CERN-NLT1 load balancing over LHCOPN and LHCONE

- Test report -

LHCONE meeting at Fermilab
31st November 2018
edoardo.martelli@cern.ch

Goals

- Proof of concept: load-balancing Tier0-Tier1 traffic over LHCOPN and LHCONE links when LHCOPN link is congested

- Long term: optimize network utilization in case of congestion of primary path

Dynamic load balancing with BGP

1st test on 4th of September

Goals:

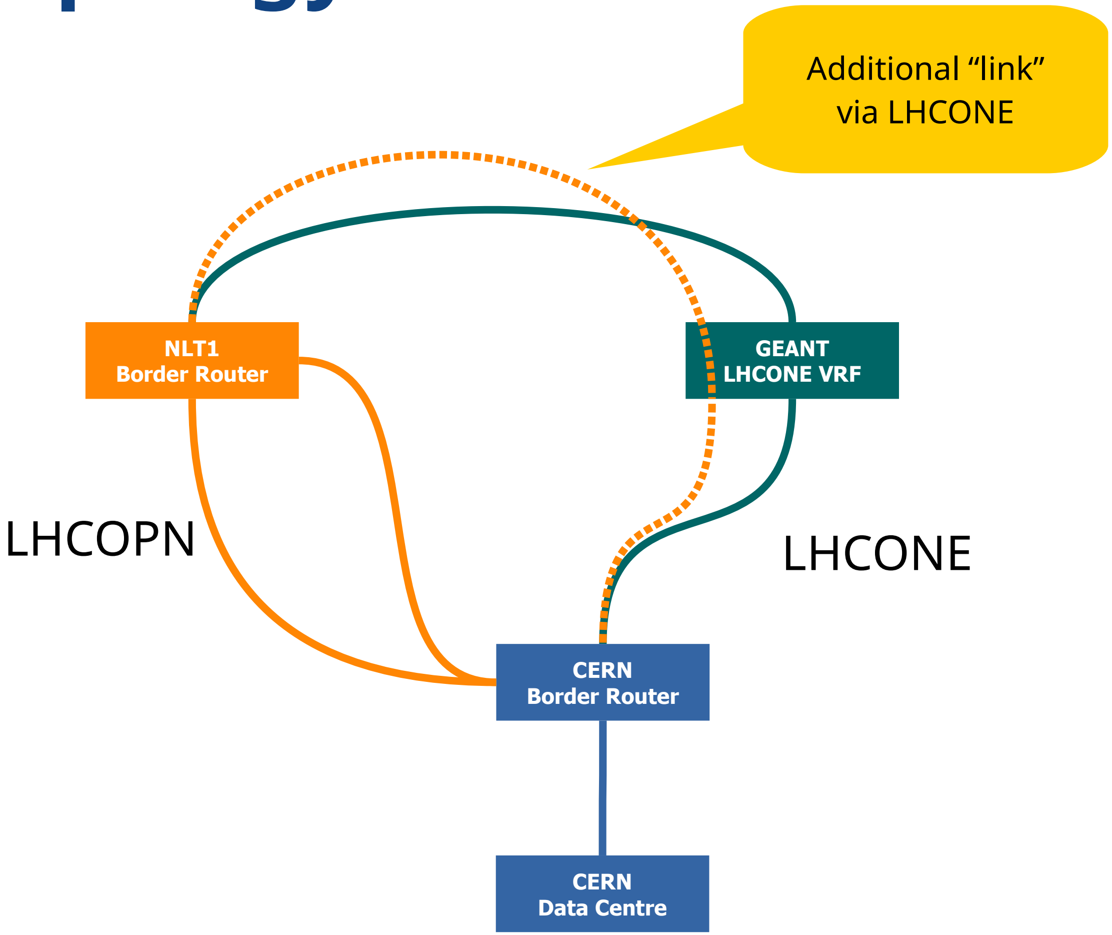
- Increase available bandwidth by manipulating routing metrics

Test:

- Adjust BGP metric on CERN router to load-balance NL-T1 prefixes over LHCOPN and LHCONE link
- No changes on the NL-T1 side: transfers keep flowing in asymmetric routing

Tested in production network during maintenance window; there was no visible effect because links were under utilized.

