Storage Scale

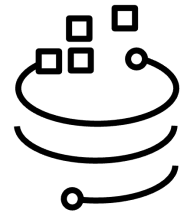
Performance Monitoring Improvements & Prometheus Exporter

IBM Storage Scale Days 2024

March 5-7, 2024 | Stuttgart Marriott Hotel Sindelfingen

Mathias Dietz

Disclaimer



IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

IBM Storage Scale Days 2024

Performance Monitoring Improvements and Prometheus Exporter

Agenda

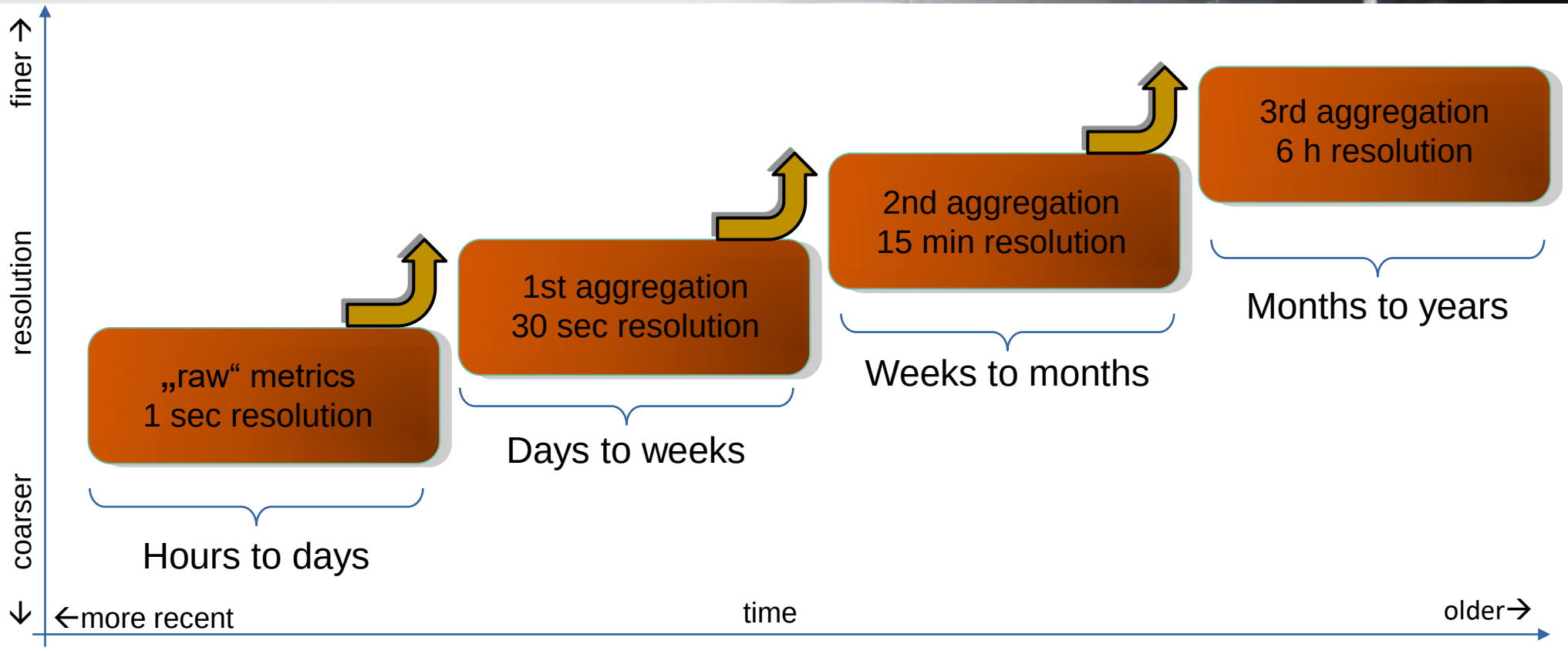
1. Performance Monitoring Overview (recap)
2. Improvements in 5.1.9/5.2.0
3. Prometheus Exporter
4. Outlook

Performance Monitoring Overview

IBM Storage Scale comes with a *powerful* performance monitoring solution. Which is essential to understand the system performance *and* to debug performance problems.

