

ATLAS Data Management in the GridKa Cloud

John Kennedy
LMU München
DESY-HH seminar



18/06/2007

Overview

- Who am I
- Cloud Overview
- DDM - Design
- DDM – OPS in the DE Cloud
- Other issues
- Conclusion

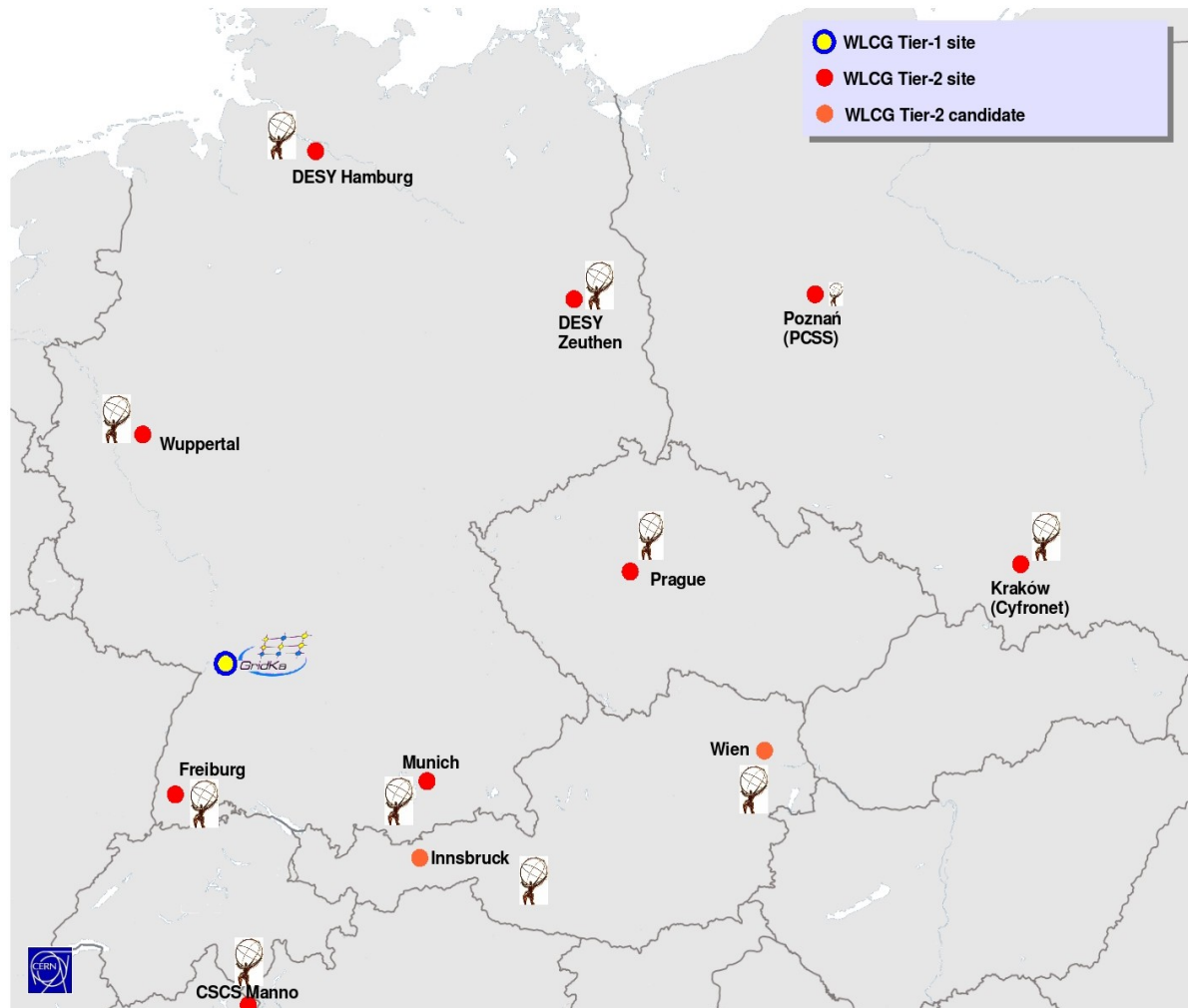
Who Am I – and what do I do

- GridKa Cloud Coordinator – Beta Version
 - Responsible for technical coordination (T1/T2's)
 - Organising functional tests
 - Atlas contact – attend many Atlas meetings
 - Organising meetings and ensuring communication
- Primary author of prod-sys monitoring
- Contribute to EGEE production shifts
- DDM-Operations T1 contact for GridKa
- Get involved in all operations areas related to our Cloud

GridKa Cloud

Cloud Overview

ATLAS GridKa Cloud - Sites



- Many sites¹³ (excl cern)
- Four^(soon five?)
- Countries
- Two ROC's
- Feed with Data
- Feed with Jobs
- Organisation is a key point to our success!

ATLAS GridKa Cloud - Services

- A Cloud is more than just a collection of sites
- Provide a service to the community
 - Data storage and distribution
 - MC production (and aggregation)
 - Distributed Analysis
 - Re-processing of RAW Data
- Central services provided by T1 are essential to the Clouds health
- T2's and their associated manpower are providing much of the driving force

