СЛЁРМ
+
Southbridge

Мониторинг Ceph

Виталий Филиппов

СЛЁРМ Ceph Видеокурс

Мониторинг Ceph

Виталий Филиппов

Разработчик-эксперт в компании CUSTIS, линуксоид. Занимаюсь разработкой на разных языках от node.js до C++, сильно упоролся по Ceph-у. :)

Автор статьи «Производительность Ceph»

Мониторинг Ceph кластера

1. Инструменты мониторинга
   → консоль
   → встроенный Dashboard
   → Grafana (официальные дашборды)
   → Grafana (народные дашборды)

2. Кроме Prometheus в Ceph есть поддержка Zabbix, InfluxDB и Telegraf

3. Однако Prometheus + Grafana — выбор по умолчанию, сами разработчики тоже используют её, так что остановимся на них

4. Показатели для мониторинга
   → метрики
   → алерты

Консоль

1. Всё самое важное собрано в статусе:
   ceph -s («status»)
   ceph -w («watch»), watch ceph -s

2. Детали по healthcheck-ам:
   ceph health
   ceph health detail

3. Состояние OSD и занятое место (на OSD и в целом):
   ceph osd df tree
   ceph df

4. Мониторинг iops-ов по отдельным RBD образам (жаль, не работает с EC :D):
   rbd perf image iotop -p <pool>
   и пулам: ceph osd pool stats

Встроенный Dashboard

1. Включение (без SSL):
```
apt-get install ceph-mgr-dashboard
ceph mgr module enable dashboard
ceph config set mgr mgr/dashboard/ssl false
ceph dashboard ac-user-create admin <пароль> administrator
```

2. Для графиков нужно поставить и настроить Prometheus и Grafana

3. ...Интегрировать Grafana с Dashboard-ом

4. ...И поставить в Grafana родные Dashboard-ы с github Ceph

5. ...Которые, разумеется, не умеют несколько кластеров в одном Prometheus-е

Prometheus & Grafana

1. Установить Grafana и Prometheus:
   `apt-get install grafana prometheus`

   --storage.tsdb.retention=5y в /etc/default/prometheus

2. Ставят, как правило, на 1 ноду, о HA не заморачиваются.
   Если надо, просто ставите одно и то же на 2 ноды.

3. Включить Prometheus Exporter (модуль ceph-mgr)
   `ceph mgr module enable prometheus`

4. Настроить Prometheus. В targets адреса всех mgr-ов, метрики отдаёт 1 из них.

```
- job_name: 'ceph'
  static_configs:
    - targets: ['172.31.1.7:9283','172.31.1.9:9283','172.31.1.16:9283',...]
  metric_relabel_configs:
    - source_labels: [instance]
      target_label: instance
      replacement: '172.31.1.5:9283'
```

Интеграция с Dashboard

1. Поставить плагины Grafana
```
grafana-cli plugins install vonage-status-panel
grafana-cli plugins install grafana-piechart-panel
```

2. Разрешить анонимного пользователя и встраивание Dashboard-ов
```
[auth.anonymous]
enabled = true
org_name = Main Org.
org_role = Viewer
[security]
allow_embedding = true
```

3. Задать URL графаны в Ceph
```
ceph dashboard set-grafana-api-url http://server
```

4. И поставить в неё родные Dashboard-ы
https://github.com/ceph/ceph/tree/master/monitoring/grafana/dashboards

Автоматизация

1. Хотя бы один раз полезно проделать это руками :-)

2. Не руками №1:
   ceph-ansible (dashboard.yml / группа хостов grafana-server)

3. Не руками №2:
   cephadm. Ставит Grafana даже по умолчанию
   Но всё-таки пока сыроват. Например, сменить адрес Grafana с ним невозможно, пароль сбрасывается к дефолтному, а alertmanager ставит без самих alert-ов.

4. На самом деле… встроенный Dashboard не так уж и нужен, консоли и Grafana достаточно

Дашборды с grafana.com

1. https://grafana.com/grafana/dashboards?search=ceph

2. 3 основных тиражируемых всеми — Ceph Cluster, Ceph Pools, Ceph OSD

3. Все довольно похожи, я тоже выкладываю свои
   https://grafana.com/grafana/dashboards/9550 9551 9552

4. Ещё CephFS — https://grafana.com/grafana/dashboards?search=cephfs

5. Ещё RadosGW — но его нет, можно взять из родных Dashboard-ов
   Плюс https://github.com/blemmenes/radosgw_usage_exporter

6. Мониторинг состояния нод — Node Exporter Full
   В т.ч. в node-exporter есть статистика bcache, но пока с багом :)

Важные метрики

1. Состояние кластера, состояние PG, состояние OSD
2. % заполнения кластера, % заполнения отдельных OSD
3. OSD apply/commit latency — пережиток прошлого
4. OSD op latency — лучше, но 4 КБ и 4 МБ операции идут вперемешку, поэтому интересно только при равномерной нагрузке
5. Slow ops — хотелось бы, но нет в exporter-е. Но можно через Graylog
6. Unfound / Degraded / Misplaced объекты, скорость бэкфилла
7. Смежные системы (сеть, NTP…)

Родные alert-ы под alertmanager

1. health error
2. health warn
3. low monitor quorum count
4. 10% OSDs down
5. OSD down
6. OSDs near full
7. flapping OSD
8. high pg count deviation
9. pgs inactive
10. pgs unclean
11. root volume full
12. network packets dropped
13. network packet errors
14. storage filling up
15. pool full
16. pool filling up

Необходимый минимум алертов

394783

На самом деле всего 3 метрики

1. Состояние кластера (OK/WARN/ERROR)
2. Состояние OSD
3. % заполнения OSD

:)

Здоровье дисков

1. Начиная с Nautilus есть сбор SMART по дискам и некое предсказание срока жизни дисков (даже облачное, тьфу на него)

2. Нужны свежие smartmontools (7.0+)

3. Список команд смотреть тут: ceph device -h

4. SMART должен выводиться в Dashboard-е, но не выводится

5. Интерпретируется он только предсказанием, а предсказание работает не очень хорошо, у меня вообще не завелось — вместо этого прилегли ceph-mgr-ы :)

Демо