Operational Intelligence

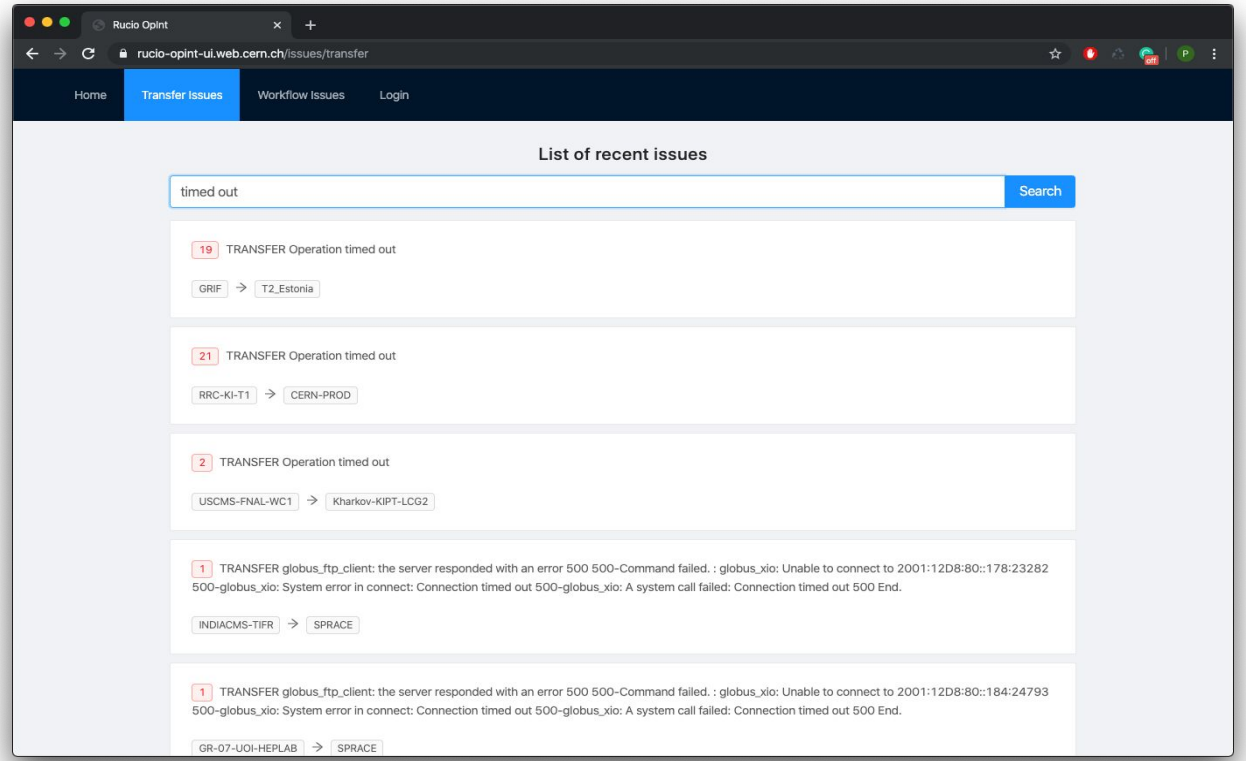
pre-GDB on operations efforts 24/2/22

Panos Paparrigopoulos on behalf of the OpInt team

Some background

- OpInt activity aims to improve resource utilization and minimise human effort on operations.
- A discussion forum and a development effort.
- Collaborative approach: cross-experiment activity.
  - Common operations for common tools can save resources.
- Data transfers (Rucio/FTS) seemed like a very good place to start.
- Multiple approaches where tried.
- Activities didn’t only limit to data transfer operations.

