ES

CERN IT Department

Data Popularity in a Federated Environment

Domenico Giordano
(CERN IT-ES)

Creating Federated Data Stores For the LHC
13/09/2012

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Outline

Introduction

Reminder of the project proposal 1 year ago

Current Status

Conclusions

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Introduction

The original concept of Data Popularity is

"to tracks the usage statistics along time of the official data accessed by the users on WLCG"

- Nowadays the experiments (ATLAS, CMS, LHCb) have adopted their own service to monitor the data popularity, mainly based on the experiment specific tools for the distributed analysis

In the context of the Federated Data Storage the Data Popularity assumes a slightly different concept:

"Data mining of the monitoring streams collected from each disk server of the federation"

- The file based information, properly aggregated, allows to answer to several monitoring needs
  - Patterns of usage for each server/client pair: local Vs remote access, kind of access (copy, direct access,...)
  - Popularity of files (and set of files - datasets) in terms of bytes read, number of accesses, number of users
  - Performance in file access for classes of users and user activities (grid copy, production, analysis, ...)

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Starting motivations

Effort started from the CMS interest in monitoring the popularity of the data accessed via Xrootd

- Monitor data popularity also for batch/interactive job submissions
  - The view that is missing in the popularity system based on distributed analysis monitoring
- Can help in managing the user space on a site:
  - Providing feedback about not only the popularity of the official datasets, but also of the user data

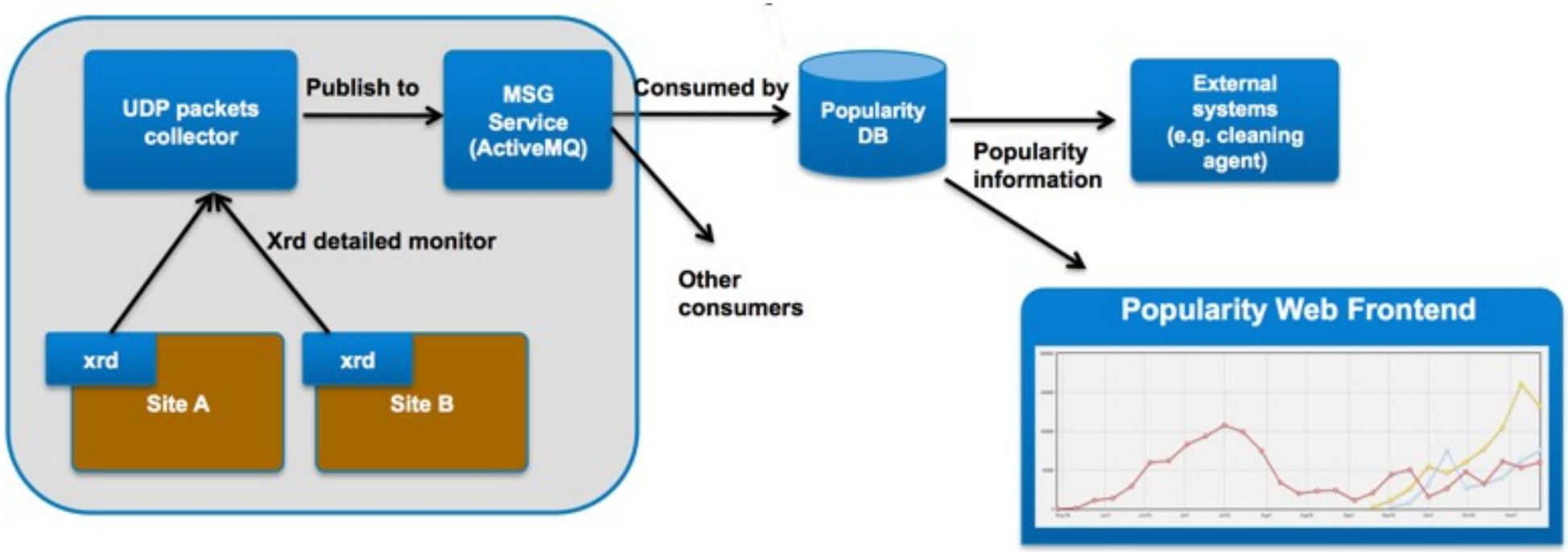
First use case: popularity of the files accessed at CERN from the CMS-EOS DataSvc

- Completed with analogous CASTOR popularity monitoring will help in balancing among archive of old data (on CASTOR) and copy of hot data (on EOS)

Extended also to EOS ATLAS, the AAA and FAX federations

- Joined other monitoring efforts (see Matevz, Julia, and Artem talks)

Architecture



Collector of the Xrootd **detailed monitoring** data see Matevz talk

- Based on UDP packets + udp2MSG service to publish in ActiveMQ
  - The next versions of the collector will directly write into ActiveMQ

Messaging System for Grid (MSG)

- Publish-subscribe model
  - Reduce the number of services collecting the UDP packets
  - Several consumers can access the MSG Broker

Popularity DB (Oracle DB):

- Collects the file based data extracted from the MSG Broker
- Allow data mining: aggregation and monitoring views exposed through a Web fronted

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

ES

Current Status

CERN IT Department

4 different instances of the "UCSD" collector are currently gathering Xrootd detailed monitoring UDP packets

