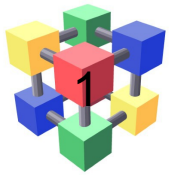
LCG context, needs and deployment strategy

Alessandra Forti

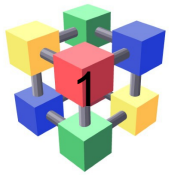
PerfSONAR Day

4th April 2013



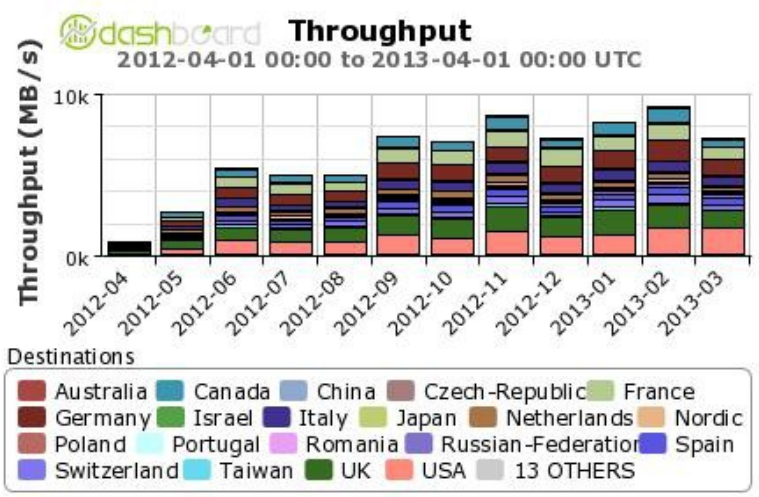
Layout

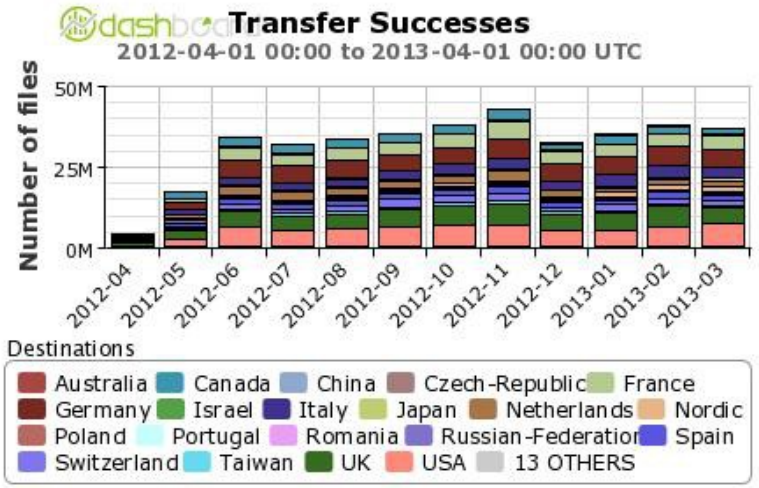
- Experiments data access
- WLCG Ops&Tools TEG recommendation
- WLCG perfSONAR Task Force
- Some real life examples from the UK



Atlas

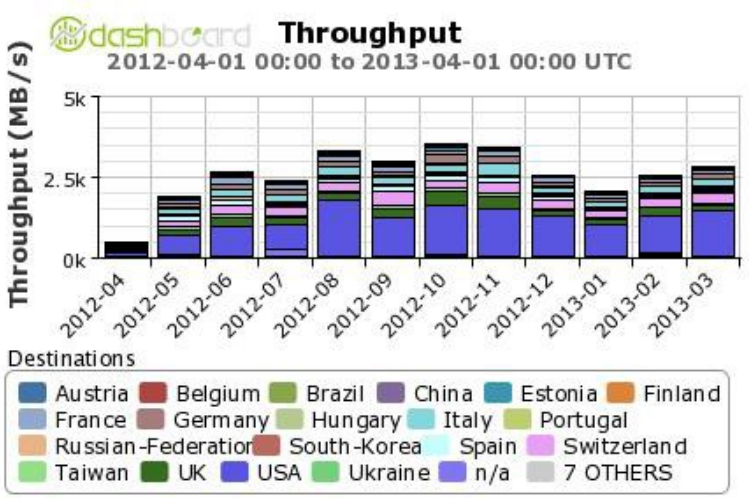
- Originally star topology with a hierarchical structure
  - Evolved towards a flatter data distribution with the introduction of T2D. T2 which can distributed data to other T2
- Runs any activity at any site
- Data transfer partially dynamic (PD2P)
- Working on federated storage based on xrootd (FAX)
  - Access to other sites storage from WNs
    - Copy to scratch
    - Direct IO





