Configuring Prometheus for High Performance

CloudNativeCon Berlin – 2017-03-30

Björn “Beorn” Rabenstein, Production Engineer, SoundCloud Ltd.

SOUNDCLOUD

Björn Rabenstein

beorn7

I am a **@Prometheus** developer and a **@SoundCloud** engineer.

@SoundCloud

Berlin

beorn@soundcloud.com

Organizations

Overview | Repositories 15 | Stars 159 | Followers 117 | Following 75

Popular repositories

Customize your pinned repositories

perks
Forked from bmizerany/perks

Effective Computation of Things

Go ★ 6 3

concurrentcount

Experiments to benchmark implementations of a concurrent counter.

Go ★ 6 2

talks

List of my public talks since 2015

★ 3

rsmod

Skeleton of moderator code for Rolling Stock

Python ★ 1

golang_protobuf_extensions
Forked from matttproud/golang_protobuf_extensions

A few Protocol Buffer extensions for the Go language (golang).

Go

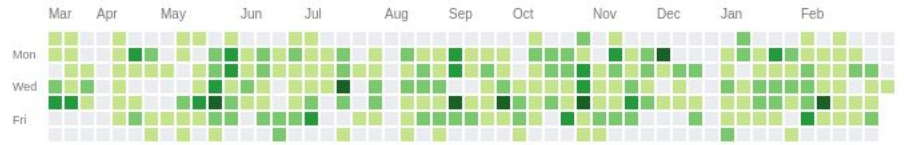
midgard-hausregel

House rules for the Midgard RPG (in German)

TeX

1,603 contributions in the last year

Contribution settings



```
$ prometheus -h
```

$ prometheus -h

```
usage: prometheus [<args>]

   -version false
      Print version information.

   -config.file "prometheus.yml"
      Prometheus configuration file name.

 == ALERTMANAGER ==

   -alertmanager.notification-queue-capacity 10000
      The capacity of the queue for pending alert manager notifications.

   -alertmanager.timeout 10s
      Alert manager HTTP API timeout.

   -alertmanager.url
      Comma-separated list of Alertmanager URLs to send notifications to.

 == LOG ==

   -log.format "\"logger:stderr\""
      Set the log target and format. Example:
      "logger:syslog?appname=bob&local=7" or "logger:stdout?json=true"

   -log.level "\"info\""
      Only log messages with the given severity or above. Valid levels:
      [debug, info, warn, error, fatal]

 == QUERY ==

   -query.max-concurrency 20
      Maximum number of queries executed concurrently.

   -query.staleness-delta 5m0s
      Staleness delta allowance during expression evaluations.

   -query.timeout 2m0s
      Maximum time a query may take before being aborted.

 == STORAGE ==

   -storage.local.checkpoint-dirty-series-limit 5000
      If approx. that many time series are in a state that would require
      a recovery operation after a crash, a checkpoint is triggered, even if
      the checkpoint interval hasn't passed yet. A recovery operation requires
      a disk seek. The default limit intends to keep the recovery time below
      1min even on spinning disks. With SSD, recovery is much faster, so you
      might want to increase this value in that case to avoid overly frequent
      checkpoints.

   -storage.local.checkpoint-interval 5m0s
      The period at which the in-memory metrics and the chunks not yet
      persisted to series files are checkpointed.

   -storage.local.chunk-encoding-version 1
      Which chunk encoding version to use for newly created chunks.
      Currently supported is 0 (delta encoding), 1 (double-delta encoding), and
      2 (double-delta encoding with variable bit-width).

   -storage.local.dirty false
      If set, the local storage layer will perform crash recovery even if
      the last shutdown appears to be clean.

   -storage.local.engine "persisted"
      Local storage engine. Supported values are: 'persisted' (full local
      storage with on-disk persistence) and 'none' (no local storage).

   -storage.local.index-cache-size.fingerprint-to-metric 10485760
      The size in bytes for the fingerprint to metric index cache.

   -storage.local.index-cache-size.fingerprint-to-timerange 5242880
      The size in bytes for the metric time range index cache.

   -storage.local.index-cache-size.label-name-to-label-values 10485760
      The size in bytes for the label name to label values index cache.

   -storage.local.index-cache-size.label-pair-to-fingerprints 20971520
      The size in bytes for the label pair to fingerprints index cache.

   -storage.local.max-chunks-to-persist 524288
      How many chunks can be waiting for persistence before sample
      ingestion will be throttled. Many chunks waiting to be persisted will
      increase the checkpoint size.

   -storage.local.memory-chunks 1048576
      How many chunks to keep in memory. While the size of a chunk is
      1kiB, the total memory usage will be significantly higher than this value
      * 1kiB. Furthermore, for various reasons, more chunks might have to be
      kept in memory temporarily. Sample ingestion will be throttled if the
      configured value is exceeded by more than 10%.

   -storage.local.num-fingerprint-mutexes 4096
      The number of mutexes used for fingerprint locking.

   -storage.local.path "data"
      Base path for metrics storage.

   -storage.local.pedantic-checks false
      If set, a crash recovery will perform checks on each serie
      This might take a very long time.

   -storage.local.retention 360h0m0s
      How long to retain samples in the local storage.

   -storage.local.series-file-shrink-ratio 0.1
      A series file is only truncated (to delete samples that h
      exceeded the retention period) if it shrinks by at least
      ratio. This saves I/O operations while causing only a lim
      space overhead. If 0 or smaller, truncation will be perfo
      single dropped chunk, while 1 or larger will effectively
      truncation.

   -storage.local.series-sync-strategy "adaptive"
      When to sync series files after modification. Possible va
      'never', 'always', 'adaptive'. Sync'ing slows down storag
      but reduces the risk of data loss in case of an OS crash.
      'adaptive' strategy, series files are sync'd for as long
      is not too much behind on chunk persistence.

   -storage.remote.graphite-address
      The host:port of the remote Graphite server to send sampl
      None, if empty.

   -storage.remote.graphite-prefix
      The prefix to prepend to all metrics exported to Graphite
      empty.

   -storage.remote.graphite-transport "tcp"
      Transport protocol to use to communicate with Graphite. '
      empty.

   -storage.remote.influxdb-url
      The URL of the remote InfluxDB server to send samples to.
      empty.

   -storage.remote.influxdb.database "prometheus"
      The name of the database to use for storing samples in In

   -storage.remote.influxdb.retention-policy "default"
      The InfluxDB retention policy to use.

   -storage.remote.influxdb.username
      The username to use when sending samples to InfluxDB. The
      corresponding password must be provided via the INFLUXDB_
      variable.

   -storage.remote.opentsdb-url
      The URL of the remote OpenTSDB server to send samples to.
      empty.

   -storage.remote.timeout 30s
```

