Building a 21st century monitoring infrastructure

Migration of the monitoring infrastructure to Prometheus & ELK at DESY, Zeuthen

Andreas Haupt, Philipp Bolle
Fall HEPiX 2019, Amsterdam

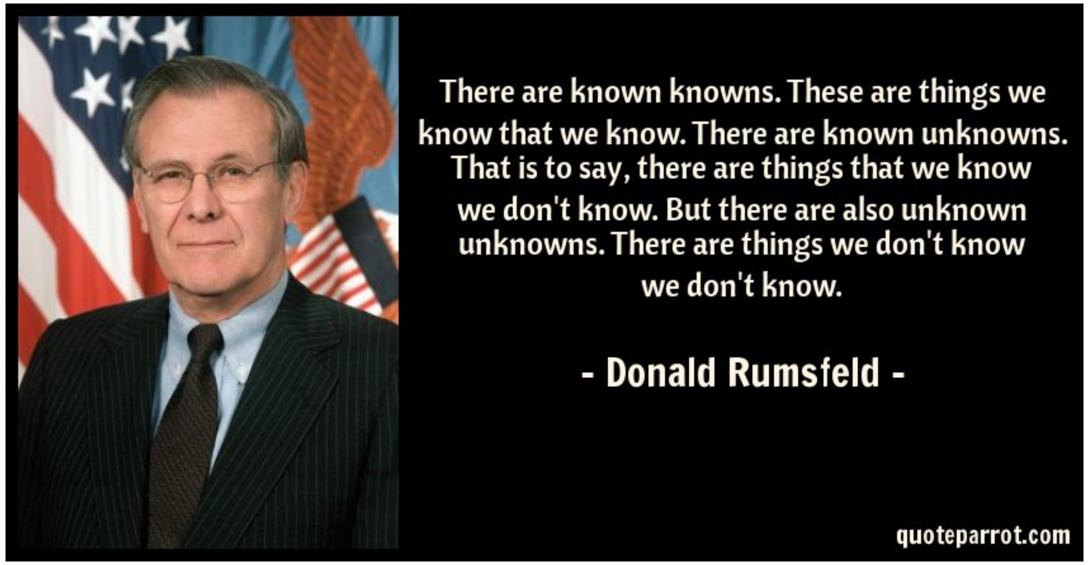
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Old monitoring landscape

in the year 2017

- Icinga – service monitoring and alarming
  - Hosts apply for being monitored by sending a list of services
    - a concept we wanted to keep
  - many self-written checks in place
- Ganglia – cluster visualisation
- Home-made computing centre overview (Comon)
- Central syslog hosts running LogSurfer (… yes!)
  - with alarming on PCRE patterns

Why change?



There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

- Donald Rumsfeld -

quoteparrot.com

The Rumsfeld Matrix

	Knowns	Unknowns
Known	**Known Knowns**	**Known Unknowns**
Unknown	**Unknown Knowns**	**Unknown Unknowns**

Source: http://www.lean-agility.de/2017/07/die-rumsfeld-matrix.html

Why change?

Getting away from monitoring the "known knowns" only ...

- Old infrastructure was aging and had some design deficites
  - Some components completely unmaintained
  - Linux-only
  - Bringing it to a current state would have resulted in a redesign anyway
- Some (independed) dashboards available but no general overview in place
  - "One solution per problem"
- Major reason: new colleagues usually do not want to maintain old, complex, grown systems ;-)
  - Starting from scratch is a good chance to rethink all decisions again