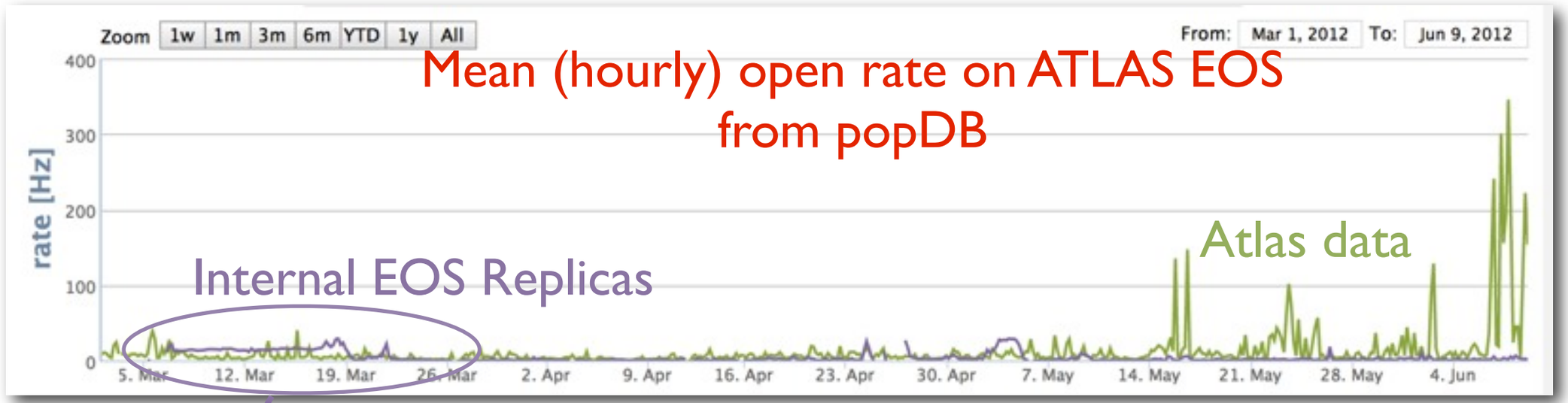
- xrootd.monitor all flush 30s mbuff 1472 window 5s dest user files

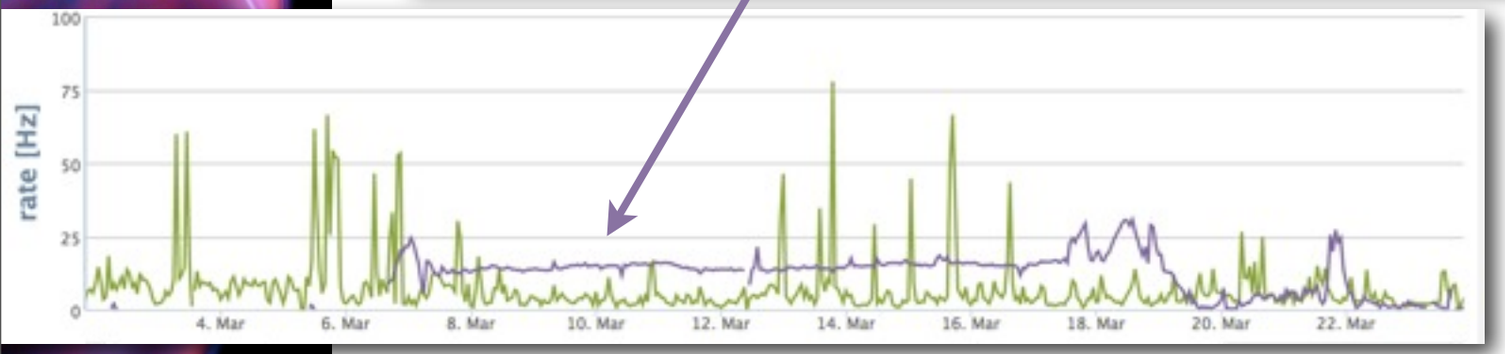
Collectors for	Running at	MSG-Broker	Topics	Consumers	Amount data	Since
EOS-CMS	CERN	gridmsg007.cern.ch	xrdpop.cms_popularity	Popularity	21 GB	03/12
EOS-Atlas	CERN	gridmsg007.cern.ch	xrdpop.eosa_popularity	Popularity	40 GB	03/12
US-CMS	UCSD	gridmsg007.cern.ch	xrdpop.uscms_popularity	- Popularity - dashb_wlcg	5 GB	06/12
FAX	FNAL	gridmsg007.cern.ch	xrdpop.fax_popularity	- Popularity - dashb_wlcg - realTime	35 MB	07/12

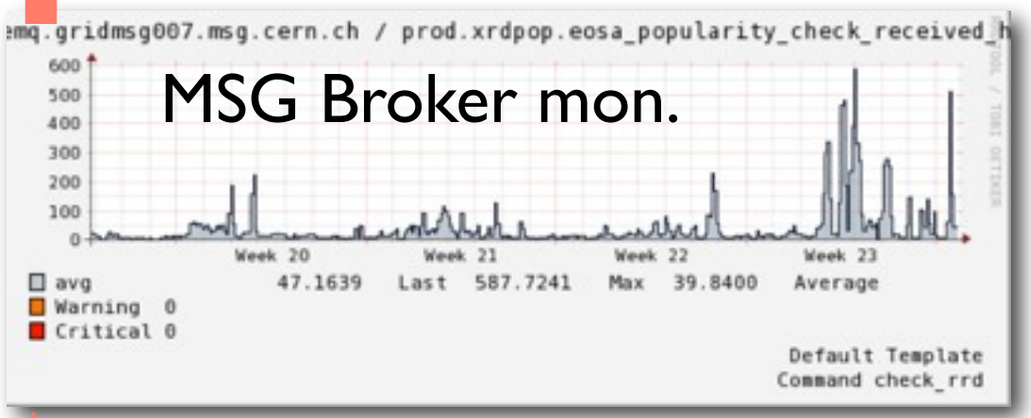
* soon collect at CERN also the detailed monitoring from ATLAS-EU sites

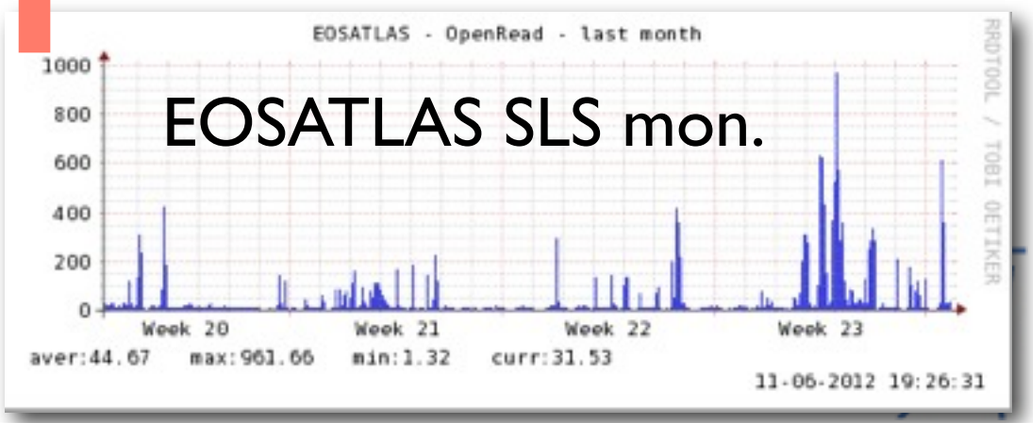
Able to cope with the current throughput