Initial approach - logs aggregation



Something smarter: Log clustering

Language Model: Word2Vec

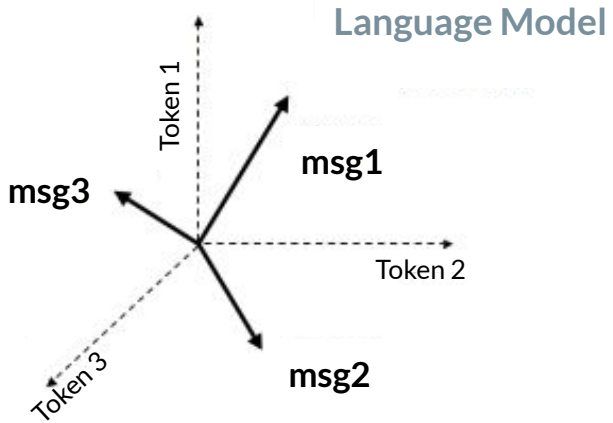
We train a language model to represent message tokens in a convenient way. The message embedding is then retrieved by combining the representations of the tokens they are made of.

Raw message

TRANSFER globus_ftp_client: the server responded with an error 500 Command failed. : Error registering globus write error: Too many links

Tokens

[“tranfer”, “globus_ftp_client”, “server”, “responded”, “error”, “500”, “command”, “failed”, “error”, “registering”, “globus”, “write”, “error”, “too”, “many”, “links”]



Clustering: K-Means

After we have the representation, then we use it for clustering.

Random initialisation of centroids

Choice of K optimised based on WSSE[1] and ASW[2] → K=12

Single Linkage with cosine distance

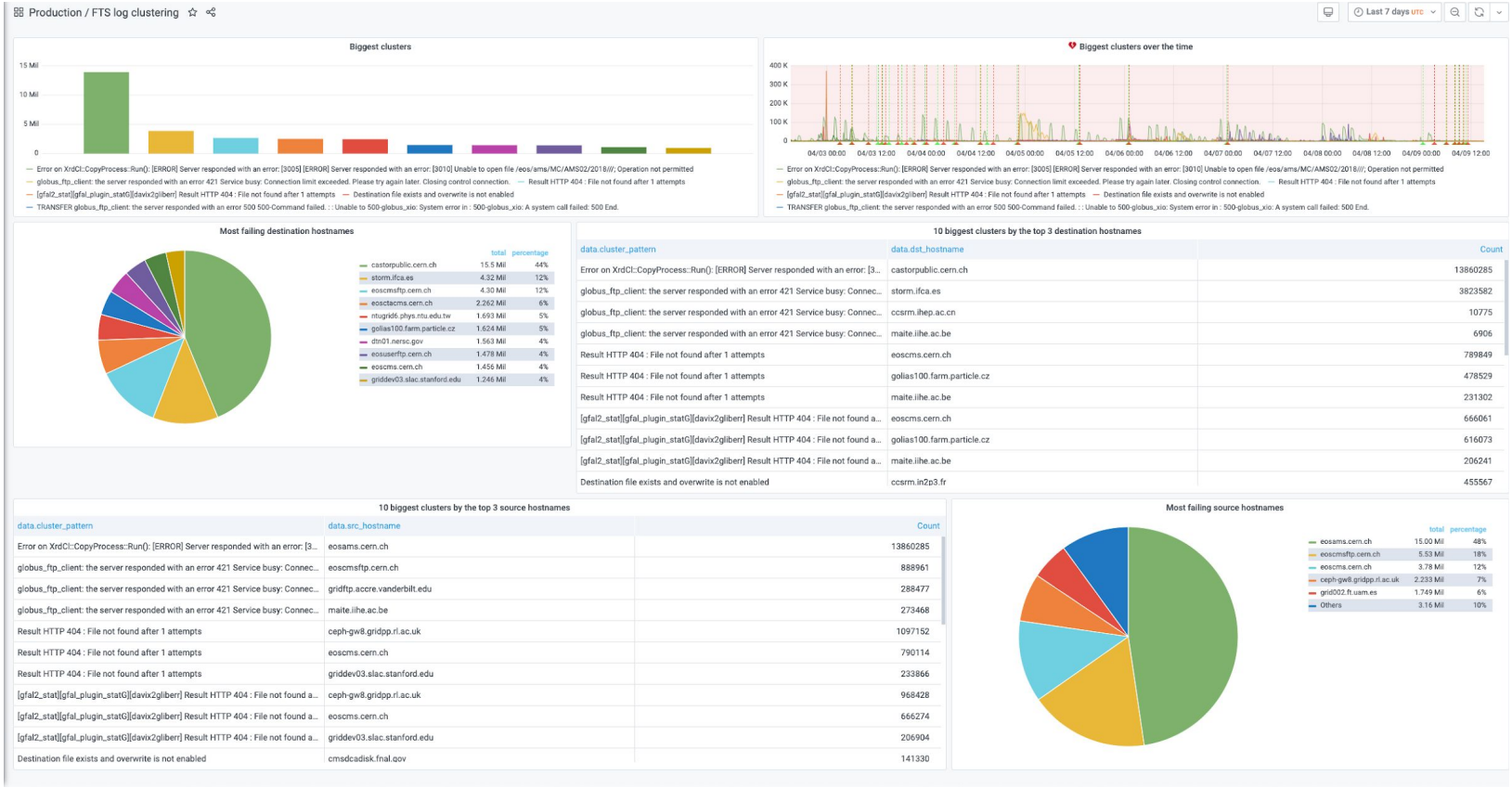
cluster_label	n_unique_mess
1	77697
0	30722
4	20122
3	17255
5	16214
2	4546
9	3500
8	1752
11	1003
6	623
10	323
7	163

