T0-T1-T2 traffic

Working group

- Mailing list: project-lhocpn-t012wg@cernNOSPAMPLEASE.ch
- Mailing list archive

Problem definition

PROBLEM

- T0, T1 and T2 sites need to transfer large amount of data between them.
- Connectivity via normal internet upstreams may be limited in bandwidth and expensive, so not suitable for LCG data transfer.
- Links between pairs may already exist, but they've a limited scope and are not well exploited.

REQUIREMENTS

- Let T0/T1/T2 sites exchange traffic in the most flexible and economical way
- maximize exploitation of links
- straightforward L3 configuration

CONSTRAINTS

- cost of connectivity
- limited L3 network know-how at Tier2s
- do not destroy the LHCOPN, since it's working fine and the LHC is going to start.
- last mile not easy, nor cheap, to implement.

Requirements from the experiments

Largest data flows are in these directions:

- T0->T1 = 10 (already addressed by the LHCOPN)
- T1->T0 = 10 (already addressed by the LHCOPN)
- T1->T1 = 10 (already addressed by the LHCOPN and CBF, but may be changed)
- T1->T2 =
- T2->T1 =
- T0->T2 =
- T2->T2 =

[Kors]

1. For all experiments the most important thing is to save a second copy of the RAW data to tape at the Tier-1's. The first copy is on tape at CERN. This makes the T0 --> T1 links different from all others. This also says something about the importance of the T0-T1 backup paths. Any additional features on the OPN should not endanger this prime functionality.
2. The latest STEP exercise has taught us that (for ATLAS, but I believe also for CMS) the peak rates between T1's are just as high or even in excess of the T0-T1 rate. It has never become a problem though during the test because it is less time critical than the T0-T1. Moreover, we have the freedom to go to round-robin rather than point to point mode because it is all about distributing the same data to all T1's.
There is a difference between the ATLAS and the CMS model for T1--> T2. (B.t.w. T1 --> T1 and T1 --> T2 should have the arrows pointing both ways: -->) Where for CMS the T2 should be able to get its data from each T1, for ATLAS the most important traffic is between the T2's and the T1 in the same "cloud". So the ATLAS model is more hierarchical and the CMS model is more of a mesh. ATLAS has some out-of-the-cloud T1 --> T2 traffic but it is less important.

4. ATLAS has only a few (4: Rome, Munich, Michigan, Geneva) calibration data streams between the T0 and T2's. These data are time critical but of moderate rate (<50 MB/s). Calibrations (muon and trigger) are done at those sites and processing in the T0 depends on the results of those calibrations being sent back to CERN in a timely fashion.

5. For ATLAS, some T2's are more important than others: we have officially ~60 T2's but 50% of our analysis gets done in the ~10 best T2 sites. We would like to have the option of having those T2's better served but must keep in mind that the list of "golden T2's" may vary with time. Not on a daily basis but some T2's may improve significantly in a matter of months. For the T1 --> T2 traffic, the same holds as for the T1 --> T1: the rate may be high, higher than any of the other channels and even more so for CMS than for ATLAS: CMS relies on cacheing the required data, whereas ATLAS relies to a first order on pre-placement. Experience will show how time critical the dataflow is for analysis in the T2's: if it is too slow, processes may time out and cpu usage may become very inefficient.

3. There is a difference between the ATLAS and the CMS model for T1--> T2. (B.t.w. T1 --> T1 and T1 --> T2 should have the arrows pointing both ways: -->) Where for CMS the T2 should be able to get its data from each T1, for ATLAS the most important traffic is between the T2's and the T1 in the same "cloud". So the ATLAS model is more hierarchical and the CMS model is more of a mesh. ATLAS has some out-of-the-cloud T1 --> T2 traffic but it is less important.

4. ATLAS has only a few (4: Rome, Munich, Michigan, Geneva) calibration data streams between the T0 and T2's. These data are time critical but of moderate rate (<50 MB/s). Calibrations (muon and trigger) are done at those sites and processing in the T0 depends on the results of those calibrations being sent back to CERN in a timely fashion.

5. For ATLAS, some T2's are more important than others: we have officially ~60 T2's but 50% of our analysis gets done in the ~10 best T2 sites. We would like to have the option of having those T2's better served but must keep in mind that the list of "golden T2's" may vary with time. Not on a daily basis but some T2's may improve significantly in a matter of months. For the T1 --> T2 traffic, the same holds as for the T1 --> T1: the rate may be high, higher than any of the other channels and even more so for CMS than for ATLAS: CMS relies on cacheing the required data, whereas ATLAS relies to a first order on pre-placement. Experience will show how time critical the dataflow is for analysis in the T2's: if it is too slow, processes may time out and cpu usage may become very inefficient.

