |
| --- |
|  |
| 测试项目：规模对比测试 |
| 测试目的：对比不同网络规模下无损网络和传统ETHERNET网络的性能 |
| 测试配置：分别进行无损和传统Ethernet无阻塞组网，网络拓扑、端口速率和节点配置均一致。测试不同规模下，两种组网下OpenFoam测试用例motorBike的运行时间。 |
| 网络拓扑：以两级CLOS、无阻塞组网，服务器规模是16台 |
| 测试步骤：  **1**、在每台服务器（Huawei 2288 v3，操作系统RedHat 7.1）上分别安装OpenFoam（版本V1606+），OFED（版本3.4-1.0.0.0），并配置所有节点间无密码互相登录。  **2**、任选一台服务器作为master（上图中最左边一台），其余服务器作为slave。1/2/4/8/16台时传统Ethernet和无损组网分别配置。  3、执行前处理；  4、执行运算；  5、记录执行时间。 |
| 预期结果：二者均无丢包，无损网络交换机任务完成时间较低 |
| 备注： |

|  |
| --- |
|  |
| 测试项目：有背景流测试 |
| 测试目的：对比有背景流时，无损网络和传统Ethernet网络的性能 |
| 测试配置：分别进行无损和传统Ethernet无阻塞组网，网络拓扑、端口速率和节点配置均一致。测试增加背景流时，两种组网下OpenFoam测试用例motorBike的运行时间。 |
| 网络拓扑：以两级CLOS、16台服务器组网为例 |
| 测试步骤：  **1**、在每台服务器（Huawei 2288 v3，操作系统RedHat 7.1）上分别安装OpenFoam（版本V1606+），OFED（版本3.4-1.0.0.0），并配置所有节点间无密码互相登录。  **2**、任选一台服务器作为master（上图中最左边一台），其余服务器作为slave。16台时传统Ethernet和无损组网分别配置。  3、添加背景流；  4、执行前处理；  5、执行运算；  6、记录执行时间。 |
| 预期结果：传统交换机存在丢包可能性，无损网络交换机不丢包，且任务完成时间少 |
| 备注： |

## 无损网络DPDK性能测试

### DPDK无丢包转发性能测试

|  |
| --- |
|  |
| 测试项目：DPDK无丢包转发性能测试 |
| 测试目的：无损网络DPDK转发性能 |
| 测试配置：   1. 服务器安装linux操作系统，安装官方网站的DPDK包 2. 测试仪分别连接被测试服务器网卡的各端口 |
| 网络拓扑： |
| 测试步骤：   1. 检查DPDK的PMD driver依赖包是否与DPDK官方网站的一致 2. 启动DPDK自带的L3fwd程序 3. 通过测试仪持续生成64B的数据包，数据包通过L3fwd到同网卡的同端口发送回测试仪 4. 通过测试仪表检测在零丢包情况的网络吞吐，记录转发性能数据 5. 将数据包大小换为128B、256B以及512B，分别重复执行上述3、4步骤 |
| 预期结果：  记录网卡在不同大小数据包的转发性能数据，在各种包长情况下，转发能力均可达到预期25GE的线速速率 |
| 备注： |

## 无损网络虚拟化场景性能测试

### 虚拟化SR-IOV场景下的带宽和延迟性能测试

|  |
| --- |
|  |
| 测试项目：时延测试 |
| 测试目的：虚拟化场景下，以太网络和IB网络交换机的时延测试 |
| 测试配置：   1. 服务器网卡开启SRIOV，分别对网卡的其中一个端口配置一定数量VF 2. 启动虚机，虚拟机配置网卡，通过SRIOV连接到物理网卡的其中一个VF 3. 分别进行Ethernet和IB组网及配置，测试通信延时 |
| 网络拓扑：单台交换机组网测试。如下图为例：  Server 1 VM1 VF1  Server 2 VM1 VF1 |
| 测试步骤：  1、配置两台服务器，使用双端口网卡。每个网卡的两个端口，端口1连到交换机，端口2直连对端服务器。对交换机做简单配置，保证连通性。  2、测试通过端口1的VF IB\_write\_latency延时，以及通过端口2的VF对应延时。  3、记录平均时延结果。 |
| 预期结果：   1. IB交换机延时增加在100ns以内； 2. 以太交换机延时300ns左右。 |
| 备注： |

