End to End ML with Natural Language processing for training and inference on vSphere

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

1 Introduction

While virtualization technologies have proven themselves in the enterprise with cost effective, scalable and reliable IT computing, Machine Learning (ML) however has not evolved and is still bound to dedicating physical resources to obtain explicit runtimes and maximum performance. VMWare and Bitfusion has developed technologies to effectively share accelerators for machine learning over the network

2 Solution Components

2.1 NVIDIA v100 GPUs for Machine Learning

With the impending end to Moore's law, the spark that is fueling the current revolution in deep learning is having enough compute horsepower to train neural-network based models in a reasonable amount of time

The needed compute horsepower is derived largely from GPUs, which Nvidia began optimizing for deep learning since 2012. One of the latest in this family of GPU processors is the NVIDIA Tesla v100.

Figure 1: The NVIDIA v100 GPU

NVIDIA® Tesla® V100 Tensor Core is the most advanced data center GPU ever built to accelerate AI, high performance computing (HPC), data science and graphics. It's powered by

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

NVIDIA Volta architecture, comes in 16 and 32GB configurations, and offers the performance of up to 100 CPUs in a single GPU. Data scientists, researchers, and engineers can now spend less time optimizing memory usage and more time designing the next AI breakthrough. (Source: NVIDIA)

2.2 Bitfusion FlexDirect

Bitfusion FlexDirect that will evolve to become part of vSphere and NVIDIA GPU accelerators can now be part of a common infrastructure resource pool, available for use by any virtual machine in the data center in full or partial configurations, attached over the network. The solution works with any type of GPU server and any networking configuration such as TCP, RoCE or InfiniBand. GPU infrastructure can now be pooled together to offer an elastic GPU as a service, enabling dynamic assignment of GPU resources based on an organization's business needs and priorities. Bitfusion FlexDirect runs in the user space and doesn't require any changes to the OS, drivers, kernel modules or AI frameworks. (Source: Bitfusion)

2.3 Natural Language Processing

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the *natural language*. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. (Source: BecomingHuman.ai)

2.4 The Machine Learning Process:

In machine learning, we are able to leverage existing data to learn from and predict an outcome without using human judgement or manual rules. The processing steps in machine learning include:

- Data gathering
- Data preparation
- Choice of model
- Training
- Model Validation
- Hyperparameter tuning

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

- Inference

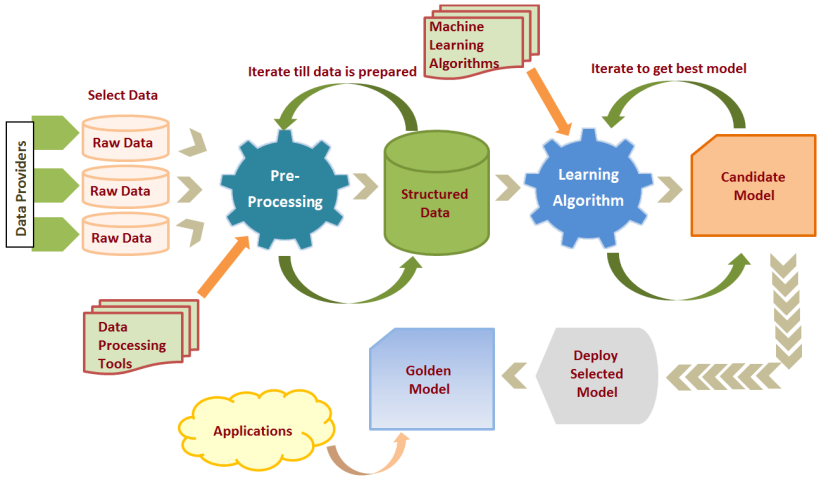

Figure 2: Machine Learning Pipeline. (Source: eLearning Industry)

3 The Solution:

This solution showcases the use of VMware Essential PKS for end to end machine learning. Machine learning will use natural language processing to develop an accurate model for movie reviews in this solution. The data processing and collection phases are already accomplished by IMDB and hence the solution will address the training, validation, tuning and the inference phases of the ML process pipeline.

