geminiopencloud
雙子星雲端運算

將IaaS導入進K8S-淺談KubeVirt經驗

顏安劭 (Eddie Yen)
系統架構FE

About Me

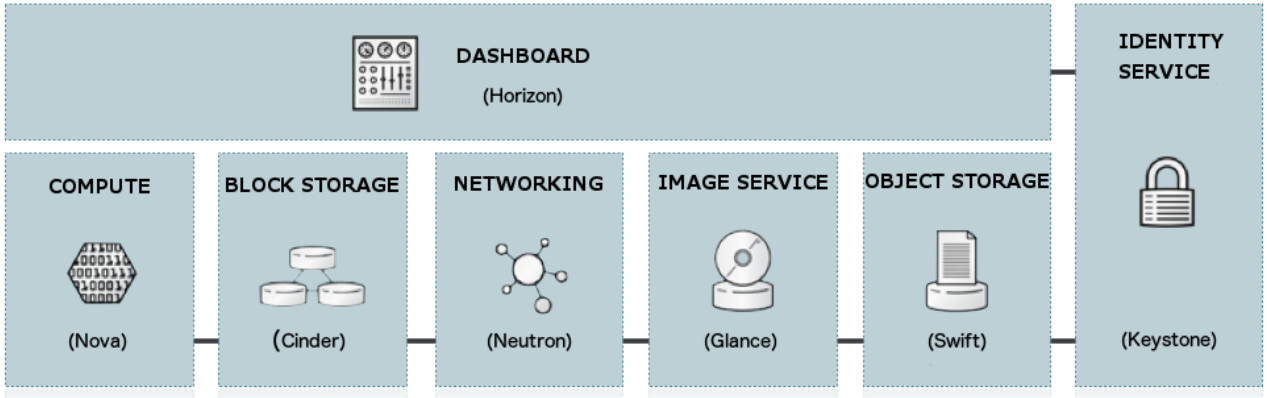
- 在雙子星雲端擔任系統架構建置與維護工程師，公司為CNCF認證之Kubernetes服務提供商(KCSP)
- 對電腦硬體、虛擬化技術、網路架構等有興趣與研究
- 建置與維運OpenStack、Ceph與K8S的數年經驗
- 參與不少專案，為數個客戶建置與協助維護私有雲環境

Agenda

- KubeVirt專案簡介
- IaaS在維運及轉型上的難處
- KubeVirt所解決的問題
- KubeVirt架構與相關功能

傳統IaaS平台在維運上的難處

- 部分IaaS服務平台並非都適合所有中小型環境
  - 內部服務的相依性太高
  - 建置作業太長，需考量不少因素
  - 事後改動難度高，缺乏彈性
  - 架構龐大，增加維護與學習成本
  - 簡單的目標都需要不少操作



將IaaS導入到容器的難處

- 服務在各方面評估下還是VM為佳，不適合容器化
- 轉移可能耗時，因成本等考量，不想要長時間維持多平台
- 需對服務重新規劃完全不一樣的安全設計
- 相依的服務無容器可用或不合需求