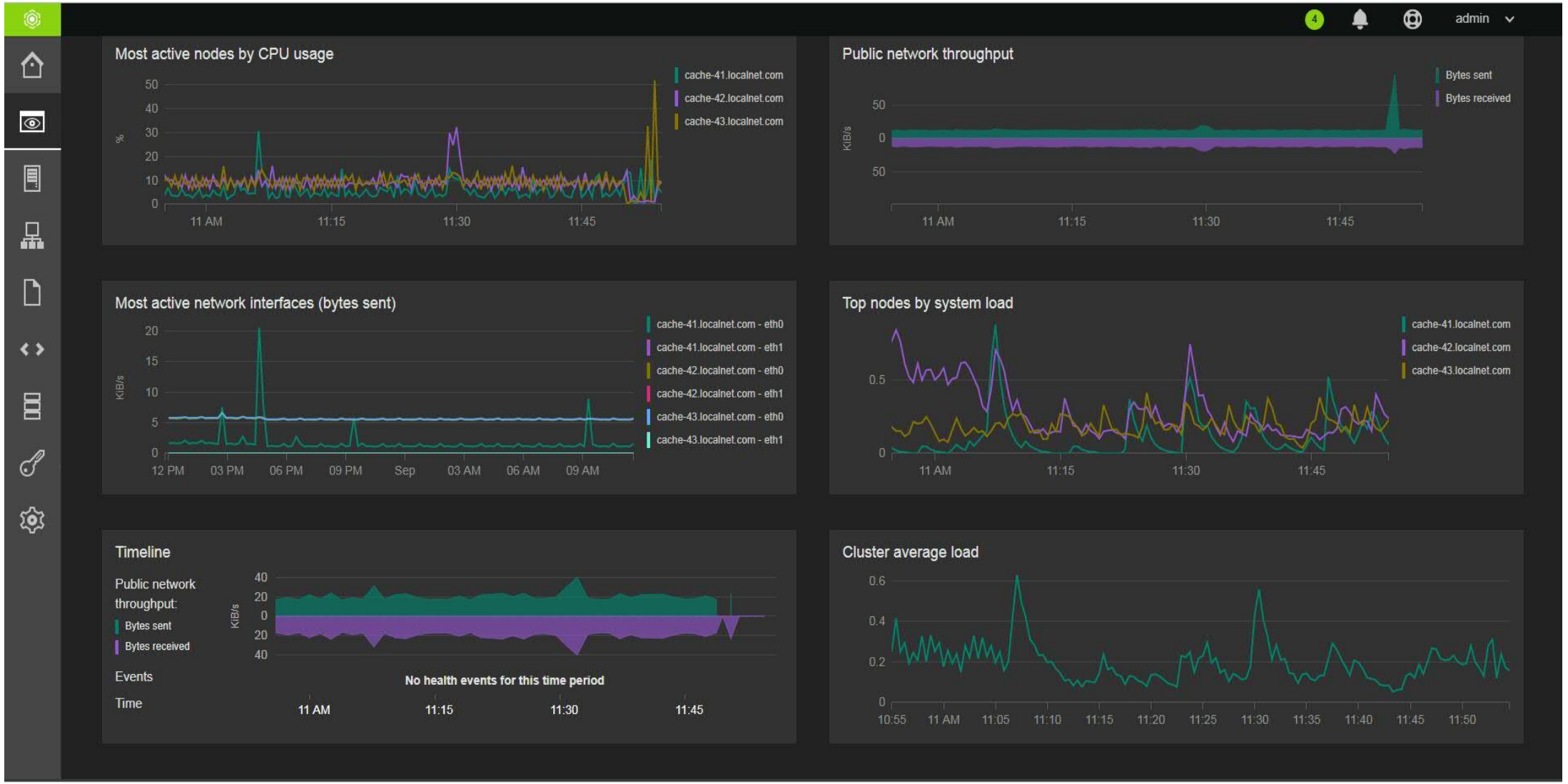
- In-Memory Timeseries Database (pmcollector) with aggregation
  - Flexible query language (perfmon query)
  - Time based aggregation to keep long term history
  - Federation for high scalability
- Performance sensors for many Scale components and subsystems
  - More than 1500 metrics from different Spectrum Scale components
  - Performance sensors for Linux, GPFS IO, AFM, GNR, SMB, NFS , etc.
  - Collect capacity/usage information
- Visualization of performance data and APIs to query data
  - Visualize performance data in Spectrum Scale GUI
  - Use REST API or cmd line interface to query raw data
  - Use Grafana Bridge to integrate performance data in your Grafana dashboard
- Alerting: Use custom thresholds to detect certain conditions and raise mmhealth events

