DPM DOME

Petr Vokáč

HEPiX

27th March 2019

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

DPM overview

- Popular grid storage system
  - multi-protocol
    - SRM, GridFTP, XRootD WebDAV
  - multi-vo
    - X.509 / VOMS and token based authentication, ACL
  - scalable to medium size sites ~ 10PB
    - headnode – maintains metadata, namespace and ACL + redirect clients
    - disknodes – data storage
  - built on top of common software
    - XRootD, Apache, MySQL, Globus
  - supported by CERN IT
    - relatively easy to deploy (puppet)
    - integrated with WLCG workflows

DPM

Disk Pool Manager

BDII registered

- 120 endpoints
  - 80 srmping reach.
- 100PB

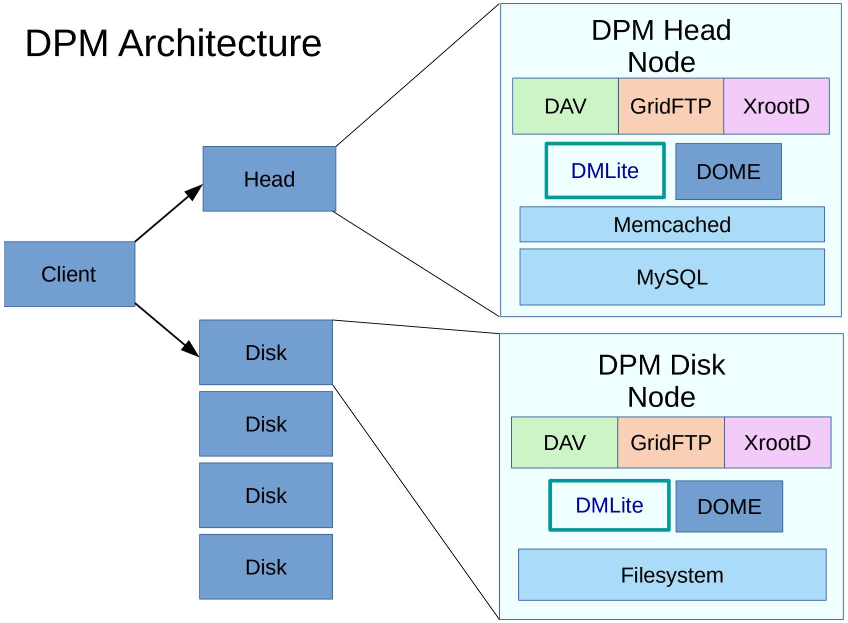
ATLAS Storage software overview

DPM	40
dCache	34
StoRM	18
Other	20
Σ	112

DPM overview

- Popular grid storage system
  - multi-protocol
    - SRM, GridFTP, XRootD WebDAV
  - multi-vo
    - X.509 / VOMS and token based authentication, ACL
  - scalable to medium size sites ~ 10PB
    - headnode – maintains metadata, namespace and ACL + redirect clients
    - disknodes – data storage
  - built on top of common software
    - XRootD, Apache, MySQL, Globus
  - supported by CERN IT
    - relatively easy to deploy (puppet)
    - integrated with WLCG workflows

