CERN

A Dashboard is Worth a Thousand Words

Better Monitoring for Better Ops

Luca Magnoni – Computing Engineer/CERN IT
@lucamag

About myself

- Software Engineer
- > 10 years in distributed systems & data intensive applications
- Service Manager & Project Architect

CERN

a
Worldwide
collaboration

CERN's primary mission:

SCIENCE

Fundamental research on particle physics, pushing the boundaries of knowledge and technology

CERN

World's largest particle physics laboratory

Image credit: CERN

The Large Hadron Collider: LHC

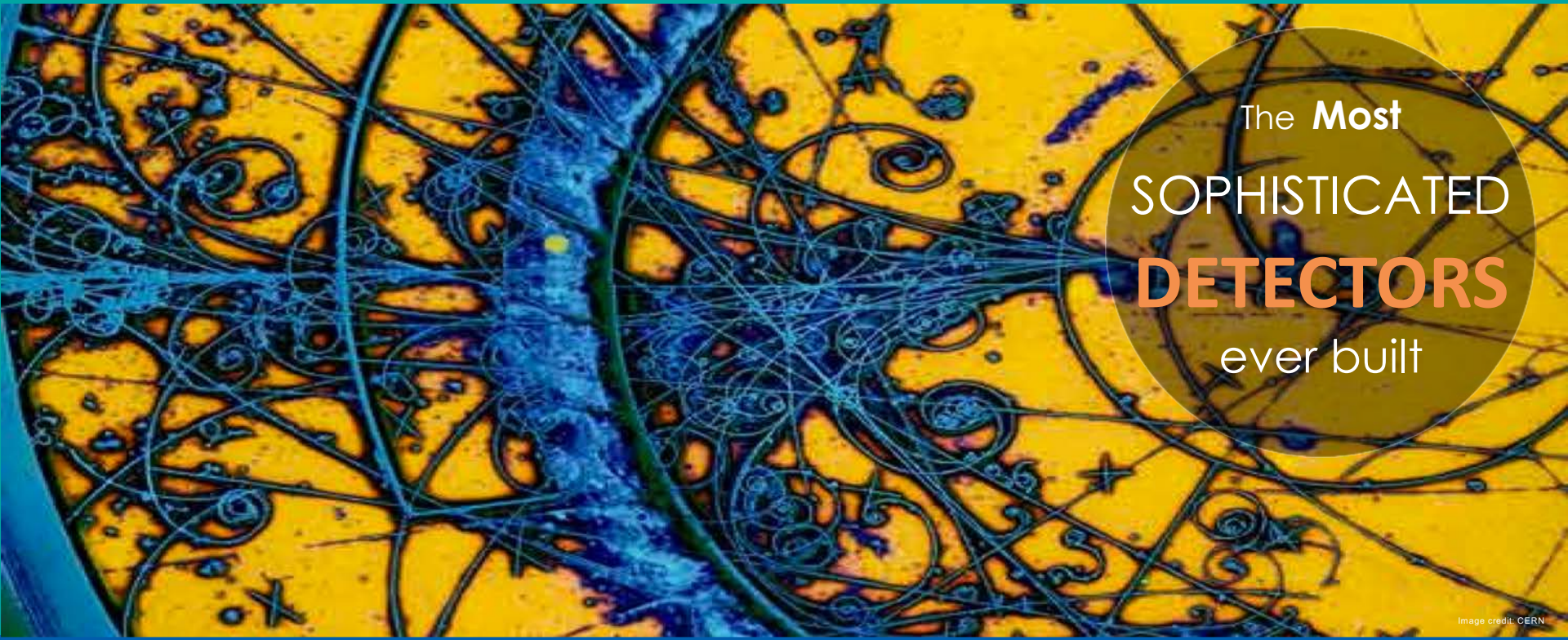
LHC: World's Largest Cryogenic System (1.9 K)