message abstraction

cluster_label	n_unique_mess
8	123
3	74
0	65
2	62
4	60
9	52
11	49
7	31
1	24
10	16
6	12
5	8

Progressive ID	Cluster ID	n_messages	unique_strings	unique_patterns	top_5_msg	top_5_src	top_5_dst
2	11	89777	5144	101	msg': 'transfer globus_ftp_client the server responded with an error 500 500 command failed 500 an unknown error occurred 500 end', 'n': 79454, 'n_perc': 0.885	src': 'UKI-NORTHGRID-LANCS-HEP', 'n': 19807, 'n_perc': 0.2206	dst': 'UKI-NORTHGRID-LIV-HEP', 'n': 79537, 'n_perc': 0.8859
					msg': 'transfer globus_ftp_client the server responded with an error 451 general problem problem while connected to \\$ADDRESS input/output error', 'n': 1689, 'n_perc': 0.0188	src': 'UKI-LT2-QMUL', 'n': 19024, 'n_perc': 0.2119	dst': 'AGLT2', 'n': 1828, 'n_perc': 0.0204
					msg': 'transfer globus_ftp_client the server responded with an error 451 general problem failed to connect \\$IPv6 connection timed out', 'n': 1661, 'n_perc': 0.0185	src': 'UKI-NORTHGRID-MAN-HEP', 'n': 18768, 'n_perc': 0.2091	dst': 'RAL-LCG2', 'n': 1264, 'n_perc': 0.0141
					msg': 'transfer globus_ftp_client the server responded with an error 451 internal timeout', 'n': 963, 'n_perc': 0.0107	src': 'UKI-LT2-RHUL', 'n': 16030, 'n_perc': 0.1786	dst': 'NDGF-T1', 'n': 1050, 'n_perc': 0.0117
					msg': 'transfer globus_ftp_client the server responded with an error 500 command failed ipc failed while attempting to perform request', 'n': 822, 'n_perc': 0.0092	src': 'BNL-ATLAS', 'n': 2759, 'n_perc': 0.0307	dst': 'BNL-ATLAS', 'n': 997, 'n_perc': 0.0111

