2022.08.19 陕煤云主机MySQL性能低问题

时间

2022.08.19 下午

耗时

排查问题半个小时左右

记录排查过程一个小时左右

客户

陕煤

环境

ZStack 4.2.20

物理机规格：80C754G，超融合环境，2 NUMA nodes(0：0-19,40-59；1：20-39，60-79)；

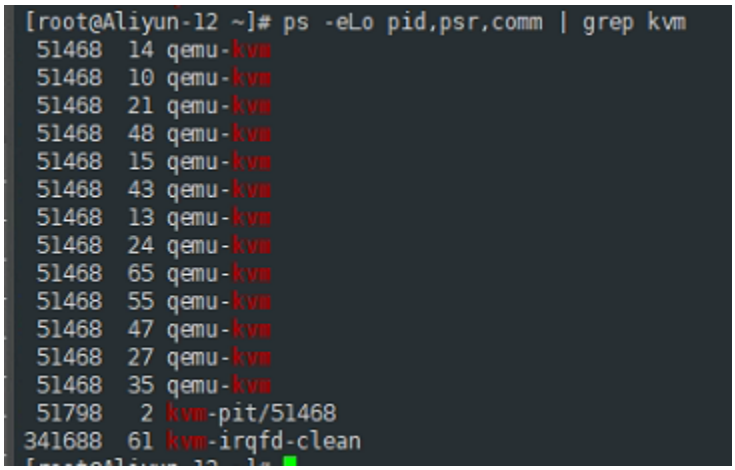
云主机规格：128C256G，启用了规格热修改；

其他信息：

1. 物理机上运行了其他的任务，例如docker、ceph等。
2. 云主机上运行了客户业务，不能执行重启关机等操作，只能查看；

```
[root@Aliyun-12 ~]# lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                80
On-line CPU(s) list:   0-79
Thread(s) per core:    2
Core(s) per socket:    20
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Gold 6230N CPU @ 2.30GHz
Stepping:              7
CPU MHz:               2899.987
CPU max MHz:           3500.0000
CPU min MHz:           1000.0000
BogoMIPS:              4600.00
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              1024K
L3 cache:              28160K
NUMA node0 CPU(s):     0-19,40-59
NUMA node1 CPU(s):     20-39,60-79
Flags:                 fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx pdpe1gb rdtscp lm
p_tsc aperfmperf eagerfpu pni pclmulqdq dtes64 ds_cpl vmx smx est tm2 ssse3 sdbg fma cx16 xtpr pdcm pcid dca sse4_1 sse4_2 x2apic movbe popcnt tsc_deadline_timer aes xsave
n intel_pt ssbd mba ibrs ibpb stibp ibrs_enhanced tpr_shadow vnmi flexpriority ept vpid fsgsbase tsc_adjust bmi1 hle avx2 smep bmi2 erms invpcid rtm cqm mpx rdt_a avx512f
aveopt xsavec xgetbv1 cqm_llc cqm_occup_llc cqm_mbm_total cqm_mbm_local dtherm ida arat pln pts hwp hwp_act_window hwp_epp hwp_pkg_req pku ospke avx512_vnni md_clear spec_
[root@Aliyun-12 ~]# docker ps
CONTAINER ID        IMAGE               COMMAND                  CREATED             STATUS              PORTS               NAMES
fbd25857efe4        sds/postgres        "docker-entrypoint..."   2 weeks ago         Up 2 weeks                              sds-postgres
db071d50bdb6        sds/pgbouncer       "./entrypoint.sh /..."   2 weeks ago         Up 2 weeks                              sds-pgbouncer
2624aadecb12        sds/elasticsearch   "/tini -- /usr/loc..."   2 weeks ago         Up 2 weeks                              sds-elasticsearch
a879af3c635f        sds/openresty       "/usr/local/openre..."   2 weeks ago         Up 2 weeks                              sds-prom-nginx
816768b6c678        sds/prometheus      "/bin/wrapper.sh -..."   2 weeks ago         Up 2 weeks                              sds-prometheus
2b7f45eae8df        sds/node            "npm run start:prod"     2 weeks ago         Up 2 weeks                              sds-xmd-api
a8a0336436bb        sds/openresty       "/usr/local/openre..."   2 weeks ago         Up 2 weeks                              sds-nginx
[root@Aliyun-12 ~]#
```

```
[root@Aliyun-12 ~]# free -h