Design decisions

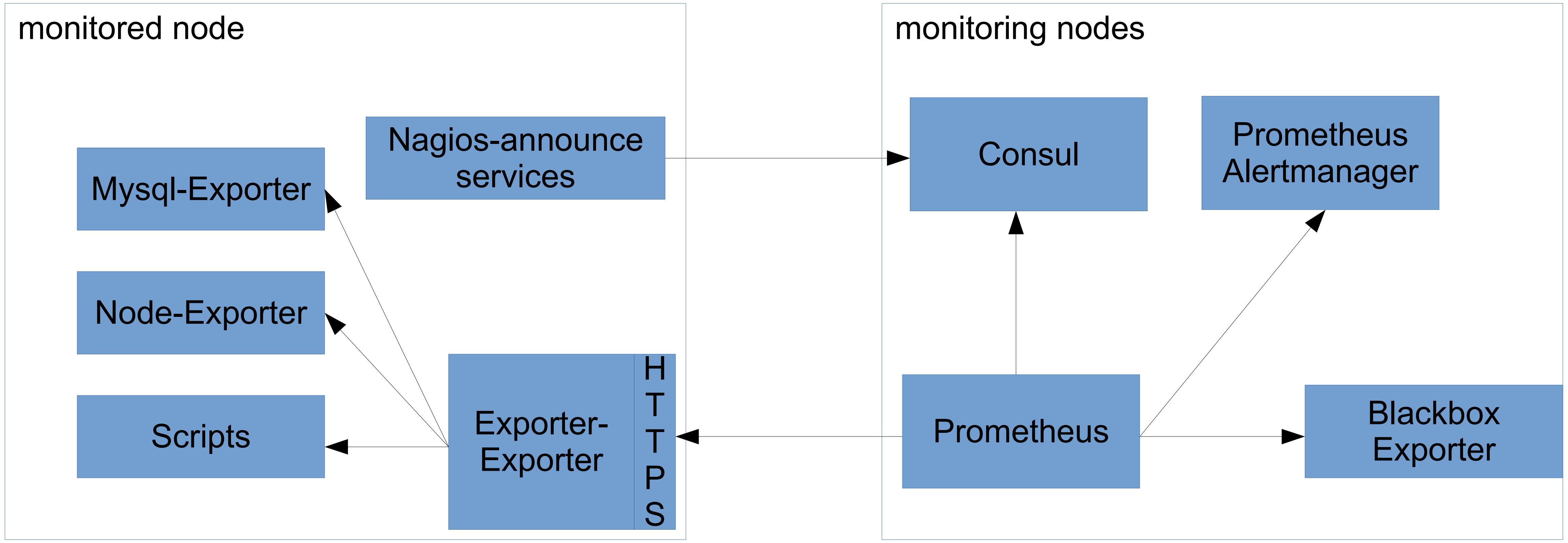
or: products to be used

- Distinguish between metric-based and event-based monitoring pipelines
  - Metrics: Prometheus
    - Dashboard: Grafana
  - Events: ELK
    - Dashboard: Kibana
- Hashicorp Consul acts as service registry
- Evaluate as common, platform-independent monitoring infrastructure
  - Monitor Windows servers, too …
  - … and later maybe even network devices?