[Question] Are the T2s dedicated to a single experiment, or are they more like the T1s, which in general are serving more than one experiments?
[Answer] A handful ~10 of our T2's also serve other VO's. We share very few with CMS, I can think of only ~5, Manno, Taipei, UC London are the big ones. The T2 federations often support all VO's but if you look inside you discover that individual sites mostly support single VO's ... with a few exceptions like Grif and Lyon.

[Question] you only mention Atlas and CMS. Why? Is it because Alice and Lhcb fall in one of the two models? Or their traffic is negligible compared to the CMS and Atlas one? Or?
[Answer] For the network bandwidth indeed ATLAS and CMS make the bulk. The numbers for LHCb are negligible on this scale. Alice has more impressive numbers but only for Heavy Ion running. They don't distribute real time so it is less critical. Moreover H.I. running comes much later and during H.I. running the other experiments are mostly off so it becomes a special case. For the moment, if you get it to work for CMS and ATLAS it will also work for the others.

[Harvey]
Ken tells me a typical baseline might be to consider transferring 200 Terabytes a few times per year, and taking up to two weeks to do it. That would be 1.5 Gbps as a long term average, although as we've seen peaks have been much higher while doing that, and in several cases reach the 7-9 Gbps range repeatedly. In addition, the time to complete a dataset transfer can be quite important. I build a couple of scenarios in what follows: Within the overall usable disk space specified, which is 200 - 400 Tbytes now if I recall correctly, the typical dataset sizes mentioned are 30 Terabytes. In discussing the data forms and samples of interest to a single physics group, order of magnitude numbers are 100 Terabytes.

Scenarios:

1. A 30 Terabyte transfer
   • To do this in a 24 hour day, the required rate needs to average ~3 Gbps (and as above, to get this in practice will require larger peaks)
   • It's tolerable perhaps, but labor intensive, to do this at 1 Gbps avg.
   • much below 1 Gbps it's just too expensive in terms of manpower and wasted disk space to be useful

2. Once several datasets are in flight, say up to the level of 100 Tbytes then
   • To be agile and respond to new versions or urgent needs you need most of a 10 Gbps link (leaving 1 Gbps or more for other uses like local analysis-related and general network
activities). So a goal could be 6-8 Gbps average: 100 Tbytes then arrives in ~30-40 hours.

- You can get by with about 3 Gbps net - again that is going to be manpower intensive and some sites will find this burdensome due to the multi-day length of the transfer; ~ 100 hours for 100 Tbytes.
- At 1 Gbps - 100 Tbytes takes ~12 days solid. This is extremely manpower consuming when you think of the manual operations required to ensure that 100% of a given dataset has arrived intact.

3. Once you have a few times 100 Tbytes of "active" datasets (by 2011 ?) then you had better plan on a level capability of "several Gbps and up".

Another dimension is that of a shared infrastructure. It is operationally better to have transfers be scheduled, proceed, and be completed, than drag on for weeks. The scenarios with shorter transfers are more efficient by far - both in terms of working efficiency of everyone concerned, and also for good use of the available network resources.

--

To add to the discussion I have this from Shawn McKee regarding ATLAS Tier2s. They regularly exercise sending/receiving 20*3.6 GByte files:

*We want 10GE sites to regularly do 400MB/sec. We track fastest/slowest/average file transfer time as well as problematic transfers.*

While it remains to be seen if all ATLAS Tier2s will have all that they need from the local Tier1, this sets the scale (namely ~3 Gbps). I note that when transferring fairly small individual files this way, with no buffer-filling and shipping a la FDT, then as we've observed 3 Gbps average means 7-9 Gbps peaks. You can also have a look at a 2007 presentation of Eli Dart (ESnet) on network requirements, from which I extracted the summary information (enclosed). On slides 7 and 8 you will see some (circa 2007 and 2010) bandwidth requirements estimates for Tier1 - Tier2. Also for aggregate Tier3 requirements.

**Comments on Requirements**

[Artur]

T1-T1 is only partially addressed in the OPN. The CBFs are the infrastructure for and between those T1s which do have them, but then it's up to scavenger service on the Inks dedicated to T0-T1. The fact that some T1s expressed desire to have additional resources dedicated to T1-T1 (in this case through CERN) indicates that this might not be adequate.

CBFs are also a good thing in Europe, but not really an option once we leave the continent. So an exchange point makes good sense. But that's running ahead of myself.

The other question raised is about Tier2s. Their importance in the experiments' model is rising, the resources there are growing every time you look at the WLCG tables. If we are concerned about the success of the LHC, then neglecting them is not good. But admittedly this will be much more challenging than infrastructure for Tier1s.