$ prometheus -h

```
usage: prometheus [<args>]

   -version false
      Print version information.

   -config.file "prometheus.yml"
      Prometheus configuration file name.

 == ALERTMANAGER ==

   -alertmanager.notification-queue-capacity 10000
      The capacity of the queue for pending alert manager notifications.

   -alertmanager.timeout 10s
      Alert manager HTTP API timeout.

   -alertmanager.url
      Comma-separated list of Alertmanager URLs to send notifications to.

 == LOG ==

   -log.format "\"logger:stderr\""
      Set the log target and format. Example:
      "logger:syslog?appname=bob&local=7" or "logger:stdout?json=true"

   -log.level "\"info\""
      Only log messages with the given severity or above. Valid levels:
      [debug, info, warn, error, fatal]

 == QUERY ==

   -query.max-concurrency 20
      Maximum number of queries executed concurrently.

   -query.staleness-delta 5m0s
      Staleness delta allowance during expression evaluations.

   -query.timeout 2m0s
      Maximum time a query may take before being aborted.

 == STORAGE ==

   -storage.local.checkpoint-dirty-series-limit 5000
      If approx. that many time series are in a state that would require
      a recovery operation after a crash, a checkpoint is triggered, even if
      the checkpoint interval hasn't passed yet. A recovery operation requires
      a disk seek. The default limit intends to keep the recovery time below
      1min even on spinning disks. With SSD, recovery is much faster, so you
      might want to increase this value in that case to avoid overly frequent
      checkpoints.

   -storage.local.checkpoint-interval 5m0s
      The period at which the in-memory metrics and the chunks not yet
      persisted to series files are checkpointed.

   -storage.local.chunk-encoding-version 1
      Which chunk encoding version to use for newly created chunks.
      Currently supported is 0 (delta encoding), 1 (double-delta encoding), and
      2 (double-delta encoding with variable bit-width).

   -storage.local.dirty false
      If set, the local storage layer will perform crash recovery even if
      the last shutdown appears to be clean.

   -storage.local.engine "persisted"
      Local storage engine. Supported values are: 'persisted' (full local
      storage with on-disk persistence) and 'none' (no local storage).

   -storage.local.index-cache-size.fingerprint-to-metric 10485760
      The size in bytes for the fingerprint to metric index cache.

   -storage.local.index-cache-size.fingerprint-to-timerange 5242880
      The size in bytes for the metric time range index cache.

   -storage.local.index-cache-size.label-name-to-label-values 10485760
      The size in bytes for the label name to label values index cache.

   -storage.local.index-cache-size.label-pair-to-fingerprints 20971520
      The size in bytes for the label pair to fingerprints index cache.

   -storage.local.max-chunks-to-persist 524288
      How many chunks can be waiting for persistence before sample
      ingestion will be throttled. Many chunks waiting to be persisted will
      increase the checkpoint size.

   -storage.local.memory-chunks 1048576
      How many chunks to keep in memory. While the size of a chunk is
      1kiB, the total memory usage will be significantly higher than this value
      * 1kiB. Furthermore, for various reasons, more chunks might have to be
      kept in memory temporarily. Sample ingestion will be throttled if the
      configured value is exceeded by more than 10%.

   -storage.local.num-fingerprint-mutexes 4096
      The number of mutexes used for fingerprint locking.

   -storage.local.path "data"
      Base path for metrics storage.

   -storage.local.pedantic-checks false
      If set, a crash recovery will perform checks on each seri
      This might take a very long time.

   -storage.local.retention 360h0m0s
      How long to retain samples in the local storage.

   -storage.local.series-file-shrink-ratio 0.1
      A series file is only truncated (to delete samples that h
      exceeded the retention period) if it shrinks by at least
      ratio. This saves I/O operations while causing only a lim
      space overhead. If 0 or smaller, truncation will be perfo
      single dropped chunk, while 1 or larger will effectively
      truncation.

   -storage.local.series-sync-strategy "adaptive"
      When to sync series files after modification. Possible va
      'never', 'always', 'adaptive'. Sync'ing slows down storag
      but reduces the risk of data loss in case of an OS crash.
      'adaptive' strategy, series files are sync'd for as long
      is not too much behind on chunk persistence.

   -storage.remote.graphite-address
      The host:port of the remote Graphite server to send sampl
      None, if empty.

   -storage.remote.graphite-prefix
      The prefix to prepend to all metrics exported to Graphite
      empty.

   -storage.remote.graphite-transport "tcp"
      Transport protocol to use to communicate with Graphite. '
      empty.

   -storage.remote.influxdb-url
      The URL of the remote InfluxDB server to send samples to.
      empty.

   -storage.remote.influxdb.database "prometheus"
      The name of the database to use for storing samples in In

   -storage.remote.influxdb.retention-policy "default"
      The InfluxDB retention policy to use.

   -storage.remote.influxdb.username
      The username to use when sending samples to InfluxDB. The
      corresponding password must be provided via the INFLUXDB_
      variable.

   -storage.remote.opentsdb-url
      The URL of the remote OpenTSDB server to send samples to.
      empty.

   -storage.remote.timeout 30s
```