DDM

Distributed Data Management

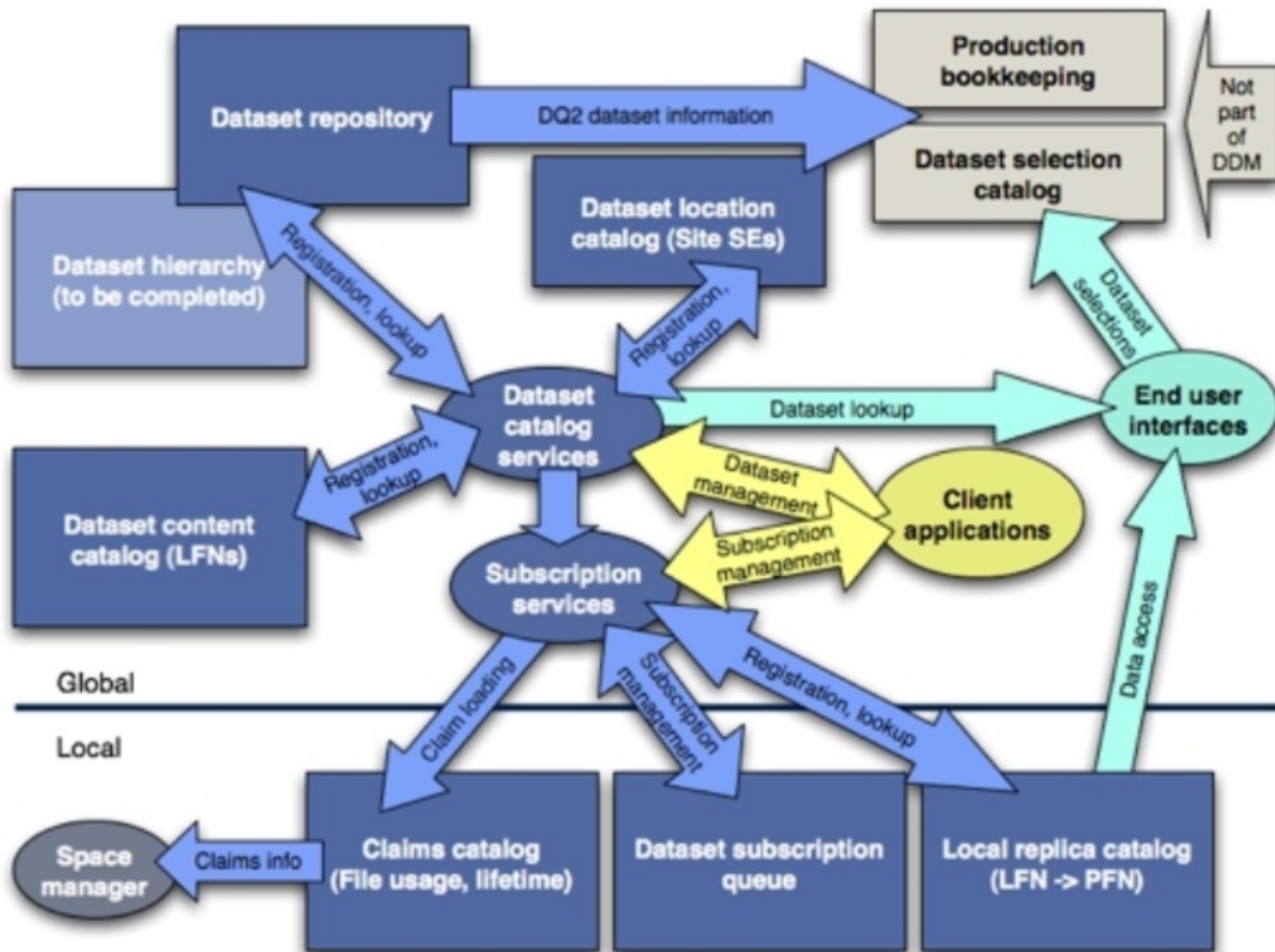
DDM Intro

- Data Management on a File level cannot scale
- Data Management on a **Dataset** level rather than a **File** Level - Sounds Promising
- **Central Dataset Catalogs**
- **Distributed File Catalogs**
- By using DDM we expect to make things easier by a few orders of magnitude

DDM Architecture

- Two Main Components
- Dataset Catalogs
 - Contain dataset definitions, locations
 - Single set of central catalogs hosted at cern (currently)
- Site Services
 - Cataloging of data at each site
 - Movement of Data to sites
 - Scheduled transfers (FTS)
 - Removal of old data at sites

DDM Architecture



Datasets

- Datasets – what are they:
 - **Contain files**
 - Logical File Name, GUID, metadata
 - **Have Versions**
 - Files can be added/removed between versions
 - **Has Unique Identifiers**
 - Per dataset, per Version
 - **Hierarchies**
 - Datasets which contain other Datasets

Datasets Catalogs

- Dataset Repository Catalog
 - What Data exists in the system
- Dataset Content Catalog
 - What files are in a given dataset (version)
- Dataset Location Catalog
 - Where is the Data located
- Dataset Subscription Catalog
 - Keeps track of all requests for data (sites)
- Dataset Selection Catalog Not Implemented/Not Part of DQ2
 - What Datasets match a certain query
- Dataset Hierarchy Catalog Not Implemented
 - Records hierarchical organisation between datasets

Dataset states/versions

- **States**
 - Open
 - Files can be added
 - Closed
 - Files cannot be added (new version can be made)
 - Frozen
 - No more files/Versions (REAL DATA)
- **Versions**
 - Can track changes in data
 - Files can be added removed between versions
 - Subscriptions will get latest version unless otherwise instructed.

