20220726--南京工业大学v2v迁移云主机数据库(mariadb)性能问题

2022.07.26处理情况：

一、来源

来自于 ZSTAC-48375 - [南京工业大学]v2v迁移过来的虚拟机MySQL性能低 已关闭

二、分析

1. 从jira上的描述来看，客户v2v迁移过来的云主机(32C32G)与新创建的云主机，运行mariadb的性能差距过大(约一倍)。由于jira上只有数据库mariadb的截图等信息，缺少云主机和物理机等相关信息，与 技术支持@任银坤 沟通， 获取云主机xml定义文件以及物理机NUMA nodes信息。

此外，迁移的云主机上运行的是mariadb server version：10.4.14-MariaDB-log MariaDB Server。云主机使用的是同一个主存储(**ceph存储**)。

客户反馈新创建的云主机可以cover住用户的性能需求，使用同样的主存储，同样的规格，只有运行的物理机不一样。

2. 根据反馈回来的云主机xml以及物理机NUMA nodes信息，有以下分析：

    1. **开启了规格热修改**，这个会跟numa功能冲突；
    2. **未做cpu绑定、emulator pin以及vnuma**，这部分需要物理机numa node信息再配置；
    3. 猜测未启用mariadb的innodb_numa_interleave (如果Mariadb这个版本有这个参数)；
    4. 在此基础上，如果性能还未达到预期，考虑叠在上述配置下加以下配置：
        a. 集群内存大页；
        b. 数据云盘iothreadpin；
    5. 额外的，如果物理机上存在其他得vm或者进程任务在运行，考虑CPU隔离；

3. 分析过程中还获取了其他的信息：

    1. 当前使用mariadb server 版本**不支持innodb_numa_interleave**，即无法配置numa交织(执行show global variables like '%numa%';)，参考https://jira.mariadb.org/browse/MDEV-18860；
    2. 迁移的云主机和新创建的云主机运行的物理机**不是同一个物理机**，并且**物理机上还存在其他的负载**，无法进行直接的性能对比，这可能是性能差距过大的原因之一；
    3. 使用的ZStack 版本为**4.2.16nangong**，ui上只支持cpu绑定，并无vnuma等功能；
    4. 从物理机capabilities信息来看，不支持iommu；

4. 沟通后决定今天先整理信息，第二天一起协同支持这个问题。

5. 优先考虑的优化如下，待明天执行后确认效果：

    1. 关闭规格热修改；
    2. 考虑未做CPU隔离，先配置CPU绑定到NUMA node 1，并配置vnuma拓扑；
    3. 配置emulatorpin，使用NUMA node 1上未绑定使用的CPU；
    4. 配置iothreadpin，使用NUMA node 1上未绑定使用的CPU；

6. 如果步骤5执行后效果未及预期，继续叠加以下优化：

    1. CPU隔离，隔离NUMA node 1 - 优先级中；
    2. 启用集群内存大页(先确认当前zstack版本是否支持集群内存大页、当前集群下物理机是否可以重启)，配置vm内存大页 - 优先级低；

3. 隔离物理机上kworker以及cpu中断 - 优先级低；

2022.07.27处理情况：

需要先确定以下要点：

1. **vm是否可以执行重启、变更xml等操作；**
2. **物理机是否可以重启以配置CPU隔离等；**
3. **当前集群下的物理机可以执行的操作等(可能涉及到集群内存大页，需要重启相关物理机)；**

