

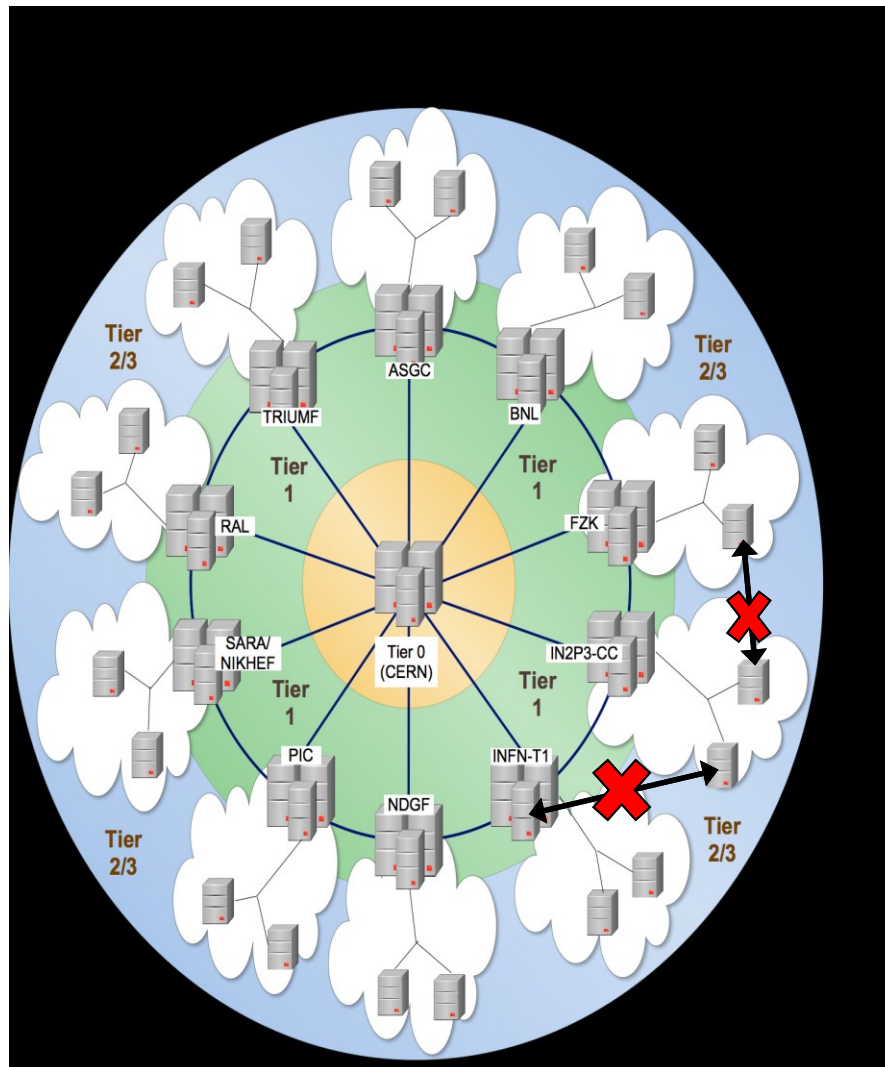
# ATLAS clouds and T2Ds

Simone Campana  
CERN-IT/ES

Several slides and material are courtesy of

- Fernando Barreiro
- Shawn McKee
- Hironori Ito





Hierarchical Model based on the Monarc network topology

Possible communications:

- T0-T1
- T1-T1
- Intra-cloud T1-T2

Forbidden communications:

- Inter-cloud T1-T2
- Inter-cloud T2-T2

- Consolidation of User Analysis outputs problematic
  - Analysis runs over many cloud
  - Consolidation needs to “hop” the data through the T1s
- MonteCarlo production must confine one task to one cloud
  - To facilitate the output aggregation at T1
- PD2P more inflexible
  - Need to replicate from T1 to T2s of the same cloud
  - Or “hop” through a T1
- T2s can not really be used as storage of “primary” data
  - Issues in creating secondary copies at other T1s

