基于Kubernetes的DeepSpeed方案

一. DeepSpeed

1. 源代码

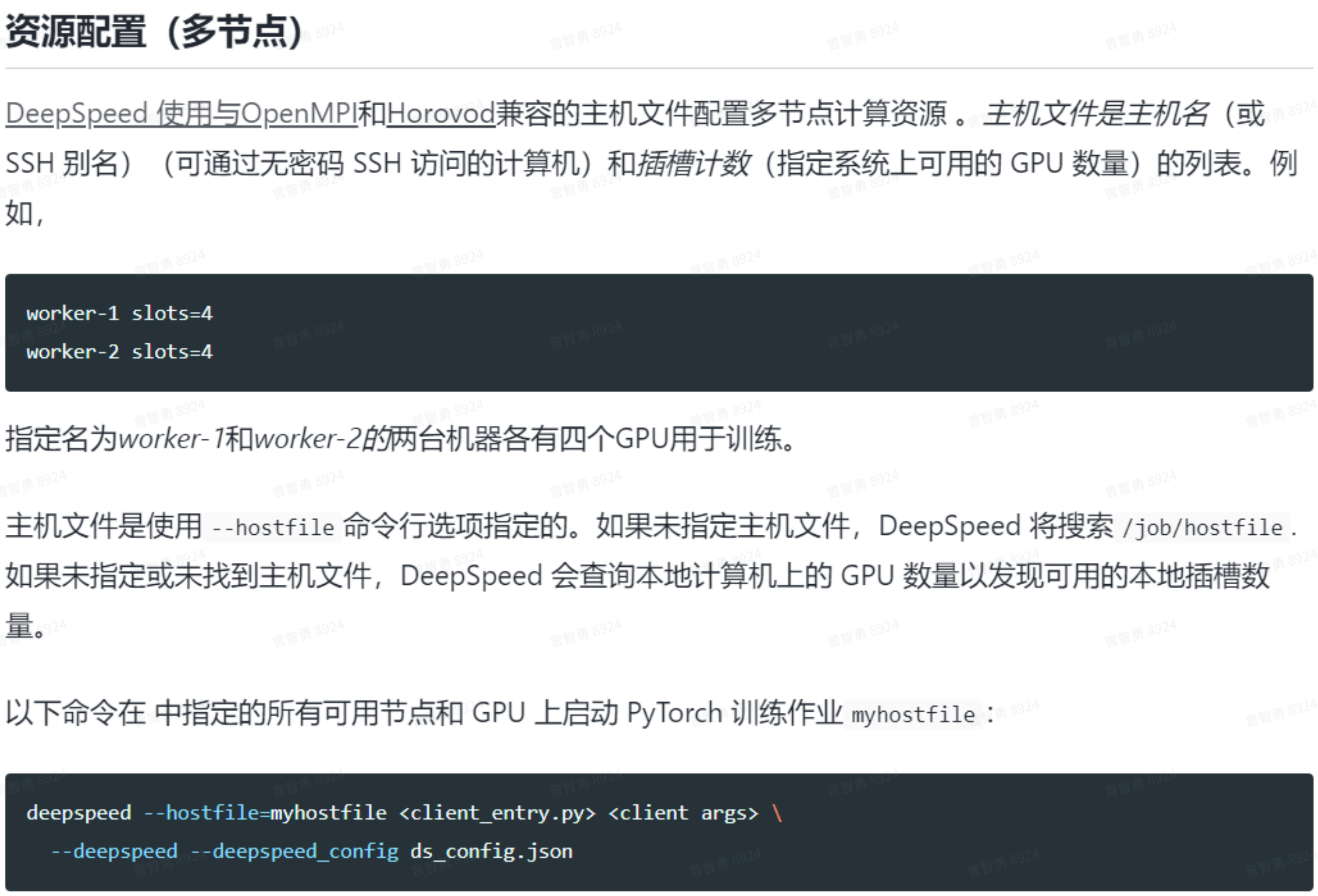
https://github.com/microsoft/DeepSpeed

2. 示例仓库

https://github.com/microsoft/DeepSpeedExamples

3. 使用文档

https://www.deepspeed.ai/getting-started/#mpi-compatibility



资源配置（多节点）

DeepSpeed 使用与OpenMPI和Horovod兼容的主机文件配置多节点计算资源。*主机文件是主机名*（或 SSH 别名）（可通过无密码 SSH 访问的计算机）和*插槽计数*（指定系统上可用的 GPU 数量）的列表。例如，

```
worker-1 slots=4
worker-2 slots=4
```