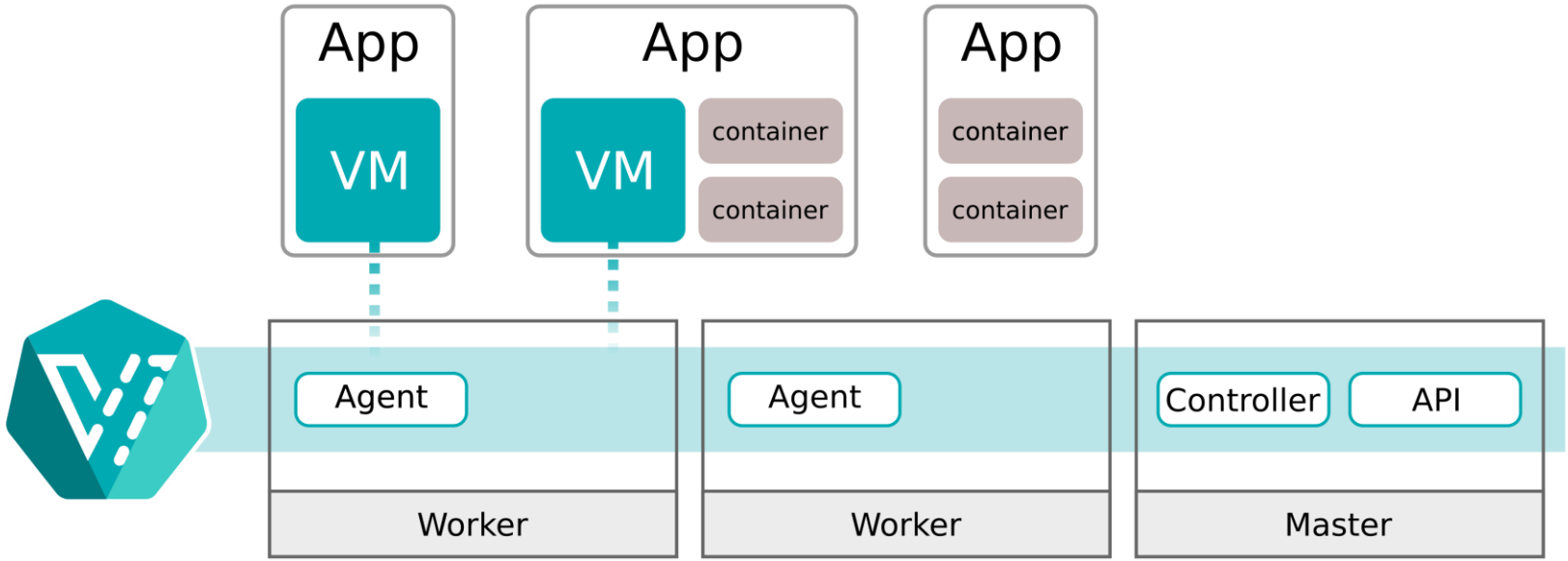
KubeVirt專案簡介

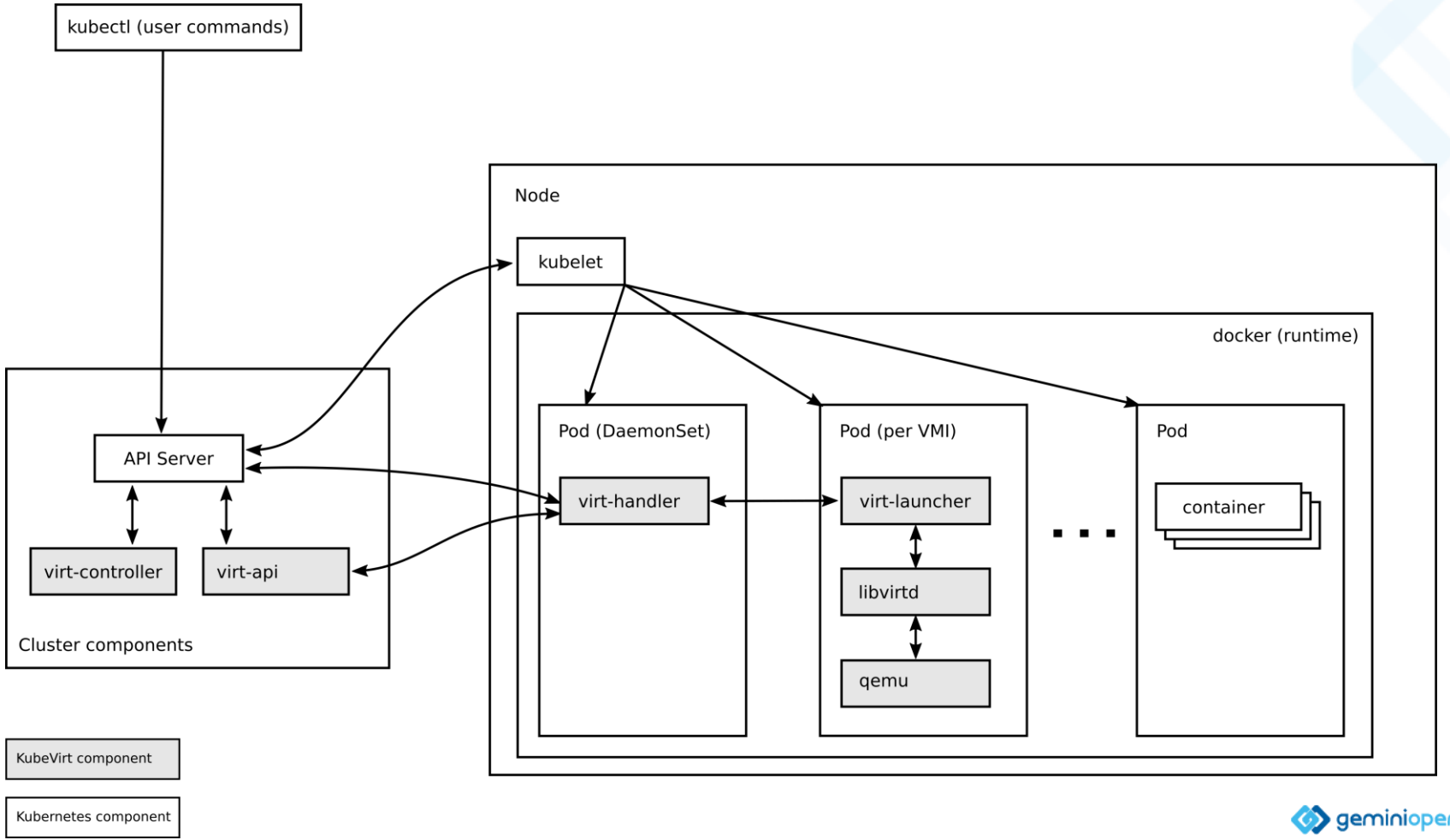
- 由RedHat於2017年建立的K8S專案
- 原先是讓自家雲端整合平台建立可以開VM的K8S服務，後來開源開放
- 同時是CNCF的Incubating階段專案
- 目前已發展到0.57版，支援K8S 1.20+
- 部分平台如Platform9也加入了此專案作為服務的一部份