3.1 VMware Essentials PKS:

There are hundreds of tools in the cloud native ecosystem, and new ones are rapidly emerging from the open-source community. Building flexibility into infrastructure is key to ensure adoption of new technologies to run workloads anywhere: on premises, in public clouds, or in a hybrid-cloud environment. VMware Essential PKS provides upstream **Kubernetes** binaries and expert support, enabling enterprise to build a flexible, cost-effective cloud native platform.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

VMware Essentials PKS offers the following core advantages:

- Provides access to the energy and innovation stemming from the open source community
- Provides the flexibility to run Kubernetes across public clouds and on-premises infrastructure
- Can grow the cluster count from tens to hundreds to thousands without fear of spiraling costs

4 Solution Architecture:

Typically, in enterprise all the data is aggregated into a data lake or database within an on-premises centralized datacenter that has advanced computing capabilities. To simulate this situation, in this solution we will do the training, evaluation and tuning phases in an on-premise virtualized datacenter with NVIDIA GPU compute capability.

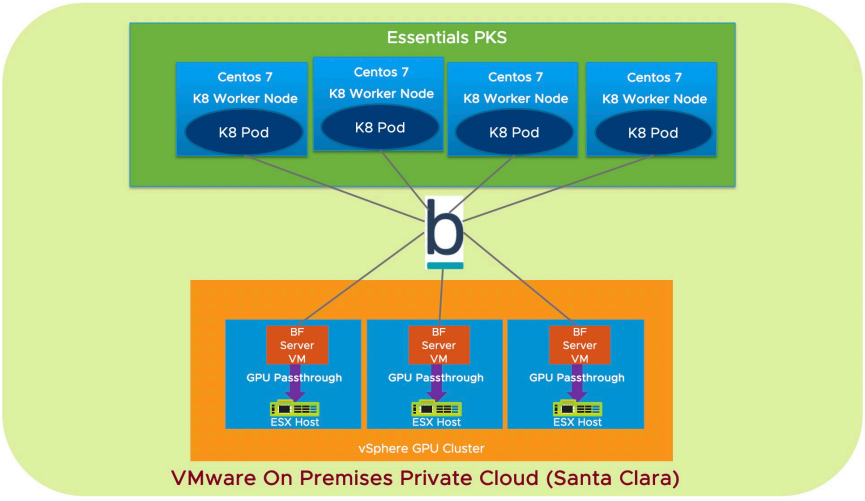

Figure 3: Logical Schematic of the Training Infrastructure

The inference process for a big enterprise happens in edge locations which typically have limited compute capabilities. In this solution, we will use a VMware cloud on AWS remote location with CPU only virtual machines as the edge location for inference.

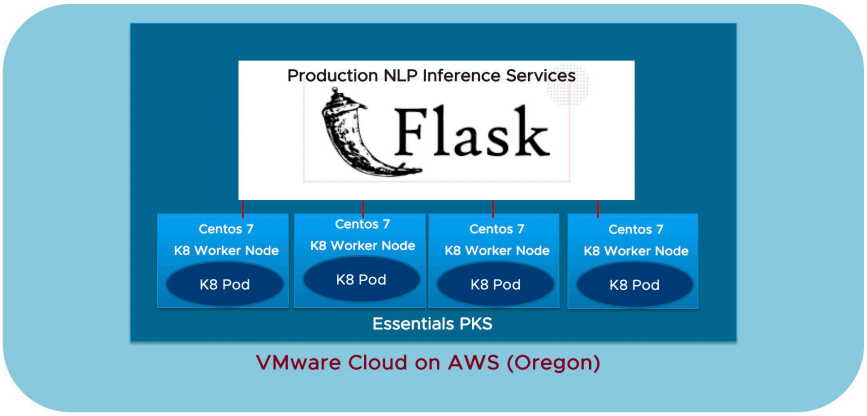

Figure 4: Logical Schematic of the Inference Infrastructure

The newer more accurate models are transferred from the training site to the Inference site regularly. Data that has been inaccurately classified in the production inference site is transferred to the training site to learn from and make the model more accurate. The components of the solution include the following:

4.1.1 Virtual Infrastructure Components:

The virtual infrastructure used to build the solution is shown below:

HW Components	
Cluster	**6 X Dell R740 node cluster**
Storage	**Pure M50 Fibrechannel for VMFS**
Storage Fabric	**Brocade Fibrechannel**
Network	**Extreme VDX Switches**
GPU	**NVIDIA V100**

Table 1: HW components of the solution

The VMware SDDC and other SW components used in the solution are shown below:

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Software	Version
ESXi	**6.7 U2**
vCenter	**6.7 U2**
OS	**Centos 7.6**
Bitfusion FlexDirect	**1.11.7**
MLPERF	**--**
Essentials PKS	**1.13.2**

Table 2: SW components of the solution

4.2 Logical Architecture of Solution Deployed:

Essentials PKS provided the framework to create Kubernetes clusters seamlessly working with the VMware SDDC components. A logical schematic of the Kubernetes clusters for training on-premises and Inference in the Cloud are shown.