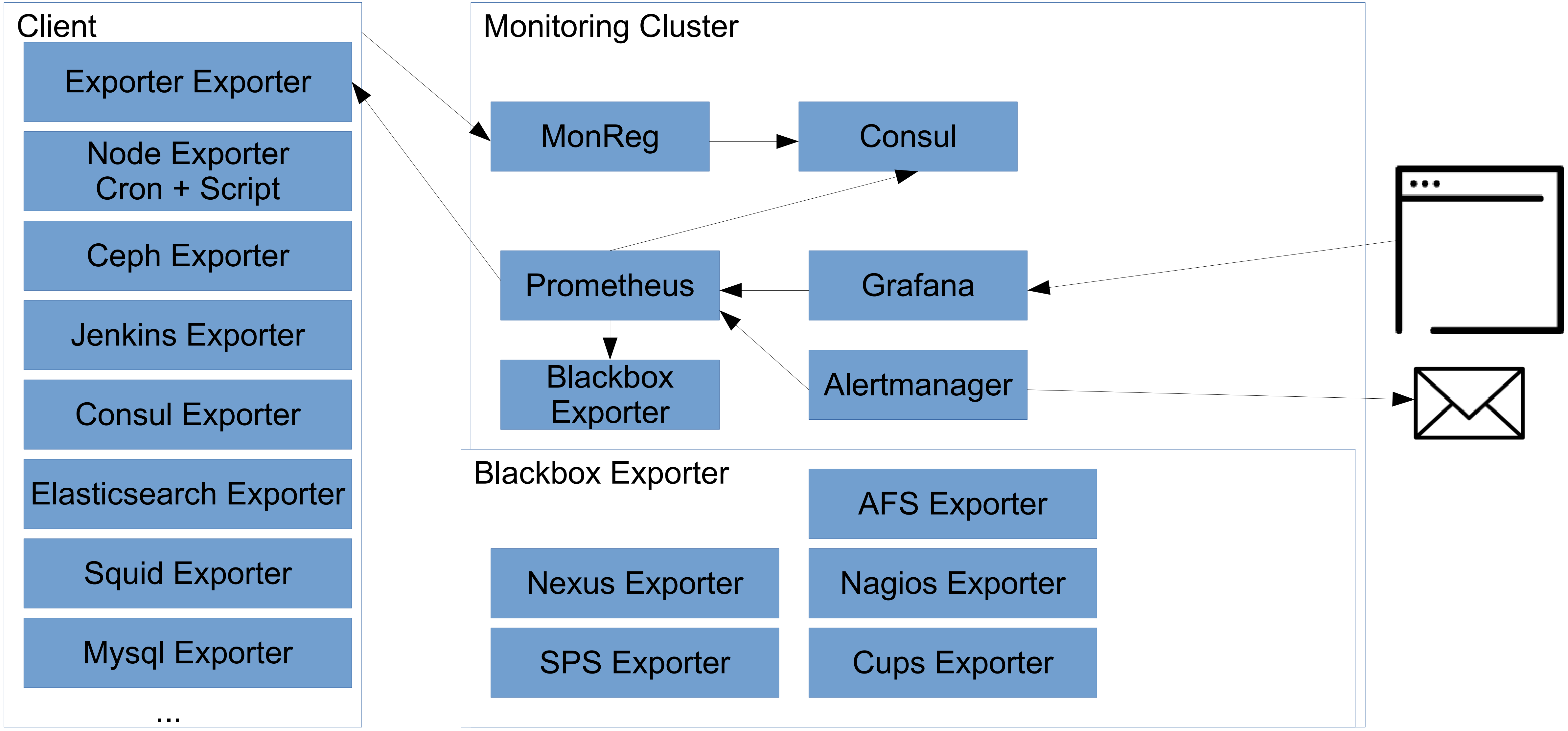
Implementation

First milestone

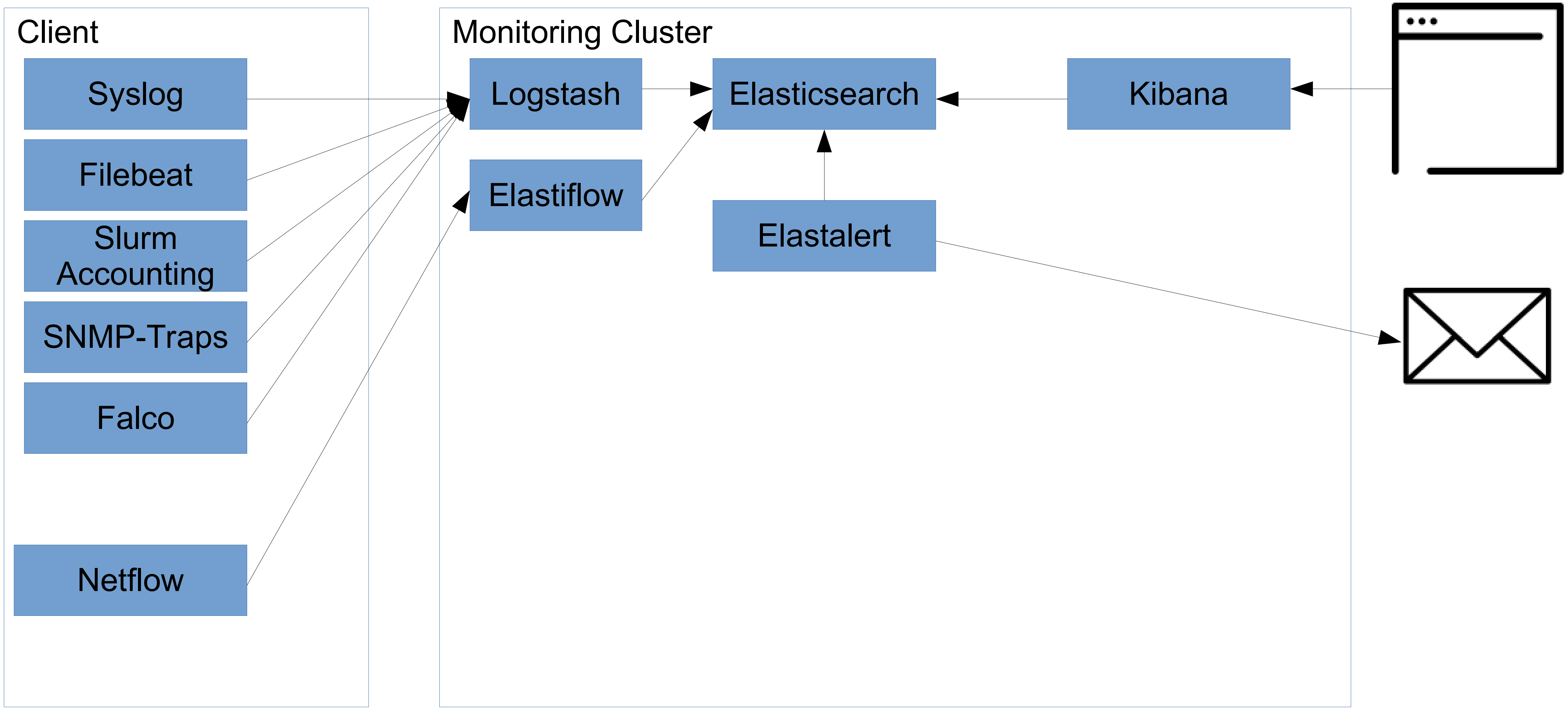
Minimal-Setup



Metric


Client
Exporter Exporter
Node Exporter Cron + Script
Ceph Exporter
Jenkins Exporter
Consul Exporter
Elasticsearch Exporter
Squid Exporter
Mysql Exporter
...
Monitoring Cluster
MonReg
Consul
Prometheus
Grafana
Blackbox Exporter
Alertmanager
Blackbox Exporter
AFS Exporter
Nexus Exporter
Nagios Exporter
SPS Exporter
Cups Exporter

Event


Client
Syslog
Filebeat
Slurm Accounting
SNMP-Traps
Falco
Netflow
Monitoring Cluster
Logstash
Elastiflow
Elasticsearch
Elastalert
Kibana

Implementation challenges

or: how to enable a container-based monitoring in an rpm-only system landscape