DPM
Disk Pool Manager



DPM flavours

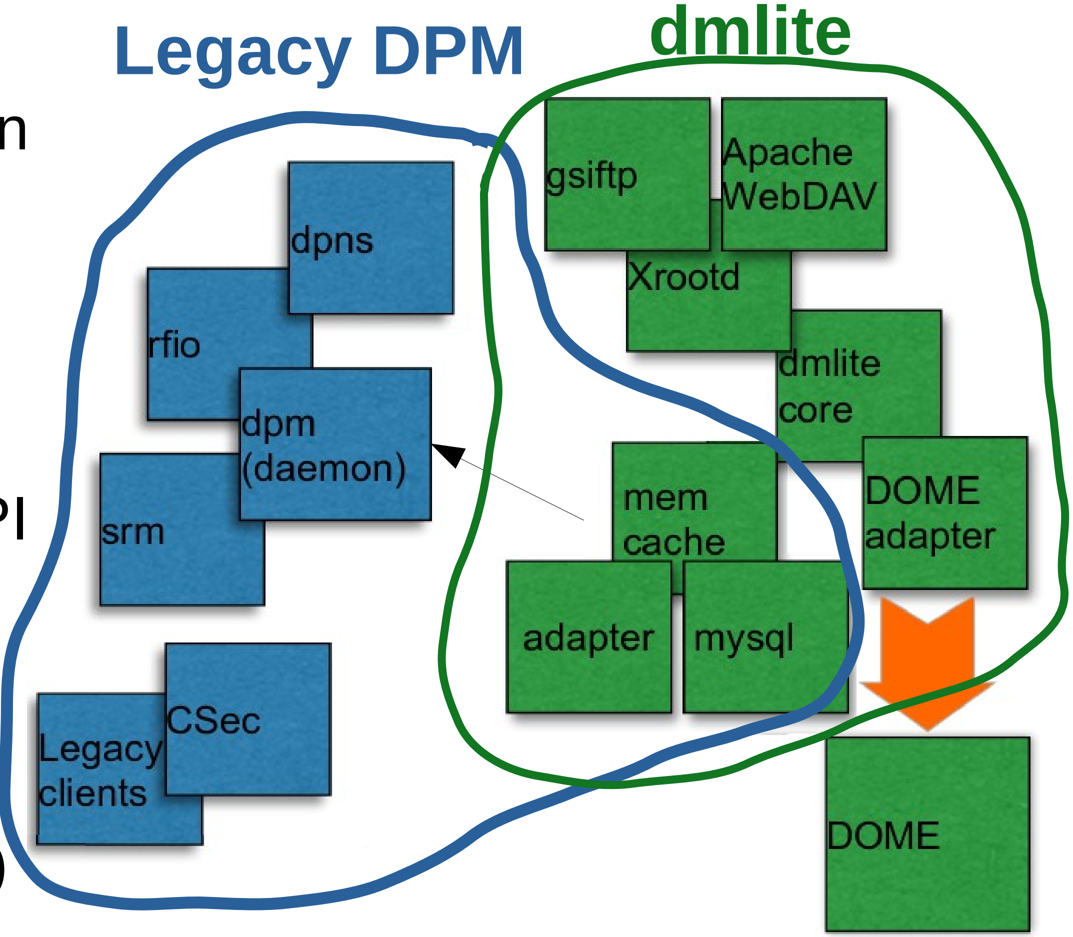
Legacy

vs.

DOME

DPM flavours

- **Legacy DPM** (lcgdm)
  - old design and codebase – development frozen
  - dpm daemon does coordination
    - every process loads dmlite libraries
    - per process db connections
  - performance bottlenecks (cachin
  - sub-optimal resource usage
- **DOME** (Disk Operations Management Engine)
  - coordination daemon (REST API
    - called by dmlite library
    - DPM 1.9 fastcgi (2016)
    - DPM 1.10 XrdHTTP (2018)
  - dpm daemon for SRM (adapter)



DPM flavours – protocols

DPM

SRM (GridFTP), RFIO

Legacy DPM + dmlite

SRM (GridFTP), RFIO

XRootD, WebDAV

SRM (GridFTP), RFIO

GridFtp, XRootD, WebDAV

Legacy DPM + dmlite DOME

GridFTP

XRootD, WebDAV

dmlite DOME

LCGDM EOL

Legacy DPM support ends with 1st June 2019

LCGDM EOL

~~DPM~~

~~SRM (GridFTP), RFIO~~

~~Legacy DPM + dmlite~~

~~SRM (GridFTP), RFIO~~

~~XRootD, WebDAV~~

~~SRM (GridFTP), RFIO~~

~~GridFtp, XRootD, WebDAV~~

~~Legacy DPM + dmlite DOME~~

GridFTP

XRootD, WebDAV

dmlite DOME

LCGDM EOL

~~**DPM**~~

~~**SRM (GridFTP), RFIO**~~

~~**Legacy DPM + dmlite**~~

~~**SRM (GridFTP), RFIO**~~

~~**XRootD, WebDAV**~~

- announced at the beginning of 2018
- it doesn’t mean legacy DPM disappears
  - no more updates in legacy codebase
  - available in repositories while it compiles
    - mostly c code that should not break
- Standard answer for legacy components:

“upgrade to DOME flavour”

