HPC-IO Analyse & Visualisierung mit Grafana

Simon Rosenberger

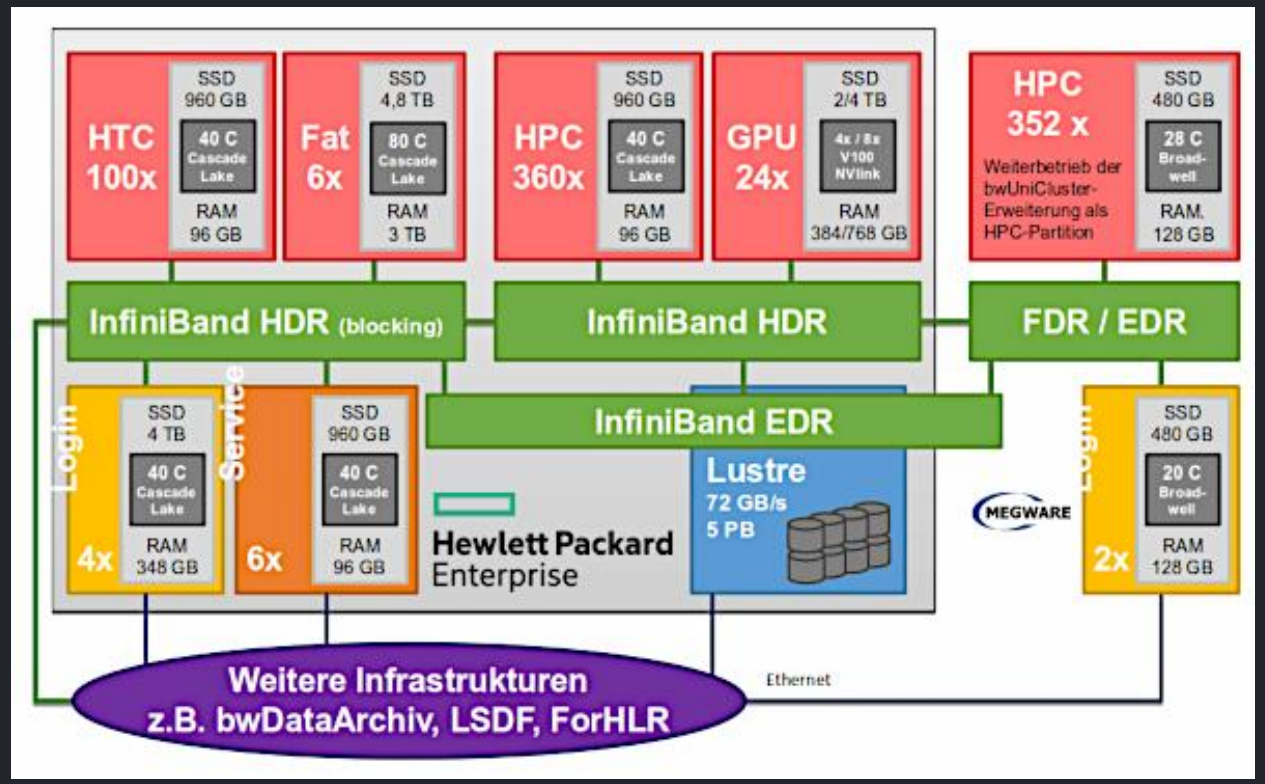
Gliederung

1. Motivation
2. Libiotrace
3. Was sind InfluxDB / Grafana?
4. Grafana Plugins
5. Grafana Dashboard

Motivation

Systeme wie BwUniCLuster:

- Große Mengen an Speicher & Rechenkapazität
- Parallele Anbindung über InfiniBand
- Disc-IO Anbindung über IB-Netzwerk ist langsam