Network topology



Before any change – CERN LHCOPN router

```
telnet@L513-E-RBRXL-2#sh ip bgp 145.100.32.0/22
Number of BGP Routes matching display condition : 6
Status codes: s suppressed, d damped, h history, * valid, > best, i internal  x:best-external
Origin codes: i - IGP, e - EGP, ? - incomplete
    Network              Next Hop        MED    LocPrf     Weight Path
*>x 145.100.32.0/22    192.16.166.74   10         100        100     1162 I         <== LHCOPN NLT1 peer 1
*   145.100.32.0/22    192.16.166.82   10         100        100     1162 I         <== LHCOPN NLT1 peer 2
*   145.100.32.0/22    192.16.166.86   10         100        100     39590 1162 i
*   145.100.32.0/22    62.40.126.217   100        100        100     20965 1162 I   <== GEANT LHCONE
*   145.100.32.0/22    198.124.80.69   100        100        100     293 20965 1162 i
*   145.100.32.0/22    192.65.183.5    100        100        100     20641 17579 20965 1162 i
       Last update to IP routing table: 16h50m20s, 2 path(s) installed:
       Route is advertised to 19 peers:                                           <== T1-T1 transit vie CERN
        192.16.166.90(43)                        192.16.166.86(39590)              192.16.166.178(2875)
        192.16.166.162(59624)                    192.16.166.158(59624)             192.16.166.142(43475)
        192.16.166.70(43475)                     192.16.166.66(43475)              192.16.166.98(24167)
        192.16.166.150(24167)                    192.16.166.82(1162)               192.16.166.42(789)
        192.16.166.38(58069)                     192.16.166.30(3152)               192.16.166.18(137)
        192.16.166.10(43)                        172.24.46.2(513)                  172.24.46.3(513)
        172.24.46.4(513)

telnet@L513-E-RBRXL-2#sh ip route 145.100.32.0/22
Type Codes - B:BGP D:Connected I:ISIS O:OSPF R:RIP S:Static; Cost - Dist/Metric
BGP  Codes - i:iBGP e:eBGP
ISIS Codes - L1:Level-1 L2:Level-2
OSPF Codes - i:Inter Area 1:External Type 1 2:External Type 2 s:Sham Link
STATIC Codes - d:DHCPv6
        Destination        Gateway         Port           Cost          Type Uptime src-vrf
1       145.100.32.0/22    192.16.166.74   ve 3503        20/10         Be   16h50m -
        145.100.32.0/22    192.16.166.82   ve 2904        20/10         Be   16h50m -
```

Load balancing of the two BGP routes

Change of LHCOPN route-map

====> Added entry to LHCOPN-IN route map (used for all peerings with Tier1s) to catch NLT1 prefixes

```
route-map LHCOPN-IN permit 5
 match ip address prefix-list PL-NLT1
 match as-path ASP-NLT1
 set weight 100
 set local-preference 1000
 set metric 5
```

AS path not changed after this first step

```
telnet@L513-E-RBRXL-2#sh ip bgp 145.100.17.0/28
    Network            Next Hop         MED    LocPrf     Weight Path
*>x 145.100.17.0/28    192.16.166.74    5          1000       100    1162 i
*   145.100.17.0/28    192.16.166.82    5          1000       100    1162 i
*   145.100.17.0/28    192.16.166.86    10         100        100    39590 1162 i
*   145.100.17.0/28    62.40.126.217    100        100        100    20965 1162 i
*   145.100.17.0/28    198.124.80.69    100        100        100    293 20965 1162 i
*   145.100.17.0/28    192.65.183.5     100        100        100    20641 17579 20965 1162 i
       Last update to IP routing table: 0h1m16s, 2 path(s) installed:
       Route is advertised to 19 peers:
        192.16.166.90(43)                        192.16.166.86(39590)              192.16.166.178(2875)
        192.16.166.162(59624)                    192.16.166.158(59624)             192.16.166.142(43475)
        192.16.166.70(43475)                     192.16.166.66(43475)              192.16.166.98(24167)
        192.16.166.150(24167)                    192.16.166.82(1162)               192.16.166.42(789)
        192.16.166.38(58069)                     192.16.166.30(3152)               192.16.166.18(137)
        192.16.166.10(43)                        172.24.46.2(513)                  172.24.46.3(513)
        172.24.46.4(513)
```