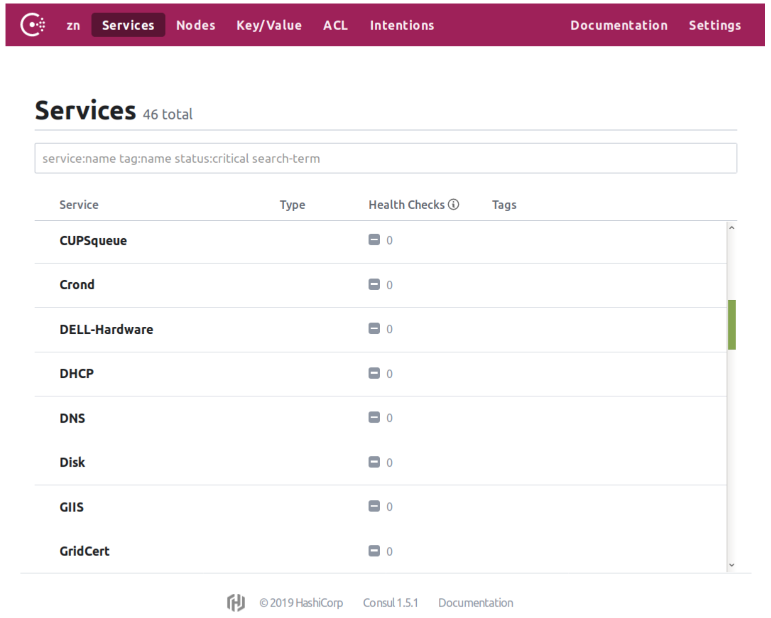
- The monitoring infrastructure is based on bleeding-edge software …
  - Release cycles of 3 months only
  - You might want to argue whether that's a good idea ;-)
- Container / microservice-based architecture
  - To keep things up-to-date you will need to establish an automatic container build pipeline
- There is a rather steep learning curve in understanding / learning the new possibilites

A closer look at some components

Consul service registry

Keeping track of monitored services

- Used as plain service registry
- Distributed over monitoring nodes for redundancy
- Was the best choice in our case due to good integration with Prometheus
  - Integrated with old registration workflow

