Full Scalable Media Cloud Solution with Kubernetes Orchestration

Zhenyu Wang, Xin (Owen) Zhang
Agenda

- Media in the Network and Cloud
- Intel Media Server Reference Software Stack
- Container with MSS enablement
- Kubernetes with Container integration
- Kubernetes with Container enabling on VCA2
- Kubernetes device plugin/Intel GPU plugin
- Use Case(1080p): VCA transcoding & k8s scheduling on VCA nodes
Media in the Network and Cloud

**Visual Understanding**
Object Recognition & Tracking
Indexing / Search
Smart Cities, Security and Surveillance

**Video Delivery**
Cloud and Comms:
Ingest / Storage / Edge
Transcode / Trans-size / Trans-rate
Video Conferencing

**Graphics in the Cloud**
Remote Desktop
Remote Workstation
Cloud Gaming
Rendering
Container with MSS enablement

- More Containers can be run than VMs
- Almost same performance with Native
- Package application and dependencies integrated
- Share same kernel as the host
- No need providing hardware based on the isolation

<table>
<thead>
<tr>
<th>MSS16.5 &amp; APP</th>
<th>MSS16.5 &amp; APP</th>
<th>MSS16.5 &amp; APP</th>
<th>...</th>
<th>MSS16.5 &amp; APP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docker1</td>
<td>Docker2</td>
<td>Docker3</td>
<td></td>
<td>Docker20</td>
</tr>
</tbody>
</table>

Native Kernel & I915 backports driver

Intel E3 Server

I915 device node

720P transcode on 20 dockers
Kubernetes with Container integration

- Orchestrates intelligent scheduling, self-healing, and horizontal scaling for rapid delivery of apps.
- Easy to customize the worker nodes.
- CPU/memory resources control.

![Kubernetes components diagram]

**Master**
- Kube-Scheduler
- etcd
- Controller Manager
- Proxy
- Kubectl
- K8s Dashboard

**Node**
- Kubelet
- Docker
- Proxy
- Streams Pool

**Container Registry**

**Media transcoding services**
Kubernetes with Container enabling on VCA2

• Intel® Visual Compute Accelerator 2 (Intel® VCA 2) (VCA1585LMV)
  – Intel® Xeon® Scalable processor and Intel® Xeon® processor E5-based platforms with Iris® Pro Graphics and Intel® Quick Sync Video media transcode capabilities

• Applications include:
  – Broadcast – Ultra-high channel density with high visual quality.
  – Remotely rendered graphics – High video quality, low latency graphics for enterprise productivity and anytime anywhere gaming.
  – Multi-party communication – Video-enabled B2B, B2C, and C2C communication with massive scaling
Kubernetes with Container enabling on VCA2
Kubernetes with Container enabling on VCA2

- 2-4 VCA2 cards
- 12 k8s nodes
Kubernetes device plugin

- k8s version 1.8 above
- DevicePlugins feature gated
- Default enabled after k8s 1.10
Intel GPU device plugin

- **Enumerate all DRM devices for Intel GPU**
  - `/dev/dri/card[0-9]*`
  - `/dev/dri/renderD[0-9]*`

- **Create server and sock**
  - `/var/lib/kubelet/device-plugins/intelGPU.sock`

- **Register resource to kubelet**
  - Resource name: “intel.com/gpu”

- **Serve “ListAndWatch” request**
  - Send device id and healthy

- **Serve “Allocate” request**
  - Return device id’s HostPath/ContainerPath to assign
Use of Intel GPU plugin

- A yaml file for pod
- [https://github.com/intel/intel-device-plugins-for-kubernetes](https://github.com/intel/intel-device-plugins-for-kubernetes)

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: intelgpu-demo-pod
spec:
  containers:
    -
      name: intelgpu-demo-container-1
      image: ubuntu-demo:latest
      imagePullPolicy: IfNotPresent
      command: [ "..." ]
      resources:
        limits:
          intel.com/gpu: 1
```
Intel GPU plugin in Kubernetes on VCA2
Use Case (1080p): Media transcoding on VCA2

- the performance comparison of number of real time streams between per VCA card and per Iris P580

**MEDIA TRANSCODING**

<table>
<thead>
<tr>
<th>Number of real time streams per VCA card @30fps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1080p</td>
</tr>
<tr>
<td>H.264 -&gt; H.264</td>
</tr>
<tr>
<td>H.264 -&gt; H.265</td>
</tr>
<tr>
<td>H.265 -&gt; H.265</td>
</tr>
</tbody>
</table>
Use Case (1080p): k8s scheduling on VCA2