K8S CPU管理策略介绍

1 背景

官方目前，只支持静态的配置cpuset(通过cpu-manager-policy)，并不支持动态的配置cpuset：

- https://github.com/kubernetes/kubernetes/issues/10983
- https://github.com/kubernetes/kubernetes/issues/10570

2 使用场景

对CPU敏感性的任务，如上下问切换，cache miss等

3 机制

kubelet的cpu-manager-policy默认配置为none，如果配置为static，主要注意如下：

- 需要为kubelet设置保留资源，如kube-reserved和system-reserved
- Pod QoS为Guaranteed 且 cpu request为整数cpu
- 不满足独占cpu的pod，将使用Shared Pool中的CPU集(CPUCapacity - ReservedCPUs - ExclusiveCPUs)
- 在开启超线程的机器上，CPU Manager会把所有core都作为可调度的core(这样在HT打开情况下，密集型的job可能由于大量的上下文切换导致性能暴跌)
- 优先在同一个物理core/socket上分配CPU
- CPU Manager当前不支持isolcpus

4 使用

如果当前环境的的cpu-manager-policy策略为none，要想调整为static，步骤如下：

- drain节点： kubectl drain node $node_name
- 删除/var/lib/kubelet/cpu_manager_state文件
- 修改/var/lib/kubelet/config.yaml配置文件，调整cpu-manager-policy策略为static
- 重启kubelet服务
- uncordon节点： kubectl uncordon node $node_name

5 验证

创建如下3种类型pod：

- 不配置resources， 为BestEffort类型
- 配置resources，但request和limit不相等，为Burstable类型
- 配置resources，request和limit相等，且CPU为整数，为Guaranteed类型

可以发现：

- 不管哪种类型pod，共享池中资源都会被正常更新：

```
 Namespace                  Name                               CPU Requests  CPU Limits  Memory
Requests  Memory Limits  AGE
  ---------                  ----                               ------------  ----------
---------------  -------------  ---
  default                    centos-besteffort-74bf9d9d57-dnhzs   0 (0%)        0 (0%)      0
(0%)           0 (0%)         57m
  default                    centos-guaranteed-6945548f79-78dsd   2 (16%)       2 (16%)     200Mi
(0%)       200Mi (0%)     52m
  default                    centos-burstable-7487ff7949-dwrb8    3 (25%)       4 (33%)     100Mi
(0%)       200Mi (0%)     25s
  kube-system                calico-node-hbhkb                  250m (2%)       0 (0%)      0
(0%)           0 (0%)         2d17h
  kube-system                kube-proxy-smzgc                   0 (0%)          0 (0%)      0
(0%)           0 (0%)         2d17h
  skydiscovery-system        license-manager-worker-9qbqt       100m (0%)    500m (4%)   100Mi (0%)
200Mi (0%)         2d5h
Allocated resources:
  (Total limits may be over 100 percent, i.e., overcommitted.)
  Resource           Requests     Limits
  --------           --------     ------
  cpu                5350m (44%)  6500m (54%)
  memory             400Mi (0%)   600Mi (0%)
  ephemeral-storage  0 (0%)       0 (0%)
Events:              <none>
```

- cpu_memory_state文件只扣除满足static策略类型pod所分配的资源：

```
[root@skyaxe-computing-1 kubelet]# cat cpu_manager_state
{"policyName":"static","defaultCpuSet":"0-7,9,11-15","entries":
{"2c42f25b380dadba4e75e289dc68a8d515a84073784fc4ae8380c676289a21c1":"8,10"},"checksum":1770764076}
[root@skyaxe-computing-1 kubelet]#
```

当把Guaranteed类型的pod删除后，共享资源和state文件都会被更新：

- 共享池：

```
 Namespace                  Name                               CPU Requests  CPU Limits  Memory
Requests  Memory Limits  AGE
  ---------                  ----                               ------------  ----------
---------------  -------------  ---
  default                    centos-74bf9d9d57-dnhzs            0 (0%)        0 (0%)      0
(0%)           0 (0%)         80m
  default                    centos-request-7487ff7949-dwrb8    3 (25%)       4 (33%)     100Mi
(0%)       200Mi (0%)     23m
  kube-system                calico-node-hbhkb                  250m (2%)     0 (0%)      0
(0%)           0 (0%)         2d17h
  kube-system                kube-proxy-smzgc                   0 (0%)        0 (0%)      0
(0%)           0 (0%)         2d17h
  skydiscovery-system        license-manager-worker-9qbqt       100m (0%)     500m (4%)   100Mi
(0%)       200Mi (0%)     2d5h
Allocated resources:
  (Total limits may be over 100 percent, i.e., overcommitted.)
  Resource           Requests     Limits
  --------           --------     ------
  cpu                3350m (27%)  4500m (37%)
  memory             200Mi (0%)   400Mi (0%)
  ephemeral-storage  0 (0%)       0 (0%)
Events:              <none>
```