~~**SRM (GridFTP), RFIO**~~

~~**GridFtp, XRootD, WebDAV**~~

~~**Legacy DPM + dmlite DOME**~~

GridFTP

XRootD, WebDAV

dmlite DOME

LCGDM EOL

~~**DPM**~~

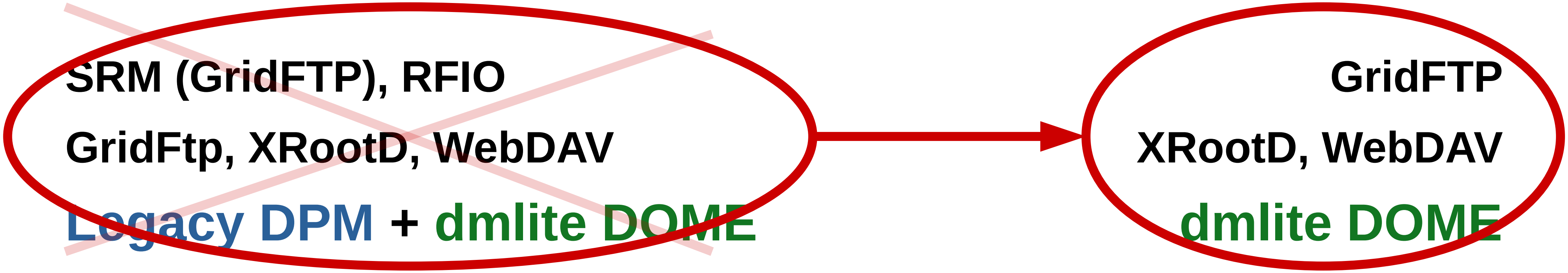
~~**SRM (GridFTP), RFIO**~~

~~**Legacy DPM + dmlite**~~

~~**SRM (GridFTP), RFIO**~~

~~**XRootD, WebDAV**~~

- plan suitable for most sites with DPM
  - upgrade to the latest DPM + dmlite version
  - enable DOME configuration
  - try to reduce SRM usage
    - WLCG no longer needs SRM
    - not effortless for other users / experiments
  - remove legacy DPM (SRM)



~~**SRM (GridFTP), RFIO**~~

~~**GridFtp, XRootD, WebDAV**~~

~~**Legacy DPM + dmlite DOME**~~

GridFTP

XRootD, WebDAV

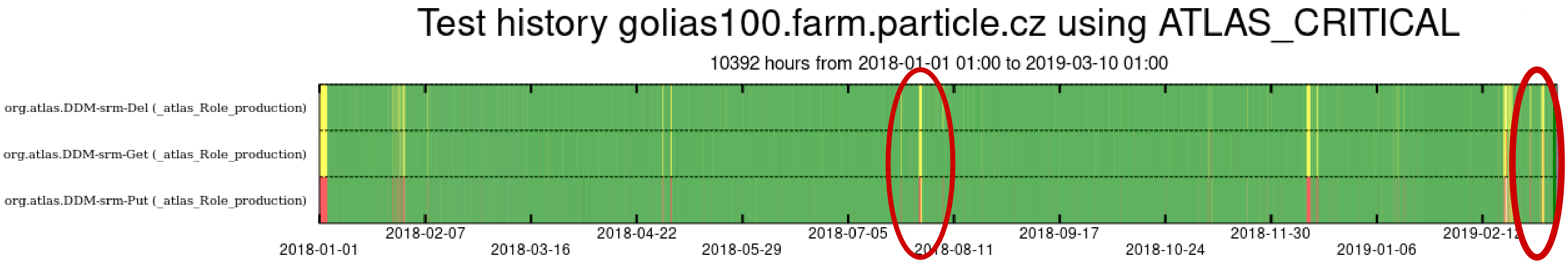
dmlite DOME

Upgrade motivation

- EOL June 1st 2019 for legacy DPM
- Poor performance caused by ancient DPM design
  - cluster size increasing → number of jobs higher → more transfer requests → DPM headnode can easily become overloaded (stuck)
- No new features in the legacy codebase and for DPM running without DOME enabled
  - no checksums
  - no TPC
  - poor support for non-SRM protocols
- Issues that can’t be really solved within legacy DPM
  - empty spacetoken quotas
  - FTS/gfal supports spacetokens only for SRM
- Software with latest security fixes