Motivation

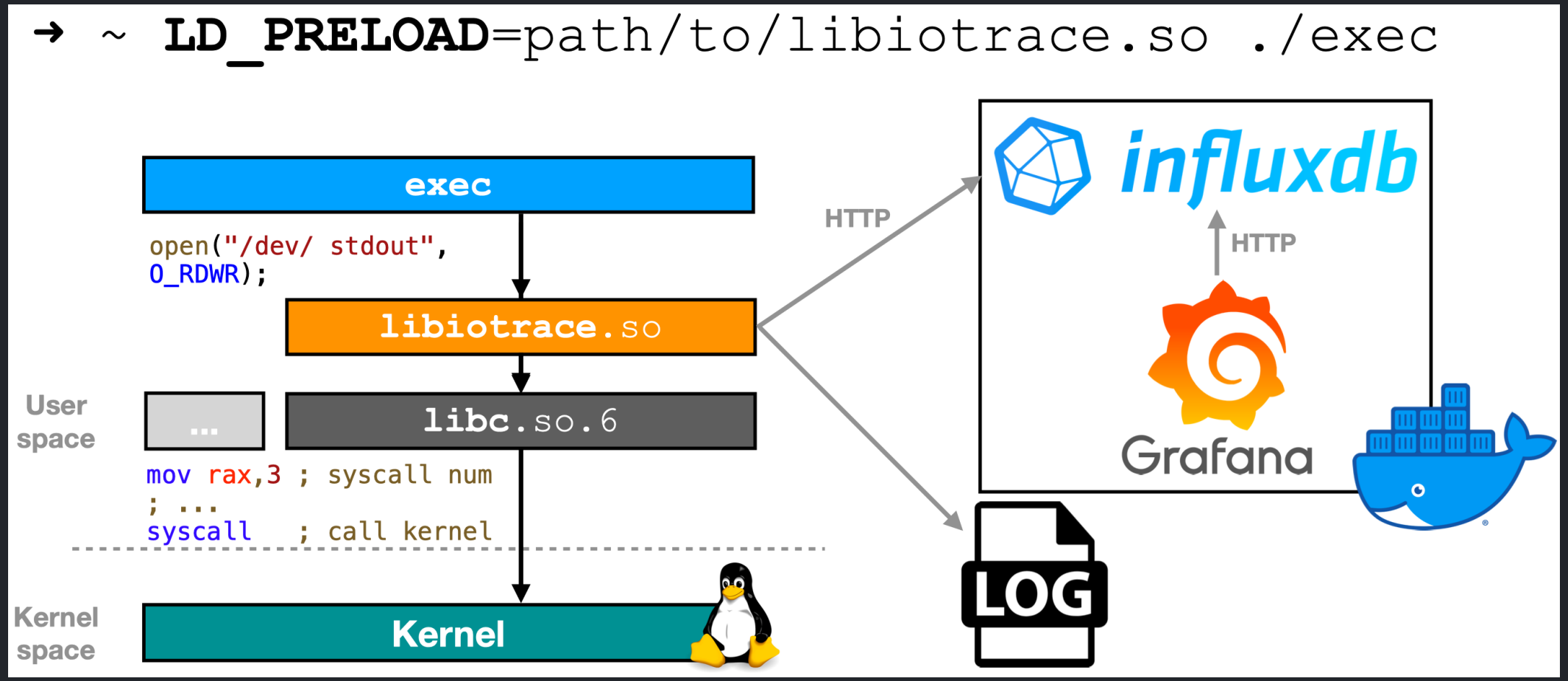
Problem Livetracing

- Nicht erkennbar welche Prozesse und Threads im System viel schreiben & lesen und die Performance bremsen
  ➔ unerkannte Bottlenecks

