API 接收优化建议

一、服务器相关优化

1. 机器性能需要达到要求

我们推荐的机器配置：

CPU: Intel(R) Xeon(R) Gold 6244 CPU @ 3.6GHz
内存：64G
硬盘：500G(最好固态)
网卡：Solarflare x2522 10G 网卡 X2
OS: CentOS 7.6 X64
主板类型: DELL R740

客户行情接收机器配置可以跟推荐配置进行比较，性能自然是越高越好

2. BIOS 设置

系统 BIOS 设置，主要为了使 CPU、内存运行于最佳性能方式下，不考虑节能。以 dell 服务器为例，调整为如下表设置

System Setup screen	Setting	Recommended alternative for low-latency environments
System Profile Settings	System Profile	Custom
System Profile Settings	CPU Power Management	Maximum Performance
System Profile Settings	Memory Frequency	Maximum Performance
System Profile Settings	Turbo Boost	Disabled
System Profile Settings	C1E	Disabled
System Profile Settings	C States	Disabled
System Profile Settings	Monitor/Mwait	Disabled
System Profile Settings	Memory Patrol Scrub	Enabled
System Profile Settings	Memory Refresh Rate	1x
Memory Settings	Memory Mode	Optimizer
Memory Settings	Node Interleaving	Disabled
Memory Settings	Correctable Memory ECC	Disabled
Processor Settings	Logical Processor	Disabled
Processor Settings	Virtualization Technology	Disabled
Processor Settings	CPU Interconnect Speed	Maximum Data Rate
Processor Settings	Adjacent Cache Line Prefetch	Enabled
Processor Settings	Hardware Prefetcher	Enabled

Processor Settings	DCU Streamer Prefetcher	Enabled
Processor Settings	DCU IP Prefetcher	Enabled

3. 内核参数配置

内核参数配置的目的，也是让 CPU 运行于最高性能下。

以推荐的 Centos/Redhat 7.x 系统为例：
使用 root 用户，编辑 /etc/default/grub 文件

找到GRUB_CMDLINE_LINUX=”xxxxxx”这行，没有就自行创建一行。
在引号内尾部，加上 “nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce idle=poll”
如需进行cpu隔离，可采用添加isolcpus=3,5,7
如需进行 cpu 中断隔离，可采用添加 rcu_nocbs=3,5,7 nohz_full=3,5,7
编辑完成后保存文件。

保存完成后更新内核配置，在 shell 下执行
grub2-mkconfig -o /boot/grub2/grub.cfg
grub2-mkconfig –o /boot/efi/EFI/centos/grub.cfg
两个语句实际根据安装方式，其中一种会生效。执
行后，重启服务器。
服务器重启后，使用命令行：cat /proc/cmdline，如果打印的内容中已经存在刚才增加的一串
“nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce idle=poll”，则说明修改已经生效。
注意：cpu隔离和中断隔离不能把所有cpu都隔离！