In any case, the existing OPN should not be touched, I believe there is no question about this. And I believe this is also not needed. Adding T1-T1 resources can be done in various ways, which we should investigate, such as simply adding and dedicating capacity in existing topology, or creating a double-star with a T1-T1 hub, or ... What concerns T2s, that's an extension on the periphery. The OPN core should remain unchanged.

What concerns me more is the case of non-European sites. While on European footprint, one could argue, costs of adding a 10G here and there are mostly reasonable, that might be a bigger problem to sites further away. So we will need to consider efficiency here as well.
Network models

A - Dynamic-Circuits

The Tier centres co-locate their border routers in their NREN PoPs. They buy one or two access circuits (primary and backup) from their premises to their border routers, then from there the NRENs take care of provisioning dynamic circuits to any other co-located Tier router.

It doesn't address the L3 issue, but maybe in the end every Tier centre will need connectivity to a small set of other Tier centres. In this case the routers may be always running with all the necessary addresses and routing commands, running idle while the circuits are down.

Pro
- best use of NREN bandwidth
- Tiers pay for long distance links when needed

Cons
- Awkward/not scalable L3 configs
- lot of coordination needed among NRENs
- may not be possible to connect any pair of Tiers

B - Internet-exchange

The WLCG community builds and maintain a distributed exchange point infrastructure with access switches in a few strategic locations (I may think of Starlight, CERN and Surfnet in Amsterdam, plus another location in the US and/or in Asia) with enough bandwidth to interconnect them.

Then the Tier centres buy one or more circuits from their premises to the access switches and connect them to their border routers. With this model, all the routers will reside in the same IP network, being able to reach any other IX member and to establish ad hoc routing policy with any of them.

Pro
- effective and scalable L3 configuration
- best use of Tier access bandwidth
- easy to connect pair of Tiers independently of their location

Cons
- cost of IX infrastructure
- permanent cost of longer links for Tiers

C - Lightpath exchange

One way around that would be to merge the two concepts, and use static long tail to lightpath exchanges. Foreseen to be dynamic when and where applicable.

Layer 3 is a non-issue (or at least not a big one). Dynamic circuits work as temporary layer 2 connections end-to-end. The only place you have to set IP routes are at the end-hosts (or border router). This can be static, by assigning addresses in a /24 subnet, very much like the existing LHC OPN, or dynamically by user/end-host agents performing the task of requesting the circuit (among other things).
Comments on Models

[Edoardo]
Looking at the requirements gathered so far, it seems the best model would be to have the possibility to establish temporary direct high bandwidth Layer2 links between pairs of Tier centres. But while dynamic provision of circuit may be already possible on the carrier parts in US and Europe, I see it a bit difficult to implement this scenario at Layer1 and Layer3 in the last mile:
L1: Tier centre are not usually located at the NREN PoPs and they may have to buy a permanent access circuit from a commercial carrier. Apart the cost, it will hardly take part in a dynamic circuit provisioning system.
L3: addressing, routing and security policies may be difficult to achieve in a dynamic way; I don't think anyone will let a central entity configure their routers.

[Harvey]
Yes we can concatenate static circuits at the edges with dynamic circuits in the "core". And we will need to use dynamic circuits (automatically, to serve the number of sites foreseen) to conserve bandwidth.

To quote Ian Fisk, Tier2s will transfer 1 - 50 Tbytes on a weekly, and sometimes a daily basis. This means we have to allocate and manage the channels created.

We are pursuing this in earnest with Internet2 in the US. Many groups including us are applying to the US NSF to support dynamic circuits across the Atlantic. Indeed the NSF proposal solicitation specifically mentions dynamic circuits.

Given that grid middleware has evolved (in spite of our efforts) to be non-aware of network configuration or topology in any real sense, we are not considering letting such middleware control the network.

However, what has been conceived is to start with network-resident services that, after initial experience, can build IDCs and switch authenticated, authorized flows onto the circuits.

In order to access these services a simple protocol would be needed wherein there is a request from a grid application for a transfer, with the transfer parameters (volume, source(s), destination(s), etc.), then it is allocated and scheduled. This is to allow us to make good use of the available network resources in the presence of many requests. Then the use during the transfer would be checked and a record kept of good/less-good clients in terms of their actual use of the requested network bandwidth.

An additional (very useful) option for networks with VCAT and LCAS is to squeeze network allocations when a transfer is not using the requested bandwidth.

[Hanno]
I like the IXP model. We as NL-T1 would be able to make use of this model to exchange traffic with all sites connected to NL-T1 either directly (e.g. T0 and some T1's), via the exchange (e.g. other T1's and T2's) or via the Internet (T2's if they don't have a direct or IXP connection to NL-T1).