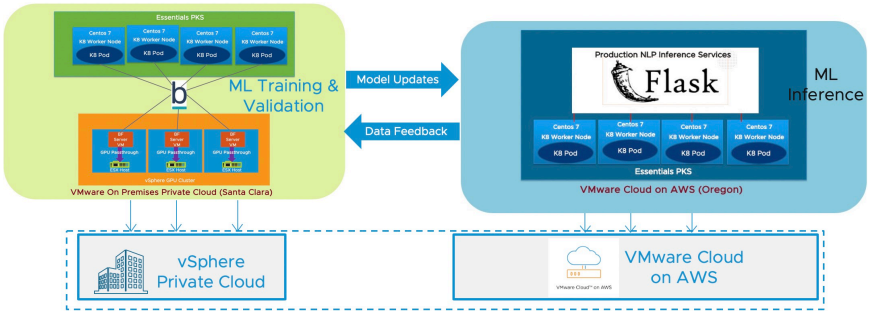

Figure 5: Logical Schematic of Solution

4.3 Summary of steps for Solution Deployment:

- VMware Essentials PKS was installed on a vSphere Cluster with six Dell R740 servers.
- Two of these nodes had one NVIDIA GPU V100 cards each.
- FlexDirect server was deployed on two Linux virtual machines, each attached to an NVIDIA GPU.
- A Kubernetes cluster with four worker nodes and one master was created with Essentials PKS

- Docker images were created with the components shown in Appendix A
- Paddle (Parallel Distributed Deep Learning) an easy-to-use, efficient, flexible and scalable deep learning platform was deployed
- Flask is a micro web framework written in Python was installed in the inference site
- MLPerf is a benchmark suite for measuring how fast systems can train models to a target quality metric. Each MLPerf Training benchmark is defined by a Dataset and Quality Target.
- Sentiment Analysis is a binary classification task. It predicts positive or negative sentiment using raw user text. The IMDB dataset is used for this benchmark.
- IMDB Data set obtained from http://ai.stanford.edu/~amaas/data/sentiment/

4.4 Sentiment Analysis basics:

Sentiment Analysis is a binary classification task. It predicts positive or negative sentiment using raw user text. The IMDB dataset is used for this benchmark. The methodology and code used was from the following GitHub site.
https://github.com/mlperf/training/tree/master/sentiment_analysis

The model used is a convolution neural network (CNN) based on "Effective use of word order for text categorization with convolutional neural networks"

5 Modeling, validation and inference:

The IMDB Dataset which provides 50000 movie reviews for sentiment analysis was used for the training and evaluation phase. Below are some of the aspects of the dataset and it processing parameters.

- **Data preprocessing**
  - The dataset isn't preprocessed in any way.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

- **Training and test data separation**
  - The entire dataset is split into training and test sets. 25000 reviews are used for training and 25000 are used for validation. This split is pre-determined and cannot be modified.
- **Training data order**
  - Training data is traversed in a randomized order.
- **Test data order**
  - Test data is evaluated in a fixed order.
- **Quality target**
  - Average accuracy of 90%
- **Inference**
  - Independent dataset was used to showcase Inference (evaluation)

5.1 Training:

The training infrastructure leveraged the following:

- Kubernetes cluster based on VMware Essentials PKS on CentOS run on-premises in the Santa Clara Datacenter
- Training leveraged Bitfusion based access to remote GPU resources over the network

The initial runs compared CPU against GPU based model training. The training using GPU enabled workers were seen to be approximately 160X faster than the CPU based training for the CNN model used for sentiment analysis. This clearly indicated the need for GPU for these models. The benchmark code for sentiment analysis from mlperf.org, was leveraged to create a trained and evaluated model for sentiment analysis based on IMDB dataset. The model was saved on disk. This part of the solution was run on VMware Essentials PKS on CentOS run on-premises in the Santa Clara Datacenter.