Clustering results in Grafana



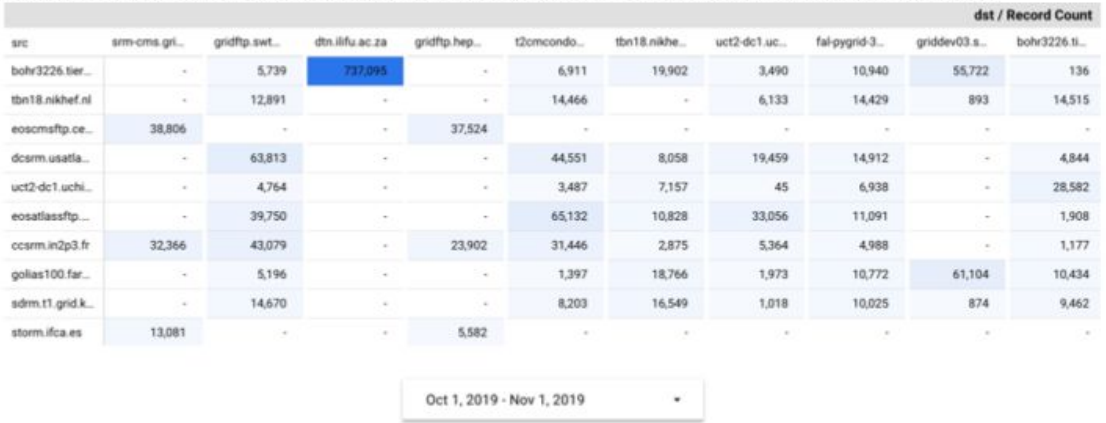
https://monit-grafana.cern.ch/d/Zx_bXneWz/fts-log-clustering?from=now-7d&orgId=11&to=now

Anomaly detection on transfers

An interesting find was that error distribution not only varied over time, but also over the interconnections between nodes.

Given the observed changes in error distribution across time, connection graph and content (as represented by the error categories), we investigated graph anomaly detection algorithms as a possible way to identify patterns in the logs.

MIDAS (MIcrocluster-based Detector of Anomalies in Streams) seemed a good fit:
- It finds anomalies in dynamic graphs (such as those generated by file transfers, but also
intrusions)
- It detects micro-clusters (sudden “burst” of connections between nodes, such as those
that may occur with multiple retrials, but also denials of service)
- Memory usage is constant and independent of graph size
- Update time in streaming scenarios is also constant



dst / Record Count											
src	srm-cms.gri...	gridftp.swt...	dtn.ilifu.ac.za	gridftp.hep...	t2cmcondo...	tbn18.nikhe...	uct2-dc1.uc...	fal-pygrid-3...	griddev03.s...	bohr3226.ti...
bohr3226.tier...	-	5,739	737,095	-	6,911	19,902	3,490	10,940	55,722	136
tbn18.nikhef.nl	-	12,891	-	-	14,466	-	6,133	14,429	893	14,515
eoscmsftp.ce...	38,806	-	-	37,524	-	-	-	-	-	-
dcsrm.usatla...	-	63,813	-	-	44,551	8,058	19,459	14,912	-	4,844
uct2-dc1.uchi...	-	4,764	-	-	3,487	7,157	45	6,938	-	28,582
eosatlassftp....	-	39,750	-	-	65,132	10,828	33,056	11,091	-	1,908
ccsrm.in2p3.fr	32,366	43,079	-	23,902	31,446	2,875	5,364	4,988	-	1,177
golias100.far...	-	5,196	-	-	1,397	18,766	1,973	10,772	61,104	10,434
sdrm.t1.grid.k...	-	14,670	-	-	8,203	16,549	1,018	10,025	874	9,462
storm.ifca.es	13,081	-	-	5,582	-	-	-	-	-	-

