OpenStack In Production (CERN)

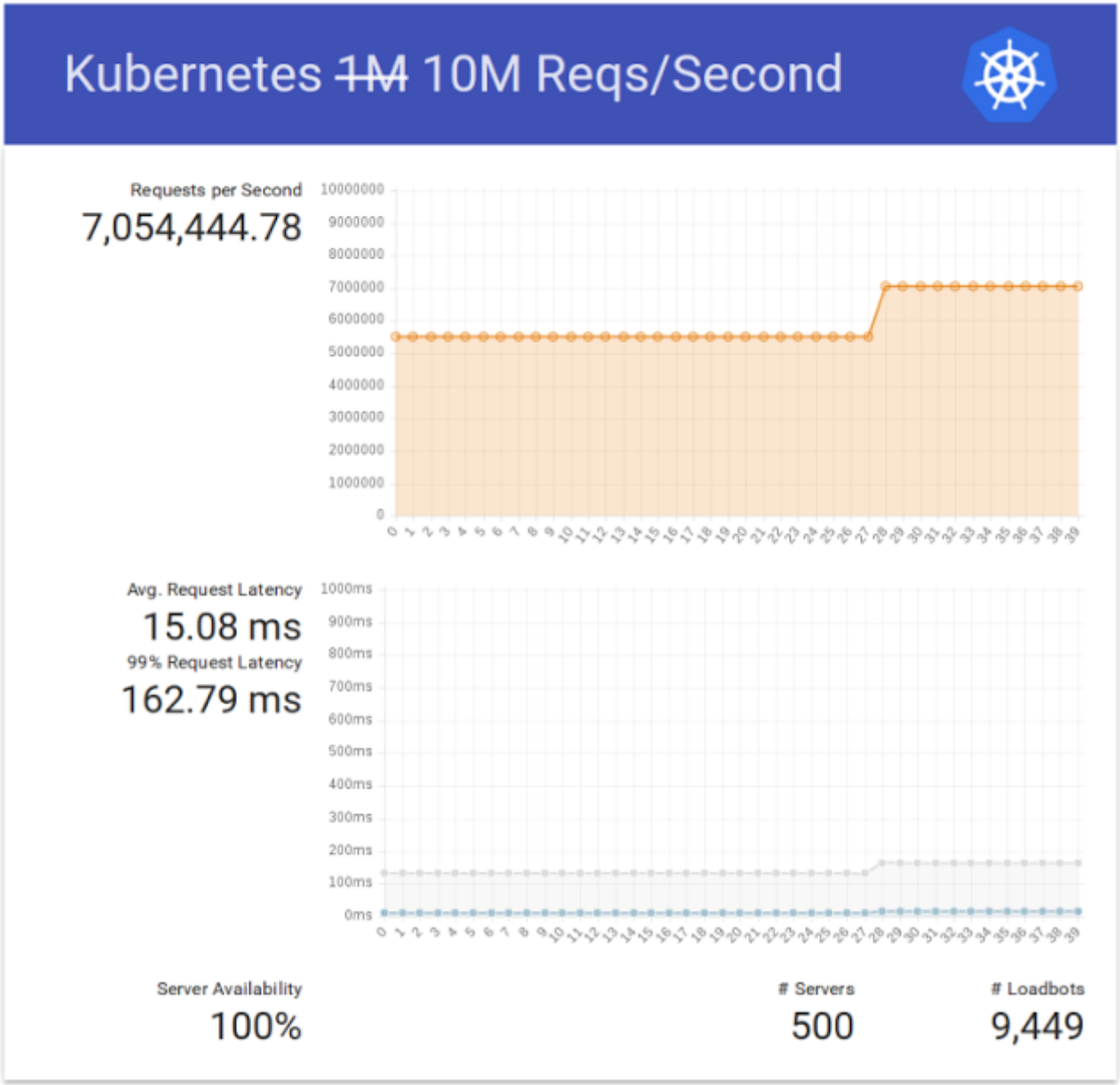
Containers on the CERN cloud

We have recently made the Container-Engine-as-a-Service (Magnum) available in production at CERN as part of the CERN IT department services for the LHC experiments and other CERN communities. This gives the OpenStack cloud users Kubernetes, Mesos and Docker Swarm on demand within the accounting, quota and project permissions structures already implemented for virtual machines.