The models were tuned over multiple iterations of tuning until the quality target of 90% accuracy was achieved.

```
ime API version: 9.0
W0725 07:15:53.896489    36 device_context.cc:267] device: 0, cuDNN Version: 7.4.
2019-07-25 07:15:53.914494 : Begin train_loop!
2019-07-25 07:15:53.914537 : Directory name for saving model is: understand_sentiment_conv.inference.model
2019-07-25 07:15:53.914552 : Target accuracy value is is: 90.6
2019-07-25 07:15:53.914563 : Start epoch #: 0
2019-07-25 07:16:34.164285 : End of epoch #: 0
2019-07-25 07:16:34.164351 : Stats:
Epoch =0, train-accuracy =83.12818879375652, train-loss =0.3750325126611457, validation-accuracy =89.34470664481728, validatio
n-loss =0.2564125956245223

2019-07-25 07:16:34.164366 : Start epoch #: 1
2019-07-25 07:17:14.494292 : End of epoch #: 1
2019-07-25 07:17:14.494346 : Stats:
Epoch =1, train-accuracy =93.33466200196014, train-loss =0.1776895931332695, validation-accuracy =90.45758928571429, validatio
n-loss =0.23244875846240592

2019-07-25 07:17:14.494360 : Start epoch #: 2
2019-07-25 07:17:55.084813 : End of epoch #: 2
2019-07-25 07:17:55.084868 : Stats:
Epoch =2, train-accuracy =97.80691965502135, train-loss =0.07917609841239695, validation-accuracy =90.6688456632653, validatio
n-loss =0.23725571504280885

2019-07-25 07:17:55.084882 : Target accuracy goal (b)reached!
2019-07-25 07:17:55.084891 : Saving model into directory: understand_sentiment_conv.inference.model
ENDING TIMING RUN AT 2019-07-25 07:17:55 AM
RESULT,sentiment,30,164,,2019-07-25 07:15:11 AM
root@mlperfsc9:/workspace/sentiment_analysis/paddle#
```

Figure 6: Output from a training run

5.2 Inference:

The inference infrastructure leveraged the following:

- Inference leveraged Kubernetes cluster based on Essentials PKS running on VMware Cloud on AWS in the Oregon region
- Inference was run on a pod that had only CPU capability – there was no GPU access

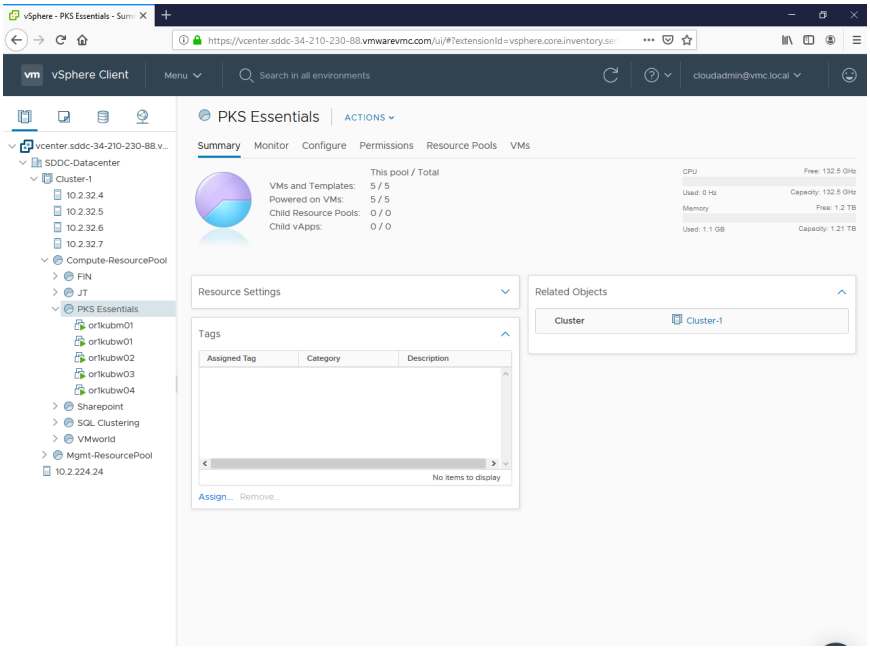

Figure 7: VMC on AWS with Essentials PKS used for Inference

6 Results:

The resulting model from the training was transferred to Essentials PKS running on VMware Cloud on AWS at an Oregon Datacenter. The inference is run on a CPU only VMware Cloud AWS infrastructure on Essentials PKS worker nodes.