Julius Volz

juliusv

Unfollow

Block or report user

Berlin, Germany

http://juliusv.com

Organizations

Overview | Repositories 41 | Stars 9 | Followers 126 | Following 3

Popular repositories

prometheus_workshop

Example client/server app used for a Prometheus workshop

Go ★ 32 8

ne-statsd-backend

Network-efficient preaggregating StatsD backend for StatsD

JavaScript ★ 28 4

cli_exercises

★ 4 2

go_link_redirector

Go Link Redirector

Ruby ★ 3 2

prometheus_demo_service

A demo server that exports synthetic bogus Prometheus metrics

Go ★ 2 2

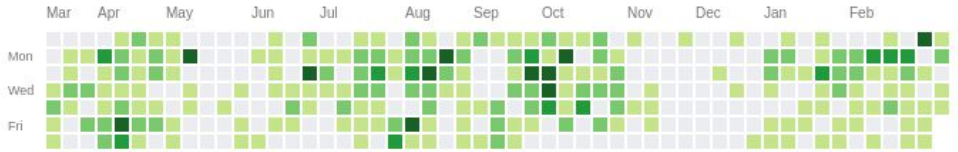
OwnTube

Forked from Piratenfraktion-Berlin/OwnTube

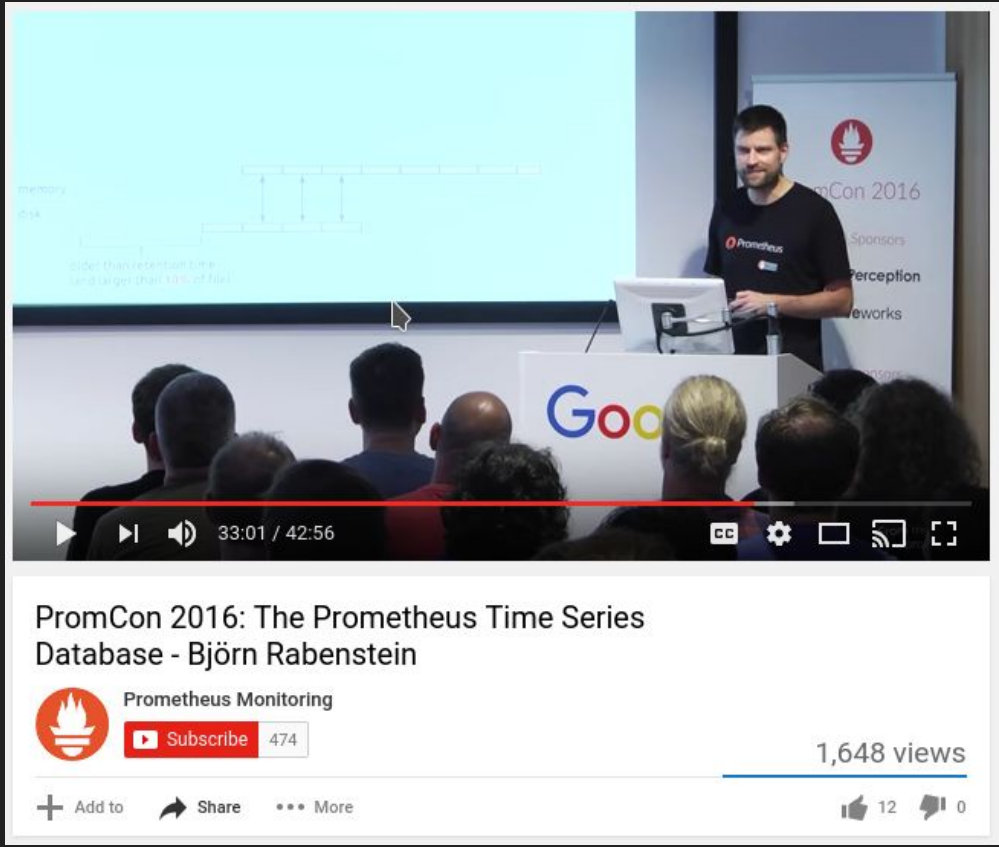
OwnTube is your personal video portal based on Django

Python

1,032 contributions in the last year



https://youtu.be/HbnGSNEjhUc



https://prometheus.io/docs/operating/storage/

INTRODUCTION
CONCEPTS
QUERYING
VISUALIZATION
INSTRUMENTING
OPERATING
Configuration
Storage
Federation

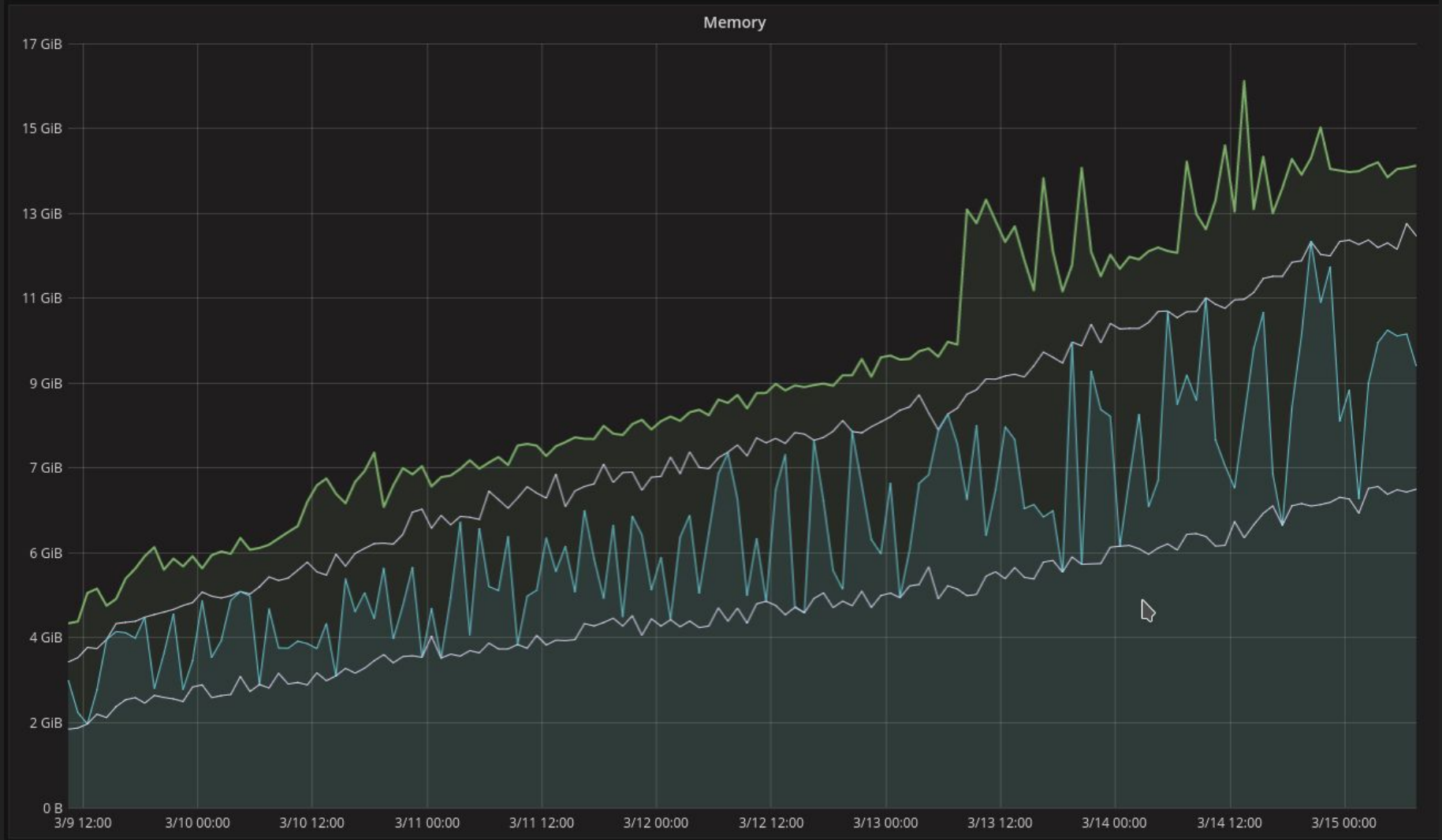
STORAGE

Prometheus has a sophisticated local storage subsystem. For indexes, it uses LevelDB. For the bulk sample data, it has its own custom storage layer, which organizes sample data in chunks of constant size (1024 bytes payload). These chunks are then stored on disk in one file per time series.