Change of LHCONE route-map

```
====> Added entry to LHCONE-IN route map to catch NLT1 prefixes

route-map LHCONE-IN permit 5
 match ip address prefix-list PL-NLT1
 match as-path ASP-NLT1
 set weight 100
 set local-preference 1000
 set metric 5
 set community  20641:20641


telnet@L513-E-RBRXL-2#sh ip bgp 145.100.32.0/22
    Network            Next Hop        MED    LocPrf     Weight Path
*>x 145.100.32.0/22    192.16.166.74   5          1000       100    1162 i
*   145.100.32.0/22    192.16.166.82   5          1000       100    1162 i
*   145.100.32.0/22    62.40.126.217   5          1000       100    20965 1162 i
*   145.100.32.0/22    192.16.166.86   10         100        100    39590 1162 i
*   145.100.32.0/22    198.124.80.69   100        100        100    293 20965 1162 i
*   145.100.32.0/22    192.65.183.5    100        100        100    20641 17579 20965 1162
       Last update to IP routing table: 0h2m22s, 2 path(s) installed:
       Route is advertised to 19 peers:


telnet@L513-E-RBRXL-2#sh ip route 145.100.32.0/22
        Destination        Gateway         Port           Cost          Type Uptime src-vr
1       145.100.32.0/22    192.16.166.74   ve 3503        20/5          Be   3m49s  -
        145.100.32.0/22    192.16.166.82   ve 2904        20/5          Be   3m49s  -
```

The NLT1 route from GEANT has now the same metrics, but still longer AS path

GEANT route not used for loadbalancing, yet

Configured multipath multi-as

```
====> Modified BGP behavior to loadbalance over routes with different AS paths

router bgp
   multipath multi-as


telnet@L513-E-RBRXL-2#sh ip route 145.100.32.0/22
        Destination        Gateway          Port           Cost          Type Uptime src-vrf
1       145.100.32.0/22    192.16.166.74    ve 3503        20/5          Be   0m40s  -
        145.100.32.0/22    192.16.166.82    ve 2904        20/5          Be   0m40s  -
```

Only 2 routes still.
Multipath-multi-as effective only on routes with the same AS path length

Same AS path

```
====> Modified LHCOPN-IN route map 5 to prepend 1 to NLT1 prefixes

route-map LHCOPN-IN permit 5
 [...]
 set as-path prepend 1162

telnet@L513-E-RBRXL-2#sh ip bgp 145.100.32.0/22
    Network            Next Hop        MED    LocPrf     Weight Path
*>x 145.100.32.0/22    62.40.126.217   5          1000       100    20965 1162 i
*   145.100.32.0/22    192.16.166.74   5          1000       100    1162 1162 i
*   145.100.32.0/22    192.16.166.82   5          1000       100    1162 1162 i
*   145.100.32.0/22    192.16.166.86   10         100        100    39590 1162 i
*   145.100.32.0/22    198.124.80.69   100        100        100    293 20965 1162 i
*   145.100.32.0/22    192.65.183.5    100        100        100    20641 17579 20965 1162 i
       Last update to IP routing table: 0h0m53s, 3 path(s) installed:
       Route is advertised to 3 peers:
        172.24.46.2(513)                         172.24.46.3(513)                         172.24.46.4(513)

====> Bad side effect: LHCONE prefix is now the best because of Next-Hop IP address, NL-T1 prefixes are no
longer advertised to the other LHCOPN Tier1s. No transit to other Tier1s via CERN for NLT1



telnet@L513-E-RBRXL-2#sh ip route 145.100.32.0/22
        Destination        Gateway         Port           Cost          Type Uptime src-vrf
1       145.100.32.0/22    62.40.126.217   ve 111         20/5          Be   1m20s  -
        145.100.32.0/22    192.16.166.74   ve 3503        20/5          Be   1m20s  -
        145.100.32.0/22    192.16.166.82   ve 2904        20/5          Be   1m20
```

Now there are 3 entries
with equal metrics
and AS path length

Success! GEANT route
now used for
load balancing!