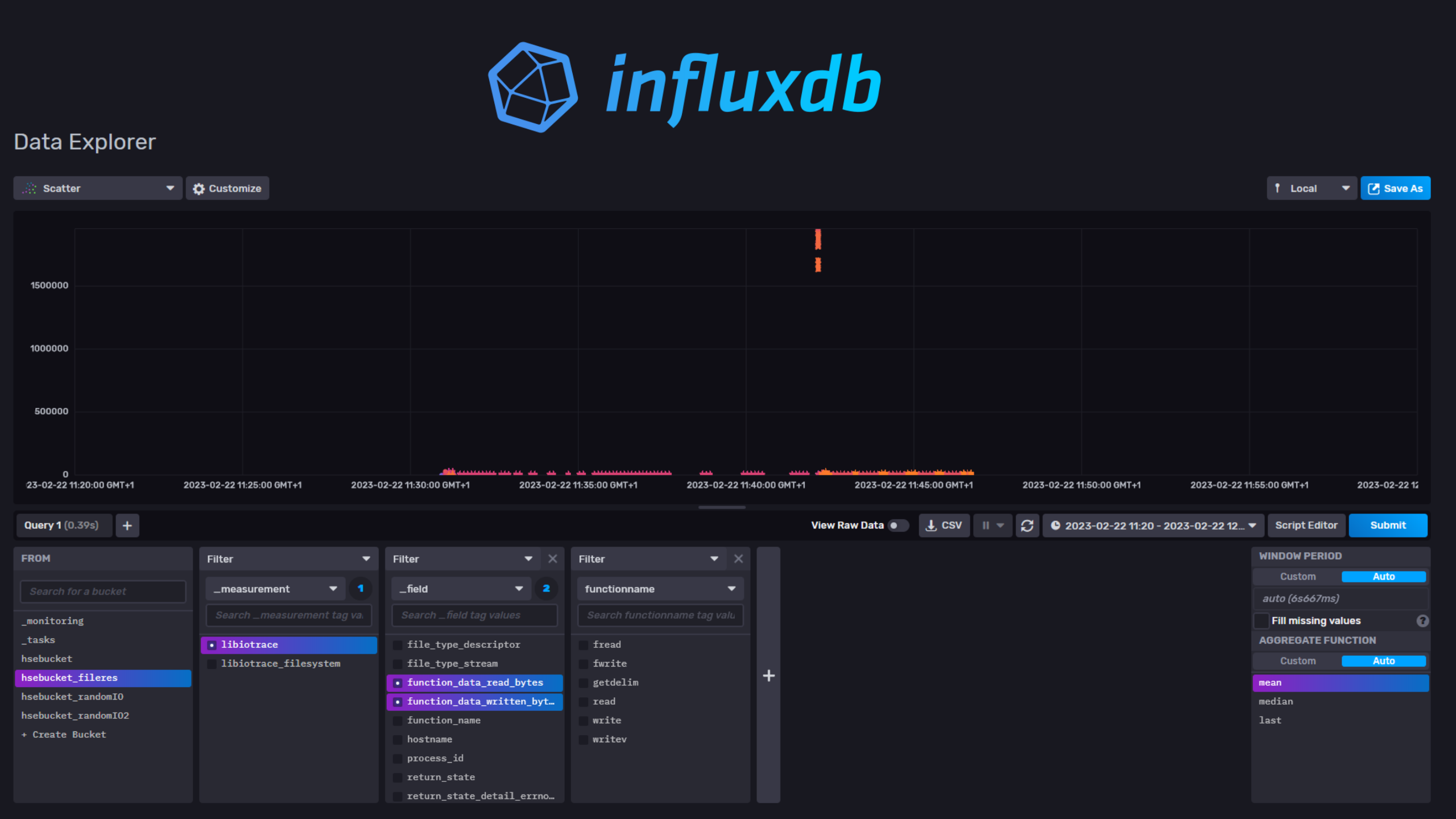
Lösungsansatz

- „HeatMap“ mit Gesamtübersicht die die Auslastung von Prozessen und unterlagerter Threads anzeigt
- Detailliertere Daten mithilfe eines Force-Graphen

Libiotrace


➜ ~ LD_PRELOAD=path/to/libiotrace.so ./exec
exec
open("/dev/ stdout", O_RDWR);
libiotrace.so
...
libc.so.6
mov rax,3 ; syscall num
; ...
syscall ; call kernel
User space
Kernel space
Kernel
HTTP
HTTP
influxdb
Grafana
LOG

- Opensource, nicht relationale Datenbank
- Datenabfrage über eigene Sprache: „Flux“
- Einfache Integration in Visualisierung über Queries
- Optimiert für High Performance Computing mit vielen Daten


influxdb
Data Explorer
Scatter
Customize
Local
Save As
1500000
1000000
500000
0
23-02-22 11:20:00 GMT+1
2023-02-22 11:25:00 GMT+1
2023-02-22 11:30:00 GMT+1
2023-02-22 11:35:00 GMT+1
2023-02-22 11:40:00 GMT+1
2023-02-22 11:45:00 GMT+1
2023-02-22 11:50:00 GMT+1
2023-02-22 11:55:00 GMT+1
2023-02-22 12
Query 1 (0.39s)
View Raw Data
CSV
2023-02-22 11:20 - 2023-02-22 12...
Script Editor
Submit
FROM
Search for a bucket
_monitoring
_tasks
hsebucket
hsebucket_fileres
hsebucket_randomIO
hsebucket_randomIO2
+ Create Bucket
Filter
_measurement
1
Search _measurement tag va
libiotrace
libiotrace_filesystem
Filter
_field
2
Search _field tag values
file_type_descriptor
file_type_stream
function_data_read_bytes
function_data_written_byt...
function_name
hostname
process_id
return_state
return_state_detail_errno...
Filter
functionname
Search functionname tag valu
fread
fwrite
getdelim
read
write
writev
WINDOW PERIOD
Custom
Auto
auto (6s667ms)
Fill missing values
AGGREGATE FUNCTION
Custom
Auto
mean
median
last

Grafana

- Open Source Software für Visualisierung von Zeitreihen
- Daten aus Influx, SQL, etc. verarbeitbar
- Dashboards können aus vorgefertigten und selbst erstellten Plugins zusammengestellt werden

Plugins – Flux Queries

Written Bytes *(InfluxDB-LiveTracing)*

```
from(bucket: "hsebucket_fileres")
  |> range(start: v.timeRangeStart, stop: v.timeRangeStop)
  |> filter(fn: (r) => r["_measurement"] == "libiotrace")
  |> filter(fn: (r) => r["_field"] == "function_data_written_bytes")
  |> filter(fn: (r) => r["functionname"] == "fwrite" or r["functionname"] == "write" or r["functionname"] == "writev")
  |> group(columns: ["thread", "processid"])
  |> aggregateWindow(every: inf, fn: sum, createEmpty: false)
  |> yield(name: "sum")
```