- The network model today does not really resemble the Monarc model
  - Many T2s are connected very well with many T1s
  - Many T2s are not that well connected with their T1
- So it makes sense to break cloud boundaries.
  - Let DDM freely transfer from every site to every site
- Not quite there
  - Some links simply have limited bandwidth
  - In those cases, several hops will anyway be needed
- Defining T2Ds is an attempt to break cloud boundaries “for the cases where it makes sense”

- “D” stays for “Directly connected”
  - In DDM, every T1 is topologically closed with the other T1s and its T2s
  - In the new model, every T1 is topologically close with the other T1s, its T2s and all the T2Ds
- Requirements for being/becoming a T2D
  - Need satisfy transfer metrics with all T1s
    - Strictly quantifiable, see later
  - Need to provide a certain level of commitment and reliability
    - Not quantifiable at the moment, up to ATLAS central ops to decide



- The sonar test runs weekly
  - Transfer submission smeared from Monday to Friday
  - Cancelled 5 days later
  - Statistics collected once per week
- 10 to 15 FTS transfers between each pair of sites:  $O(10,000)$  pairs
  - 1 datasets with 5 small files: 20 MB each
  - 1 datasets with 5 medium files: 200 MB each
  - 1 datasets with 5 big files: 2 GB each
    - For a subset of pairs only
- Source datasets are pre-placed at sites
  - Destination datasets cleaned as mentioned above

[http://bourricot.cern.ch/dq2/ftsmon/sonar\\_view/cached/](http://bourricot.cern.ch/dq2/ftsmon/sonar_view/cached/)

Only DATADISK to DATADISK transfers are shown

Show  entries Search all columns:

Prio	Source	SCloud	Destination	DCloud	SMALL FILES			MEDIUM FILES			LARGE FILES		
					MB/s	MB	#Ev	MB/s	MB	#Ev	MB/s	GB	#Ev
10	CERN-PROD	CERN - T0	FZK-LCG2	DE - T1	0.76+0.45	16.68+8.13	6	8.08+2.26	230.4+40.7	209	18.21+6.83	3.36+0.72	844
10	CERN-PROD	CERN - T0	NDGF-T1	NG - T1	1.5+0.23	20.0+0.0	5	10.72+3.15	200.0+0.0	5	19.71+17.58	2.0+0.0	5
10	CERN-PROD	CERN - T0	PIC	ES - T1	1.13+0.08	20.0+0.0	5	8.03+3.74	200.0+0.0	5	20.02+6.26	2.0+0.0	5
10	CERN-PROD	CERN - T0	NIKHEF-ELPROD	NL - T1	0.78+0.27	20.0+0.0	5	6.27+2.46	200.0+0.0	5	38.95+12.7	2.0+0.0	5
10	CERN-PROD	CERN - T0	INFN-T1	IT - T1	0.86+0.02	20.0+0.0	5	7.36+1.54	234.6+40.42	171	38.45+4.35	2.0+0.0	5
10	CERN-PROD	CERN - T0	TAIWAN-LCG2	TW - T1	0.27+0.03	20.0+0.0	5	1.96+0.2	200.0+0.0	5	11.38+2.39	2.0+0.0	5
10	CERN-PROD	CERN - T0	BNL-OSG2	US - T1	0.61+0.2	20.0+0.0	5	5.14+1.38	200.0+0.0	5	10.44+4.03	2.0+0.0	5
10	CERN-PROD	CERN - T0	RAL-LCG2	UK - T1	0.45+0.02	20.0+0.0	5	2.87+1.27	200.0+0.0	5	14.31+7.36	2.0+0.0	5
10	CERN-PROD	CERN - T0	TRIUMF-LCG2	CA - T1	0.66+0.54	25.0+21.47	24	5.91+2.5	414.56+176.47	33	11.01+4.31	2.98+1.09	298
10	CERN-PROD	CERN - T0	IN2P3-CC	FR - T1	1.73+1.41	32.33+27.23	1976	8.96+4.74	273.6+202.71	1620	11.21+5.58	1.12+0.19	252
10	CERN-PROD	CERN - T0	SARA-MATRIX	NL - T1	1.18+0.33	20.0+0.0	5	8.63+3.85	200.0+0.0	5	27.64+18.49	2.0+0.0	5
9	FZK-LCG2	DE - T1	NDGF-T1	NG - T1	0.08+0.07	1.15+0.96	1920	5.7+3.0	204.46+10.93	6	15.34+11.71	2.0+0.0	5
9	FZK-LCG2	DE - T1	PIC	ES - T1	0.23+0.74	5.43+17.09	2200	6.64+2.51	210.11+69.6	1679	13.25+10.24	3.57+0.93	218
9	FZK-LCG2	DE - T1	NIKHEF-ELPROD	NL - T1	1.0+0.0	20.0+0.0	5	9.29+0.53	200.0+0.0	5	22.48+5.23	2.0+0.0	5
9	FZK-LCG2	DE - T1	CERN-PROD	CERN - T0	0.58+0.72	24.65+30.82	6435	5.36+3.45	263.22+192.81	2608	21.3+4.14	1.39+0.08	540
9	FZK-LCG2	DE - T1	INFN-T1	IT - T1	0.2+0.48	4.94+12.2	2271	6.75+1.82	228.19+88.55	1416	19.8+5.3	3.95+1.02	408
9	FZK-LCG2	DE - T1	TAIWAN-LCG2	TW - T1	0.02+0.01	1.15+0.96	1943	1.85+0.31	200.0+0.0	5	10.91+1.6	2.0+0.0	5
9	FZK-LCG2	DE - T1	BNL-OSG2	US - T1	0.04+0.04	1.15+0.99	1819	5.68+1.1	200.0+0.0	5	25.84+4.02	2.0+0.0	5
9	FZK-LCG2	DE - T1	RAL-LCG2	UK - T1	0.06+0.18	2.78+9.17	2086	3.63+1.11	267.67+140.67	765	8.45+1.74	3.44+1.07	253
9	FZK-LCG2	DE - T1	TRIUMF-LCG2	CA - T1	0.22+0.48	6.38+14.96	2356	5.74+2.2	251.63+100.41	550	17.06+6.47	3.29+0.66	548
9	FZK-LCG2	DE - T1	IN2P3-CC	FR - T1	1.69+1.93	29.47+33.26	7795	9.92+5.74	252.08+150.63	4899	29.67+10.83	2.52+1.53	859
9	FZK-LCG2	DE - T1	SARA-MATRIX	NL - T1	0.11+0.16	1.63+2.84	2071	29.29+9.21	701.71+238.52	428	54.93+19.99	3.09+1.25	251
9	NDGF-T1	NG - T1	FZK-LCG2	DE - T1	0.84+1.29	16.19+21.33	1561	12.04+4.29	234.43+71.49	1421	23.02+18.04	3.87+0.87	335
9	NDGF-T1	NG - T1	PIC	ES - T1	0.32+0.78	7.97+20.94	1179	4.16+3.24	212.64+84.89	1278	15.7+10.3	3.92+0.82	544
9	NDGF-T1	NG - T1	NIKHEF-ELPROD	NL - T1	1.32+0.04	20.0+0.0	5	11.15+0.72	200.0+0.0	5	47.69+13.42	2.0+0.0	5

Filter pr  Filter source  Source clou  Filter dest  Dest cloud

Showing 1 to 25 of 7,657 entries First Previous 1 2 3 4 5 Next Last

### Number of files transferred

<b>SMALL</b>	≤3	4	≥5
<b>MEDIUM</b>	≤2	3	≥4
<b>LARGE</b>	≤1	2	≥3

### Avg(ByteRate)+StD(ByteRate)

<b>SMALL</b>	<0.05MB/s	<0.1MB/s	≥0.1MB/s
<b>MEDIUM</b>	<1MB/s	<2MB/s	≥2MB/s
<b>LARGE</b>	<10MB/s	<15MB/s	≥15MB/s





<http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview?view=Sonar>

Show 25 entries Copy Print Save Search...