Image credit: CERN

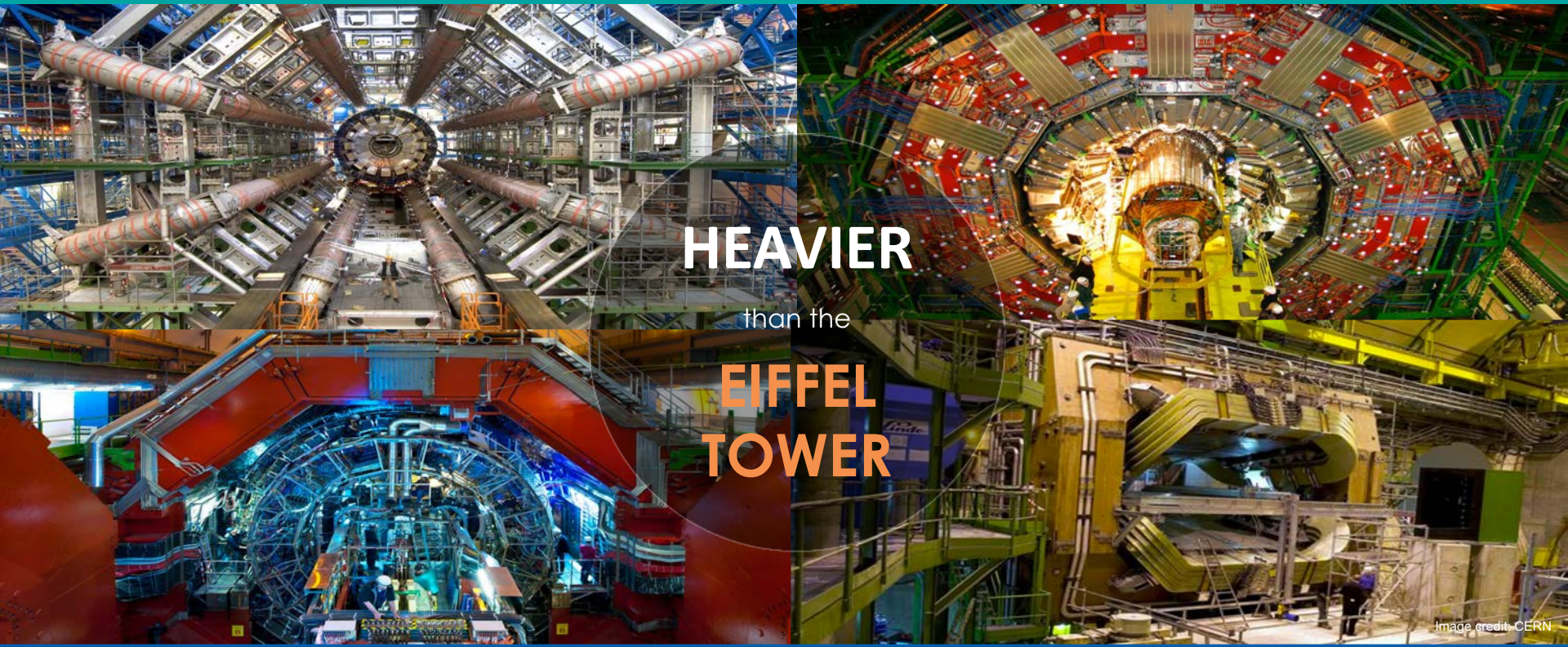
LHC: Highest Vacuum

Image credit: CERN

LHC Detectors



ATLAS, CMS, ALICE and LHCb



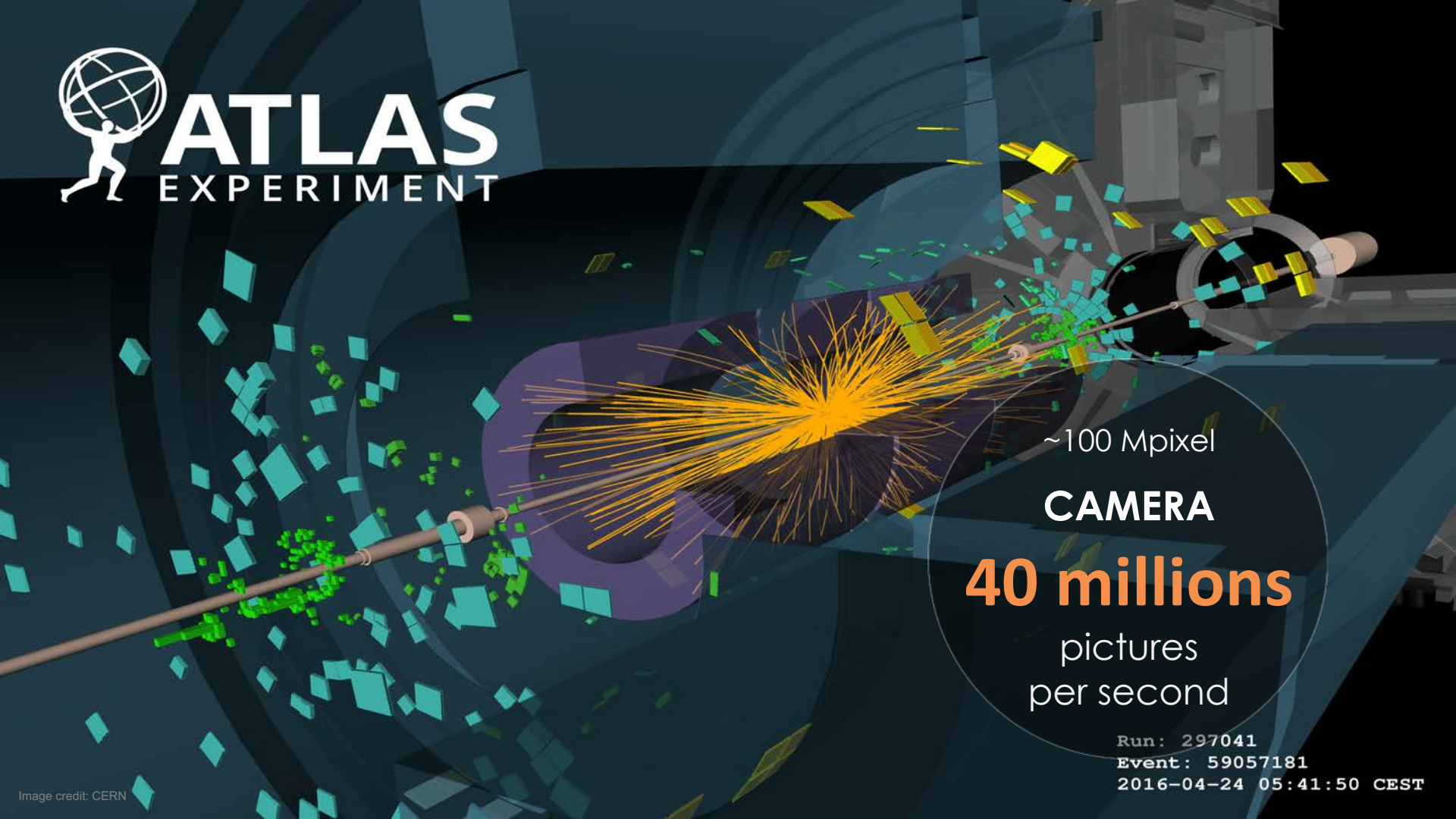
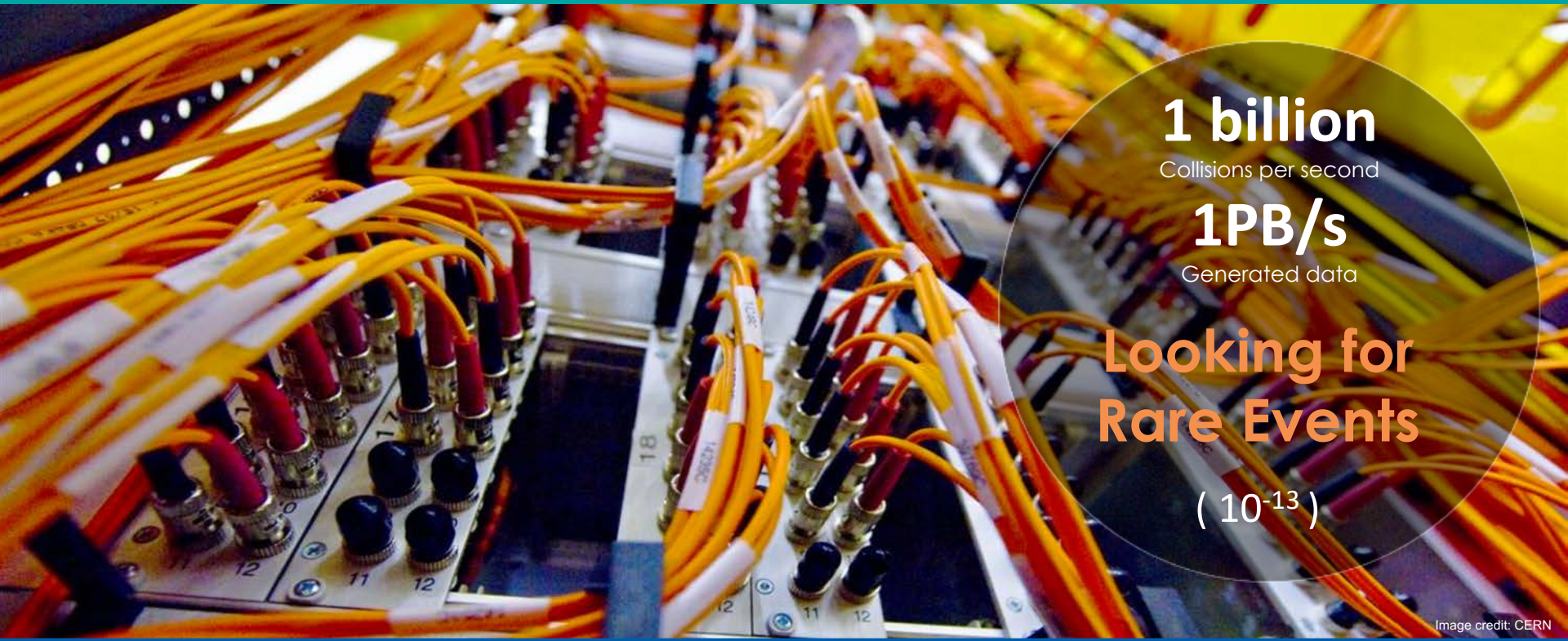
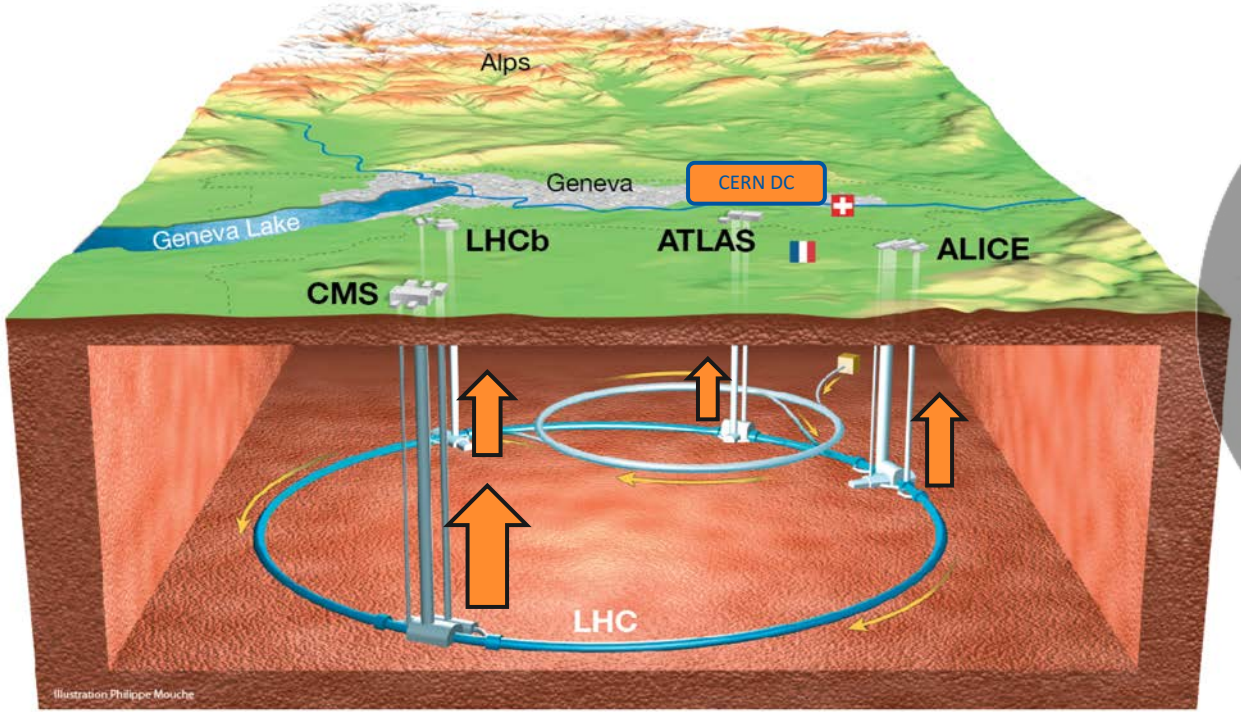


Image credit: CERN

Data Acquisition: What to Record?



Data Flow to Storage and Processing



RUN 2

ALICE: 4GB/s
ATLAS: 1GB/s
CMS: 600MB/s
LHCB: 750MB/s

CERN Data Centre: Primary Copy of LHC Data

WLCG: LHC Computing Grid

About WLCG:

- A community of 10,000 physicists
- ~250,000 jobs running concurrently
- 600,000 processing cores
- 15% of the resources are at CERN
- 700 PB storage available worldwide
- 20-40 Gbit/s connect CERN to Tier1s

Tier-0 (CERN)

- Initial data reconstruction
- Data distribution
- Data recording & archiving

Tier-1s (13 centres)

- Initial data reconstruction
- Permanent storage
- Re-processing
- Analysis

Tier-2s (>150 centres)

- Simulation
- End-user analysis

Image credit: CERN

opendata CERN

About

Explore more than **1 petabyte** of open data from particle physics!

Start typing...