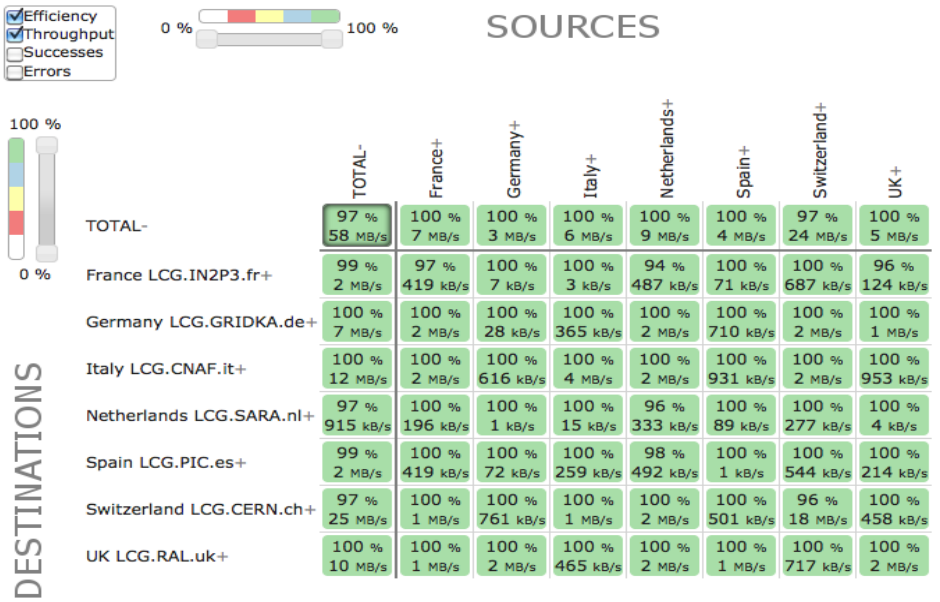
CMS

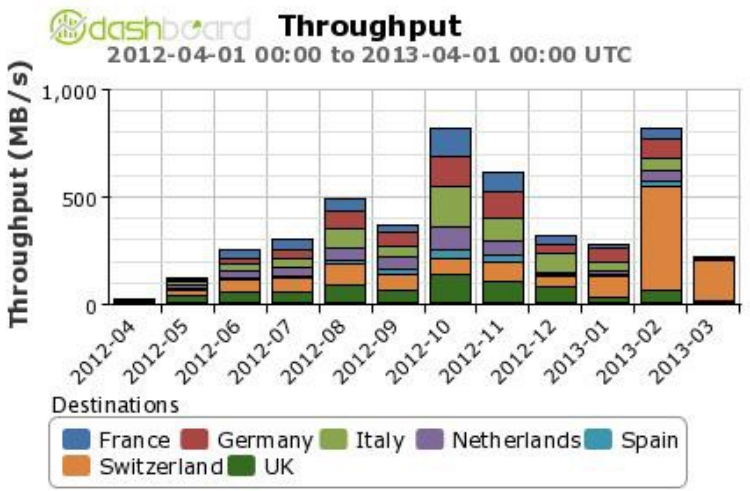
- Also originally hierarchical structure
- Runs any activity at any Tier level
  - More flexible than atlas in the atlas transfer cohices
- Actively working on federated storage based also on xrootd (AAA)
  - “Any data, anywhere, anytime”
  - Starting from a regional approach
    - Eventually going global with finer grained level of redirectors