Time-Based Aggregation



- Each storage domain is limited in the amount of memory
- Older metrics are aggregated and pushed to next aggregation level
- Eventually, metrics data is “forgotten”

Scale GUI Performance Charts



Flexible Query Language

Get metrics (mmperfmon cli, REST API):
<operation>(<metric or measurement or key>) [start end time] <bucket_size> <filters> <number of records> [--csv]

Operations: sum, avg, rate, sumrate, min, max

Metric or measurement or key: >1500 metrics available.
Measurements = calculated from metrics (e.g. metric A / metric B)
Example key: myhost-11|GPFSFilesystem|scale-cluster-1|gpfs0|gpfs_fs_bytes_written

Start|End Time|Duration|Number of samples: Range to query data from

Bucket_size: Aggregate data to the given granularity
e.g. bucket_size=100 -> automatically aggregate 10x10s values to one 100s value
Average for quantity , Sum for delta counters

Filters: Filter by any key component
e.g. cluster, node name, filesystem, fileset

Output: human readable, csv, json (rest api only)

```
#> mmperfmon query "rate(gpfs_fs_bytes_written)„ –filter „node=host-22“ -b 10 -n 3
Legend:
 1:     host-22|GPFSFilesystem|scale-cluster-12|gpfs0|gpfs_fs_bytes_written

Row          Timestamp gpfs_fs_bytes_written
  1 2024-02-15-23:43:10       0.000000
  2 2024-02-15-23:43:20       10733977.600000
  3 2024-02-15-23:43:30       114615255.000000
```

Powerful Alerting: Thresholds

Put thresholds on any performance metric and get notified by a mmhealth event.

Create new threshold rules for the specified metric or measurement through the mmhealth cli or GUI.

mmhealth thresholds add { metric[:sum|avg|min|max|rate]|measurement

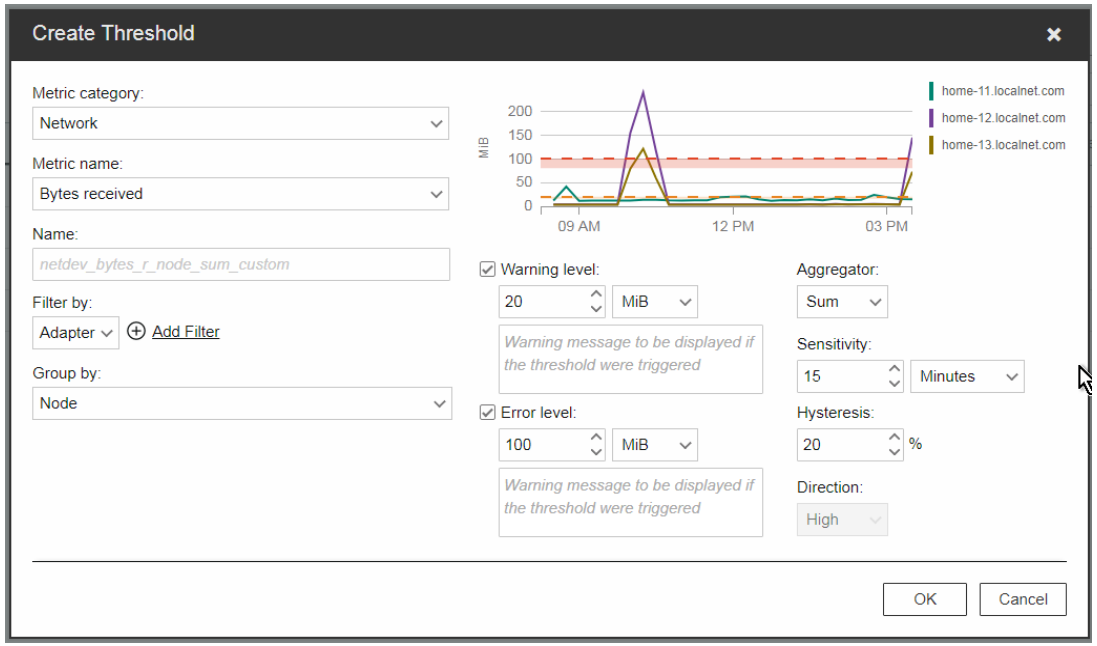
[--errorlevel{threshold error limit}
[--warnlevel {threshold warn limit}] | --direction {high|low}}

