25. hroug
godišnja konferencija

YugabyteDB:

PostgreSQL-compatible database On OKE

Franck Pachot, Developer Advocate

yugabyteDB

Franck Pachot

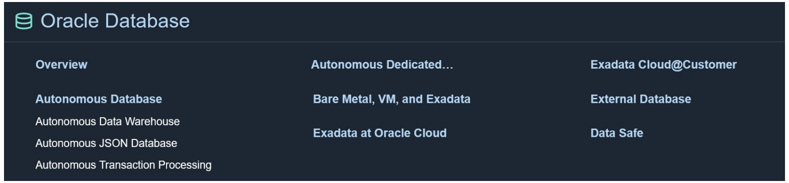
Developer Advocate on YugabyteDB
(PostgreSQL-compatible distributed database)

Past:

20 years in databases, dev and ops
Oracle ACE Director, AWS Data Hero
Oracle Certified Master, AWS Database Specialty
...

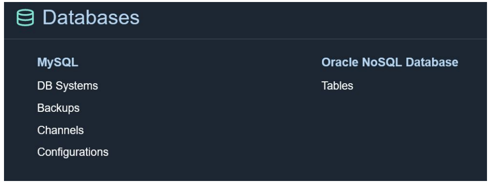
fpachot@yugabyte.com
dev.to/FranckPachot
@FranckPachot

Databases in Oracle Cloud



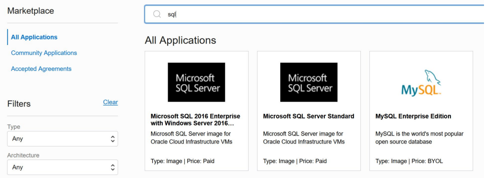
Oracle Database

- **Overview**
- **Autonomous Database**
  - Autonomous Data Warehouse
  - Autonomous JSON Database
  - Autonomous Transaction Processing
- **Autonomous Dedicated…**
- **Bare Metal, VM, and Exadata**
- **Exadata at Oracle Cloud**
- **Exadata Cloud@Customer**
- **External Database**
- **Data Safe**



Databases

- **MySQL**
  - DB Systems
  - Backups
  - Channels
  - Configurations
- **Oracle NoSQL Database**
  - Tables



Marketplace

sql

- All Applications
- Community Applications
- Accepted Agreements

Filters Clear

Type: Any

Architecture: Any

All Applications

Microsoft SQL 2016 Enterprise with Windows Server 2016…
Microsoft SQL Server image for Oracle Cloud Infrastructure VMs
Type: Image | Price: Paid

Microsoft SQL Server Standard
Microsoft SQL Server image for Oracle Cloud Infrastructure VMs
Type: Image | Price: Paid

MySQL Enterprise Edition
MySQL is the world's most popular open source database
Type: Image | Price: BYOL

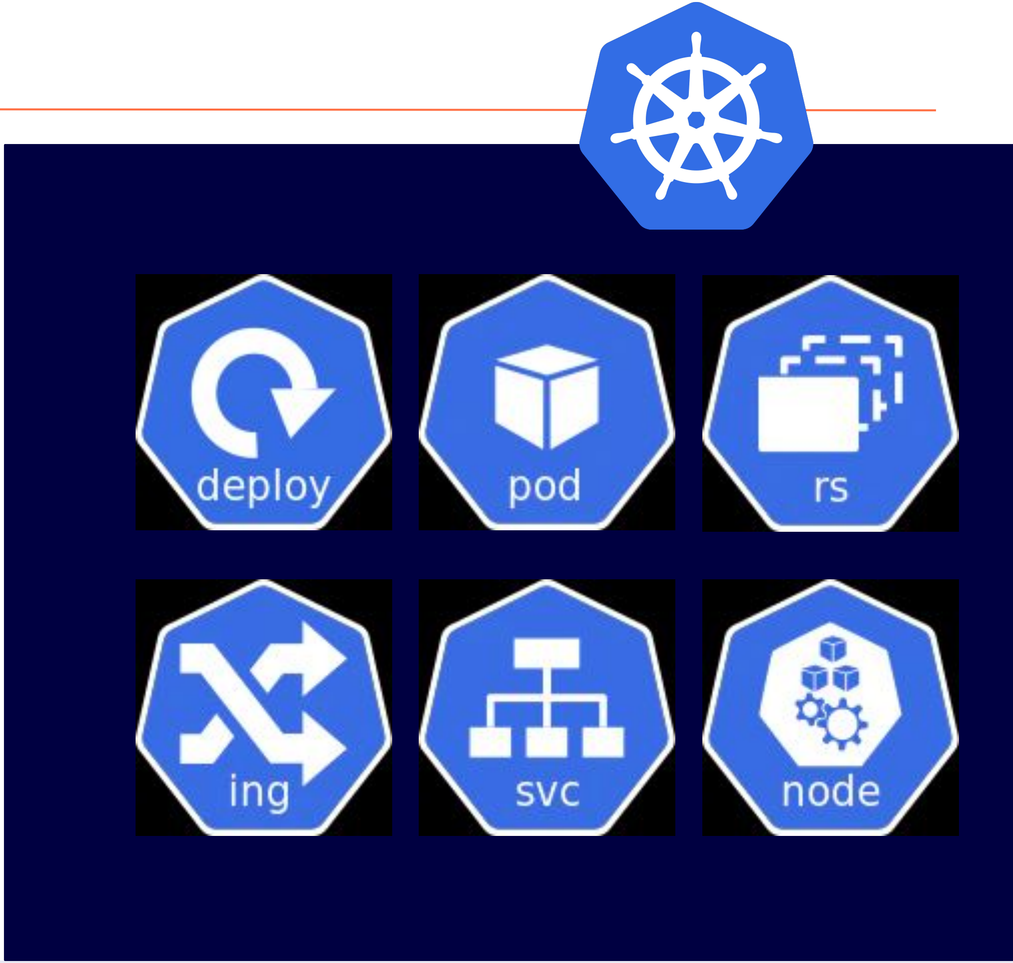
What about applications build for PostgreSQL?

