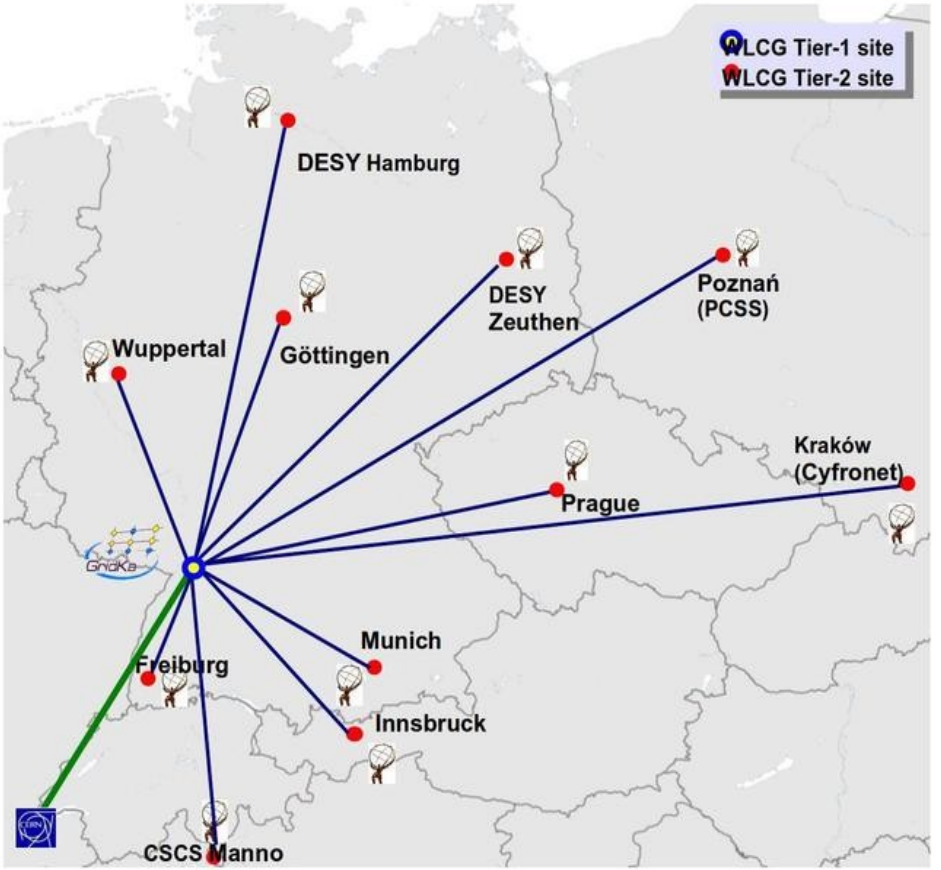
Data transfer and access for ATLAS DE cloud

Jan 24, 2011
LHC Networking
Goettingen

- Overview ATLAS data distribution
- How to analyze transfers
- Results for 2010
- Implications for networking reqts



GridKa cloud Resources

Pledges	July 2010	April 2011	July 2010	April 2011
Site	CPU(HS06)		DISK(TB)	
GridKa-T1	**21600**	***28250***	**2190**	***3125***
DESY-HH/ZN	4800	*6200*	740	*1050*
Göttingen	2280	*3800*	250	*400*
Wup	3440	*4633*	370	*633*
Freiburg	3440	*4609*	370	*518*
LMU/LRZ	3440	*4609*	370	*518*
MPI/RZG	3440	*4609*	370	*518*
UIBK/A	1850	*1857*	94	*120*
PL Federation	3180	*4000*	285	*385*
CSCS/CH	3063	*5420*	364	*469*
FZU/CZ	1783	*3500*	150	*400*
Sum T2	**30716**	**43237**	**3363**	**5011**

- Status of pledges accdg WLCG:
  - http://lcg.web.cern.ch/LCG/Resources/WLCGResources-2010-2012_04OCT2010.pdf
- substantial increase (30-40%) of resources for 2011

Overview ATLAS data distribution

- Original (pre-2010) ATLAS Computing model:
  - Standard data-flow to T2s goes via T1 of same cloud
    - All centrally produced data (both LHC and MC) go via T1 to T2s
    - Data produced in simulation jobs at T2s goes back to cloud T1
  - Exceptions:
    - Muon calibration stream directly from Cern (MPPMU/LRZ-LMU sites)
    - Re-distribution of data available on other T2 in same cloud can go directly to T2
      - Rarely the case, group data or recovery from file loss
  - In addition un-managed data download by users to arbitrary sites, not directly controlled by ATLAS .
  - Very little T2-T2 traffic in cloud expected

Overview ATLAS data distribution - 2

- Substantial CompModel modifications under discussion/in progress:
  - Main issue is how to identify data really used and get it to sites …
  - Dynamic Data distribution (PD2P) in operation since Oct 2010:
    - Datasets case get replicated from GridKa to Tier-2 Sites in case corresponding analysis jobs go to GridKa (or any other Tier-1)
    - Still most traffic directed via cloud T1 site
  - Further changes/optimizations for Group data
  - Direct transfers between arbitrary T1/T2 ↔ T2 sites will be next step and are currently tested
    - Potentially changes data transfer pattern

How to analyze data transfers/access

- Many different sources available:
  - ATLAS DQ2 DDM system
    - Standard monitoring pages
    - Detailed history of all transfers in DB
    - Caveat: not full picture, doesn't include
      - local/job up-/down-loads dq2-get/put, lcg-cp, …
      - Analysis job access of local storage system
  - dCache storage has detailed 'billing logs' of each transfer
    - Includes peer site of gridftp transfer
    - Can get huge – dedicated analysis script (J.Schultes) to generate report
  - Network router logs
- Cross-check/combine to get full & consistent picture

Questions to answer ...

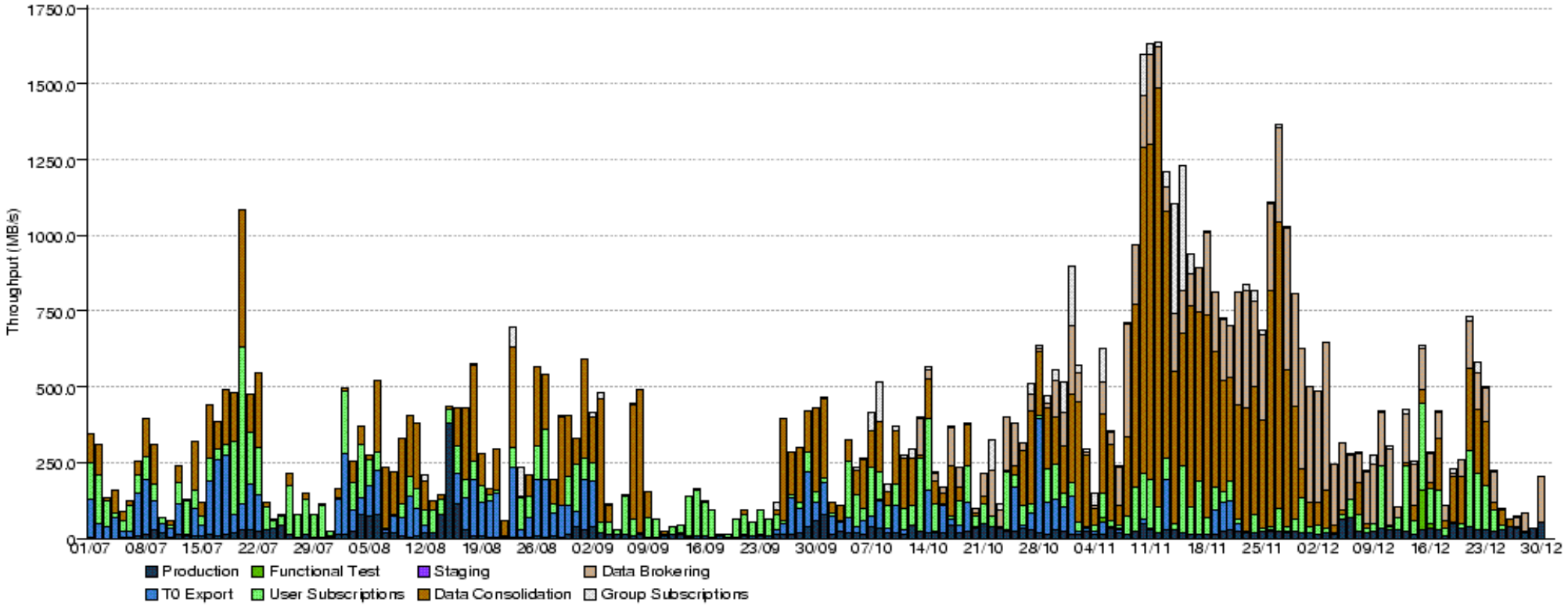
- Transfer rates/used bandwidth
  - Per link, per site, whole cloud
  - Averaged over hour, day, month, year, …
  - Per operations use-case:
    - Data export
    - MC production
    - Analysis jobs
    - DaTri (user replication requests)
    - Direct upload/download
  - ....

