ETH zürich

Achieving High Service Availability for HPC

Urban Borštnik, Diego Moreno
Logging and Monitoring hpc-ch Forum
3 October 2019
WSL

IT Services

Outline

- Introduction
- Overview of our compute clusters
- Logging
- Metrics
- Cluster Monkey
- Future outlook

Introduction

- How do you achieve high service availability?
  - Fix it when it breaks (isolate it or repair it)
  - Know what will break and when → fix it before it breaks
- How can you know when something is broken?
  - Collect logs and metrics
  - Recognize that something is broken
- How does this scale?

Overview of our compute clusters

We operate compute clusters in two data centers

Euler in Lugano

- >2300 compute nodes with over 40.000 cores
- 3 separate Infiniband fabrics
- Lustre storage
- NetApp storage
- Integrated into the ETH network
- Remote!

Leonhard in Zürich

- >130 compute nodes with over 4.000 cores, 650 GPUs
- >5 tenants
- Lustre storage
- NetApp storage
- Integrated into the ETH network

High availability: our principles

Automation!

- Many “recognition” tasks can be automated:
  - Alarms
  - Is a service running? Is some value amiss? Did a known event happen?
- Many “fixing” tasks can be automated:
  - Frequent cases with well-defined actions→scriptable
  - Rebooting a node, restarting a service
  - 1st level triage: collecting diagnostics, opening a ticket

Logging services

- Central log servers for syslogs, per datacenter
- Standard Elastic stack:
  - rsyslogd→filebeat→logstash→elasticsearch→kibana

- Correlate events between servers and clients
- Some services are verbose (not enough vs. too much)
- Compliance issues

Metrics services: collecting data

- External functionality monitoring:
  - Does SSH into the cluster work?
- Consul, collectd
- Prometheus: gathers metrics from
  - Compute nodes (node_exporter)
  - Lustre filesystem (lustre_exporter + node_exporter)
  - System-wide (batch system)
  - k8s admin cluster

Metrics: what we collect

- Typical node metrics: CPU, memory usage, …
- Specific storage metrics: device saturation, queue size, …
- Lustre metrics (HPE's lustre_exporter)
  - BW, IOPS
  - Stats per job
- Batch system: dispatch/completion rates, job pressure, …
- Software metrics: running services, versions, …

Metrics: visualization

- Grafana
- Dashboards for
  - The cluster as a whole
  - Individual nodes
  - Batch system
  - Lustre file system
  - k8s admin cluster
- Four Lustre dashboards published in Grafana repos
  - Lustre Overview, Detailed, Jobs Stats, Specific Job Stats

Monitoring data storage: typical “one-sight” usage

- *Harmful* workloads: Servers’ CPU or storage devices load
- Network issues: # connected nodes to the storage system
- Planning: capacity, inode usage and future requirements
- Job workload improvement:
  - average size per operation
  - # metadata IOPS
- Overall system health

Alerting

- alertmanager
  - Webhooks to inspect & open tickets
  - Metrics-based
- Creates tickets that need to be addressed by operators, such as
  - # connected nodes to Lustre dropping
  - High filesystem capacity used
  - Batch system not running any jobs
  - Detection of harmful patterns in workloads

Cluster Monkey

- Cluster Monkey: does what we train it to do
- A technical account and a set of scripts run as dæmons and cronjobs

- It monitors nodes and system services and takes actions if needed

Cluster Monkey: monitoring and actions

- Hardware errors on a compute node?
  - Isolate it, get diagnostics data, open a ticket for a human operator
- Is the software on a node healthy (OS version, IB firmware, IB speed, requisite dæmons, mounts, local disk, ….)?
  - Isolate it for reinstallation or open a ticket
- Is the node up for too long?
  - Isolate it for reboot

Cluster Monkey: monitoring reboots

- Rebooting or reinstalling is a multi-stage processes
  - Done by Cluster Monkey
- Performance check of node before it's put in production
  - Quick check with STREAM, HPL, and PingPong over IB to catch obvious performance problems
  - Opens a ticket if degradation noticed
  - Spectre fixes: needed to adjust expectations
- All nodes are reinstalled about once a month
  - Driven by Cluster Monkey
  - Creates rack/enclosure reservation, reinstalls, performance check

Future outlook

- Further automation
- Early detection of damaging workloads
- Acting upon abuse of login nodes
- Dashboards and alerting for users