Site Name	SrcSite	SrcCloud	SrcTier	DstSite	DstCloud	DstTier	AvgBRS(MB/s)	EvS	AvgBRM(MB/s)	EvM	AvgBRL(MB/s)	EvL	Prio
<a href="#">AGLT2 to AUSTRALIA-ATLAS</a>	AGLT2	US	T2D	<a href="#">AUSTRALIA-ATLAS</a>	CA	T2	0.49+/-0.01	5	3.59+/-0.38	5	0.00+/-0.00	0	2
<a href="#">AGLT2 to BEIJING-LCG2</a>	AGLT2	US	T2D	<a href="#">BEIJING-LCG2</a>	FR	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
<a href="#">AGLT2 to BNL-OSG2</a>	AGLT2	US	T2D	<a href="#">BNL-OSG2</a>	US	T1	1.04+/-0.05	5	7.89+/-0.35	5	38.15+/-10.40	5	8
<a href="#">AGLT2 to CA-ALBERTA-WESTGRID-T2</a>	AGLT2	US	T2D	<a href="#">CA-ALBERTA-WESTGRID-T2</a>	CA	T2	0.75+/-0.08	5	4.95+/-0.51	5	0.00+/-0.00	0	2
<a href="#">AGLT2 to CA-SCINET-T2</a>	AGLT2	US	T2D	<a href="#">CA-SCINET-T2</a>	CA	T2	0.83+/-0.01	5	5.57+/-0.32	5	0.00+/-0.00	0	2
<a href="#">AGLT2 to CA-VICTORIA-WESTGRID-T2</a>	AGLT2	US	T2D	<a href="#">CA-VICTORIA-WESTGRID-T2</a>	CA	T2	0.94+/-0.05	5	6.84+/-0.83	5	0.00+/-0.00	0	2
<a href="#">AGLT2 to CERN-PROD</a>	AGLT2	US	T2D	<a href="#">CERN-PROD</a>	CERN	T0	0.65+/-0.09	5	4.63+/-0.33	5	8.87+/-5.17	5	7
<a href="#">AGLT2 to CSCS-LCG2</a>	AGLT2	US	T2D	<a href="#">CSCS-LCG2</a>	DE	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
<a href="#">AGLT2 to CSTCDIE</a>	AGLT2	US	T2D	<a href="#">CSTCDIE</a>	NL	T3	0.76+/-0.02	5	5.32+/-1.17	5	0.00+/-0.00	0	0
<a href="#">AGLT2 to CYFRONET-LCG2</a>	AGLT2	US	T2D	<a href="#">CYFRONET-LCG2</a>	DE	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
<a href="#">AGLT2 to DESY-HH</a>	AGLT2	US	T2D	<a href="#">DESY-HH</a>	DE	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5
<a href="#">AGLT2 to DESY-ZN</a>	AGLT2	US	T2D	<a href="#">DESY-ZN</a>	DE	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5
<a href="#">AGLT2 to FZK-LCG2</a>	AGLT2	US	T2D	<a href="#">FZK-LCG2</a>	DE	T1	0.00+/-0.00	0	0.00+/-0.00	0	10.14+/-9.02	390	7
<a href="#">AGLT2 to GOEGRID</a>	AGLT2	US	T2D	<a href="#">GOEGRID</a>	DE	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
<a href="#">AGLT2 to GRIF-IRFU</a>	AGLT2	US	T2D	<a href="#">GRIF-IRFU</a>	FR	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
<a href="#">AGLT2 to GRIF-LAL</a>	AGLT2	US	T2D	<a href="#">GRIF-LAL</a>	FR	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5
<a href="#">AGLT2 to GRIF-LPNHE</a>	AGLT2	US	T2D	<a href="#">GRIF-LPNHE</a>	FR	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5

Showing 1 to 25 of 7,310 entries DB query took 4.4681 s First Previous 1 2 3 4 5 Next Last

Offers at the moment the same functionality as the previous one  
Will offer in the next release “History” views

T2Ds in DDM as of  
TODAY:

INFN-NAPOLI- ATLAS  
 IFIC-LCG2  
 GRIF-LPNHE  
 GRIF-LAL  
 DESY-HH  
 DESY-ZN  
 MWT2  
 SLACXRD  
 AGLT2

Same list as one  
month ago

## Proto (tentative) T2Ds

INFN-MILANO-ATLASC  
 INFN-ROMA1\*\*  
 UK UKI-LT2-QMUL  
 UKI-NORTHGRID-LANCS-HEP  
 UKI-NORTHGRID-MAN-HEP  
 UKI-SCOTGRID-GLASGOW\*\*  
 TOKYO-LCG2\*\*  
 LRZ-LMU\*\*  
 MPPMU\*\*  
 IFAE\*\*  
 UAM-LCG2\*\*  
 NET2\*\*  
 SWT2\_CPB\*\*

