Understanding the T2 traffic in CMS during Run-1

Wildish T¹, on behalf of the CMS Collaboration

¹ Princeton University, New Jersey, USA
E-mail: awildish@princeton.edu

Abstract.
In the run-up to Run-1 CMS was operating its facilities according to the MONARC model, where data-transfers were strictly hierarchical in nature. Direct transfers between Tier-2 nodes was excluded, being perceived as operationally intensive and risky in an era where the network was expected to be a major source of errors. By the end of Run-1 wide-area networks were more capable and stable than originally anticipated. The original data-placement model was relaxed, and traffic was allowed between Tier-2 nodes.

Tier-2 to Tier-2 traffic in 2012 already exceeded the amount of Tier-2 to Tier-1 traffic, so it clearly has the potential to become important in the future. Moreover, while Tier-2 to Tier-1 traffic is mostly upload of Monte Carlo data, the Tier-2 to Tier-2 traffic represents data moved in direct response to requests from the physics analysis community. As such, problems or delays there are more likely to have a direct impact on the user community.

Tier-2 to Tier-2 traffic may also traverse parts of the WAN that are at the 'edge' of our network, with limited network capacity or reliability compared to, say, the Tier-0 to Tier-1 traffic which goes the over LHCOPN network.

CMS is looking to exploit technologies that allow us to interact with the network fabric so that it can manage our traffic better for us, this we hope to achieve before the end of Run-2. Tier-2 to Tier-2 traffic would be the most interesting use-case for such traffic management, precisely because it is close to the users' analysis and far from the 'core' network infrastructure.

As such, a better understanding of our Tier-2 to Tier-2 traffic is important. Knowing the characteristics of our data-flows can help us place our data more intelligently. Knowing how widely the data moves can help us anticipate the requirements for network capacity, and inform the dynamic data placement algorithms we expect to have in place for Run-2.

This paper presents an analysis of the CMS Tier-2 traffic during Run 1.

1. Introduction
The CMS computing model[1] has evolved from the hierarchical MONARC structure, where Tier-2s only transfered data to/from their associated Tier-1, to a fully interconnected mesh where any site can transfer data to any other site. This was a direct result of the network performance and reliability greatly exceeding the expectations of a decade ago, when the computing models were written. Then, the network was expected to be unreliable and over-subscribed. Instead, it has turned out to be the most reliable part of the CMS computing grid, with more capacity than needed to sustain activities during Run-1.

The traffic flowing between our Tier-2 sites has grown organically, out of operational considerations arising from day to day, rather than being explicitly planned like the Tier-0/Tier-1 traffic. Differences in the way we use individual Tier-2 sites may arise from their geographical proximity to other sites, the mix of their local physicist communities, their network capabilities, their overall size and network capacity, or other factors.
CMS is also actively investigating ways to improve our interaction with the network[2], as well as considering expanding our data-management capabilities beyond the static Tier-topology of our current network to include opportunistic, volunteer, or private computing resources[3].

For all these reasons, it is therefore important to understand the nature of the Tier-2 traffic. We need that understanding in order to predict how our system will behave during Run-2, and to estimate the effect of any changes we make in the way we distribute and analyze data.

2. Summary of Run-1 traffic
Run-1 of the LHC can be loosely defined as the period from April 2010 to the end of 2014. Although the LHC accelerator was shut down for upgrading in February 2013, analysis activities have continued unabated, in much the same vein as before.

During that time, a total of 163 PB of data was transferred between the various sites making up the CMS computing grid. The annual breakdown (see table 1) varies somewhat, with a peak in 2012 - the year of the Higgs discovery - but also with significant activity in 2014, when the long shutdown was well under way. The breakdown in terms of source vs. destination tier is shown in table 2. The discrepancy between the total amount of data transferred shown in table 1 compared to table 2 is because of transfers to or, more rarely, from Tier-3 sites, which are included in the yearly totals but not are not considered further in this analysis.

Table 2 shows that Tier-2 sites acted as sources for 31% of the data during Run-1, and as destinations for 59% of the data. This is consistent with their intended roles as production centres for Monte Carlo data and as the primary location of data for end-user analysis. The table also shows that some 30% of the data received by Tier-2 sites was sourced by other Tier-2 sites, so half the traffic they received (26.4 PB) came from other Tier-2s. It is this traffic that goes beyond the original computing model.