Flux language syntax | Sample Query | Help

Read Bytes *(InfluxDB-LiveTracing)* *from(bucket: "hsebucket_fileres") |> range(start: v.timeRangeStart, stop: v.timeRangeStop) |> filter(fn: (r) => r["_measurement"] == "libiotrace") |> filter(fn: (r) => r["_fi...*

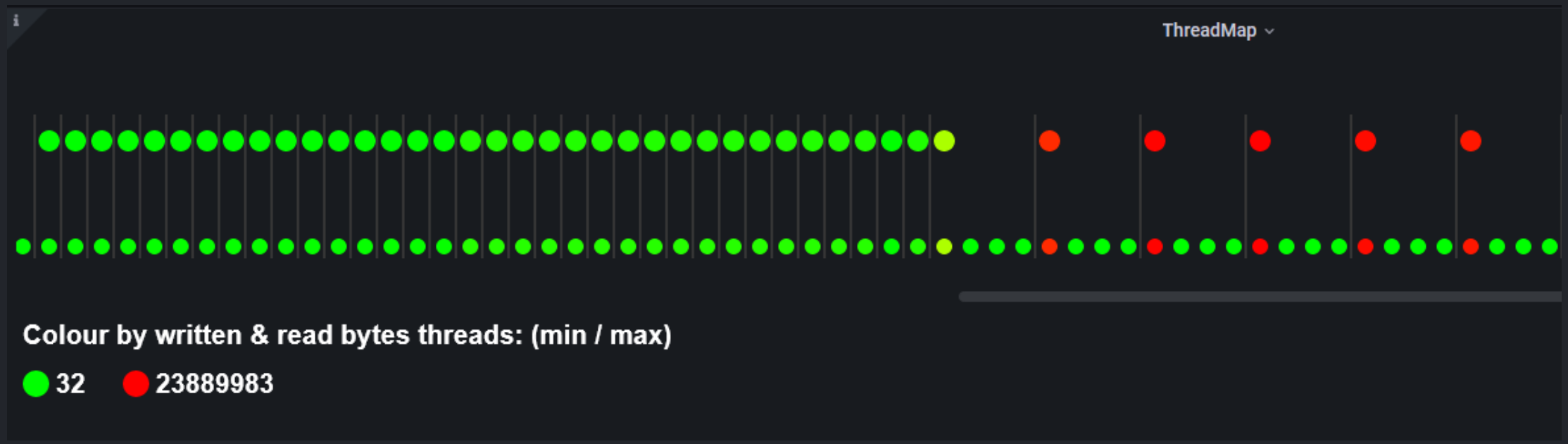
Filename Write *(InfluxDB-LiveTracing)*

```
from(bucket: "hsebucket_fileres")
  |> range(start: v.timeRangeStart, stop: v.timeRangeStop)
  |> filter(fn: (r) => r["_measurement"] == "libiotrace")
  |> filter(fn: (r) => r["_field"] == "traced_filename")
  |> filter(fn: (r) => r["functionname"] == "fwrite" or r["functionname"] == "write" or r["functionname"] == "writev")
  |> group(columns: ["thread", "processid"])
  |> aggregateWindow(every: inf, column:"_value", fn:distinct, createEmpty: false)
  |>yield(name: "distinct")
```

Flux language syntax | Sample Query | Help

Filename Read *(InfluxDB-LiveTracing)* *from(bucket: "hsebucket_fileres") |> range(start: v.timeRangeStart, stop: v.timeRangeStop) |> filter(fn: (r) => r["_measurement"] == "libiotrace") |> filter(fn: (r) => r...*

