OCI Marketplace - HPC Cluster Stack 사용 지침

Policies to deploy the stack:

allow service compute_management to use tag-namespace in tenancy

allow service compute_management to manage compute-management-family in tenancy

allow service compute_management to read app-catalog-listing in tenancy

allow group user to manage all-resources in compartment <compartmentName>

Policies for autoscaling or resizing:

변수를 지정할 때 설명한 대로 노드를 인증하는 방법으로 instance-principal 을 선택하는 경우 동적 그룹을 생성하고 여기에 다음 정책을 제공해야 함.

Allow dynamic-group instance_principal to read app-catalog-listing in tenancy

Allow dynamic-group instance_principal to use tag-namespace in tenancy

또 다른 Policy 설정 :

Allow dynamic-group instance_principal to manage compute-management-family in compartment <compartmentName>

Allow dynamic-group instance_principal to manage instance-family in compartment <compartmentName>

Allow dynamic-group instance_principal to use virtual-network-family in compartment <compartmentName>

or:

Allow dynamic-group instance_principal to manage all-resources in compartment <compartmentName>

Autoscaling

오토스케일링은 "cluster per job" 접근 방법을 사용.

즉, 대기열에서 대기 중인 작업의 경우 해당 작업에 대해 특별히 새 클러스터를 시작합니다. Autoscaling 은 클러스터 스핀다운도 처리합니다. 기본적으로 클러스터는 종료되기 전에 10 분 동안 유휴 상태로 유지됩니다. Autoscaling 은 cronjob 을 사용하여 한 스케줄러에서 다음 스케줄러로 빠르게 전환할 수 있습니다.

스택을 통해 배포된 초기 클러스터는 스핀다운되지 않습니다.

/opt/oci-hpc/autoscaling/queues.conf 에 구성 파일이 있고 /opt/oci-hpc/autoscaling/queues.conf.example 에 여러 대기열과 여러 인스턴스 유형을 추가하는

방법을 보여주는 예제가 있습니다. HPC, GPU 또는 Flex VM 에 대한 예제가 포함되어 있습니다.

인스턴스 유형 이름을 작업 정의의 기능으로 사용하여 올바른 종류의 노드를 실행/생성하는지 확인할 수 있습니다.

Queue 당 하나의 기본 인스턴스 유형과 하나의 기본 Queue 만 가질 수 있습니다. 기본이 아닌 Queue 에 submit 하려면 SBATCH 다음 줄을 파일에 추가합니다.

#SBATCH --partition compute

또는 명령줄에 sbatch -p queuename job.sh 를 추가합니다.

키워드 영구 허용은 클러스터를 스핀업하지만 false 로 설정될 때까지 삭제하지 않습니다. 해당 값을 변경한 후 slurm 을 재구성할 필요가 없습니다.

/opt/oci-hpc/autoscaling/queues.conf 를 수정한 후 /opt/oci-hpc/autoscaling/slurm_config.sh 를 실행해야 합니다.

Autoscaling 을 시작하려면: crontab -e 명령을 입력 후 아래 라인의 주석을 해제합니다.

* * * * * /opt/oci-hpc/autoscaling/crontab/autoscale_slurm.sh >> /opt/oci-hpc/autoscaling/logs/crontab_slurm.log 2>&1

Submit

Submit 방법: Slurm 작업은 언제나 제출할 수 있지만 /opt/oci-hpc/autoscaling/submit/의 예: 처럼 몇 가지 제약 조건을 더 설정할 수 있습니다.