Search

search examples: collision datasets, keywords:education, energy:7TeV

Explore

datasets
software
environments
documentation

Focus on

ATLAS
ALICE
CMS
LHCb
OPERA

Get started

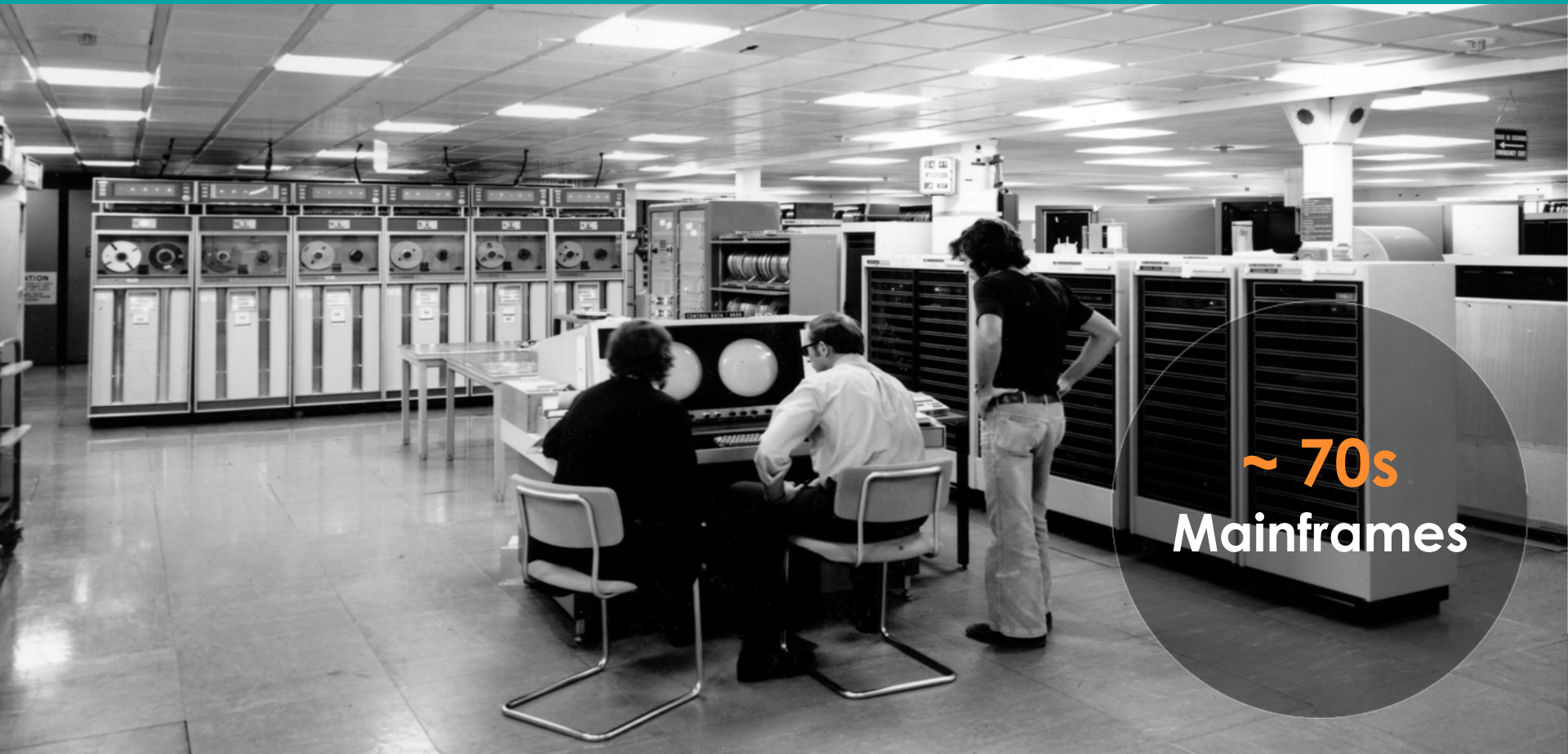
CERN IT

Image credit: CERN

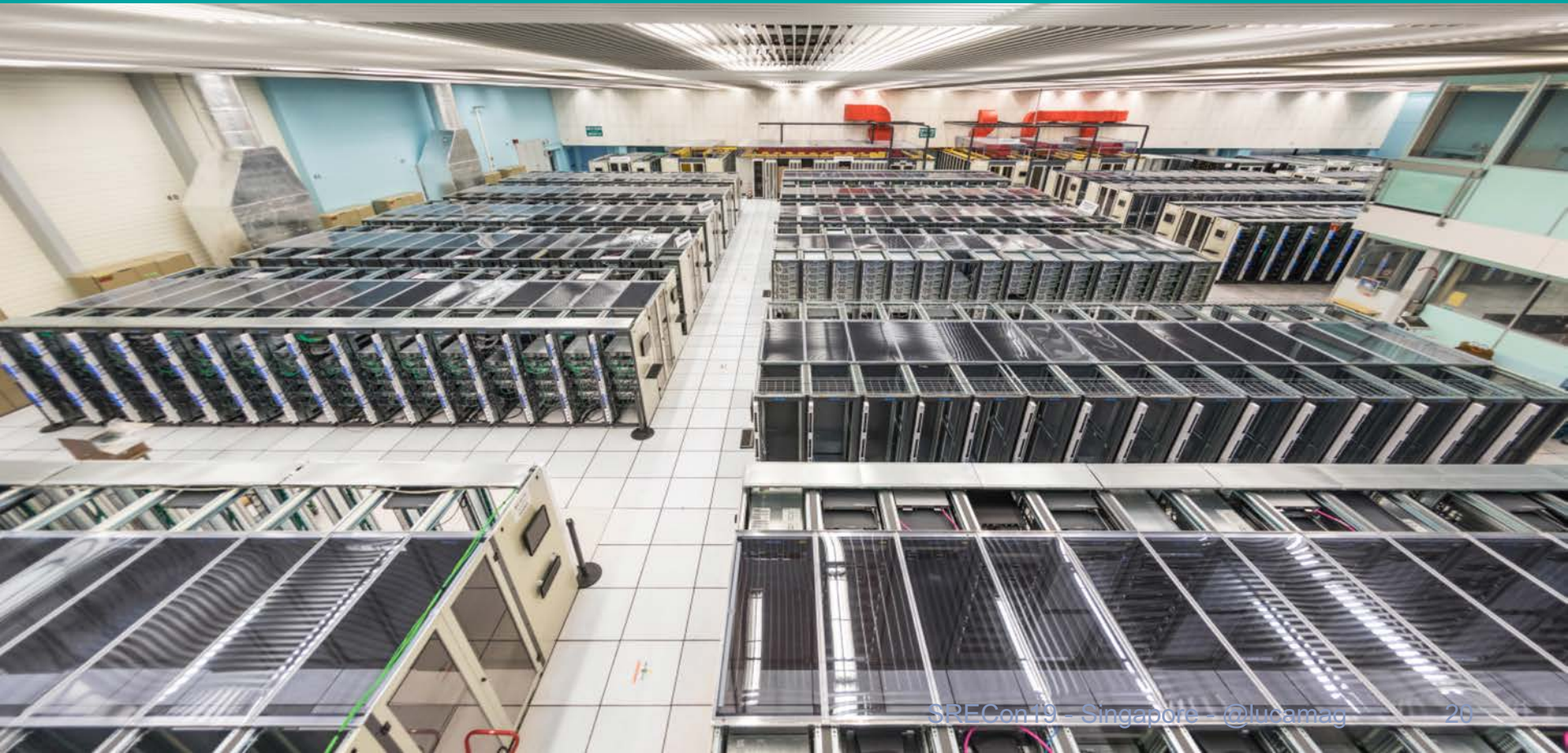
About IT Department

- Over 300 people
- Enable the laboratory to fulfill its mission
- Data Centre and more
  - supports IT Services (Batch, Storage, Network, DB, Web Servers, etc.), Experiments Services (SW builds), Engineering (Chip design), Infrastructure (hotel, bikes), Administration

IT Infrastructure / Early Days



IT Infrastructure / ~ Recent Days



~ 10 years of Data Taking

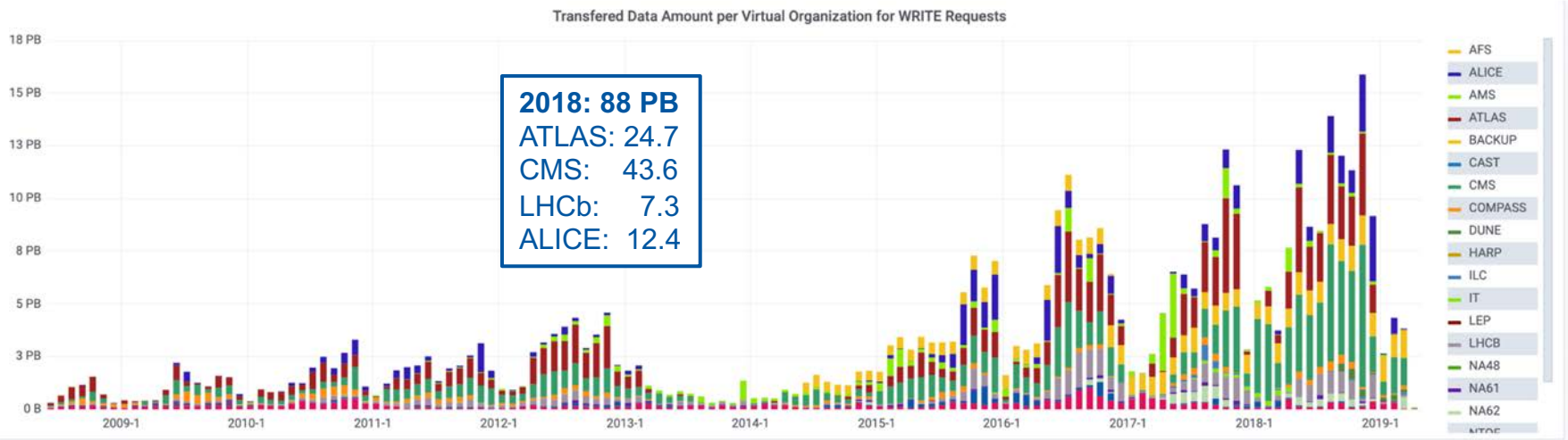


Image credit: CERN

~ 2001 / Custom Fabric

- EU funded developments
- Scale and experience for LHC was special
- Custom tools had to be developed to manage infrastructures at scale

~ 2013 / Opensource Tools

- LHC requirements kept growing (on flat budget)
- But CERN scale no longer special (e.g. Google, Facebook, Rackspace,…)
- The rise of the Clouds

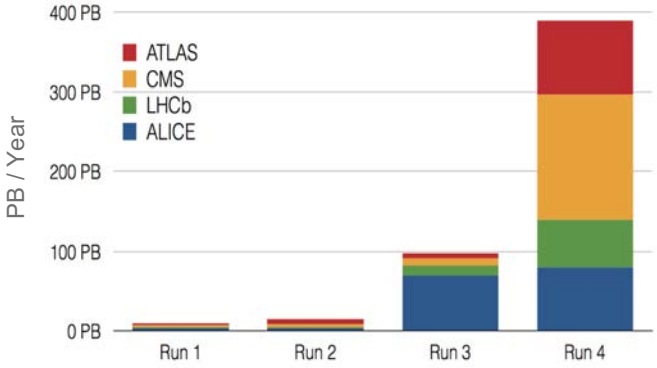


Image credit: CERN

~ 2013 / Opensource Tools

- Tool-Chain approach
- Embrace Opensource Communities
- Focus on Resource Provisioning, Configuration and Monitoring

CERN Data Centre: Private Openstack Cloud

Openstack overview

Zoom Out | Last 2 days | Refresh every 15m

Cloud resources

Available	Used	Available	Used	Available	Used
312.9 K cores	290.6 K cores	913.0 TiB RAM	793.7 TiB RAM	15.0 PiB disk	7.3 PiB disk

Openstack services stats

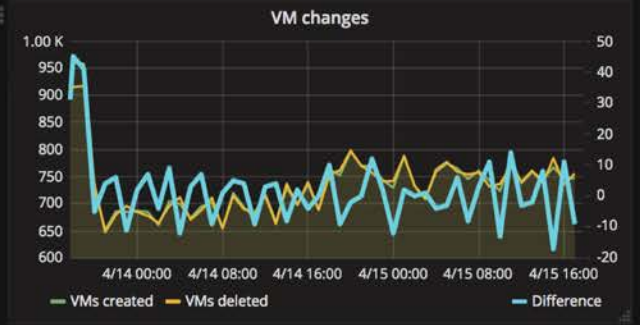
Users	Projects	VMs	Magnum clusters	Hypervisors	Images
3237	4255	37680	253	9020	2870

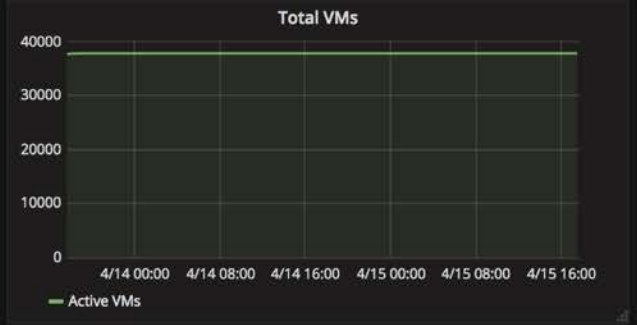
Volumes	Volume size	Fileshares	Fileshares size
5644	1.77 PiB	96	[illegible]

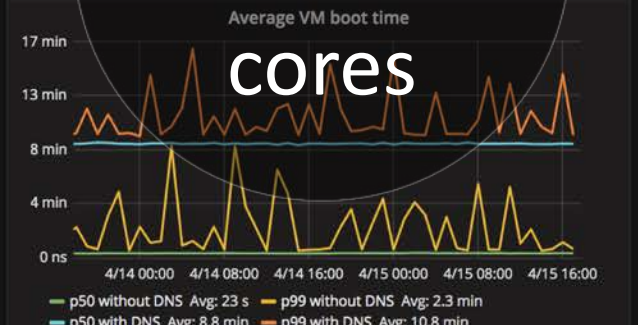
More Than

300 000

cores







2019 / Even more *Tools* …

- Containers / Kubernetes
  - New deployment models
- More Clouds
  - Hybrid workflows
- *SRE* ?