ATLAS DQ2 DDM for DE cloud

	FZK-LCG2	CSCS-LCG2	CYFRONET-LCG2	DESY-HH	DESY-ZN	LRZ-LMU	MPPMU	GOEGRID	HEPHY-UIBK	PRAGUELCG2	UNI-FREIBURG	UNI-DORTMUND	PSNC	WUPPERTALPROD	TUDRESDEN-ZIH	UNI-SIEGEN-HEP	CERN-PROD	outsidecloud	
FZK-LCG2	1.90E+011	1.03E+012	4.13E+012	6.78E+012	2.26E+012	4.87E+012	4.47E+012	4.16E+012	2.33E+012	4.04E+012	1.94E+012	5.54E+011	1.02E+012	2.62E+012	5.24E+011	0.00E+000	1.01E+014		1.20E+015
CSCS-LCG2	9.21E+013	0.00E+000	2.68E+011	1.11E+012	6.31E+012	2.39E+012	7.32E+011	4.00E+012	1.00E+009	2.06E+011	8.39E+012	1.00E+009	1.00E+009	3.42E+010	1.00E+009	1.00E+009	1.00E+009		7.60E+010
CYFRONET-LCG2	6.35E+013	1.01E+009	2.83E+011	1.19E+009	1.00E+009	1.00E+009	1.00E+009	1.33E+009	1.00E+009	1.31E+009	1.39E+009	1.00E+009	1.50E+009	1.48E+009	1.03E+009	1.00E+009	1.00E+009		7.60E+010
DESY-HH	1.54E+014	1.28E+012	1.52E+010	7.35E+011	1.74E+013	3.74E+011	7.32E+011	2.35E+012	5.12E+010	1.38E+011	3.30E+012	1.00E+009	1.00E+009	8.96E+011	1.00E+009	1.00E+009	1.00E+009		7.56E+010
DESY-ZN	1.26E+014	3.55E+012	3.42E+012	8.23E+012	1.64E+012	4.06E+012	2.60E+011	7.63E+011	4.41E+011	8.41E+011	2.82E+012	1.00E+009	1.00E+009	8.14E+011	1.00E+009	1.00E+009	1.00E+009		7.60E+010
LRZ-LMU	6.38E+013	1.00E+009	1.00E+009	1.00E+009	1.14E+012	0.00E+000	1.00E+009	1.00E+009	1.00E+009	1.07E+012	9.24E+011	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	4.72E+011		7.35E+011
MPPMU	3.40E+013	1.27E+011	6.14E+010	9.99E+011	1.00E+009	4.99E+012	0.00E+000	9.70E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	5.17E+011		7.70E+010
GOEGRID	1.01E+014	9.73E+011	1.00E+009	1.52E+010	1.28E+011	1.00E+009	3.90E+011	0.00E+000	1.00E+009	3.08E+011	2.92E+011	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009		7.60E+010
HEPHY-UIBK	3.00E+012	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	0.00E+000	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009		7.60E+010
PRAGUELCG2	5.72E+013	1.00E+009	9.81E+011	8.94E+010	3.95E+009	1.00E+009	7.11E+011	1.00E+009	1.00E+009	0.00E+000	1.00E+009	1.00E+009	1.00E+009	1.30E+012	1.00E+009	1.00E+009	1.00E+009		7.60E+010
UNI-FREIBURG	9.14E+013	2.44E+011	7.48E+011	6.15E+011	1.53E+010	1.58E+010	2.31E+009	5.37E+011	1.00E+009	4.55E+010	1.06E+010	1.00E+009	1.00E+009	1.33E+011	1.00E+009	1.00E+009	1.00E+009		7.56E+010
UNI-DORTMUND	7.95E+011	1.00E+009	1.28E+009	1.43E+009	1.41E+009	1.56E+009	1.05E+009	1.14E+009	1.09E+009	1.77E+009	1.34E+009	0.00E+000	1.54E+009	1.00E+009	1.31E+009	1.00E+009	1.00E+009		7.52E+010
PSNC	8.39E+011	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	0.00E+000	1.00E+009	1.00E+009	1.00E+009	1.00E+009		7.60E+010
WUPPERTALPROD	4.95E+013	1.02E+009	3.19E+009	1.09E+012	3.53E+011	4.81E+010	6.66E+011	9.88E+009	1.00E+009	1.24E+011	1.96E+011	1.00E+009	1.00E+009	5.18E+009	1.00E+009	1.00E+009	1.00E+009		2.71E+011
TUDRESDEN-ZIH	1.56E+011	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	0.00E+000	1.00E+009	1.00E+009		7.60E+010
UNI-SIEGEN-HEP	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	1.00E+009	0.00E+000	1.00E+009		7.60E+010

- Can get (from C.Serfon) full transfer matrix of between all sites in cloud (+external)
  - Number of transfers, data volume, ..., for arbitrary periods

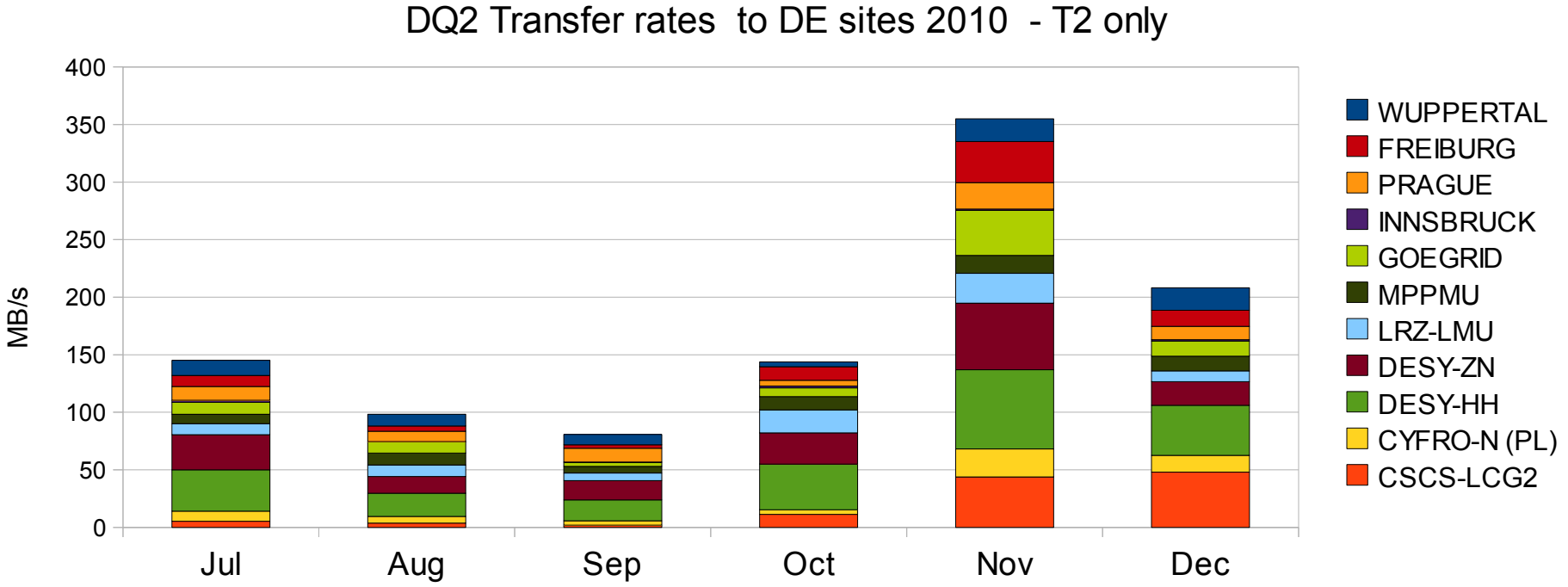
ATLAS DQ2 DE cloud data distribution overall July10-Dec10



- Includes T0 export and T1 ↔ T1 data exchange via LHC-OPN
- Peak rates exceeded 1 GB/s
  - Peak in Nov due to reprocessing
- replication of user data contributes significantly

ATLAS DQ2 data distribution in DE cloud -monthly

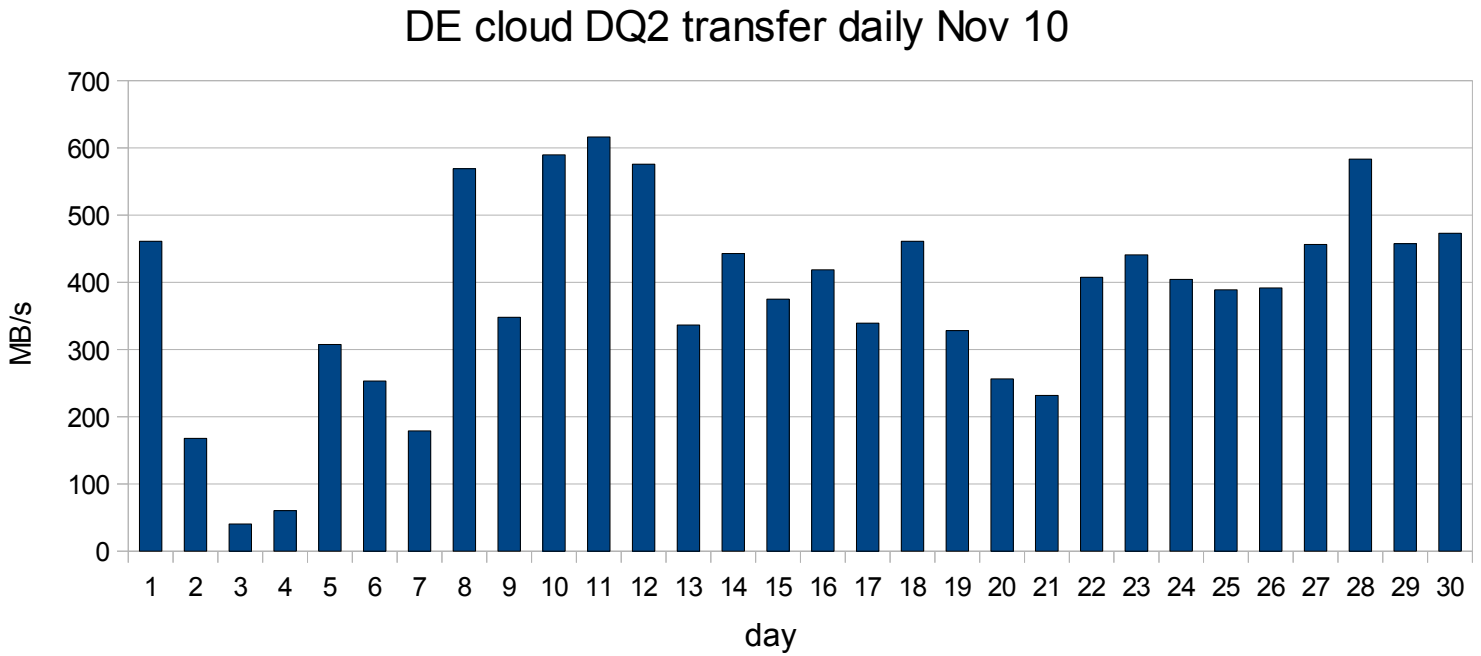
Exclude T0 export & T1 ↔ T1 traffic



- All DQ2 transfers to Tier2 sites in DE cloud
  - Average around 190 MB/s
    - DESY sites take larger data share (50% each vs 16% for most other T2s)
    - Peak (350 MB/s) in Nov caused by PD2P (dynamic data replication) + AOD reprocessing

