Handout NOTED Presentation

This is a summary of the status report on the NOTED project at the LHCONE/LHCOPN meeting in Umeå 2019, and meant as complementary material to the slides.

The following sections present an executive summary, the motivation, an overview, completed steps, next steps, open questions (some of which are directed at the audience) and ideas for the future.

Executive Summary

NOTED (Network-Optimized Transfer of Experimental Data) seeks to facilitate and utilize SDN to optimize the file transfers in WLCG by i) aggregating information on the currently running and upcoming file transfers, ii) providing said information on a website and via a REST API/json file, iii) allowing network operators to act on the information as well as implementing actions within our scope (i.e. within CERN).

We are working on a prototype and thus actively seeking input from future users to optimize the system for their wishes and needs (see: Open Questions).

Motivation

In the following, we show the motivation behind NOTED.

Due to the luminosity upgrade, we expect even higher bandwidth usage (peaks) for file transfers. To provide smooth and fast file transfers in the future, we need to traffic engineer the network beneath, e.g. via Bandwidth-on-Demand (BoD) mechanisms or by load balancing over several links with automatically tuned BGP metrics.

This approach was named NOTED (Network-Optimized Transfer of Experimental Data).

The next section gives an overview over this data-driven approach.

Overview

NOTED seeks to facilitate and utilize SDN to optimize the file transfers in WLCG by i) aggregating information on the currently running and upcoming file transfers, ii) providing said information via a human and machine interface, i.e. on a website and via a REST API/json file, iii) allowing network operators to act on the information as well as implementing actions within our scope (i.e. within CERN).

The following two paragraphs explain the completed and next steps for each of the mentioned three parts: i) the information aggregation, ii) the interface, iii) the actions.

Completed Steps

This section presents the completed steps for the aggregated file transfer information, the interface to access the information, and the actions to optimize the network, respectively.

Information Aggregation

We need two types of information to facilitate the desired traffic engineering: a) an estimate of the current/upcoming total file transfer size for each source-destination pair, and b) what IP prefixes are associated with each source and destination. The combined data – how much data will be/is being transferred between what IP prefixes – can then be stored in a separate storage instance. Each of those three points are explained in more detail below.

File Transfer Information We programmatically collect monitoring statistics from the FTS monitoring instance via a Grafana proxy [*fts_grafana*].

IP Prefix Information We use the information in LHCONE and LHCOPN twiki tables [*lhcone_ip_prefixes*] [*lhcopn_ip_prefixes*] as a starting point. To allow for a more computer-friendly access, a database called AGIS [*agis*] now contains the same information. AGIS is the ATLAS Grid Information System, and, as the name suggests, is a central information system from the ATLAS experiment. In the future, it will be put into CRIC [*cric*]: At the moment, CRIC is the CMS equivalent of AGIS, but in the future, it is intended for wide, experiment-agnostic usage. Since both CRIC and AGIS already contain information on storage endpoints, the IP prefix information fits well in there.
The *IP prefix information will have to be verified and updated by the sites*, since it is currently partially out of date. The incentive behind keeping these entries up to date is the possibility of optimized file transfers.

Combined Information To store the combined file transfer and IP prefix information, we create a new database; how its content can be accessed is described in the next paragraph.

Interface

The goal is to provide two interfaces: One for humans and one for machines.

Human Interface The human interface is a website plotting the current file transfer information; a prototype for this part already exists in form of Grafana and Kibana dashboards [*grafana*] [*kibana*]. At the moment, these dashboards show FTS information; as a next step, we plan on integrating the IP prefix information. The dashboards plot the expected file transfer size for each source-destination pair, based on the FTS jobs submitted in the last 10 minutes, and variations thereof (per VO, per FTS server, only jobs in SUBMITTED state; bar plots, heat maps, stacked plots). The dashboards automatically update themselves every few minutes, this can be adapted by the user as desired. Furthermore, Kibana and Grafana natively offer alarm triggers; as an example, this could generate a warning/logging e-mail (depending on whether the following action is manual or SDN-based) whenever there is a transfer above a custom threshold.

Machine Interface For the machine interface, please see Interface in Next Steps.

Traffic Engineering Actions

Last year, Edoardo Martelli presented two Proof of Concepts for manual traffic engineering to improve file transfers with two techniques: load balancing with BGP metric adaptations [*bgp_te*] and Bandwidth-on-Demand [*bod_te*]. This motivated the construction of a more widely deployable prototype for automated traffic engineering via SDN.

Next Steps

This section outlines future steps, again to aggregate, display and act on file transfer information.

Information Aggregation

The InfluxDB database to store the aggregated information now exists; the Python implementation to edit its entries automatically is ongoing.

Interface

We have some concrete next steps both for the human and machine interface:

Human Interface For the human interface, the next step is to enrich the plots with the IP prefix information. Furthermore, we plan on adding plots for currently ongoing transfers and what percentage of them is completed; this avoids the uncertainty of predictions, and still allows enough time to act, since the big transfers take a lot of time.

Machine Interface For the machine interface, we are working on a design where the information is publicly accessible (also by network operators who are not associated with CERN), without risking that the database is overloaded. For the case that we still need authentication for some aspects of the functionality, we are seeking input on preferred systems (see Open Questions). After this, the goal is to choose a sensible interface specification (with active input from future users) and to document it for users. Currently, we are considering a REST API or simply a json file to provide the information.

Traffic Engineering Actions

We intend to implement SDN for automatic load balancing within CERN, as an example application of the information provided by the transfer broker. We have not selected a specific technique yet, this part is still very open at the moment, see SDN in Ideas for the Future.