What about HA with shared nothing?

What about cloud-native scale-out SQL databases?

Why Kubernetes?

- High Availability
- Scalability
- Disaster Recovery
- + declarative



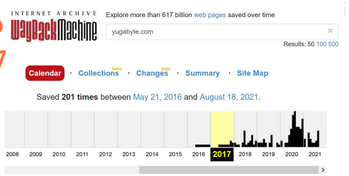
And Databases?

K8s StatefulSets

Alpha Jul. 2016

Beta Dec 2016

Stable Dec 2017



INTERNET ARCHIVE WayBackMachine

Explore more than 617 billion web pages saved over time

yugabyte.com

Results: 50 100 500

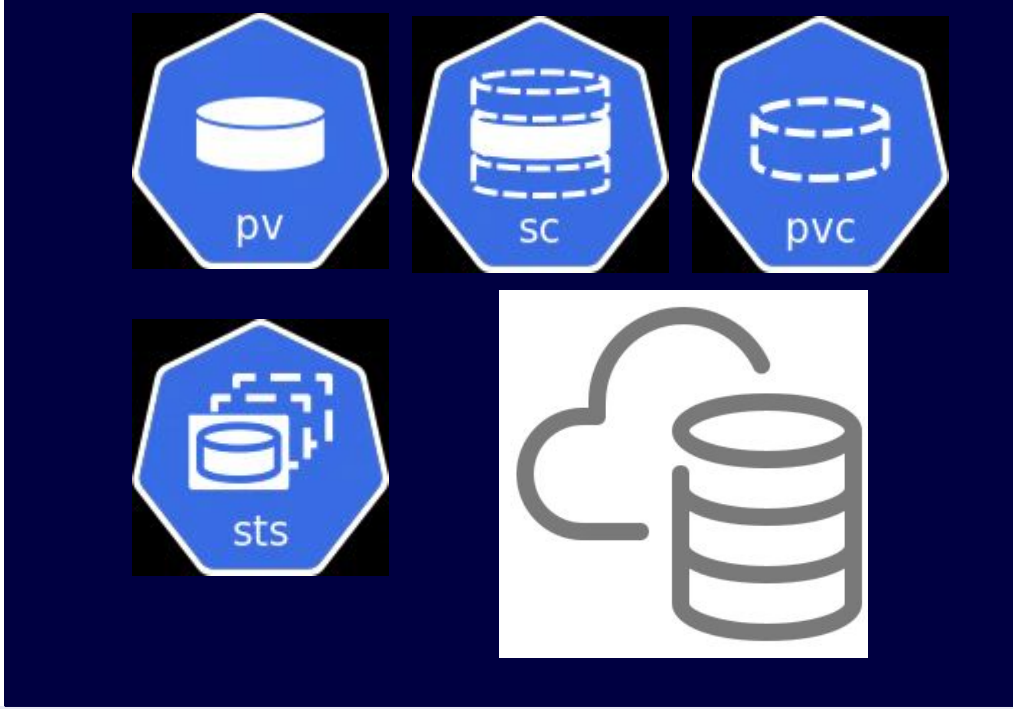
Calendar · Collections beta · Changes beta · Summary · Site Map

Saved 201 times between May 21, 2016 and August 18, 2021.

Persistent

Shared

Replicated



Distributed SQL Database

yugabyteDB

A cluster is called 'universe'