CERN IT Department









The designed workflow scales properly w.r.t. the current highest load (EOS-ATLAS)

- O(1k) messages can be consumed in few seconds by the MSG system
- Peaks of activity well correlated with other monitoring systems (EOS SLS, MSG Nagios)

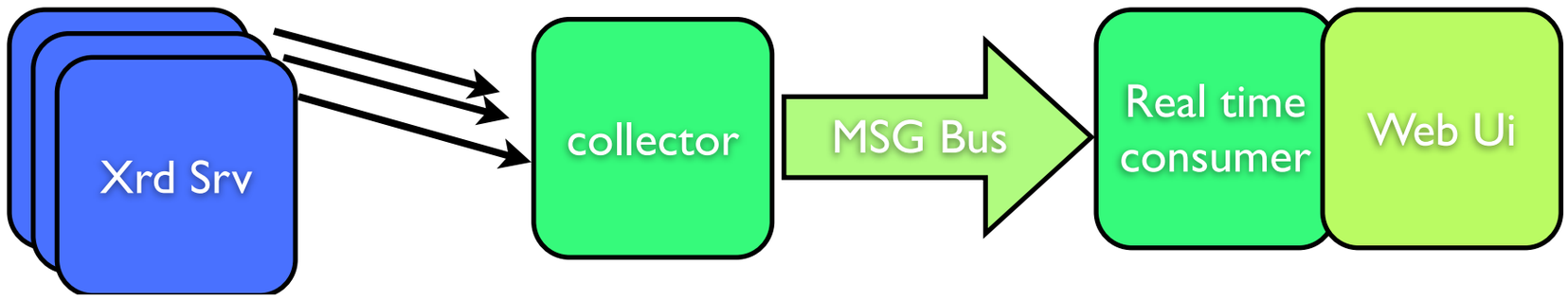
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Possibility of a real time monitoring

CERN IT Department

The prompt consumption of the messages allows a realtime monitoring of the accessed files

- Exposed through a web UI, can show per file open/closed any monitoring attribute collected and published
  - Useful for debugging



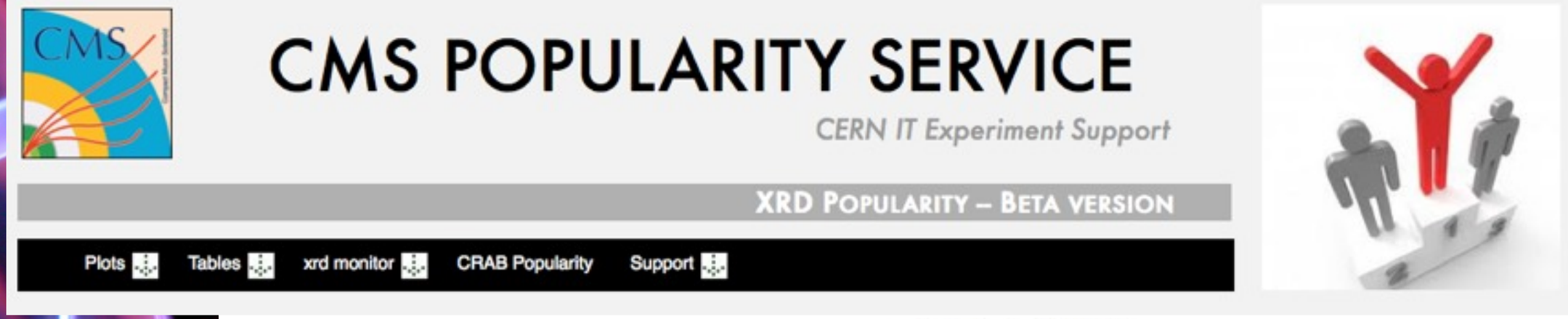
Here for FAX:Open messages in the last 10 min

starttime	endtime	server domain	client domain	read bytes	write bytes	file size	filename
2012-09-12 13:22:42	2012-09-12 13:24:37	bu.edu	aglt2.org	0	0	797152257	/atlas/dq2/user/lijav/HCtest /user.lijav.HCtest.1 /group.test.hc.NTUP_SMWZ.root
2012-09-12 13:21:44	2012-09-12 13:24:33	slac.stanford.edu	iu.edu	797152257	0	797152257	/atlas/dq2/user/lijav/HCtest /user.lijav.HCtest.1 /group.test.hc.NTUP_SMWZ.root
2012-09-12 13:21:44	2012-09-12 13:24:24	atlas-swt2.org	iu.edu	797152257	0	797152257	/atlas/dq2/user/lijav/HCtest /user.lijav.HCtest.1 /group.test.hc.NTUP_SMWZ.root
2012-09-12 13:22:41	2012-09-12 13:24:05	slac.stanford.edu	rc.fas.harvard.edu	797152257	0	797152257	/atlas/dq2/user/lijav/HCtest /user.lijav.HCtest.1 /group.test.hc.NTUP_SMWZ.root
2012-09-12 13:21:48	2012-09-12 13:23:28	iu.edu	iu.edu	0	0	797152257	/atlas/dq2/user/lijav/HCtest /user.lijav.HCtest.1 /group.test.hc.NTUP_SMWZ.root
2012-09-12 13:21:47	2012-09-12 13:23:27	bu.edu	iu.edu	0	0	797152257	/atlas/dq2/user/lijav/HCtest /user.lijav.HCtest.1

