Serverless Kubernetes boosts AI Business

Jian Huang, Container architect at Huawei Cloud
What ‘serverless’ really means is that, as a developer you don’t have to think about those servers. You just focus on code.
Kubernetes becomes the most suitable container base platform to build the whole serverless ecosystem.
Native Serverless Kubernetes:
- Lightweight architecture upon pure Kubernetes.
- Build on baremetal machines directly.
- Connect to Huawei Cloud’s vpc network and storage service.
- 95%+ API Compatible, except HostPath, etc.
- Kubernetes Service Support by IPVS

Tenant Isolation:
1. Using Kata/RunV Container to improve container security.
2. Support VPC network and traffic QoS.
3. Namespace level authorization.
Advantages

• **No Care About Building Kubernetes Clusters and Nodes**

• **Just Use it!**

• **Focus on Business Code!**

• **Pretty Good for Job Workloads**
  • Pay as you use and save money!
  • Run Immediately! No need to create and destroy clusters when you run your jobs, such as **AI training workloads**, human genome, etc.
Huawei Cloud AI Services over Kubernetes

**General APIs**
- ASR
- TTS
- Image
- NLP
- OCR
- Face
- Moderation
- AIS

**Advanced APIs**
- CBS
- ImageSearch
- VCM
- VCT
- IDS
- VGS
- VCC
- VAR

**Pre-integrated Solutions**
- City
- Internet
- Home
- Vehicle
  - Logistics
  - Healthcare
  - Campus
  - Manufacturing

**Basic Platforms**
- ModelArts
  - ExeML
  - GES
  - DLS
  - MLS
  - Batch
  - UPredict
  - RLS
- MindSpore

**Kubernetes (first-class citizen)**
- Ascend
- GPU
What we do for AI?

**AI Computing:**
- Support NVidia GPU (P4, P100, V100)
- Support Huawei Ascend 310

**Network:**
- Support VPC Network
- Support 100G InfiniBand Network

**Storage:**
- Support Cloud Storage: Huawei Cloud EVS, OBS, SFS
- Support Local Storage

**Scheduling:**
- Batch job scheduler, gang scheduler
- High-volume workloads placement
We Support NVIDIA device direct pass through both on traditional Docker and kata containers. In Kata Container, it has already support VFIO for PCI device pass-through.
AI Inference Applications usually need to share NVidia P4 GPU devices in order to improve utilization.
We didn’t really use it as NVIDIA Grid in production as the license is expensive. Looking forward to NVidia T4!
AI Computing – Huawei Ascend AI Chip

Ascend 310
Half-Precision (FP16): 8 TeraFLOPS
Integer-Precision (INT8): 16 TeraOPS
Suitable for Inference

Ascend 910
2019Q2
Half-Precision (FP16): 256 TeraFLOPS
Integer-Precision (INT8): 512 TeraOPS
Suitable for Training
Currently CCI has already support Ascend 310 for AI inference scenarios.

Each machine has 8/16 Ascend 310 chips. Each Ascend 310 has 4 cores. So we can support over 32 multi-tenant containers on a single machine now.
AI Computing – Distributed AI Training Models

Tensorflow Parameter Server

Horovod Ring AllReduce
In order to improve AI training jobs’s linear speedup:

- Support 100G inter-pod RDMA Network
- Support multiple container on a single node by InfiniBand SR-IOV
- Support Namespace level InfiniBand network isolation

We will support Huawei’s ROCE network card in 2019.
Scheduler – Batch Scheduler?

Kubernetes official job workload is EASY.

AI workloads need batch job scheduling:
• Gang Scheduling: Distributed AI training job usually need to be started together. Only Kubernetes Scheduler know about the resource allocation.
• One or more Kubernetes jobs consist of AI training workload, we need some CRD objects to manage that.
• Large amount of jobs need to be start together.
• Multiple Tenant Priorities.

Scheduler

KubeCon
CloudNativeCon

China 2018

KubeBatch
https://github.com/kubernetes-sigs/kube-batch

Poseidon
https://github.com/kubernetes-sigs/poseidon
Scheduler – Poseidon

Poseidon: A Firmament-based Kubernetes scheduler
Firmament: A high quality scheduler based on flow network, implement MCMF algorithm (Min Cost – Max Flow).

https://github.com/kubernetes-sigs/poseidon
BENCHMARKING OF SCHEDULING TIME OF PODS WITH CPU/MEM REQUIREMENTS ONLY

Firmament
Default

number of nodes/number of pods

Scheduling time in milliseconds

0 10000 20000 30000 40000 50000 60000 70000 80000

200/3420 400/6840 600/10260 800/13680 1000/17100

9785 347 715 2095 2416 3235 9785 11408 40521 66926

30245

Scheduler – Benchmark of Poseidon
Huawei Serverless AI infrastructure base on K8S

KUBE BATCH Controller (Batch Job Scheduler For ML/DL/BigData/HPC Workload)

K8S MASTER (CCI)

Queue Jobs

AI Node
- NVIDIA GPU Container
  - Tensorflow/Caffe2/OPENMPI
  - CuDNN
  - CUDA

AI Node
- Ascend Container
  - Tensorflow/Caffe2/OPENMPI
  - Ascend Driver

High Speed Inter-Network (100G)

High Performance Storage
- OBS
- EVS
- EFS
- Local SSD Temporary Storage

Opensource AI Frameworks
- TensorFlow
- Caffe2
- mxnet
- PyTorch
- OpenMPI
- Tensorflow/Caffe2/OPENMPI

Huawei ModelArts
- ExeML
- GES
- DLS
- MLS
- Batch
- UPredict
- RLS

Huawei HiLens
- Huawei ModelArts
- ExeML
- GES
- DLS
- MLS
- Batch
- UPredict
- RLS

K8S AI Framework
- KubeCon
- CloudNativeCon
- China 2018
Accelerate Huawei Cloud AI Products

Huawei MoXing Distributed Deep Learning Framework

- Hybrid parallelism
- Optimized ASGD
- Deep Gradient Compression
- Convolutional Neural Networks
- Auto Speed up

1000 GPU Linear SpeedUp 0.8+

Kubernetes for AI

Training ImageNet 1.28 million dataset with 128 GPU Card, 12 minutes