Private or public (internet) network
On-premises, cloud, hybrid, open-source

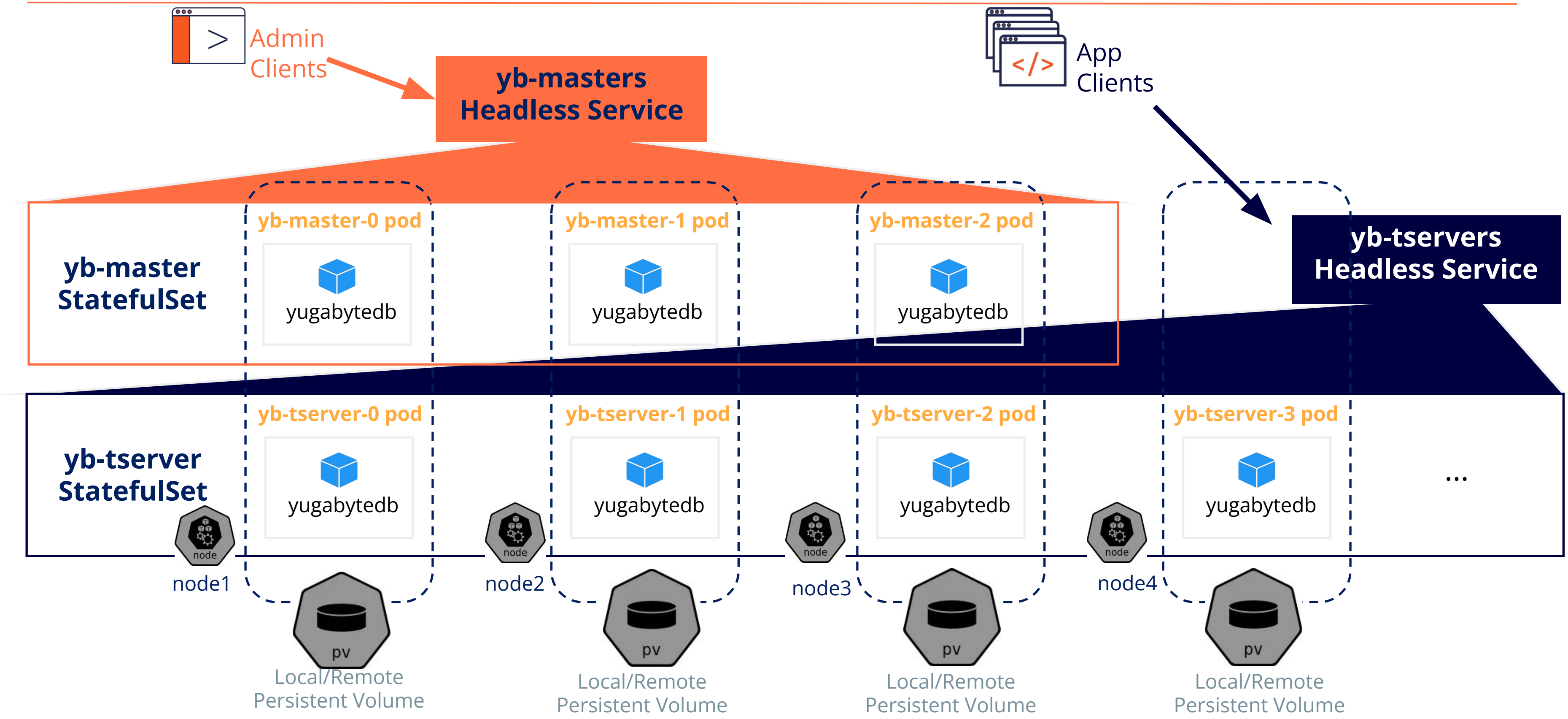
PostgreSQL API

PG compatible with read/write on all nodes

No shared storage, only network

Write on all nodes (Raft protocol)

YugabyteDB on Kubernetes



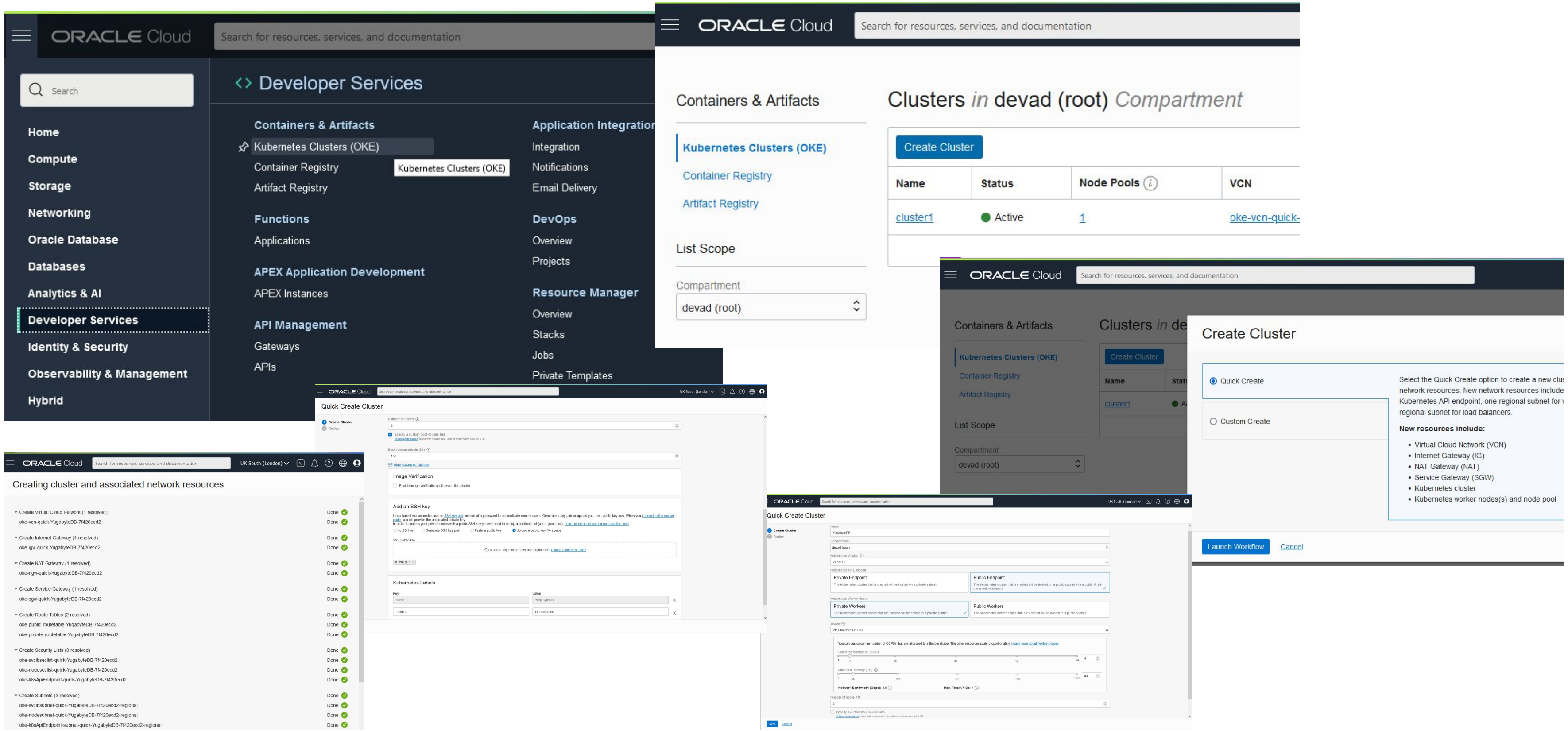
Demo

Oracle Cloud

OKE

YugabyteDB

Kubernetes on Oracle Cloud



ORACLE Cloud — Search for resources, services, and documentation

Home
Compute
Storage
Networking
Oracle Database
Databases
Analytics & AI
Developer Services
Identity & Security
Observability & Management
Hybrid

Developer Services

Containers & Artifacts
- Kubernetes Clusters (OKE)
- Container Registry
- Artifact Registry

Functions
- Applications

APEX Application Development
- APEX Instances

API Management
- Gateways
- APIs

Application Integration
- Integration
- Notifications
- Email Delivery

DevOps
- Overview
- Projects

Resource Manager
- Overview
- Stacks
- Jobs
- Private Templates

Containers & Artifacts
- Kubernetes Clusters (OKE)
- Container Registry
- Artifact Registry

List Scope
Compartment: devad (root)

Clusters *in* devad (root) *Compartment*

Create Cluster

Name	Status	Node Pools	VCN
cluster1	● Active	1	oke-vcn-quick-

Create Cluster

● Quick Create

Select the Quick Create option to create a new clus network resources. New network resources include Kubernetes API endpoint, one regional subnet for w regional subnet for load balancers.