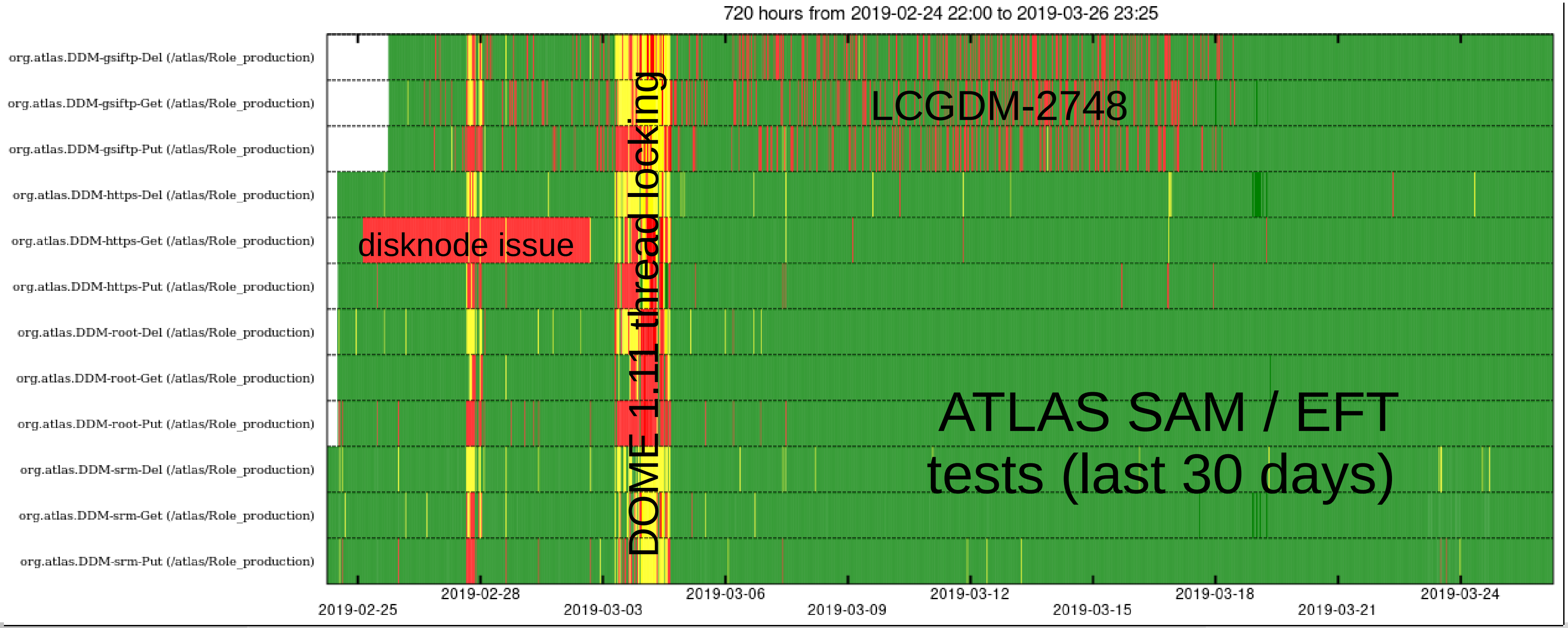
Upgrade status

- With a few exceptions all DPM sites still run legacy mode
  - only ~ ½ sites upgrade DPM packages regularly
    - regardless of site size
  - only fraction moved also to DOME since its release ~ 2 years ago
    - DPM 1.9 with dmlite 0.8.8 – only on testbeds – never in production
    - DPM 1.10 with dmlite 1.10
      - in production at PRAGUELCG2 since summer 2018
      - transfers still mostly SRM (legacy stack) + XRootD (DOME stack)
      - since March 5th 2019 ATLAS transfers with pure GridFTP
      - site and service availability monitor for this DPM looks relatively fine



Upgrade status

- With a few exceptions all DPM sites still run legacy mode
  - only ~ ½ sites upgrade DPM packages regularly
    - regardless of site size
  - only fraction moved also to DOME since its release ~ 2 years ago
    - DPM 1.9 with dmlite 0.8.8 – only on testbeds – never in production
    - DPM 1.10 with dmlite 1.10
      - in production at PRAGUELCG2 since summer 2018



