

Performance and Service Monitoring for GPFS

Stefan Dietrich, Martin Gasthuber, Jürgen Hannappel
Hamburg, 2022-07-14

Service Monitoring: Icinga




- Central Icinga 2 Installation
 - Provided by IT-Operating group
 - Monitoring for all hosts in datacenter
 - Handles hardware- and service monitoring
 - Alerting via mail or pager for on-call shift
- Hardware issues handled by Operating staff
 - Service issues partially handled by Operating
- Standard RHEL 7/8 on ESS/DSS-G
 - Majority of already available Linux checks for free
 - IBM POWER slightly cumbersome

A screenshot of the Icinga 2 web interface. The top navigation bar includes tabs for 'Host', 'Services', and 'History'. The main content area shows details for host 'nsd-gl20' (nsd-gl20.desy.de), which is 'UP' since Jan 10. It lists 17 services with status indicators (1 orange, 1 purple, 15 green). Below this, there's a 'Plugin Output' section showing 'OK - 131.169.251.72: rta 0.064ms, lost 0%'. A 'Problem handling' section includes links for 'Add comment', 'Schedule downtime', and a note with a link to a Confluence page. An 'Actions' section lists links like 'Business Impact', 'Elasticsearch Events', 'Show Host Documentation', 'Show Host Metrics In Grafana (IT)', and 'Show logs in Kibana'. At the bottom, a 'Performance data' table is displayed.

	Label	Value	Warning	Critical
●	rta	64.00 µs	600.00 ms	900.00 ms
	pl	0%	60%	80%
	rtmax	85.00 µs	-	-
	rtmin	58.00 µs	-	-

GPFS Service Checks

- Several GPFS specific service checks
 - Written in Python 3, partially with nagiosplugin <https://github.com/mpounsett/nagiosplugin>
 - Checks for pool & inode utilization, cluster export services (CES), deadlock, GPFS native RAID etc.
- Checks call mm* commands, parse output
 - Check script running on GPFS node
 - Icinga triggers check through remote agent
- Performance output of checks available in Grafana
- Checks currently not public, could be shared with KIT