LHCb

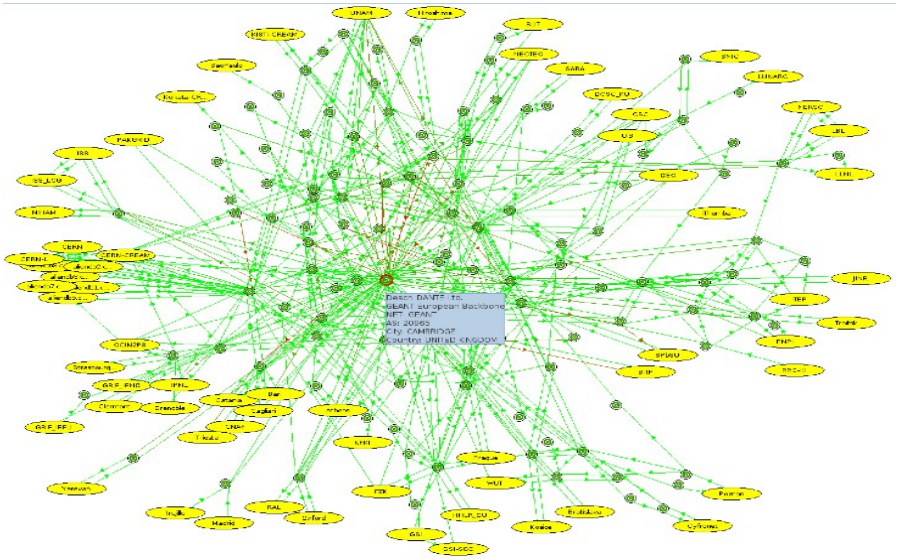
- Lhcb data access model is evolving
  - Most of the activities still at T1s.
  - Several T2s that have been and will be used as "co-processing" sites
    - During that time the RAW file will be downloaded from a close T1 and the output uploaded again to T1 storage. Therefore a network monitoring between those sites is essential for the proper operation.
  - In the future it is possible that a selected number of T2s will be even tighter integrated with the execution of more workflows (analysis)
  - The possibility to use federated storage, which will further extend the usage and needs for/of network monitoring.
- Model is hierarchical





Alice

- Jobs go where the data are
  - Access the closest SE
    - 216 PB read in 2012
- Use xrootd only
  - Other protocols supported
- Network monitoring provided by Monalisa
  - Information from Monalisa already used to broker jobs.
- Perfsonar might simplify this scheme
  - http://tinyurl.com/cl3ds73

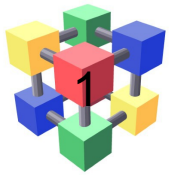


No.	ID	Site	Speed (Mbps)	Hops	RTT (ms)	Streams
1.	1253813	CCIN2P31	938.64			1
2.	1384342	CERN-CREAM	930.25	1	2.75	1
3.	1384587	CCIN2P3	913.49			1
4.	1384122	Bari	912.58			1
5.	1387530	Legnaro	846.45	10	164.08	1
6.	1386099	IPNL	822.13			1
7.	1386953	GRIF_IRFU	771.02			1
8.	1384303	Clermont	737.50			1
9.	1386930	Subatech	705.31			1
10.	1385151	GRIF_IPNO	653.69			1
11.	1387353	CSC	595.62			1
12.	1384472	CNAF	505.66			1
13.	1498080	GSI-SE2	159.39			1
14.	631699	Dortmund	64.55			1
15.	1384871	FZK	25.17			1
16.	1386040	Athens				1
17.	1375150	Birmingham				1
18.	1387115	BITP				1
19.	1384352	Bologna				1
20.	1385604	Bratislava				1
21.	1386161	Cagliari				1
22.	1388335	Catania				1

No.	ID	Site	Speed (Mbps)	Hops	RTT (ms)	Streams
1.	1384708	Legnaro	939.57	12	11.54	1
2.	1387396	GRIF_IPNO	938.64			1
3.	1385081	IPNL	938.64	9	5.41	1
4.	1384421	Bari	930.25	12	22.46	1
5.	1385211	CERN-CREAM	930.25	1	27.97	1
6.	1386451	GRIF_IRFU	930.25			1
7.	1383856	SARA	904.21			1
8.	1384894	Subatech	880.85	9	18.32	1
9.	1388581	Oxford	838.07	17	19.82	1
10.	1387315	Trieste	838.07	17	15.02	1
11.	1385813	Athens	654.35	13	34.71	1
12.	1387621	Kosice	612.40			1
13.	1267784	PDC	603.41	17	38.67	1
14.	1384835	UIB	569.89	18	51.41	1
15.	1388355	NIPNE	544.74	19	47.11	1
16.	1384516	LUNARC	494.46			1
17.	1386944	DCSC_KU	444.18			1
18.	1384264	JINR	377.51	13	63.23	1
19.	1385671	ISS_LCG	352.34	20	38.74	1
20.	1384799	Poznan	352.34			1
21.	1388044	CNAF	226.50	7	14.72	1
22.	1386166	Trujillo	226.50			1