New resources include:
- Virtual Cloud Network (VCN)
- Internet Gateway (IG)
- NAT Gateway (NAT)
- Service Gateway (SGW)
- Kubernetes cluster
- Kubernetes worker node(s) and node pool

○ Custom Create

Launch Workflow   Cancel

Quick Create Cluster

1 Create Cluster
2 Review

Name: YugabyteDB
Compartment: devad (root)
Kubernetes Version: v1.19.12

Kubernetes API Endpoint
- Private Endpoint — The Kubernetes cluster that is created will be hosted on a private subnet
- Public Endpoint ✓ — The Kubernetes cluster that is created will be hosted on a public subnet with a public IP address auto-assigned

Kubernetes Worker Nodes
- Private Workers ✓ — The Kubernetes worker nodes that are created will be hosted in a private subnet
- Public Workers — The Kubernetes worker nodes that are created will be hosted in a public subnet

Shape: VM.Standard.E3.Flex

You can customize the number of OCPUs that are allocated to a flexible shape. The other resources scale proportionately. Learn more about flexible shapes.

Select the number of OCPUs: 4
Amount of Memory (GB): 64
Network Bandwidth (Gbps): 4.0   **Max. Total VNICs:** 4

Number of nodes: 3
□ Specify a custom boot volume size
Volume performance varies with volume size. Default boot volume size: 46.6 GB

Next   Cancel

Quick Create Cluster

Number of nodes: 3
☑ Specify a custom boot volume size
Volume performance varies with volume size. Default boot volume size: 46.6 GB
Boot volume size (in GB): 150
Hide Advanced Options

Image Verification
□ Enable image verification policies on this cluster

Add an SSH key
Linux-based worker nodes use an SSH key pair instead of a password to authenticate remote users. Generate a key pair or upload your own public key now. When you connect to the worker node, you will provide the associated private key.
In order to access your private nodes with a public SSH key you will need to set up a bastion host (a.k.a. jump box). Learn more about setting up a bastion host

○ No SSH key ○ Generate SSH key pair ○ Paste a public key ● Upload a public key file (.pub)

SSH public key: A public key has already been uploaded. Upload a different one?
id_rsa.pub