指定名为*worker-1*和*worker-2的*两台机器各有四个GPU用于训练。

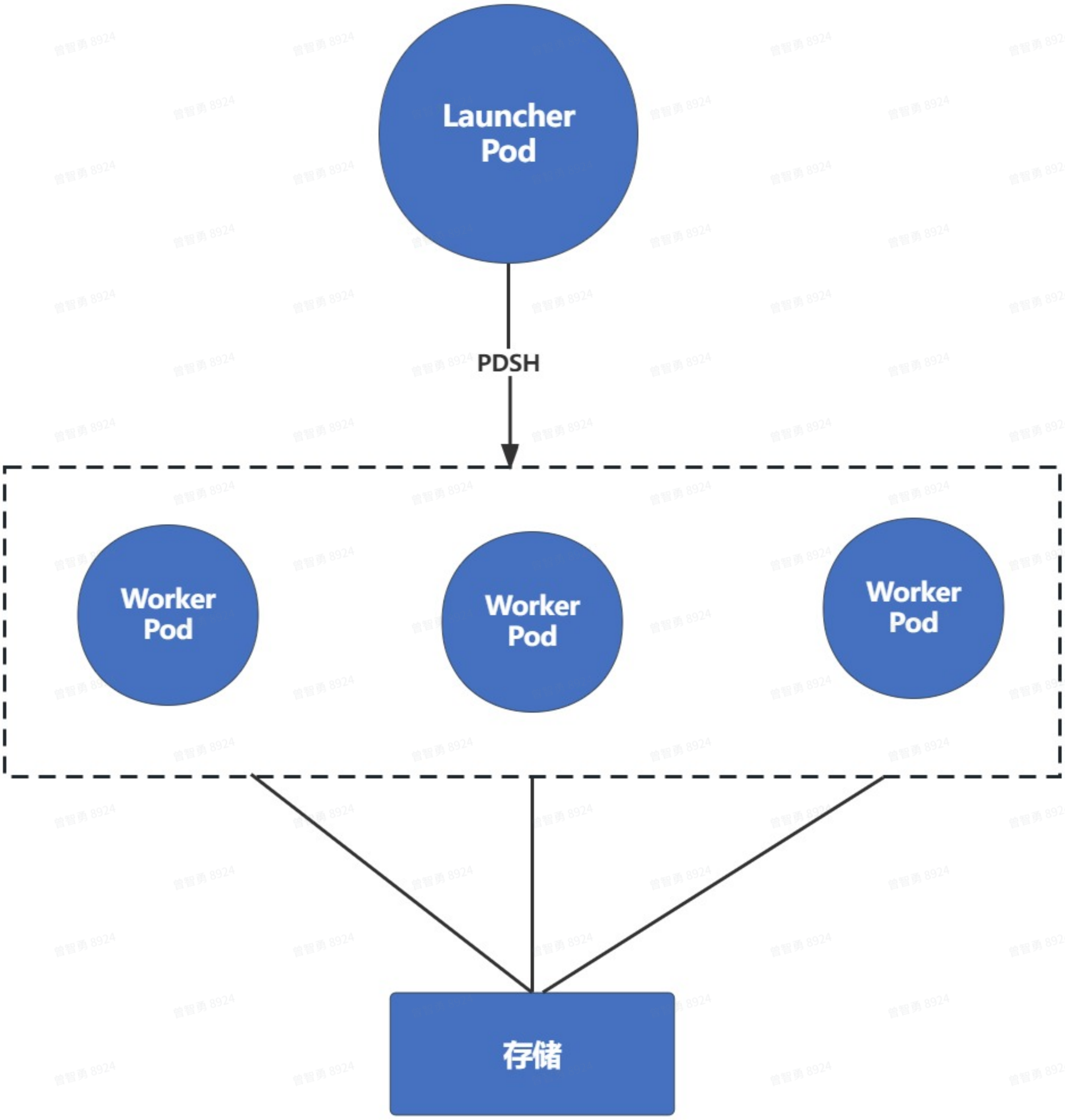
主机文件是使用 `--hostfile` 命令行选项指定的。如果未指定主机文件，DeepSpeed 将搜索 `/job/hostfile`。如果未指定或未找到主机文件，DeepSpeed 会查询本地计算机上的 GPU 数量以发现可用的本地插槽数量。