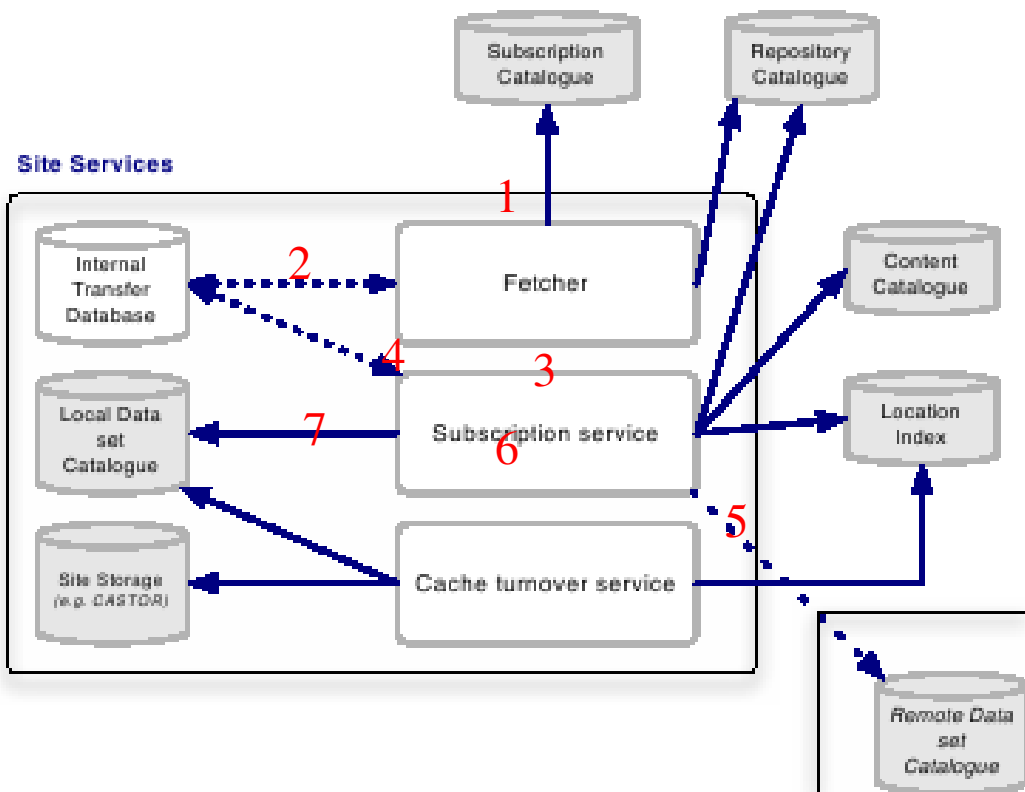
Site Services

- **Subscription handling**
 - Ensure that contents of a subscribed dataset are moved to the site
 - Automatic handling of dataset updates
- **Cache turnover**
 - Allow for deletion of unwanted data
- Cron Jobs run at T1's providing services for associated centers, you need to be supported to use DDM

A subscription Explained

- Subscribe a dataset to a site
- Services at T1 will pickup subscription and submit transfer request (FTS)
- T1 services monitor transfer status doing retries if needed
- If dataset is closed once completely at site marked as COMPLETE, otherwise remains INCOMPLETE (even though all files there)

A subscription Explained



- 1) Fetcher- picks up new subscriptions from subscription catalog
- 2) Queues requests to local database of ongoing transfers
- 3) Subscription service finds missing files (repo,content cat,local cat)
- 4) Adds missing files to local database
- 5) Finds replicas of missing files
- 6) Interacts with data transfer layer (FTS), moves files in blocks
- 7) Bookkeeping of transferred files, registering in local dataset catalog

Client and End User are Tools

- Client tool exists for interface to central catalogs
 - Registering subscriptions, select source/destination...
 - Getting info about datasets
 - Listing subscriptions at sites
- A Rich Sample of end user tools.
 - dq2_ls
 - dq2_get
 - dq2_register
 - dq2_put
 - dq2_cleanup
 - dq2_sample
- End user tools may be extended

DQ2 0.3

- Central catalogs moved to Oracle, supported by several servers!
- Site services smarter
 - Lower loads
 - Better error treatment
 - rpms for deployment
 - Voboxes may be based at CERN
- Monitoring much improved
- End user tools being re-written
- 0.2->0.3 Migration set for 19-21 June