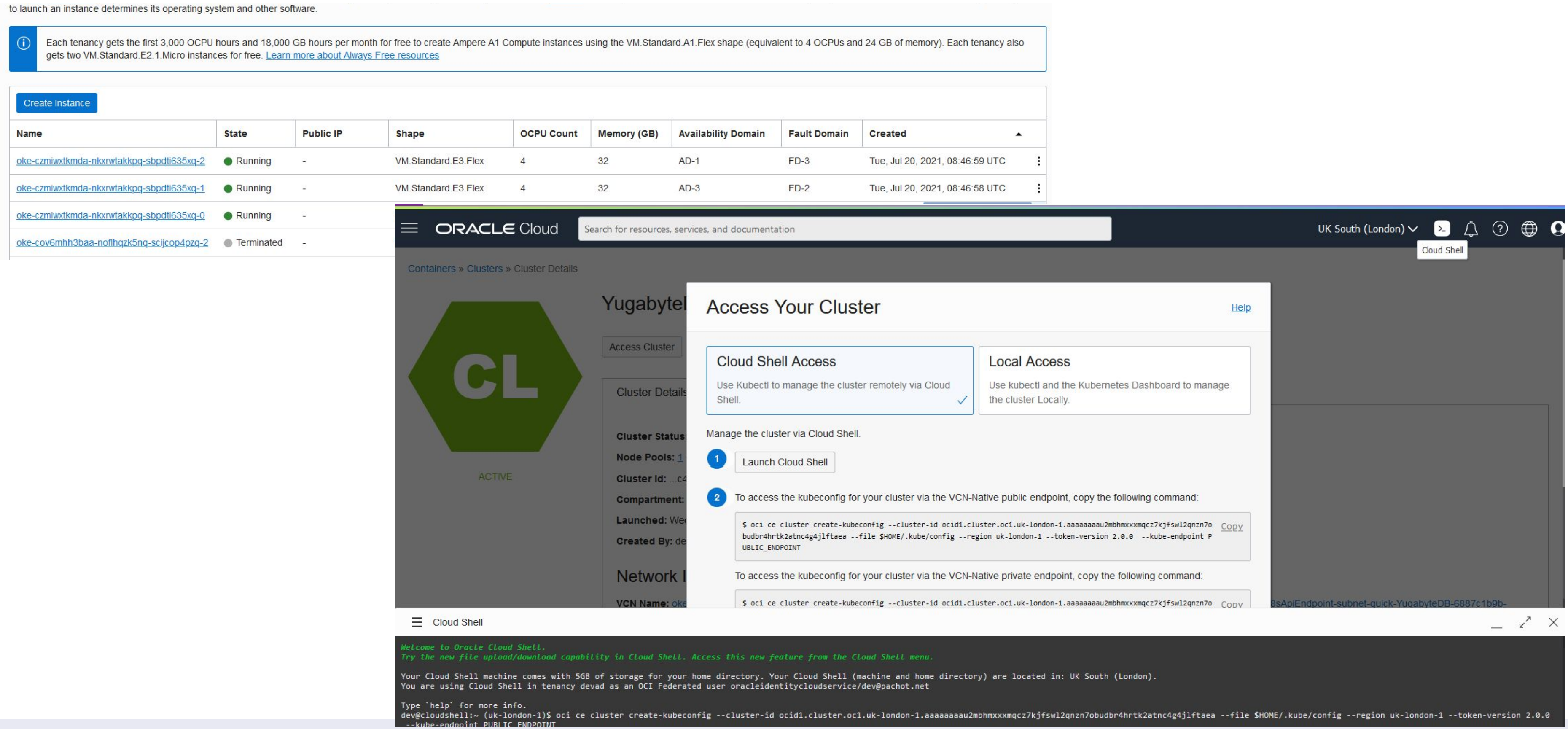
Kubernetes Labels

Key	Value
name	YugabyteDB
License	OpenSource

Creating cluster and associated network resources

```
Create Virtual Cloud Network (1 resolved)                      Done ✓
  oke-vcn-quick-YugabyteDB-7f420ecd2                           Done ✓
Create Internet Gateway (1 resolved)                           Done ✓
  oke-igw-quick-YugabyteDB-7f420ecd2                           Done ✓
Create NAT Gateway (1 resolved)                                Done ✓
  oke-ngw-quick-YugabyteDB-7f420ecd2                           Done ✓
Create Service Gateway (1 resolved)                            Done ✓
  oke-sgw-quick-YugabyteDB-7f420ecd2                           Done ✓
Create Route Tables (2 resolved)                               Done ✓
  oke-public-routetable-YugabyteDB-7f420ecd2                   Done ✓
  oke-private-routetable-YugabyteDB-7f420ecd2                  Done ✓
Create Security Lists (3 resolved)                             Done ✓
  oke-svclbseclist-quick-YugabyteDB-7f420ecd2                  Done ✓
  oke-nodeseclist-quick-YugabyteDB-7f420ecd2                   Done ✓
  oke-k8sApiEndpoint-quick-YugabyteDB-7f420ecd2                Done ✓
Create Subnets (3 resolved)                                    Done ✓
  oke-svclbsubnet-quick-YugabyteDB-7f420ecd2-regional          Done ✓
  oke-nodesubnet-quick-YugabyteDB-7f420ecd2-regional           Done ✓
  oke-k8sApiEndpoint-subnet-quick-YugabyteDB-7f420ecd2-regional  Done ✓
```

Kubernetes on Oracle Cloud



to launch an instance determines its operating system and other software.

Each tenancy gets the first 3,000 OCPU hours and 18,000 GB hours per month for free to create Ampere A1 Compute instances using the VM.Standard.A1.Flex shape (equivalent to 4 OCPUs and 24 GB of memory). Each tenancy also gets two VM.Standard.E2.1.Micro instances for free. Learn more about Always Free resources

Name	State	Public IP	Shape	OCPU Count	Memory (GB)	Availability Domain	Fault Domain	Created
oke-czmiwxtkmda-nkxrwtakkpq-sbpdti635xq-2	● Running	-	VM.Standard.E3.Flex	4	32	AD-1	FD-3	Tue, Jul 20, 2021, 08:46:59 UTC
oke-czmiwxtkmda-nkxrwtakkpq-sbpdti635xq-1	● Running	-	VM.Standard.E3.Flex	4	32	AD-3	FD-2	Tue, Jul 20, 2021, 08:46:58 UTC
oke-czmiwxtkmda-nkxrwtakkpq-sbpdti635xq-0	● Running	-						
oke-cov6mhh3baa-noflhqzk5nq-scijcop4pzq-2	● Terminated	-						

Access Your Cluster

Cloud Shell Access ✓
Use Kubectl to manage the cluster remotely via Cloud Shell.

Local Access
Use kubectl and the Kubernetes Dashboard to manage the cluster Locally.

Manage the cluster via Cloud Shell.

1 Launch Cloud Shell

2 To access the kubeconfig for your cluster via the VCN-Native public endpoint, copy the following command:

```
$ oci ce cluster create-kubeconfig --cluster-id ocid1.cluster.oc1.uk-london-1.aaaaaaaau2mbhmxxxmqcz7kjfswl2qnzn7o
budbr4hrtk2atnc4g4jlftaea --file $HOME/.kube/config --region uk-london-1 --token-version 2.0.0  --kube-endpoint P
UBLIC_ENDPOINT
```

To access the kubeconfig for your cluster via the VCN-Native private endpoint, copy the following command:

```
$ oci ce cluster create-kubeconfig --cluster-id ocid1.cluster.oc1.uk-london-1.aaaaaaaau2mbhmxxxmqcz7kjfswl2qnzn7o
```

```
Welcome to Oracle Cloud Shell.
Try the new file upload/download capability in Cloud Shell. Access this new feature from the Cloud Shell menu.

Your Cloud Shell machine comes with 5GB of storage for your home directory. Your Cloud Shell (machine and home directory) are located in: UK South (London).
You are using Cloud Shell in tenancy devad as an OCI Federated user oracleidentitycloudservice/dev@pachot.net

Type `help` for more info.
dev@cloudshell:~ (uk-london-1)$ oci ce cluster create-kubeconfig --cluster-id ocid1.cluster.oc1.uk-london-1.aaaaaaaau2mbhmxxxmqcz7kjfswl2qnzn7obudbr4hrtk2atnc4g4jlftaea --file $HOME/.kube/config --region uk-london-1 --token-version 2.0.0
 --kube-endpoint PUBLIC_ENDPOINT
```

yugabyteDB

Helm charts

```
helm repo add yugabytedb \
 https://charts.yugabyte.com
helm repo update

kubectl create namespace yb-demo

helm install yb-demo  \
 yugabytedb/yugabyte  \
 --namespace yb-demo
 --set replicas.{master,tserver}=3
```

allows 1 node failure

StatefulSets

```
kind: StatefulSet
metadata:
  name: yb-tserver
  namespace: yb-demo
spec:
  replicas: 3
  serviceName: yb-tservers
  podManagementPolicy: parallel
  updateStrategy:
   type: RollingUpdate
  containers:
   - name: yb-tserver
     image: 'yugabytedb/yugabyte:2.7.2.0-b216'
     command:
      - exec /home/yugabyte/bin/yb-tserver \
        --replication_factor=3 --enable_ysql=true
```

ordered or parallel
for faster scale-up

no downtime upgrade

Services

```
kind: Service
metadata:
  name: yb-tserver-service
  namespace: yb-demo
spec:
  ports:
  - name: tcp-ysql-port
    protocol: TCP
    port: 5433
    targetPort: 5433
    nodePort: 31874
  - name: tcp-yql-port
    port: 9042
  selector:
    app: yb-tserver
  clusterIP: 10.244.147.170
  type: LoadBalancer
```