云主机可以重启，不可以删除数据；

物理机上存在其他业务，操作需要申请；

支持过程：

1. 查看了物理机的一些信息：

```
top - 11:38:32 up 64 days, 18 min,  2 users,  load average: 1.09, 1.23, 1.42
Tasks: 1225 total,   1 running, 1224 sleeping,   0 stopped,   0 zombie
%Cpu0  :  1.7 us,  1.1 sy,  0.0 ni, 96.1 id,  0.0 wa,  0.0 hi,  1.1 si,  0.0 st
%Cpu1  :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu2  :  1.7 us,  1.7 sy,  0.0 ni, 96.6 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu3  :  0.6 us,  1.1 sy,  0.0 ni, 98.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu4  :  0.6 us,  0.6 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu5  :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu6  :  2.3 us,  0.6 sy,  0.0 ni, 97.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu7  :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu8  :  2.2 us,  0.6 sy,  0.0 ni, 97.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu9  :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu10 :  0.6 us,  0.6 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu11 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu12 :  1.1 us,  1.1 sy,  0.0 ni, 97.8 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu13 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu14 :  0.6 us,  0.6 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu15 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu16 :  0.6 us,  1.1 sy,  0.0 ni, 98.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu17 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu18 :  0.6 us,  0.6 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu19 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu20 :  0.6 us,  1.1 sy,  0.0 ni, 98.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu21 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu22 :  0.6 us,  0.6 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu23 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu24 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu25 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu26 :  2.3 us,  0.0 sy,  0.0 ni, 97.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu27 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu28 :  2.8 us,  0.0 sy,  0.0 ni, 97.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu29 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu30 :  2.3 us,  0.0 sy,  0.0 ni, 97.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu31 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu32 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu33 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu34 :  1.1 us,  0.0 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu35 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu36 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu37 :  0.0 us,  0.6 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu38 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu39 :  0.0 us,  0.6 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu40 :  1.7 us,  0.6 sy,  0.0 ni, 97.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu41 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu42 :  2.2 us,  0.6 sy,  0.0 ni, 97.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu43 :  0.6 us,  0.0 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu44 :  0.6 us,  0.6 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu45 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu46 :  1.1 us,  0.0 sy,  0.0 ni, 98.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu47 :  0.0 us,  0.6 sy,  0.0 ni, 99.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu48 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu49 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu50 :  0.6 us,  1.7 sy,  0.0 ni, 97.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu51 :  1.1 us,  2.2 sy,  0.0 ni, 96.6 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu52 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu53 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
```

```
[root@zstack-9 ~]# cat /sys/devices/system/node/node1/meminfo
Node 1 MemTotal:       402653184 kB
Node 1 MemFree:        342521148 kB
Node 1 MemUsed:         60132036 kB
Node 1 Active:          49104556 kB
Node 1 Inactive:          553116 kB
Node 1 Active(anon):    48679460 kB
Node 1 Inactive(anon):     65884 kB
Node 1 Active(file):      425096 kB
Node 1 Inactive(file):    487232 kB
Node 1 Unevictable:         6248 kB
Node 1 Mlocked:             6248 kB
Node 1 Dirty:                  4 kB
Node 1 Writeback:              0 kB
Node 1 FilePages:         984720 kB
Node 1 Mapped:            115484 kB
Node 1 AnonPages:       48679868 kB
Node 1 Shmem:              67552 kB
Node 1 KernelStack:        31952 kB
Node 1 PageTables:        155568 kB
Node 1 NFS_Unstable:           0 kB
Node 1 Bounce:                 0 kB
Node 1 WritebackTmp:           0 kB
Node 1 Slab:              517120 kB
Node 1 SReclaimable:      124348 kB
Node 1 SUnreclaim:        392772 kB
Node 1 AnonHugePages:   36780032 kB
Node 1 HugePages_Total:     0
Node 1 HugePages_Free:      0
Node 1 HugePages_Surp:      0
[root@zstack-9 ~]# cat /sys/devices/system/node/node0/meminfo
Node 0 MemTotal:       401212464 kB
Node 0 MemFree:        290726364 kB
Node 0 MemUsed:        110486100 kB
Node 0 Active:          97611440 kB
Node 0 Inactive:         1768032 kB
Node 0 Active(anon):    96241116 kB
Node 0 Inactive(anon):     25380 kB
Node 0 Active(file):     1370324 kB
Node 0 Inactive(file):   1742652 kB
Node 0 Unevictable:         1828 kB
Node 0 Mlocked:             1828 kB
Node 0 Dirty:                 72 kB
Node 0 Writeback:              0 kB
Node 0 FilePages:        3140480 kB
Node 0 Mapped:            145156 kB
Node 0 AnonPages:       96240532 kB
Node 0 Shmem:              26008 kB
Node 0 KernelStack:        26768 kB
Node 0 PageTables:        153360 kB
Node 0 NFS_Unstable:           0 kB
Node 0 Bounce:                 0 kB
Node 0 WritebackTmp:           0 kB
Node 0 Slab:              950684 kB
Node 0 SReclaimable:      514208 kB
Node 0 SUnreclaim:        436476 kB
Node 0 AnonHugePages:   30855168 kB
Node 0 HugePages_Total:     0
```

```
[root@zstack-9 ~]# virsh version
Compiled against library: libvirt 4.9.0
Using library: libvirt 4.9.0
Using API: QEMU 4.9.0
Running hypervisor: QEMU 2.12.0

[root@zstack-9 ~]#
```

