OpenStack in Production

Hints and tips from the CERN OpenStack cloud team

Friday, 17 June 2016

Scaling Magnum and Kubernetes: 2 million requests per second

Two months ago, we described in this blog post how we deployed OpenStack Magnum in the CERN cloud. It is available as a pre-production service and we're steadily moving towards full production mode as a standard part of the CERN IT service offerings to give Containers-as-a-Service.

As part of this effort, we've started testing the upgrade procedures, the latest being to the final Mitaka release. If you're here to see some fancy load tests, keep reading below, but some interesting details on the upgrade:

- We build our own RPMs to include a few patches from post-Mitaka upstream (the most important being the trustee user to support lifecycle operations on the bays) and some CERN customizations (removal of neutron LBaaS and floating ips which we don't yet have, adding the CERN Certificate Authority, ...). Check here for the patches and build procedure
- We build our Fedora Atomic 23 image to get more recent versions of docker and kubernetes (1.10 and 1.2 respectively), plus support for an internal distributed filesystem called CVMFS. We do use the upstream disk-imagebuilder procedure with a few additional elements available here

While discussing how we could further test the service, we thought of this kubernetes blog post, achieving 1 million requests per second against a service running on a kubernetes cluster. We thought we could probably do the same. Requirements included:

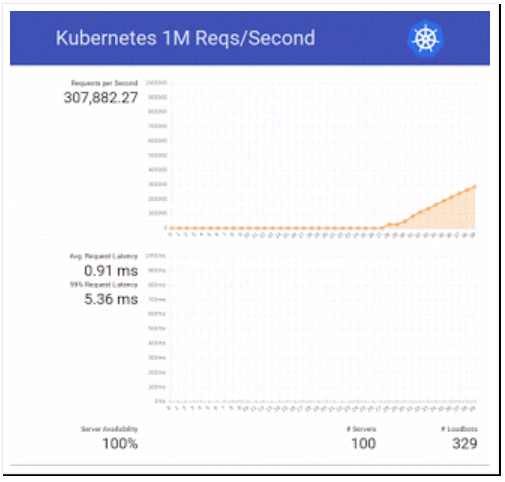
- kubernetes 1.2, which our recent upgrade offered
- available resources to deploy the cluster, and luckily we were installing a new batch of a few hundred physical nodes which could be used for a day or two

So along with the upgrade, Bertrand and Mathieu got to work to test this setup and we quickly got it up and running.

Quick summary of the setup:

- 1 kubernetes bay
- 1 master node, 16 cores (not really needed but why not)
- 200 minions, 4 cores each

In total there are 800 cores, which matches the cluster used in the original test. How did our test go?



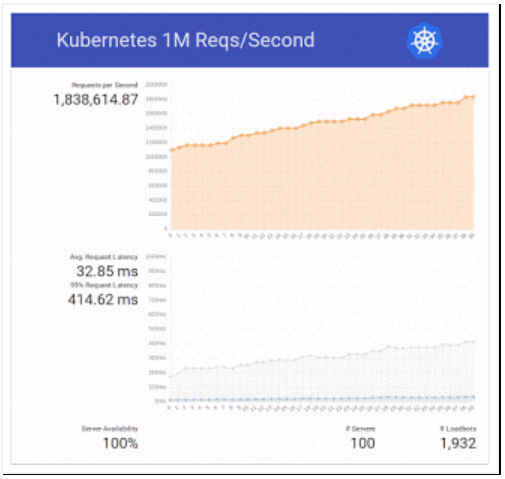
We ended up trying a bit more and doubled the number to 2 million requests per second :)

Blog Archive

▼ 2016 (4)
▼ June (1)
Scaling Magnum and Kubernetes: 2 million requests ...
► April (3)
► 2015 (17)
► 2014 (6)
► 2013 (6)

Contributors

- Arne Wiebalck
- Belmiro Moreira
- Daniel van der Ster
- Jan van Eldik
- Jose Castro Leon
- Marcos Fermín Lobo
- Ricardo Rocha
- Thomas Oulevey
- Tim Bell



We learned a few things on the way:

- set Heat's *max_resources_per_stack* to something big. Magnum stacks create a lot of these, and with bays of hundreds of nodes the value gets high enough that unlimited (-1) is tempting and we have it like that now. It leaves the option for people to deploy a stack with so many resources that Heat could break, so we'll investigate what the best value is
- while creating and deleting many large bays, Heat shows errors like 'TimeoutError: QueuePool limit of size ... overflow ... reached' which we've seen in the past for other OpenStack services. We'll contribute the patch to fix it upstream if not there yet
- latency values get high even before the 1 million barrier, we'll check further the demo code and our setup (using local disk, in this case SSDs instead of the default volume attachment in Magnum should help)
- Heat timeout and retrial configuration values need to be tuned to manage very large stacks. We're still not sure what are the best values, but will update the post once we have them
- Magnum shows 'Too many files opened' errors, we also have a fix to contribute for this one
- Nova, Cinder (bay nodes use a volume), Keystone and all other OpenStack services scaled beautifully, our cloud usually has a rate of ~150 VMs created and deleted per hour, here's the plot for the test period, we eventually tried bays up to 1000 nodes

And what's next?

- **Larger bays:** at the end of these tests we deployed a few bigger bays with 300, 500 and 1000 nodes. And in just a couple weeks there will be a new batch of physical nodes arriving, so we plan to upgrade Heat to Mitaka and build on the recent upstream work (by Spyros together with Ton and Winnie from IBM) adding Magnum scenarios to Rally to run additional scale tests and see where it breaks
- **Bay lifecycle:** we stopped at launching a large number of requests in a bay, next we would like to perform bay operations (update of number of nodes, node replacement) and see which issues (if any) we find in Magnum
- **New features:** lots of upstream work going on, so we'll do regular Magnum upgrades (cinder support, improved bay monitoring, support for some additional internal systems at CERN)

And there's also Swarm and Mesos, we plan on testing those soon as well. And kubernetes updated their test, so stay tuned...

Acknowledgements

- Bertrand Noel, Mathieu Velten and Spyros Trigazis from CERN IT, for the work upstream and integrating Magnum at CERN, and on getting these demos running
- Rackspace for their support within the CERN Openlab on running containers at scale
- Indigo Datacloud building a platform as a service for e-science in Europe

- Kubernetes for an awesome tool and the nice demo
- All in the CERN OpenStack Cloud team, for a great service (especially Davide Michelino and Belmiro Moreira for all the work integrating Neutron at CERN)
- The upstream Magnum team, for building what is now looking like a great service, and we look forward for what's coming next (bay drivers, bare metal support, and much more)
- Tim, Arne and Jan for letting us use the new hardware for a few days

Posted by Ricardo Rocha at 06:50

G+1 +2 Recommend this on Google

Labels: cern, kubernetes, magnum, openstack

1 comment

Add a comment

Top comments

Five Minutes of Cloud via Google+ 2 weeks ago - Shared publicly
http://openstack-in-production.stfi.re/2016/06/scaling-magnum-and-kubernetes-2-million.html

1 · Reply

Home Older Post

Subscribe to: Post Comments (Atom)