ALICE internal and external network requirements @T2s

WLCG Workshop 2017
University of Manchester

20 June 2017

Latchezar Betev

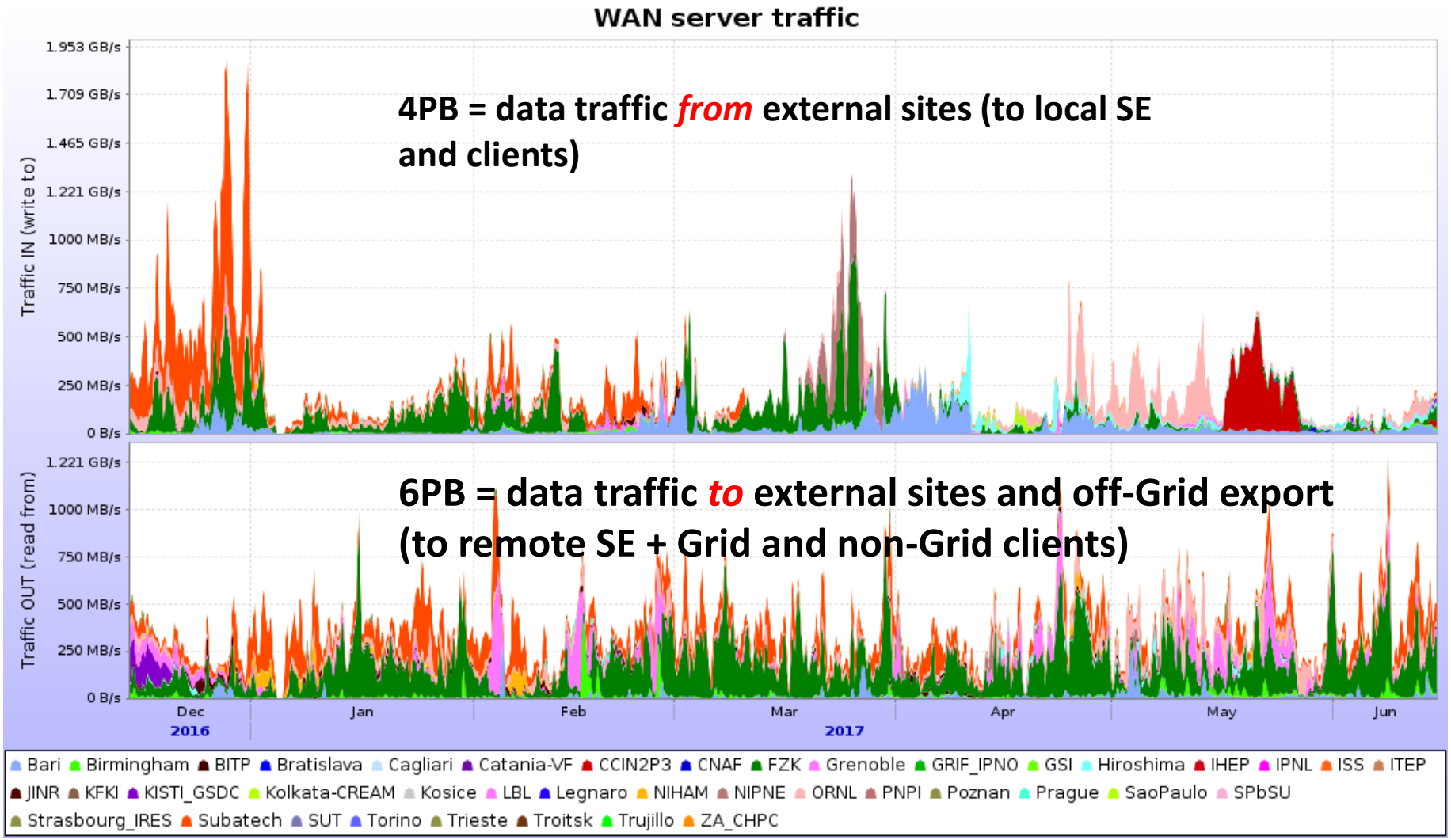
The ALICE Grid today

Average CPU count

~130 CPU cores - April 2017

60% @T0+T1s, 40% @T2s

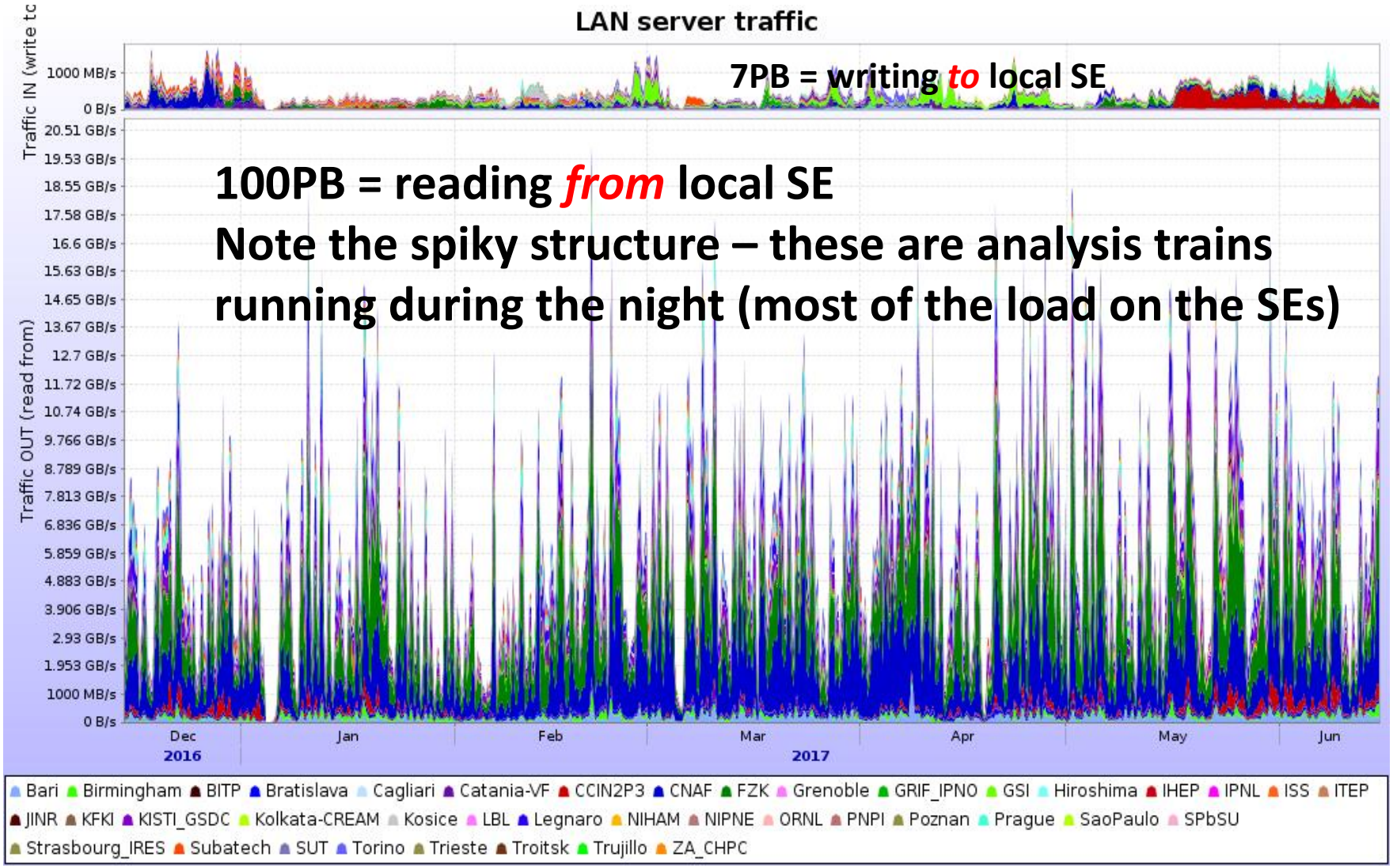
WAN @T2s


WAN server traffic
Traffic IN (write to)
1.953 GB/s
1.709 GB/s
1.465 GB/s
1.221 GB/s
1000 MB/s
750 MB/s
500 MB/s
250 MB/s
0 B/s
4PB = data traffic *from* external sites (to local SE and clients)
Traffic OUT (read from)
1.221 GB/s
1000 MB/s
750 MB/s
500 MB/s
250 MB/s
0 B/s
6PB = data traffic *to* external sites and off-Grid export (to remote SE + Grid and non-Grid clients)
Dec
2016
Jan
Feb
Mar
2017
Apr
May
Jun
Bari Birmingham BITP Bratislava Cagliari Catania-VF CCIN2P3 CNAF FZK Grenoble GRIF_IPNO GSI Hiroshima IHEP IPNL ISS ITEP
JINR KFKI KISTI_GSDC Kolkata-CREAM Kosice LBL Legnaro NIHAM NIPNE ORNL PNPI Poznan Prague SaoPaulo SPbSU