Monitoring *“all the things”*

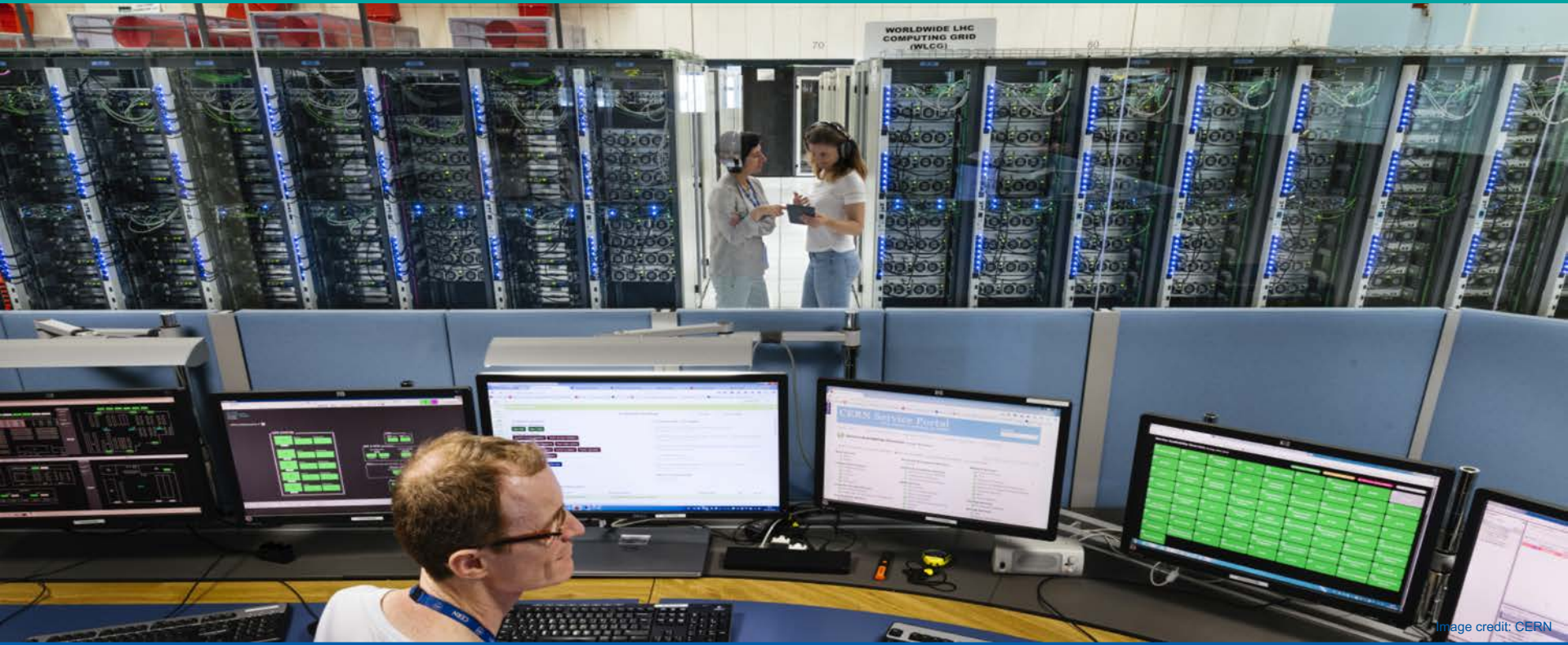


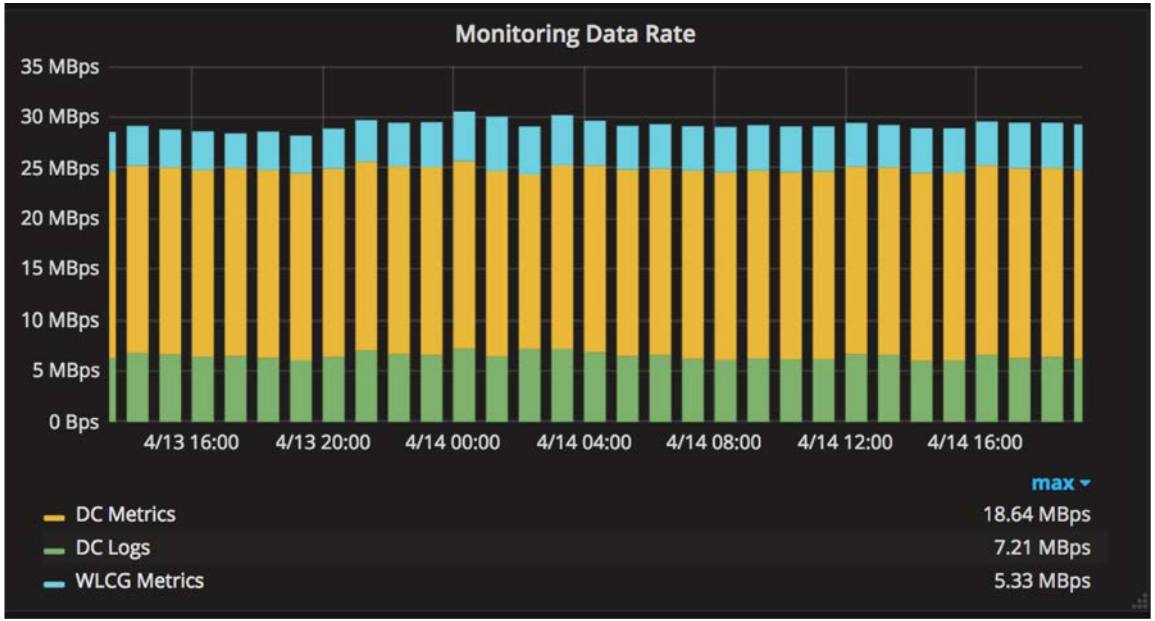
Image credit: CERN

Monitoring Mission

- Provide **Monitoring as a Service** for **CERN Data Centre (DC), IT Services** and the **WLCG collaboration**
  - e.g. Dashboards, Alarms, Search, Archive
- Collect, transport, store and process metrics and logs for applications and infrastructure