              total        used        free      shared  buff/cache   available
Mem:           754G        455G        238G        226M         60G        289G
Swap:            0B          0B          0B
[root@Aliyun-12 ~]# cat /sys/devices/system/node/node
node0/ node1/
[root@Aliyun-12 ~]# cat /sys/devices/system/node/node0/meminfo
Node 0 MemTotal:       401246352 kB
Node 0 MemFree:         31621088 kB
Node 0 MemUsed:        369625264 kB
Node 0 Active:         327371300 kB
Node 0 Inactive:        19849084 kB
Node 0 Active(anon):   323043708 kB
Node 0 Inactive(anon):    63952 kB
Node 0 Active(file):    4327592 kB
Node 0 Inactive(file): 19785132 kB
Node 0 Unevictable:      328948 kB
Node 0 Mlocked:          329004 kB
Node 0 Dirty:               424 kB
Node 0 Writeback:             0 kB
Node 0 FilePages:       24364048 kB
Node 0 Mapped:           6245568 kB
Node 0 AnonPages:       323184664 kB
Node 0 Shmem:              66552 kB
Node 0 KernelStack:        58752 kB
Node 0 PageTables:        642684 kB
Node 0 NFS_Unstable:           0 kB
Node 0 Bounce:                 0 kB
Node 0 WritebackTmp:           0 kB
Node 0 Slab:             1614896 kB
Node 0 SReclaimable:      864020 kB
Node 0 SUnreclaim:        750876 kB
Node 0 AnonHugePages:  128729088 kB
Node 0 HugePages_Total:     0
Node 0 HugePages_Free:      0
Node 0 HugePages_Surp:      0
[root@Aliyun-12 ~]# cat /sys/devices/system/node/node1/meminfo
Node 1 MemTotal:       402653184 kB
Node 1 MemFree:        218438364 kB
Node 1 MemUsed:        184214820 kB
Node 1 Active:         141384548 kB
Node 1 Inactive:        19962004 kB
Node 1 Active(anon):   126146824 kB
Node 1 Inactive(anon):   167820 kB
Node 1 Active(file):   15237724 kB
Node 1 Inactive(file): 19794184 kB
Node 1 Unevictable:     2519684 kB
Node 1 Mlocked:         2519684 kB
Node 1 Dirty:              3676 kB
Node 1 Writeback:             0 kB
Node 1 FilePages:       35205196 kB
Node 1 Mapped:           5495500 kB
Node 1 AnonPages:       128661568 kB
Node 1 Shmem:             169756 kB
Node 1 KernelStack:        42368 kB
```



问题描述

问题来自 TIC-84 - 正在获取问题细节。。。 状态

客户反馈迁移后的云主机MySQL性能低，并未提及具体的性能衡量方法，只反馈性能差

排查步骤、过程

该物理机上只有一台云主机，不能执行重启关机等操作，所以只能查看相关信息。

1. 先查看物理机的基本信息：lscpu、free -h；
2. 技术支持反馈负载高，所以使用top命令后按1查看了CPU的负载情况，显示每个CPU均有不同us，us的数值均在0-35之间且分布比较乱，只能得出**物理机上除了云主机外还有其他负载**；
3. 然后使用ps -eLo pid,psr,comm | grep kvm命令查看云主机运行在哪些CPU上，发现**云主机虽然定义了128个CPU，但是实际上只有截图中所示的十几个vCPU线程**；
4. 使用ps auxw | grep kworker命令发现物理机上还运行docker、ceph等负载；
5. 执行docker ps命令发现物理机上运行了7个容器；
6. 然后去查看物理机上NUMA node 0和1的内存使用情况，发现**node0可用内存只剩下30G左右，但是node1上的可用内存还有200~300G**，怀疑是整个物理机上运行的负载均运行在node0上，所以相关程序的内存大部分均在node0上，导致node0负载偏高；
7. 由于云主机不能重启关机，所以无法通过关机操作来确定云主机实际使用内存是哪个NUMA node上；