4. SolarFlare 网卡相关优化（可选）

如果系统安装了 SolarFlare 网卡，可以通过以下配置使之运行于较好性能下并能够正常支持Tcp Direct 相关的程序。

1）开启系统的大页功能：

编辑 /etc/sysctl.conf 文件中，增加一行：

vm.nr_hugepages=2048

如果该行已经存在无需改动，重启服务器生效。

2）为防止 SolarFlare 网卡的 onload 驱动报出 PIO Buffer不足的错误设置

在/etc/modprobe.d/sfc.conf 里增加一句

options sfc piobuf_size=0

重启服务器生效。

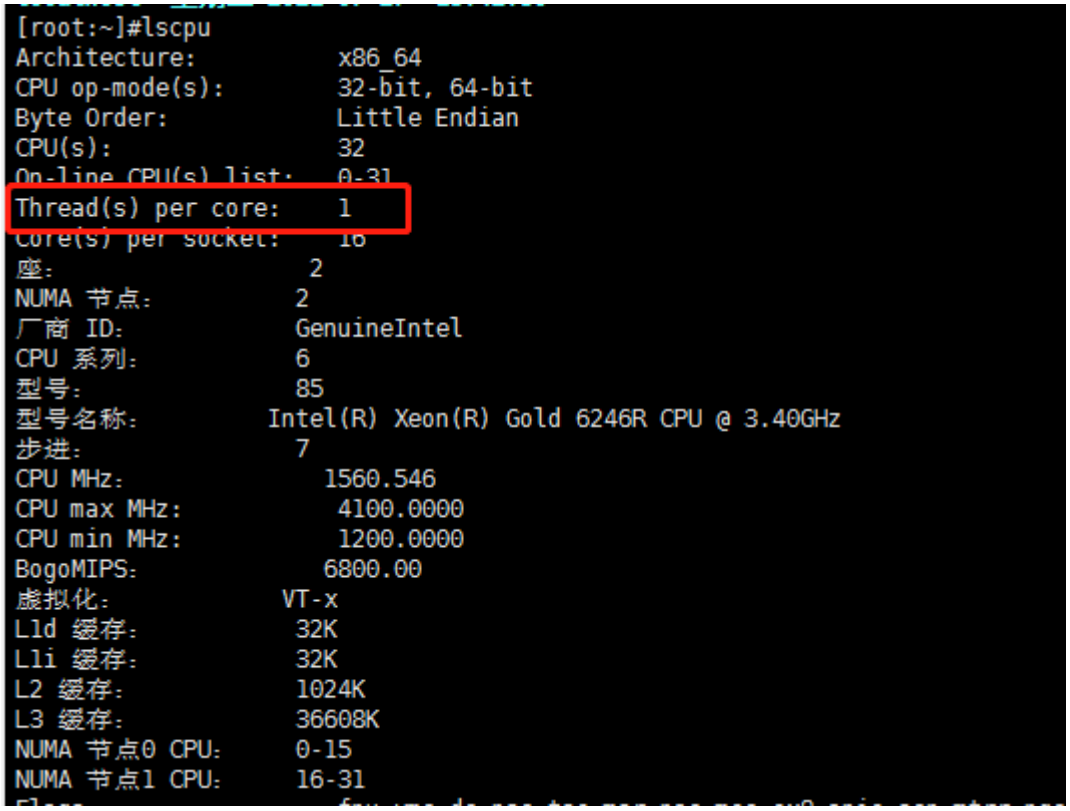
5. 超频机 CPU 性能模式检查

如果机器是超频机，需要检查模式是否为 latency-performance，如下图

```
[root:~]#tuned-adm profile
Available profiles:
- balanced                    - General non-specialized tuned profile
- desktop                     - Optimize for the desktop use-case
- latency-performance         - Optimize for deterministic performance at the cost of increased power consumption
- network-latency             - Optimize for deterministic performance at the cost of increased power consumption, focused o
n low latency network performance
- network-throughput          - Optimize for streaming network throughput, generally only necessary on older CPUs or 40G+ ne
tworks
- powersave                   - Optimize for low power consumption
- throughput-performance      - Broadly applicable tuning that provides excellent performance across a variety of common ser
ver workloads
- virtual-guest               - Optimize for running inside a virtual guest
- virtual-host                - Optimize for running KVM guests
Current active profile: latency-performance
```

如果不是性能模式，需要修改，修改命令：tuned-adm profile latency-performance 保证在超频机下采用的 Cpu 主频为超频。

6. 选择关闭超线程



如果 Thread(s) per core=1 说明超线程已经被关闭，如果=2 则说明未关闭

7. 服务器是否有做定时重启

尽量保证服务器运行一段时间（建议 1 周）后做个定时重启，避免服务器长时间运行后 cache 过大的问题。

二、程序相关优化

1. 建议采用 solarflare 的 efvi 模式接收行情

相比于使用普通 socket 接收行情方式，efvi 接收行情能较大幅度提升行情接收能力，保证行情能收全

2. 程序里接收的缓冲区是否设小了，可以设置一个稍大一点的缓冲区

以 efvi 行情接收 demo 为例，缓冲区最大可以设置为 4096（demo 默认为 2048）

```
bool udp_quote_tree::init(sock_udp_param& shfe)
{
    m_udp_param = shfe;
    m_receive_quit_flag = false;

    m_res = (resources*)calloc(1, sizeof(struct resources));
    if (!m_res)
    {
        exit(1);
    }

    /* Open driver and allocate a VI. */
    TRY(ef_driver_open(&m_res->dh));

    TRY(ef_pd_alloc_by_name(&m_res->pd, m_res->dh, (m_udp_param.m_eth_name).c_str(), EF_PD_DEFAULT));

    TRY(ef_vi_alloc_from_pd(&m_res->vi, m_res->dh, &m_res->pd, m_res->dh, -1, 2048, 0, NULL, -1, EF_VI_FLAGS_DEFAULT));

    m_res->rx_prefix_len = ef_vi_receive_prefix_len(&m_res->vi);
```