- state文件：

```
[root@skyaxe-computing-1 kubelet]# cat cpu_manager_state
{"policyName":"static","defaultCpuSet":"0-15","checksum":2019817980}
[root@skyaxe-computing-1 kubelet]#
```

6 实现原理

cpu-manager-policy底层是通过cgroup的cpuset来实现：

- 在没有创建满足static类型的pod时，burstable pod的cpuset为0-15：

```
[root@skyaxe-computing-1 kubepods]# pwd
/sys/fs/cgroup/cpuset/kubepods
[root@skyaxe-computing-1 kubepods]# cat burstable/podb8934f1b-6716-4987-83af-451567276865
/caf808e62307a8e45f6e8b673a85f5a151c4518488a468827371b03a13193cef/cpuset.cpus
0-15
[root@skyaxe-computing-1 kubepods]#
```

- 当创建一个满足static类型pod时，burstable pod的cpuset为0-7,9,11-15

```
[root@skyaxe-computing-1 kubepods]# pwd
/sys/fs/cgroup/cpuset/kubepods
[root@skyaxe-computing-1 kubepods]# cat pod395b834e-400c-4318-9b88-e969dade9af5
/c32334f4eda76211c4a4c89faac1c1309876dce54e928558d9704a9a0f686f7d/cpuset.cpus
8,10
[root@skyaxe-computing-1 kubepods]# cat burstable/podb8934f1b-6716-4987-83af-451567276865
/caf808e62307a8e45f6e8b673a85f5a151c4518488a468827371b03a13193cef/cpuset.cpus
0-7,9,11-15
[root@skyaxe-computing-1 kubepods]#
```

- 在删除满足static类型的pod时，burstable pod的cpuset又变回为0-15

```
[root@skyaxe-computing-1 kubepods]# pwd
/sys/fs/cgroup/cpuset/kubepods
[root@skyaxe-computing-1 kubepods]# cat burstable/podb8934f1b-6716-4987-83af-451567276865
/caf808e62307a8e45f6e8b673a85f5a151c4518488a468827371b03a13193cef/cpuset.cpus
0-15
[root@skyaxe-computing-1 kubepods]#
```

注：系统服务(如：sshd、runtime、kubelet）总是可在所有CPU上运行，不受是否有static pod影响（即使设置kube-reserved 或者 system-reserved）

```
[root@skyaxe-computing-1 cpuset]# pwd
/sys/fs/cgroup/cpuset
[root@skyaxe-computing-1 cpuset]# ls
cgroup.clone_children  cgroup.sane_behavior  cpuset.effective_cpus  cpuset.mem_hardwall     cpuset.
memory_pressure_enabled  cpuset.mems                       kubepods       notify_on_release
cgroup.event_control   cpuset.cpu_exclusive  cpuset.effective_mems  cpuset.memory_migrate   cpuset.
memory_spread_page       cpuset.sched_load_balance        kube.slice     release_agent
cgroup.procs           cpuset.cpus           cpuset.mem_exclusive   cpuset.memory_pressure  cpuset.
memory_spread_slab       cpuset.sched_relax_domain_level  machine.slice  tasks
[root@skyaxe-computing-1 cpuset]# cat cpuset.cpus
0-15
[root@skyaxe-computing-1 cpuset]# wc -l tasks
651 tasks
[root@skyaxe-computing-1 cpuset]#
```

7 注意点

- 当满足static策略的pod被分配指定CPU后，当前运行在这些CPU上的POD会由于每10s(cpuManagerReconcilePeriod参数控制)一次的Reconcile进行CPU进行迁移，因此最坏情况下，会有10s时间static策略的pod和非static策略的pod共享这些CPU

- cpu_manager_state文件不会扣除保留资源，即在设置保留资源，同时也没有满足static策略pod时，显示机器上的所有CPU

8 参考

- https://cloud.tencent.com/developer/article/1402119
- https://kubernetes.io/docs/tasks/administer-cluster/cpu-management-policies/