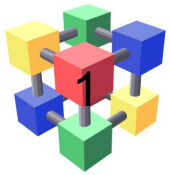
Motivations for Monitoring

- LHC collaborations are:
  - Data intensive
  - Globally distributed
  - Rely upon the network as a critical part of their infrastructure
- Finding and debugging LHC network problems can be difficult and, in some cases, take months.
  - How can we quickly identify when problems are network problems and help isolate their locations?
  - Experiments might want to blacklist
- We don't want to have a network monitoring system per VO!



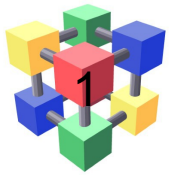
WLCG Ops&Tools TEG R5

- R5: WLCG Network Monitoring: deploy a WLCG-wide and experiment independent monitoring system for network connectivity
  - It is suggested that the PerfSONAR network monitoring system is deployed at all WLCG sites (two boxes, one for throughput and one for latency tests). This should help debug and resolve network-related problems which in the past have sometimes taken a very long time to resolve (many months) and for which the responsibilities have not easily been agreed. [...] The network monitoring metrics should be exposed both programmatically and through a dashboard-like interface. Commonalities with the FTS monitoring should be leveraged in order to provide a unique and complete network and transfers monitoring system.
    - WLCG Ops&Tools TEG final report



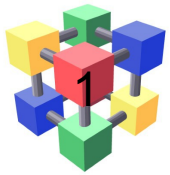
perfSONAR TF

- Main goal assure that most WLCG sites install perfsonar
- Put together a deployment scenario from experiment models and priorities
  - ATLAS 3 categories of sites: OPN (including T0 and T1s), T2D, T2 (including T2 and T3)
    - Priorities
      - Priority 1: OPN-OPN links
      - Priority 2: Tx-Tx links in the same cloud
      - Priority 3: T1-T2D links (different cloud)
      - Priority 4: T2D-T2D links (different cloud)
      - Priority 5: all other links
  - Experiments deployment scenarios



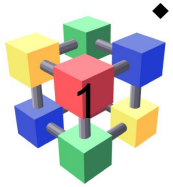
perfSONAR TF

- Main goal assure that most WLCG sites install perfsonar
- Put together a deployment scenario from experiment models and priorities
  - ATLAS 3 categories of sites: OPN (including T0 and T1s), T2D, T2 (including T2 and T3)
    - Priorities
      - Priority 1: OPN-OPN links
      - Priority 2: Tx-Tx links in the same cloud
      - Priority 3: T1-T2D links (different cloud)
      - Priority 4: T2D-T2D links (different cloud)
      - Priority 5: all other links
  - Experiments deployment scenarios