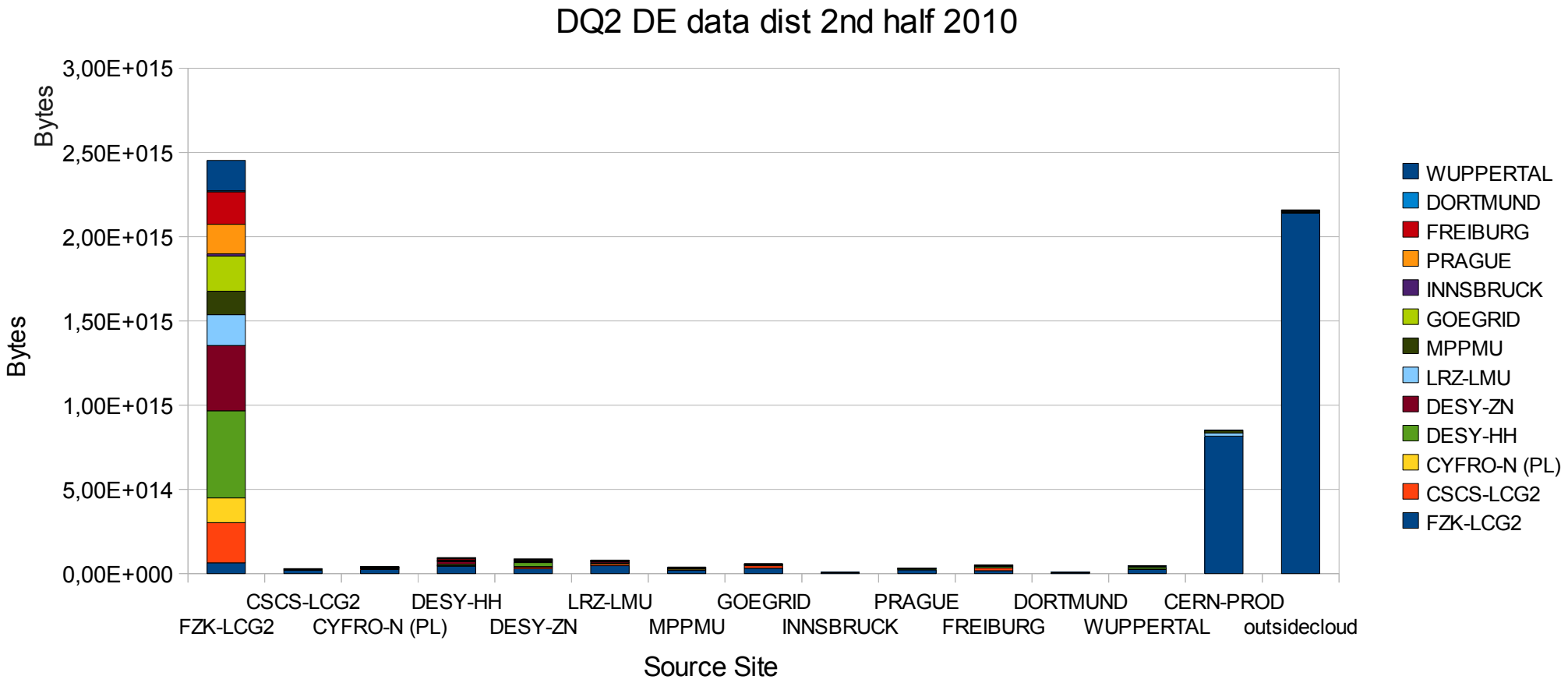
ATLAS DQ2 data distribution in DE cloud -daily Nov 10

Exclude T0 export & T1 ↔ T1 traffic



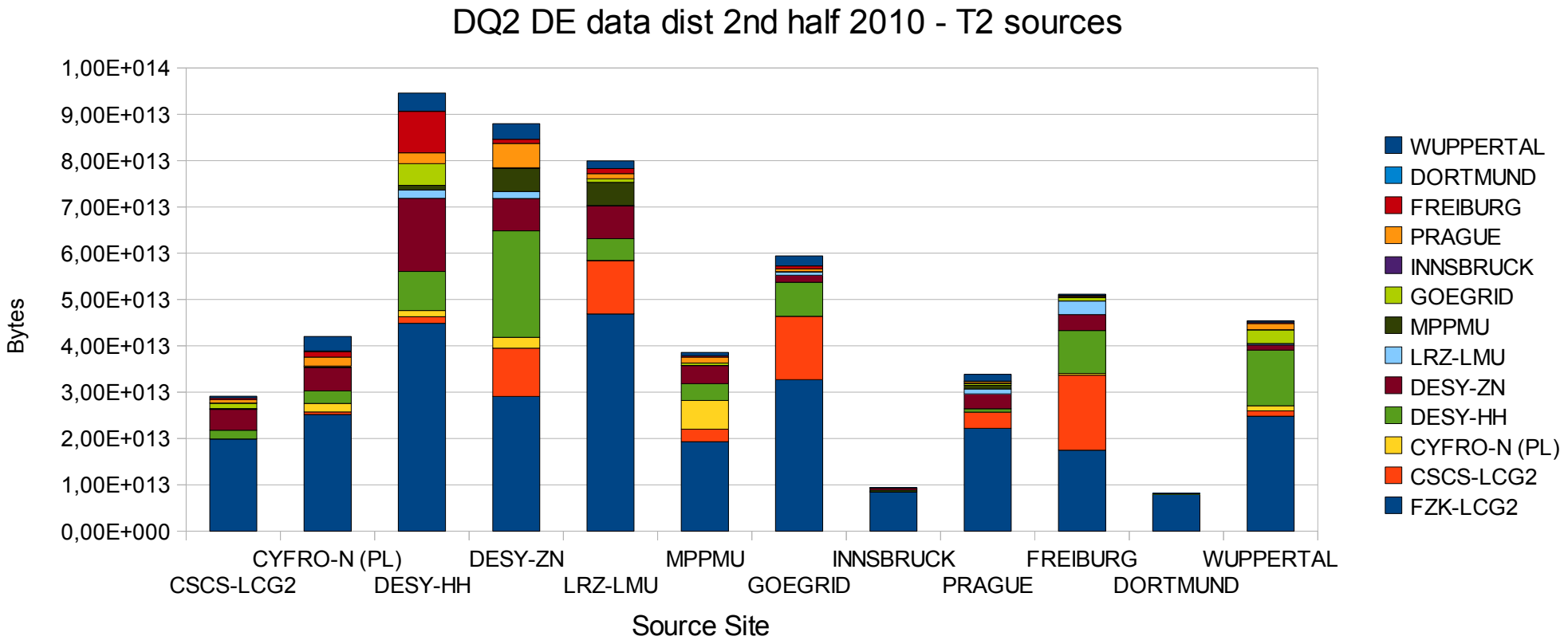
- All DQ2 transfers to Tier2 sites in DE cloud
  - Daily peaks up to 600 MB/s
    - Bandwidth saturation ??

ATLAS DQ2 data distribution in DE cloud - by source site -



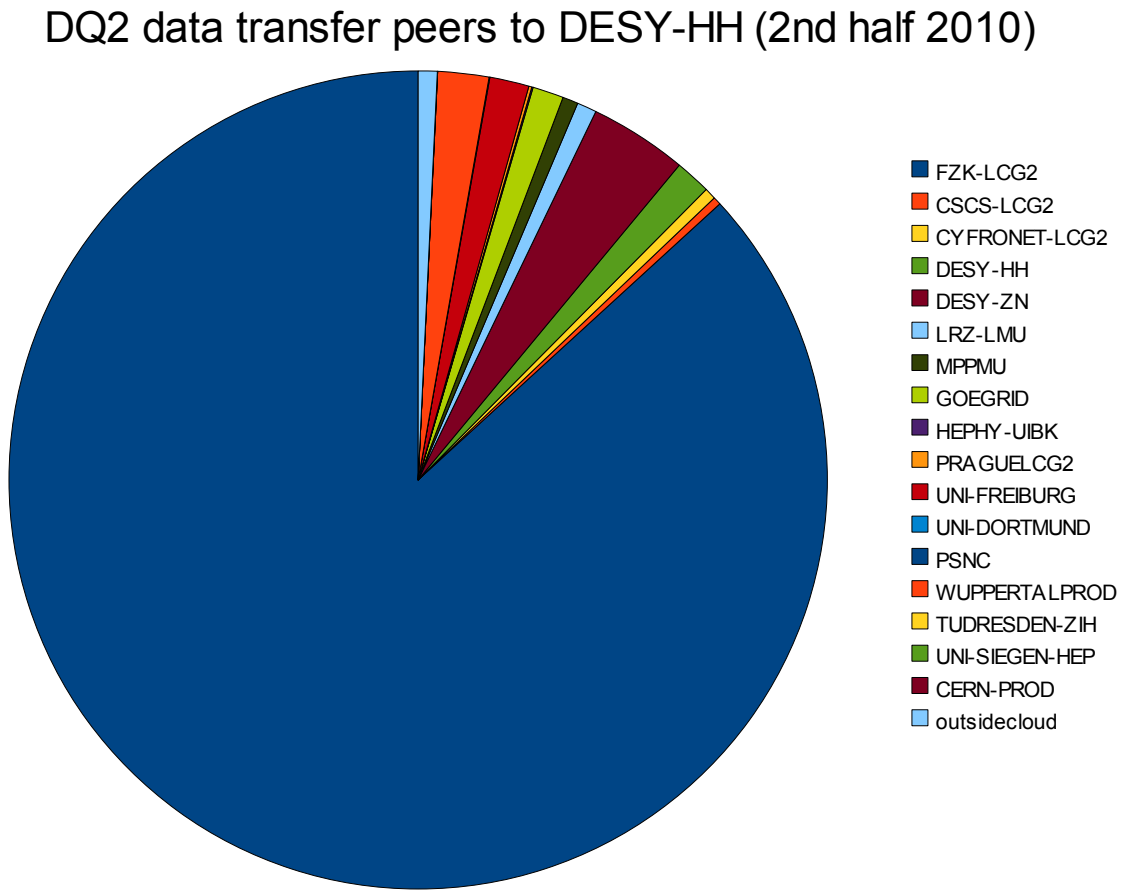
- DQ2 transfers: > 90% from GridKa to T2s (apart from T0/T1 to GridKa)