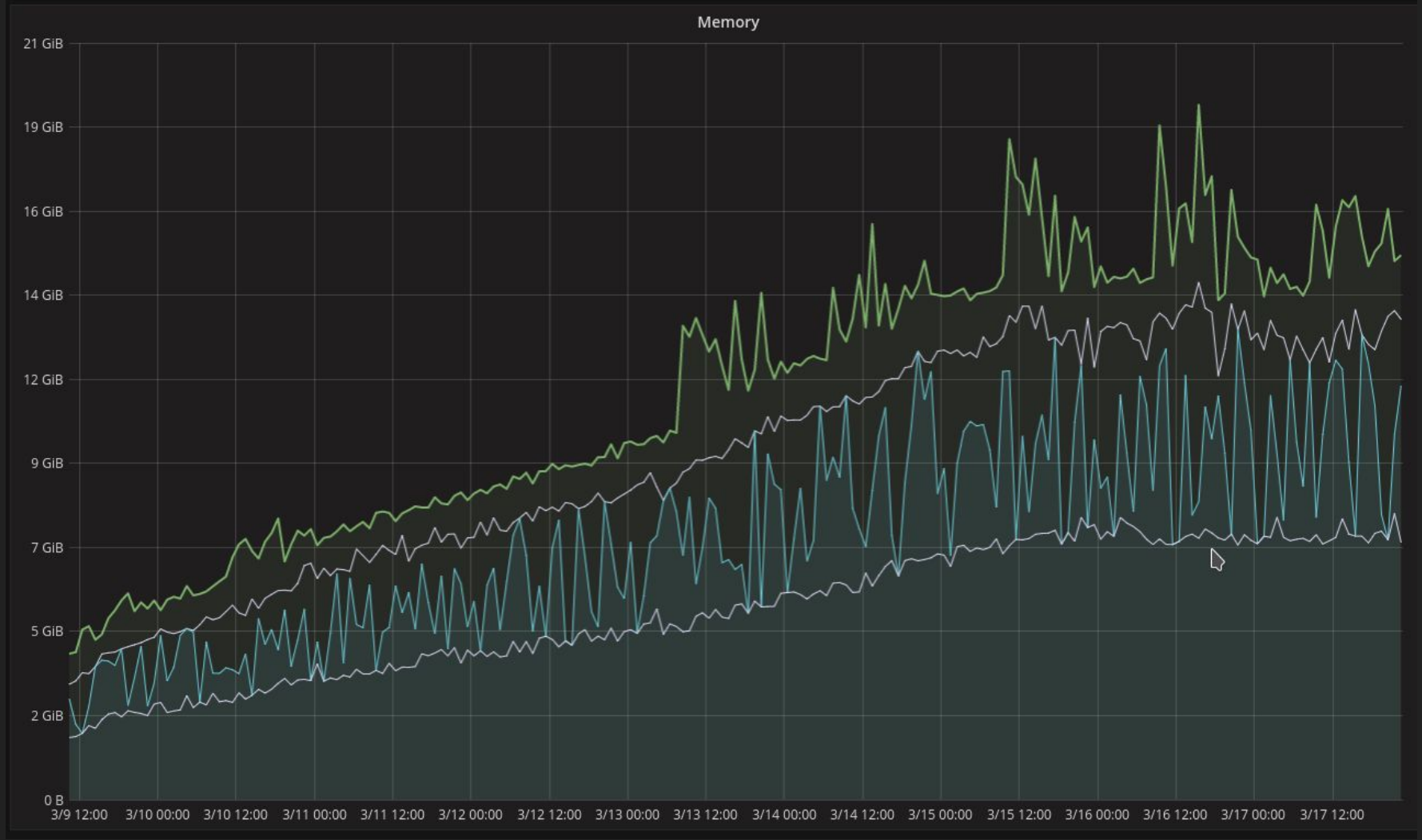
- Memory usage
- Disk usage
- Chunk encoding
- Settings for high numbers of time series
- Persistence pressure and "rushed mode"
- Settings for very long retention time
- Helpful metrics
- Crash recovery
- Data corruption

Memory usage

Prometheus keeps all the currently used chunks in memory. In addition, it keeps the most recently used chunks in memory up to a threshold configurable via the `storage.local.memory-chunks` flag. If you have a lot of RAM available, you might want to increase it above the default value of 1048576 (and vice versa, if you run into RAM problems, you can try to decrease it). Note that the actual RAM usage of your server will be much higher than what you would expect from multiplying `storage.local.memory-chunks` by 1024 bytes. There is inevitable overhead for managing the sample data in the storage layer. Also, your server is doing many more things than just storing samples. The actual overhead depends on your usage pattern. In extreme cases, Prometheus has to keep more chunks in memory than configured because all those chunks are in use at the same time. You have to experiment a bit. The metrics


Memory
17 GiB
15 GiB
13 GiB
11 GiB
9 GiB
7 GiB
6 GiB
4 GiB
2 GiB
0 B
3/9 12:00
3/10 00:00
3/10 12:00
3/11 00:00
3/11 12:00
3/12 00:00
3/12 12:00
3/13 00:00
3/13 12:00
3/14 00:00
3/14 12:00
3/15 00:00

OOM


Memory
21 GiB
19 GiB
16 GiB
14 GiB
12 GiB
9 GiB
7 GiB
5 GiB
2 GiB
0 B
3/9 12:00
3/10 00:00
3/10 12:00
3/11 00:00
3/11 12:00
3/12 00:00
3/12 12:00
3/13 00:00
3/13 12:00
3/14 00:00
3/14 12:00
3/15 00:00
3/15 12:00
3/16 00:00
3/16 12:00
3/17 00:00
3/17 12:00

-storage.local.memory-chunks 1048576

```
-storage.local.memory-chunks         1048576
-storage.local.max-chunks-to-persist 524288
```

prometheus_local_storage_persistence_urgency_score
prometheus_local_storage_max_chunks_to_persist
prometheus_local_storage_chunks_to_persist
prometheus_local_storage_max_memory_chunks
prometheus_local_storage_memory_chunks