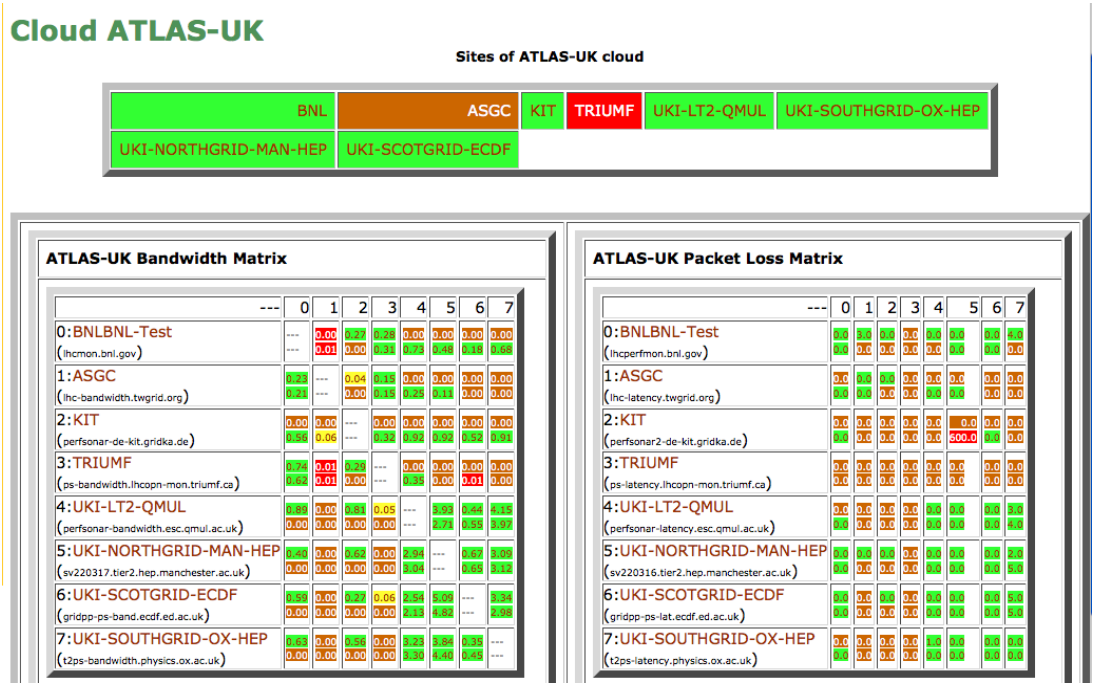
PerfSONAR TF (2)

- Recommend hardware and setup
  - Location: perfSONAR instances useful if they are local to the storage
  - Networking: network config & hardware should be similar as much as possible to the storage one
    - If you use bonding on one use it also on the other
  - OS: different OS might behave differently
- Simplify perfSONAR configuration for sites
  - At the moment mostly manual and painful
  - Introduced concept of centralized mesh tests, i.e. machines read one or more central configurations.
  - Each experiment can have a set of meshes the manage centrally
    - US, IT, UK already have at least a centralised meshes



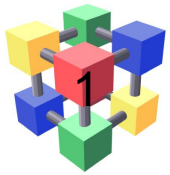
BNL dashboard

- Each instance of perfsonar gives a site view from that site
- Global view needed
  - Different sites can be arranged in different views
    - Example Atlas UK sites vs some problematic T1



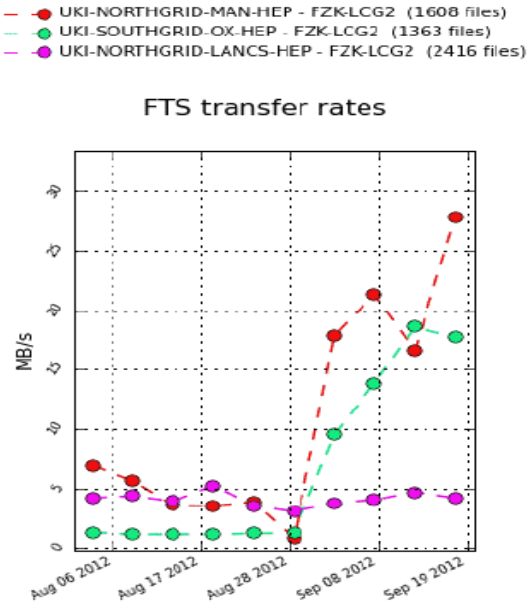
PerfSONAR TF (3)

- Simplify also the installation as much as possible to a out-of-the box style
- Get the perfSONAR services properly handled
  - Publication of each service in GOCDB
    - How to publish perfsonar in GOCDB
    - Handling of downtimes
  - Monitoring of services in nagios/sum tests
    - Only to check services are working
    - Several tests currently used to blacklist or downgrade sites depending on the tests
    - There are no proper low level network tests