The Tier-2 to Tier-2 traffic per year is shown in table 3. Again we see evidence of a peak of activity in 2012, with perhaps a drop after then.

Looking at the amount of data transferred per site, it is not surprising that most of the activity is in the USA or Europe. CMS has fewer Tier-2 centres outside these regions, and they have much poorer network connectivity in general, so they account for very little of the total traffic. Table 4 shows the traffic sourced/received when the source is in one or the other region. The volume received is comparable, though the volume sent differs by a factor two. This is explained by figure 1, which shows the traffic for each site. The outlier at the top of the plot is the CERN Tier-2 (T2,CH,CERN), which had a special role in Run-1 and during the shutdown. In Run-1 it was used for time-critical operations related to the Tier-0 activity, such as calibrations and cross-checks that were not yet automated. During the shutdown, it was essentially operated as a Tier-1 with a large Tier-2.

Figure 1 also shows that although the amount of data sent/received varies per site, the majority lie broadly along the same line, sending and receiving data in a similar ratio.

### Table 1. Total volume of data (PB) transferred per year by CMS during Run-1.

| Year | Total volume (PB) |
|------|-------------------|
| 2010 | 26.1              |
| 2011 | 28.7              |
| 2012 | 40.1              |
| 2013 | 30.9              |
| 2014 | 38.0              |

### Table 2. Volume of data (PB) transferred between different computing Tiers by CMS during Run-1.

| Destination T0 T1 T2 Total | Source T0 2.4 14.8 5.4 22.6 |
|----------------------------|-------------------------------|
| Source T1 1.8 22.8 57.4 82.0 |
| Source T2 1.7 19.8 26.4 47.9 |
| Total 5.9 57.4 89.2 152.5    |

### Table 3. Tier-2 to Tier-2 traffic per year.

| Year | Tier-2 to Tier-2 Traffic (PB) |
|------|-------------------------------|
| 2010 |                               |
| 2011 |                               |
| 2012 |                               |
| 2013 |                               |
| 2014 |                               |
### Tier-2 to Tier-2 traffic

| Year | Volume (PB) | Percentage |
|------|-------------|------------|
| 2010 | 5.2         | 21.2       |
| 2011 | 5.5         | 20.6       |
| 2012 | 6.9         | 37.4       |
| 2013 | 4.8         | 17.0       |
| 2014 | 4.0         | 11.2       |

Table 3. Total volume of data transferred per year between Tier-2 sites only.

Interestingly, one site (Purdue) actually sent more data than it received, and another (DESY) came within a couple of percent of the same achievement. These are two of our biggest Tier-2 sites, and therefore produce a large amount of Monte Carlo data. If their local analysis communities are not proportionally large, they may have less call to transfer data locally for analysis, hence the low received/sent ratio.

### 3. The Tier-2 Duty Cycle

It is of interest to see how often a given Tier-2 is active as either a source or a destination. Figure 2 shows this, where the duty-cycle is calculated as the fraction of the 1736 days under consideration the site acted as either a source or a destination. No threshold is applied, so even the transfer of a single file in a day is enough to count that day as active.

A large number of sites were active as sources for more than 80% of the time, with a somewhat greater spread on the fraction of their time as a destination. Although the majority of the traffic was to/from US or European sites, it is notable that SPRACE (Brazil), Taiwan and JINR all feature in this plot as being significantly active. It is also notable that 7 out of 8 US Tier-2 sites are extremely active, with the European sites showing a greater spread. The remaining US Tier-2 (Vanderbilt) is another specialised site, primarily serving the heavy-ion community, so it is not surprising that it is less active than the others.

Figure 3 shows the duty cycle for all sites with an applied cut at 100 GB per day, which averages to about 10 MB per second. This changes the picture somewhat. The US sites still maintain a high level of activity, both as sources and destinations, while the European sites now show a greater spread. The sites clustered at higher duty-cycles as both sources and destinations still tend to be the larger sites.

The CERN Tier-2, despite receiving such a large volume of data, has an apparently low duty-cycle as a destination. This is consistent with its activities being closely coupled to the Tier-0, it’s extremely busy when it’s active but is not active all the time for analysis, unlike other Tier-2s.

The fact that sites are more often active as sources than as destinations, yet receive more data than they send, is mostly due to them receiving about two-thirds of their data from Tier-1s.