[--sensitivity {bucketsize}] [--hysteresis {percentage}]

[--filterBy] [--groupBy] [--name {ruleName}]

[--errormsg {user defined action description}]

[--warnmsg {user defined action description}]



Examples:

mmhealth thresholds add cpu_idle:avg --errorlevel 60 --direction high --name cpu_avg_bynode --groupby node'

mmhealth thresholds add MetaDataPool_capUtil --errorlevel 90 --direction high --groupby gpfs_fs_name --name myRule'

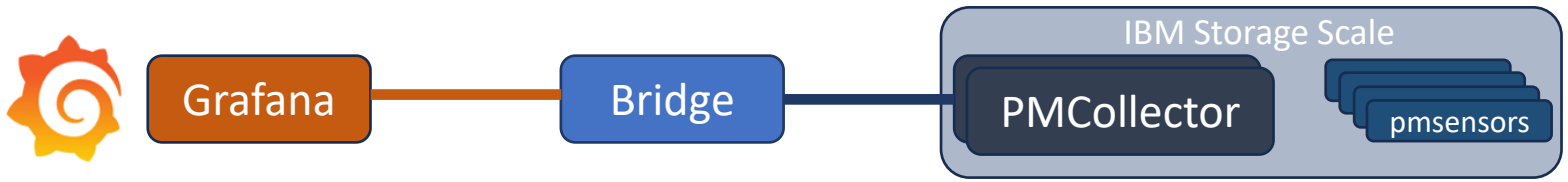
Bridge for Grafana

Grafana – Open-source performance data graphical visualizer

- Provides a powerful and elegant way to create, explore, and share dashboards and data with your team and the world.

IBM Spectrum Scale Performance Monitoring Bridge

- Makes IBM Storage Scale performance data available to Grafana
  - Uses „openTSDB“ data exchange format
- Predefined dashboards for common Scale use-cases (AFM, ESS, Waiters, etc.)



IBM Storage Scale Days 2024

Performance Monitoring Improvements and Prometheus Exporter

Agenda

1. Performance Monitoring Overview (recap)
2. Improvements in 5.1.9/5.2.0
3. Prometheus Exporter
4. Outlook

Improvements in 5.1.9/5.2.0

- Dynamic Page Pool metrics
- Remote filesystem capacity reporting
- Custom Measurements for simplified mmhealth thresholds
  - IBM support can add custom measurements
- New predefined measurements
  - fileset quota
  - vdisk usage
  - vfs latencies
- Threshold events readability. Limit number of keys displayed
- Increased threshold limit
- Pmcollector Scalability and Stability improvements
- Support for ARM platform

Hints & Tips

- **Sizing recommendations**
  - ~150 client nodes per collector
  - 3-5 Storage Building Blocks per collector
  - 16GB (up to 10 nodes) , 32GB (up to 100 Nodes), the more the better !!

- **Configure sensors and sensor period individually to match your needs**
  - Run *„mmpermon config show“* to see a list of available sensors
  - Run *„mmperfmon query --list metrics“* to see a list of available metrics (incl. disabled sensors)
    - https://www.ibm.com/docs/en/storage-scale/5.1.9?topic=tool-list-performance-metrics
    - https://www.ibm.com/docs/en/ess-p8/6.1.9?topic=gui-performance-metrics-available-in
  - Keep in mind: more sensors / lower periods cause additional system load and memory usage

- **Common field issues**
  - Large number of keys slow down pmcollector
    - Monitor number of keys using *„mmperfmon query --list expiredKeys“*
    - Delete expired keys *„mmperfmon delete“*
  - Time needs to be in sync
    - The pmcollector will drop sensor data with time stamps in the future

IBM Storage Scale Days 2024

Performance Monitoring Improvements and Prometheus Exporter

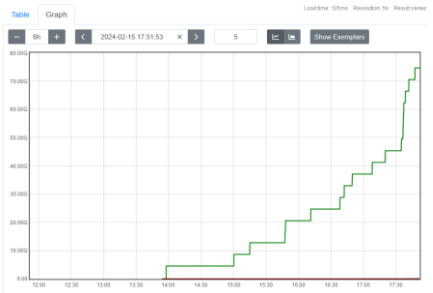
Agenda

1. Performance Monitoring Overview (recap)
2. Improvements in 5.1.9/5.2.0
3. Prometheus Exporter
4. Outlook

What is Prometheus ?

Citing prometheus.io:
From metrics to insights ! Power your metrics and alterting with the leading open-source monitoring solution.