\*\* Ready to be Approved



- Looks like we are doing pretty good
  - Many T2Ds commissioned in less than 2 months
  - Probably representing  $\gg 80\%$  of resources
- But it is not really as good ...
  - Some clouds have no T2Ds
  - Less than 50% of T2s in the two lists
  - We are very permissive in the commissioning metrics
- The currently used network is general purpose
  - LHC experiments could suffer from other communities activities
  - Other communities could suffer from LHC experiments activities
  - LHC experiments could suffer from their own activity
- See next talk about LHCONE

- Disclaimer: nothing on this slide has been agreed
  - But IMHO it would make sense
- T2Ds should be the only “multi cloud” T2s
  - TBD in Panda and Prodsys session on Wed
- PD2P should be enabled for T1-T2Ds regardless the cloud
  - TBD in Panda and Prodsys session on Wed
- T2s hosting “GROUPDISK” should be T2Ds
  - TBD somewhere, but please discuss
- Surely, local users on T2Ds will have an advantage
  - Faster dq2-get, faster (no multi-hops) DaTRI requests

- If we could commission T2Ds to T2Ds channels we could explore more possibilities
- Using T2Ds for “primary replicas”
  - This would relieve a lot of pressure from T1, especially in term of providing disk space
- More flexibility for GROUP data at T2s
  - Replication from T2D groupdisk to T2D groupdisk across cloud
  - Wider PD2P for groupdisk

- It is in the site interest to become T2D
- It is in the cloud interest to have as many sites as possible flagged as T2D
- It is in the T1 interest to support as many T2Ds as possible
- The commissioning effort cannot be taken by central operations
  - Central operations can do the overall coordination
  - But it is up to squads (and sites) to do the commissioning
- Several sites/squads have been very reactive
  - Some other, less reactive



- perfSONAR in a nutshell:
  - Memory-to-Memory test between two perfSONAR boxes.
    - 9 sites: BNL + 8 T2s
  - Boxes located close to the gridftp/storage in terms of the network
  - Two types of active tests: throughput and latency.
- Site throughput test
  - runs to every other site each 4-hour window in both directions.
  - Simple 'iperf' (memory-to-memory) test between sites measuring achievable bandwidth.
  - Test results and average over all the measurements in the last 24 hours:
  - Metrics:
    - Any result > 100 Mbits/sec as "OK" (for now).
    - Between 10-100 Mbits/sec is "WARNING"
    - below 10 Mbits/sec is "CRITICAL".
- Latency tests
  - run between nodes running 'ntpd'
    - 600 packets sent every minute between all sites both directions
  - Metrics not yet based on the measured one-way delay values
  - Metrics on average packet loss integrated over 30 mins
    - "OK" If the loss is < 2 packets, "WARNING" if the loss is >=2 and <10 and "CRITICAL" if >=10 packets lost.

- Disk-To-Disk test between two ATLAS storages
  - in between US T1s/T2s/T3s as wells to/from foreign T1s
- Input files are pre-placed at source sites.
  - Files at the destination deleted once transfers are completed
- Transfer initiated by BNL FTS
- Tests are done once to twice a day.
  - Depending on the pairs
- Results are measured using the values from FTS and stored in the FTS monitor
  - <http://www.usatlas.bnl.gov/dq2/throughput>
  - The monitor is independent from the actual test: any other results can be inserted into the same monitor via http Restful way.
  - The inclusion of perfSONAR results to the same monitor are in the plan.

- Network commissioning and monitoring should be a common and coherent effort
  - Inside the experiment
  - Between experiments
  - With network providers, LHCOPN, LHCONE
- Common effort on a Network Monitoring Task Force proposed by Shawn McKee
  - Extend iperSonar
  - Merge FTS test and Sonar tests
  - Extend to/integrate with/import from other experiments
  - Integrate into the WLCG infrastructure
- ATLAS Distributed Computing Management unanimously agree it is a good idea.