以下命令在 中指定的所有可用节点和 GPU 上启动 PyTorch 训练作业 `myhostfile`：

```
deepspeed --hostfile=myhostfile <client_entry.py> <client args> \
  --deepspeed --deepspeed_config ds_config.json
```

二. Kubernetes方案

1. 架构



架构要求

1. 共享存储
2. SSH直接访问
3. 获取hostfile
4. 分配 Worker Pod到GPU节点

2. 准备内容

Deepspeed镜像：docker.dm-ai.cn/public/deepspeed:v0.9.5-3

3. Deepspeed 示例

创建PVC和PV（存储）

https://gitlab.dm-ai.cn/devops/model/deepspeed-demo/-/blob/master/deepspeed-test-pv.yml

创建secret（用于ssh访问）

https://gitlab.dm-ai.cn/devops/model/deepspeed-demo/-/blob/master/deepspeed-test-secret.yml

创建configmap (hostfile)

https://gitlab.dm-ai.cn/devops/model/deepspeed-demo/-/blob/master/deepspeed-test-configmap.yml

创建Worker Pod

https://gitlab.dm-ai.cn/devops/model/deepspeed-demo/-/blob/master/deepspeed-test-pod.yml

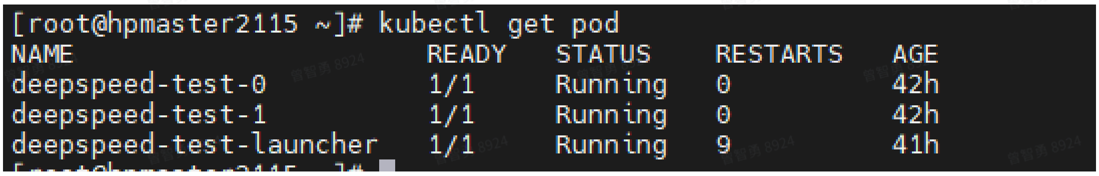
创建Headless Service（用于名称解析）

https://gitlab.dm-ai.cn/devops/model/deepspeed-demo/-/blob/master/deepspeed-test-service.yml

创建launcher Pod（执行任务）

https://gitlab.dm-ai.cn/devops/model/deepspeed-demo/-/blob/master/deepspeed-test-launcher-pod.yml

效果：



```
[root@hpmaster2115 ~]# kubectl get pod
NAME                         READY   STATUS    RESTARTS   AGE
deepspeed-test-0             1/1     Running   0          42h
deepspeed-test-1             1/1     Running   0          42h
deepspeed-test-launcher      1/1     Running   9          41h
[root@hpmaster2115 ~]#
```

4. 训练平台需解决问题

问题一：创建基于hostfile的configmap

问题二：如何获取到训练任务状态?

问题三：是否在任务结束后清理Pod实例?