ATLAS SAM / EFT tests (last 30 days)

WLCG DPM Upgrade TF

- Scope of DPM upgrade task force activities
  - coordinate upgrade of the DPM sites
    - DPM 1.12.1 released in EPEL testing
    - upgrade to the DOME mode (legacy EOL in summer 2019)
    - enable WLCG Storage Resource Reporting (SSR)
  - provide guidance and support for upgrade and reconfiguration
  - validate SSR reporting and later integrate to CRIC and WLCG Storage Space Accounting
- Upgrade plan
  - phase 1 (end of 2018)
    - get experience DPM DOME experience with small number of sites
  - phase 2
    - by summer 2019 80% of DPM storage (in terms of capacity)
    - by the end of 2019 80% of DPM sites

DOME packages and configuration

- Follow official documentation
  - upgrading spacetokens → quotatokens
    - dmlite-shell quotatokenmod
    - dmlite-mysql-dirspaces.py
  - puppet configuration options (configure_legacy, configure_dome, configure_comeadapter)
- Minimum recommended software versions
  - DPM version 1.12.0 + dmlite version 1.12.x + dpm-dsi 1.15
  - XRootD 4.9.x
  - davix 0.7.3 (not very important – prevent overloaded DOME crash)
- Additional puppet configuration to enable new features
  - enable gridftp redirection (gridftp_redirect)
    - add globus-GridFTP to GOCDB for Argus monitoring
  - enable XRootD checksum (configure_dpm_xrootd_checksum)
  - enable XRootD TPC (headnode configure_dpm_xrootd_delegation)

ATLAS DOME configuration

- AGIS SE configuration
  - GridFTP for wan_read, wan_write, third_party_copy
  - XRootD for lan_read, lan_write
  - can take up to ~ 2 hours till Rucio pick new configuration
    - rucio-admin rse info RSE_NAME
    - https://rucio-ui.cern.ch/rse?rse=RSE_NAME
- AGIS Panda queues
  - rucio mover
- SAM tests for non-SRM sites
  - update test templates to consider GridFTP protocol
- Publish WLCG SSR information with cron job every 15 minutes

```
/usr/bin/dpm-storage-summary.py --path /dpm/fqdn/home
```