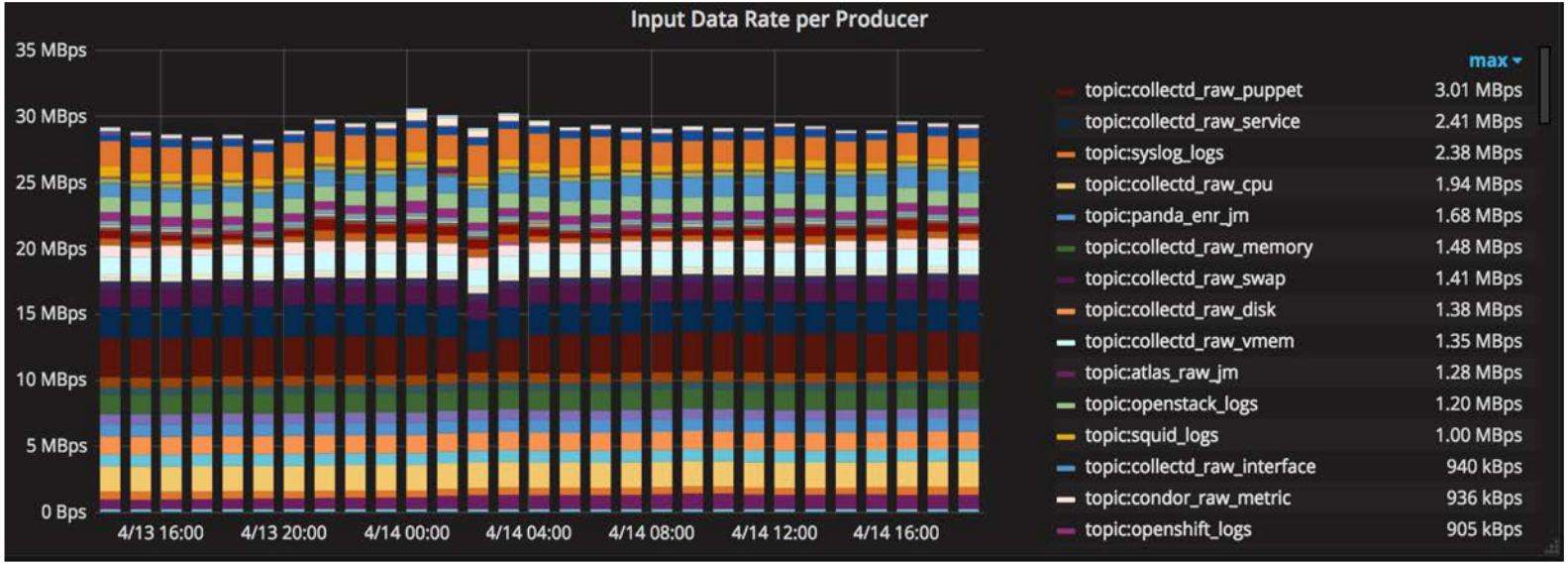
Challenges / Rate & Volume

- from ~ 40k machines
- > 3 TB/day (compressed)
- ~ 100 kHz



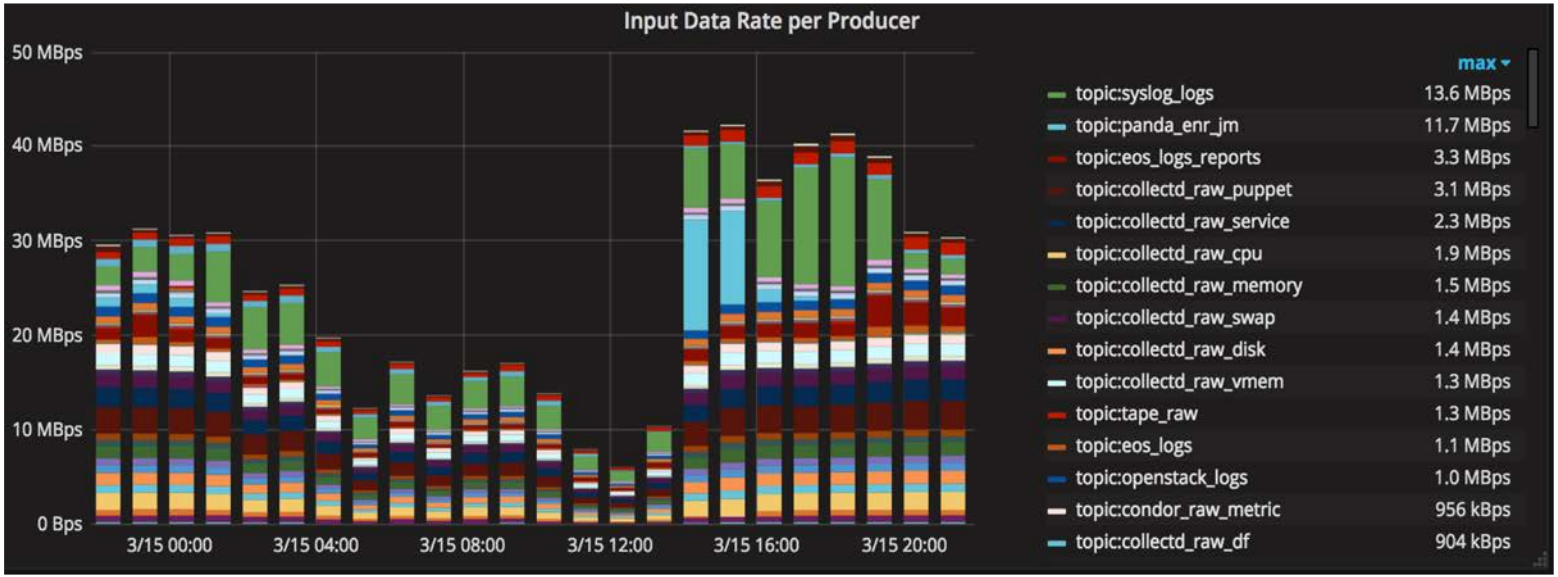
Challenges / Variety

- > 150 producers



Challenges / Reliability

- spikes in rate and volume



Non-Technical Challenges

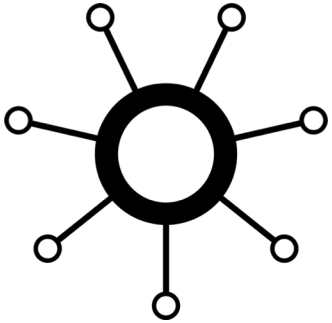
- Migrate people from legacy (custom) dashboards and tools
- Stay up to date with upstream tools & trends
- Build community, internal and external

How to provide **better** monitoring?

- ~ 2016 a new project started to provide a new central monitoring infrastructure to CERN IT
- Goal:
  - Effective
  - Scalable
  - Sustainable

Easy Data Integration / Telemetry

- Collectd
  - lightweight / plugin based
  - ~ 40k machines
- JSON/HTTP gateways
  - custom metrics and logs
- Prometheus
  - Kubernetes



Responsive / Multiple Backends

- Elasticsearch
  - search and discovery
  - 3 clusters, ~ 100 TB
- InfluxDB
  - time-series data
  - > 30 instances, 60 kHz
- HDFS
  - long-term archive

Data Integration is ***hard***

- **Metrics** → TSDB
- **Logs** → search/index
- **All** data → archive
- *Some* Metrics → search/index
- *Some* Logs → TSDB
- *Users* : *" btw, where can I tap in to get my data? "*

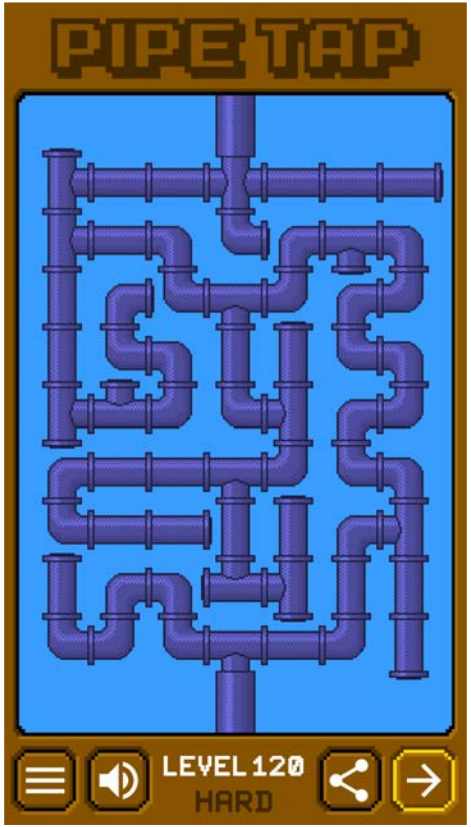
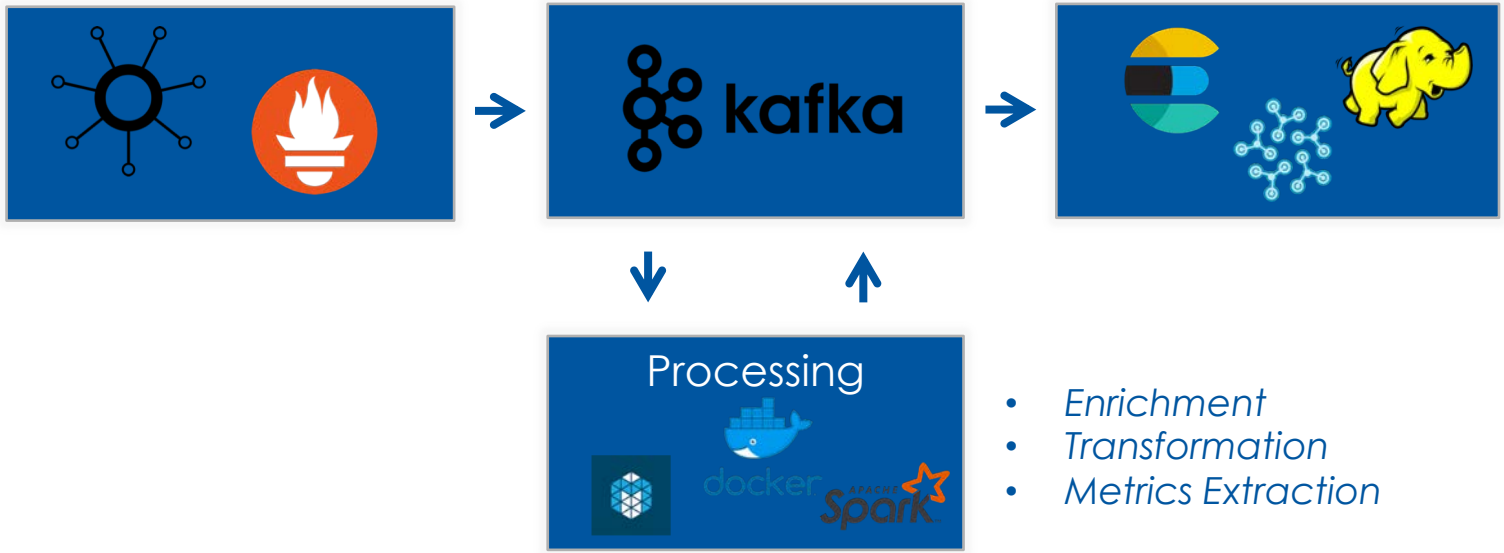


Image: © 2014 - 2017 Cool Wolf Games

Monitoring Pipeline



- *Enrichment*
- *Transformation*
- *Metrics Extraction*

On the pipeline approach

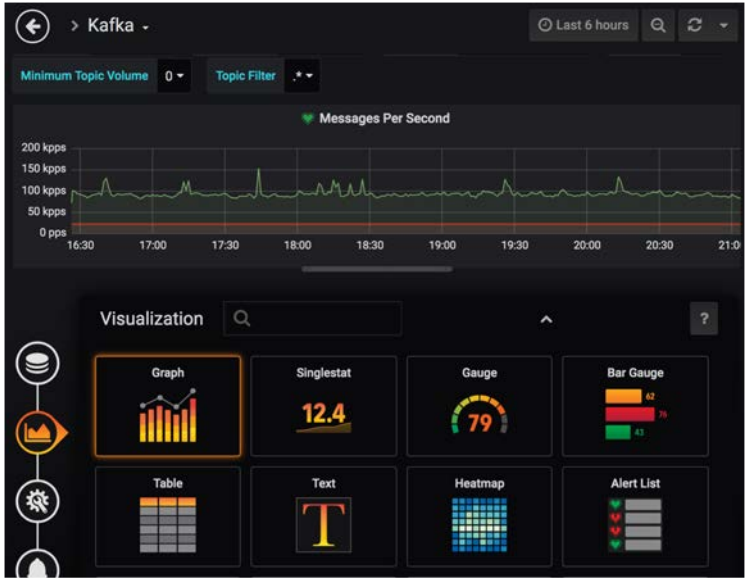
- Provides key functionalities:
  - decouples producers / consumers
  - enables stream processing
  - resilient (72 hours data retention)
- Kafka cluster:
  - on-premises (v 1.0.2)
  - Openstack VMs with Ceph volumes
  - ~ **15k** partitions in total