DDM-OPS

DDM Operations in our Cloud

DDM Ops Overview

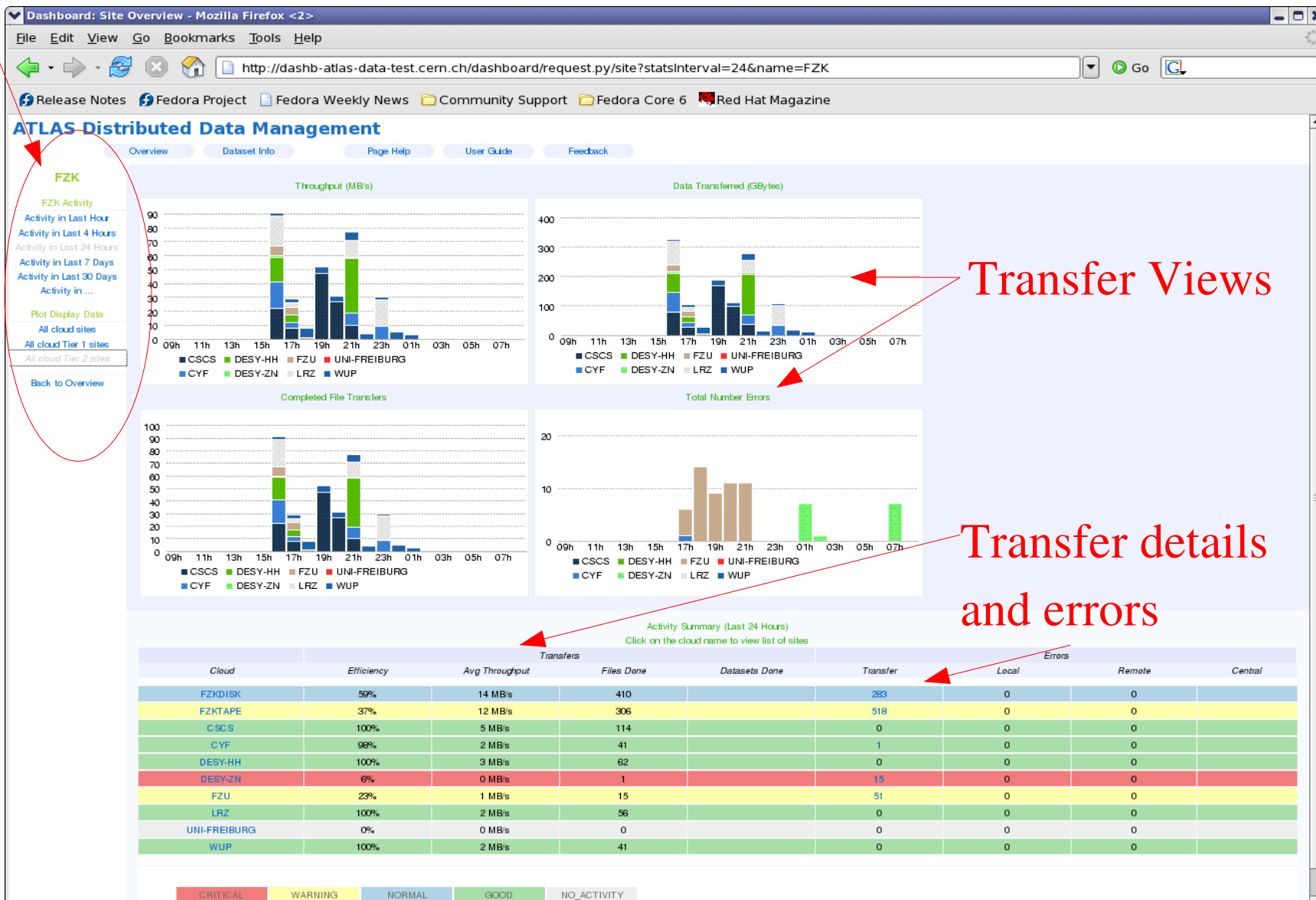
- Monitoring
- Data in our Cloud
- T0 Tests (CERN->GridKa->T2's)
- AOD Replication [Data for Physics Analysis](#)
- Production Data Aggregation
- Tasks of ddm-ops group

DDM Monitoring

- ARDA dashboard is ever improving
 - Error reporting
 - Transfer rates
 - Dataset/file status (full history)
 - Views per Cloud
- With dq2 0.3 new pages available
- Developers are very open to suggestions
- Makes many things easy
- Also several ddm-ops monitoring pages

DDM Monitoring

Cloud View



Transfer Views

Transfer details and errors

DE Data summary

Summary of disk occupancy and number of files for the DE cloud.
Only files belonging to the central MC production are taken into account.

WARNING: Only datasets whose project name (character chain before the first point) includes one of the following list are treated in this page :
- csc11 mc11 mc12 calib0 calib1 testIdeal testMisal mcMisal stream valid1

This report was generated: 11 June 2007 - 09:32

Size

Site	AOD	ESD	TAG	SAN	HPTV	CBNT	EVNT	RDO	HITS	NTUP	HIST	log	Total
FZKDISK	16011.3	10601.7	0.5	84.5	55.3	8.7	3541.6	20383.5	2398.1	4848.4	0.0	67.5	58001.1
FZKTAPE	2787.0	1559.1	0.0	0.0	0.0	0.0	207.7	9466.3	5170.9	4.7	0.0	1.1	19196.8
FZU	2177.3	30.6	0.0	4.1	2.8	0.1	34.3	38.8	12.2	177.4	0.0	0.4	2478.0
CSCS	34.5	862.5	0.0	5.2	4.1	0.0	2.4	73.5	38.0	12.8	0.0	1.0	1034.0
CYF	5190.2	447.9	0.1	7.8	5.5	0.9	51.5	1251.9	184.3	1639.3	0.0	1.8	8781.2
DESY-HH	10420.6	20.1	0.0	0.0	0.0	0.0	0.2	4.8	1.8	3457.9	0.0	0.1	13905.5
DESY-ZN	1382.0	68.1	0.1	12.4	8.6	0.2	197.8	316.2	154.1	438.7	0.0	3.4	2581.6
UNI-FREIBURG	71.1	267.8	0.0	5.6	3.7	0.0	3.7	229.7	20.3	18.4	0.0	0.7	621.0
WUP	3145.8	94.1	0.0	6.4	5.0	0.1	84.4	152.4	94.1	966.4	0.0	2.1	4550.8
LRZ	1572.1	98.8	0.1	8.4	5.8	0.0	0.8	86.1	46.1	105.7	0.0	1.2	1925.1

