High Available + Scalable Prometheus with Thanos in Alibaba

Guo’an Qin & Tao Li, Alibaba Group
About US

秦国安

花名 炎烈，高级技术专家
十二年工作经历，先后在百度、奇虎360、豌豆荚从事运维与开发工作
2016年加入阿里巴巴，主导完成数据库容器化及资源调度、存储计算分离
2018年加入容器平台团队，先后负责 kubelet & 数据生态及基础技术 SRE 相关工作

李涛

花名 吕风，高级开发工程师
2018年加入容器平台团队，负责数据生态相关工作
Background
Use Case

- Policy Engine
- Center Scheduler
- Algorithm
- Reliability
- UI
- Resource operation
- Data Analysis
- HPA
- VPA
Our Plan

Monitoring dashboard

Systems (data points trend/aggregation)
- Autoscaling
- Alerting
- Scheduler

Operational dashboard

Human (interactive arbitrary query focused)
- System operator
- Algorithm engineer
- Data analyst

- Hardware metrics
- OS metrics
- Application metrics
- Sigma system metrics
- ...
Architecture

**Infrastore**
- Timeseries API & Event API
  - OLTP
  - OLAP
  - Object Storage
  - Collector
  - Config

**Metrics Server**
- WebUI
- VPH/HPA
- Resource Management
- ...

**Node Structure**
- Wall
  - Infrastore Metrics and Event Handler
  - tsz engine
  - Event Store
  - Data
  - query interface
  - collector Interface
    - kernel
    - cgroup
    - storage
    - event log

**Kubelet**
- Metrics API & Manage API
  - Sketch
  - Autopolit
  - Pod Priority
  - Kubelet core features
  - Pouch Container Runtime Interface
  - Pods
    - Pod1
    - Pod2
    - Pod3
    - Pod4
Node Agent
Design Principles

- **7 * 24 Service**
  - Metrics standalone
  - Decoupling
  - Always Alive

- **Low Overhead**
  - Low Resource Usage
  - Unique Metrics Collection

- **Real-time Query & Predict**
  - High Performance
  - Predict

- **Flexible**
  - Custom Scrape Interval
  - Support Aggregation Rules
  - Independent Release
Agent Architecture
Implementation

TSDB Engine
- High Performance & Low Overhead
- Built-in Predict Function
- TSZ Compression

Plugins
- gRPC Plugins
- Standardized Interface
- Host + Plugins Pattern

HTTP API
- Support Local Query
- Prometheus Exporter

Event Storage
- OS
- Kernel
- Hardware
- Troubleshooting
Standing on the shoulders of giants

https://golang.org/
https://github.com/hashicorp/go-plugin
https://github.com/prometheus/tsdb
https://github.com/prometheus/prometheus/tree/master/promql
Data Achievement
Prometheus & Thanos in Practice
Prometheus – Architecture overview

1. Target discovery
2. Kubernetes Nodes
   - Kube State Metric
   - Node Exporter
   - Kube Component
3. Prometheus
4. Alert Manager
5. Notification
   - PromQL
Prometheus at Scale
Prometheus Splitting By Use

- Prometheus for DB
- Prometheus for etcd
- Prometheus for others
Prometheus Hierarchy

Cluster 1

Cluster 2

Cluster N

Global Prometheus

/federeate
Prometheus Horizontal Sharding

- Master
- Slave 1
- Slave 2
- Slave N
- Shard 0
- Shard 1
- Shard M
- Kubernetes Nodes

Prometheus

Kubernetes

Node

Shard
Prometheus High Availability

Cluster 1

Cluster 2

Cluster N

Global Prometheus

/federeate
Hello, Thanos

Highly available Prometheus setup with long term storage capabilities.

https://github.com/improbable-eng/thanos
Global Querier

Prometheus Query API → Querier → gRPC

Slaves 1, 2, N

Shards 0, 1, M

Kubernetes Nodes
Federated Querier

Region A (master)

Region B

Region C

Store API
Custom Target Discovery

1. Sync
2. Verify
3. Write To File
4. File Watch

Prometheus
Configuration Management

1. Control API → Flint Server → Sync API → Flint Client → SIGHUP → Prometheus
2. Save to File
3. Reload
04

Thinking Future
Enhancement

• Authentication
• Query cache
• Query safeguards
• Multi Tenancy
• Integrated with Apache Flink
Challenge

• Data migration & Recovery
• Scaling along with Kubernetes
Thanks!