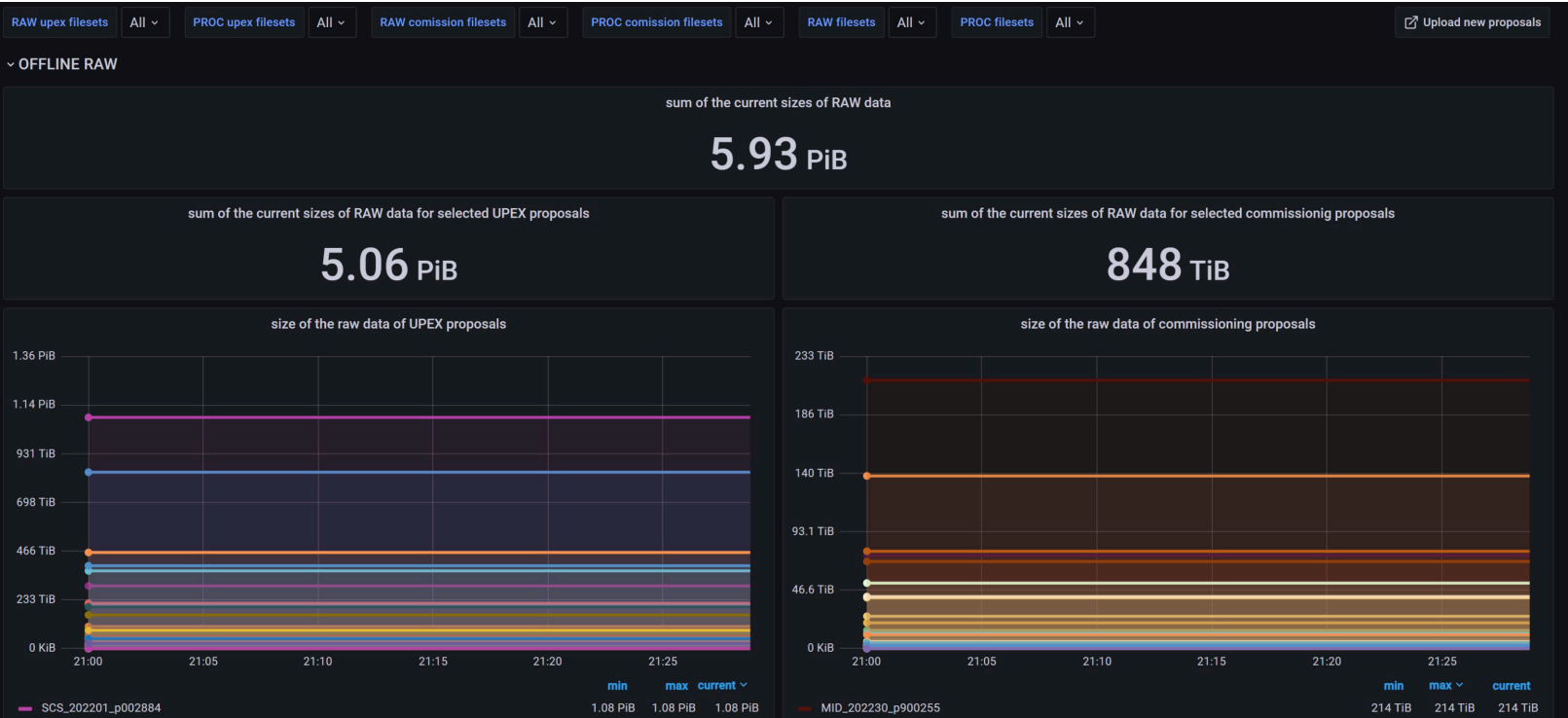
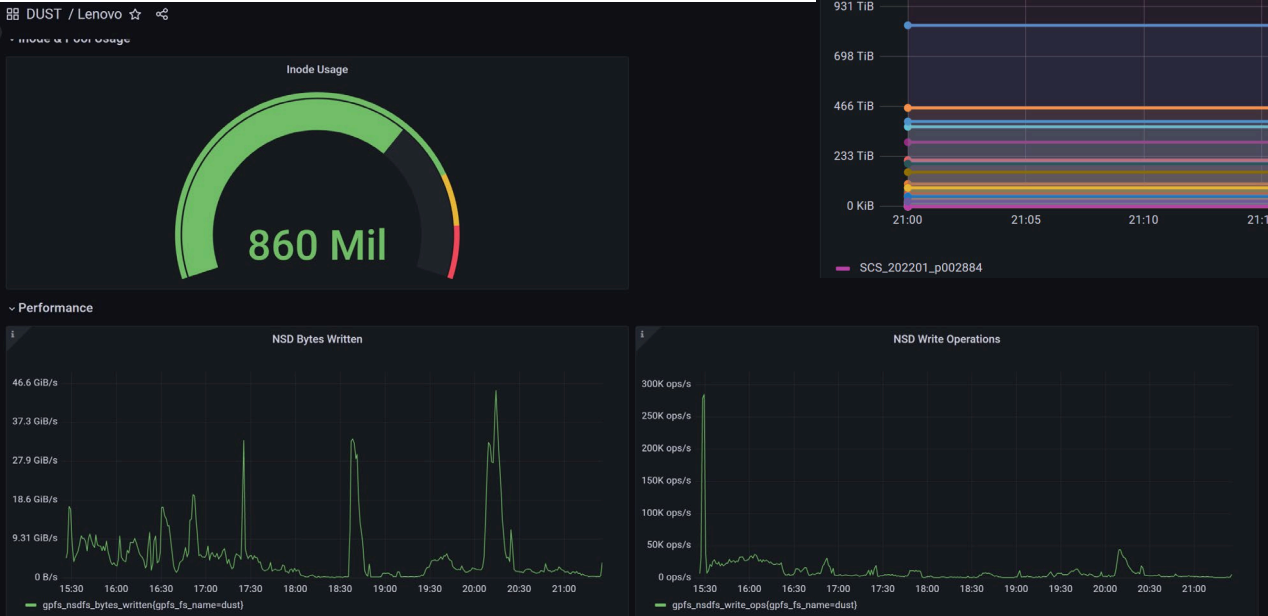
OK	GPFS cache1 inode Usage
Feb 18	CHECK_GPFS_INODES OK - cache1_quotaInDoubt is 0.01%
OK	GPFS cache1 Usage
May 17	FILESPACE OK - Usage for filesystem "cache1" below threshold
OK	GPFS core1 inode Usage
Jun 26	CHECK_GPFS_INODES OK - core1_quotaInDoubt is 0.05%
CRITICAL	GPFS core1 Usage
Jul 1	FILESPACE CRITICAL - core1: data pool usage is 92% (outside range 0:89)
OK	 GPFS Deadlock
May 18	OK - No deadlock detected