Oct 1, 2019 - Nov 1, 2019

Figure 3: Count of errors over connection pairs

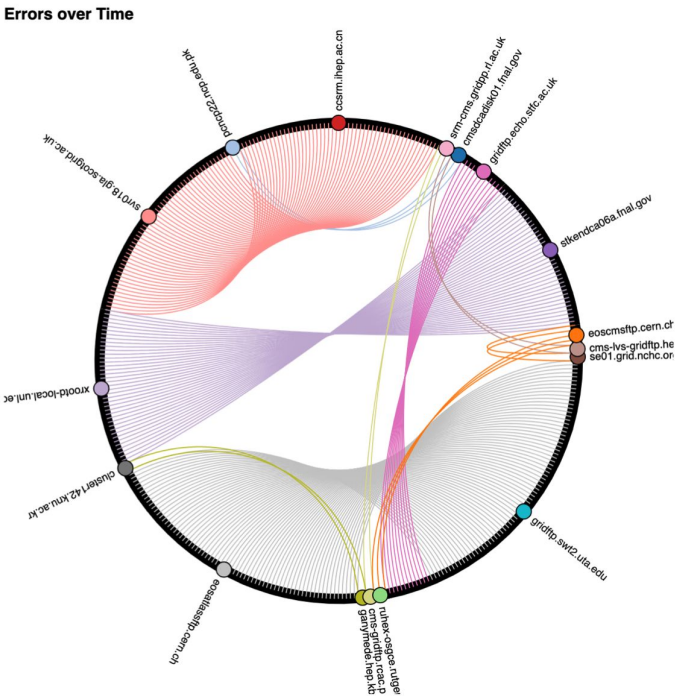
Start_Hour / Record Count											
Top 10 - dat...	Top 10 - data...	201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 10, 201...	Oct 1(
bohr3226.tier...	dtn.ilifu.ac.za	4,106	3,450	3,511	4,215	4,636	3,411	3,155	3,782	4,600	
	griddev03.sla...	2	-	-	-	-	-	-	-	-	
	serv02.hep.p...	183	163	143	171	207	155	171	210	195	
	tbn18.nikhef.nl	50	55	51	49	43	211	20	7	25	
	fal-pygrid-30.l...	32	38	34	29	27	25	14	26	62	
	f-dpm000.gri...	27	32	26	28	25	398	3	2	5	
	ftp1.ndgf.org	26	29	26	28	23	395	3	-	-	
	sdrm.t1.grid.k...	25	28	27	28	25	201	3	-	-	
	dcache-atlas-...	26	29	26	26	26	323	3	-	-	
	xrootd.echo.s...	23	29	29	26	21	202	-	-	-	

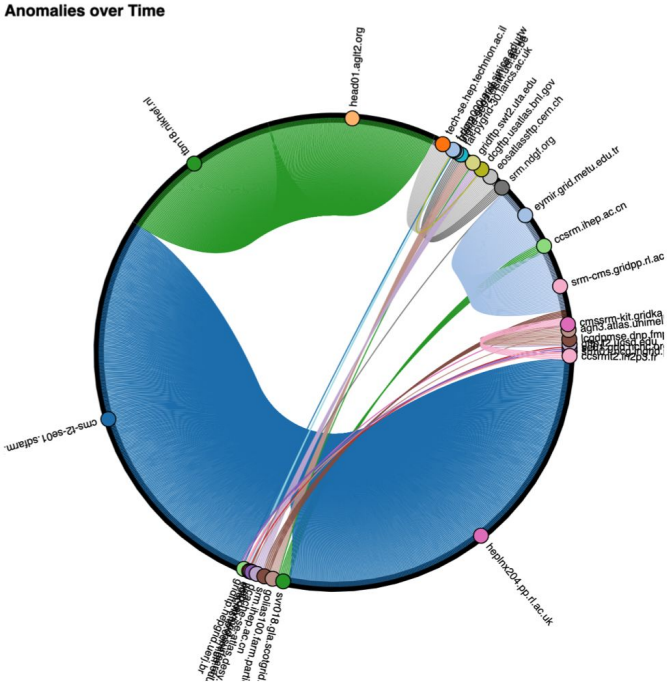
Figure 4: Variation over time for a given connection pair