KubeVirt所解決的問題

- 幫助服務轉型
  - 讓容器與VM同時共存單一平台
  - 可直接測試與微服務的相容性與負載
  - 無須依賴容器版套件/服務
- 增加安全與穩定性
  - 提升服務安全與資料保護
  - 對資源用量變化很大的服務，可降低干擾或服務中斷風險
- 將IaaS改為以K8S為基底服務
  - 減少以往IaaS的維運難處
  - 享有部分K8S自有的功能或特性
  - 搭配容器可做出混合型應用

KubeVirt架構


App
VM
App
VM
container
container
App
container
container
Agent
Agent
Controller
API
Worker
Worker
Master



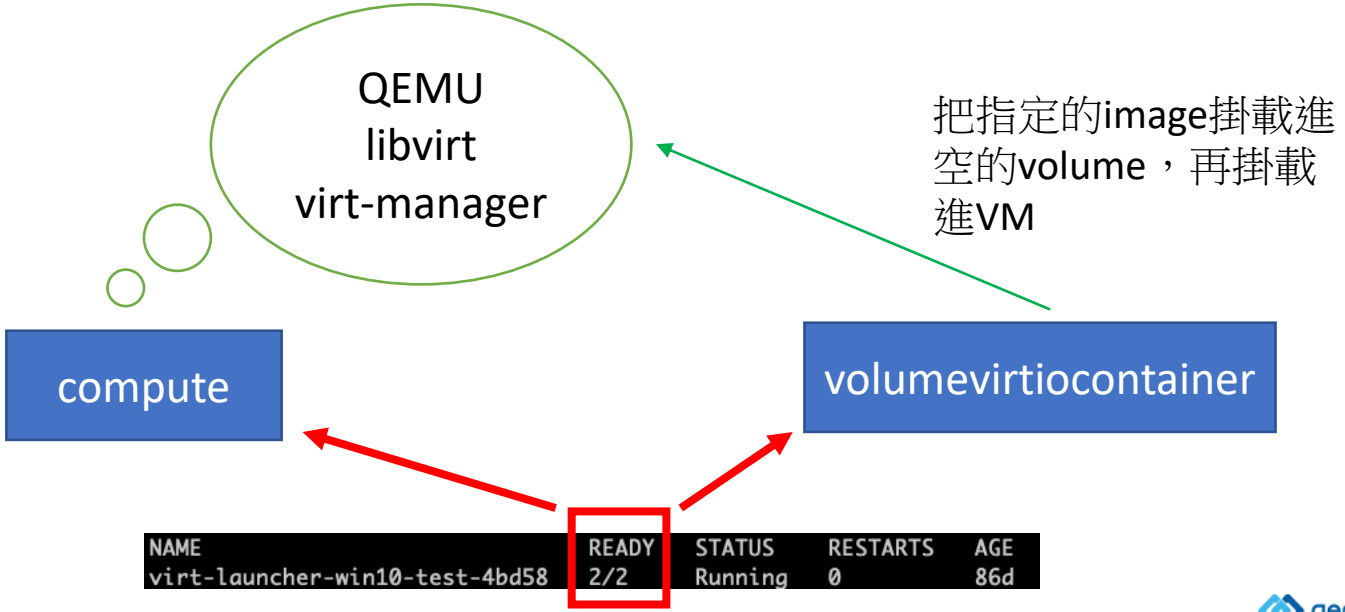
VM的建立

- 透過kubevirt.io的K8S API建立
  - 格式與一般K8S相同
  - 可建立一次性、永久或多個備援VM
  - 根據需求設定VM的參數
  - 可預先建立資源模板
- VM的映像檔與儲存空間來源
  - 以PVC為主
  - 可直接存取本機端的映像檔
  - 也可直接使用包成容器image的映像檔(containerDisk)

```
apiVersion: kubevirt.io/v1
kind: VirtualMachine
metadata:
  name: win10-test
spec:
  running: false
  template:
    metadata:
      labels:
        kubevirt.io/domain: win10-test
    spec:
      domain:
        cpu:
          cores: 4
        features:
```

```
volumes:
- name: ins-iso
  persistentVolumeClaim:
    claimName: iso-win10
- name: harddrive
  persistentVolumeClaim:
    claimName: win10-drive
- name: virtiocontainer
  containerDisk:
    image: kubevirt/virtio-container-disk
```

- KubeVirt在建立VM時會使用兩個容器：
  - compute：運行VM用容器
  - volumevirtiocontainer：將image掛載進環境內的容器



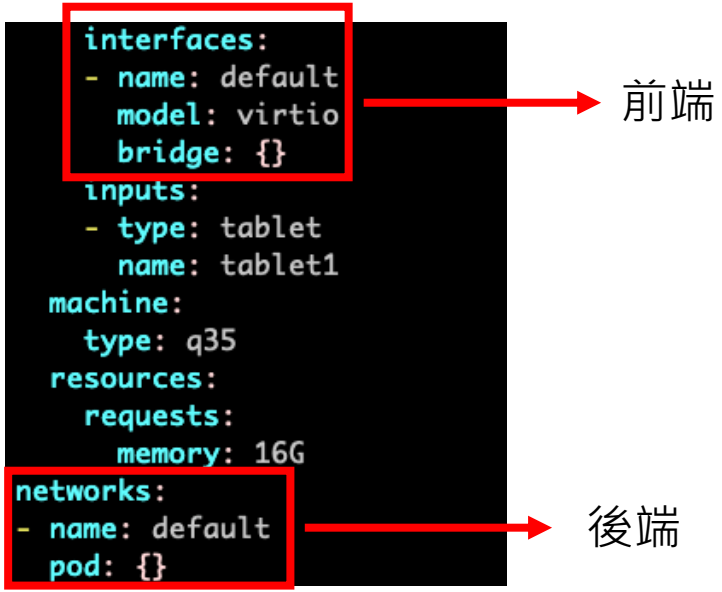
Virtctl指令工具

- 專門使用在KubeVirt上的指令工具
- 簡化部分需要修改YAML或打API的操作
- 可開關VM、掛載Volume、上傳映像檔、開連接埠等操作
- 需要加入label在YAML內，才能被辨識