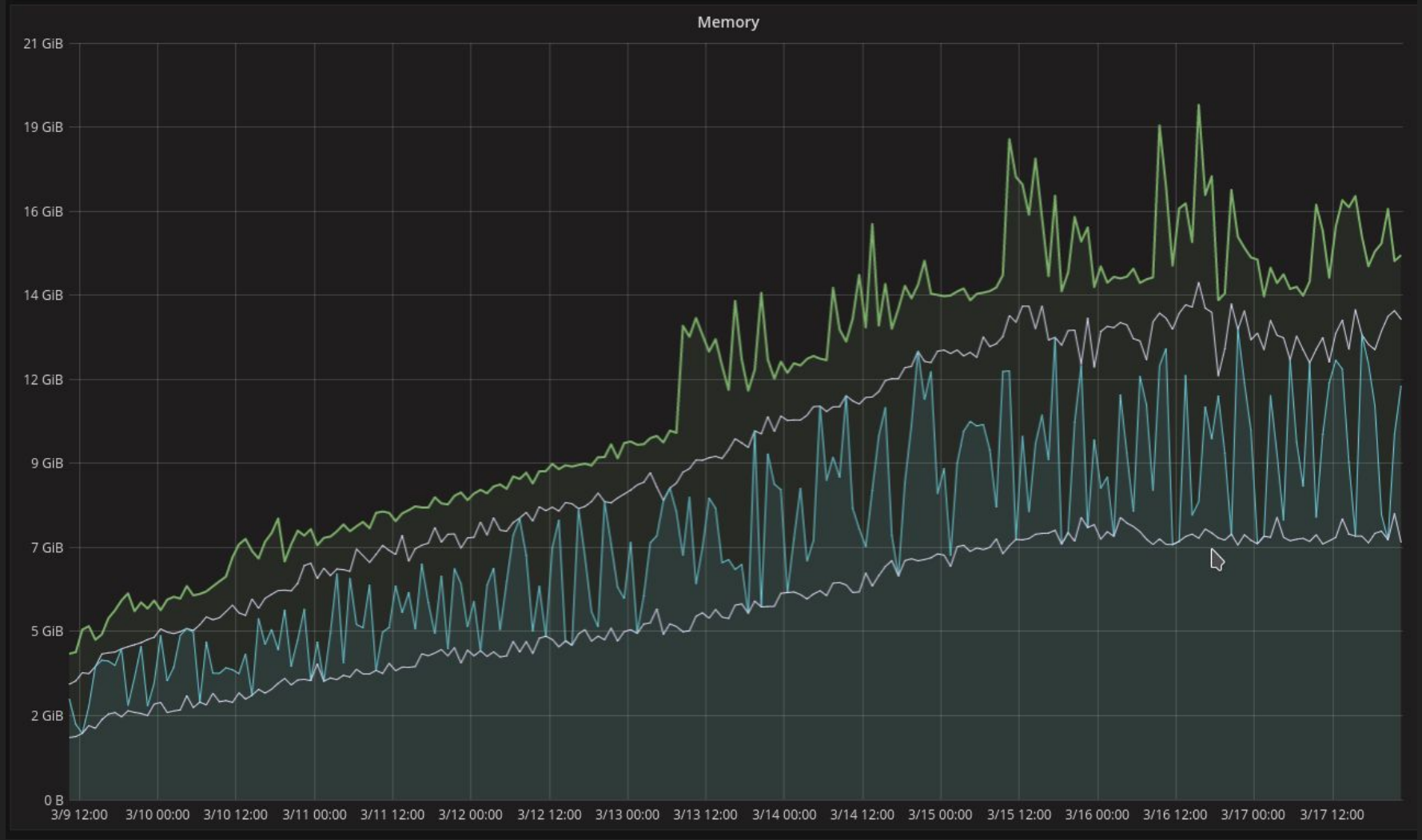
```
WARN[4948] Storage has entered rushed mode.
chunksToPersist=12000120 maxChunksToPersist=15000000
maxMemoryChunks=19000000 memoryChunks=16800306
source=storage.go:1660 urgencyScore=0.800008

INFO[5294] Storage has left rushed mode.
chunksToPersist=10499991 maxChunksToPersist=15000000
maxMemoryChunks=19000000 memoryChunks=17911345
source=storage.go:1647 urgencyScore=0.6999994
```

-storage.local.series-sync-strategy adaptive → never | always

```
time="2016-11-27T01:03:35Z" level=error msg="Storage needs
throttling. Scrapes and rule evaluations will be skipped."
chunksToPersist=524468 maxChunksToPersist=524288
maxToleratedMemChunks=3460300 memoryChunks=3152841
source="storage.go:908"
```

```
-storage.local.memory-chunks         1048576
-storage.local.max-chunks-to-persist 524288
```


Memory
21 GiB
19 GiB
16 GiB
14 GiB
12 GiB
9 GiB
7 GiB
5 GiB
2 GiB
0 B
3/9 12:00
3/10 00:00
3/10 12:00
3/11 00:00
3/11 12:00
3/12 00:00
3/12 12:00
3/13 00:00
3/13 12:00
3/14 00:00
3/14 12:00
3/15 00:00
3/15 12:00
3/16 00:00
3/16 12:00
3/17 00:00
3/17 12:00

```
# The value for memory_chunks is a conservative setting for normal operation,
# but with a lot of load or a lot of time series, you might want to tune it
# even further down. On the other hand, in most cases, you can probably set it
# to higher values if you know what you are doing.
default['prom']['memory_chunks'] = node['memory']['total'].to_i / 6
# Half of memory_chunks for max_chunks_to_persist is a rule of thumb. With
# many time series, you might want to set it to a lower value. With only a few
# series but high ingestion rate, you might want to set it to a higher value.
default['prom']['max_chunks_to_persist'] = node['prom']['memory_chunks'] / 2
```

1.5.2

Brian Brazil

brian-brazil

Unfollow

Block or report user

Robust Perception

brian.brazil@robustperception.io

http://www.robustperception.io

Organizations

Overview | Repositories 40 | Stars 27 | Followers 127 | Following 0

Pinned repositories

prometheus/prometheus

The Prometheus monitoring system and time series database.