https://arxiv.org/pdf/1911.04464.pdf

Anomaly detection on transfers

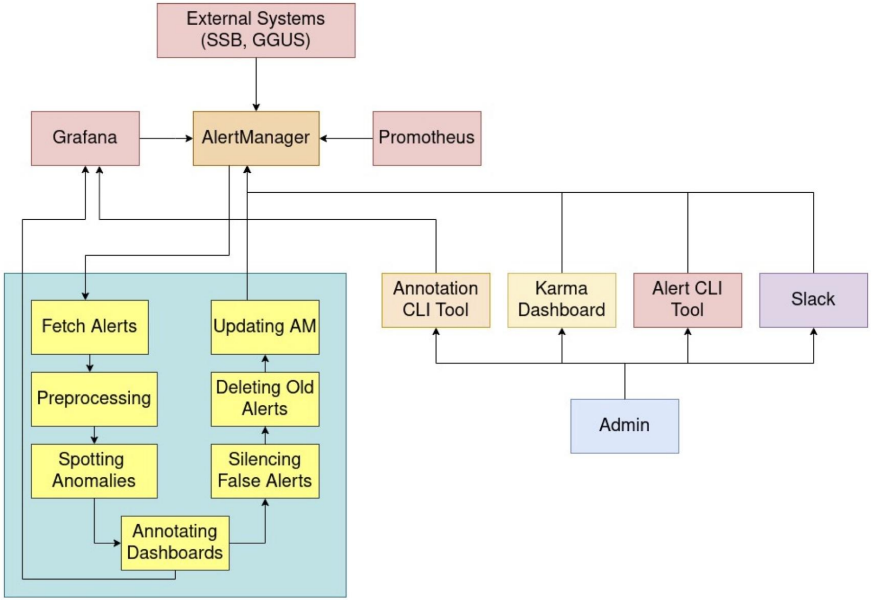
With MIDAS we were able to find the most anomalous connections on a graph in a given time window and monitor the evolution of those connections over time.





Intelligent Alert system

- CMS developed an intelligent layer in their infrastructure to **detect, analyze and predict abnormal system behaviors** using the **alerts** produced by the infrastructure.



- The alert manager fetches the existing alerts, filters them, and **annotates Grafana** dashboards based on the alert tag
- SSB and GGUS are also integrated into the Alert Manager
- The system provides useful insights about when outages happen and how they affect the productivity reported by various systems in CMS dashboards
- Using **open source tools** makes this effort experiment-agnostic

More info: http://cern.ch/go/cxq8

Jobs buster

- ATLAS " Jobs Buster" tries to spot **operational problems** in submitted jobs
- **Machine Learning** is used to cluster the errors and then find the **common denominator** between failed jobs in the cluster (could be software version, site name, transfer src/dst etc)

More info: http://cern.ch/go/8qwC

What we learned - Pros

- A lot of work/brainstorming was done.
- People from different experiments were brought together and open source technologies and community standards were used.
- We now know that we can definitely improve operations and we have to make sure that we can scale them as our resources grow.
- We have the infrastructure to analyse and present the information. This is a very solid ground to build on top.

What we learned - Cons

- Building trust in ML solutions and implement them in production is not easy.
- The lack of annotated datasets strongly limits the capability to validate our solutions.
- We didn’t manage to involve the experiment operations teams in the efforts.
- We should probably take a step back and start simpler.

What we learned - Outlook points

- Are operation teams interested in spending some time/effort to build common operation models/strategy/tools?
- Are experiments interested in shared operations? For example at shared sites, or for shared frameworks like Rucio/FTS?
    - If not shared operations, shared tools on which we can evolve concrete automation?
- If yes, then OpInt could be the starting point for co-develop solutions which could be concretely useful to the experiments.
- We could adapt the format of the forum as needed, e.g. from a general one to one that focus on some common aspects (monitoring, k8s, agile, dev ops etc…)

What next?

- We know that problems will appear as the infrastructure grows and we want to make sure that the efficiency of operations will scale accordingly.
- We are questioning whether we should continue, and if yes, how?
- We can probably find manpower but we need some commitment from the experiments, at least providing some ideas and guidance.
- We had a chat with FTS in some possible developments, more details in the next talk.
- We hope you will help us reevaluate our strategy and decide what we should do next.