Understanding the T2 traffic in CMS during Run-1

T.Wildish, Princeton University

CMS data-transfers at the beginning of Run-1 followed the MONARC model, with strictly hierarchical transfers between sites. The network was expected to be a point of failure, and debugging such failures was thought to be difficult.

During Run-1 the networks performed very well, so rules were relaxed to allow Tier-2 sites to transfer data directly to/from any site in CMS. This new model has never been quantified or modeled, nor, therefore, optimized.

The T2 sites serve data for analysis, so any change in their network behaviour may have a direct impact on how rapidly CMS produces physics results. This paper presents an analysis of the T2 traffic during Run-1, which can serve as a starting point for further modeling and optimization of operations during Run-2 and beyond.

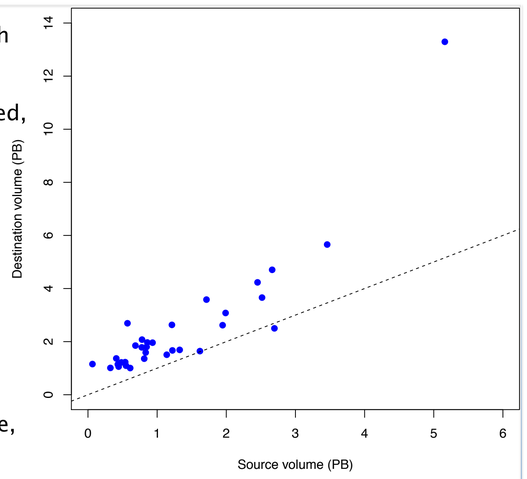
Defining Run-1 to be the period from April 2010 to the start of 2015, the total volume of data transferred between different tiers is summarised in this table.

Src \ Dst	T0	T1	T2	Total
T0	2.4	14.8	5.4	22.6
T1	1.8	22.8	57.4	82.0
T2	1.7	19.8	26.4	47.9
Total	5.9	57.4	89.2	152.5

The grand total of 152.5 PB represents, on average, almost exactly 1 GB/sec, aggregated CMS-wide.

T2s received about twice as much data as they sourced, and the total T2->T2 traffic (26.4 PB) was about 1/6 of the total traffic in Run-1. Notably, T2->T2 traffic exceeded T2->T1 or T1->T1 traffic.

As either sources or destinations, T2s took part in almost 3/4 of the data transfers during Run-1, only slightly less than the T1s. On the other hand, the T0 took part in only 1/6 of the traffic.



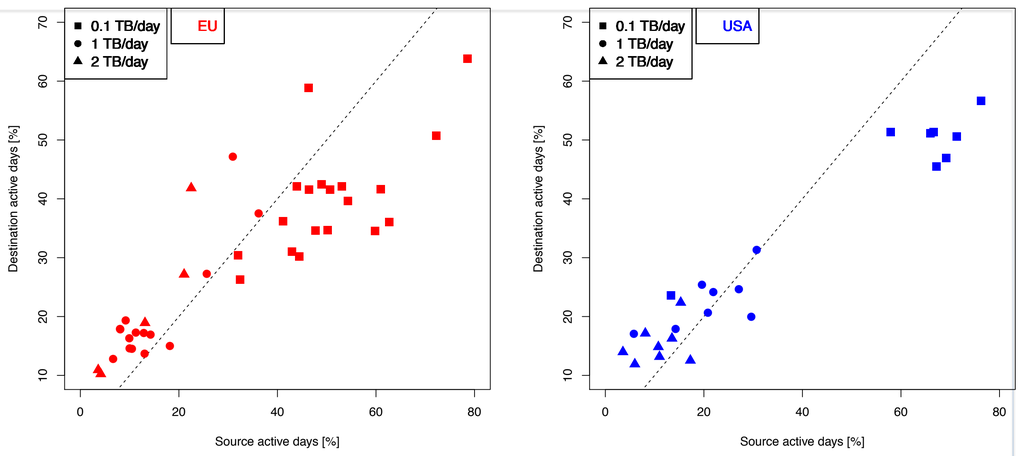
This plot shows the volume transferred by each site as a source vs. as a destination.

Most sites received more data than they sourced, about 1.6-2.5 times as much. This is to be expected, they source data by running monte-carlo production from time to time, but they regularly receive fresh data for analysis.

The dashed line shows equal volumes of data sourced/received, as a guide.

The outlier at the top is the CERN T2, which received an impressive 13.3 PB of data during Run-1. This doubtless reflects its special nature, being closely coupled to the T0.

The site below the line, Purdue, produced more data than it received for analysis. The site exactly on the line (RWTH) almost achieved the same. Not surprisingly, these are two of our largest T2 sites, with an impressive capacity for monte-carlo production.