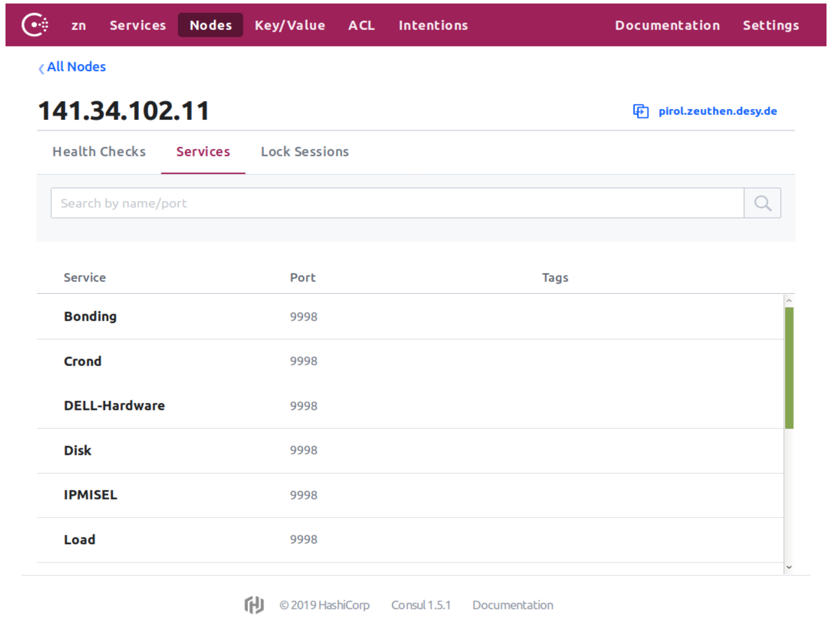


Services 46 total

service:name tag:name status:critical search-term

Service	Type	Health Checks	Tags
CUPSqueue		0	
Crond		0	
DELL-Hardware		0	
DHCP		0	
DNS		0	
Disk		0	
GIIS		0	
GridCert		0	

© 2019 HashiCorp Consul 1.5.1 Documentation



141.34.102.11 pirol.zeuthen.desy.de

Service	Port	Tags
Bonding	9998	
Crond	9998	
DELL-Hardware	9998	
Disk	9998	
IPMISEL	9998	
Load	9998	

© 2019 HashiCorp Consul 1.5.1 Documentation

Prometheus alerts

- Prometheus stores metrics in a time series database
- Alerts are actually just a list of YAML-based definitions describing out-of-order metrics
  - Prometheus brings its own query language: PromQL
  - Many of them are provided together with the exporter (e.g. CEPH, …)
- Simple example: alert if a certificate expires soon