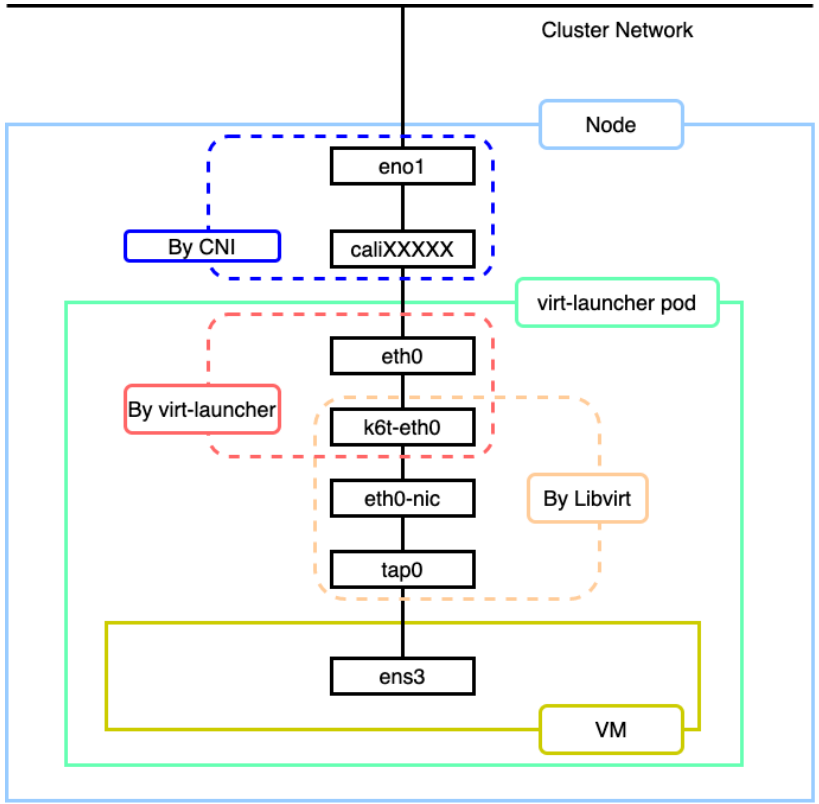
```
root@kubevirt-host:~/kubevirt_yaml# virtctl
Available Commands:
  addvolume     add a volume to a running VM
  console       Connect to a console of a virtual machine instance.
  expose        Expose a virtual machine instance, virtual machine,
 or virtual machine instance replica set as a new service.
  fslist        Return full list of filesystems available on the gu
est machine.
  guestosinfo   Return guest agent info about operating system.
  help          Help about any command
  image-upload Upload a VM image to a DataVolume/PersistentVolumeC
laim.
  migrate       Migrate a virtual machine.
  pause         Pause a virtual machine
  removevolume remove a volume from a running VM
  restart       Restart a virtual machine.
  start         Start a virtual machine.
  stop          Stop a virtual machine.
  unpause       Unpause a virtual machine
  userlist      Return full list of logged in users on the guest ma
chine.
  version       Print the client and server version information.
  vnc           Open a vnc connection to a virtual machine instance
.
```

VM的網路

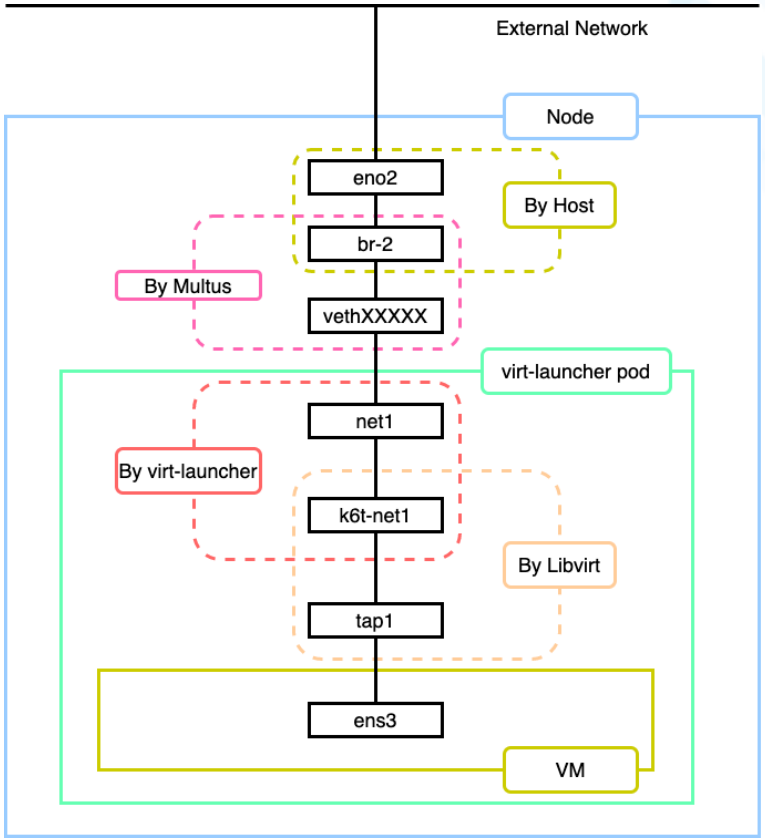
- 可設定VM要使用的前端與後端網路
  - 前端：VM網卡端的連接模式
  - 後端：VM Pod的網路連接
- 後端網路綁定
  - 直接使用K8S預設的CNI (Pod)
    - 可套用NetworkPolicy、Service等進行連線管理
    - 運作上與一般對Pod方式無異
  - 使用Multus CNI
    - 可不透過K8S預設CNI，實現傳統連接方式
    - 必須另外調整網路防護相關設定



