Lustre arm64 初探

Lustre 是 HPC 中最为广泛应用的存储后端，经常活跃在 IO500 榜单中。随着 HPC 国产化的推进，Arm64 服务器在 HPC 后端存储获得了广泛应用到机会。为了能够更好的推动 Lustre 在 Arm64 平台上的广泛应用，Huawei，Linaro 和 Lustre 社区合作，推动 Lustre 服务端对 Arm64 的支持，同时发布了 Lustre Arm64 社区版本，维护 Arm64 CI。在此基础上，我们针对 Lustre 在 Arm64 平台进行了非常充分和完整的功能测试，性能测试和性能优化分析。希望通过这些工作，同社区同仁一道，能够更好的丰富 Arm64 的 HPC 生态。

1. lustre 版本和 CI

1.1. Lustre 社区对 Arm64 的支持情况

Lustre 客户端官方版本已经支持 ARM64，服务端也已经基本完成了对 Arm64 的支持，其版本信息维护在 Lustre 官网中。当前 Lustre 官网的 Arm64 的开发版本指向由 Linaro 所提供的维护的源中，具体链接请参考如下地址：https://uk.linaro.cloud/repo/lustre/。目前，社区的 Arm64 的 CI 和发行版均由 Linaro 负责维护，每日对 Lustre 最新的社区代码进行构建。在构建完成后，Arm64 的 CI 任务会对当前的 Lustre 版本进行集成测试。

Lustre 的测试集非常丰富，基本涵盖了所有产品测试场景。目前，在 Linaro 维护的 Arm64 CI 中，在日构建之后，均会进行 Lustre 多节点测试，测试集所有信息请参考链接。在 Lustre 使能和支持的过程中，我们基本解决了绝大多数 Arm64 上碰到的问题，主要集中在 64K 页的支持，Arm64 上的元数据对齐等问题，同时针对一些测试 case 在 Arm64 上进行了适配改造。目前，绝大多数测试 case 已经能够在 Arm64 上通过，仅有的少于 10 个测试用例正在修复中。

我们基于 openEuler 22.03LTS 版本做了 Lustre 适配正在上传 Lustre 上游社区，源码包仓库地址为 https://gitee.com/src-openeuler/lustre。技术 review 版本 RPM 下载路径 https://uk.linaro.cloud/repo/devel/lustre/v2.15/oe2203/

1.2. Lustre Arm64 社区版本计划

由 Linaro 提供和发布的 Lustre Arm64 版本目前基于 CentOS 8.5，内核版本为 4.18。目前，Arm64 的 CI 任务和版本测试均基于此内核和版本。后续跟 openEuler 社区商讨出 Lustre 版本，同时完成相应的测试 CI 覆盖。

后续我们会针对 Lustre 2.15 版本进行测试 CI 的覆盖和 RPM 构建，为社区提供一个基于 Arm64 的稳定版本。同时，也会针对增加根据网络情况，提供支持 InfiniBand(IB)的 Lustre 版本。目

前，针对 IB 的编译测试均已完成，有关步骤请参考 1.3 节。

1.3. Lustre 版本编译

Lustre 编译顺序是：先编译 Lustre 内核(可选)，安装重启后，编译安装 OFED 驱动，安装 zfs 和 e2fsprogs 包后再编译 Lustre 版本。

注意：

1）CentOS8

CentOS 8.4.2105 ARM64 IB 驱动的下载地址， e2fsprogs 包下载地址, zfs 请使用源码 zfs-2.1-release 编译

2）openEuler 22.03

openEuler 22.03 Arm64 IB 驱动包下载地址, e2fsprogs 包下载地址, zfs 请使用源码 zfs-2.1-release 编译

openEuler 22.03 lustre master 分支支持已经上传社区，v2.15 LTS 分支支持上传中， 源码 clone 地址分支 b2_15-openeuler （gitee 仓地址：https://gitee.com/xin3liang/lustre-src, https://gitee.com/xin3liang/lustre-e2fsprogs.git)。

3）openEuler 20.03/KylinOS V10

openEuler 20.03 lustre master 分支支持上传社区中，v2.15 LTS 源码 clone 地址分支 b2_15-openeuler（gitee 仓地址：https://gitee.com/xin3liang/lustre-release, https://gitee.com/xin3liang/lustre-e2fsprogs.git)。e2fsprogs 包下载地址, zfs 请使用源码 zfs-2.1-release 编译

Need to install 22.03's kmod scripts rpm

```
1. # For oe20.03
sudo dnf install https://repo.openeuler.org/openEuler-22.03-LTS-
SP2/everything/aarch64/Packages/kernel-rpm-macros-30-
35.oe2203sp2.aarch64.rpm

2. # For KylinOS V10
For KylinOS get kmodtool and macros.kmp from oe22.03.sp2+, or
here
https://uk.linaro.cloud/devel/tmp/

sudo sed -i 's/openEuler/kylin/g'
/usr/lib/rpm/macros.d/macros.kmp
sudo cp /usr/lib/rpm/openEuler/kmodtool
/usr/lib/rpm/kylin/kmodtool
sudo echo "%openEuler 2" >> /etc/rpm/macros.dist
```