Strasbourg_IRES Subatech SUT Torino Trieste Troitsk Trujillo ZA_CHPC

WAN Numbers @T2s

- Total WAN traffic are 1/11 of LAN traffic
  - WAN total T2s = 650MB/sec => **12.5KB/sec/core** (52K cores)
- Our 'advisory rule' for T2s is **"100Mb WAN network per 1K cores"**
  - Usually the available bandwidth is superior to the above
  - ~Independent of T2 role – a diskless T2 has to export ~2x the produced data; T2 with an SE exports one copy and accepts copies from other centres.

LAN @T2s


LAN server traffic
Traffic IN (write to
1000 MB/s
0 B/s
7PB = writing *to* local SE
Traffic OUT (read from)
20.51 GB/s
19.53 GB/s
18.55 GB/s
17.58 GB/s
16.6 GB/s
15.63 GB/s
14.65 GB/s
13.67 GB/s
12.7 GB/s
11.72 GB/s
10.74 GB/s
9.766 GB/s
8.789 GB/s
7.813 GB/s
6.836 GB/s
5.859 GB/s
4.883 GB/s
3.906 GB/s
2.93 GB/s
1.953 GB/s
1000 MB/s
0 B/s
100PB = reading *from* local SE
Note the spiky structure – these are analysis trains running during the night (most of the load on the SEs)
Dec
2016
Jan
Feb
Mar
2017
Apr
May
Jun
Bari Birmingham BITP Bratislava Cagliari Catania-VF CCIN2P3 CNAF FZK Grenoble GRIF_IPNO GSI Hiroshima IHEP IPNL ISS ITEP
JINR KFKI KISTI_GSDC Kolkata-CREAM Kosice LBL Legnaro NIHAM NIPNE ORNL PNPI Poznan Prague SaoPaulo SPbSU
Strasbourg_IRES Subatech SUT Torino Trieste Troitsk Trujillo ZA_CHPC

LAN Numbers @T2s

- LAN sustains by far the highest traffic (jobs go to data)
  - About 2% of data is read remotely (for example, local SE issue)
  - 90% of the LAN traffic is analysis, 10% is local + remote clients write (for a T2 with SE)
  - Diskless T2s - LAN=WAN traffic
- Figure of merit (for 85%+ CPU efficiency): **2.5MB/sec/core** from local SE
- Remote access penalty
  - RTT is introducing efficiency penalty of ~15% per 20ms
  - Remote read is feasible for small data samples and as failover