Summary of configuration changes

```
! match only NL-T1 prefixes
ip prefix-list PL-NLT1 seq 5 permit 145.100.17.0/28
ip prefix-list PL-NLT1 seq 10 permit 145.100.32.0/22

! match only direct links (AS1162) and GEANT LHCONE (AS20965)
ip as-path access-list ASP-NLT1 seq 5 permit ^1162$|^20965 1162$

! allow load-balancing also on different AS paths
router bgp
   multipath multi-as

! best metrics on LHCOPN links and same AS path of the LHCONE access
route-map LHCOPN-IN permit 5
 match ip address prefix-list PL-NLT1
 match as-path ASP-NLT1
 set weight 100
 set local-preference 1000
 set metric 5
 set as-path prepend 1162

! LHCONE metrics match those of the LHCOPN route-map
route-map LHCONE-IN permit 5
 match ip address prefix-list PL-NLT1
 match as-path ASP-NLT1
 set weight 100
 set local-preference 1000
 set metric 5
 set community  20641:20641
```

Conclusions

- It is possible to load-balance traffic by only adjusting BGP metrics

- Load balancing can be applied to one side only; asymmetry on two network domains is acceptable (if not crossing statefull firewalls)

Load balancing stress test

2nd test on 18th of October

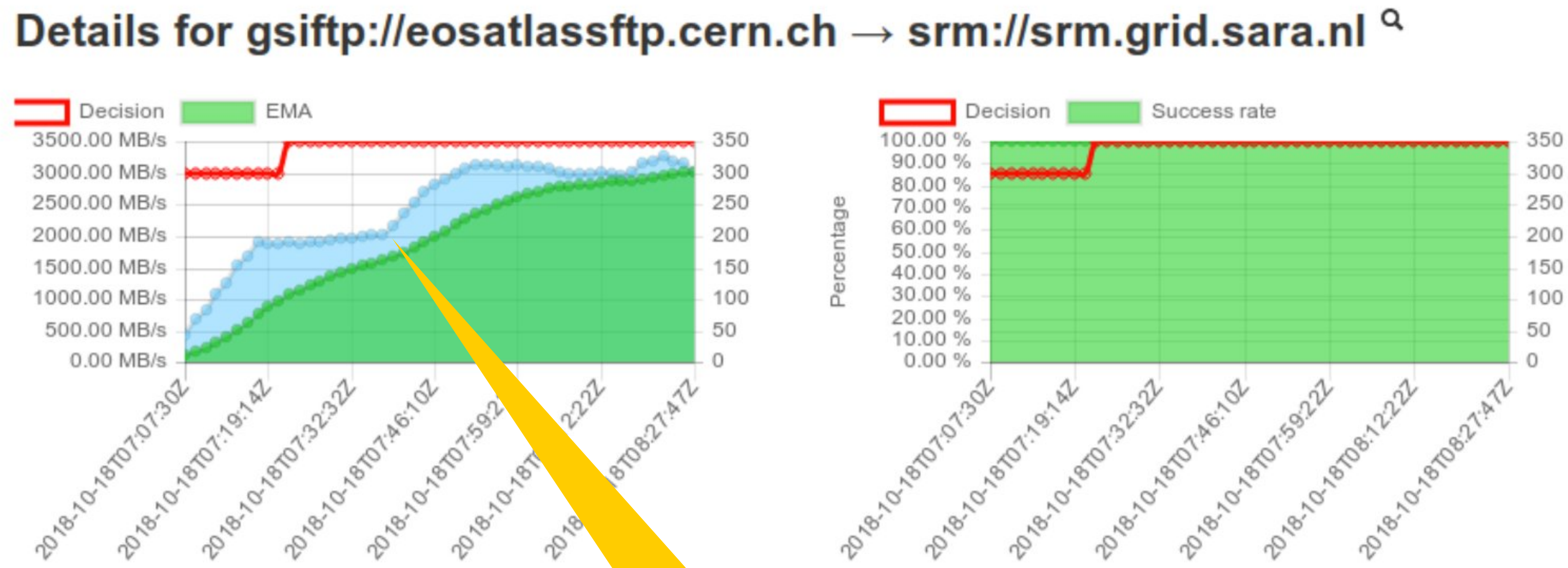
Goal:

- Apply load balancing in situation of LHCOPN links saturated and observe the effects

Test:

- ATLAS generated 30 TB of data to transfer from CERN to NL-T1 using Rucio (which relies on FTS, which uses EOS).
- After saturating the direct LHCOPN link, a third path via LHCONE was added to the load-balancing

Bandwidth gain for FTS

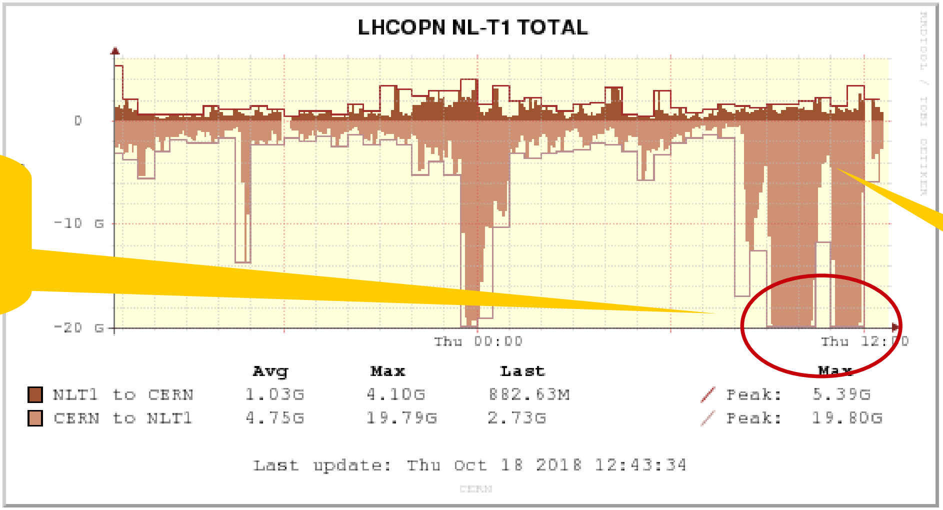


Details for gsiftp://eosatlassftp.cern.ch → srm://srm.grid.sara.nl

FTS transfer speed increase when added LHCONE link

CERN network side

NL-T1 2x10G LHCOPN links saturated

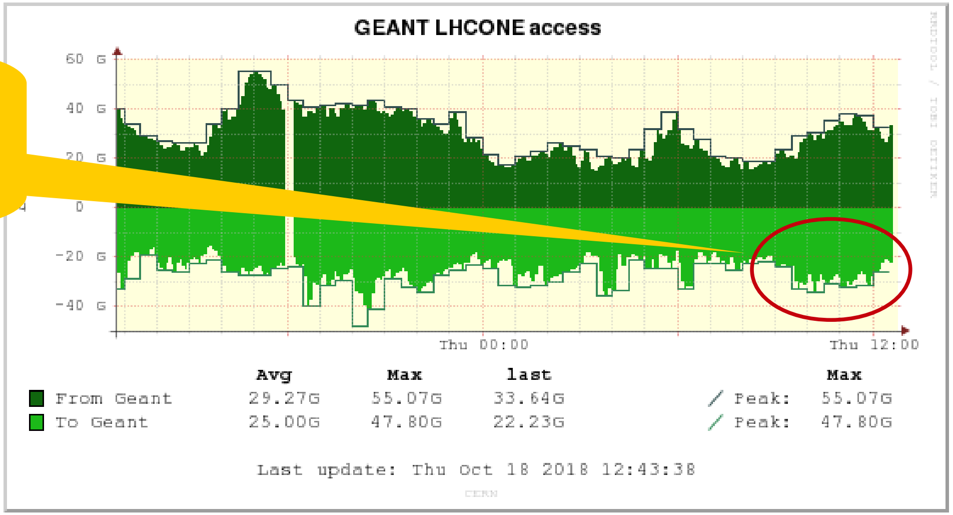


LHCOPN NL-T1 TOTAL

	Avg	Max	Last		Max
🟫 NLT1 to CERN	1.03G	4.10G	882.63M	Peak:	5.39G
🟫 CERN to NLT1	4.75G	19.79G	2.73G	Peak:	19.80G

Last update: Thu Oct 18 2018 12:43:34