With the same approach other real time monitors can be put in place

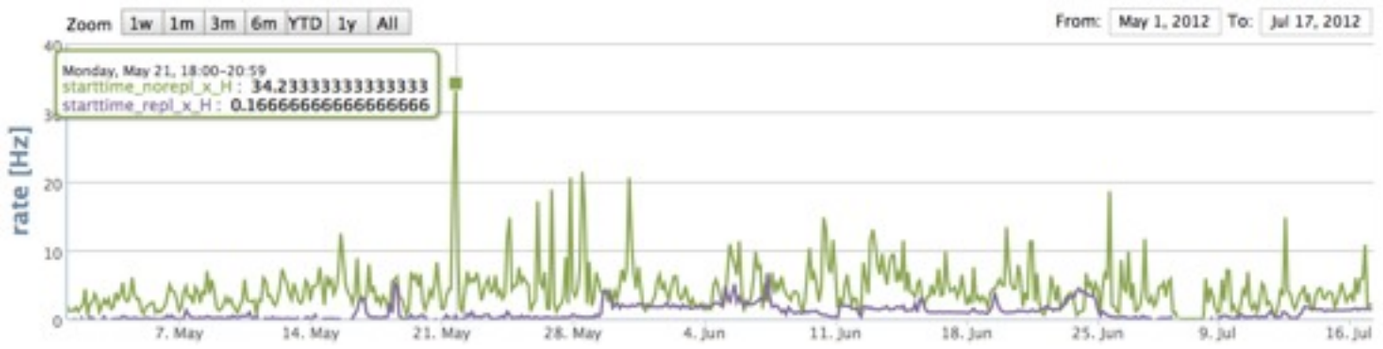
- Monitor of service's availability, load, etc

Web UI



XRD monitor Plots

- Monitor insertion rates



Popularity Metrics

- Most Popular DataTiers, DataSets (DS), Processed DS (campaigns) : PLOT & TABLE
- User Collections: PLOT & TABLE
  - Popularity of the user areas (eg. /store/user/giordano/*) in terms of accesses, users/day, ReadBytes , CPU time
- Breakdown by Local/GRID users
- User Stats TABLE (breakdown by DataTiers)

Similar views can be provided to ATLAS

- Some examples in the next slides

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

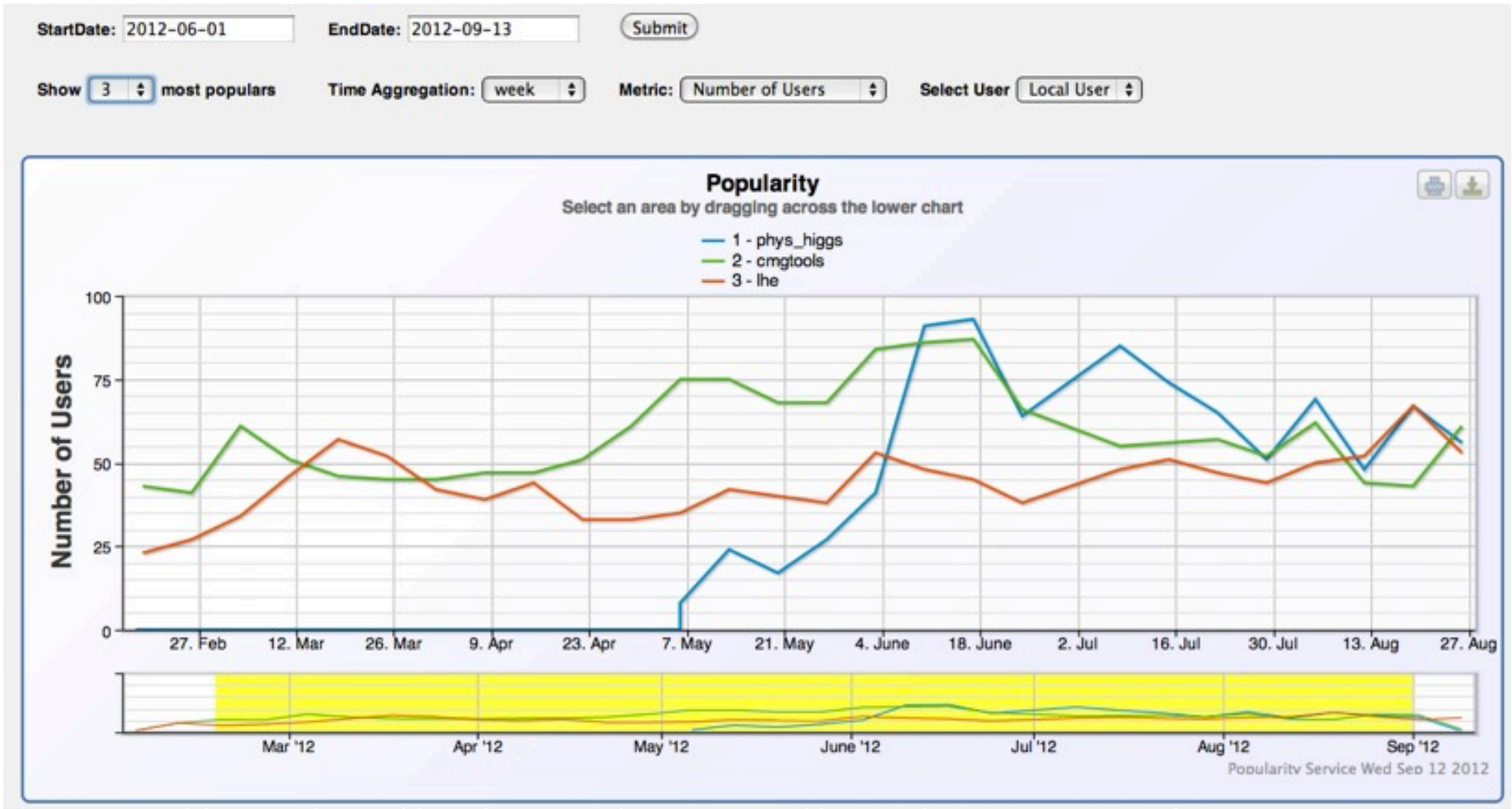
ES

Example: user data popularity

CERN IT Department

Monitor access patterns for data in the user areas

- `/eos/cms/store/*/user/giordano/*`
- Aggregate per Local / Grid user



Weekly number of local users of CMS-EOS files in the 3 most popular user areas

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Example: Project/DataType on EOS AtlasDataDisk

CERN IT Department

Breakdown per activities

- GridProduction (atlprd*) , GridAnalysis (atlplt*), DDM (central - atlas003, user -atlas001), users