Default CNI



Multus CNI



geminiopencloud

VM的Volume

- 永久性Volume
  - PersistentVolumeClaim (PVC)
  - DataVolume (DV)
  - hostDisk
- 暫時性Volume
  - emptyDisk
  - containerDisk
  - ephemeral
- 設定性質的Volume
  - configMap
  - Secret
  - cloudInitConfig

containerDisk

- 全名為Container Disk Image
- 本質為把VM映像檔存放進容器映像檔內
- 適用於大量部署且不需要保留資料的使用情境
- 一個容器映像檔可包含多個VM映像檔
- 不可動態調整映像檔大小

```
root@v172-ad:~/image-compare# tree -L 2 -h container-disk/
container-disk/
└── [4.0K]  disk
    └── [355M]  downloaded

1 directory, 1 file
root@v172-ad:~/image-compare# qemu-img info container-disk/disk/downloaded
image: container-disk/disk/downloaded
file format: raw
virtual size: 355M (371732480 bytes)
disk size: 355M
```

把containerDisk解開後的內部檔案架構

DataVolume (DV)

- 建立於PVC的上層，資料儲存也是基於PVC
- 提供KubeVirt關於Volume狀態及呼叫的版本化API
- 具自動化功能，可指定來源映像檔來建立DV
- 也可寫進VM的YAML內，開機時自動獲取建立

```
root@kubevirt-host:~/kubevirt_yaml# virtctl image-upload dv cirros \
> --size=50Mi --image-path=cirros-0.4.0-x86_64-disk.img \
> --uploadproxy-url=https://10.233.53.137 --insecure
PVC default/cirros not found
DataVolume default/cirros created
Waiting for PVC cirros upload pod to be ready...
Pod now ready
Uploading data to https://10.233.53.137

 12.13 MiB / 12.13 MiB [==============================] 100.00% 0s

Uploading data completed successfully, waiting for processing to com
plete, you can hit ctrl-c without interrupting the progress
Processing completed successfully
Uploading cirros-0.4.0-x86_64-disk.img completed successfully
```

```
root@kubevirt-host:~/kubevirt_yaml# kubectl get dv
NAME      PHASE       PROGRESS   RESTARTS   AGE
cirros    Succeeded   N/A                   22h
```

KubeVirt上的IaaS相關功能

- Cloud-Init & Sysprep
  - 為可在VM建立時進行環境客製化的程序
  - 適用於大量部署或使用cloud image
  - 可直接在YAML內帶入cloud-init或Sysprep資訊
  - 初始化內容可直接手動輸入，或者從configMap/Secret取得
- Containerized Data importer (CDI)
  - 將ISO或VM Disk檔案轉換成PVC的工具
  - 屬於PVC的上層，動態建立與管理PVC
  - 可從不同類型的來源匯入映像檔
  - 進行轉換並存成PVC或DataVolume

```
    - name: cloudinitdisk
      disk:
        bus: virtio
volumes:
  - name: containerdisk
    containerDisk:
      image: kubevirt/cirros-container-disk-demo:latest
  - name: cloudinitdisk
    cloudInitNoCloud:
      userData: |
        #cloud-config
        ssh_authorized_keys:
          - ssh-rsa AAAAB3NzaK8L93bWxnyp test@test.com
```

```
root@kubevirt-host:~/kubevirt_yaml# virtctl image-upload \
> dv cirros \
> --size=50Mi \
> --image-path=cirros-0.4.0-x86_64-disk.img \
> --uploadproxy-url=https://10.233.53.137 \
> --insecure
PVC default/cirros not found
DataVolume default/cirros created
Waiting for PVC cirros upload pod to be ready...
Pod now ready
Uploading data to https://10.233.53.137

 12.13 MiB / 12.13 MiB [=============================================================
==============================================================================
==============================================================================
==============] 100.00% 0s

Uploading data completed successfully, waiting for processing to complete,
you can hit ctrl-c without interrupting the progress
Processing completed successfully
Uploading cirros-0.4.0-x86_64-disk.img completed successfully
```