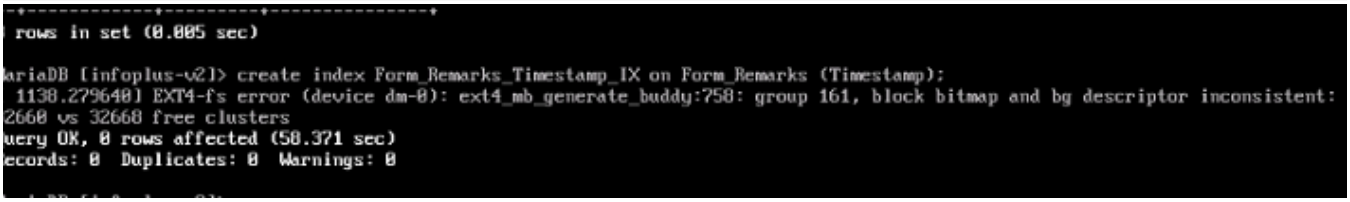
2. 按照昨天准备的方案，关闭规格热修改，配置vnuma、iothreadpin、emulatorpin， 配置好后重启云主机：

具体的配置如下：

云主机配置

```
<iothreads>4</iothreads>
<iothreadids>
    <iothread id='3'/>
</iothreadids>

<cputune>
  <vcpupin vcpu='0' cpuset='1'/>
  <vcpupin vcpu='1' cpuset='3'/>
...
  <vcpupin vcpu='31' cpuset='63'/>
  <iothreadpin iothread='3' cpuset='65'/>
  <emulatorpin cpuset='67'/>
</cputune>

<numatune>
    <memnode cellid='0' mode='preferred' nodeset='1'/>
</numatune>

<cpu mode='host-passthrough' check='none'>
    <topology sockets='1' cores='32' threads='1'/>
    <numa>
      <cell id='0' cpus='0-31' memory='33554432' unit='KiB'>
      <distances>
        <sibling id='0' value='10' />
      </distances>
    </numa>
</cpu>

<disk type='network' device='disk'>
      <driver name='qemu' type='raw' iothread='3' />
      <auth username='zstack'>
        <secret type='ceph' uuid='f4ef6892-fb4d-418f-8592-1da0244dc26a'/>
      </auth>
      <source protocol='rbd' name='pool-35fe9718a70242fea18db351fed1a9ef/26ff8e83e27e4406b328cc06ac064717'>
        <host name='1.1.1.23' port='6789'/>
        <host name='1.1.1.21' port='6789'/>
        <host name='1.1.1.22' port='6789'/>
      </source>
      <target dev='vdb' bus='virtio'/>
      <serial>26ff8e83e27e4406b328cc06ac064717</serial>
      <alias name='virtio-disk1'/>
      <address type='pci' domain='0x0000' bus='0x00' slot='0x0b' function='0x0'/>
    </disk>
```

3. 配置后重启云主机：使用create index Form_Remarks_Timestamp_IX on Form_Remarks(Timestamp);语句的执行时间来表达性能指标，该语句据技术支持和客户描述，在30s-60s内都是符合预期的，配置后的测试结果为58s，在预期范围内。



到这里就表示优化生效了，未优化时该语句执行时间为120s左右，优化后为在预期内的58s。

问题分析

在查看了物理机上的信息，并与技术支持沟通后，认为有以下原因：

1. 迁移云主机后，云主机运行的物理机上node0资源竞争较node1更为激烈；
2. 物理机上存在别的云主机(约4~5台其他云主机)会影响到云主机；
3. 不存在资源绑定(vcpu绑定、emulatorpin，iothreadpin)，导致内存访问以及cpu调度没有限定；
4. 云主机规格(32C32G)较小，不能充分使用资源；

需要注意的是：配置iothreadpin前后，执行create index语句的执行时间并无显著变化，表示iothreadpin这个优化并未带来显著的效果，可能是因为io压力不够；

如果后续需要进一步优化的建议：

1. 做CPU隔离、kworker以及中断亲和性设置，用来隔离物理机上其他负载带来的影响，让云主机独占资源；
2. 如果需要固定修改的云主机xml配置，建议使用xml-hook功能；
3. 启用内存大页；
4. 优化内核参数，例如vm.dirty_ratio，vm.swappness，block read_ahead_kb等；

需要改进的点：

1. 当前版本(4.2.160nangong)只支持cpu绑定，不支持vnuma，手工配置容易出错；
2. 只是执行create index语句来查看执行时间，手段单一；

其他：

会议录制：https://alidocs.dingtalk.com/i/team/gBVzxynBj92qXOao/docs/gBVzxdW9vMpNlzOa