6.1 Sentiment Analysis application in production:

A Flask based web interface is used to frontend the sentiment analysis model that was created during training. The production application can be used analyze movie reviews individually to gauge the sentiment of the reviewer

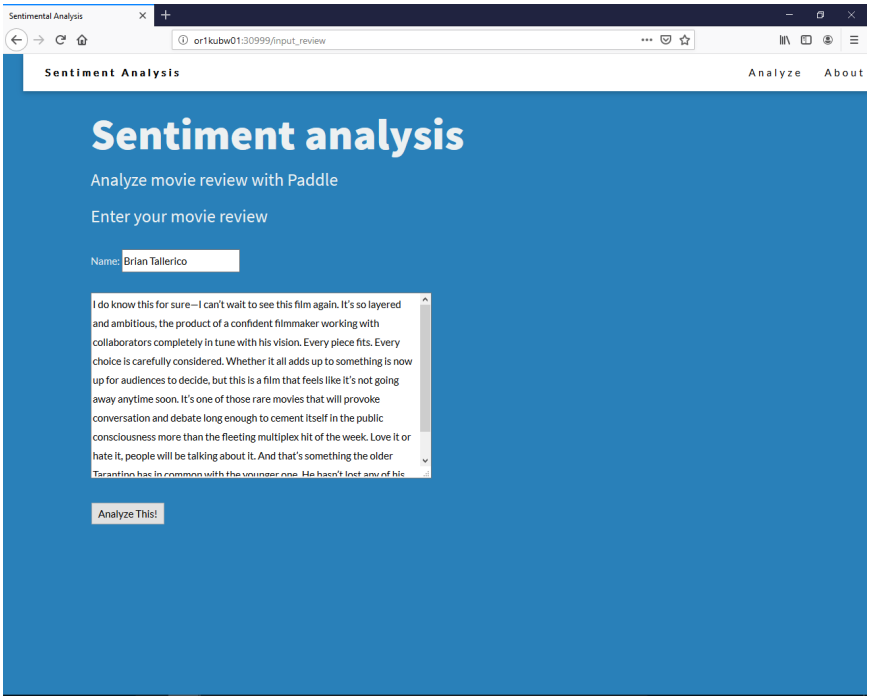

Figure 8: Data Entry screen for sentiment analysis

1. The website shown takes movie review as input
2. It applies the model to do machine learning inference on the review

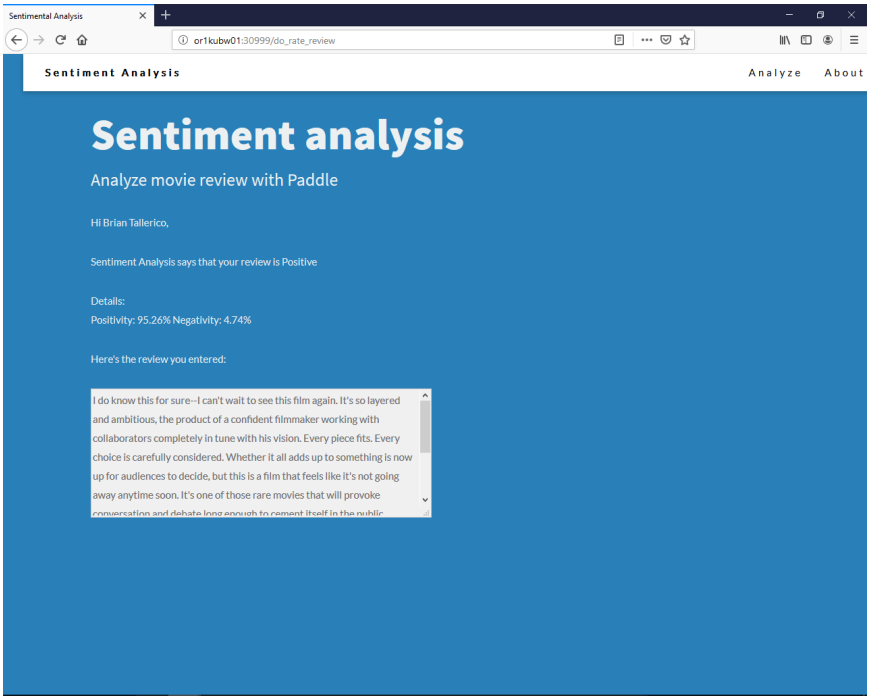

Figure 9: Screen showing results of the sentiment analysis from the inference engine