PROJECT/DATATYPE POPULARITY

Popularity of Project/DataType in terms of # Accesses, read TBytes, in a specific time window, for the file accesses in **atlasdatadisk**

StartDate: 2012-06-01 **EndDate:** 2012-09-13 Submit Query

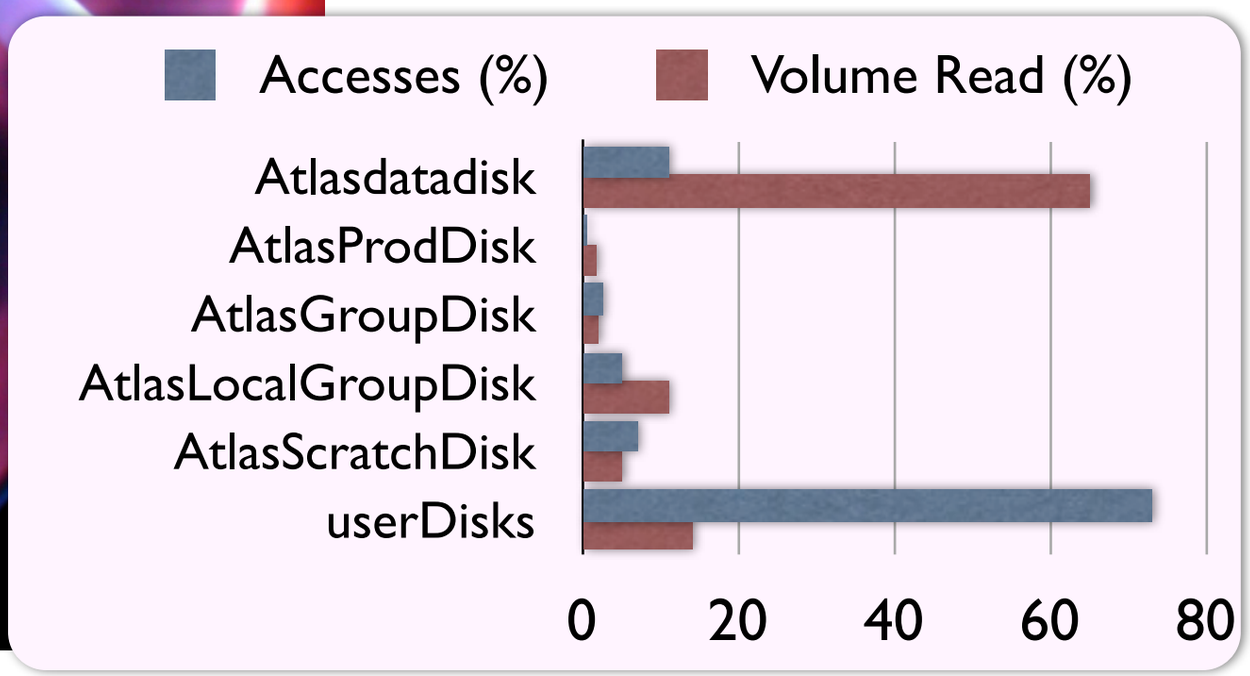
Show 10 entries Search:

Project/DataType	Username	Accesses		read TBytes	
		[N]	[%]	[TB]	[%]
mc12_8TeV/HITS	GridProduction	1777856	12.6	767.7	12.9
mc11_7TeV/HITS	GridAnalysis	1189270	8.4	716.2	12.0
data12_8TeV/NTUP	DDMCentralActivity	758982	5.4	643.6	10.8
data12_8TeV/AOD	GridProduction	269612	1.9	531.5	8.9
data12_8TeV/AOD	DDMCentralActivity	195711	1.4	427.8	7.2
data12_8TeV/AOD	GridAnalysis	996858	7.1	360.6	6.1
data12_8TeV/NTUP	GridAnalysis	1571935	11.2	298.4	5.0
data12_8TeV/RAW	GridProduction	183159	1.3	291.7	4.9
data12_8TeV/AOD	DDMactivity	85222	0.6	205.8	3.5
data12_8TeV/AOD	users	478656	3.4	113.2	1.9
Shown Sum		**7507261**	**53.3**	**4356.5**	**73.2**
Total Sum		**14086503**	**98.69**	**5950.8**	**99.09**

ES Example: Eos-Atlas Disks usage

CERN IT Department

DiskName	UserName	Accesses		read TBytes	
		[N]	[%]	[TB]	[%]
userDisks	users	92247392	73.4	1299.2	14.3
GridDisks	users	18110823	14.4	1208.1	13.3
GridDisks	GridAnalysis	7342024	5.8	2072.6	22.9
GridDisks	GridProduction	3933903	3.1	2194.7	24.2
GridDisks	DDMCentralActivity	3245828	2.6	1892.4	20.9
GridDisks	DDMactivity	810117	0.6	399.7	4.4
userDisks	DDMactivity	2093	0.0	1.4	0.0
userDisks	GridAnalysis	4	0.0	0.0	0.0
Shown Sum **Total Sum**		**125692184** **125692184**	**99.9** **99.9**	**9068.1** **9068.1**	**100** **100**
Filter DiskName	Filter UserName	Filter Accesses N	Filter Accesses %	Filter read TBytes	Filter read TBytes %