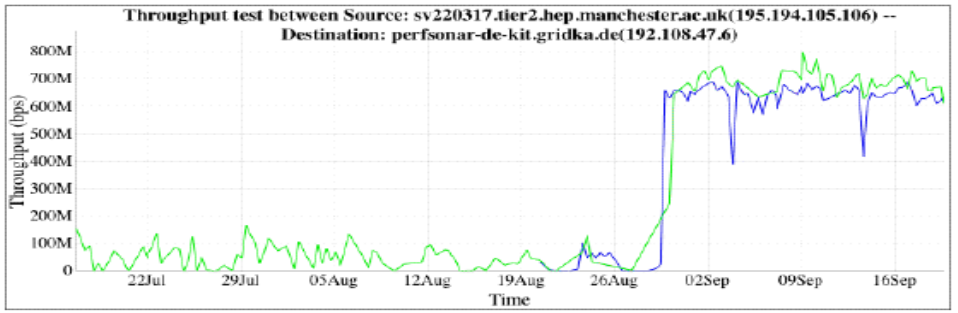


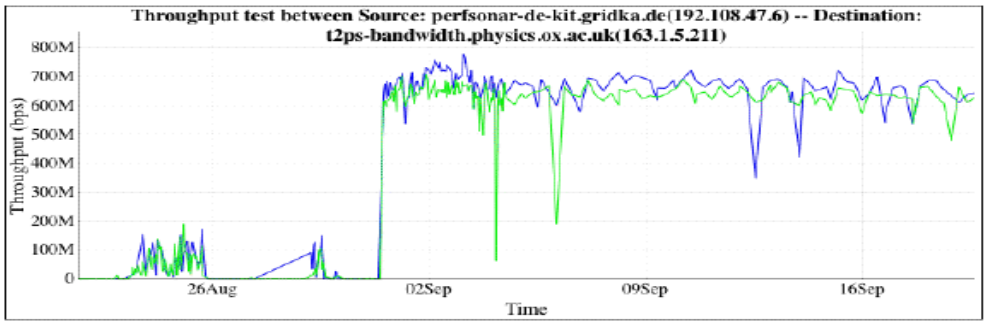
UK T2s → FZK

- Many UK sites had a problem in the Atlas sonar tests with FZK for several months
- Most UK sites installed perfsonar and perfSonar throughput was also really poor
  - Diagnosed problem with FZK firewall
  - Few sites bypassed firewall and there was a dramatic improvement