|  |
| --- |
|  |
| 测试项目：带宽测试 |
| 测试目的：虚拟化场景下，以太网络和IB网络交换机的带宽测试 |
| 测试配置：   1. 服务器网卡开启SRIOV，分别对网卡的其中一个端口配置一定数量VF 2. 启动虚机，虚拟机配置网卡，通过SRIOV连接到物理网卡的其中一个VF 3. 分别进行Ethernet和IB组网及配置，测试通信带宽 |
| 网络拓扑：单台交换机组网测试。如下图为例：  Server 1 VM1 VF1  Server 2 VM1 VF1 |
| 测试步骤：   1. 配置两台服务器，使用双端口网卡。每个网卡的两个端口，端口1连到交换机，端口2直连对端服务器。对交换机做简单配置，保证连通性。 2. 测试通过端口1的VF IB\_write\_bandwidth带宽，以及通过端口2的VF对应带宽。 3. 记录平均带宽结果。 |
| 预期结果：   1. IB交换机带宽为25Gb线速； 2. 以太交换机带宽接近25Gb线速。 |
| 备注： |

### 虚拟化SR-IOV场景下的DPDK吞吐性能和不丢包测试

|  |
| --- |
|  |
| 测试项目：虚拟化场景的DPDK不丢包测试 |
| 测试目的：测试被测设备虚拟化的转发性能 |
| 测试配置：   1. 服务器网卡开启SRIOV，分别对两个网卡的其中一个端口配置一定数量VF 2. 启动虚机，虚拟机配置两个网卡，分别通过SRIOV连接到两个不同物理网卡的其中一个VF 3. 测试仪分别连接被测试服务器的两个网卡 |
| 网络拓扑： |
| 测试步骤：   1. 检查DPDK的PMD driver依赖包是否与DPDK官方网站的一致 2. 在虚拟机内启动DPDK自带的testPMD程序 3. 通过测试仪持续生成64B的数据包，数据包通过一个网卡的VF到虚拟机的L3fwd到另一个网卡VF发送回测试仪。 4. 通过测试仪表检测在零丢包情况的网络吞吐，记录转发性能数据。 5. 将数据包大小换为128B、256B以及512B，分别重复执行上述3、4步骤。 |
| 预期结果：  记录网卡在使用SR-IoV在不同大小数据包时的转发性能数据，在各种包长情况下，转发能力均可接近线速速率 |
| 备注： |

## 无损网络传统存储业务性能测试

### iSCSI over TCP/RDMA随机读写性能对比测试(FIO)

|  |
| --- |
|  |
| 测试项目：8K数据块7：3测试 |
| 测试目的：测试存储系统8K数据块混合读写的性能 |
| 测试配置：   1. 系统环境搭建完成，处于正常运行状态 2. 每个卷100G,卷的数量不限，以能发挥最大性能为标准 3. 限制存储系统使用2个cpu 4. 测试卷已使用fio工具进行全盘初始化写 5. 计算节点已安装FIO测试工具，对裸盘进行性能测试 6. 以iSCSI模式挂载方式、两副本模式测试性能，分别进行无损和传统Ethernet无阻塞组网，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑：  单台交换机组网测试。如下图为例：  Server 1  Server 2 |
| 测试步骤：   1. 在计算节点上启动fio测试工具，设置测试模型为：8KB数据块大小，7：3比例混合读写。fio线程数（numjobs=1参数）可自行调整，以发挥存储最大性能为准。 2. 记录fio运行结束后输出的IOPS |
| 预期结果：   1. fio正常运行 2. 记录fio运行结束后输出的IOPS。 |
| 备注： |

|  |
| --- |
|  |
| 测试项目：1024k 数据块随机读写性能 |
| 测试目的：测试存储系统1024K数据块随机读写的性能 |
| 测试配置：   1. 系统环境搭建完成，处于正常运行状态 2. 每个卷100G,卷的数量不限，以能发挥最大性能为标准 3. 限制存储系统使用2个cpu 4. 测试卷已使用fio工具进行全盘初始化写 5. 计算节点已安装FIO测试工具，对裸盘进行性能测试。 6. 以iSCSI模式挂载方式、两副本模式测试性能，分别进行无损和传统Ethernet无阻塞组网，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑：  单台交换机组网测试。如下图为例：  Server 1  Server 2 |
| 测试步骤：   1. 在计算节点上启动fio测试工具，设置测试模型为：分别测试1024K随机读和写的性能。fio线程数（numjobs=1参数）可自行调整，以发挥存储最大性能为准。 2. 记录fio运行结束后输出的IOPS |
| 预期结果：   1. fio正常运行 2. 记录fio运行结束后输出的IOPS |
| 备注： |