ATLAS DQ2 data distribution in DE cloud - by source Tier2 only -



- Tier-2 sources: most goes to GridKa (=output from MC production)

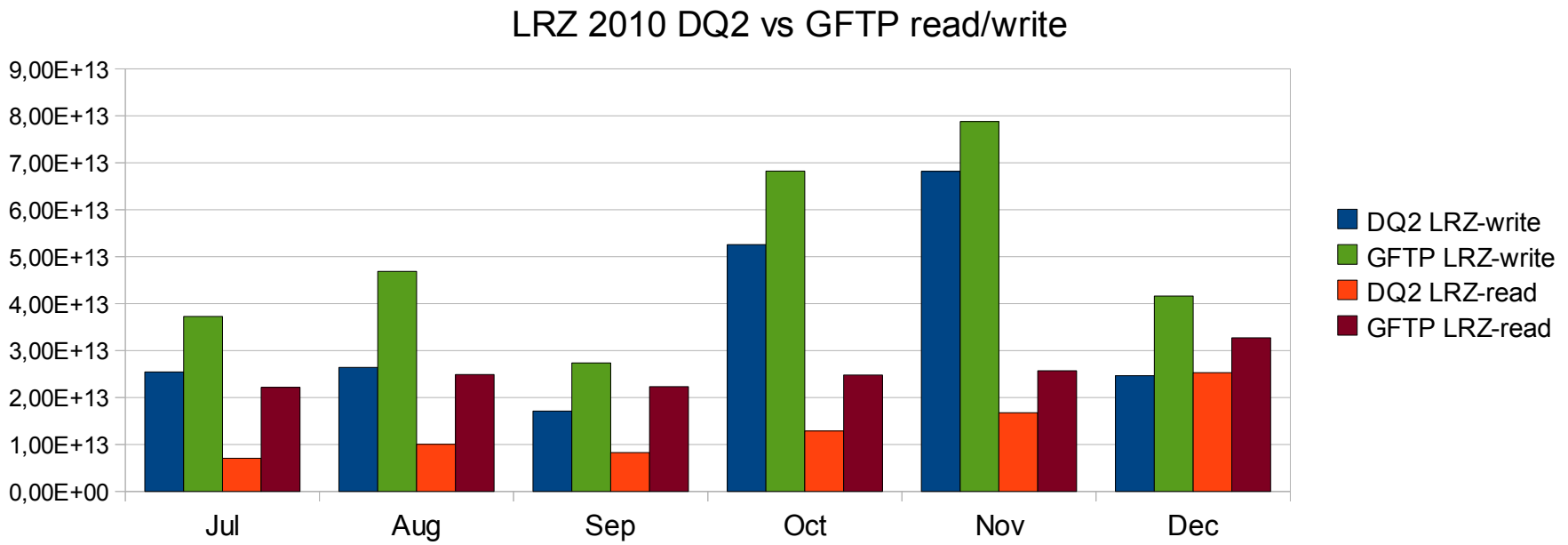
ATLAS DQ2 data distribution – Desy-HH example



600 TB total input

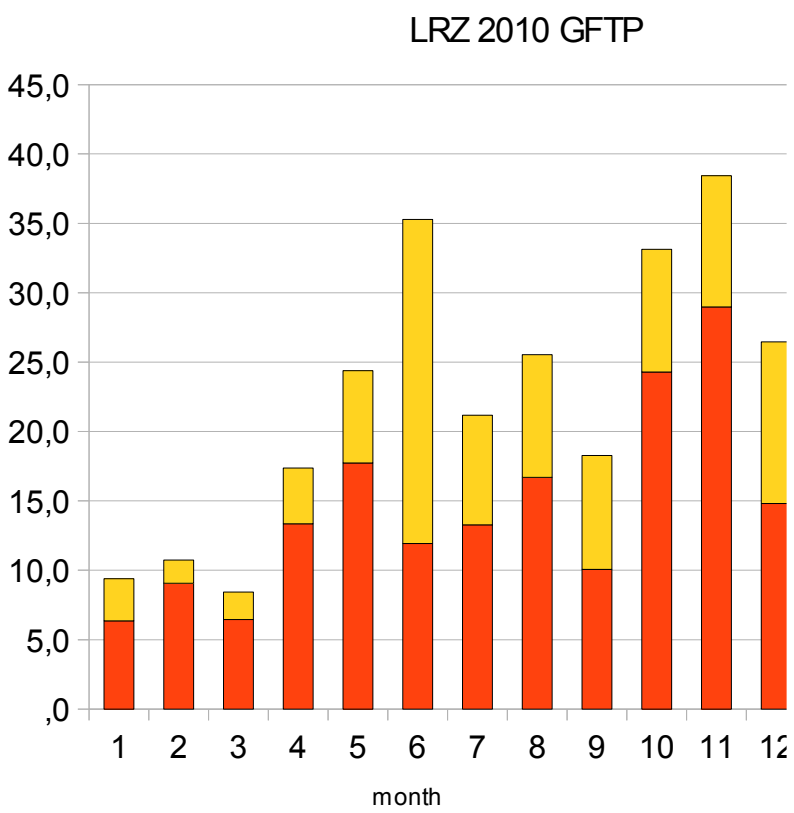
> 85% from GridKa

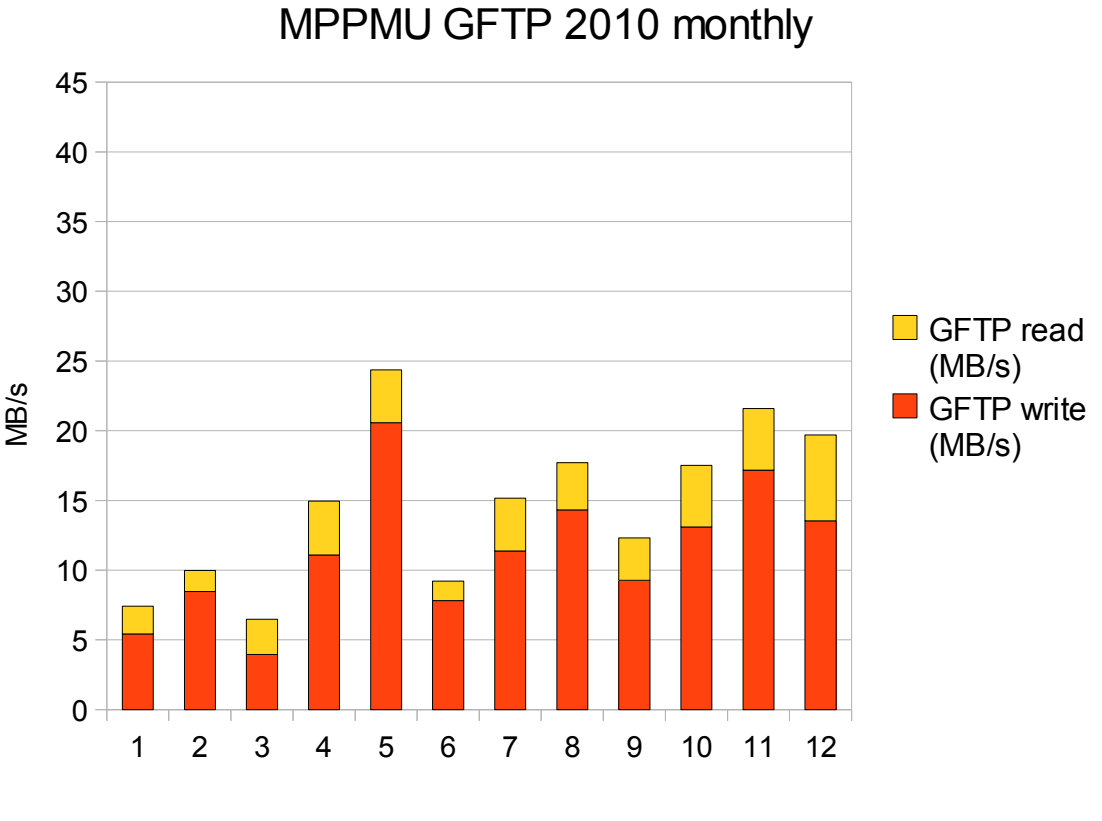
Compare ATLAS DQ2 vs Cache billing



- Discrepancy DQ2 read/write vs billing logs:
  - DQ2 misses transfers from production jobs (file upload to SE) and individual downloads/uploads (dq2-get/put, lcg-cp)
    - About factor **1.5** for writing (not really an issue, local GFTP upload from MC production)
    - About factor **2** for reading – **substantial** traffic from individual downloads
    - Checked in detail for LRZ Nov-2010 → consistent