Open Questions

There are open questions that are project-internal, and open questions that are directed at the future users; the next paragraphs describe both types in more detail.

Project-Internal Open Questions

- How much added value is there in file transfer *predictions* compared to information on *ongoing* transfers?
  - If they would be very useful, *how* can we achieve accurate predictions?
  - What accuracy can we achieve, and what guarantees can we make?
- What type of SDN methods are the ideal choice for the CERN network? What impact do they have?

Questions for Future Users

- What sort of IP prefix storage management and authentication system do you like?
- What should the machine interface look like?
- What information are you interested in?
  For example, through the FTS monitoring, we can access information on the experiment, the FTS server, the maximum number of retries, the task name, the job state, user, timestamps for staging and submission, the client, file state, request type... if there is some other information you would like to have, please let us know and we will investigate the options.
- What sort of traffic engineering actions would you like to perform?

Ideas for the Future

We have a collection of ideas for the steps after the transfer broker is fully deployed – that is, the Software-Defined Networking aspect, refinements for the predictions of the file transfers, and possibilities to enhance the statistics and plots with more information and formats.

SDN

For the Software-Defined Networking, we consider segment routing [*seg_route*], implementing some part of the SDN in the data plane (e.g. with P4 [*p4*]), generalizing the approaches that were used in the manual traffic engineering PoCs, and are looking for even more options.

Refined Predictions of File Transfers

Additionally to information on currently ongoing transfers, we seek to have accurate predictions about future transfers. There are two types of prediction refinements that play a key role in effective traffic engineering measures: a) the precision of what IP prefixes are going to be affected, and b) when and for what duration they will be affected.

IP Prefix Prediction To narrow down the range of IP prefixes that might be affected by a transfer, we consider to estimate the probability with which each storage element will be used, based on its storage history (idea from Mario Lassnig). As an example, if a lot of space was freed on a specific storage element, this storage element is more likely to be used for an upcoming transfer compared to more full storage elements.

Transfer Time Predictions To know for what timespan the network should be optimized, we need a prediction of the time a file spends in the Rucio [*rucio*] queue, FTS queue and on the network. Joaquín Bogado is doing a PhD project on this topic.

More Statistics, more Plots

There are a lot of interesting approaches to visualize the data that remain to be explored; both for short-term and long-term monitoring coverage. Also, the data could be further enriched, e.g. with additional FTS statistics.

Short-Term Plots To add the geographical information intuitively, one could display the upcoming transfers on a world map. Moreover, if we also plot information on *future* transfers (additionally to *ongoing* transfers), plots for each file state – in the Rucio queue, FTS queue and on the network – could be of interest.

Long-Term Plots If we also use file transfer predictions, plots to quantify the prediction quality would be interesting (input from Joe Mambretti at the presentation). Further ideas for future work are plots for what source-destination combinations are common, how bursty the traffic is, and especially how well the traffic engineering measures worked.

Additional FTS statistics The FTS monitoring is rich in statistics which could be furhter leveraged in the future. One example are pre-staging statistics: These statistics affect file transfers where the file has to be first fetched from tape, which affects the expected time to complete.

Contact

We are delighted to receive questions, feedback and comments:

edoardo.martelli@cern.ch

coralie.busse-grawitz@cern.ch

Sources

[fts_grafana] *https://monit-grafana.cern.ch/d/000000670/fts-servers-dashboard?orgId=25&from=1560204000000&to=1560290399999&var-group_by=vo&var-vo=All&var-vo_es=All&var-source_se=All&var-source_se_es=All&var-src_country=All&var-dst_country=All&var-src_site=All&var-dst_site=All&var-dest_se=All&var-dest_se_es=All&var-fts_server=All&var-fts_server_es=All&var-bin=1h&var-dest_hostname=All&var-source_hostname=All*, last visited: 11. June 2019

[lhcone_ip_prefixes] *https://twiki.cern.ch/twiki/bin/view/LHCOPN/LhcopnIpAddresses#AS_numbers_and_LHC_prefixes*, last visited: 12. June 2019

[lhcopn_ip_prefixes] *https://twiki.cern.ch/twiki/bin/view/LHCOPN/LhcopnIpAddresses#AS_numbers_and_LHC_prefixes*, last visited: 12. June 2019

[agis] *http://atlas-agis.cern.ch/agis/*, last visited: 11. June 2019

[cric] *http://cms-cric.cern.ch/*, last visited: 11. June 2019

[bod_te] *https://indico.cern.ch/event/825233/contributions/3451803/attachments/1853951/3044480/NOTED-20190311-NLT1-GEANT-BoD-test.pdf*, last visited: 12. June 2019

[bgp_te] *https://indico.cern.ch/event/725706/contributions/3169200/attachments/1744659/2824103/LHCOPNE-20181031-FNAL-CERN-NLT1-test.pdf*, last visited: 12. June 2019

[kibana] *https://monit-kibana.cern.ch/kibana/app/kibana::/dashboard/Awq_fXKqNqzpRjqCuVZ_?_g*=(), last visited: 12. June 2019

[grafana] *https://monit-grafana.cern.ch/d/G5oAZRZZk/fts-noted?orgId=25&refresh=5s&from=1560345236986&to=1560345836986&var-group_by=vo&var-vo=All*, last visited: 12. June 2019

[seg_route] *https://www.segment-routing.net/*, last visited: 12. June 2019

[p4] *https://p4.org/*, last visited: 17. June 2019

[rucio] *https://rucio.cern.ch/*, last visited: 17. June 2019