CERN router went in software forwarding while trying some tricks to push more traffic on the LHCONE link



GEANT LHCONE access

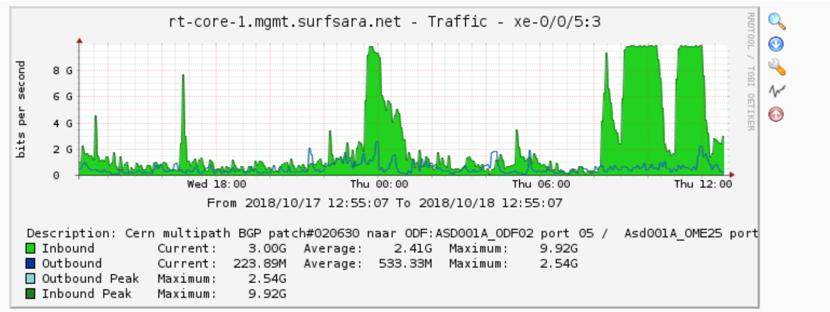
	Avg	Max	last		Max
🟩 From Geant	29.27G	55.07G	33.64G	Peak:	55.07G
🟩 To Geant	25.00G	47.80G	22.23G	Peak:	47.80G

Last update: Thu Oct 18 2018 12:43:38

No major impact on CERN LHCONE 100G access

NL-T1 network side

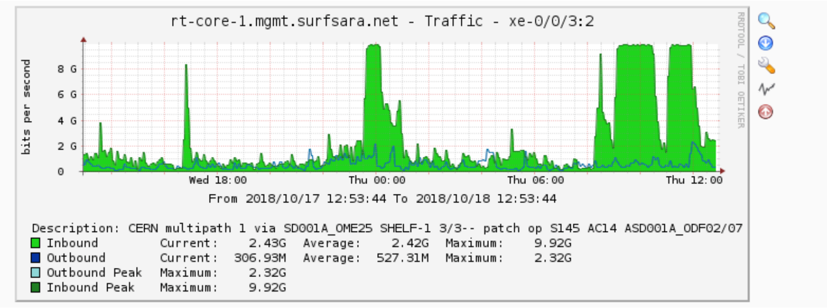
LHCOPN 10Gb link 1



```
From 2018/10/17 12:55:07 To 2018/10/18 12:55:07

Description: Cern multipath BGP patch#020630 naar ODF:ASD001A_ODF02 port 05 /  Asd001A_OME25 port
🟩 Inbound         Current:    3.00G  Average:    2.41G  Maximum:    9.92G
🟦 Outbound        Current:  223.89M  Average:  533.33M  Maximum:    2.54G
🟦 Outbound Peak   Maximum:    2.54G
🟩 Inbound Peak    Maximum:    9.92G
```