Activity in the past 3 months for two categories of disks: GridDisks, userDisks

- userDisks ⇔ AtlasCernGroupDisk and user

ES

Understand the disk usage

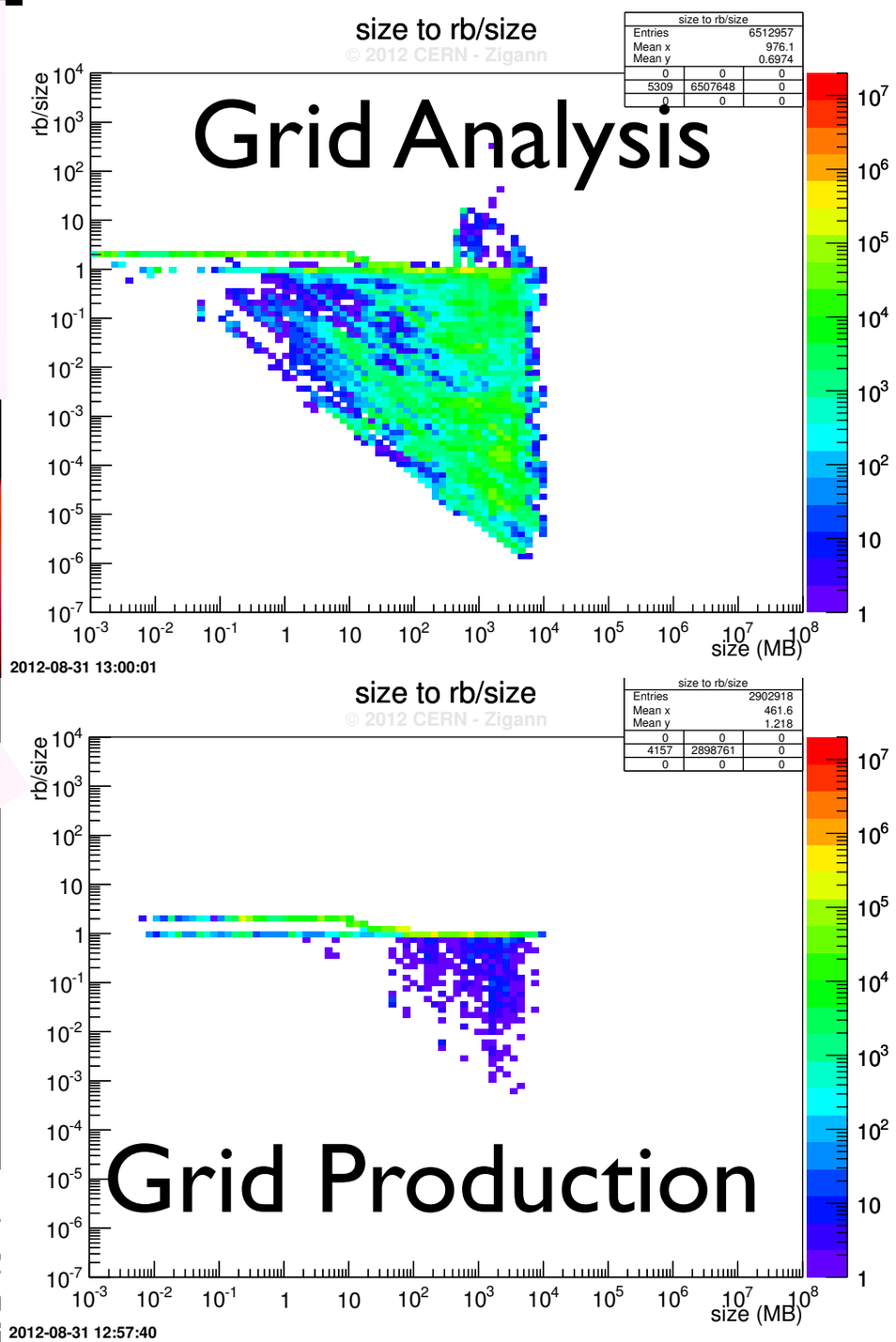
CERN IT Department

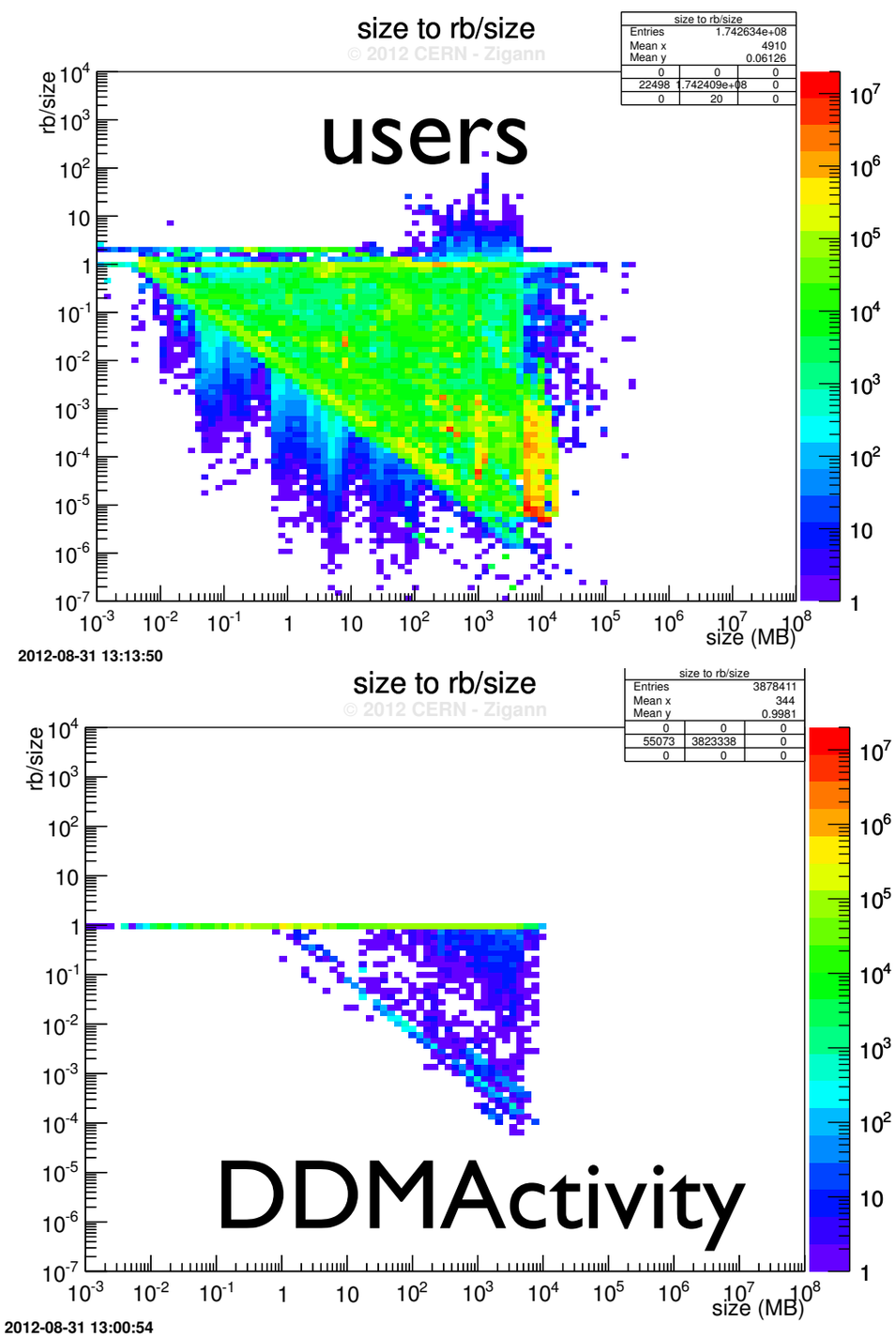
Correlate with classes of activities

Can help in optimizing the file distribution as well as the analysis approaches

Example: Correlate the readBytes Vs file size for different activities in EOS ATLAS