3. It displays inference results – i.e., show if the review is positive or negative
4. In the example a movie review from "Once Upon a Time in Hollywood (Movie 2019)" from a popular reviewer is pasted into the web page
5. The analysis shows that the sentiment is overwhelmingly positive (95%)

7 Conclusion:

The solution clearly demonstrated end to end machine learning on the vSphere platform. VMware Essentials PKS was successfully deployed with NVIDIA GPU and Bitfusion to provide for a high-performance container based machine learning platform. vSphere solutions can be leveraged during different stages of the ML Pipeline with Natural Language Processing. Training, evaluation and inference processes in the ML workflow on vSphere were effectively demonstrated.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Appendix A: Docker File used in Solution

```
#
# This example Dockerfile illustrates a method to install
# additional packages on top of Ubuntu 16.0.4 container image.
#
# To use this Dockerfile, use the `docker build` command.
# See https://docs.docker.com/engine/reference/builder/
# for more information.
#

#Download sidgoyal78/paddle:benchmark12042018
#FROM sidgoyal78/paddle:benchmark12042018
FROM paddlepaddle/paddle:latest-gpu-cuda9.0-cudnn7

RUN apt-get update && apt-get install -y --no-install-recommends \
   vim \
   && \
   rm -rf /var/lib/apt/lists/

# Install flexdirect
#RUN cd /tmp && wget -O installfd    getfd.bitfusion.io && chmod +x installfd &&
./installfd  -v fd-1.11.2  -- -s -m binaries
# Use flexdirect v1.11.7 instead of v1.11.2
RUN cd /tmp && wget -O installfd    getfd.bitfusion.io && chmod +x installfd &&
./installfd  -v fd-1.11.7  -- -s -m binaries

#
# IMPORTANT:
# Build docker image from the dir: sc2k8cl3:/data/tools/dev
#

RUN mkdir -p /workspace
RUN mkdir -p /root/.cache/paddle/dataset/imdb

COPY ./sentiment_analysis/ /workspace/sentiment_analysis
COPY  ./aclImdb_v1.tar.gz /root/.cache/paddle/dataset/imdb

WORKDIR /workspace/sentiment_analysis/
```

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Appendix B: YAML File used for the Kubernetes pods

```
apiVersion: v1
kind: Pod
metadata:
  name: mlperfsc9
spec:
  volumes:
    - name:  bitfusionio
      hostPath:
        path: /etc/bitfusionio
    # - name: nfs
      # nfs:
        # # FIXME: use the right name
        # #server: nfs-server.default.kube.local
        # server: "172.16.35.40"
        # path: "/GPU_DB"
        # readOnly: false
  containers:
  - name: mlperfsc9
    #image: ubuntu:16.04
    #image: sc2harbor1.vslab.local/library/mlperf:18.06-py3
    #image: sc2harbor1.vslab.local/library/mlperf:senti
    #
    # Ash comments: CUDA9-cudnn7 - required for this GPU
    #
    #image: sc2harbor1.vslab.local/library/mlperf:senti_cuda9
    image: pmohan77/mlperf:senti_cuda9
    "imagePullPolicy": "Always"
    command: ["/bin/bash", "-ec", "while :; do echo 'My pod name is
${MY_NODE_NAME} .'; sleep 300 ; done"]
    #
    env:
      # - name: MY_NODE_NAME
        # valueFrom:
          # fieldRef:
            # fieldPath: metadata.name
      - name: POD_UID
        valueFrom:
          fieldRef:
            apiVersion: v1
            fieldPath: metadata.uid
      - name: IMAGENET_HOME
        "value": "/tmp/understand_sentiment_conv.inference.model"
        #value: "/gpu_data/imagedata/tiny-imagenet-200/tiny-imagenet-200"
    volumeMounts:
      # name(s) must match the volume name(s) above
      - name: bitfusionio
        mountPath: /etc/bitfusionio
        readOnly: true
```

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

```
    # - name: nfs
      # mountPath: "/gpu_data"
      # mountPath: "/gpu_data/imagedata/logs"
      # mountPath: "/logs"
      # subPath: $(POD_UID)
  restartPolicy: Never
```