Num Files

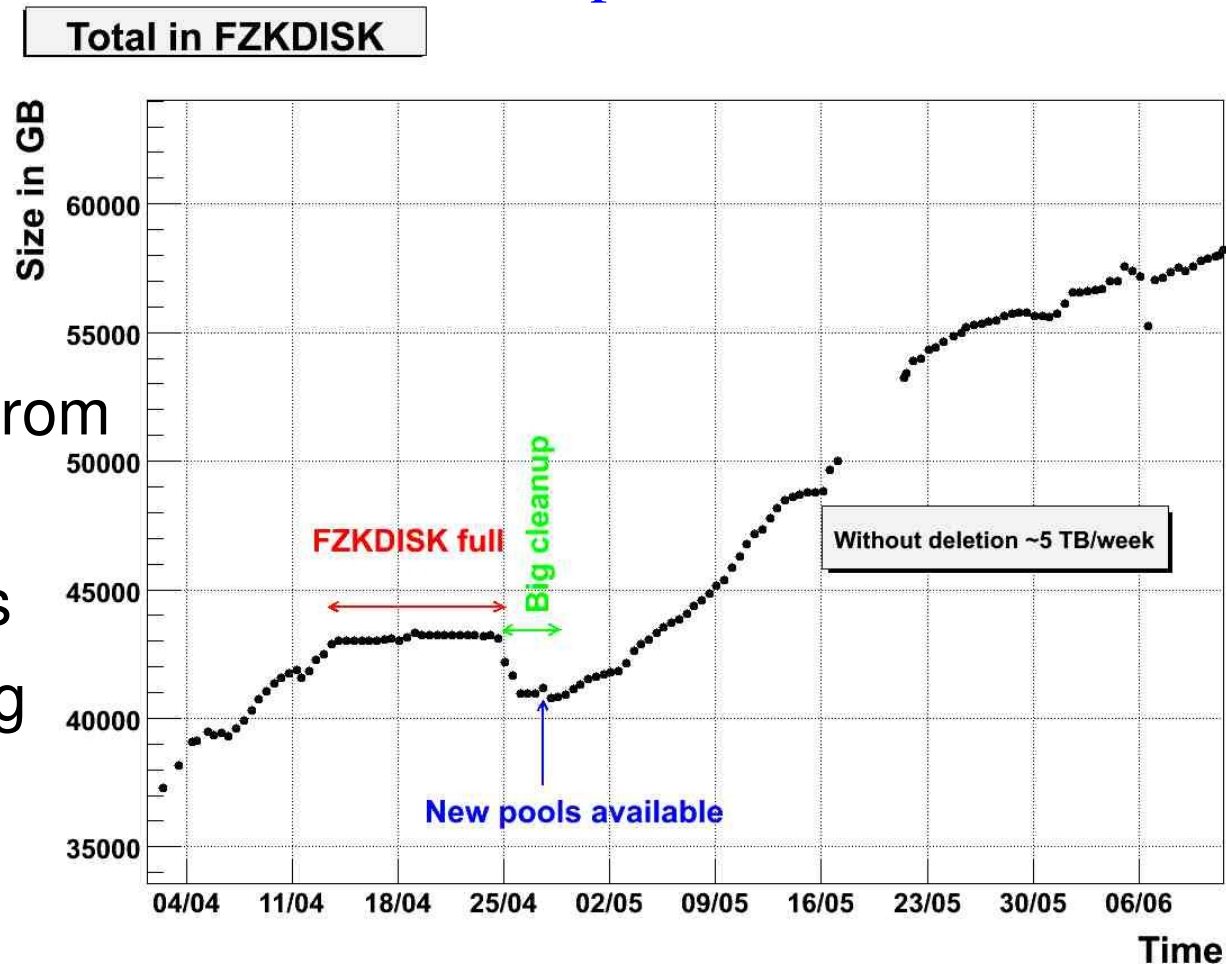
Site	AOD	ESD	TAG	SAN	HPTV	CBNT	EVNT	RDO	HITS	NTUP	HIST	log	Total
FZKDISK	281771	22615	598	607	604	364	32069	239005	135886	259926	125	215007	1188577
FZKTAPE	28837	2261	0	0	0	1	932	92683	134305	206	6	15505	274736
FZU	25459	61	45	40	37	3	607	347	329	8614	6	3475	39023
CSCS	450	2656	42	42	44	0	45	712	664	549	5	3623	8832
CYF	81265	599	71	62	63	48	501	11763	2872	73563	12	11340	182159
DESY-HH	196626	35	0	0	0	0	7	47	47	188529	0	677	385968
DESY-ZN	22372	146	110	106	108	7	1559	3042	3041	21062	33	27943	79529
UNI-FREIBURG	1006	461	48	41	45	0	39	1732	504	789	4	5886	10555
WUP	65141	122	67	54	53	2	889	1619	2008	60274	18	16680	146927
LRZ	17552	166	76	69	66	0	18	865	1083	3004	7	4729	27635

Comments and questions : cedric.serfon at physik.uni-muenchen.de . Script original idea : Stephane Jezequel (jezequel at lapp.in2p3.fr)

GridKa Disk

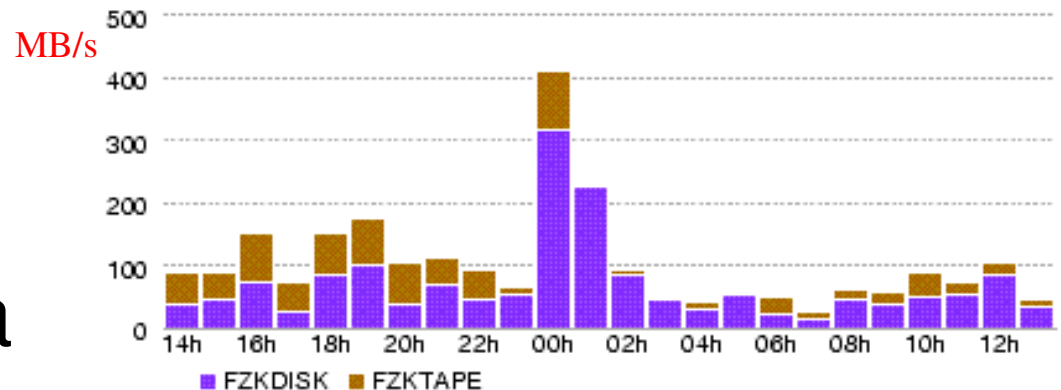
- Disks filled in April
- Approx 2 weeks stopped
 - Production
 - AOD replication
- Cleanup of data
- New Disks
 - Fast as possible from Gridka
- Commissioning disks incrementally causing probs