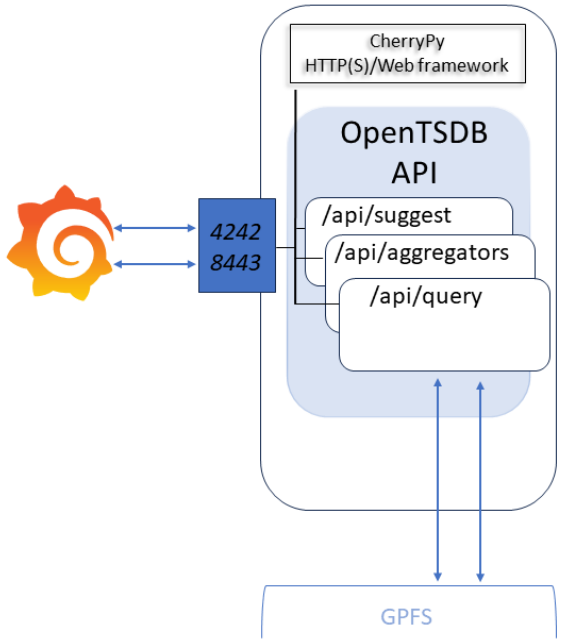
Dimensional Data
Timeseries data is identitied by metric name and key value pair
Powerful Queries
PromQL allows slicing and dicing of collected time series data
Great Visualization
Offers multiple modes to visualize data + Grafana integration
Efficient Storage
Stores time series in memory and on local disk in an efficient format. Federation and sharding.
Precise alerting
Alerts based on flexible PromQL queries. Altermanager handles notifications and silencing



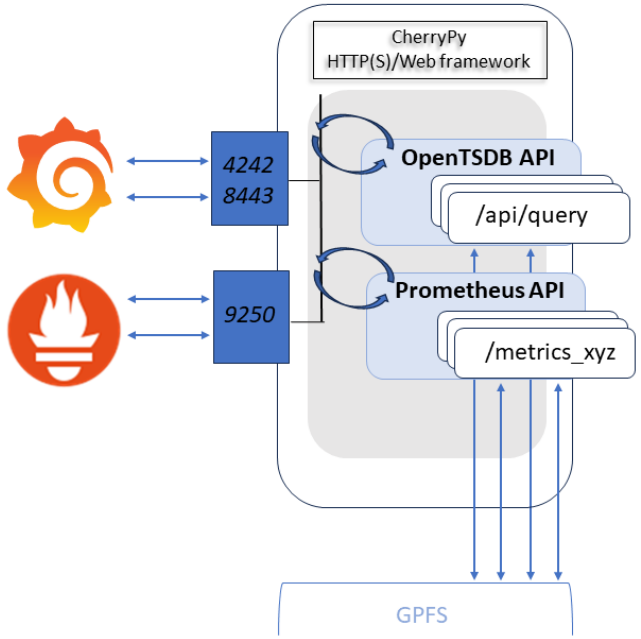
Prometheus Exporter for Scale

- Prometheus exporter for Scale is a software component that exposes Scale performance metrics in a format that Prometheus timeseries database can scrape.
- IBM Storage Scale customers frequently request the capability to store and analyze performance metrics from GPFS devices in the Prometheus database.
- Prometheus Exporter plugin have been integrated in the IBM Storage Scale bridge for Grafana version 8.0.0