Storage:

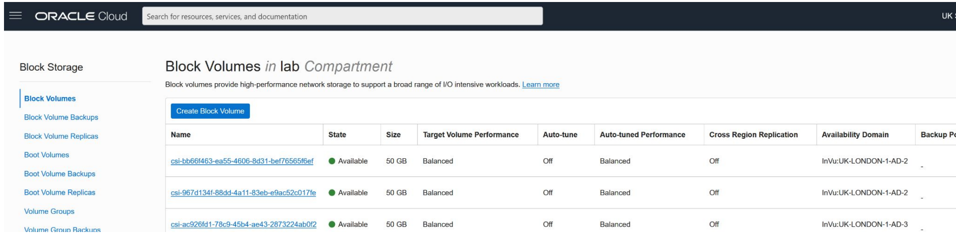
Ephemeral storage:

if a container dies, data must be read from the other nodes:

- quorum must still be there or data is lost (RPO>0)
- Full availability is up when all data has been transferred (RTO>0)

Shared storage

A shared remote storage (NFS)
is **not** necessary in a distributed DB



Block Volumes *in* lab *Compartment*

Block volumes provide high-performance network storage to support a broad range of I/O intensive workloads. Learn more

Name	State	Size	Target Volume Performance	Auto-tune	Auto-tuned Performance	Cross Region Replication	Availability Domain	Backup Po
csi-bb66f463-ea55-4606-8d31-bef76565f6ef	● Available	50 GB	Balanced	Off	Balanced	Off	InVu:UK-LONDON-1-AD-2	-
csi-967d134f-88dd-4a11-83eb-e9ac52c017fe	● Available	50 GB	Balanced	Off	Balanced	Off	InVu:UK-LONDON-1-AD-2	-
csi-ac926fd1-78c9-45b4-ae43-2873224ab0f2	● Available	50 GB	Balanced	Off	Balanced	Off	InVu:UK-LONDON-1-AD-3	-

Local or Persistent Volume

Can be pre-provisioned in the worker node or outside (cloud block storage)

Persistent Volumes:

- ✔ provisioned dynamically by K8s from block storage
- ✔ resilient to node failures (without reconstruction)

```
"volumeClaimTemplates": [
 {
  "kind": "PersistentVolumeClaim",
  "name": "datadir0"
  },
  "spec": {
    "accessModes": [ "ReadWriteOnce" ],
    "resources": { "requests": { "storage": "10Gi" }  },
    "storageClassName": "bv",
    "volumeMode": "Filesystem"
```

mounted once
= not shared

Anti-affinity:

✔ One pod per node (privileges durability over HA)

preferred or required

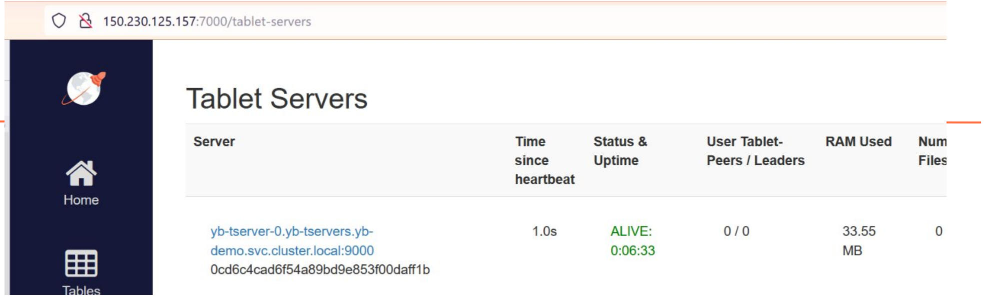
```
spec:
 affinity:
  preferredDuringSchedulingIgnoredDuringExecution:
  - weight: 100
   podAffinityTerm:
    labelSelector:
     matchExpressions:
     - key: app
      operator: In
      values:
      - yb-tserver
    topologyKey: kubernetes.io/hostname
```

Headless service

- ✔ new nodes discovered and added to DNS
- ✔ direct connection (placement aware smart clients)

```
apiVersion: v1
kind: Service
metadata:
   name: yb-tservers
   labels:
     app: yb-tserver
spec:
   clusterIP: None
```

External IP



Tablet Servers

Server	Time since heartbeat	Status & Uptime	User Tablet-Peers / Leaders	RAM Used	Num Files
yb-tserver-0.yb-tservers.yb-demo.svc.cluster.local:9000 0cd6c4cad6f54a89bd9e853f00daff1b	1.0s	ALIVE: 0:06:33	0 / 0	33.55 MB	0