Go ★ 8.6k ⑂ 871

prometheus/client_python

Prometheus instrumentation library for Python applications

Python ★ 195 ⑂ 82

prometheus/blackbox_exporter

Blackbox prober exporter

Go ★ 151 ⑂ 55

prometheus/client_java

Prometheus instrumentation library for JVM applications

Java ★ 127 ⑂ 109

prometheus/snmp_exporter

SNMP Exporter for Prometheus

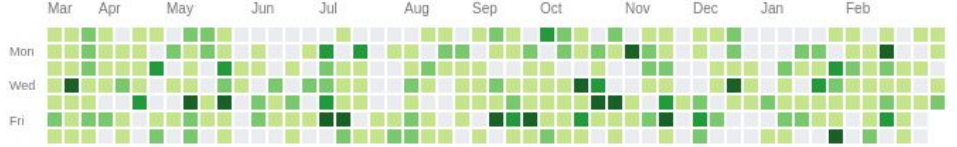
Go ★ 69 ⑂ 49

prometheus/docs

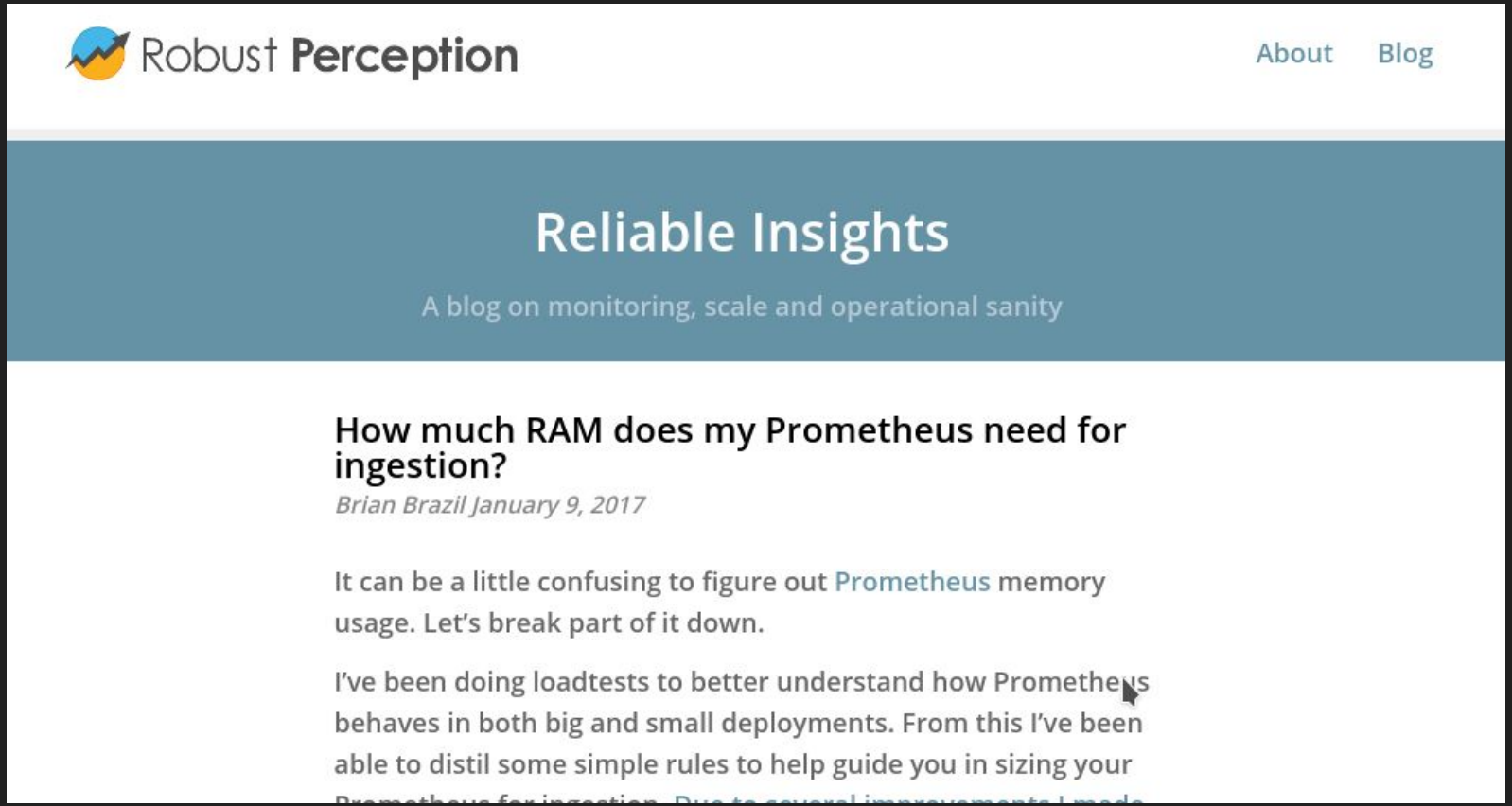
Prometheus documentation: content and static site generator

CSS ★ 50 ⑂ 159

879 contributions in the last year



www.robustperception.io/how-much-ram-does-my-prometheus-need-for-ingestion



```
(increase(prometheus_local_storage_chunk_ops_total{job="prometheus",type="create"}[6h]) / 2 / .8 * 1.6)
→ 6838483

-storage.local.max-chunks-to-persist 6838483


max_over_time(prometheus_local_storage_memory_series{job="prometheus"}[6h])
→ 4611785

-storage.local.memory-chunks 11450268
```

```
(increase(prometheus_local_storage_chunk_ops_total{job="prometheus",type="create"}[6h]) / 2 / .8 * 1.6)
→ 6838483
```

`-storage.local.max-chunks-to-persist 6838483 → 4194304`

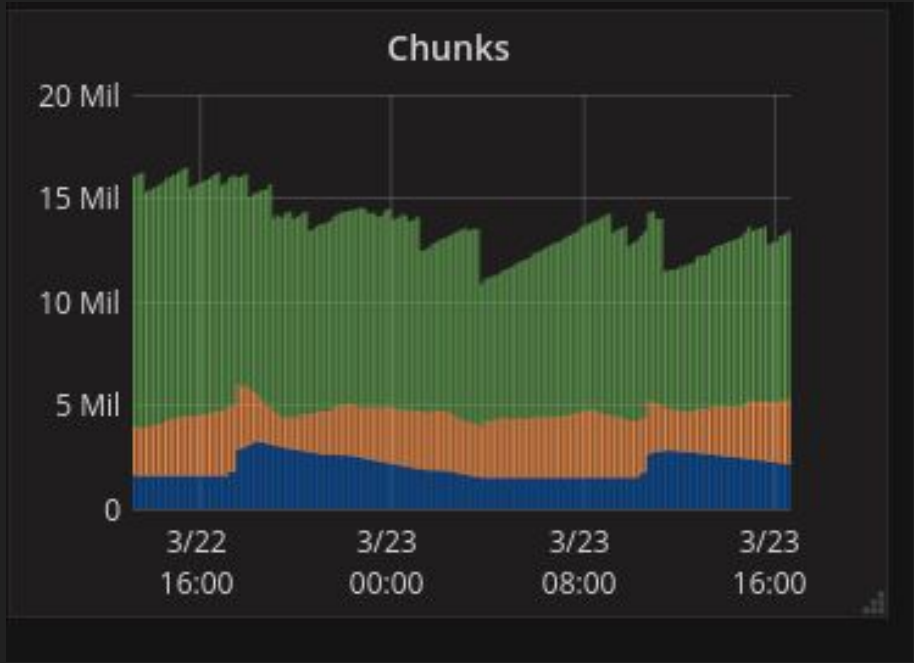
```
max_over_time(prometheus_local_storage_memory_series{job="prometheus"}[6h])
→ 4611785
```

`-storage.local.memory-chunks 11450268 → 8388608`

1.6

```
-storage.local.target-heap-size  2147483648
```


Series in memory
5.0 Mil
4.5 Mil
4.0 Mil
3.5 Mil
3.0 Mil
2.5 Mil
2.0 Mil
1.5 Mil
3/22
16:00
3/23
00:00
3/23
08:00
3/23
16:00