Both production and data movement are dependent on GridKa



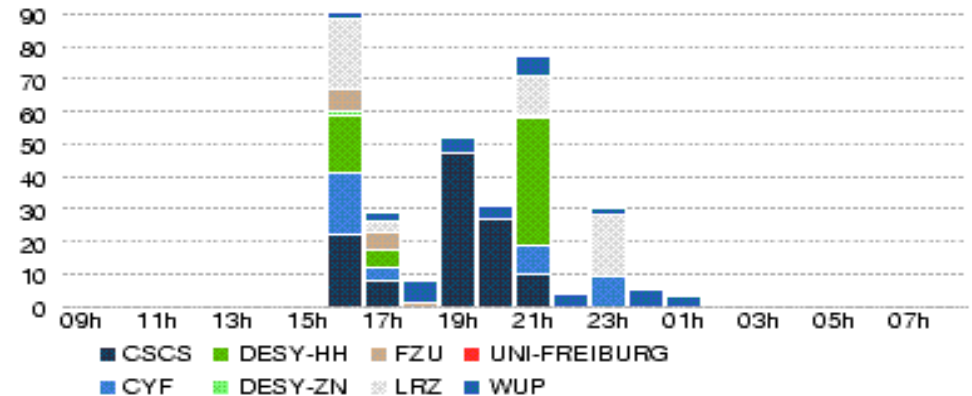
T0 -> GridKa

- T0 = Test DDM for data replication (mimic real data taking)
- T0->T1->T2's
- ~90MB/s to GridKa
- Data split realistically between DISK and TAPE
- Goal is sustained transfer with full rate and 90%+ efficiency – we are close
- Major errors been debugged



GridKa -> T2's

- Functional test^{12/06/07}
- Using dq2 0.3
- Very pleased so far
- Need to do sustained transfers!



Activity Summary (Last 24 Hours)
Click on the cloud name to view list of sites

Cloud	Transfers				Errors			
	Efficiency	Avg Throughput	Files Done	Datasets Done	Transfer	Local	Remote	Central
FZKDISK	59%	14 MB/s	410		283	0	0	
FZKTAPE	37%	12 MB/s	306		518	0	0	
CSCS	100%	5 MB/s	114		0	0	0	
CYF	98%	2 MB/s	41		1	0	0	
DESY-HH	100%	3 MB/s	62		0	0	0	
DESY-ZN	6%	0 MB/s	1		15	0	0	
FZU	23%	1 MB/s	15		51	0	0	
LRZ	100%	2 MB/s	56		0	0	0	
UNI-FREIBURG	0%	0 MB/s	0		0	0	0	
WUP	100%	2 MB/s	41		0	0	0	

CRITICAL
WARNING
NORMAL
GOOD
NO_ACTIVITY

Nice to see a few errors!
Now debugged!

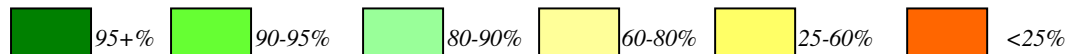
AOD Replication

- Request from physics groups to have copy of v12 AOD distributed throughout the cloud.
- **Distributed Analysis** tools (**Ganga**) can use this data in an efficient manner (**Jobs-->Data**)
- Currently we are distributing randomly according to a sites size(requested fraction)
- Test page in place to allow sites to select datasets for their sites using a reg-exp fashion.

AOD Replication

- GridKa is best in terms of completed datasets and second in terms of num of files

to \ from	ASGC	BNL	CERN	CNAF	FZK	LYON	NG	PIC	RAL	SARA	TRIUM F	%
ASGC												76
BNL												84
CERN												43
CNAF												19
FZK												78
LYON												76
NG												69
PIC												85
RAL												32
NIKHE												32
TRIUM F												30



AOD in DESY-HH

Transferred data for site DESY-HH and dataset type AOD - Mozilla Firefox <2>

File Edit View Go Bookmarks Tools Help

http://www.etp.physik.uni-muenchen.de/ddm/DE/DESY-HH/AOD/list_CC.html

Release Notes Fedora Project Fedora Weekly News Community Support Fedora Core 6 Red Hat Magazine

Atlas AODs available on site DESY-HH and registered in DDM

This report was generated: 16 June 2007 - 02:40 and took 938 seconds

Last LFC scan: 16 June 02:24 (Used to produce this page)

WARNING: Only datasets whose project name (character chain before the first point) includes one of the following list are treated in this page :
 - csc11 mc11 mc12 calib0 calib1 testIdeal testMisal mcMisal stream valid1

All informations are extracted from LFC catalog and partly DDM

[Summary on disk occupancy per Athena version \(page bottom\)](#)

Back to [summary table](#)

Clicking on dataset will open a window displaying all informations provided by [AMI](#)

- Green bar : All files from the dataset in the site
- Blue bar : Part of the files from the dataset in the site
- Red bar : More files in the site than registered in DDM

Total space occupied by the dataset [as function of time](#)