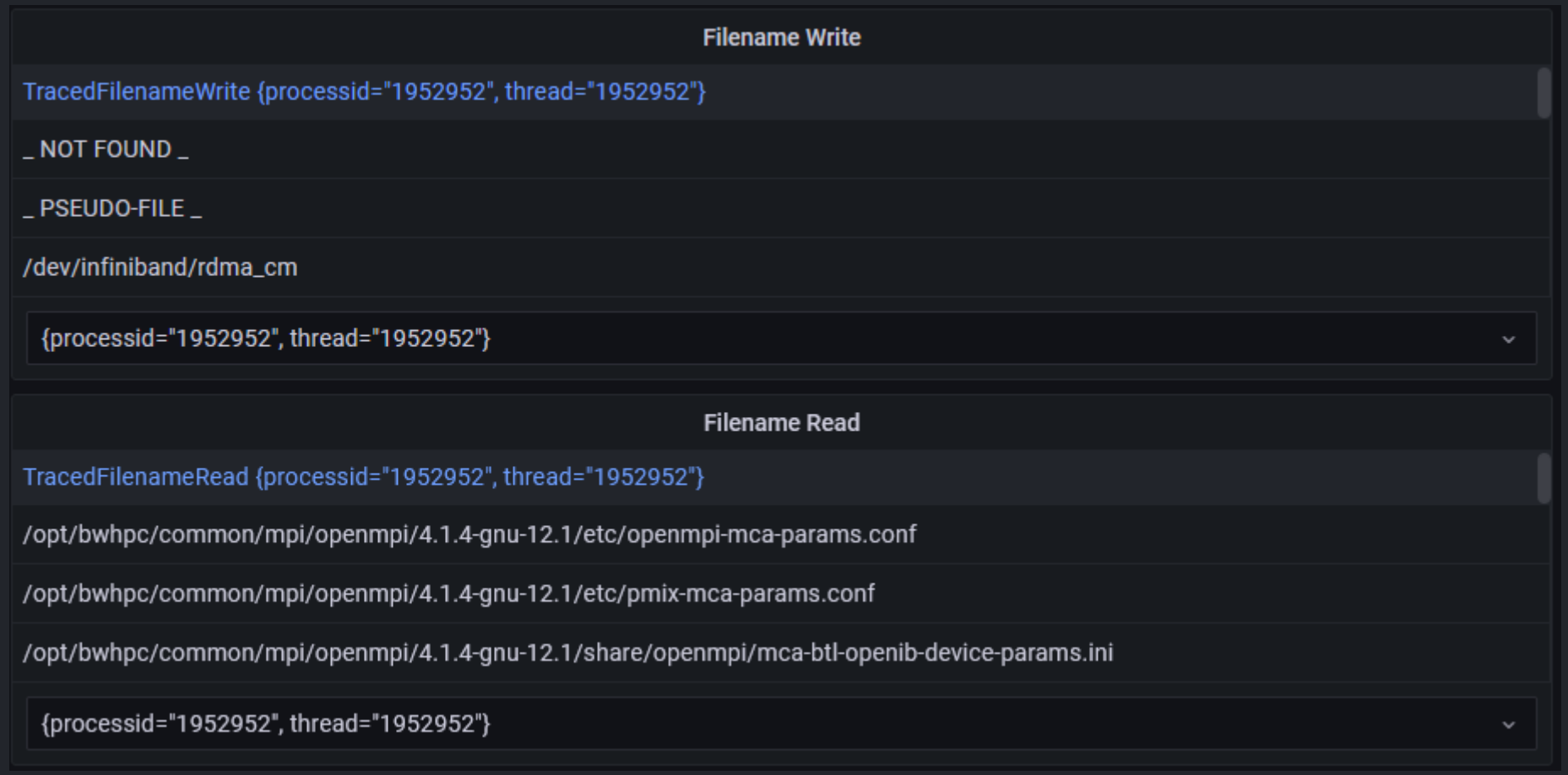
Plugins - Thread Map



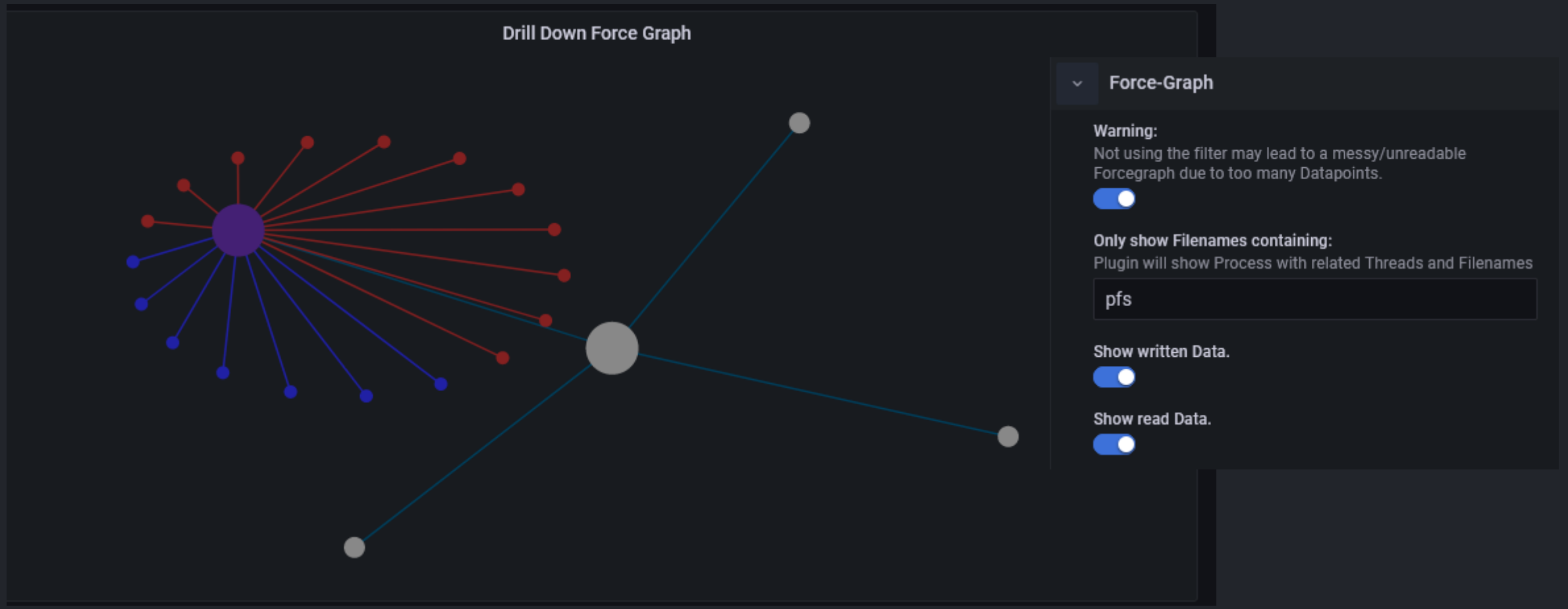
Plugins - Thread Map

Plugins - Thread Map

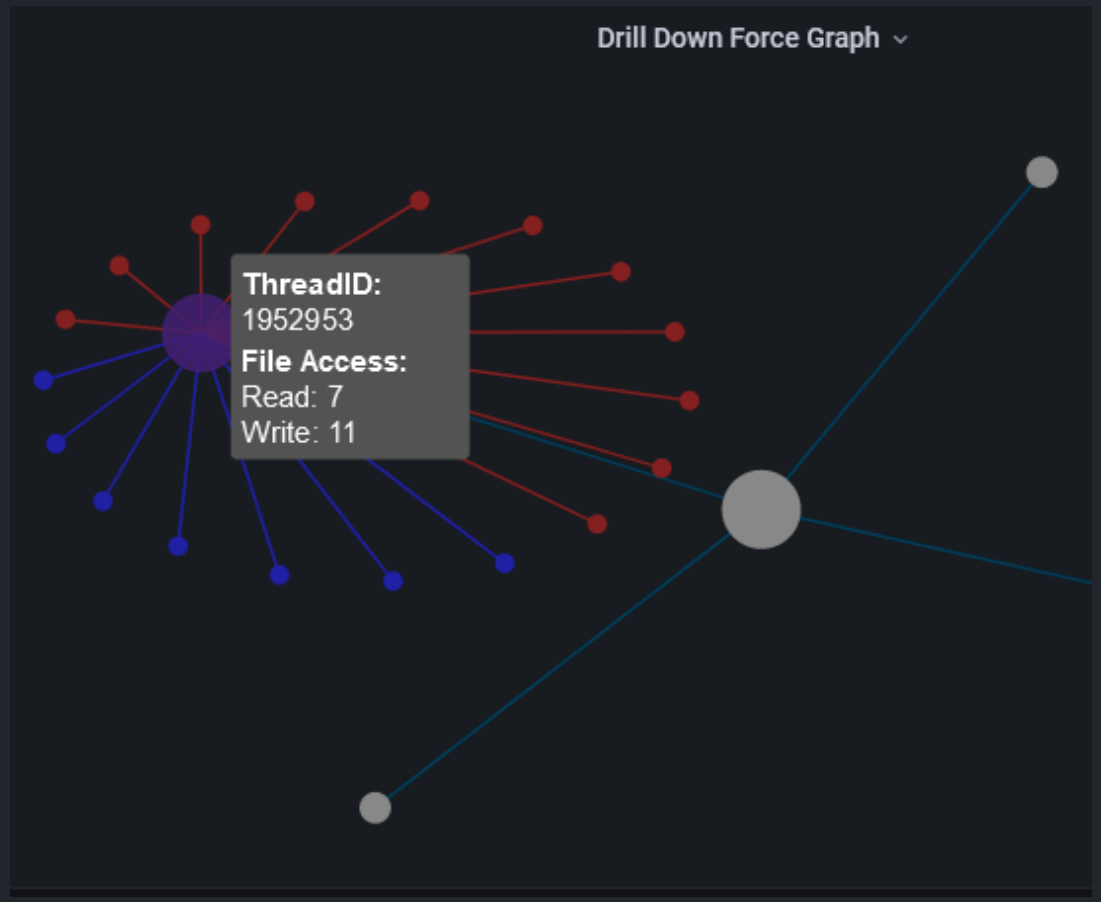
Plugins - Tables

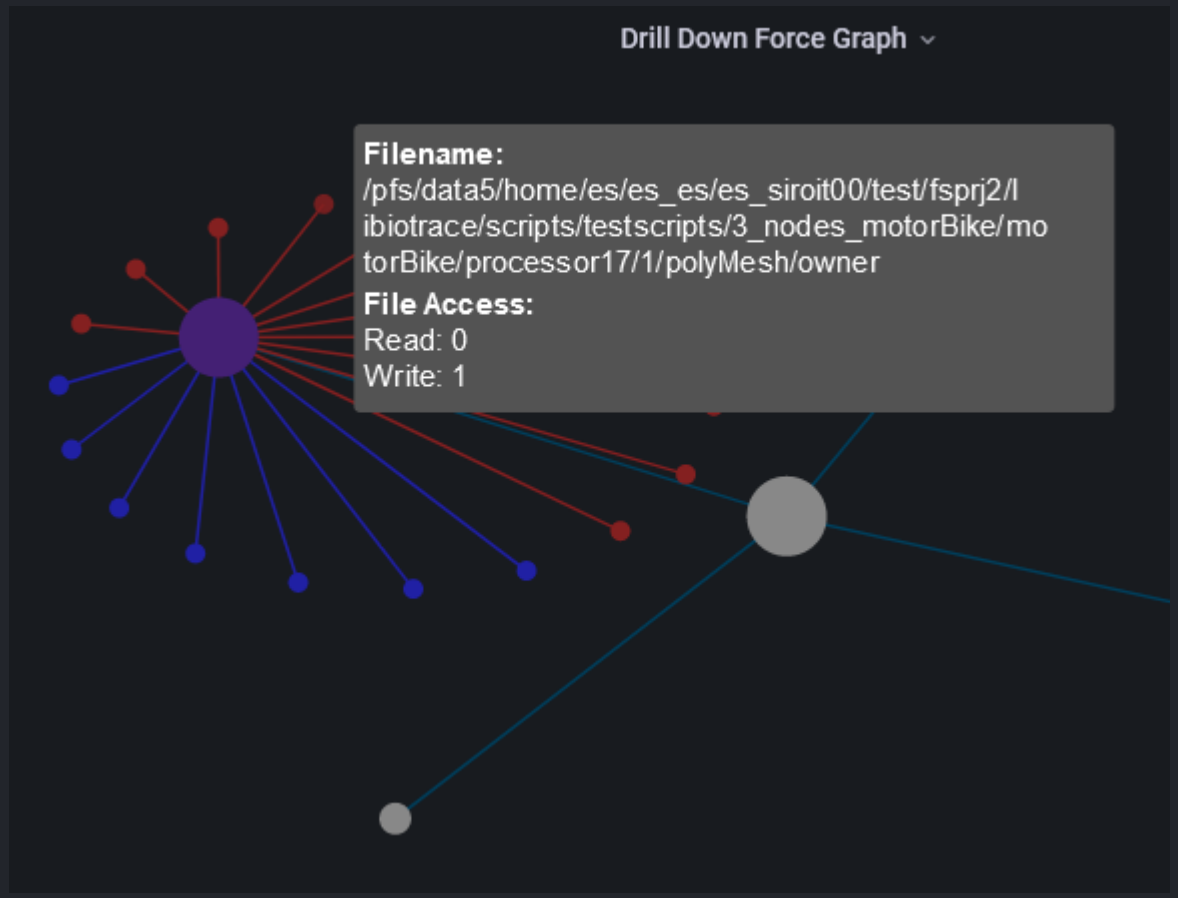


Plugins - Force Graph