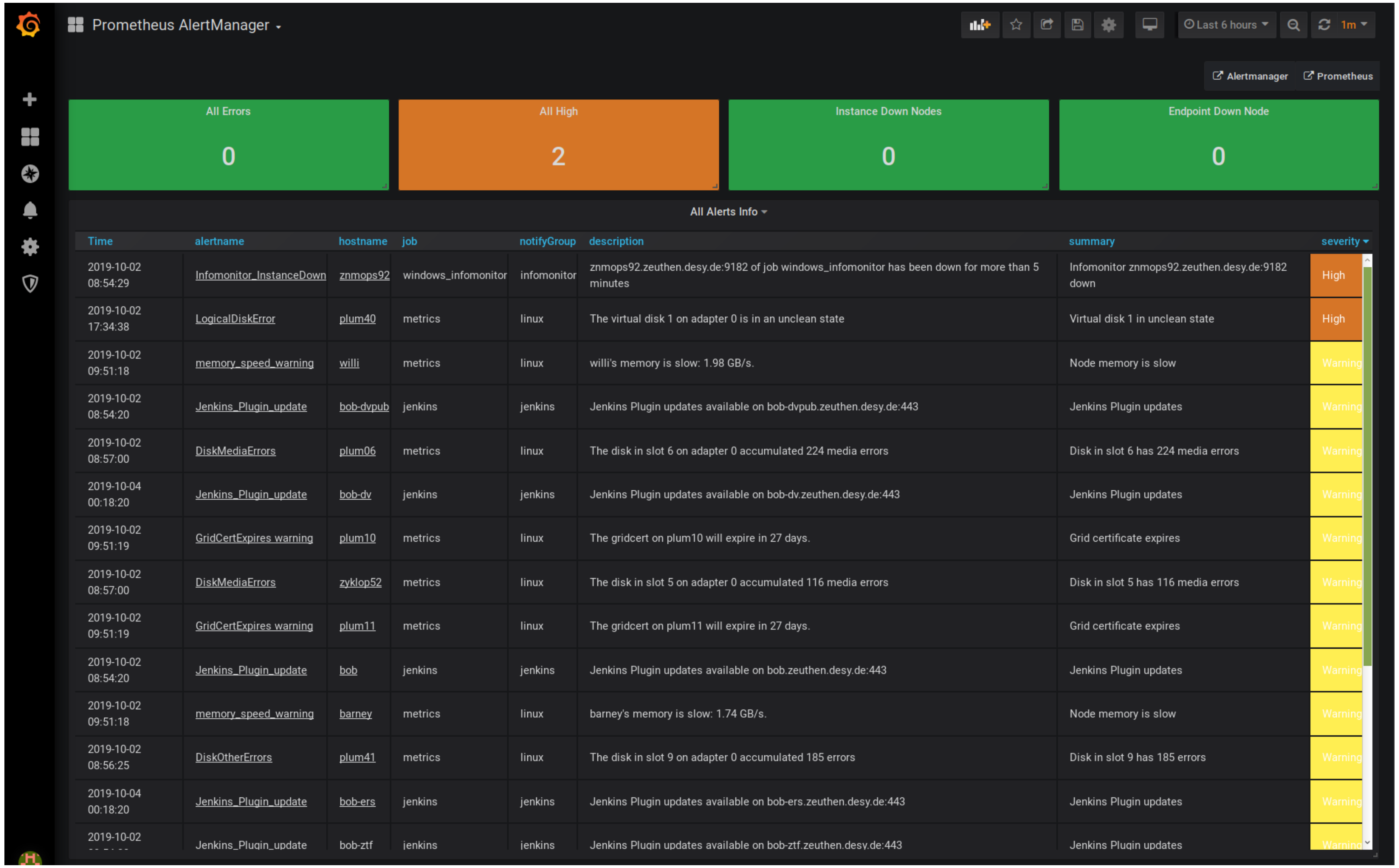
```
- name: hostcert
  rules:
  - alert: HostCertExpires warning
    expr: hostcert_cert_expires < 30
    for: 60m
    labels:
      severity: warning
      notifyGroup: linux
    annotations:
      summary: "Host certificate expires"
      description: "The hostcert on {{ $labels.hostname }} will expire in {{ humanize $value}} days."
```

Future prediction with PromQL

- Many time-based functions exist with PromQL
  - Implementing predictions is quite hard, though …
  - Just learning how to adapt all the nice anomaly detection features work
- A more advanced example:

```
  - alert: filesystem_running_full
    expr: filesystem:free:percent < 25 AND predict_linear(filesystem:free:percent[3h], 2*24*3600) < 0 AND
stdvar_over_time(filesystem:free:percent[3h]) < 0.25 AND delta(filesystem:free:percent[1h]) < -0.25
    for: 30m
    labels:
      severity: high
      notifyGroup: linux
    annotations:
      summary: "Node filesystem running full"
      description: "{{ $labels.hostname }}'s filesystem {{ $labels.mountpoint }} is likely to run full during the
next 2 days! {{ $value }}% space left."
```