File name	AODs in DDM	AODs at DESY-HH	AODs %	Total size GB	Mean size MB	Production Cloud	Transfer Subscr.
calib0_mc12.007040.singlepart_gamma_Et20.recon.AOD.v12000601_tid006813	40	40	100%	0.1	3.3	NORDIC	-
csc11.005403.SU3_jimmy_susy.atlfast.AOD.v12000601_tid005895	50	50	100%	17.4	356.2	OSG	-
csc11.005403.SU3_jimmy_susy.atlfast.AOD.v12000602_tid006878	50	50	100%	17.3	354.7	OSG	-
ideal0_mc12.007061.singlepart_e_E100.recon.AOD.v12000601_tid006666	200	199	99.5%	0.8	4.4	OSG	-
mc12.005001.pythia_minbias.atlfast.AOD.v12000601_tid005832	2	2	100%	0.2	80.0	NORDIC	-
mc12.005200.T1_McAtNlo_Jimmy.atlfast.AOD.v12000602_tid007109	200	200	100%	69.8	357.5	OSG	-
mc12.005200.T1_McAtNlo_Jimmy.atlfast.AOD.v12000603_tid007247	198	198	100%	69.0	357.1	OSG	-
mc12.005200.T1_McAtNlo_Jimmy.atlfast.AOD.v12000604_tid008247	200	158	79%	55.0	356.2	OSG	-
mc12.005202.Mcatnlo_jim_top_leptpt120.atlfast.AOD.v12000604_tid009930	10	10	100%	3.6	368.7	OSG	-
mc12.005565.T1_McAtNlo_noUE.atlfast.AOD.v12000604_tid009257	100	50	50%	9.2	188.4	FR	-
mc12.006250.AcerMCtbar.atlfast.AOD.v12000602_tid008543	100	100	100%	26.8	274.1	NL	-
mc12.006251.AcerMCtbar.atlfast.AOD.v12000602_tid008544	99	64	64.5%	14.2	227.6	ES	-

Completed fraction

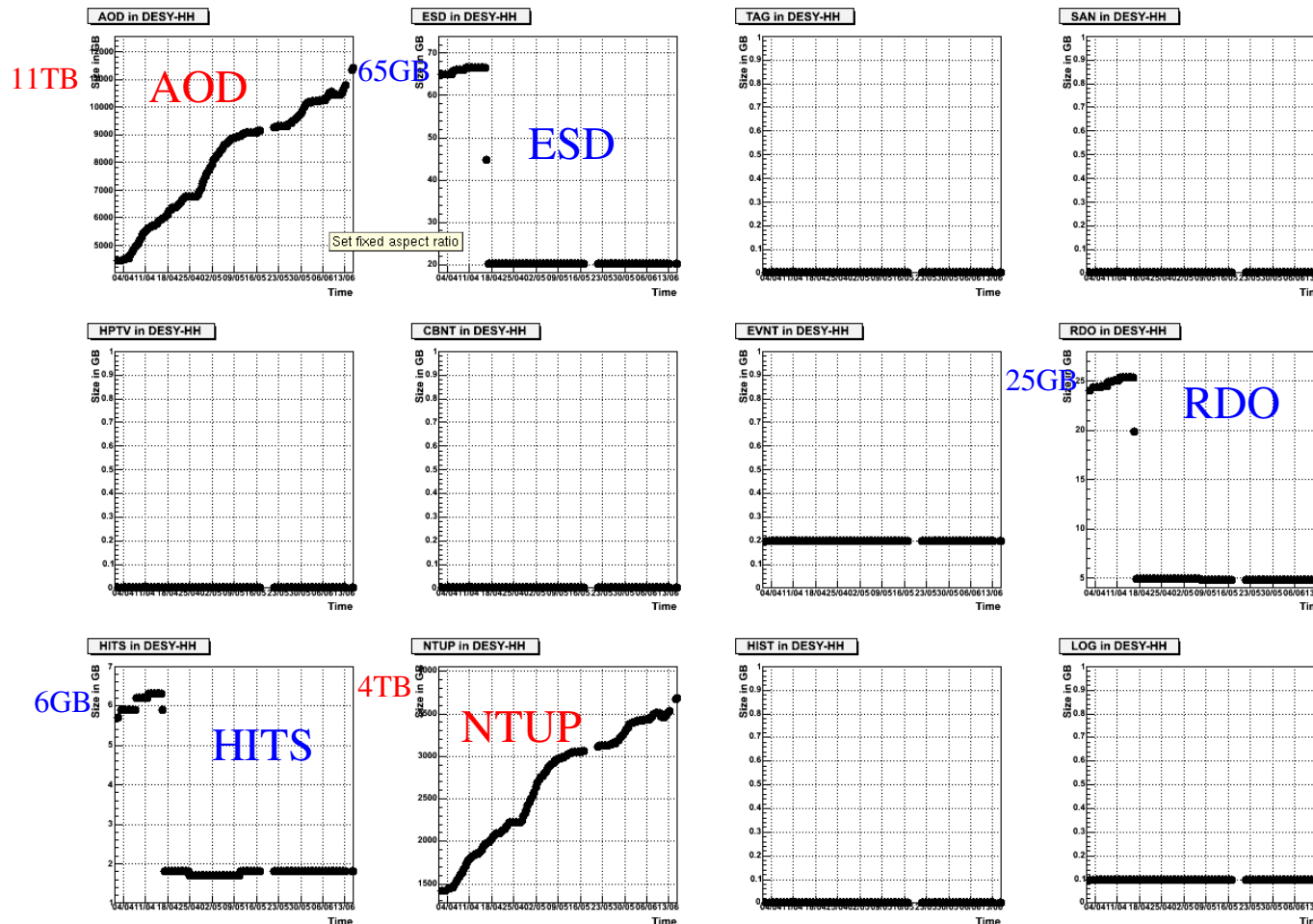
Done

Prod-Sys Aggregation

- Jobs from prod-sys can run all over the world
- Data is however assigned to a Cloud (DE)
- Output from the jobs saved to T1 or T2
- Currently Majority of our disk usage at T1 is from Production
- The aggregated AOD data is then Replicated to other T1's and T2's
- Smooth running production leads to better disk usage on our T1/T2's

DESY-HH^{what's happening}

- AOD/NTUP comes in ^{For Physics Analysis}
- Production data stored temporarily ^{Cleaned after rep to T1}