Role of tiers

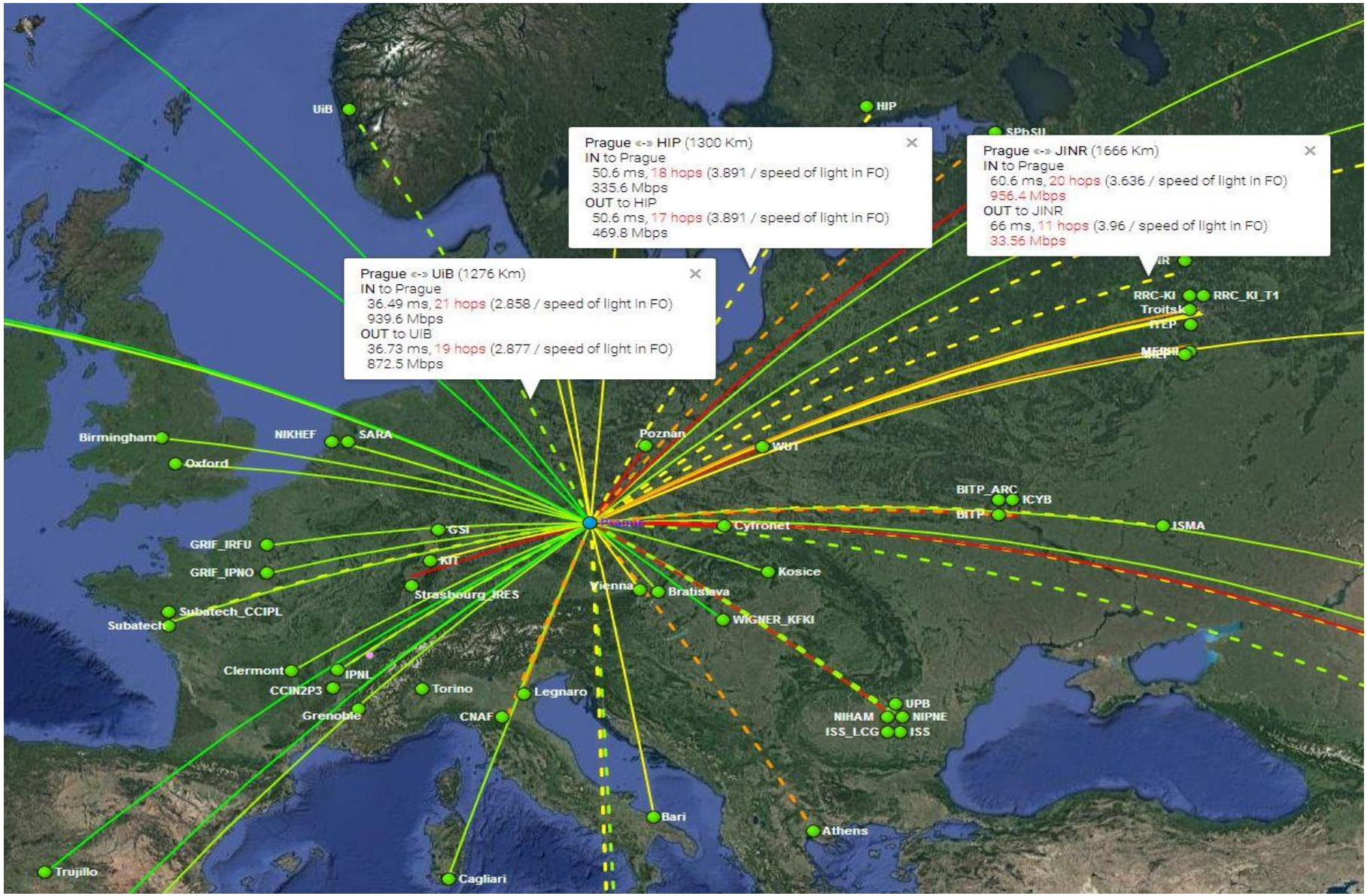
- T0/T1/T2s – only differentiation is RAW data storage and processing (T0/T1s)
- All tiers with storage do MC and analysis
- Diskless T2s can do only MC, all data is exported to close T0/T1s/T2s

Role of network monitoring

- ALICE is using in-house measured network map
  - Through ***all-to-all*** VO-box periodic bandwidth tests
- Used to determine the best placement for data from every client (Storage Auto-discovery) in real time
  - Based on network topology
  - Traceroute/tracepath for RTT (bidirectional)
  - 1 TCP stream bandwidth tests (like jobs do)

=> Distance metric function of the above measurements plus storage availability and occupancy levels

Network paths



Network tuning

- Comprehensive WAN network picture is essential for efficient data transfers and placement
- ALICE depends on and strongly encourages the full implementation of LHCONE at all sites
  - Plus the associated tool (like PerfSONAR), properly instrumented to be used in the individual Grid frameworks
- Huge progress (Thank you network experts!)
  - Number of sites in well-connected parts of the world still have to join
  - Entire regions with high local growth of computing resources, *in particular Asia*, require a lot of further work