We shared the latest news on the service with the CERN technical staff (link). This is the follow up on the tests presented at the OpenStack Barcelona (link) and covered in the blog from IBM. The work has been helped by collaborations with Rackspace in the framework of the CERN openlab and the European Union Horizon 2020 Indigo Datacloud project.

erformance

At the Barcelona summit, we presented with Rackspace and IBM regarding our additional performance tests after the previous blog post. We expanded beyond the 2M requests/s to reach around 7M where some network infrastructure issues unrelated to OpenStack limited the scaling further.



As we created the clusters, the deployment time increased only slightly with the number of nodes as most of the work is done in parallel. But for 128 node or larger clusters, the increase in time started to scale almost linearly. At the Barcelona summit, the Heat and Magnum teams worked together to develop proposals for how to improve further in future releases, although a 1000 node cluster in 23 minutes is still a good result

Cluster Size (Nodes)	Concurrency	Deployment Time (min)
2	50	2.5
16	10	4
32	10	4
128	5	5.5
512	1	14
1000	1	23

Storage

With the LHC producing nearly 50PB this year, High Energy Physics has some custom storage technologies for specific purposes, EOS for physics data, CVMFS for read-only, highly replicated storage such as applications.

One of the features of providing a private cloud service to the CERN users is to combine the functionality of open source community software such as OpenStack with the specific needs for high energy physics. For these to work, some careful driver work is needed to ensure appropriate access while ensuring user rights. In particular,

- EOS provides a disk-based storage system providing high-capacity and low-latency access for users at CERN. Typical use cases are where scientists are analysing data from the experiments.
- CVMFS is used for a scalable, reliable and low-maintenance for read-only data such as software.

There are also other storage solutions we use at CERN such as

- HDFS for long term archiving of data using Hadoop which uses an HDFS driver within the container. HDFS works in user space, so no particular integration was required to use it from inside (unprivileged) containers
- Cinder provides additional disk space using volumes if the basic flavor does not have sufficient. This Cinder integration is offered by upstream Magnum, and work was done in the last OpenStack cycle to improve security by adding support for Keystone trusts.

CVMFS was more straightforward as there is no need to authenticate the user. The data is read-only and can be exposed to any container. The access to the file system is provided using a driver (link) which has been adapted to run inside a container. This saves having to run additional software inside the VM hosting the container.

EOS requires authentication through mechanisms such as Kerberos to identify the user and thus determine what files they have access to. Here a container is run per user so that there is no risk of credential sharing. The details are in the driver (link).

Service Model

One interesting question that came up during the discussions of the container service was how to deliver the service to the end users. There are several scenarios:

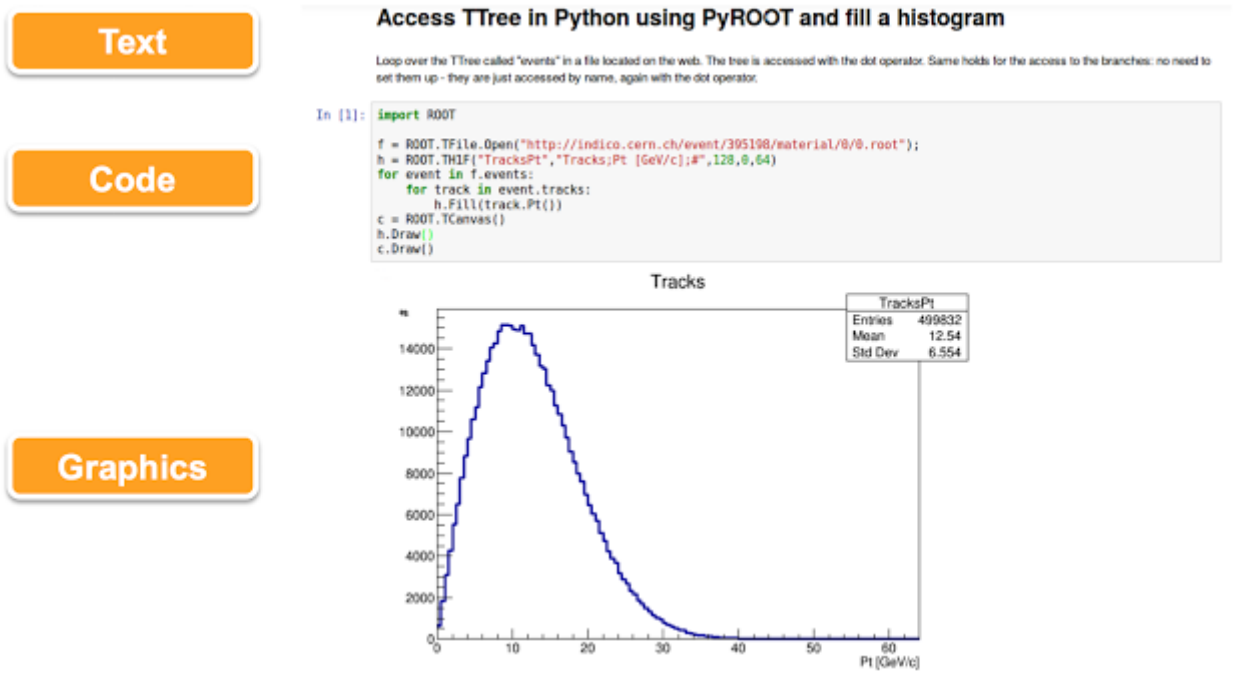
1. The end user launches a container engine with their specifications but they rely on the IT department to maintain the engine availability. This implies that the VMs running the container engine are not accessible to the end user.
2. The end user launches the engine within a project that they administer. While the IT department maintains the templates and basic functions such as the Fedora Atomic images, the end user is in control of the upgrades and availability.
3. A variation of option 2., where the nodes running containers are reachable and managed by the end user, but the container engine master nodes are managed by the IT department. This is similar to the current offer from the Google Container Engine and requires some coordination and policies regarding upgrades

Currently, the default Magnum model is for the 2nd option and adding option 3 is something we could do in the near future. As users become more interested in consuming containers, we may investigate the 1st option further

Applications

Many applications at use in CERN are in the process of being reworked for a microservices based architecture. A choice of different container engines is attractive for the software developer. One example of this is the file transfer service which ensures that the network to other high energy physics sites is kept busy but not overloaded with data transfers. The work to containerise this application was described at the recent CHEP 2016 FTS poster.

While deploying containers is an area of great interest for the software community, the key value comes from the physics applications exploiting containers to deliver a new way of working. The Swan project provides a tool for running ROOT, the High Energy Physics application framework, in a browser with easy access to the storage outlined above. A set of examples can be found at https://swan.web.cern.ch/notebook-galleries. With the academic paper, the programs used and the data available from the notebook, this allows easy sharing with other physicists during the review process using CERNBox, CERN's owncloud based file sharing solution.



Another application being studied is http://opendata.cern.ch/?ln=en which allows the general public to run analyses on LHC open data. Typical applications are Citizen Science and outreach for schools.

)ngoing Work

There are a few major items where we are working with the upstream community:

- Cluster upgrades will allow us to upgrade the container software. Examples of this would be a new version of Fedora Atomic, Docker or the container engine. With a load balancer, this can be performed without downtime (spec)
- Heterogeneous cluster support will allow nodes to have different flavors (cpu vs gpu, different i/o patterns, different AZs for improved failure scenarios). This is done by splitting the cluster nodes into node groups (blueprint)
- Cluster monitoring to deploy Prometheus and cAdvisor with Grafana dashboards for easy monitoring of a Magnum cluster (blueprint).

?eferences

- End user documentation for containers on the CERN cloud at http://clouddocs.web.cern.ch/clouddocs/containers/index.html
- CERN IT department information is at http://cern.ch/it.
- CERN openlab Rackspace collaboration on container presentations are listed here.
- Indigo Datacloud project details are here.

by Tim Bell (noreply@blogger.com) at January 09, 2017 02:54 PM

anuary 05, 2017

Daniel Berrange