Performance Monitoring



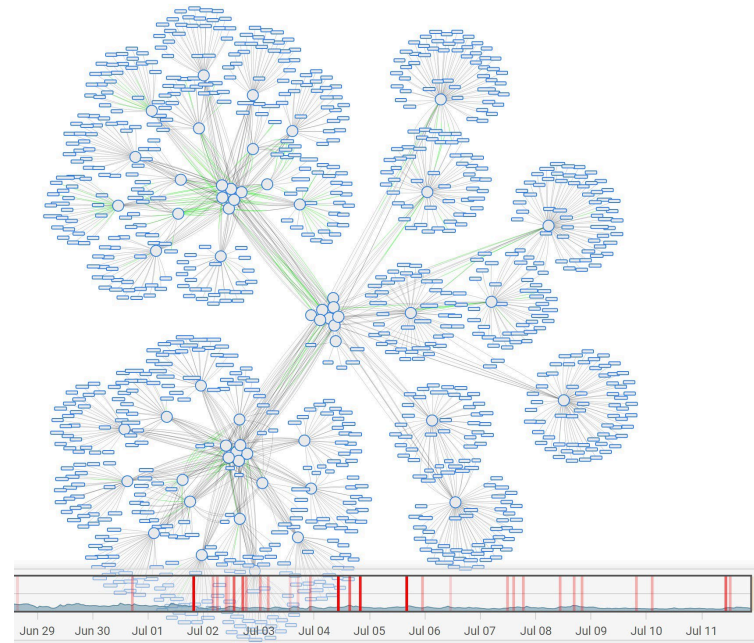
- Several metrics from different data sources
 - Basic Linux metrics
 - > Icinga Service Checks
 - GPFS specific metrics
 - > Spectrum Scale Performance Monitoring Tool aka ZIMon
 - Custom metrics, e.g. Open FDs of Ganesha
 - > Script pushing metrics to Carbon/Graphite
 - InfiniBand (ibqueryerrors)
 - > InfiniBand Radar and Prometheus
 - Visualization of metrics with Grafana
 - ...Alerting for important metrics not handled by Icinga
 - Several dashboards, primary for admins
- Grafana Bridge for ZIMon
<https://github.com/IBM/ibm-spectrum-scale-bridge-for-grafana>
 - Grafana for visualization instead of GPFS GUI
 - Works mediocre, e.g. wrong query can easily block bridge or ZIMon collector
 - IBM provides periodically updates for bridge
 - Created script to dump metrics from ZIMon and push to InfluxDB
 - > currently not in use