geminiopencloud

Disk Resize & Snapshot

- KubeVirt可直接針對PVC/DV以及本機映像檔直接進行Resize
  - PVC/DV：直接放大即可，KubeVirt會自動放大
  - 本機映像檔：使用qemu-img放大
  - 無法針對暫時性儲存類型進行放大
- 配合external-snapshotter的CRD可進行VM快照
  - 透過API進行對VM建立或還原快照
  - 只能對PVC/DV做快照
  - 並非所有的CSI都支援快照

```yaml
apiVersion: snapshot.kubevirt.io/v1alpha1
kind: VirtualMachineSnapshot
metadata:
  name: snap-larry
spec:
  source:
    apiGroup: kubevirt.io
    kind: VirtualMachine
    name: larry
```

Live Migration

- 可將VM搬移到其他K8S node運行
  - 在VM運行的情況下做搬移
  - 但限制很多
    - 不允許後端為Pod且前端使用bridge的VM作業
    - VM需開放Port 49152和49153
    - 若儲存類型為PVC/DV，存取模式需改為RWX
- 若要支援此功能，建議另建network
  - 可使用Multus CNI建立只對內的Migration通道

```
apiVersion: kubevirt.io/v1alpha3
kind: VirtualMachineInstanceMigration
metadata:
  name: migration-job
spec:
  vmiName: vmi-fedora
```

KubeVirt額外功能

- Host Device Passthrough
  - 可將主機端的PCI-E裝置assign進KubeVirt VM內
  - 可支援GPU、NVMe及其他PCI-E介面裝置
  - 須先做前置作業，包括啟用IOMMU、VFIO、設定KubeVirt等
  - 使用NVIDIA GPU可搭配KubeVirt GPU device plugin

```
configuration:
  permittedHostDevices:
    pciHostDevices:
    - pciVendorSelector: "10DE:1EB8"
      resourceName: "nvidia.com/TU104GL_Tesla_T4"
      externalResourceProvider: true
    - pciVendorSelector: "8086:6F54"
      resourceName: "intel.com/qat"
    mediatedDevices:
    - mdevNameSelector: "GRID T4-1Q"
      resourceName: "nvidia.com/GRID_T4-1Q"
```

```
kind: VirtualMachineInstance
spec:
  domain:
    devices:
      gpus:
      - deviceName: nvidia.com/TU104GL_Tesla_T4
        name: gpu1
      - deviceName: nvidia.com/GRID_T4-1Q
        name: gpu2
      hostDevices:
      - deviceName: intel.com/qat
        name: quickaccess1
```

小結

- KubeVirt在當前功能上對客戶的幫助
  - 讓VM基底服務導入進K8S內，保留需要VM的需求
  - 使用K8S網路架構與大多數功能，幫助容器轉型
  - 自動化佈署、模板、快照等功能可滿足部分IaaS需求
  - 與容器之間的混合服務有一定的潛力
  - 硬體層級的資源隔離，降低干擾與增加安全性
- KubeVirt當前的缺點
  - 對VM參數的調整還不夠彈性
  - 部分功能還有不少使用上的限制
  - 需要額外套件才能支援VM failover

KubeVirt文章Medium連結

在 K8S 上也能跑 VM！KubeVirt 簡介與建立

KubeVirt-K8S上的VM服務專案-VM的儲存空間

geminiopencloud

雙子星雲端運算

以簡馭繁・直上雲端

Thank you

www.geminiopencloud.com

contact@geminiopencloud.com

03-6590698