### iSCSI over TCP/RDMA OLTP模型读写性能对比测试(VDBENCH)

|  |
| --- |
|  |
| 测试项目：OLTP模型读写性能测试 |
| 测试目的：测试被测存储系统OLTP模型读写性能 |
| 测试配置：   1. 按照测试组网图搭建测试环境 2. 每个卷100G,卷的数量不限，以能发挥最大性能为标准 3. 限制存储系统使用2个cpu 4. 被测产品运行正常 5. 存储系统中所有已划分空间按1024k块大小写满 6. 以iSCSI模式挂载方式、两副本模式测试性能，分别进行无损和传统Ethernet无阻塞组网，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑：  单台交换机组网测试。如下图为例：  Server 1  Server 2 |
| 测试步骤：   1. 在压力机上部署Vdbench测试工具。主机定义设置如下：   hd=default,vdbench=/home/vdbench,user=root,shell=ssh  hd=hd1,system=node1  hd=hd2,system=node2  hd=hd3,system=node3  ……   1. 设置测试工具模型如表6-1所示，测试数据量为块存储总SSD缓存大小的1.5倍除以副本数。在压力机上运行Vdbench测试工具。存储定义、工作负载定义、运行定义设置如下：   sd=sd1,hd=hd1,lun=/dev/sdb,openflags=o\_direct,threads=1（厂家自行调整）  ……  wd=wd1,sd=sd\*,seekpct=100,rdpct=100,xfersize=8k,skew=20  wd=wd2,sd=sd\*,seekpct=100,rd  pct=100,xfersize=4k,skew=45  wd=wd3,sd=sd\*,seekpct=100,rdpct=0,xfersize=8k,skew=15  wd=wd4,sd=sd\*,seekpct=0,rdpct=100,xfersize=64k,skew=10  wd=wd5,sd=sd\*,seekpct=0,rdpct=0,xfersize=64k,skew=10 |
| 预期结果：  测试过程无异常，记录平均读写IOPS |
| 备注： |

### iSCSI over TCP/RDMA OLAP模型读写性能对比测试(VDBENCH)

|  |
| --- |
|  |
| 测试项目：OLAP模型读写性能测试 |
| 测试目的：测试被测存储系统OLAP模型读写性能 |
| 测试配置：   1. 按照测试组网图搭建测试环境 2. 每个卷100G,卷的数量不限，以能发挥最大性能为标准 3. 限制存储系统使用2个cpu 4. 被测产品运行正常 5. 存储系统中所有已划分空间按1024k块大小写满 6. 以iSCSI模式挂载方式、两副本模式测试性能，分别进行无损和传统Ethernet无阻塞组网，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑：  单台交换机组网测试。如下图为例：  Server 1  Server 2 |
| 测试步骤：  测试步骤：   1. 在压力机上部署Vdbench测试工具。主机定义设置如下：   hd=default,vdbench=/home/vdbench,user=root,shell=ssh  hd=hd1,system=node1  hd=hd2,system=node2  hd=hd3,system=node3  ……   1. 设置测试工具模型如表6-1所示，测试数据量为块存储总SSD缓存大小的1.5倍除以副本数。在压力机上运行Vdbench测试工具。存储定义、工作负载定义、运行定义设置如下：   sd=sd1,hd=hd1,lun=/dev/sdb,openflags=o\_direct,threads=1（厂家自行调整）  ……  wd=wd1,sd=sd\*,seekpct=100,rdpct=100,xfersize=4k,skew=15  wd=wd2,sd=sd\*,seekpct=100,rdpct=0,xfersize=4k,skew=5  wd=wd3,sd=sd\*,seekpct=0,rdpct=100,xfersize=64k,skew=70  wd=wd4,sd=sd\*,seekpct=0,rdpct=0,xfersize=64k,skew=10  rd=rd1,wd=wd1,iorate=max,elapse=604800,maxdata=数据量,interval=5,warmup=30 |
| 预期结果：  测试过程无异常，记录平均读写IOPS |
| 备注： |

## 无损网络NVMF（NVMe over Fabric）业务性能测试

针对N**VM**F业务，为了对比测试不同网络规模下无损网络、 传统Ethernet网络、InfiniBand性能，需要分别进行无损、传统Ethernet、IB组网，在相同的网络拓扑、端口速率和节点配置情况下，测试NVMF业务IOPS和尾部时延性能。