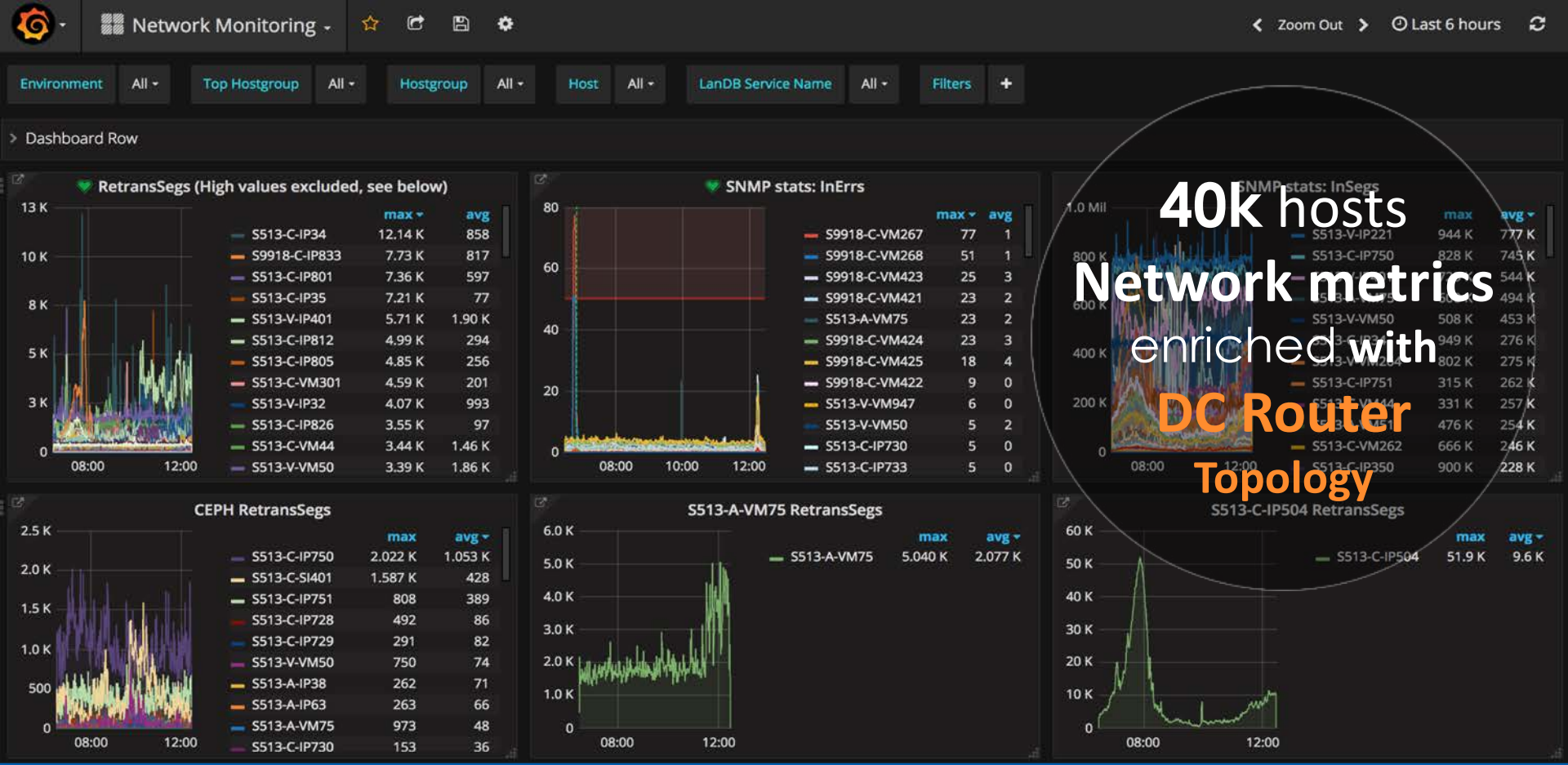
Dashboards & Visualizations

- Critical for the success of the project
- Need to delegate control to users
- Multiple tools

Dashboard / Grafana

- Recommended tool for Dashboard
- Multiple Backends
- Customizable (menus, filters, etc.)





40K hosts

Network metrics

enriched **with**

DC Router

Topology

ATLAS DDM Dashboard (Latest data)

Zoom Out | Last 4 hours UTC

Group by: dst_cloud | Binning: auto | Activity: Analysis Input + Data Brokering + Data Consolidation + Data Rebalancing + Deletion + Express + Functional Test + Production Input + Production Output + Recovery + T0 Export + T0 Tape + User Subscriptions + default + Staging

Source country: All | Source site: All | Destination country: All | Destination site: All | Filters: + | Matrix Columns: dst_country | Matrix Rows: src_country

Transfers

Matrix

Efficiency

	Australia	Canada	Chile	China	Czech Republic	France	Germany	Greece	Israel	Italy	Japan	Netherlands	Nordic
Australia	-	92%	-	100%	100%	100%	98%	-	100%	100%	100%	95%	100%
Canada	100%	67%	-	100%	90%	97%	99%	-	82%	100%	100%	100%	100%
Chile	100%	100%	-	-	100%	44%	100%	-	50%	83%	100%	20%	100%
China	-	-	-	-	-	0%	70%	-	-	100%	-	-	100%
Czech Republic	100%	100%	-	-	-	60%	100%	-	60%	100%	100%	100%	100%
France	100%	100%	100%	94%	100%	94%	99%	-	51%	100%	98%	100%	44%
Germany	100%	95%	-	97%	93%	93%	100%	-	83%	99%	100%	100%	88%
Greece	0%	0%	-	0%	0%	0%	0%	-	0%	0%	0%	0%	0%
Israel	-	95%	-	-	100%	100%	100%	-	-	100%	-	-	-
Italy	100%	90%	-	73%	100%	92%	98%	-	77%	82%	100%	89%	100%
Japan	100%	100%	100%	79%	100%	83%	100%	-	84%	99%	-	100%	67%
Netherlands	100%	100%	-	100%	100%	100%	99%	-	100%	64%	100%	31%	100%
Nordic	100%	100%	100%	100%	100%	81%	99%	-	77%	100%	100%	100%	0%
Poland	-	100%	-	-	100%	64%	100%	-	60%	100%	100%	100%	100%
Portugal	-	100%	-	-	-	100%	100%	-	-	-	-	-	-

	Australia	Canada	Chile	China	Czech Republic	France	Germany	Greece	Israel	Italy	Japan	Netherlands	Nordic
Australia	-	132 kBps	-	978 Bps	424 kBps	262 kBps	7 MBps	-	2 Bps	3 MBps	125 kBps	7 kBps	3 kBps
Canada	37 kBps	86 kBps	-	26 kBps	668 kBps	[illegible]	[illegible]	[illegible]	[illegible]	155 kBps	710 kBps	176 MBps	3 MBps
Chile	81 Bps	6 kBps	-	-	76 kBps	63 kBps	1 MBps	-	2 kBps	7 kBps	767 Bps	3 kBps	218 Bps
China	-	-	-	-	-	0 Bps	289 kBps	-	-	17 kBps	-	-	521 Bps
Czech Republic	21 kBps	192 kBps	-	[illegible]	[illegible]	[illegible] kBps	[illegible]	[illegible]	[illegible] kBps	[illegible] kBps	[illegible] kBps	79 kBps	212 kBps
France	3 kBps	676 kBps	2 kBps	47 kBps	2 MBps	47 MBps	[illegible]	[illegible]	20 MBps	775 kBps	1 MBps	2 MBps	893 kBps
Germany	443 kBps	18 MBps	-	2 MBps	8 MBps	125 MBps	62 MBps	-	2 MBps	9 MBps	17 MBps	5 MBps	7 MBps
Greece	0 Bps	0 Bps	-	0 Bps	0 Bps	0 Bps	0 Bps	-	0 Bps	0 Bps	0 Bps	0 Bps	0 Bps
Israel	-	847 kBps	-	-	539 kBps	207 kBps	4 MBps	-	-	45 Bps	-	-	-
Italy	46 kBps	4 MBps	-	3 kBps	1 MBps	2 MBps	22 MBps	-	6 kBps	304 kBps	2 MBps	176 kBps	351 kBps
[illegible]	[illegible]	871	113	3	[illegible]	264	[illegible]	[illegible]	51	376	[illegible]	[illegible]	152

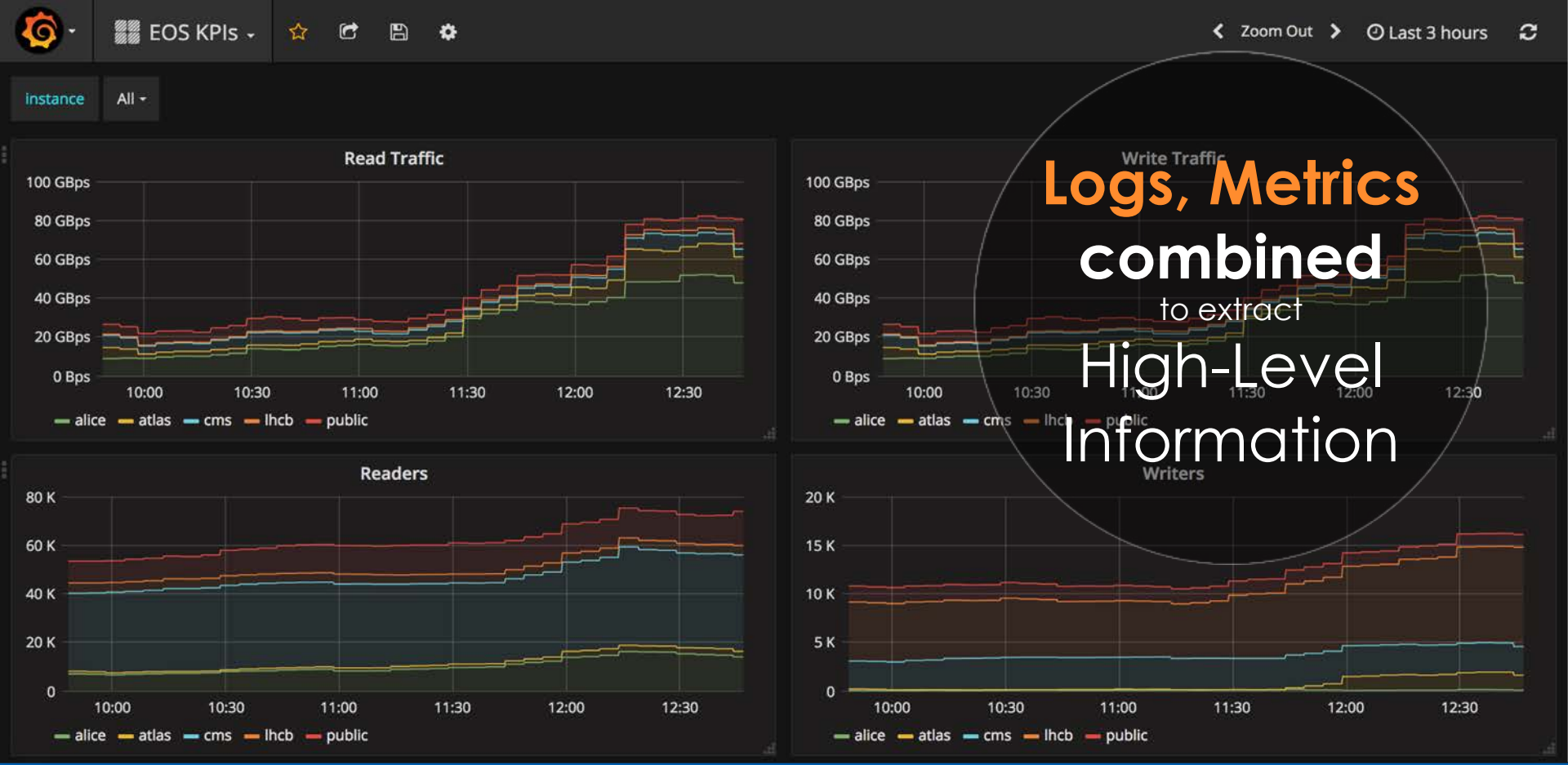
Staging

7 **Millions**/day

WLCG transfers

classified

By location, country site,…



Logs, Metrics
combined
to extract
High-Level
Information

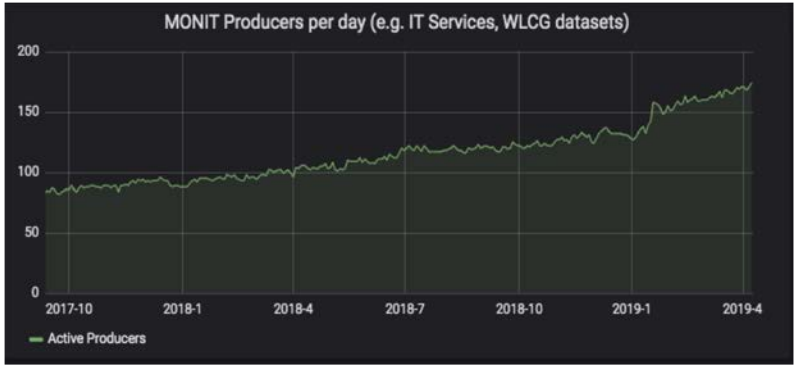
Alarming

- Local (on the machine)
  - Simple Threshold / Actuators
- On Dashboards
  - Grafana Alert Engine
- External (Spark, Spectrum, etc.)
- Integrated with ticketing system
  - ServiceNow