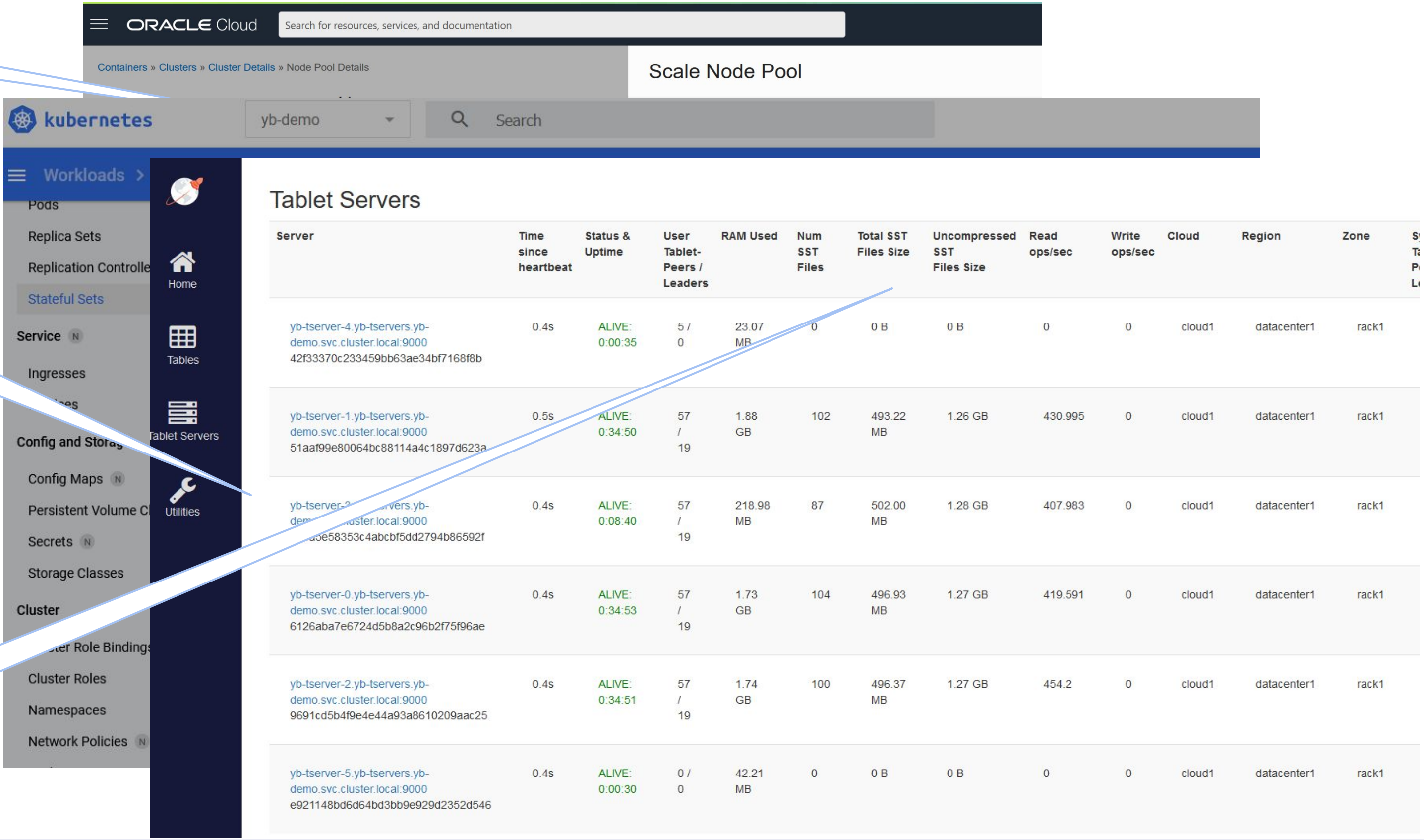
```
cloudshell$ kubectl get services -n yb-demo

NAME                  TYPE           CLUSTER-IP     EXTERNAL-IP
PORT(S)

yb-master-ui          LoadBalancer   10.96.81.113   150.230.125.157
7000:30510/TCP

yb-masters            ClusterIP      None           <none>
7000/TCP,7100/TCP

yb-tserver-service LoadBalancer 10.96.94.76    132.226.208.207
5433:31094/TCP,6379:31326/TCP,9042:32136/TCP

yb-tservers           ClusterIP      None           <none>
5433/TCP,9000/TCP,12000/TCP,11000/TCP,13000/TCP,9100/TCP,6379/TCP,9042/TCP
```

Demo: scale-out

more nodes

more replicas

autorebalance

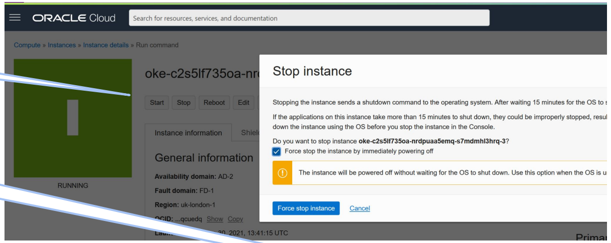


Tablet Servers

Server	Time since heartbeat	Status & Uptime	User Tablet-Peers / Leaders	RAM Used	Num SST Files	Total SST Files Size	Uncompressed SST Files Size	Read ops/sec	Write ops/sec	Cloud	Region	Zone
yb-tserver-4.yb-tservers.yb-demo.svc.cluster.local:9000 42f33370c233459bb63ae34bf7168f8b	0.4s	ALIVE: 0:00:35	5 / 0	23.07 MB	0	0 B	0 B	0	0	cloud1	datacenter1	rack1
yb-tserver-1.yb-tservers.yb-demo.svc.cluster.local:9000 51aaf99e80064bc88114a4c1897d623a	0.5s	ALIVE: 0:34:50	57 / 19	1.88 GB	102	493.22 MB	1.26 GB	430.995	0	cloud1	datacenter1	rack1
yb-tserver-[illegible]rvers.yb-dem[illegible]uster.local:9000 [illegible]a5e58353c4abcbf5dd2794b86592f	0.4s	ALIVE: 0:08:40	57 / 19	218.98 MB	87	502.00 MB	1.28 GB	407.983	0	cloud1	datacenter1	rack1
yb-tserver-0.yb-tservers.yb-demo.svc.cluster.local:9000 6126aba7e6724d5b8a2c96b2f75f96ae	0.4s	ALIVE: 0:34:53	57 / 19	1.73 GB	104	496.93 MB	1.27 GB	419.591	0	cloud1	datacenter1	rack1
yb-tserver-2.yb-tservers.yb-demo.svc.cluster.local:9000 9691cd5b4f9e4e44a93a8610209aac25	0.4s	ALIVE: 0:34:51	57 / 19	1.74 GB	100	496.37 MB	1.27 GB	454.2	0	cloud1	datacenter1	rack1
yb-tserver-5.yb-tservers.yb-demo.svc.cluster.local:9000 e921148bd6d64bd3bb9e929d2352d546	0.4s	ALIVE: 0:00:30	0 / 0	42.21 MB	0	0 B	0 B	0	0	cloud1	datacenter1	rack1

Demo: node failure

Kill node



General information

Availability domain: AD-2

Fault domain: FD-1

Region: uk-london-1

OCID: ...qcuedq Show Copy

Laun[illegible] 30, 2021, 13:41:15 UTC

Stop instance

Stopping the instance sends a shutdown command to the operating system. After waiting 15 minutes for the OS to s

If the applications on this instance take more than 15 minutes to shut down, they could be improperly stopped, resul down the instance using the OS before you stop the instance in the Console.