LHCOPN 10Gb link 2

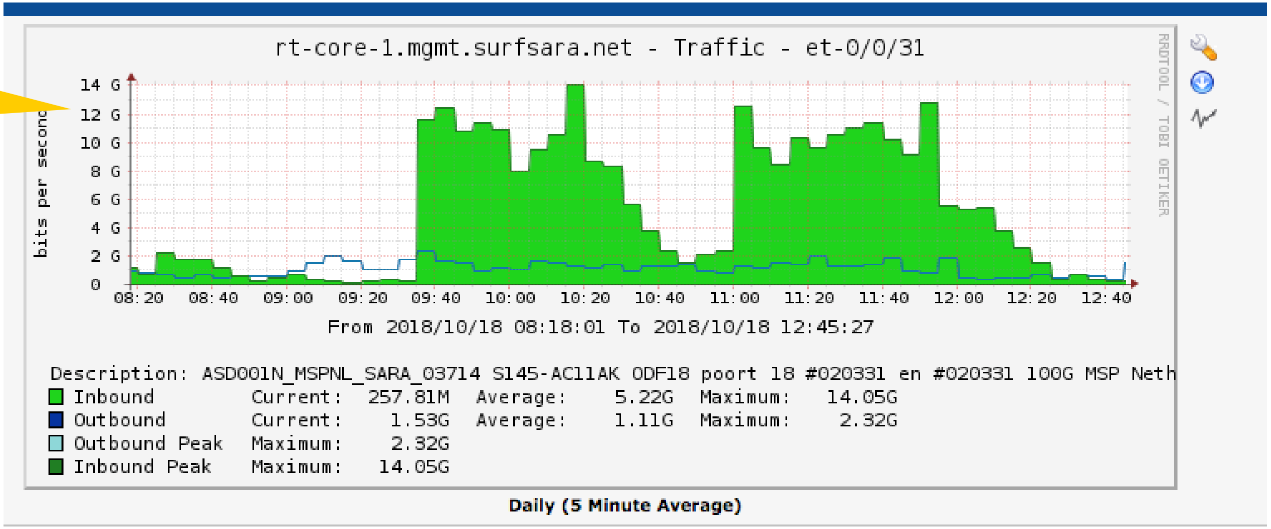


```
From 2018/10/17 12:53:44 To 2018/10/18 12:53:44

Description: CERN multipath 1 via SD001A_OME25 SHELF-1 3/3-- patch op S145 AC14 ASD001A_ODF02/07
🟩 Inbound         Current:    2.43G  Average:    2.42G  Maximum:    9.92G
🟦 Outbound        Current:  306.93M  Average:  527.31M  Maximum:    2.32G
🟦 Outbound Peak   Maximum:    2.32G
🟩 Inbound Peak    Maximum:    9.92G
```