首先配置target和client服务器，相应target和client建链；client访问target下的NVMe SSD，使用标准工具FIO进行测试，分别测试不同组网（多打一、多打多）、不同读写类型、不同块大小、不同队列深度、不同线程数、不同背景流等情况下的IOPS性能和尾部时延；最终统计测试结果并记录IOPS和尾部时延性能。也可根据用户需求对服务器CPU利用率进行比较，测试过程中，使用top命令观察，用nmon记录各核心占用率。

下面结合两个具体案例来详细说明无损网络与IB、传统Ethernet的性能测试。

### 无损网络&InfiniBand网络测试

|  |
| --- |
|  |
| 测试项目：NVMF读性能测试 |
| 测试目的：无损网络和IB网络的性能测试 |
| 测试配置： 分别进行无损和IB组网及配置，测试NOF业务IOPS和尾部时延性能，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑： CLOS网络，TOR下接服务器。如下图为例： |
| 测试步骤：  1、配置target和client服务器。例如可以采用华为RH2288v3服务器和Mellanox ConnectX-4网卡。安装linux操作系统，因需要支持NOF协议栈，需安装Ubuntu16.10或以上版本。target端需要创建和配置subsystem、namespace、port等，配置完成后client与相应target建链；建链完成后在client生成逻辑NVMe盘，可以使用标准存储测试工具FIO进行测试，使用FIO2.10或以上版本。  2、完成所有client和target建链后，使用标准工具FIO进行测试，client访问target下的NVMe SSD。本例测试2打2读操作场景（2个client和2个target服务器一一对应建链，client和target在不同的两个TOR，2个client同时访问其target下的NVMe SSD盘）、块大小128K、队列深度为1、线程数为256个、不加背景流；  3、统计FIO测试结果；  4、记录IOPS和尾部时延性能。 |
| 预期结果：  无损交换机和IB交换机均无丢包，二者IOPS相当，尾部时延相当。 |
| 备注： |

|  |
| --- |
|  |
| 测试项目：NOF写性能测试 |
| 测试目的：无损网络和IB网络的性能测试 |
| 测试配置： 分别进行无损和IB组网及配置，测试NOF业务IOPS和尾部时延性能，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑： CLOS网络，TOR下接服务器。如下图为例： |
| 测试步骤：  1、配置target和client服务器。  2、完成所有client和target建链后，使用标准工具FIO进行测试，client访问target下的NVMe SSD。  3、统计FIO测试结果；  4、记录IOPS和尾部时延性能。 |
| 预期结果：  1、配置target和client服务器。例如可以采用华为RH2288v3服务器和Mellanox ConnectX-4网卡。安装linux操作系统，因需要支持NOF协议栈，需安装Ubuntu16.10或以上版本。target端需要创建和配置subsystem、namespace、port等，配置完成后client与相应target建链；建链完成后在client生成逻辑NVMe盘，可以使用标准存储测试工具FIO进行测试，使用FIO2.10或以上版本。  2、完成所有client和target建链后，使用标准工具FIO进行测试，client访问target下的NVMe SSD。本例测试2打2读操作场景（2个client和2个target服务器一一对应建链，client和target在不同的两个TOR，2个client同时访问其target下的NVMe SSD盘）、块大小128K、队列深度为1、线程数为256个、不加背景流；  3、统计测试结果；  4、记录IOPS和尾部时延性能 |
| 预期结果：  无损交换机和IB交换机均无丢包，二者IOPS相当，尾部时延相当。 |
|  |
| 备注： |

### 与传统Ethernet网络对比测试

|  |
| --- |
|  |
| 测试项目：NOF读性能测试 |
| 测试目的：对比无损网络和传统Ethernet网络的性能 |
| 测试配置： 分别进行无损和传统Ethernet组网及配置，测试NOF业务IOPS和尾部时延性能，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑：CLOS网络，TOR下接服务器。如下图为例： |
|  |
| 测试步骤：  1、配置target和client服务器。  2、完成所有client和target建链后，使用标准工具FIO进行测试，client访问target下的NVMe SSD。  3、统计测试结果；  4、记录IOPS和尾部时延性能。 |
| 预期结果：无损网络交换机无丢包，IOPS有大幅提升，尾部时延低。 |
| 备注： |