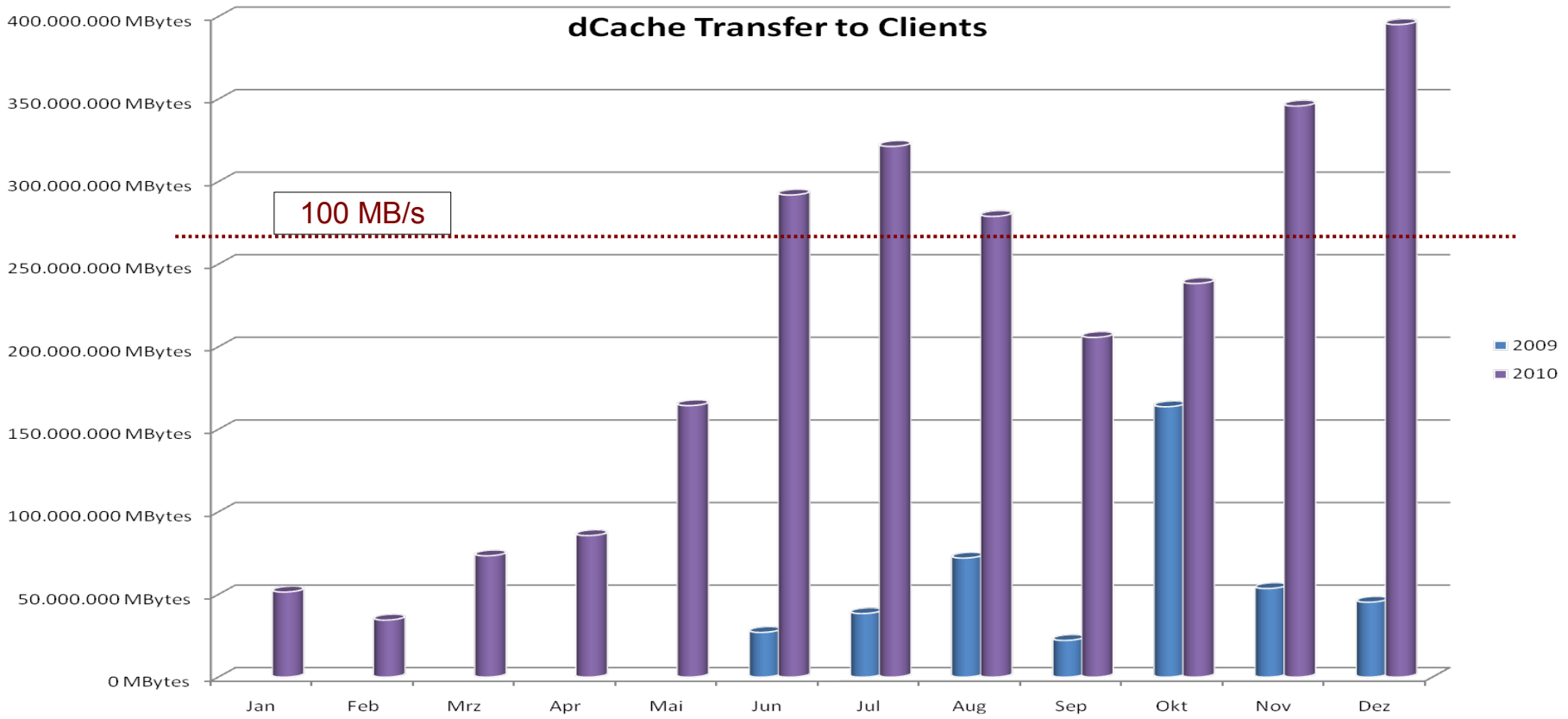
Transfer rates observed – Tier-2 (LRZ/MPPMU)





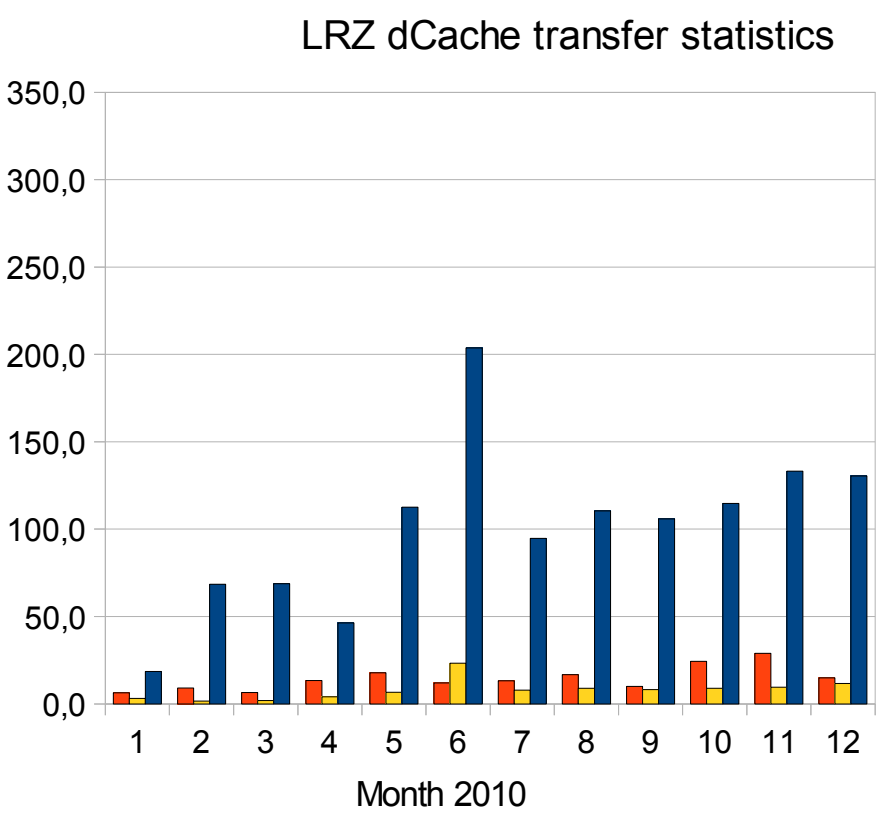
- LRZ&MPPMU GFTP monthly transfers 2010 (dCache billing)
- monthly average ~20-30 MB/s

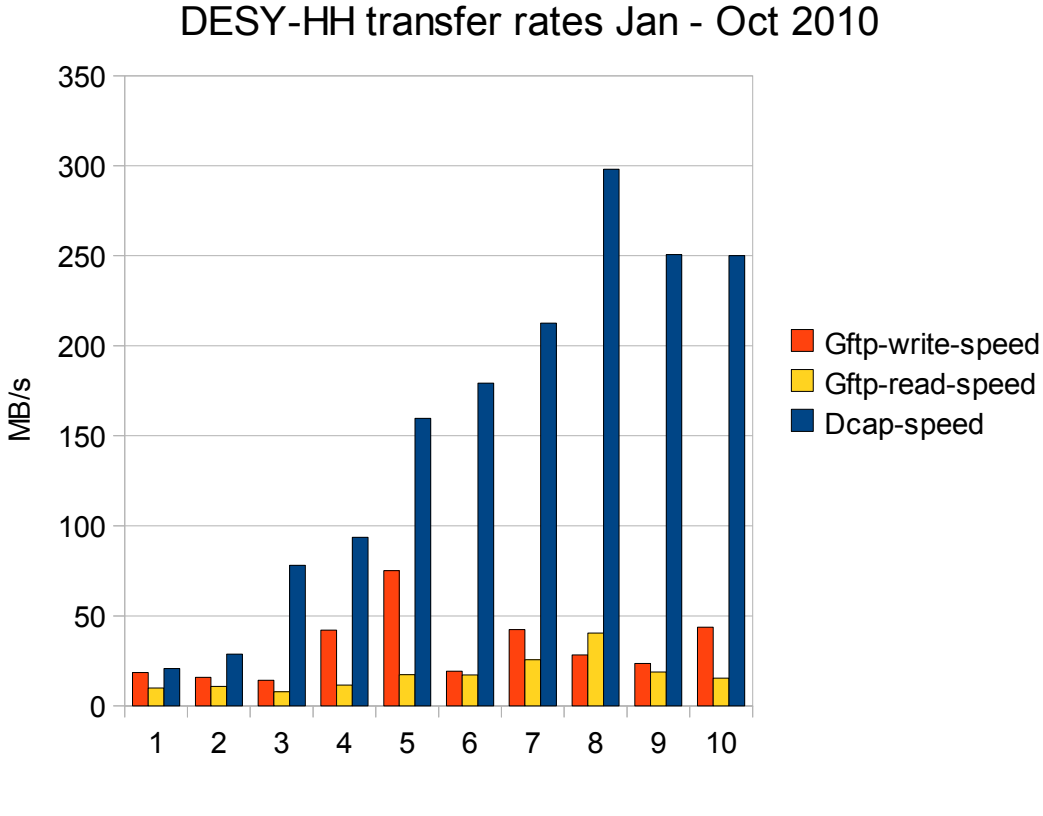
Transfer rates observed – Tier-2 (Wuppertal)