Monitoring Technologies

Successful story

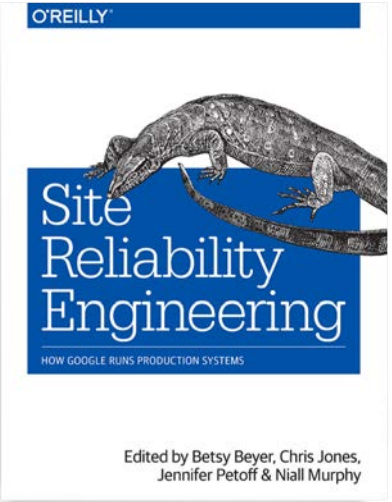
- Monitoring by Numbers:
  - ~ 900 Active Users
  - > 1000 Dashboards
  - ~ 1000000 Queries/day
- > 30 Grafana Orgs
  - service operations, debug, troubleshooting, etc.
- Next is to profit at best from all this data



Better Ops

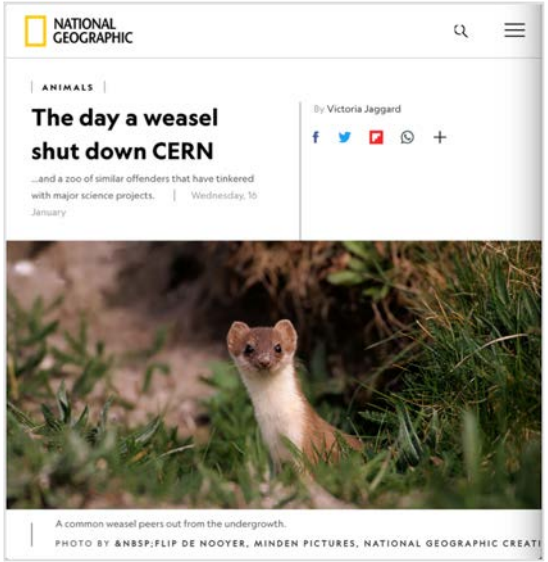
SRE Key Points of Interest

- Common framework for production-systems management
- Reduce operational load
- Formalize best-practices for software velocity and quality



SRE Practices / Good fit with CERN IT

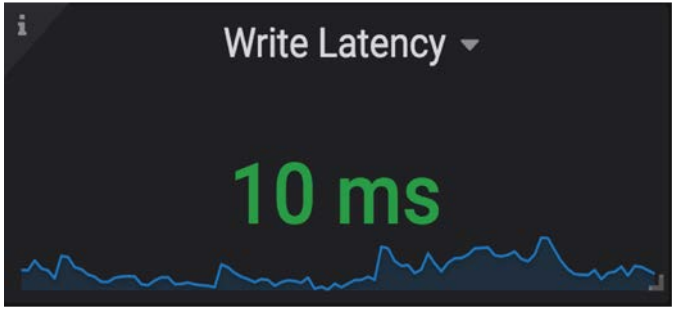
- SRE Culture
  - Openness, Sharing
- Joint-Ownerships / Accountability
  - “one person’s symptom is another person’s cause”
- Sustainability
  - Attracts skills, Career opportunities



blameless post-mortems

Monitoring like SRE

- Goal: build common language & culture
- Introducing SLI & SLO
- One Dashboard at the Time



Introducing SLI / Strategy

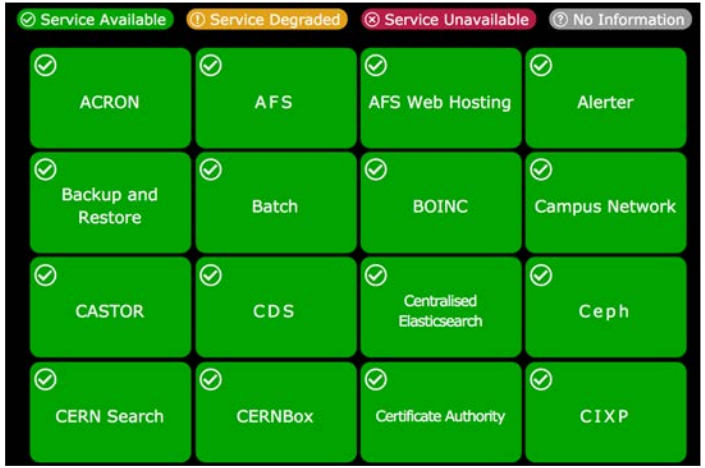
- Build critical mass of early adopters among main IT services
- Work with service managers to extract the relevant data (*some* indicators already there…)
- Try to solve a real problem to help the idea spread faster
  - i.e. improve Service Availability reporting

Tackling Service Availability

- True stories from chat / mattermost snippets:
  - Is there a problem with service *X* ?
  - Is anybody else having *issues today ?*
  - I *think* service *X* is *slow*
  - I *think* service *X* is having some *issue lately*
  - How our service *evolved*? Are we doing *better*?