|  |
| --- |
|  |
| 测试项目：NOF写性能测试 |
| 测试目的：对比无损网络和传统Ethernet网络的性能 |
| 测试配置： 分别进行无损和传统Ethernet组网及配置，测试NOF业务IOPS和尾部时延性能，网络拓扑、端口速率和节点配置均一致。 |
| 网络拓扑：CLOS网络，TOR下接服务器。如下图为例： |
|  |
| 测试步骤：  1、配置target和client服务器。  2、完成所有client和target建链后，使用标准工具FIO进行测试，client访问target下的NVMe SSD。  3、统计测试结果；  4、记录IOPS和尾部时延性能。 |
| 预期结果：预期结果：无损网络交换机无丢包，IOPS有大幅提升，尾部时延低。 |
| 备注： |

## 无损网络分布式数据处理业务性能测试

### 高性能Spark HiBench排序性能测试

|  |
| --- |
|  |
| 测试项目：高性能Spark HiBench排序性能测试 |
| 测试目的：  对需要大量Spark节点间Block Fetch读写操作的大数据计算服务器集群，通过应用高性能无损网络，降低CPU利用率，并缩短数据传输和计算时间。 |
| 测试配置：   1. 采用8个服务器节点搭建服务器集群 2. 安装Spark, Hadoop, SparkRDMA, Hibench, Moven等开源软件 3. 启动Hadoop，确认HDFS工作 4. 分别进行无损和传统Ethernet组网及配置，网络拓扑、端口速率和节点配置均一致 |
| 网络拓扑：  CLOS网络，TOR下接服务器。如下图为例： |
|  |
| 测试步骤：   1. 环境准备，测试文件路径 2. 下载Spark(<http://www-us.apache.org/dist/spark/spark-2.2.0/spark-2.2.0-bin-hadoop2.7.tgz>), Hadoop (<http://www-us.apache.org/dist/hadoop/common/hadoop-2.7.4/hadoop-2.7.4.tar.gz>), Hibench (<https://github.com/intel-hadoop/HiBench/archive/master.zip>), SparkRDMA (<https://github.com/Mellanox/SparkRDMA/releases/download/v1.0/spark-rdma-1.0.tgz>)  and Maven (<http://www-us.apache.org/dist/maven/maven-3/3.5.2/binaries/apache-maven-3.5.2-bin.tar.gz>). 3. 解压对应的文件. 4. 将所有参与测试的节点名字写到hadoop/etc/hadoop/slaves 5. 设置免密ssh登录，检查其是否工作hadoop/sbin/slaves.sh hostname –f ，能看到如下输出:   host1: host1.example.com  host2: host2.example.com   1. mount tmpfs 到指定测试路径. 例如: /mnt/user/ramdisk 2. 运行:   *hadoop/sbin/slaves.sh mkdir /mnt/user/ramdisk*  *hadoop/sbin/slaves.sh sudo mount -t tmpfs -o size=100g tmpfs /mnt/user/ramdisk*   1. 按照对应的配置修改配置文件：core-site.xml;hadoop.tmp.dir需改为第6步设置的路径. spark-defaults.conf 里面的spark.local.dir，也需要修改 2. 将配置文件中的默认主机名，更改为测试主机名。 3. 运行*hadoop/sbin/start-all.sh*. 确保log中没有出现错误提示. 确定HDFS已启动: *hadoop/bin/hdfs dfsadmin –report。* 确定Yarn已启动: hadoop/bin/yarn node –list 4. 编译 HiBench:           cd Hibench && ../apache-maven/bin/mvn -Psparkbench -Dspark=2.2 -Dscala=2.11 clean package   1. Configure HiBench. Attached examples hadoop.conf, vanilla-spark.conf (to run vanilla spark benchmark), spark-rdma.conf. Also set in conf/workloads/micro/terasort/terasort.conf hibench.workload.datasize                       1750000000 2. 生成dataset: bin/workloads/micro/terasort/prepare/prepare.sh 3. 运行 terasort with vanilla spark: put vanilla-spark.conf to conf/spark.conf   *bin/workloads/micro/terasort/spark/run.sh*   1. 将 配置文件conf/spark.conf 替换为 spark-rdma.conf. 再次运行 benchmark :   *bin/workloads/micro/terasort/spark/run.sh*   1. 在如下路径可以看到测试结果报告: HiBench/report/hibench.report |
| 预期结果：   1. 记录非RDMA时的CPU利用率和计算延时 2. 记录使用RDMA的CPU利用率和计算延时 |
| 备注： |