- Transfer volume per month (local DCAP + WAN GFTP)
  - Big increase over year

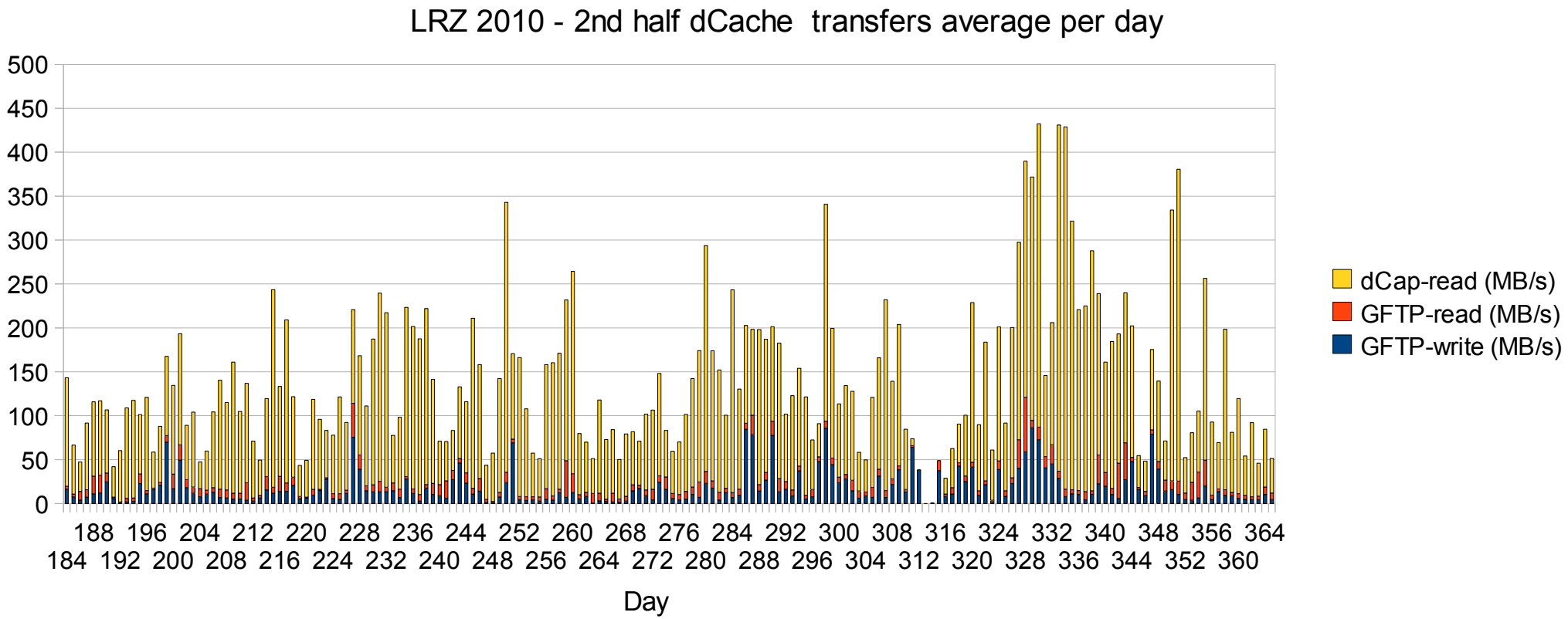
Transfer rates observed – Tier-2 (LRZ/Desy-HH)





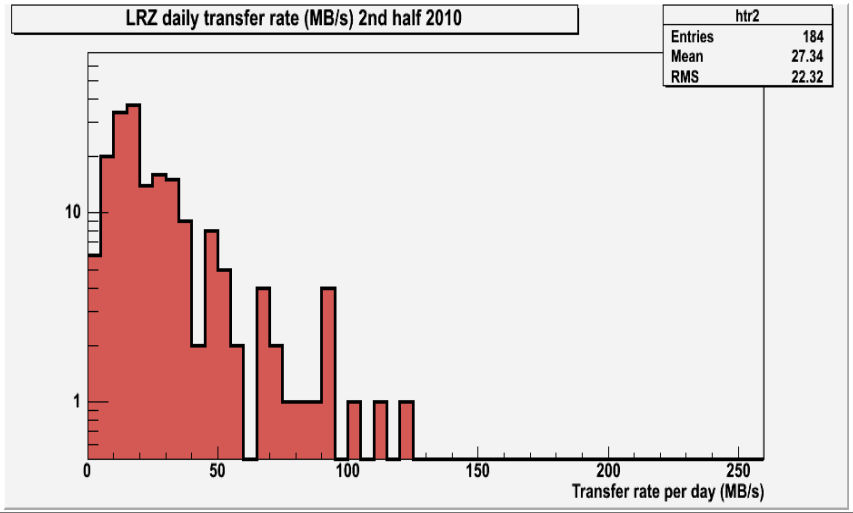
- Now DCAP transfers included = local reading from analysis jobs
  - Substantially higher ~120 MB/s vs 25 MB/s (GFTP) @ LRZ
- Desy-HH larger site/storage: DCAP~250 MB/s, GFTP~60 MB/s

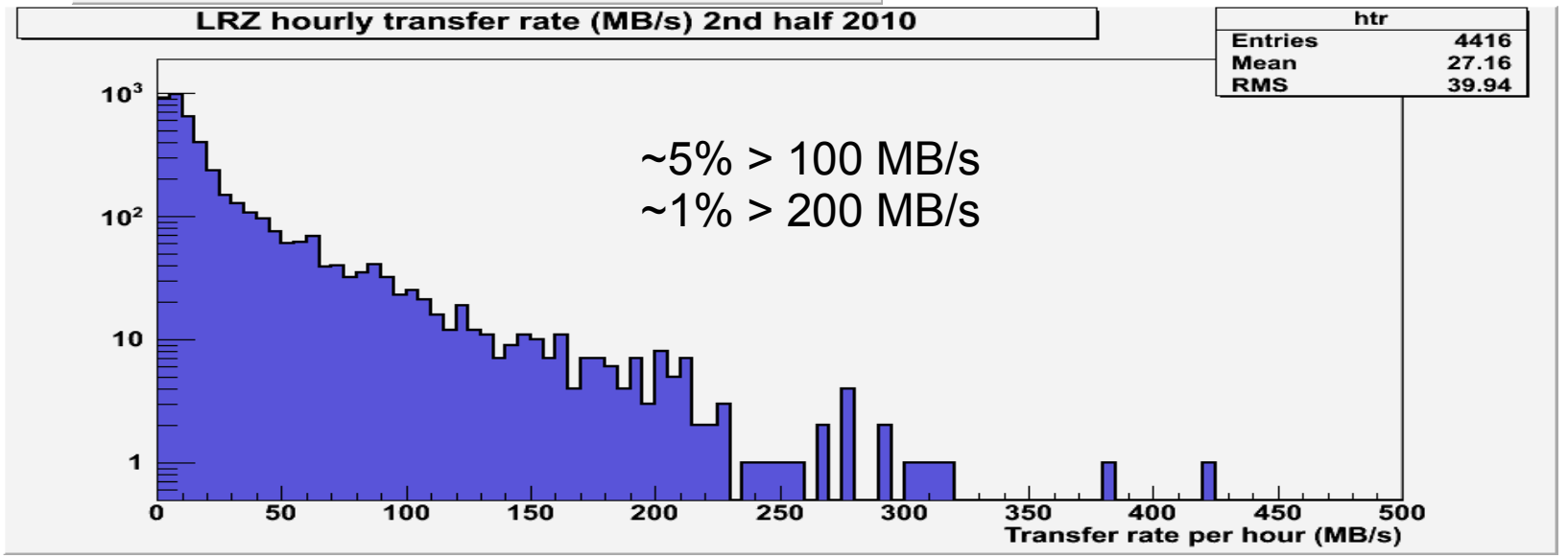
Transfer rates observed – Tier-2 (LRZ)



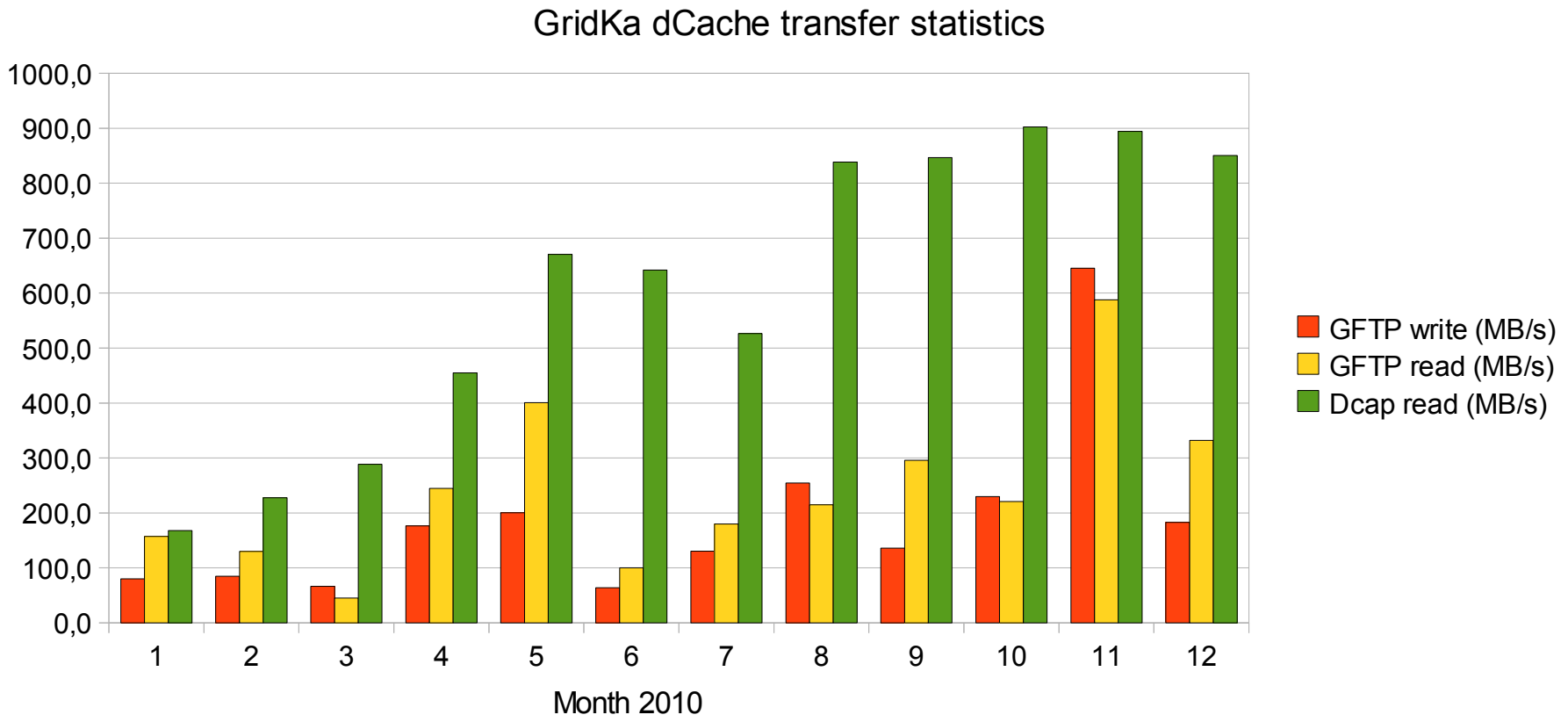
- Daily averages – large fluctuations:
  - GFTP up to 100 MB/s
  - DCAP (=local) up to 500 MB/s