### 4. The Duty-cycle by Regions

Figure 4 shows the duty-cycle for the US and European sites separately, for three different daily thresholds: 100 GB/day, 1 TB/day, and 2 TB/day. The European sites show a much greater spread at all thresholds, while the US sites are much more closely clustered for each threshold. All 8 US sites are represented as both sources and destinations at the highest threshold, 2 TB/day, while only about 1 in 4 of the European sites have managed this high rate. The percentage of active days, as either a source or a destination, drops more sharply with rising threshold for the US sites than it does for the European sites.
Figure 1. Volume of data transferred as a destination vs. as a source for Tier-2 to Tier-2 traffic in CMS in Run-1. The outlier at the top is the CERN Tier-2. Most sites received more data than they sent, as can be seen from their position above the dashed line which has a gradient of one.

These effects are probably all due to the higher network bandwidth generally available in the US compared to European sites. When a large volume of data is requested from a US site it can deliver it in a short space of time, so delivering 2 TB or more per day is not difficult. For less well-connected European sites, on the other hand, such a transfer may span two days, or more, giving less days active at high rates and a wider spread on the days of activity at intermediate thresholds.

The tighter clustering of the US sites at different thresholds also suggests that they are all engaged in similar activities (e.g. types of analysis or Monte Carlo production), despite variations in local analysis communities.

Figure 5 shows the same information, but with one line representing each site for a range of thresholds, rather than a three points per site at only three thresholds. The minimum threshold is 100 GB/day, the maximum is 4.0 TB/day. All the US sites in fact managed to transfer 4.0 TB/day on some occasion, but not all the EU sites were capable of that, so the lines for those sites end at lower thresholds.

Two outliers can be identified. The first is the uppermost line in the plot for European sites.
Figure 2. Tier-2 duty cycle, i.e. the percentage of days the site acts as a destination vs. as a source. The dashed line has a gradient of 1, i.e. sites which are equally active as sources and destination will lie on this line. Site name positions are approximate, and have been adjusted for readability in some cases. N.B. note the axes are not both on the same scale.

5. Conclusions
The CMS Tier-2 sites transferred a significant volume of data during Run-1, more than was originally anticipated. Much of that traffic was between Tier-2 sites, both as source and destination, which didn’t feature in the original computing model. Most sites transfer at least
Figure 3. Tier-2 duty cycle, i.e. the percentage of days the site acts as a destination vs. as a source. The dashed line has a gradient of 1, i.e. sites which are equally active as sources and destination will lie on this line. Site name positions are approximate, and have been adjusted for readability in some cases. N.B. note the axes are not both on the same scale.

100 GB per day on a routine basis, while some transfer up to a few TB per day at least once per week.

Tier-2s receive far more data than they source, yet are more frequently active as sources than as destinations. This probably represents the gathering of many datasets in one location, to support a specific local analysis community need.

The most significant variation observed is that the US sites systematically reach higher daily volumes than other sites, and show less spread in their pattern of activity as a function of volume. This probably reflects the performance of the network itself, the US has led the way with high-speed networking for Tier-2 sites.

Apart from that, the remarkable similarity in the behaviour of the sites suggests that they are all engaged in essentially the same activities. This is not a given, they vary considerably in size and capability, and their local physicist communities are not expected to be equivalent.

This is good news for future data-management activities in CMS, as it means that Tier-2 sites can be treated identically, regardless of location. This greatly simplifies modeling, and activities such as dynamic replication of data or initial placement of new data.
Figure 4. Percentage of days active as a destination vs. a source at different daily thresholds for European sites (left) and US sites (right). N.B. again, the axes are not both on the same scale, to improve visibility.

Figure 5. Percentage of days active as a destination vs. a source at different daily thresholds for European sites (left) and US sites (right). Each line represents a different site. The thresholds (not shown) are 0.1, 0.4, 0.7, 1.0, 1.5, 2.0, 3.0, and 4.0 TB per day. Again, the axes are not both on the same scale.

References
[1] Grandi C, Stickland D and Taylor L 2005 The CMS Computing Model CERN-LHCC-2004-35/G-083, CMS note 2004-031
[2] Integrating Network-Awareness and Network-Management into PhEDEx, International Symposium on Grids and Clouds 2015
[3] Bonacorsi D and Wildish T 2013 Challenging CMS Computing with Network-Aware Systems *J. Phys. Conf. Ser.* 513 (2014)