问题根源

初步检查后，造成该问题的的根源怀疑如下：

1. 物理机上未作资源隔离，其他负载影响了云主机；
2. 由于OS调度问题，云主机和物理机上其他负载在竞争node0上的资源，造成性能下降；
3. 云主机未配置资源绑定（vCPUpin、emulatorpin、iothreadpin等）；
4. 云主机vCPU数量严重超过了物理机CPU数量，且实际vCPU线程并未存在那么多；

问题判断方法

云主机计算、数据库性能问题排查的简单过程：

1. 看物理机基本规格；
2. 看物理机上各个CPU负载，是否存在使用率接近满了或者满了的情况；
3. 看物理机上是否存在除云主机外的其他负载；
4. 看云主机vCPU线程运行在哪些CPU上，是否有的CPU负载高；
5. 再看物理机上各个NUMA node的内存使用情况，是否存在内存使用不均衡的情况；

临时解决方法

由于不能直接操作云主机且物理机上存在其他负载，与技术支持沟通后，只能先给出优化方案，向客户申请优化测试，**并不能确定提供的优化方案的优化效果**。

基于上述的排查，决定优化方案大致为将云主机绑定到node 1，因为node1 的资源竞争并没有那么激烈。**下列的所有优化手段均涉及到重启云主机**：

1. 降低云主机规格，云主机如果运行的是MySQL，并不是vCPU数量越多越好，MySQL貌似最多支持48个CPU，建议云主机规格改为48，这个数值可以借助于top查看mysql进程的CPU使用率来决定；
2. 关闭规格热修改，vNUMA与规格热修改会有性能冲突；
3. 配置vCPU绑定，绑定到NUMA node 1上，需要注意的是，云主机规格超过了单个NUMA node，先将vCPU绑定node 1上，再将剩余的绑定到node 0 上，需要注意的是，云主机的vCPU分布一定是绝大部分在node 1上的；

4. 基于vCPU绑定的情况配置vNUMA拓扑，按照vCPU和云主机内存大小的比例来配置云主机的vNUMA node的内存大小；

如果上述优化后性能仍不及预期，可以考虑下述的优化手段：

1. （可选）为云主机的数据库数据所在的磁盘配置IOThreadPin，需要注意的是IOThreadPin需要绑定在node1，因为绝大部分vCPU在node1。之所以是可选选项，因为云主机目前的大问题在CPU和内存；
2. （可选）物理机上CPU隔离，将物理机上CPU隔离，预留一部分给系统和其他负载使用，隔离的CPU专门划分为云主机使用，隔离物理机上云主机和其他负载。需要注意的是CPU隔离要搭配vCPU绑定使用，因为不配置vCPU绑定就意味着OS来调度，OS无法自动调度被隔离的CPU；
3. （可选）使用内存大页。

由于客户版本较低，上述的操作在MN节点和UI上不支持，需要借助于xmlhook功能。

长期方案

ZStack 4.3.12开始已经支持vNUMA功能了，可以配置除了iothreadpin外的vCPU绑定、vNUMA等，对于简单的数据库性能问题，足够使用了，如果还需要更进一步的优化，设计的改动项会更多，且效果也不确定，需要手工慢慢调优。

相关的文章

存在类似的技术支持：20220726--南京工业大学v2v迁移云主机数据库(mariadb)性能问题

标签内容

在指定标签中没有内容