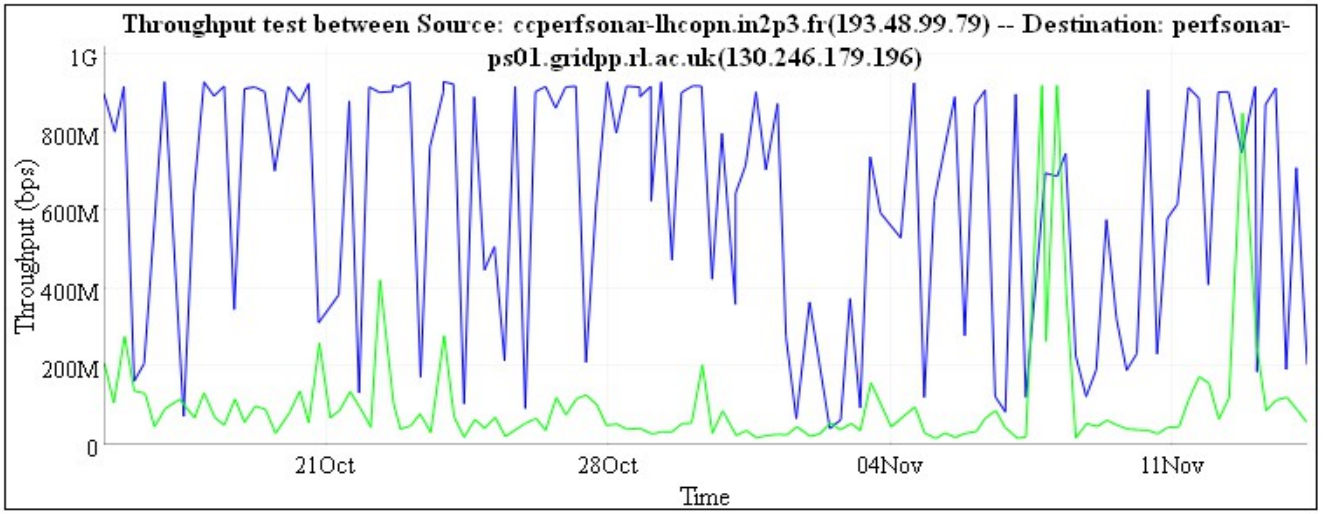
GGUS-Ticket-ID: #84008





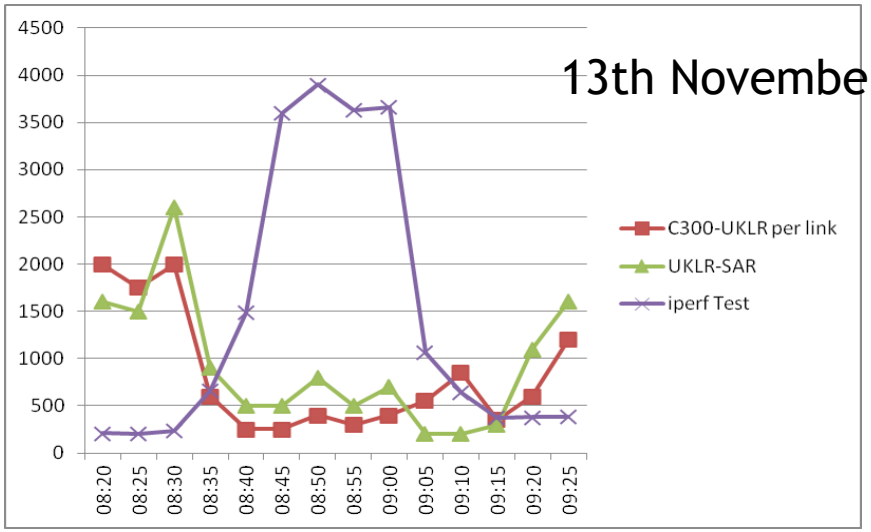
RAL T1 experience

- Background was that in October we noted that our perfsonar performance showed a considerable asymmetry between inbound and outbound rates. Was worse as distance increased.
  - First problem we found was assymetric routing from some sites on the OPN to RAL. Identified this using the perfsonar traceroute functionality.Tracked down to a number of Tier-1s not accepting our new prefixes following an enlargement of our OPN subnet. Corrected this problem after dialogue with sites concerned.



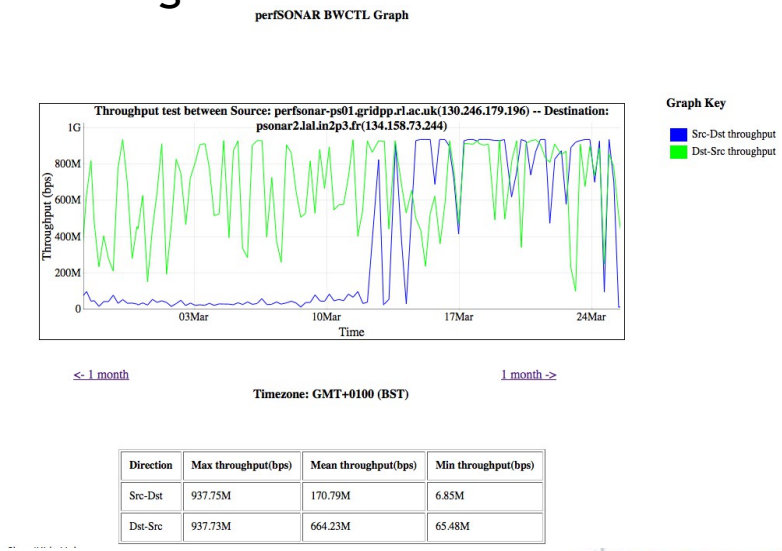
RAL T1 experience

- We verified the perfsonar results using iperf and other tests
  - Link aggregation protocol set incorrectly on Nortel to Force10 switch
  - Suspicions raised that Force10 C300 might be losing packets
- Carried out intervention replacing switch and currently running without agregation.
  - Result no packet loss and outbound performance now excellent. Indeed seems better than inbound now.



Castor off transfers draining

Mb/s perfSONAR



Direction	Max throughput(bps)	Mean throughput(bps)	Min throughput(bps)
Src-Dst	937.75M	170.79M	6.85M
Dst-Src	937.73M	664.23M	65.48M

Conclusions

- Experiment models are evolving from a hierarchical with well defined transfer paths to a mesh of transfers with different priorities.
  - Asyncronous transfers more dynamic respect to a couple of years ago
  - Experiment are extending their activities to all type of sites
    - Wide variety of file sizes and type of traffic
  - Introduction of federated storage
  - Future already talking about network on demand appication for both CMS and Atlas
    - Network needs to be monitored
    - Applications need to be instrumented