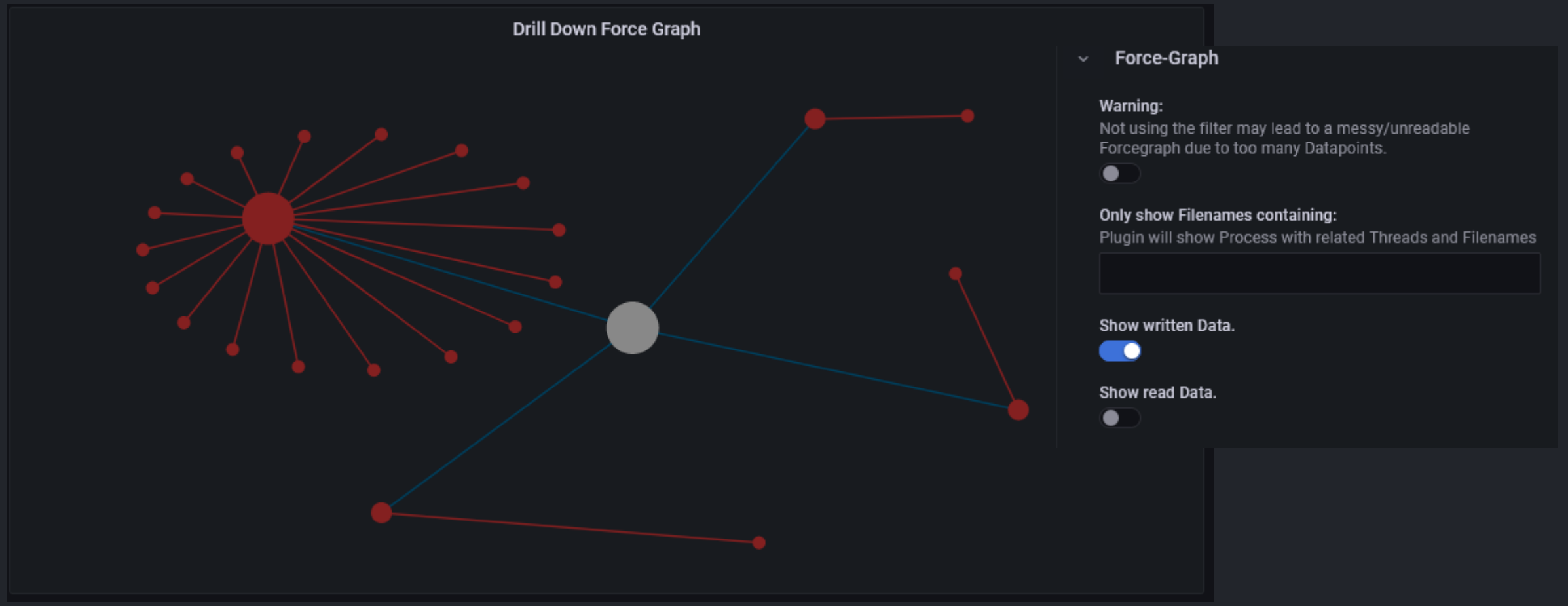


Plugins - Force Graph





Plugins - Force Graph



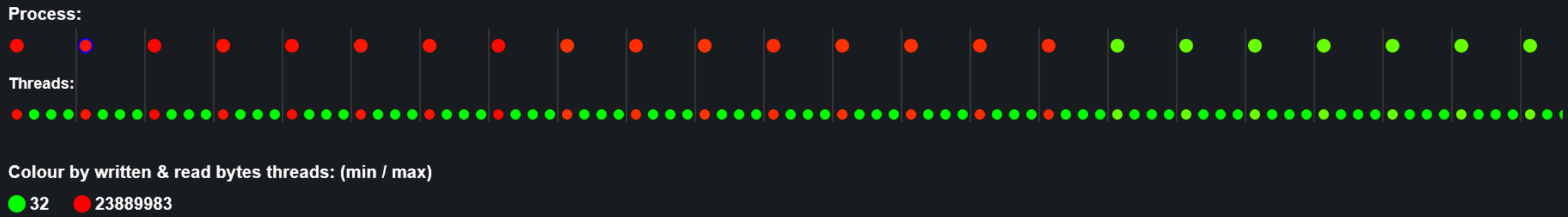
Plugins - Force Graph

Ca. 2000 Datenpunkte

Dashboard

ThreadMap & Drilldown

ThreadMap



Filename Write

TracedFilenameWrite {processid="1952952", thread="1952952"}