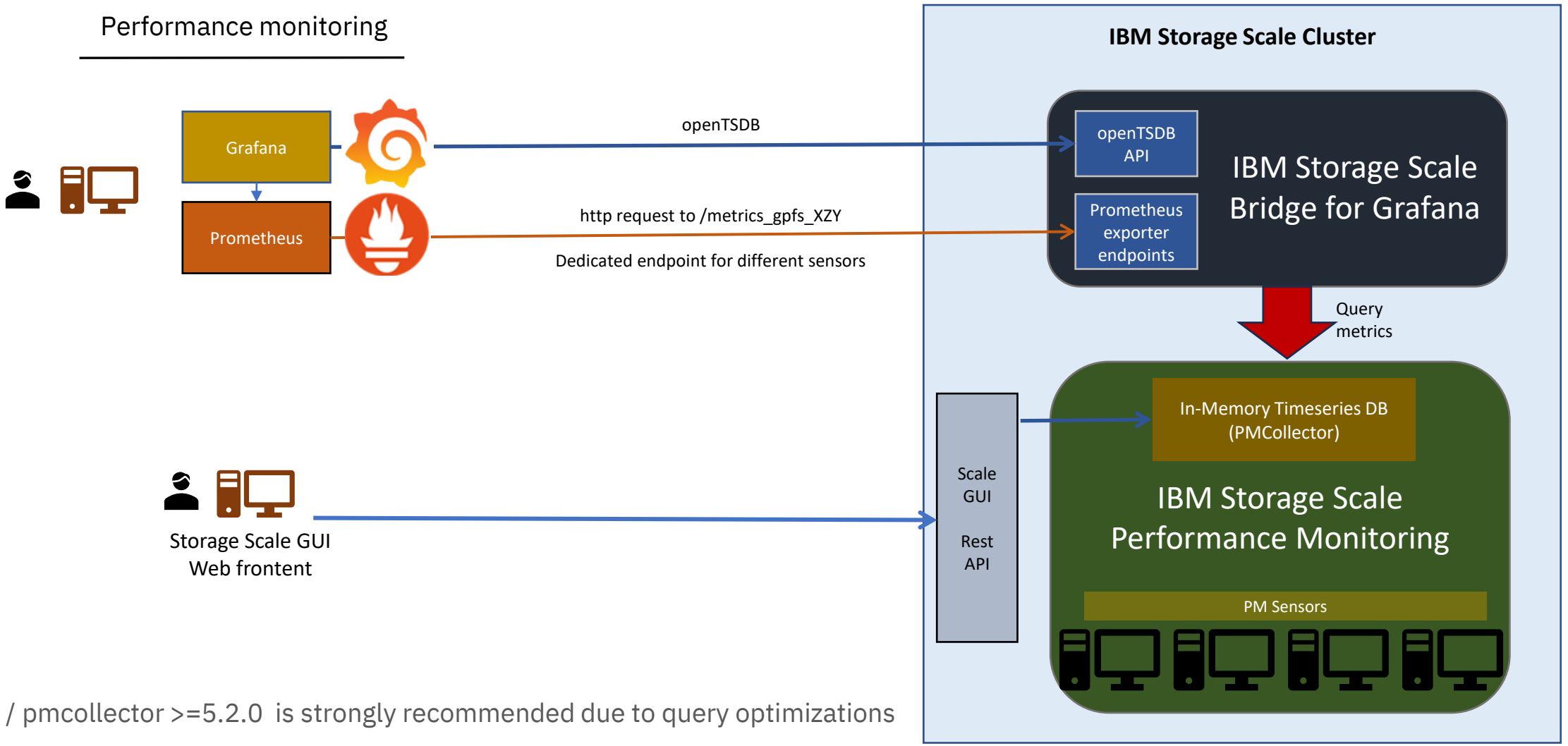
Grafana-bridge v7.x



Grafana-bridge v8.x



Prometheus Exporter for Scale



Scale / pmcollector >=5.2.0 is strongly recommended due to query optimizations

Prometheus Exporter Key Design Points

Prometheus Exporter for Scale is a frontend to pmcollector

- Data resides on pmcollector
  - Long term history of aggregated data (pmcollector keeps up to 1 year)
  - Performance charts in the GUI
  - Used for service & support (gpfs.snap, callhome)
  - mmhealth thresholds
  - Used by ESS Hospital
- Data is queried from pmcollector on demand
  - Data is queried in raw format to not have any aggregation/sampling errors
- Prometheus keeps a copy of the data (doubled)
  - Once scraped the data in prometheus is independently managed