Example Grafana Dashboards



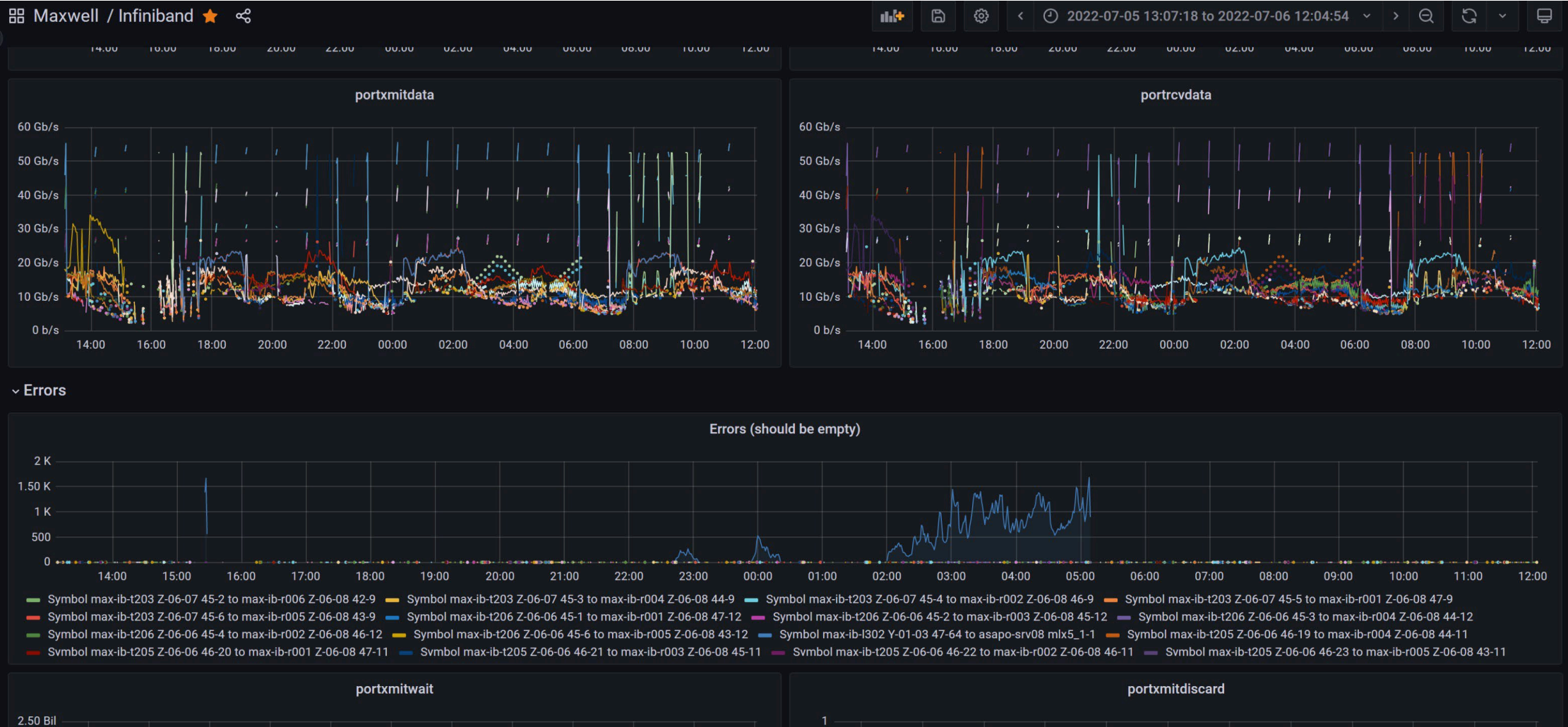
InfiniBand Monitoring

- InfiniBand Radar
 - <https://github.com/infiniband-radar/infiniband-radar-web>
 - Developed by former Bachelor Student
- InfiniBand Prometheus Exporter
 - <https://github.com/guilbaults/infiniband-exporter>
- InfiniBand has lots and lots of metrics...
 - ...documentation on metrics is sparse
 - Mellanox: „Just buy UFM“
 - Still trying to identify the key metrics for InfiniBand, e.g. congestion/overload or dropped packets



Fabric List			Last 5 minutes
Search	Connection Sort		
Hostname, GUID, SM State			
Fold all Hosts: 1031			
+ asap3-bl-prx01			
+ asap3-bl-prx02			
+ asap3-bl-prx03			
+ asap3-bl-prx04			
+ asap3-bl-prx05			
+ asap3-bl-prx06			
+ asap3-bl-prx07			
+ asap3-bl-prx08			
+ asap3-bl-prx09			
+ asap3-bl-prx20			
+ asap3-bl-prx21			
+ asap3-bl-prx22			
+ asap3-bl-prx23			
+ asap3-bl-prx24			
+ asap3-bl-prx25			
+ asap3-ems20-dep			
+ asap3-prx01			
+ asap3-prx02			
+ asap3-utl01			
+ asap3-utl02			
+ asap3-utl03			