_ NOT FOUND _
_ PSEUDO-FILE _
/dev/infiniband/rdma_cm

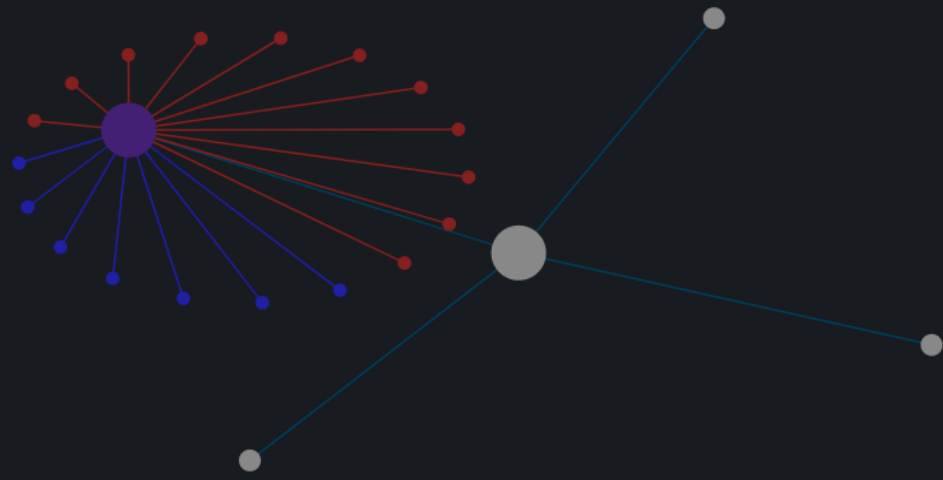
{processid="1952952", thread="1952952"}

Filename Read

TracedFilenameRead {processid="1952952", thread="1952952"}
/opt/bwhpc/common/mpi/openmpi/4.1.4-gnu-12.1/etc/openmpi-mca-params.conf
/opt/bwhpc/common/mpi/openmpi/4.1.4-gnu-12.1/etc/pmix-mca-params.conf
/opt/bwhpc/common/mpi/openmpi/4.1.4-gnu-12.1/share/openmpi/mca-btl-openib-device-params.ini

{processid="1952952", thread="1952952"}

Drill Down Force Graph



Vielen Dank für ihre Aufmerksamkeit!

Fragen?