三. Arena（待定）

1. 介绍

https://help.aliyun.com/document_detail/2249322.html

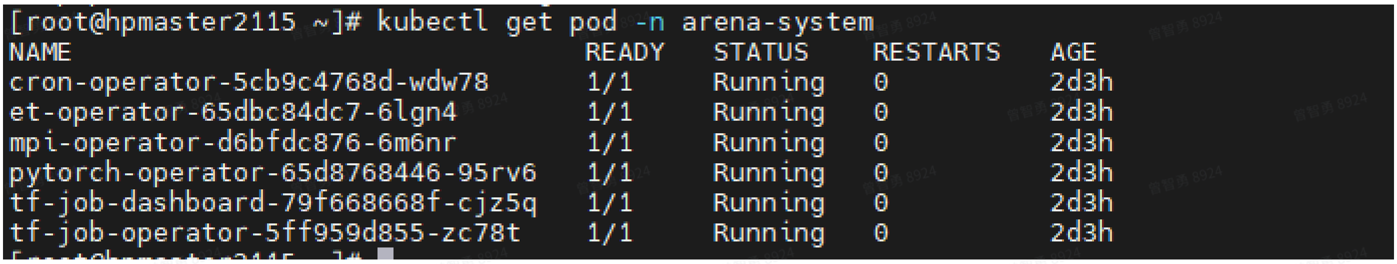
DeepSpeed分布式训练（阿里云）

2. 源码仓库

https://github.com/kubeflow/arena

3. 安装Arena

https://arena-docs.readthedocs.io/en/latest/installation/complete/



```
[root@hpmaster2115 ~]# kubectl get pod -n arena-system
NAME                                   READY   STATUS    RESTARTS   AGE
cron-operator-5cb9c4768d-wdw78         1/1     Running   0          2d3h
et-operator-65dbc84dc7-6lgn4           1/1     Running   0          2d3h
mpi-operator-d6bfdc876-6m6nr           1/1     Running   0          2d3h
pytorch-operator-65d8768446-95rv6      1/1     Running   0          2d3h
tf-job-dashboard-79f668668f-cjz5q      1/1     Running   0          2d3h
tf-job-operator-5ff959d855-zc78t       1/1     Running   0          2d3h
[root@hpmaster2115 ~]# 
```