This model looks very much like the way ISPs work for handling traffic to the global Internet so it is a well known model and it works.

In my opinion an exchange (IXP) must be very stable so dynamic links in the core of the IXP should not be my choice. I would rather choose dynamic connections from sites to the IXP if this is possible at all (to make this happen you need a lot of configuration steps on the network layer(s)).

How many T2 sites are there?

Do we have an overview of where they geographicly are?
To make this work we would need:

- stable connections between the PoP's of the IXP
- IXP PoP's on the right locations
- simple / cost effective access to the IXP for all sites

[Harvey on Model C]
Given that grid middleware has evolved (in spite of our efforts) to be non-aware of network configuration or topology in any real sense, we are not considering letting such middleware control the network.

However, what has been conceived is to start with network-resident services that, after initial experience, can build IDCs and switch authenticated, authorized flows onto the circuits.

In order to access these services a simple protocol would be needed wherein there is a request from a grid application for a transfer, with the transfer parameters (volume, source(s), destination(s),etc.), then it is allocated and scheduled. This is to allow us to make good use of the available network resources in the presence of many requests. Then the use during the transfer would be checked and a record kept of good/less-good clients in terms of their actual use of the requested network bandwidth.

An additional (very useful) option for networks with VCAT and LCAS is to squeeze network allocations when a transfer is not using the requested bandwidth.

[Wouter]
My expectation is that LCAS squeezing is as complicated as dynamically creating/deleting circuits in a heterogeneous networks. By the way I guess the number of SDH devices in the field which support LCAS is still limited.

I am not sure if this has been discussed before and dismissed this approach, but would a layer 2 network with QoS be more useful to control/prevent congestion in the network. Possibly with just a port policer for ingress traffic which can be manipulated by requests from the application or by human intervention? I think most ingredients are available on the shelf (DCN?), and if we overcome the implementation problems (STP, VLAN, MAC tables) of a large Layer 2 network it may work?

[Harvey on Wouter's proposal]
I think we can see if our working (in production) Layer 1 infrastructure for dynamic circuits that supports the Layer 2 circuits with channels with real granular bandwidth guarantees can be replaced by Layer 2 alone, and to what extent. That brings in another set of issues as you point out.

One aspect, which is part of the reason for the Layer 1 infrastructure, is that using Layer 2 VSLPs and configuring them to fail over is unstable in the presence of flapping links (as we experienced in 2006 - 2007). Once you have a number of links, and cross-links for resiliency, it is hard to maintain and restore the Layer 2 topology in the presence of link outages.

Adjusting bandwidth is complicated, but it is conceptually less complicated than just running out of bandwidth and having no facility for hitless transitions and graceful degradation (bandwidth reduction) of channels. It would be more academic if we did not also have the applications that can be used to fully use the bandwidth in a smooth and manageable way; but we do. It's all part of a self-consistent picture.

[Karin and Klaus]
DFN followed the discussion in the twiki. We agree not to open the LHCOPN for T2 sites.

Our guess is that dynamic circuits are not the appropriate means, to connect the T2 sites. Dynamic circuits are not stable enough to be used in a production network. From the economic point of view there will be no benefit because the infrastructure to build dynamic circuits has to be payed before in the same way like p2p-
DFN will establish a test with dynamic circuits ("autobahn") with T1 in GridKA via geant to check the availability and stability of such a link. We will check also the operational procedures to establish the link and the failure management via more than one domain. We think that especially these procedures are a main point to make such links usable for production.

[Enzo]
GARR doesn't plan to provide dynamic circuits, as we are confident that we can rely on a stable overprovisioned IP service network over GARR, GEANT and transatlantic circuits. On the other hand GARR is ready to provide any lightpath between italian T2 and any other T1 or T2 outside Italy, on INFN requirements and demand basis.

LHC-Tier2 sites in Italy will be connected with 10G lightpaths to the Tier1 at CNAF. (today the 10+ italian T2 are connected to the GARR network at 1G over IP).

CNAF-T1 will be connected to GEANT at 100G, as soon that this is available (2011 or before). Today CNAF-T1 is connected to GARR with a double and redundant path at multiple 10G links) CNAF-T1 is connected to CERN-T0 and to GRIDKA-T1 with 10G lightpaths. Very soon another 10G lightpath to CERN-T1 (not T0) will be set up over a third physical path.

GARR recommends that CERN be connected to the GEANT PoP in Geneva (located at CERN) with multiple 10G links or at 2*100G as soon as this technology is available. "Premium IP" could be considered for LHC purpose.