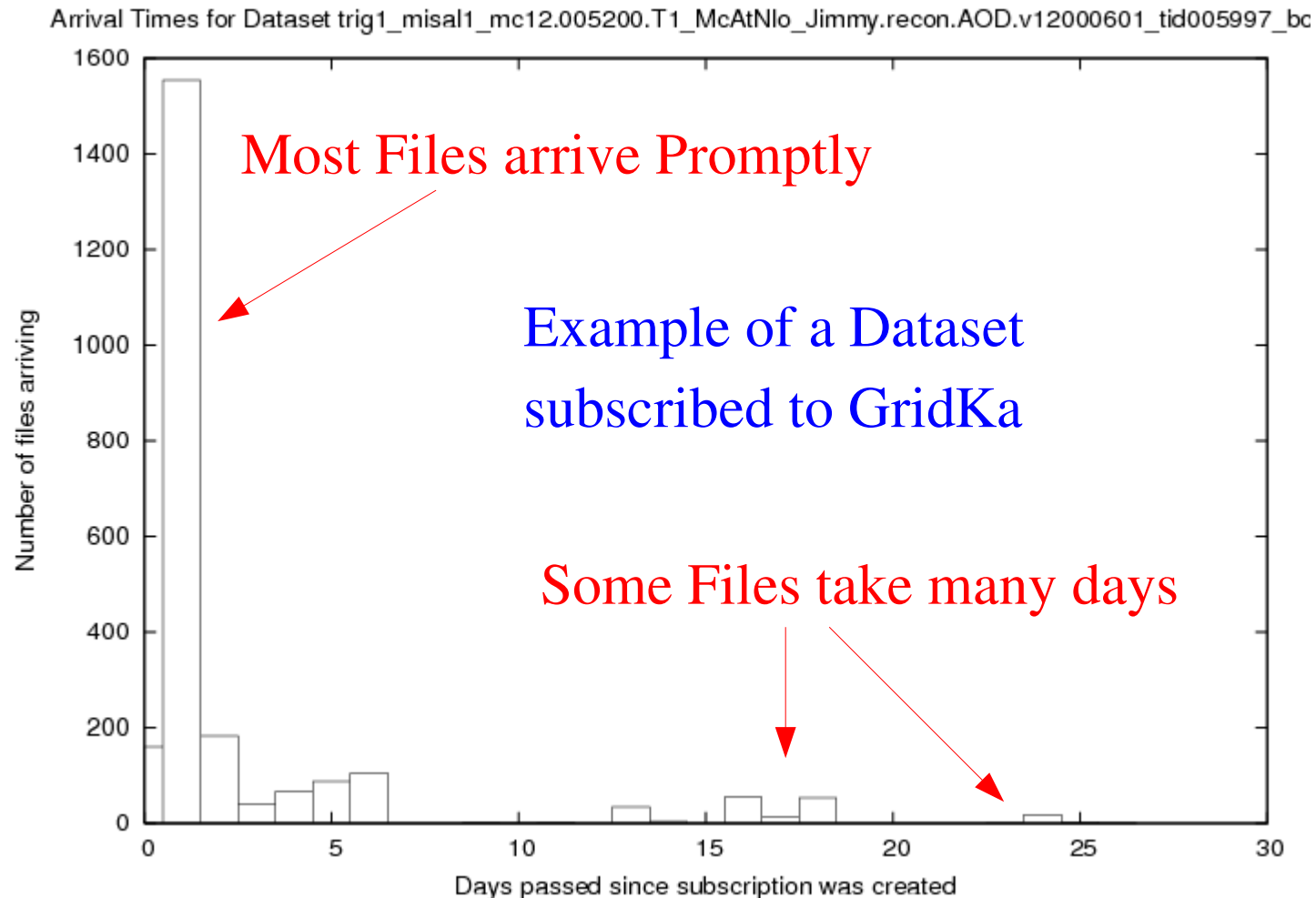


Replication

Production

DataSet Arrival Times

- Get all the Data
- Get it Promptly



DDM Ops Tasks

- Integrity checks
- Monitor site usage space/transfers – react to problems
- Ensure AOD replication working
- Ensure data aggregation working, with cleaning scripts
- Develop tools to help monitor and manage
- React to problems (trouble tickets) from central ddm ops team

Conclusion/Outlook

- **Distributed Data Management isn't easy!**
- We are doing well but need to push a little more
- Need to prepare for Distributed Analysis – Ganga – we have AOD replicated for this.
- Test sustained transfer T0->T1->T2
- Ensure that we get complete datasets and get them promptly
- Work on tools framework to optimise our cloud
 - Integrity
 - Data cleaning after use
- Keep improving knowledge in DDM
- Commit more manpower to DDM-Ops

More info – web sites

- Computing operations wiki:
<https://twiki.cern.ch/twiki/bin/view/Atlas/ComputingOperations>
- DDM wiki:
<https://twiki.cern.ch/twiki/bin/view/Atlas/DistributedDataManagement>
- DDM Central Operations wiki:
<https://twiki.cern.ch/twiki/bin/view/Atlas/DDMOperationsGroup>
- T0 test wiki:
<https://twiki.cern.ch/twiki/bin/view/Atlas/TierZero20071>
- GridKa Cloud Summary:
<http://www.etp.physik.uni-muenchen.de/ddm/DE/summary.html>
- ARDA Dashboard:
<http://dashb-atlas-data-test.cern.ch/dashboard/request.py/site>
- GridKa Cloud wiki:
<https://twiki.cern.ch/twiki/bin/view/Atlas/GridKaCloud>

More Info - People

- Central DDM Ops
 - Alexei Klimentov – Working Group leader
 - Stephane Jezequel – Data Co-ordinator
- GridKa Cloud DDM Ops (approx 1.1FTE)
 - Cedric Serfon
 - Andrzej Olszewski
 - Jiri Chudoba
 - Kai Leffhalm^{TBC}
 - John Kennedy
 - Max Klinger^{Project Student}
- Use atlas-germany-computing@desy.de to contact our ddm group

Thanks for you time and attention!