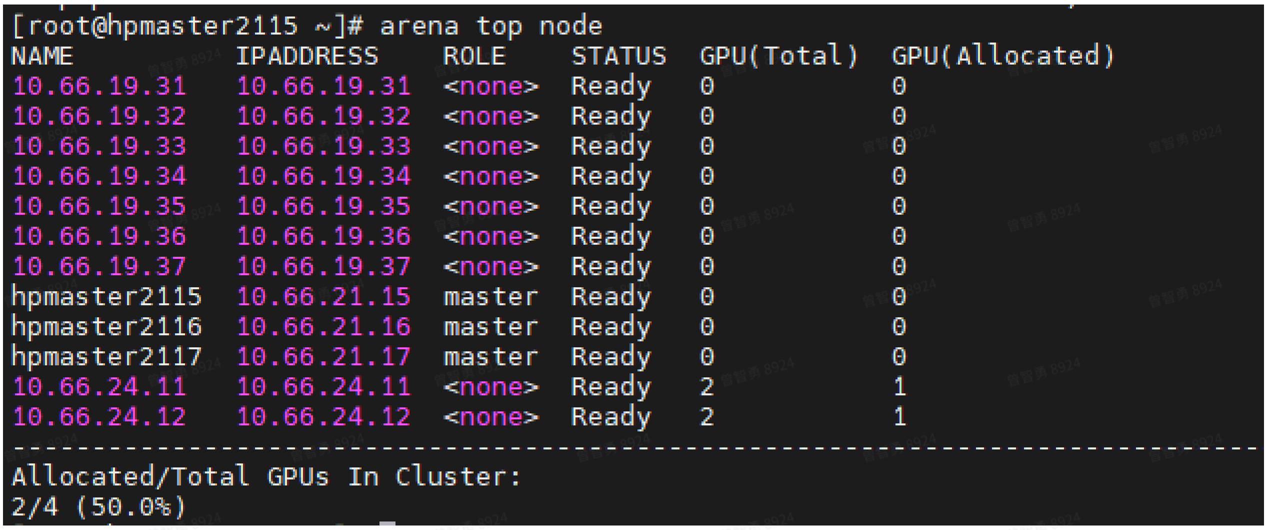
4. Arena使用手册

https://arena-docs.readthedocs.io/en/latest/training/

示例：

查看机器资源

```
$ arena top node
```



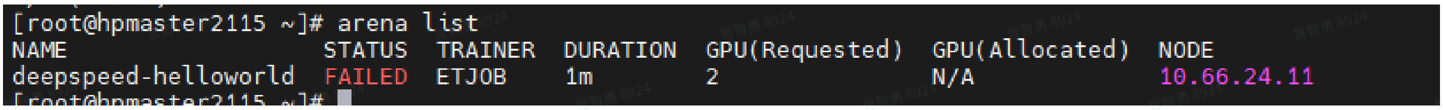
```
[root@hpmaster2115 ~]# arena top node
NAME          IPADDRESS    ROLE    STATUS  GPU(Total)  GPU(Allocated)
10.66.19.31   10.66.19.31  <none>  Ready   0           0
10.66.19.32   10.66.19.32  <none>  Ready   0           0
10.66.19.33   10.66.19.33  <none>  Ready   0           0
10.66.19.34   10.66.19.34  <none>  Ready   0           0
10.66.19.35   10.66.19.35  <none>  Ready   0           0
10.66.19.36   10.66.19.36  <none>  Ready   0           0
10.66.19.37   10.66.19.37  <none>  Ready   0           0
hpmaster2115  10.66.21.15  master  Ready   0           0
hpmaster2116  10.66.21.16  master  Ready   0           0
hpmaster2117  10.66.21.17  master  Ready   0           0
10.66.24.11   10.66.24.11  <none>  Ready   2           1
10.66.24.12   10.66.24.12  <none>  Ready   2           1
-----------------------------------------------------------------------------------
Allocated/Total GPUs In Cluster:
2/4 (50.0%)
```

创建一个etjob

```
$ arena submit etjob \
    --name=deepspeed-helloworld \
    --gpus=1 \
    --workers=2 \
    --image=docker.dm-ai.cn/devops/deepspeed:hello-deepspeed \
    --data=training-data:/data \
    --tensorboard \
    --logdir=/data/deepspeed_data \
    "deepspeed  --hostfile=/etc/edl/hostfile /workspace/DeepSpeedExamples/HelloD
```

查看任务

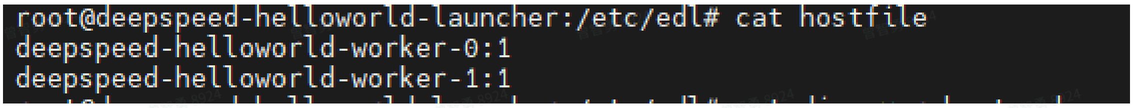
```
$ arena list
```



```
[root@hpmaster2115 ~]# arena list
NAME                  STATUS  TRAINER  DURATION  GPU(Requested)  GPU(Allocated)  NODE
deepspeed-helloworld  FAILED  ETJOB    1m        2               N/A             10.66.24.11
[root@hpmaster2115 ~]#
```

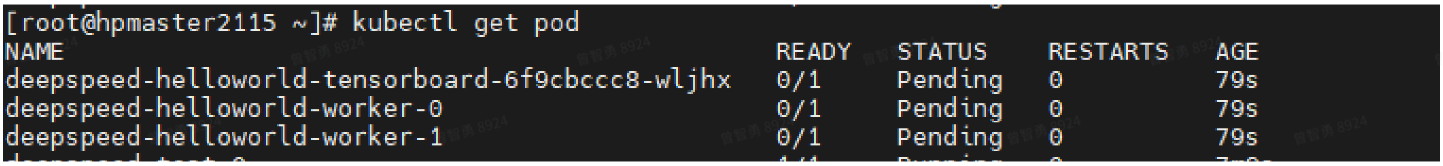
5. Arena存在问题：

问题一：hostfile格式错误



```
root@deepspeed-helloworld-launcher:/etc/edl# cat hostfile
deepspeed-helloworld-worker-0:1
deepspeed-helloworld-worker-1:1
```

问题二：有时候不会出现launcher Pod



```
[root@hpmaster2115 ~]# kubectl get pod
NAME                                        READY   STATUS    RESTARTS   AGE
deepspeed-helloworld-tensorboard-6f9cbccc8-wljhx   0/1     Pending   0          79s
deepspeed-helloworld-worker-0               0/1     Pending   0          79s
deepspeed-helloworld-worker-1               0/1     Pending   0          79s
```

5. Arena SDK (Goland)

https://arena-docs.readthedocs.io/en/latest/sdk/go/

参考资料：

1. https://zhuanlan.zhihu.com/p/256236705

超大模型分布式训练DeepSpeed教程

2. https://arena-docs.readthedocs.io/en/latest/

arena

3. https://zhuanlan.zhihu.com/p/276122469

【深度学习】— 分布式训练常用技术简介

4. https://zhuanlan.zhihu.com/p/79030485

腾讯机智团队分享--AllReduce算法的前世今生

5. https://www.youtube.com/watch?v=_NOk-mBwDYg&list=PLa85ZdUjfWS21mgibJ2vCvLziprjpKoW0&index=94

DeepSpeed 5a