Plots from Philipp Zigann (IT-DSS)



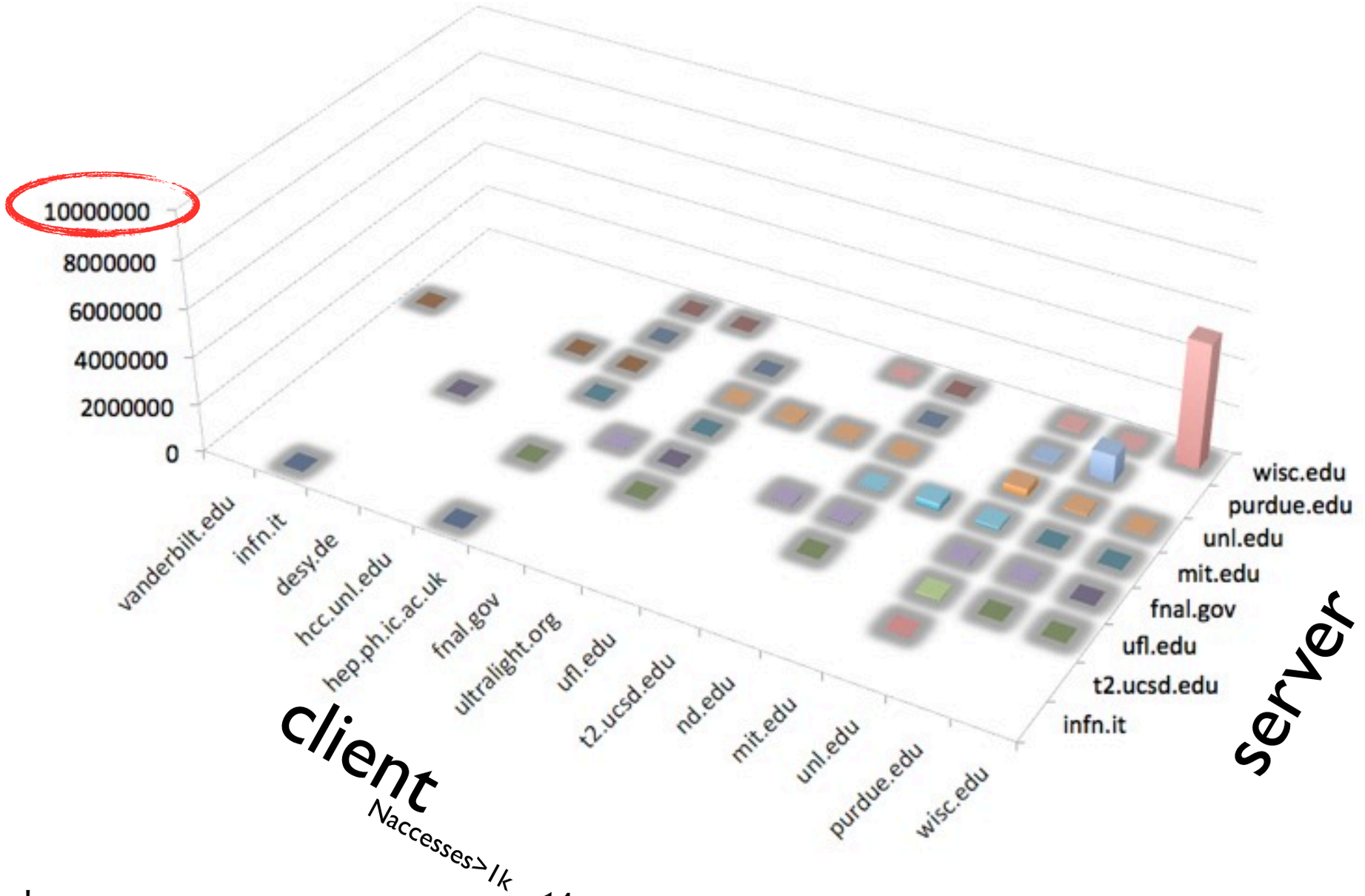


CERN IT Departmer
CH-1211 Geneva 2
Switzerlan
www.cern.ch/i

Local Vs Remote access

The client and server identifiers together with file name and read bytes allow to monitor the local Vs remote accesses

- Here displayed for the CMS Federation (for the last month)
  - Includes also the fallback of grid jobs failing while opening a file locally