Transfer rates observed – Tier-2 (LRZ) GFTP per day/hour



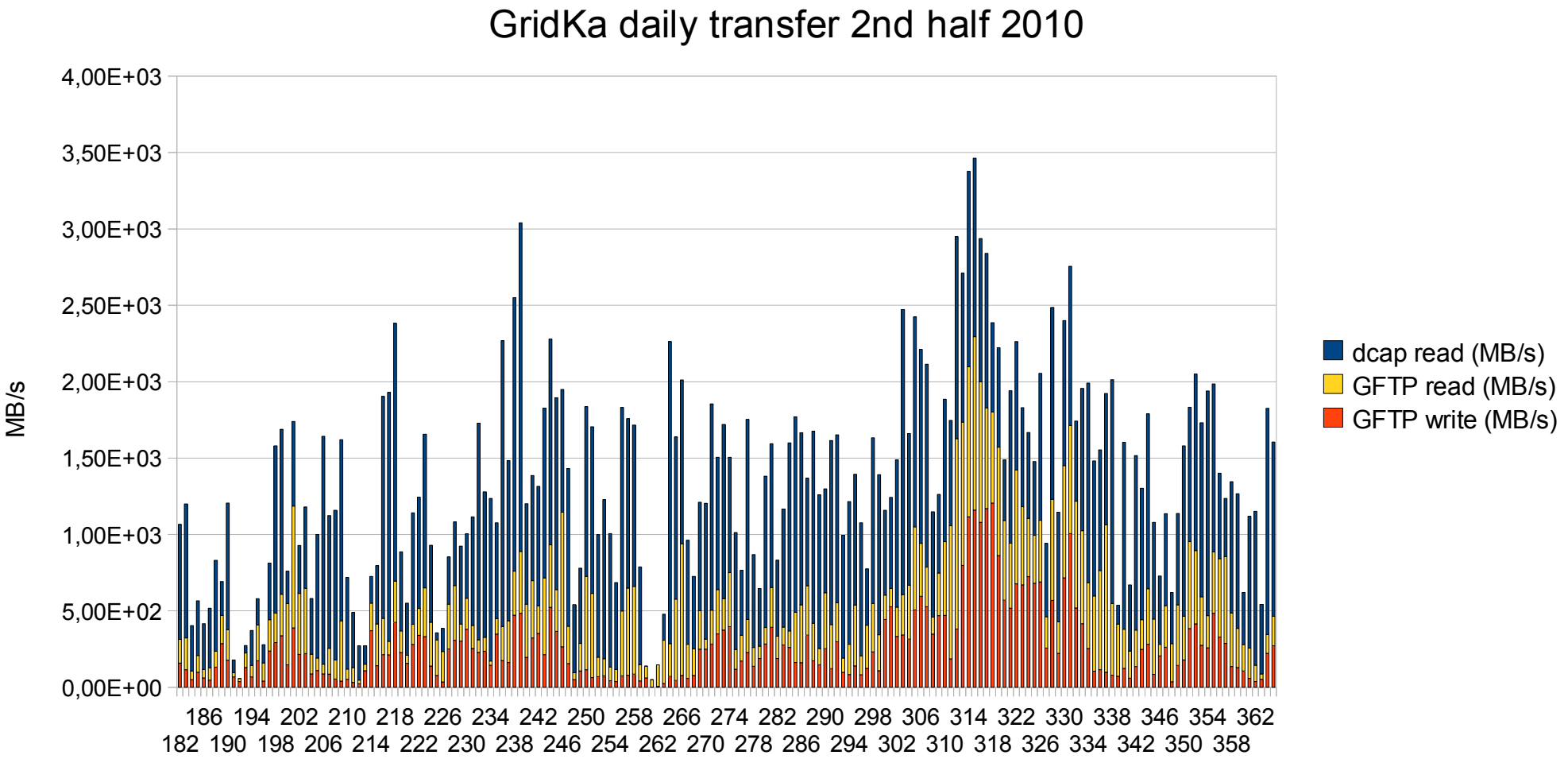


Transfer rates observed – Tier-1 (GridKa)



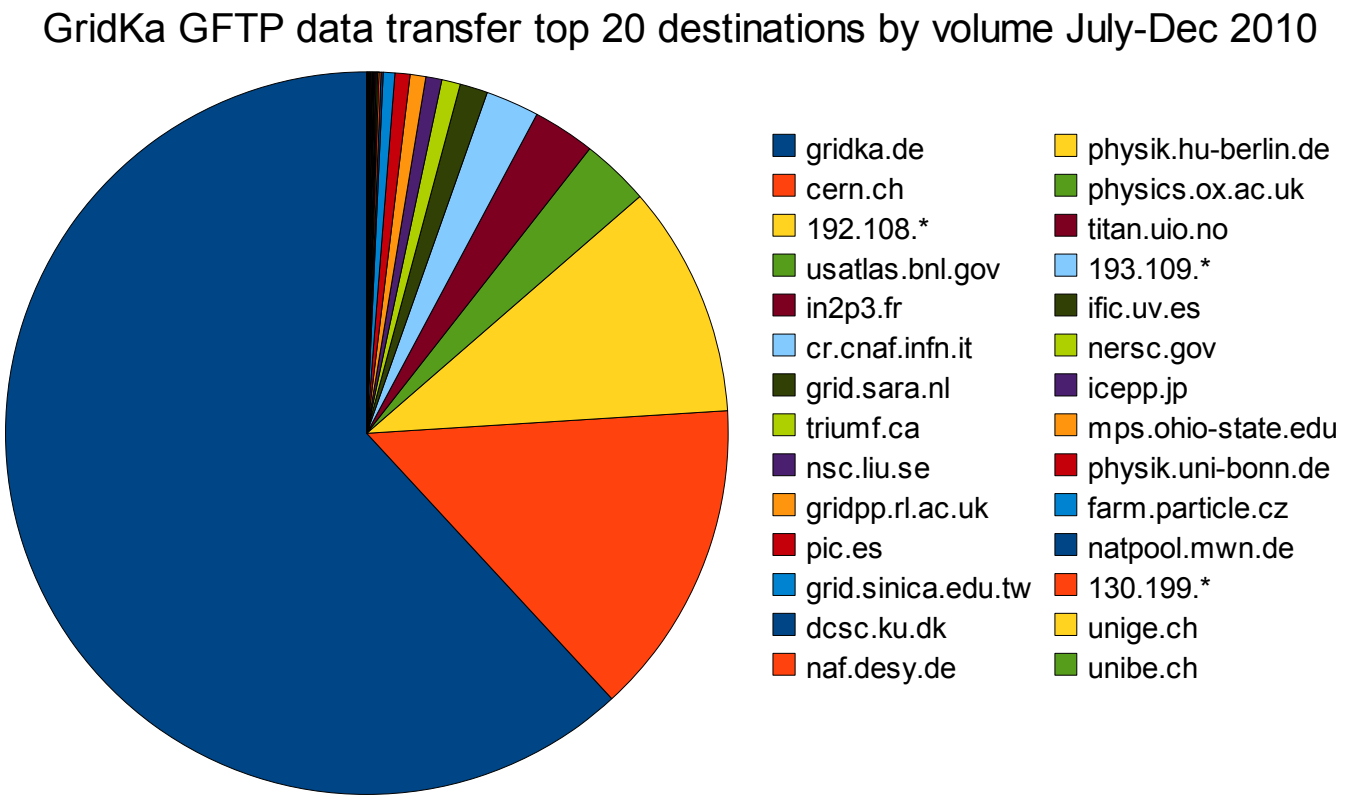
- Much higher GFTP transfer rate (420 MB/s vs 25@LRZ)
- Also DCAP (=local) transfer substantially more (600 MB/s vs 120@LRZ)

Transfer rates observed – Tier-1 (GridKa)



- Daily rates up to 2 GB/s for both GFTP and DCAP
  - Only smaller part (<1 GB/s) is WAN traffic in DE cloud

Tier-1 (GridKa) – data distribution from dCache billing



- Transfer connection to ~ 200 domains, 90% non-DQ2 connections
- 40 domains > 1 TB
- Volume wise DQ2 transfers dominate, only ~ 1% non-DQ2

Summary

- WAN transfer rates observed:
  - Tier-2 (=LRZ)
    - Other T2s similar
    - Except Desy-HH/ZN
      - factor 2-3 more
  - DE cloud
    - Excluding T0, T1 ↔ T1
  - GridKa T1:
    - Up to 2 GB/s daily
  - Data download outside ATLAS DDM add ~10%
- LAN transfers much higher:
  - several GB/s peak rates