Alert dashboard



All Errors	All High	Instance Down Nodes	Endpoint Down Node
0	2	0	0

All Alerts Info

Time	alertname	hostname	job	notifyGroup	description	summary	severity
2019-10-02 08:54:29	Infomonitor_InstanceDown	znmops92	windows_infomonitor	infomonitor	znmops92.zeuthen.desy.de:9182 of job windows_infomonitor has been down for more than 5 minutes	Infomonitor znmops92.zeuthen.desy.de:9182 down	High
2019-10-02 17:34:38	LogicalDiskError	plum40	metrics	linux	The virtual disk 1 on adapter 0 is in an unclean state	Virtual disk 1 in unclean state	High
2019-10-02 09:51:18	memory_speed_warning	willi	metrics	linux	willi's memory is slow: 1.98 GB/s.	Node memory is slow	Warning
2019-10-02 08:54:20	Jenkins_Plugin_update	bob-dvpub	jenkins	jenkins	Jenkins Plugin updates available on bob-dvpub.zeuthen.desy.de:443	Jenkins Plugin updates	Warning
2019-10-02 08:57:00	DiskMediaErrors	plum06	metrics	linux	The disk in slot 6 on adapter 0 accumulated 224 media errors	Disk in slot 6 has 224 media errors	Warning
2019-10-04 00:18:20	Jenkins_Plugin_update	bob-dv	jenkins	jenkins	Jenkins Plugin updates available on bob-dv.zeuthen.desy.de:443	Jenkins Plugin updates	Warning
2019-10-02 09:51:19	GridCertExpires warning	plum10	metrics	linux	The gridcert on plum10 will expire in 27 days.	Grid certificate expires	Warning
2019-10-02 08:57:00	DiskMediaErrors	zyklop52	metrics	linux	The disk in slot 5 on adapter 0 accumulated 116 media errors	Disk in slot 5 has 116 media errors	Warning
2019-10-02 09:51:19	GridCertExpires warning	plum11	metrics	linux	The gridcert on plum11 will expire in 27 days.	Grid certificate expires	Warning
2019-10-02 08:54:20	Jenkins_Plugin_update	bob	jenkins	jenkins	Jenkins Plugin updates available on bob.zeuthen.desy.de:443	Jenkins Plugin updates	Warning
2019-10-02 09:51:18	memory_speed_warning	barney	metrics	linux	barney's memory is slow: 1.74 GB/s.	Node memory is slow	Warning
2019-10-02 08:56:25	DiskOtherErrors	plum41	metrics	linux	The disk in slot 9 on adapter 0 accumulated 185 errors	Disk in slot 9 has 185 errors	Warning
2019-10-04 00:18:20	Jenkins_Plugin_update	bob-ers	jenkins	jenkins	Jenkins Plugin updates available on bob-ers.zeuthen.desy.de:443	Jenkins Plugin updates	Warning
2019-10-02	Jenkins_Plugin_update	bob-ztf	jenkins	jenkins	Jenkins Plugin updates available on bob-ztf.zeuthen.desy.de:443	Jenkins Plugin updates	Warning