Prometheus Exporter Setup

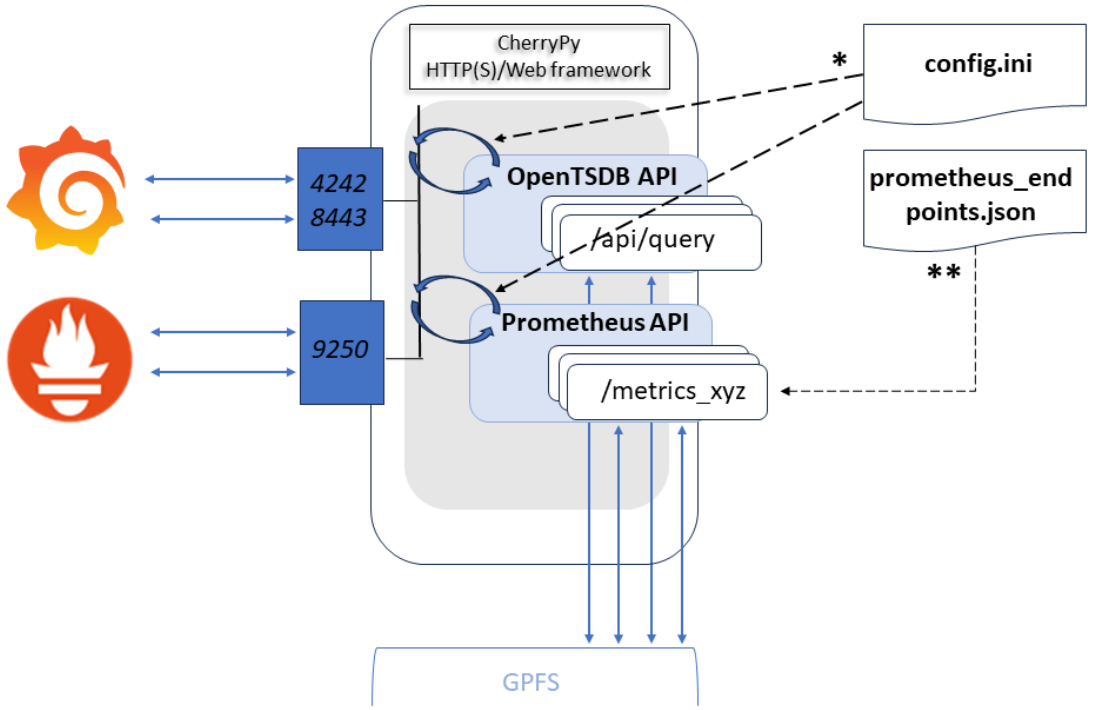
Storage Scale Bridge for Grafana has a new option to enable Prometheus Exporter endpoints.
A dedicated TCP communication port must be configured:

- Either through command line option *-e <exporter port>, or*
- Specified (commented in) in the config.ini file

```
###################### Prometheus Exporter API Connection Defaults ######
# Port number the bridge listening on for Prometheus server https requests;
# ssl cert and key configuration required
# prometheus = 9250
```

The Prometheus endpoints are managed via a JSON file and dynamically registered at startup if the Prometheus API is enabled.

```
{
   "/metrics_gpfs_filesystem": "GPFSFilesystem",
   "/metrics_gpfs_nsddisk": "GPFSNSDDisk",
   "/metrics_gpfs_fileset": "GPFSFileset",
   "/metrics_gpfs_diskcap": "GPFSDiskCap",
..
}
```



For detailed instructions, please refer to the Grafana-bridge wiki page:
Setup the IBM Spectrum Scale Performance Monitoring Bridge for classic IBM Spectrum Scale devices.

