Guide: Expanding GPU Capacity in Hosted ROSA (RHOAI)

Goal:

Scale your Red Hat OpenShift AI (RHOAI) GPU capacity from 4 GPUs -> 8 GPUs on a Hosted ROSA cluster.

1. Check Current Machine Pools:

```
rosa list machinepools -c adonheis4
```

2. Scale-Up Attempt:

```
rosa edit machinepool gpu-nodes -c adonheis4 --replicas=8
```

If stuck at 4/8 replicas -> stale ignition metadata.

3. Confirm AWS Node Activity:

```
aws ec2 describe-instances --filters "Name=tag:api.openshift.com/name,Values=adonheis4" "Name=instance-type,Values=g6e.24xlarge" --region us-east-2
```

4. Rebuild GPU Pool (Fix Ignition):

----

a) Delete old GPU pool

```
rosa delete machinepool --cluster adonheis4 --name gpu-nodes
```

Wait 5-15 minutes for cleanup:

```
rosa list machinepools -c adonheis4
```

b) Recreate GPU pool fresh

```
rosa create machinepool --cluster adonheis4 --name gpu-nodes --instance-type g6e.24xlarge --replicas 8 --disk-size 300GiB --autorepair
```

5. Watch Provisioning:

```
watch -n 30 "rosa describe machinepool --cluster adonheis4 --name gpu-nodes"
```

6. Confirm Node Registration:

```
oc get nodes -L node.kubernetes.io/instance-type
```

7. Verify GPU Detection:

```
oc get nodes -o jsonpath='{range .items[*]}{.metadata.name}{"\t"}{.status.allocatable.nvidia\.com/gpu}{"\n"}{end}'
```

If GPUs missing:

```
oc rollout restart ds/nvidia-device-plugin-daemonset -n nvidia-gpu-operator
```

8. Enable Autoscaling (optional):

```
rosa edit machinepool --cluster adonheis4 --name gpu-nodes --enable-autoscaling --min-replicas=4 --max-replicas=8
```

Automation Script: rosa-rebuild-gpu-pool.sh

See full markdown guide for code implementation.