60-70% load on the 20Gb NL-T1 LHCONE access

LHCONE 20G

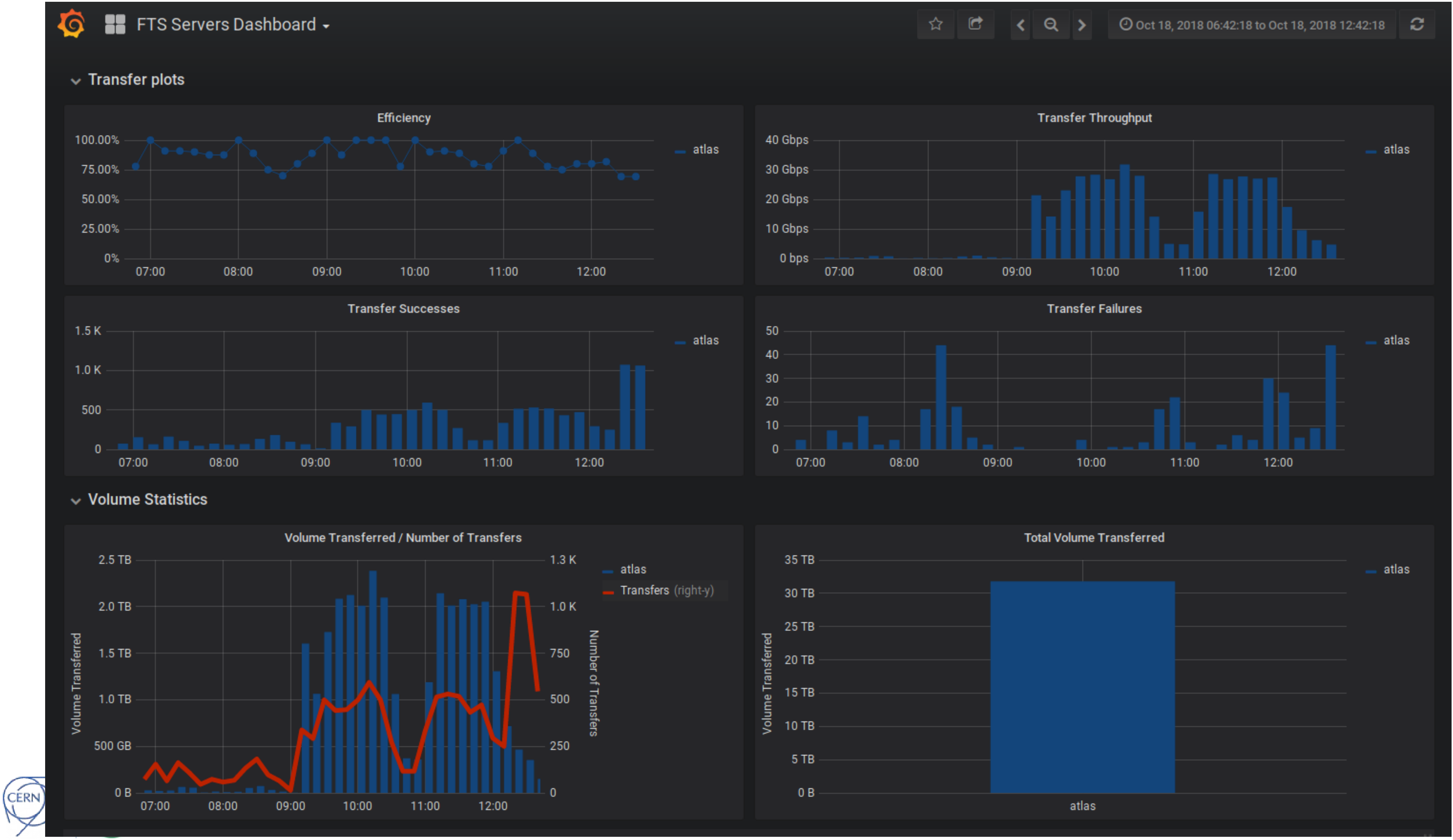


```
From 2018/10/18 08:18:01 To 2018/10/18 12:45:27

Description: ASD001N_MSPNL_SARA_03714 S145-AC11AK ODF18 poort 18 #020331 en #020331 100G MSP Neth
🟩 Inbound         Current:  257.81M  Average:    5.22G  Maximum:   14.05G
🟦 Outbound        Current:    1.53G  Average:    1.11G  Maximum:    2.32G
🟦 Outbound Peak   Maximum:    2.32G
🟩 Inbound Peak    Maximum:   14.05G
```

Daily (5 Minute Average)

FTS dashboard



Conclusions

- Successfully added additional ~10Gbps from underutilized LHCONE link to the saturated 20Gbps primary LHCOPN links

- FTS automatically exploited the additional bandwidth

- Routers' load-balancing not quite capable of fully exploit links with different speeds

Side notes

- The 2nd test was first tried on the 4th of October, but didn't succeed because EOS ATLAS service was saturated with other transfers

- On the 18th , the EOS ATLAS instance was reserved for the test to exploit all the bandwidth

=> Network bandwidth seems to be more abundant than file transfer capabilities

Next steps

Next steps

Load balance with dynamic circuit:

- Create temporary circuit on GEANT SDN BoD infrastructure using API

- Traffic engineering with Segment Routing on CERN-SURFnet link

Questions?

edoardo.martelli@cern.ch