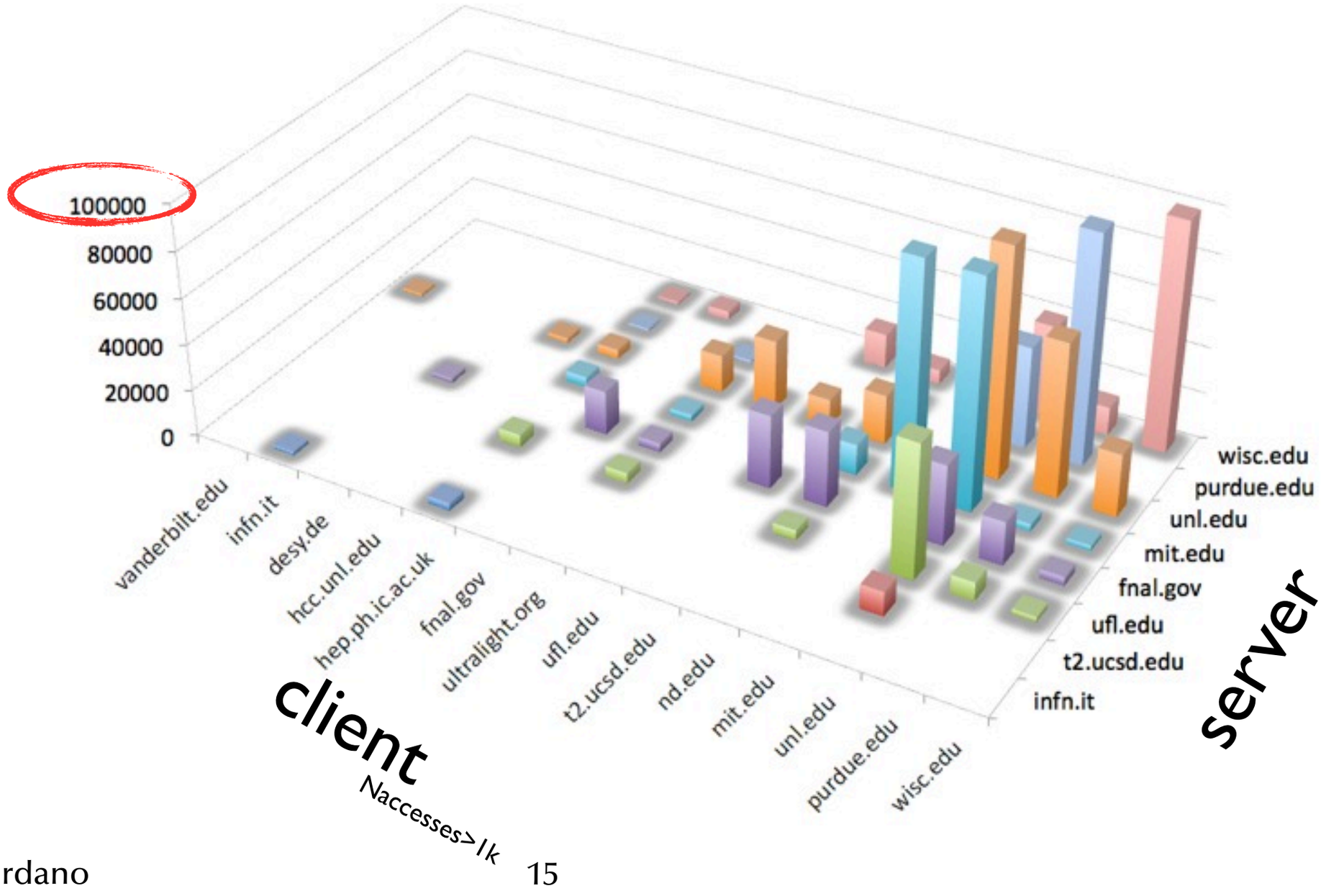


CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Local Vs Remote access

The client and server identifiers together with file name and read bytes allow to monitor the local Vs remote accesses

- Here displayed for the CMS Federation (for the last month)
  - Includes also the fallback of grid jobs failing while opening a file locally



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Some considerations

There are already many metrics accessible in the collector crucial for data mining

- Single read Vs Vector read, averaged quantities, max quantities

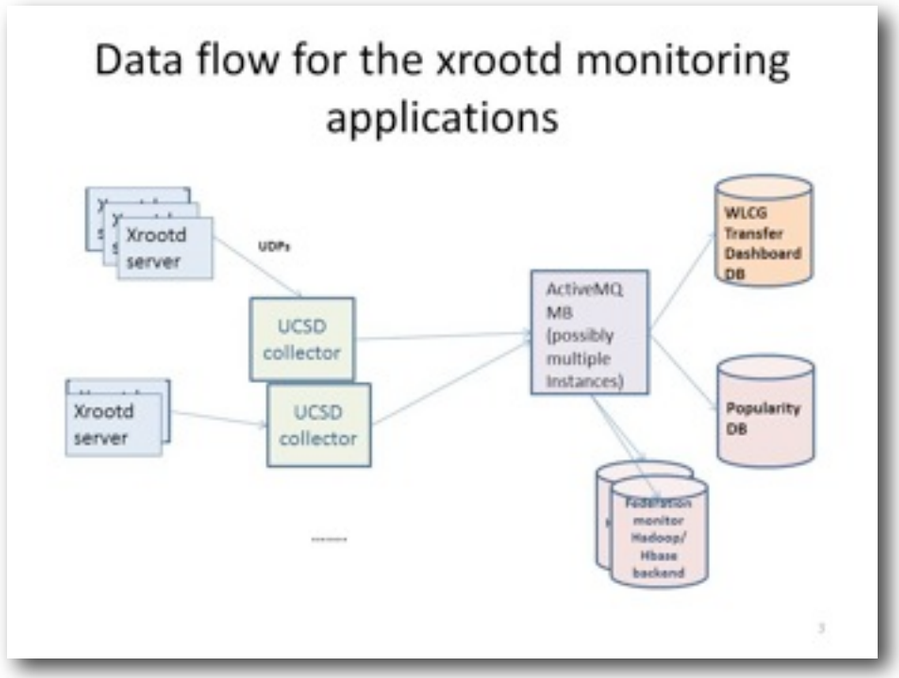
It would be useful to have the following additional information

- Site Name of the diskServers
  - To avoid to solve with other services the topology problem (convert server domains to a specific site name)
- Guarantee that the attribute file_lfn contains the lfn and not the pfn on disk
  - Would imply to query LFC to account the popularity of the same file in two different sites
- Application Info
  - Fill the dedicated field of the detailed monitor report with the identifier of the applications accessing files via xrootd
    - ✳ Instrument the applications to provide this information
  - Helps in distinguishing what is direct access, what is copy, what is official analysis/reconstruction fw, etc

Conclusions

The designed workflow for the Xrootd data popularity has proved to be reliable

- Collector + MSG System + DB
- Allows easy access of the same information to other systems and consumers
  (described in Julia & Artem talks)



If the "popularity" word is confusing because already identifies other tools, let's think at it as inclusive mining of the Federation Monitoring Data

- Can provide a deep insight into the performance of the Federated storage as well as into the patterns of usage of the official and user data

BACKUP

CMS Xrootd popularity workflow