3. 行情接收程序是否有绑核和隔离

客户需要在程序里进行 CPU 绑核，CPU 隔离的方法可以参考章节 1 中服务器相关优化的第 3 点配置。隔离之后保证该组播通道的接收独享一个 CPU 核。

4. 绑核是否有采用亲密度高的 CPU 组

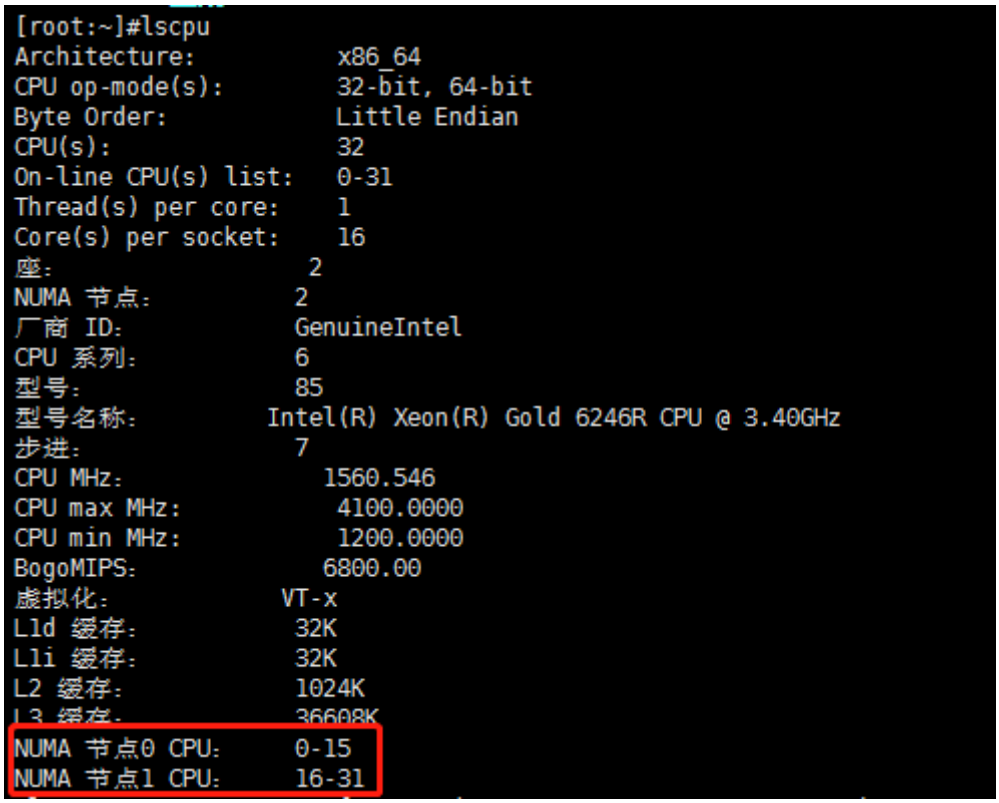
在极速环境下，发挥 solarflare 网卡的最佳性能，需要让网卡与 cpu 保持一个良好的绑定关系。

1) 查看网卡所在 cpu 组号

命令：cat /sys/class/net/网卡名/device/numa_node

得出某网卡与 cpu 组的亲密度为 0

2）lscpu 查看 cpu 分组情况



则程序绑核以及隔离的时候，用组别 0 里面对应的 cpu 核心即可。

5. 程序中是否有落日志

尽量减少程序中因为打印日志而产生的性能损耗，如果需要打印日志的话，如需记录日志可采用异步的方式。

6. 行情接收处理部分有没有使用虚函数

虚函数在使用过程中会查询虚函数表，所以这部分可能会稍微影响性能。

7. 避免使用系统命令工具

在调用 API 的同进程中,不建议使用 system(“ifconfig”)或 system(“netstat -s”)等系统调用，会影响 EFVI 的接收

三、问题处理

1. Linux socket 模式风险项

普通 socket 不要使用不同网卡接收同组播 IP 会接收，否则会接收到两份一样的数据。

2. windows 高速模式问题处理

Solarflare 网卡下，高速模式个别通道无数据。
由于 solarflare 网卡的硬件限制， 在使用 windows 高速模式时，匹配数据包使用的是对 ip port 四要素的 hash 摘要， 而非原始值， 在这个转换过程中， 存在低概率的 hash 碰撞，导致部分通道的的配置被无效化，最终出现程序接收不到数据的情况。
对此的解决方法，可以通过修改 组播 ip 端口来规避 hash 碰撞。

3. ubuntu 系统下 x25 模式问题

ubuntu 20.04 版本不支持低版本的 exanic 驱动编译安装， 目前测试成功的驱动版本为 2.7.2。
api 的 exanic x25 模式在 ubuntu 上需要使用 root 权限启动，对于使用 sudo 的场合， 环境变量可能会失效， 可以在使用程序时候先完成环境变量配置，命令 sudo LD_LIBRARY_PATH=./:$LD_LIBRARY_PATH ./路径/可执行程序 形式的命令运行程序。