安装编译环境依赖包：

```
sudo dnf -y groupinstall "Development Tools"
sudo dnf -y install audit-libs-devel binutils-devel elfutils-
devel kabi-dw ncurses-devel newt-devel numactl-devel openssl-
devel pciutils-devel perl perl-devel python3-docutils xmlto xz-
devel elfutils-libelf-devel libcap-devel libcap-ng-devel libyaml
libyaml-devel kernel-rpm-macros libblkid-devel libtirpc-devel
libnl3-devel mpich libmount-devel llvm-devel clang
sudo dnf install -y  gcc make autoconf automake libtool rpm-build
libtirpc-devel libblkid-devel libuuid-devel libudev-devel
openssl-devel zlib-devel libaio-devel libattr-devel kernel-devel-
$(uname -r) kernel-debugsource-$(uname -r) python3 python3-devel
python3-setuptools python3-cffi libffi-devel git ncompress
libcurl-devel
sudo dnf install -y  python3-packaging dkms
```

编译安装 Lustre patched kernel 包，然后重启机器（**可选，for rhel family OS only，只用于支持 SCSI T10-PI feature**）

编译参考“Lustre 编译安装 2）”

```
git clone -b zfs-2.1-release https://github.com/openzfs/zfs
cd zfs
sh autogen.sh && ./configure && make rpms
sudo dnf install ./*aarch64.rpm
```

IB 驱动的编译安装

```
./mlnxofedinstall --all --force --without-kmod-iser --without-xpmem-modules --without-libxpmem --add-kernel-support
```

pdsh 工具编译安装，pdsh 工具用于部署拉起 Lustre 集群 (for openEuler)

```
git clone https://github.com/chaos/pdsh
cd pdsh
./bootstrap
./configure --with-ssh
make
make install
```

Lustre 编译安装：

存在 ARM64、x86 混合部署，务必使用 4K 页。具体配置如下表：

处理器	2x 鲲鹏 920 5250 2.6GHz	x86 10 台
内存	512GB 2666MHz	376GB 2666MHz
网络	1x Infiniband ConnectX-5 100GE-双端口	
磁盘	4x ES3000 V6 3.2TB NVMe SSD	
操作系统	CentOS 8.4.2105/kernel-4.18.0-372.9.1.el8 使用 4K PAGESIZE openEuler 22.03 LTS/kernel-5.10.0-60.18.0.50	
lustre 版本	2.15.52	
OFED 版本	OFED-5.6-2.0.9	
zfs 版本	2.1.5-1	

2.3. 功能验证

2.3.1. 功能测试

功能测试由 IO500 测试、数据一致性测试两部分组成，具体如下：

- IO500 测试，在 10 客户端上并发 100~900 线程测试集群 12 小时，检查服务状态，不出现挂死，持续一周。
- 数据一致性测试，通过复制大量不同大小的文件到 Lustre 文件系统，重启 Lustre 集群后，再复制回来，检查前后的 md5 值是否一致。前后 md5 值是一致的，确认数据一致性没有问题。

基于 CentOS 8.4、openEuler 22.03 系统测试，上述功能测试均通过。

2.3.2. 混合部署

服务端 ARM64 与 x86 混合部署，客户端 ARM64 与 x86 混合跑 IO500 测试，无异常。

openmpi 测试需要确保打开异构计算支持，在 configure 里增加--enable-heterogeneous，编译出来的版本执行 ompi_info | grep Heterogeneous 查看，返回 yes 代表支持异构计算，才能混合跑 IO500 测试。

2.3.3. kvm 虚拟化

将 Lustre mds、oss、客户端用虚拟机部署，采用虚拟机绑核绑 numa、直通网卡、NVMe

盘等优化手段，跟物理机性能持平。在一台服务器上，使用主机+虚拟机分管资源，可以避免跨片数据流动，提升性能。用 IO500 对比测试物理机、主机+虚拟机部署方式，单服务端能提升 27%，单客户端提升 31%.元数据性能提升显著。

2.3.4. 故障切换

使用 NVMe-oF 远端盘，在主 MGS、MDS、OSS 上格式化文件系统，挂载 MGT、MDT、OST 后，拷贝一批文件到 lustre 文件系统，记录文件属性、md5 值。主卸载后，切换到备节点，挂载后，查看文件属性、md5 值，跟主记录的值进行对比，一致。
分别验证了 ldiskfs、zfs，结果均一致。zfs 在备挂载前需要额外将 zpool 从主导出，导入到备节点。

2.3.5. PCC 功能

在客户端配置 PCC 功能后，用 dd 命令测试不同大小数据块，在 512 字节到 4MB 之间，PCC 开启相比关闭能显著提升带宽。

2.3.6. barreleye 仪表盘

源码编译出包 barreleye，在 ARM64 上部署，配合 influxdb、grafana，能观测 lustre 统计指标，方便调测性能参考。

2.4. 性能测试

参考剑桥大学的优化文档 https://www.eofs.eu/_media/events/lad19/03_matt_raso-barnett-io500-cambridge.pdf，进行了 multirail、条带、DoM、DNE 等优化配置，用 posix.odirect 优化 ior easy 读写，用 mpiio 优化 ior hard 测试。在单台服务器上测试不同 ost、mdt 个数，找到了单台服务器最佳性能配置。对比测试了 ib、tcp 网络栈，ib 比 tcp 高 10 分左右。后端存储文件系统 ldiskfs 比 zfs 高 13 分左右。 我们还对比了 ARM64、x86 做服务端的性能，基本持平。

2.5. Arm 相关问题

LU-15978: fix striped directory deletion fails for 64K PAGE_SIZE (Fixed in master)
LU-15722: fix write stuck for 64K PAGE_SIZE(Fixed in master)
LU-16246: NULL pointer at lod_lookup+0x24/0x38
LU-16245: ASSERTION(iobuf->dr_elapsed_valid == 0)

3. 总结展望

当前 Lustre 已有 ARM64 版本可用，且通过了我们的小规模集群验证，欢迎大家试用。我们会继续修复测试用例失败项改善版本，争取能直接在 lustre 官网发布 ARM64 版本。