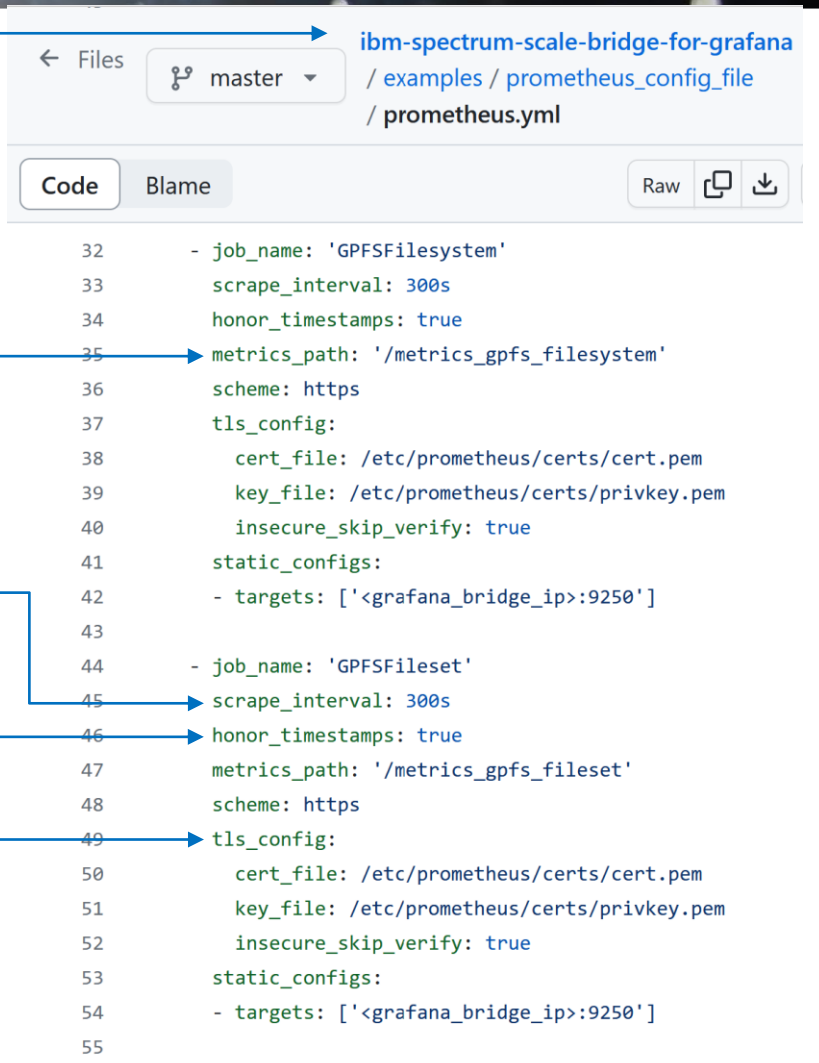
Prometheus Scraper Configuration

Prometheus server setup

You need to add an individual scrape job to the Prometheus configuration file (prometheus.yaml) for each IBM Storage Scale performance data collection sensor that you want to export to Prometheus

- Only sensors listed in the grafana-bridge prometheus-endpoints.json file are supported
- The scrape job metrics_path must match the endpoint name specified in the grafana-bridge prometheus-endpoints.json file
- The scrape job interval must also match the sensor period configured by the IBM Storage Scale performance monitoring tool
- Honor_timestamps must be enabled to ensure that Prometheus is storing the original timestamp of the metric
- The SSL key and certificate configured for the Prometheus API within the IBM Storage Scale bridge for Grafana must be accessible to the Prometheus server

An example of the prometheus.yaml file can be found at https://github.com/IBM/ibm-spectrum-scale-bridge-for-grafana/tree/master/examples/prometheus_config_file



IBM Storage Scale Days 2024

Performance Monitoring Improvements and Prometheus Exporter

Agenda

1. Performance Monitoring Overview (recap)
2. Improvements in 5.1.9/5.2.0
3. Prometheus Exporter
4. Outlook

Outlook

Outlook

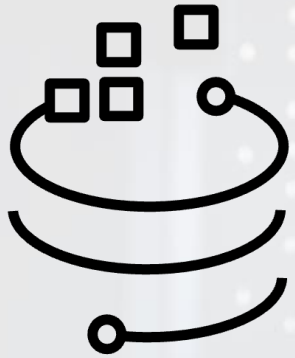
- Faster identification of ESS3500/6000 performance issues (Use case driven)
- Export „Health“ data to pmcollector & Prometheus
  - Data can be scraped by Prometheus the same way as the performance metrics
  - Implement pmsensor to collect mmhealth data -> Convert status into numerical representation (e.g. HEALTHY = 1)
  - Keep long term history of health status -> For use by GUI, Grafana dashboards and IBM service
- Fileset and job level IO statistics (based on Qos)
- ESS hardware metrics (temperatures, power, etc.)
- Pmcollector Scalability improvements
- Automatic generation of prometheus.yaml file
- IBM Storage Scale bridge for Grafana reporting on its own performance
- Prometheus Exporter API support on CNSA

Your feedback and ideas are welcome !

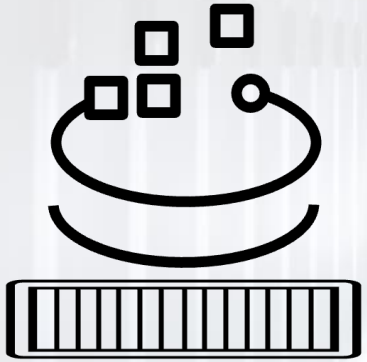
To share your thoughts with the development team, open a new issue in the grafana-bridge repository and describe the topic you are interested in.

https://github.com/IBM/ibm-spectrum-scale-bridge-for-grafana/issues

Thank you for using



Storage Scale



Storage Scale System

Mathias Dietz

Mail: mdietz@de.ibm.com

LinkedIn: www.linkedin.com/in/mathias-dietz-42465892