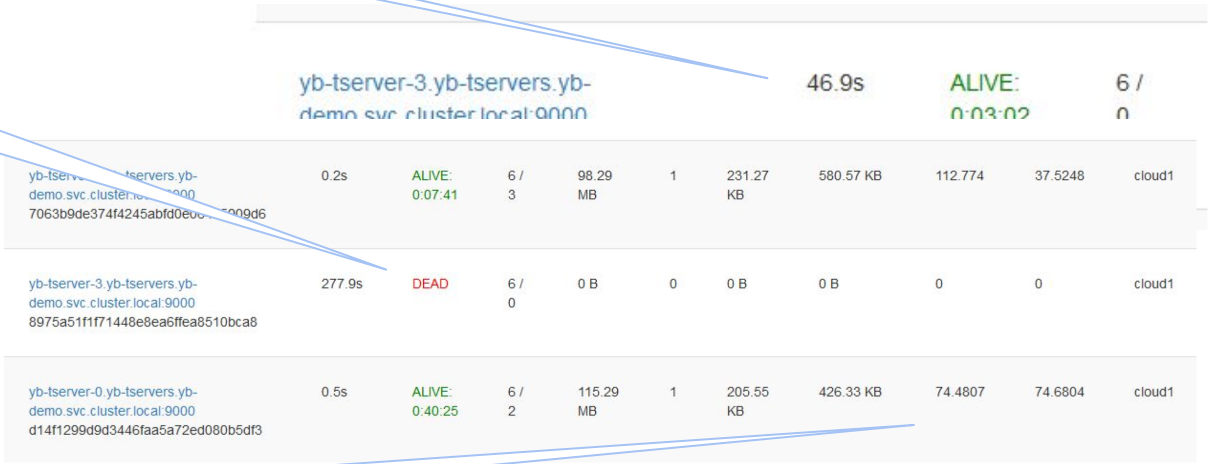
Do you want to stop instance **oke-c2s5lf735oa-nrdpuaa5emq-s7mdmhl3hrq-3**?

☑ Force stop the instance by immediately powering off

The instance will be powered off without waiting for the OS to shut down. Use this option when the OS is u

timeout

dead node detected



yb-tserver-3.yb-tservers.yb-demo.svc.cluster.local:9000 46.9s ALIVE: 0:03:02 6 / 0

yb-tserv[illegible]tservers.yb-demo.svc.cluster.l[illegible]00 7063b9de374f4245abfd0e0[illegible]5009d6	0.2s	ALIVE: 0:07:41	6 / 3	98.29 MB	1	231.27 KB	580.57 KB	112.774	37.5248	cloud1
yb-tserver-3.yb-tservers.yb-demo.svc.cluster.local:9000 8975a51f1f71448e8ea6ffea8510bca8	277.9s	DEAD	6 / 0	0 B	0	0 B	0 B	0	0	cloud1
yb-tserver-0.yb-tservers.yb-demo.svc.cluster.local:9000 d14f1299d9d3446faa5a72ed080b5df3	0.5s	ALIVE: 0:40:25	6 / 2	115.29 MB	1	205.55 KB	426.33 KB	74.4807	74.6804	cloud1

Other nodes still working (new leaders elected)

Automating Day 2 Operations

HANDLING FAILURES

K8s: Pod failure is automatic
ops: Node failure: manually add new workers

ops: Local storage failure: manually add new volume
YB: Automatic re-sharding

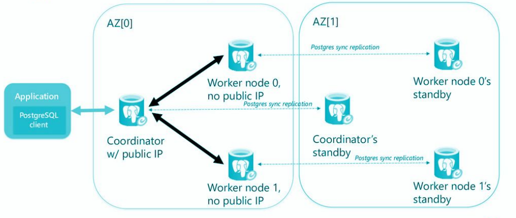
ROLLING UPGRADES

K8s: onDelete or rollingUpdate
(pod spawned with same network id / storage)
YB: can run with nodes in newer version

BACKUP & RESTORE

YB: distributed snapshots and backup
ops: restore to existing or new cluster

Distributed vs. Streaming Replication

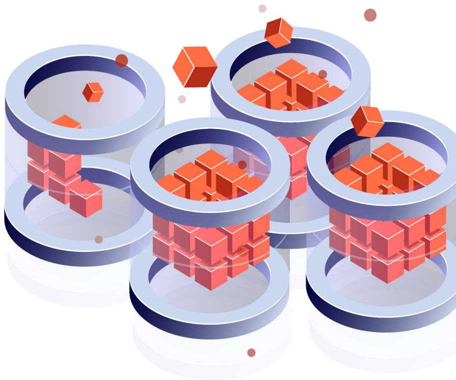

Application
PostgreSQL client
AZ[0]
Coordinator w/ public IP
Worker node 0, no public IP
Worker node 1, no public IP
AZ[1]
Postgres sync replication
Postgres sync replication
Postgres sync replication
Worker node 0's standby
Coordinator's standby
Worker node 1's standby

Streaming replication and sharding:

The primary is still a SPOF:

- Failover may take time (RTO)
- Failover may miss transactions (RPO)
- Manual or complex automation (rolling upgrade require many failovers)

PostgreSQL on K8s at Zalando: Two years in production https://av.tib.eu/media/52142



Distributed with replication factor

All nodes are equal

- Leaders are balanced over all nodes
- Followers are ready to be elected in few seconds (Raft protocol)

Thank You

Join us on Slack:
www.yugabyte.com/slack
Star us on GitHub:
github.com/yugabyte/yugabyte-db

fpachot@yugabyte.com
dev.to/FranckPachot
@FranckPachot

yugabyteDB

Core message:

- A PostgreSQL database active on multiple nodes
- Operations fully automated (cloud, K8s, PaaS)
- Distributed to provide: Resilience, High Availability, Geo Distribution, Elasticity