LRZ average Jul-Dec 10:	27 MB/s
Max month (Nov)	40 MB/s
Max daily	120 MB/s
Max hourly	420 MB/s
5% (1%)	> 100 (200) MB/s

Cloud average Jul-Dec 10:	188 MB/s
Max month (Nov)	350 MB/s
Max daily	600 MB/s

Tier-2 feedback

- LRZ/Muenchen:
  - 5.5 Gb currently → plan to double (trunking) in ~weeks
    - problems if LHC fills bandwidth (Web, Video suffer)
      - could use virtual link to limit b/w
      - but first we should try to limit w/ FTS settings
  - Dedicated link possible (done for DEISA), but cost-effective?
- RZG/Muenchen:
  - 3 Gb currently (10 Gb wire but contract limited)
  - ATLAS peak transfer periods fill capacity but no complaints so far

Tier-2 feedback

- Bergische Universität Wuppertal:
  - technische Daten:
    - X Win Kernknoten in Wuppertal (~5m vom Uni Zentralrouter):
    - ca 400 GBit/s (?)
    - Uni-Anschluss 1Gbit/s FD - Upgrade von 450 (2009-2011) bzw. 600 (ab 2011) Mbit/s laufend bezahlt aus Helmholtz Förderung
    - physikalische Anbindung Uni Zentral Router ./. Tier-2:
    - 10 Gbit/s FD LR Fibre,
    - 1 Gbit/s FD LR Fibre Backup,
    - weitere 3 „dark Fibre“ (10 GBit/s LR fähig)
  - Kosten:
    - DFN: I7 (433/600 MBit/s) 103.800€
    - Sondervereinbarung:
      - Upgrade auf 1GBit/s („I8“) 30.000€ (statt 51.900€)

Tier-2 feedback

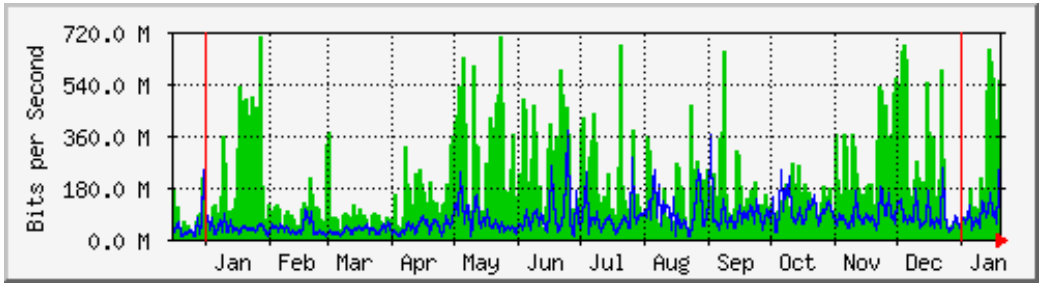
- Desy-HH/ZN:
  - dedicated P2P link (10 Gb/s) Desy-HH ↔ GridKa
    - also for Desy-ZN traffic
  - but only used for direct GridKa – Desy transfers
    - traffic from other T1/T2 sites go via shared X-Win

What to conclude for future networking requirements … for discussion

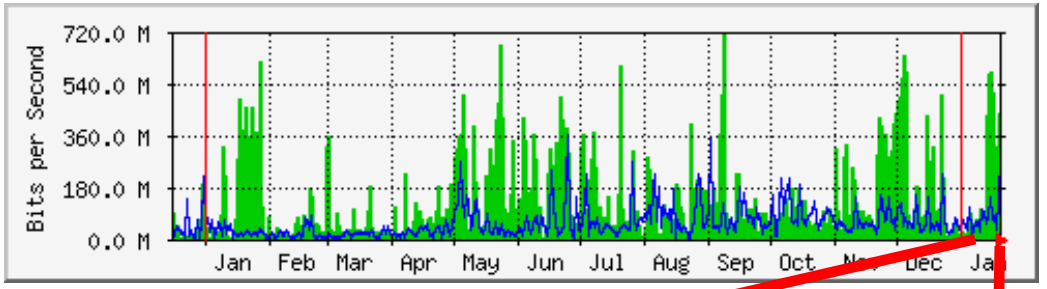
- Current connectivity was ok for 2010 operation
  - No real bottlenecks/delays, some headroom left
- Increase of used bandwidth over last months
  - Real trend? Or threshold effects by new/changed activities ??
- Changes in ATLAS data distribution model under way
  - More dynamic data placement, direct transfer between T1/T2 sites, etc
  - Presumably increases network use but hard to quantify, many unknowns
    - Connectivity T2 ↔ all T1 will become important
- DE cloud T2 connectivity:
  - Several sites with > 2 Gb/s – should be ok
  - Sites w/ 1 Gb/s might reach saturation occasionally
  - What about integral traffic, can we get >10 Gb/s total in ATLAS-DE cloud?

Backup slides

Netzauslastung BU Wuppertal

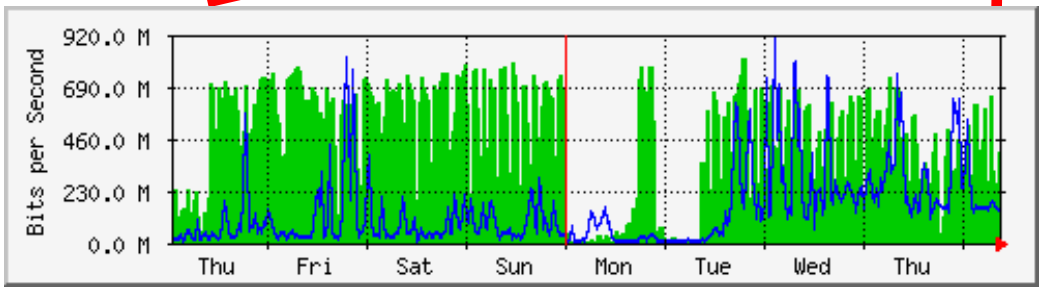


2010 WAN



2010 Tier-2

Beispiel High-Traffic Phase 2 KW 2010



Nahezu gesamter Netzverkehr der Hochschule bedingt durch Tier-2 Betrieb

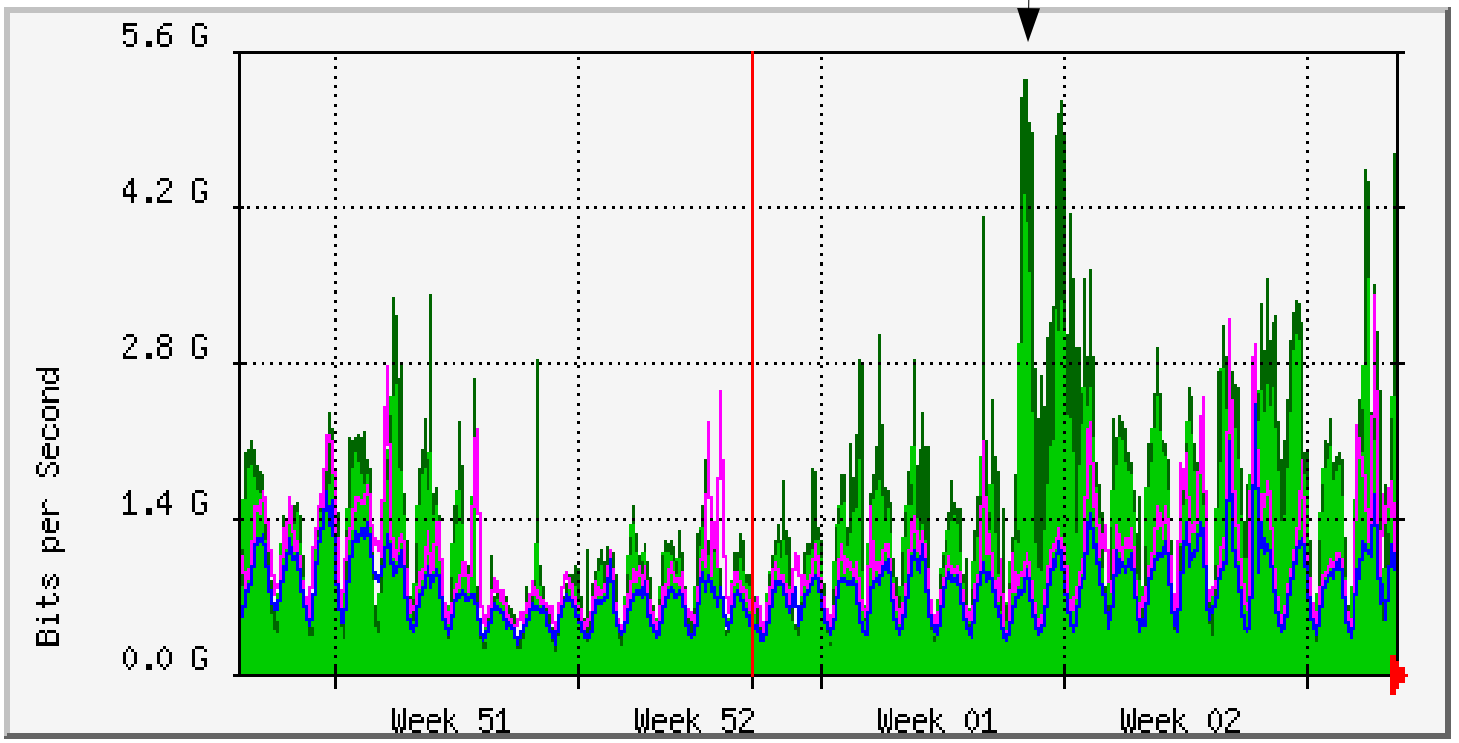
LRZ b/w limit reached early Jan 2011

LRZ operates “Muenchner Wissen-schafts Netz”:

> ~150 k students & scientists

Large baseline traffic: 1-2 Gb/s

MCDISK → DATADISK move



MPPMU/RZG – site vs T2 traffic