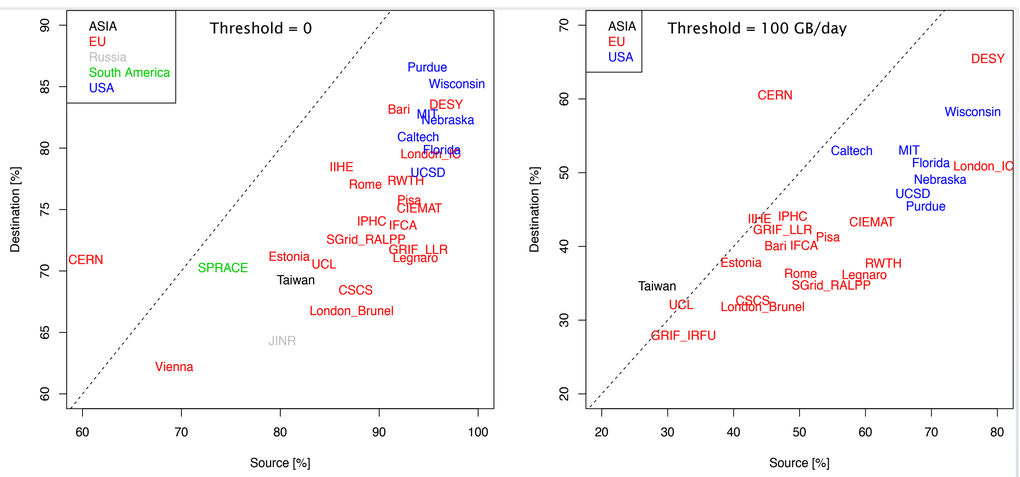


This plot shows the source/destination days of activity at different thresholds, 100 GB/day, 1 TB/day, and 2 TB/day. EU sites are on the left, US sites on the right.

As the threshold increases, sites become much less active, both as a source and a destination. Despite this, transfers of more than 2 TB per day on a given link happen a reasonable fraction of the time. It's noteworthy that all 8 US T2 sites managed to transfer 2 TB or more per day, while only a handful of the 30 or so EU T2s achieved that.

The US sites are much more active at low daily volumes, but become quickly less active as the volume grows. In the EU the decline in activity is slower. This may reflect the higher network bandwidth available in the US compared to the EU, which means that large data volumes transfer in short time periods instead of contributing to more than one day of activity.

Higher bandwidth in the US may also explain the tighter clustering of the US sites than the EU sites. If the sites are all doing more or less the same thing, the spread could come from variations in connectivity alone.



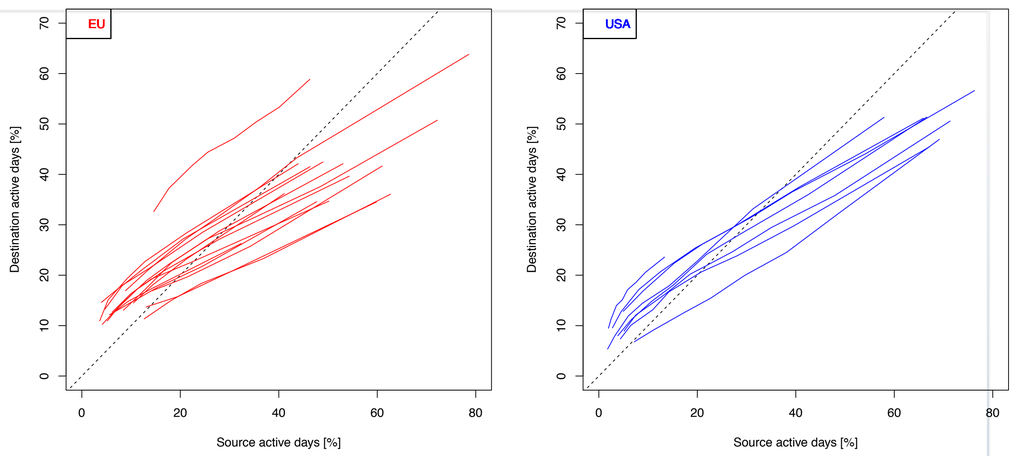
These plots show the number of days a site is active as a destination vs. the number of days it is active as a source, counting days where the data sent/received had no threshold (left) or a threshold of 100 GB per link (right). As a reference, 100 GB in one day is about 10 MB/sec on average. The dashed line shows equal activity as both source and destination. The positions of sites are approximate, they have been adjusted for readability.

At zero threshold, we see that many sites are very active as data-sources, sourcing data on close to 100% of the days during Run-1. They are less active as destinations, but still very busy. US sites are the most active, with a cloud of European sites following close behind.

Even sites with relatively poor connectivity are present, albeit at lower data-volumes (SPRACE, JINR, Taiwan).

The CERN T2, despite receiving so much data, has a relatively low duty-cycle. Again, this is consistent with its role, closely coupled to the operations of the T0.

The fact that sites are more often active as sources than as destinations, but receive more data than they send, is mostly due to them receiving about 2/3 of their data from T1s.



These plots show the source/destination active days per site for a number of different thresholds, with one line representing one site. The lower threshold (upper rightmost end of the line) is 100 GB per day, as usual, and the upper threshold (lower leftmost end of the line) is at 4 TB/day.

Only 3 EU sites actually managed to transfer that much data in one day, so for the other sites their lines end at lower thresholds, but all US sites are still represented.

The most prominent outlier is the upper curve for the EU, which is the CERN T2 again. The other unusual curve is the short trace for the US, in the bottom-left of the plot. This is Vanderbilt, which is associated with heavy-ion physics, not with p-p physics.

Apart from those two sites, all the trajectories are very closely clustered, with the same gradient, and the same curvature at higher transfer volumes. This reinforces the claim that all sites are engaged in essentially the same activities, and are not showing differences reflecting their geographical location, local physics community, hardware capability, or anything else.

The CMS T2s transferred a significant volume of data during Run-1, more than was originally anticipated, and via routes that were explicitly excluded from the original computing models.

Despite considerable variations in size and capability among our T2s, we see that they almost all show essentially the same characteristics.

They receive far more data than they source, yet are more frequently active as a source than as a destination. Most of the sites transfer volumes of the order of 100 GB per day on a routine basis, some transfer much more, up to a few TB per day, at least once per week.

The most significant variation among sites is that the US sites systematically reach higher daily volumes more often than EU sites, and show less spread in their pattern of activity as a function of volume. This is probably due to the performance of the network itself, rather than to anything the sites themselves are doing.

awildish@princeton.edu