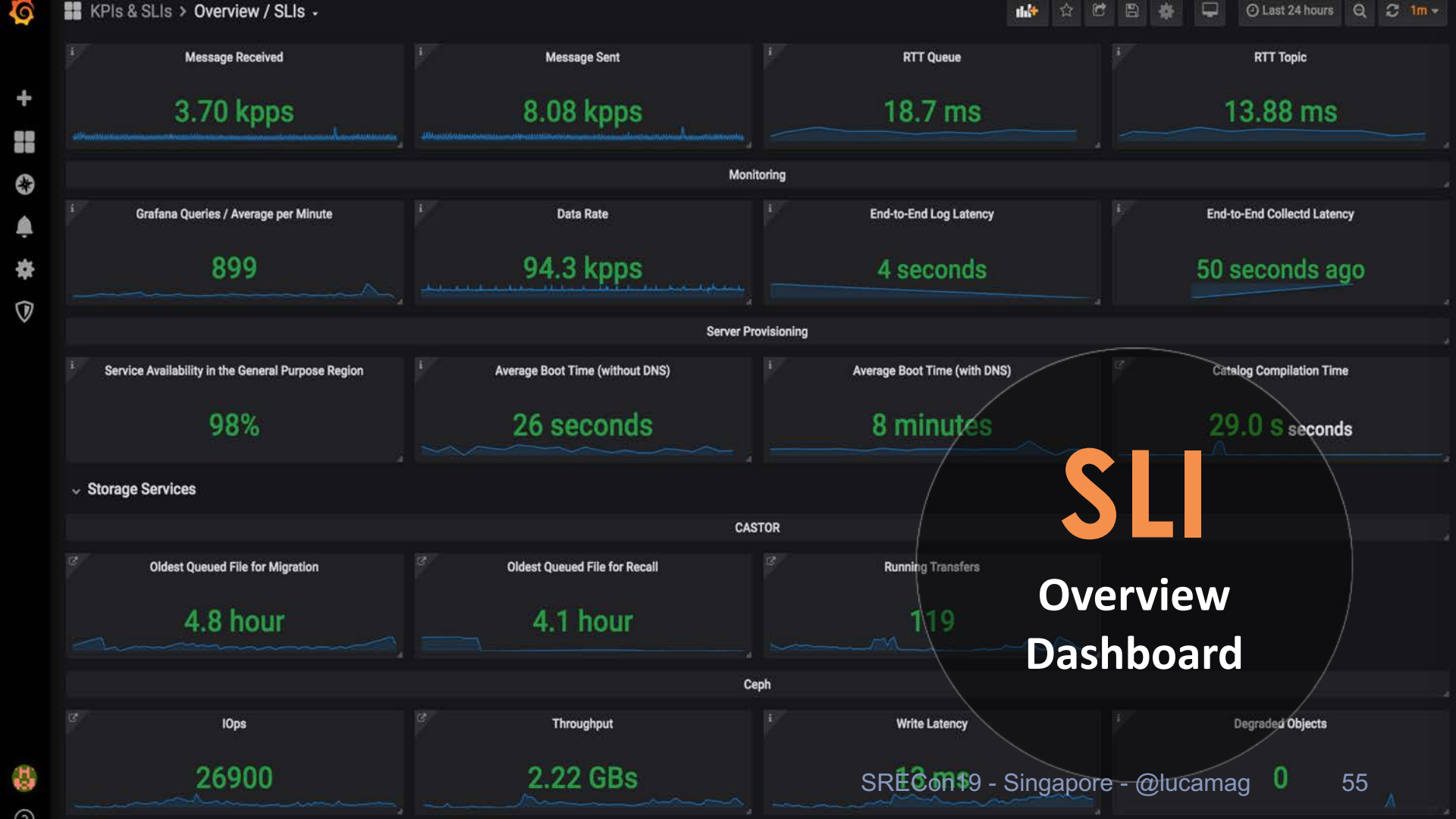
Tackling Service Availability / 2

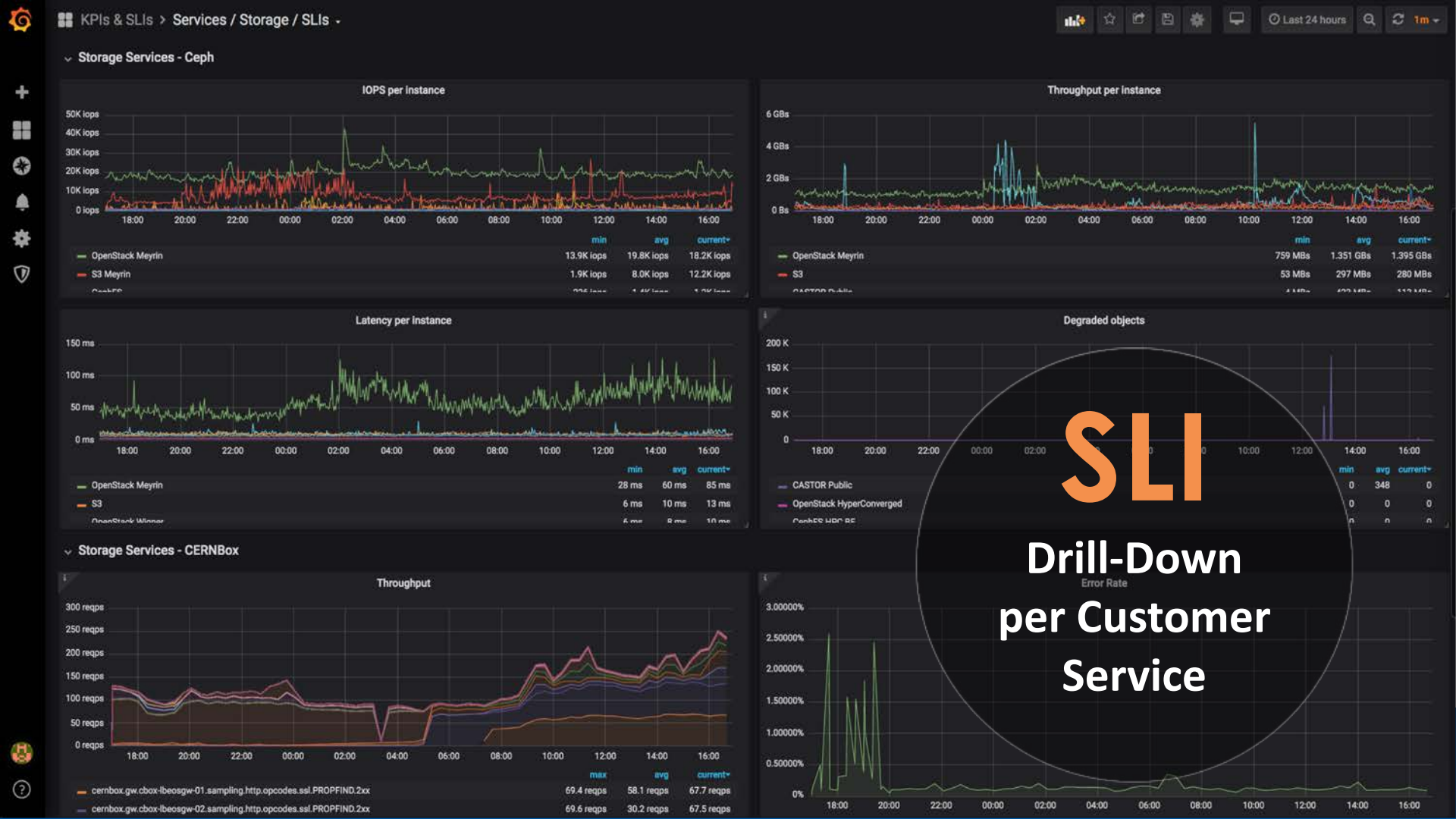
- Availability Metric exists
- Not easy to get actual health of services today
- Would benefit from more precise and quantitative measurement



SLI Overview Dashboard

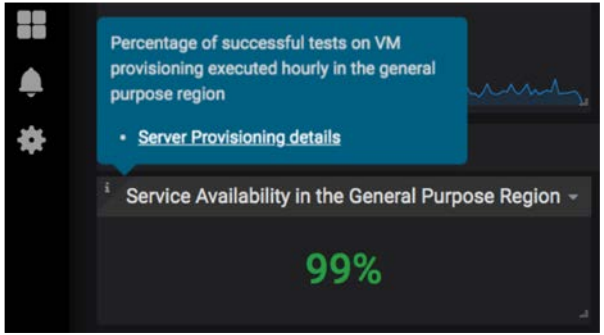
- SLI as user-facing metrics
- Focus on Golden signals, such as
  - Write/Read Latency for Storage systems
  - Rate of cloud API requests
  - Rate of batch server occupancy
  - Catalog compilation time
- Dashboard: visual feedback for users


KPIs & SLIs > Overview / SLIs
Last 24 hours
1m
Message Received
3.70 kpps
Message Sent
8.08 kpps
RTT Queue
18.7 ms
RTT Topic
13.88 ms
Monitoring
Grafana Queries / Average per Minute
899
Data Rate
94.3 kpps
End-to-End Log Latency
4 seconds
End-to-End Collectd Latency
50 seconds ago
Server Provisioning
Service Availability in the General Purpose Region
98%
Average Boot Time (without DNS)
26 seconds
Average Boot Time (with DNS)
8 minutes
Catalog Compilation Time
29.0 s seconds
Storage Services
CASTOR
Oldest Queued File for Migration
4.8 hour
Oldest Queued File for Recall
4.1 hour
Running Transfers
119
Ceph
IOps
26900
Throughput
2.22 GBs
Write Latency
Degraded Objects
0
SLI
Overview
Dashboard


KPIs & SLIs > Services / Storage / SLIs
Last 24 hours
1m
Storage Services - Ceph
IOPS per instance
50K iops
40K iops
30K iops
20K iops
10K iops
0 iops
min avg current
OpenStack Meyrin 13.9K iops 19.8K iops 18.2K iops
S3 Meyrin 1.9K iops 8.0K iops 12.2K iops
Throughput per instance
6 GBs
4 GBs
2 GBs
0 Bs
min avg current
OpenStack Meyrin 759 MBs 1.351 GBs 1.395 GBs
S3 53 MBs 297 MBs 280 MBs
Latency per instance
150 ms
100 ms
50 ms
0 ms
min avg current
OpenStack Meyrin 28 ms 60 ms 85 ms
S3 6 ms 10 ms 13 ms
Degraded objects
200 K
150 K
100 K
50 K
0
min avg current
CASTOR Public 0 348 0
OpenStack HyperConverged 0 0 0
SLI
Drill-Down
per Customer
Service
Storage Services - CERNBox
Throughput
300 reqps
250 reqps
200 reqps
150 reqps
100 reqps
50 reqps
0 reqps
max avg current
cernbox.gw.cbox-lbeosgw-01.sampling.http.opcodes.ssl.PROPFIND.2xx 69.4 reqps 58.1 reqps 67.7 reqps
cernbox.gw.cbox-lbeosgw-02.sampling.http.opcodes.ssl.PROPFIND.2xx 69.6 reqps 30.2 reqps 67.5 reqps
Error Rate
3.00000%
2.50000%
2.00000%
1.50000%
1.00000%
0.50000%
0%

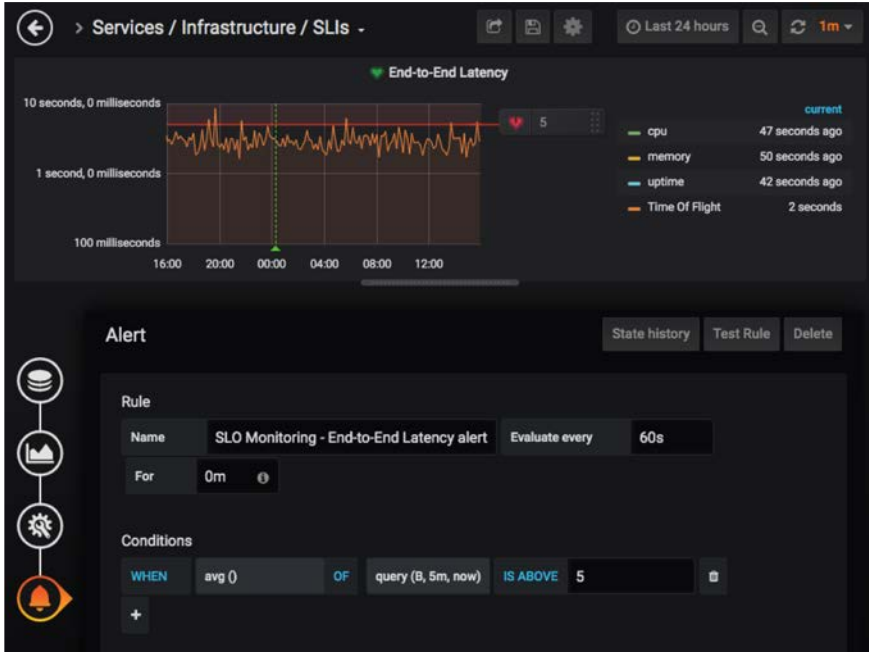
Dashboard / Grafana

- *Single Stat* panels
- Drill-down on details dashboards per category
- Threshold as SLO



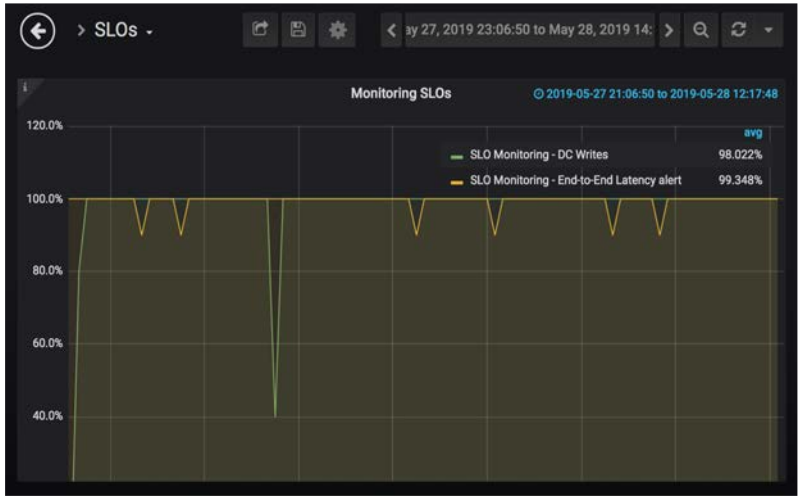
Tracking SLO

- Check SLI vs defined SLO
- Grafana Rules to generate alarms
- Webhook to HTTP



SLO Driven Operations

- Alert on SLO miss
  - Care about symptoms first
- Build Performance Trend
- SLO-Driven Availability



Technical Challenges

- Grafana Alert Engine
  - Rely on TSDB capabilities
  - Prometheus fairly advanced, InfluxQL has some limitation, Flux should solve
- Black-box vs White-box
  - white-box fits the usual metrics flow, black-box may benefits from common framework for probing

Non-technical challenges

- Service dependencies
  - *“Not my fault”*
- Big debate on user-related metrics as SLI
- Bottom-Up approach

Lessons Learned

- Dashboard-first approach works
- SLI & SLO are good starting points
- Cultural change
  - target people more than technology

Conclusion

- Successful migration to modern opensource monitoring stack and common practises
- SRE framework and culture proved to be a good direction for service operations evolution
- Just at the beginning of the SRE journey, looking forward for the next steps

CERN