#!/bin/sh #SBATCH -n 72

#SBATCH --ntasks-per-node 36

#SBATCH --exclusive

#SBATCH --job-name sleep_job

#SBATCH --constraint cluster-size-2,hpc

cd /nfs/scratch mkdir $SLURM_JOB_ID

cd $SLURM_JOB_ID MACHINEFILE="hostfile"

호스트가 동일한 위치에 있도록 mpi 용 Machinefile 생성

srun 을 통해 실행되는 것처럼 order

srun -N$SLURM_NNODES -n$SLURM_NNODES hostname > $MACHINEFILE sed -i 's/$/:36/' $MACHINEFILE

cat $MACHINEFILE

생성된 Machine 파일을 사용하여 실행

sleep 1000

cluster-size: 클러스터를 재사용할 수 있으므로 정확히 올바른 크기의 클러스터만 사용하도록 결정할 수 있습니다. 생성된 클러스터에는 cluster-size-x 기능이 있습니다. cluster-size-x 제약 조건을 설정하여 이것이 일치하는지 확인하고 16 노드 클러스터를 사용하는 1 노드 작업을 방지할 수 있습니다. 큰 클러스터에서 작은 작업을 실행하는 것이 마음에 들지 않으면 이 기능을 설정할 필요가 없습니다.

Instance Type: 제약 조건으로 실행하려는 OCI 인스턴스 유형을 지정할 수 있습니다. 이렇게 하면 올바른 모양에서 실행하고 올바른 클러스터도 생성할 수 있습니다. 인스턴스 유형은 yml 형식의 /opt/oci-hpc/autoscaling/queues.conf 파일에 정의됩니다. 사용하지 않더라도 모든 필드를 그대로 두십시오. 각 대기열에 여러 대기열과 여러 인스턴스 유형을 정의할 수 있습니다. 작업을 생성할 때 인스턴스 유형을 선택하지 않으면 기본 유형이 사용됩니다.

Clusters folders:

/opt/oci-hpc/autoscaling/clusters/clustername

Logs:

/opt/oci-hpc/autoscaling/logs

각 클러스터에는 이름이 create_clustername_date.log 및 delete_clustername_date.log 인 고유한 로그가 있습니다. crontab 의 로그는 crontab_slurm.log 에 있습니다.

Manual clusters:

클러스터를 수동으로 만들고 삭제할 수 있습니다.

Cluster 생성:

/opt/oci-hpc/autoscaling/create_cluster.sh NodeNumber clustername instance_type queue_name

Example:

```
/opt/oci-hpc/autoscaling/create_cluster.sh 4 compute2-1-hpc HPC_instance compute2
```

클러스터 이름은 queueName-clusterNumber-instanceType_keyword 여야 합니다.

키워드는 Slurm 에 등록할 /opt/oci-hpc/autoscaling/queues.conf 의 키워드와 일치해야 합니다.

Cluster Deletion:

```
/opt/oci-hpc/autoscaling/delete_cluster.sh clustername
```

삭제하는 동안 문제가 발생하면 다음을 사용하여 강제로 삭제할 수 있습니다.

```
/opt/oci-hpc/autoscaling/delete_cluster.sh clustername FORCE
```

클러스터가 이미 Destroy 되면 /opt/oci-hpc/autoscaling/clusters/clustername/currently_destroying 파일이 존재하게 됩니다.

Autoscaling Monitoring

Autoscaling 모니터링을 선택한 경우 실행 중인 작업과 대기 중인 작업 뿐만 아니라 어떤 노드가 동작하고 있는지 확인할 수 있습니다.

Grafana API 의 문제로 인해 Grafana 에서 대시보드 가져오기를 제외하고 모든 것이 자동으로 실행됩니다.

수동으로 수행하려면 선택한 브라우저에서 bastionIP:3000 으로 이동합니다.

ID 와 password 는 admin/admin 이며 최초 로그인 시 변경할 수 있습니다. 왼쪽 메뉴 모음에서 + 기호를 클릭하고 가져오기를 선택합니다. JSON 파일 업로드를 클릭하고 /opt/oci-hpc/playbooks/roles/autoscaling_mon/files/dashboard.json 에 있는 파일을 업로드합니다. 데이터 소스로 자동 크기 조정(MySQL)을 선택합니다. (mysql id/password 는 Stack 참조)

대시보드가 표시됩니다.

LDAP

HPC Cluster 에 의해 설치된 bastion host는 클러스터의 LDAP 서버 역할을 합니다. 기본 공유 홈 디렉토리를 그대로 두는 것이 좋습니다. 사용자 관리는 클러스터 명령을 사용하여 bastion에서 수행할 수 있습니다.