Alert workflow

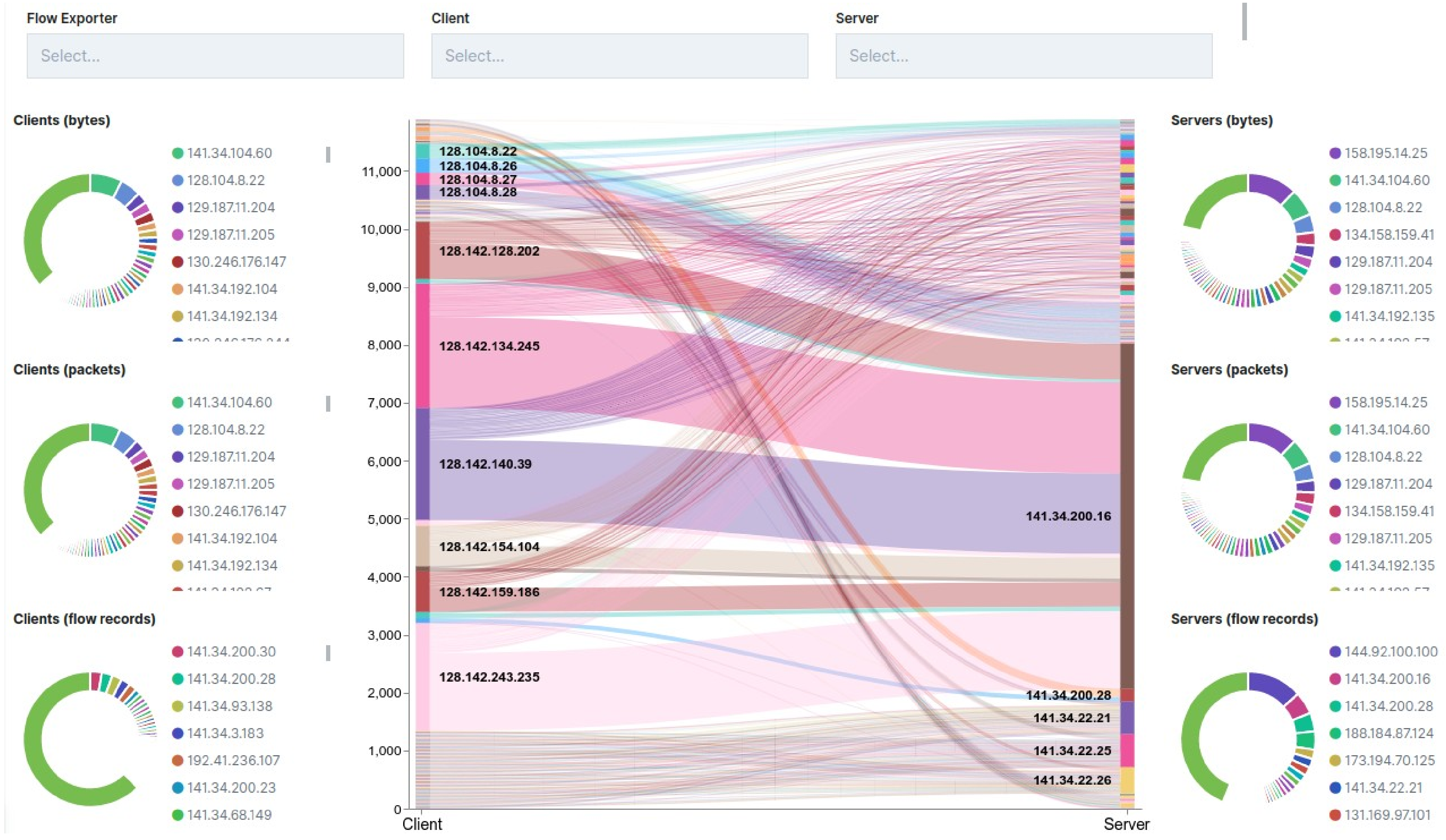
- Alerts with severity level > 'warning' cause mail to recipients defined in 'notifyGroup'
  - Own template
  - Brief error description in subject (which is not the default for some reason …)
  - Mail body contains links to Dashboard, Alertmanager to easily silence (“acknowledge”) the problem
- Once acknowledged, admins are advised to 'silence' alerts that cannot be fixed immediately
  - … otherwise alert mails will be sent again and again

ElastiFlow

Netflow analysis with ElasticSearch

- Not going to talk much about event-based monitoring
  - just started event-based monitoring with ELK, so all is rather new …
- One application: ElastiFlow: Monitor and visualise network flows
  - Our NTop installation has been out of service for some time, so we were looking for a replacement
  - Traffic from/to internet only at the moment
  - Unfortunately not all network devices are capable of providing unsampled data
    - Was not a criteria during purchase decision
    - Devil is in the detail as usual ...

ElastiFlow Report



Final remarks

Issues

a personal's opinion rant

- Ansatz: Collect and store as much data as you can get
  - „Maybe there is a need for it later”
  - But is there really a need for all of it? – It's not easy to filter
  - You have to scale your monitoring infrastructure for high data volume
- The whole infrastructure is quite complex
  - But as long as we are able to manage it, no problem
- Many exporters already exist (https://prometheus.io/docs/instrumenting/exporters/)
  - but you cannot just 'yum install' them …
  - usually implemented in GO-lang
  - packaging GO-lang rpms is still not trivial (as long as you do not just package the final binary)
  - … and: why do all exporters need to be implemented as daemon?

Status quo

where are we now?

- More than 1000 systems (bare metal, vms) and their services monitored
  - hardware, disc failures, computing centre cooling, AFS, CEPH, DNS, ...
- Many pre-defined and self-written alerts implemented
  - but fine tuning is a challenge
  - … and time-consuming
- Prometheus:
  - 43 jobs, 4617 targets, 132 alerts
  - collects almost 3 million metrics every minute!
- ELK contains more than 5 billion entries

What's next?

- Finalise the event-based setup
  - Implement missing alerts
  - Improve / adapt dashboards
- Integrate messenger-based alerts on mobile phones?
  - Considered at the moment: Mattermost or Telegram
- Convince the Windows group, they do not need to take care for monitoring their services any longer ;-)

Questions?