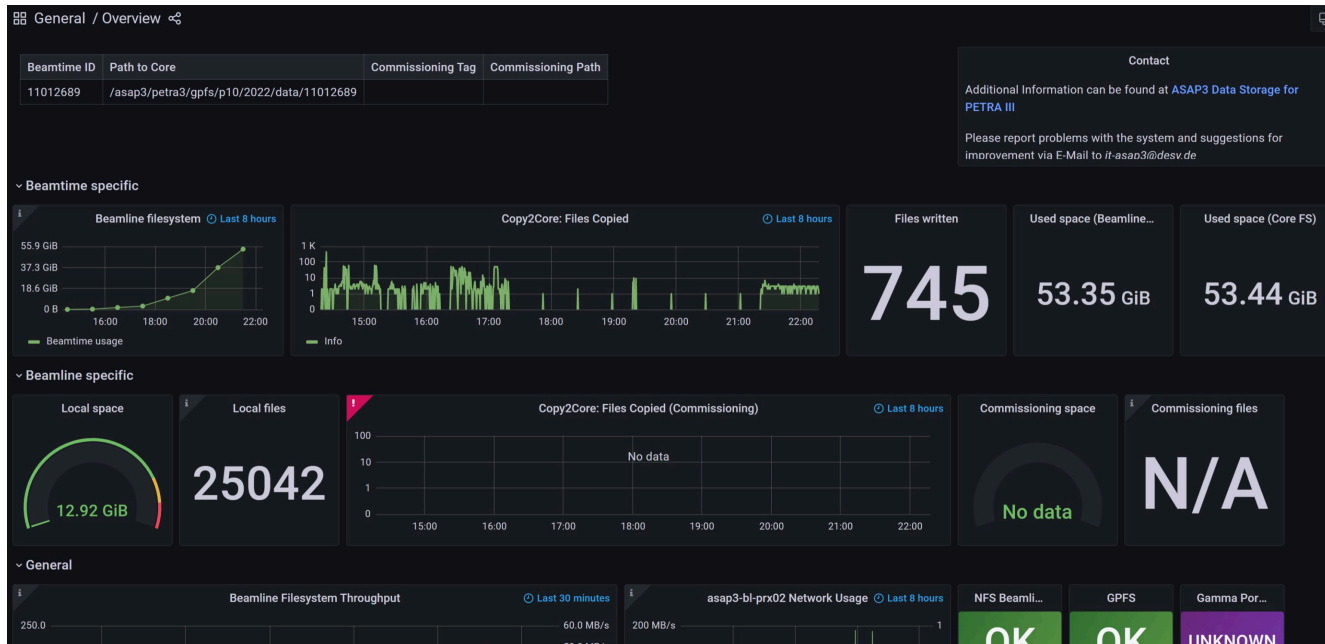
InfiniBand Prometheus Exporter



Monitoring for end-users

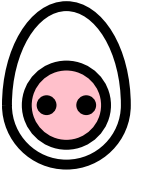
...won't somebody please think of the users!

- Specific to ASAP3 storage system
- Admins have all of the important metrics and service states, but end-users?
- Provide simple view to the beamline
 - Only important metrics, aggregate service states
- Dashboard uses ZIMon Bridge for GPFS metrics
- InfluxDB for additional data
 - Aggregate service states from Icinga (traffic light)
 - Text data, e.g. active beamtime
- Elasticsearch for log entries



Copy Tools - ewmscp

<https://gitlab.desy.de/ewmscp>



- ewmscp: Fast and versatile copy program
- Written in C++, partially with GPFS support
 - Support for multiple threads
 - Batteries included, like chown or setting XATTR
 - fixGpfsAcls: Recursively change NFS4 ACLs!
- ASAP3: Used with inotify to copy data beamline -> core filesystem

Thank you