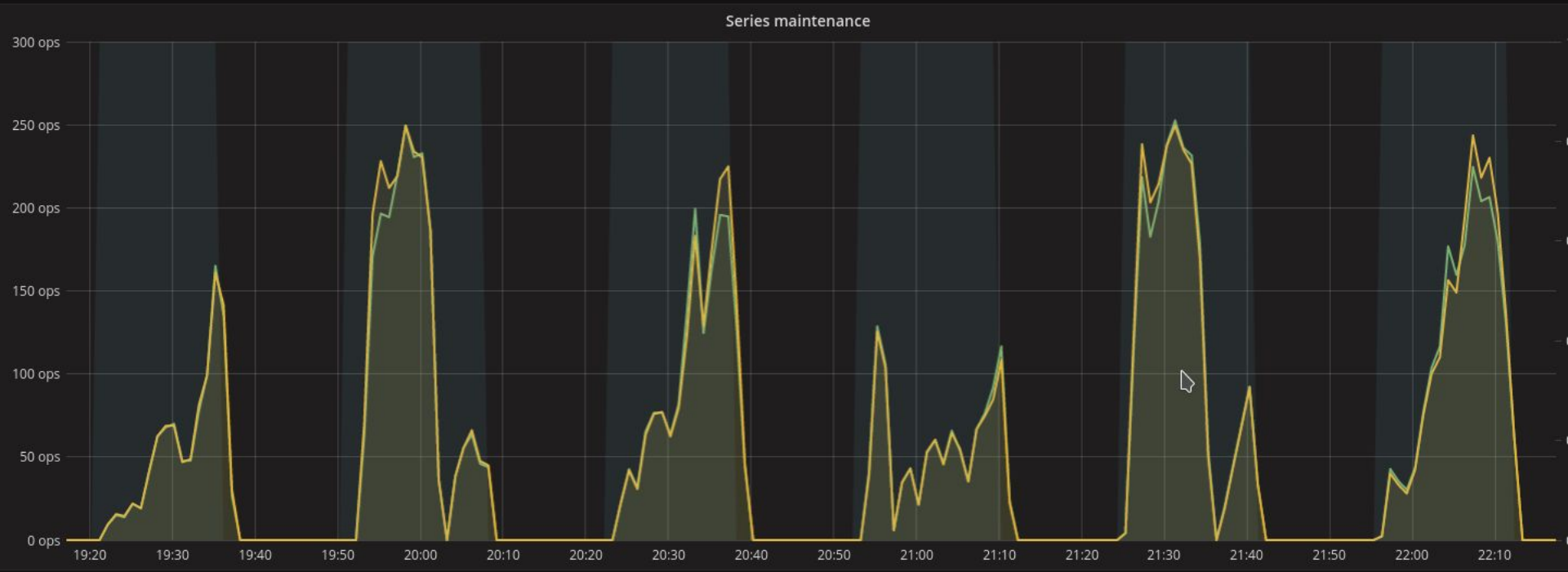

Chunks
20 Mil
15 Mil
10 Mil
5 Mil
0
3/22
16:00
3/23
00:00
3/23
08:00
3/23
16:00

Unchanged: Checkpointing *Ugh!*

```
-storage.local.checkpoint-interval 5m0s
```

```
-storage.local.checkpoint-dirty-series-limit 5000
```

(set way higher with SSD)


Series maintenance
300 ops
250 ops
200 ops
150 ops
100 ops
50 ops
0 ops
19:20
19:30
19:40
19:50
20:00
20:10
20:20
20:30
20:40
20:50
21:00
21:10
21:20
21:30
21:40
21:50
22:00
22:10

2.0

Fabian Reinartz

fabxc

Unfollow

Block or report user

CoreOS, Inc.

Berlin

Organizations

Overview | Repositories 26 | Stars 41 | Followers 71 | Following 2

Popular repositories

jiralerts

Prometheus Alertmanager integration for JIRA

Python ★ 15 ⑂ 4

tindex

Go ★ 9 ⑂ 1

tsdb

Go ★ 5

etcd_exporter

Go ★ 4 ⑂ 1

prom_sd_example

An etcd bridge for Prometheus service discovery

Go ★ 3 ⑂ 3

pagebuf

Go ★ 3

1,698 contributions in the last year

prometheus.io

Meet Julius & me 15:30 at CNCF booth G9 for more Q&A.

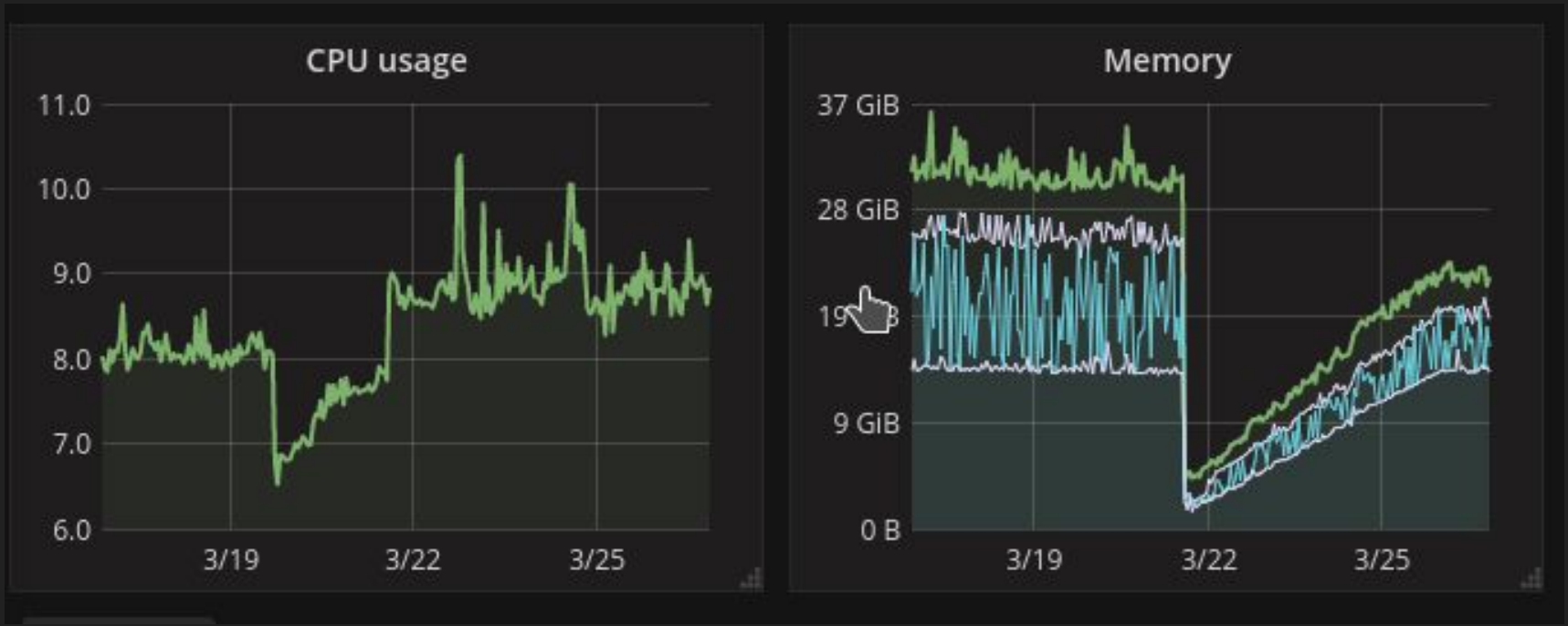
Bonus slides

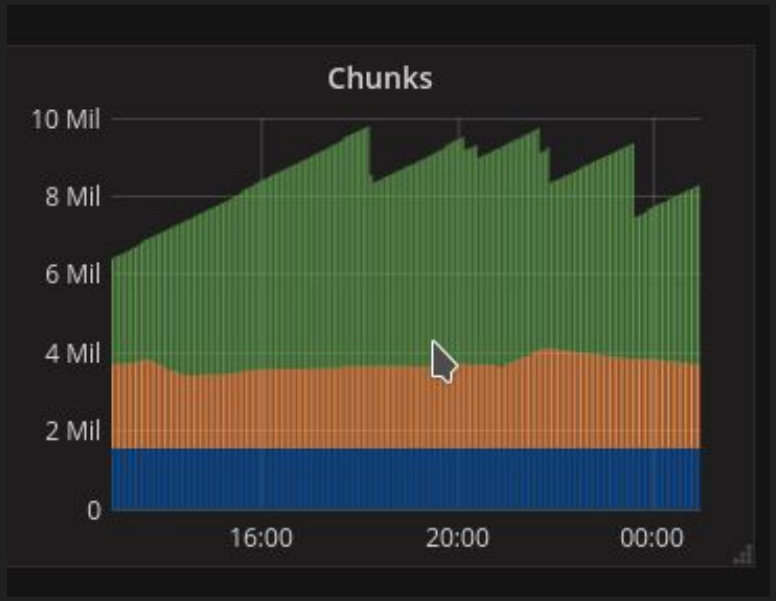
Further tricks

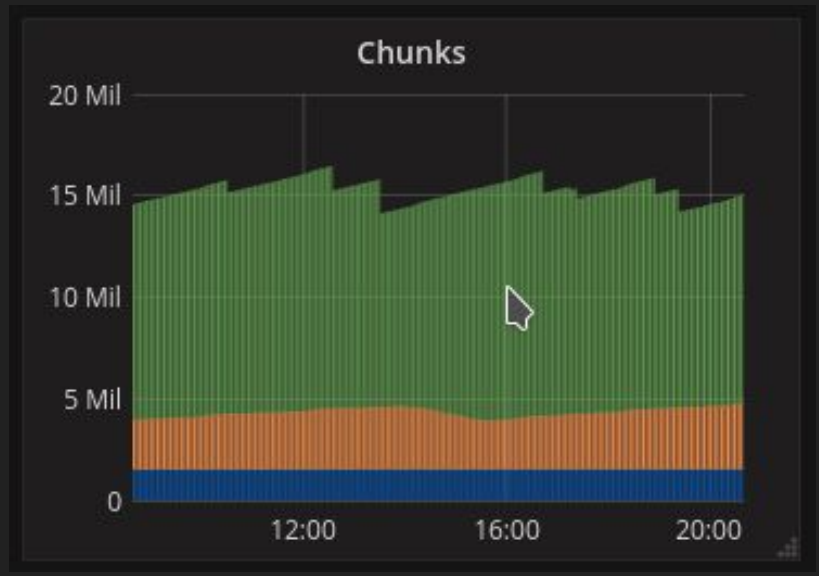
```
-storage.local.index-cache-size.*

-storage.local.num-fingerprint-mutexes 4096
```

`export GOGC=100` → `export GOGC=40` *(default in 1.6)*


CPU usage
11.0
10.0
9.0
8.0
7.0
6.0
3/19
3/22
3/25
Memory
37 GiB
28 GiB
9 GiB
0 B
3/19
3/22
3/25


Chunks
10 Mil
8 Mil
6 Mil
4 Mil
2 Mil
0
16:00
20:00
00:00


Chunks
20 Mil
15 Mil
10 Mil
5 Mil
0
12:00
16:00
20:00

disk full

```
-storage.local.retention 360h
```

```
rate(prometheus_local_storage_ingested_samples_total[10m]
)
```

```
-storage.local.chunk-encoding-version 1
```

Chunk type	bytes per sample	cores	rule evaluation duration
1	3.3	1.6	2.9s
2	1.3	2.4	4.9s

```
-storage.local.series-file-shrink-ratio 0.1
```

Matthias Rampke

matthiasr

Unfollow

Block or report user

SoundCloud
Berlin, Germany
matthias@rampke.de
http://rampke.de/

Overview | Repositories 45 | Stars 66 | Followers 35 | Following 8

Popular repositories

monitoring-workshop
Monitoring Workshop at ContainerDays HH 2016
Go ★ 6

logformat
Forked from erlehmann/logformat
converts zweipktfkts IRC plaintext logs to HTML5
Ruby ★ 5 1

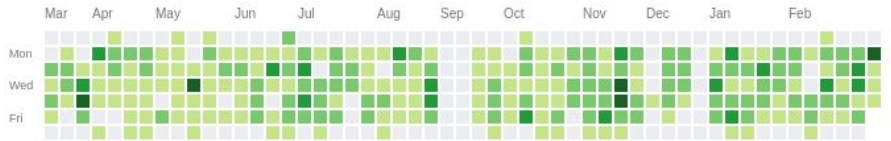
teco
Tape Editor And Corrector
C ★ 4 1

summon-arm-toolchain
Forked from eFrane/summon-arm-toolchain
A very simple build script for bare metal arm toolchain. NO LINUX!
Shell ★ 4 2

twitter
scripts and tools for twitter
Python ★ 3

mk
mkhybrid+cdrecord frontend for BeOS R5. needs fixing to run on Haiku. Original coding by Lukas F. Hartmann
C++ ★ 3 1

3,353 contributions in the last year



Mitsuhiro Tanda

mtanda

Follow

Block or report user

Tokyo, Japan

mitsuhiro.tanda@gmail.com

Overview | Repositories 51 | Stars 17 | Followers 20 | Following 2

Pinned repositories

grafana
Forked from grafana/grafana

Grafana - A Graphite & InfluxDB Dashboard and Graph Editor

Go

grafana-histogram-panel

JavaScript ★ 51 10

grafana-heatmap-epoch-panel

JavaScript ★ 28

prometheus
Forked from prometheus/prometheus

The Prometheus monitoring system and time series database.

Go

fluent-plugin-rds-mysql-log
Forked from shinsaka/fluent-plugin-rds-pgsql-log

fluentd plugin for Amazon RDS for PostgreSQL log input

Ruby

803 contributions in the last year