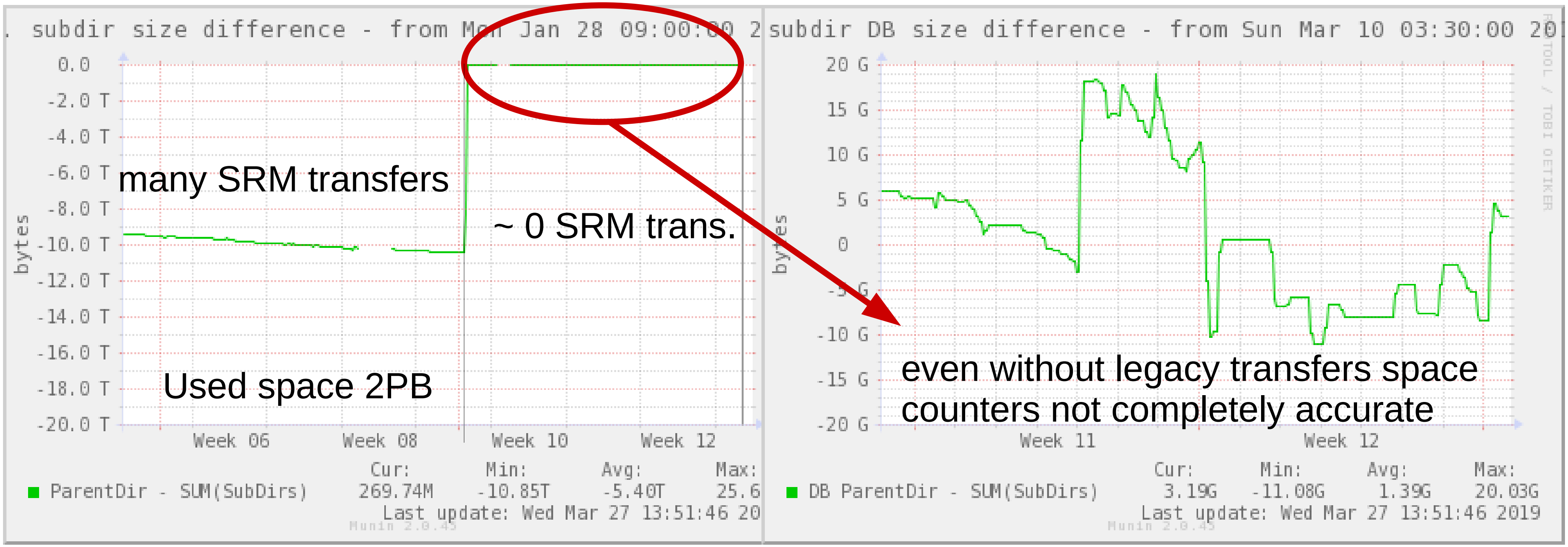
Upgrade issues

- puppet (re)configuration is relatively simple
  - can be done without puppet server
    - locally with minimal head/disk server puppet configuration file
  - a lot of variations in site DPM configurations
    - issues caused by migration to CentOS7 (vs. SL6 cfg)
    - database on dedicated machine – open file handle limit
  - most of configuration options should be now fine for production
  - necessary to follow DPM tuning hints
    - common issue – limit on number of open file handles
    - local DNS caching service (nscd or local DNS)
      - should become less important once SRM is completely dropped
- Manual configuration is error prone
  - even with our small number of upgraded sites
  - problems in puppet modules gets promptly fixed
- SSR reporting robustness vs. SRM lcg-stdm

Upgrade issues

- GridFTP redirection
  - GridFTP still used exclusively for TPC
  - necessary for pure GridFTP without SRM
  - enabled just recently on three sites
    - LCGDM-2748 can cause troubles in case one of diskserver is down (workaround – manual removal from gridftp.conf)
  - SRM performance lower ~ 20% with this configuration
    - only metadata operations that mostly affect small file transfers
- Spacetokens vs. quotatoken
  - tag based spacetoken can be design diverge from path based quotatokens (SRM transfers; should not happen with Rucio)
    - enforcing wrong quota
  - issues with (slowly) diverging counters – concurrency issue
  - physical disk space issues with SRM transfers (fixed in 1.12)
- Failing apache graceful restarts (CRL reloads; logrotate; 2.4.35)

Upgrade issues


many SRM transfers

~ 0 SRM trans.

Used space 2PB

even without legacy transfers space counters not completely accurate

- Spacetokens vs. quotatoken
  - tag based spacetoken can be design diverge from path based quotatokens (SRM transfers; should not happen with Rucio)
    - enforcing wrong quota
  - issues with (slowly) diverging counters – concurrency issue
  - physical disk space issues with SRM transfers (fixed in 1.12)
- Failing apache graceful restarts (CRL reloads; logrotate; 2.4.35)

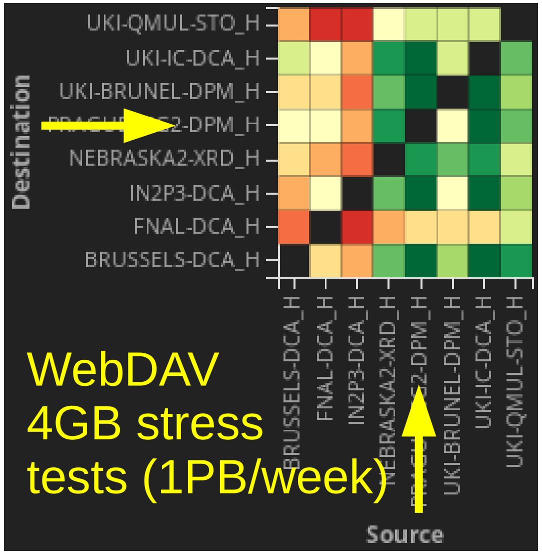
Upgrade issues

- Almost hundred tickets addressed since summer 2018
  - DPM & dmlite 1.10 → 1.12
  - most of problems found & fixed this year
    - found thanks to participating pioneering sites within WLCG TF
    - issues fixed also in software that is used by DOME DPM
    - no critical ticket currently opened for normal operation
    - several sites with PB+ storage use DOME (few with gridftp redirection)

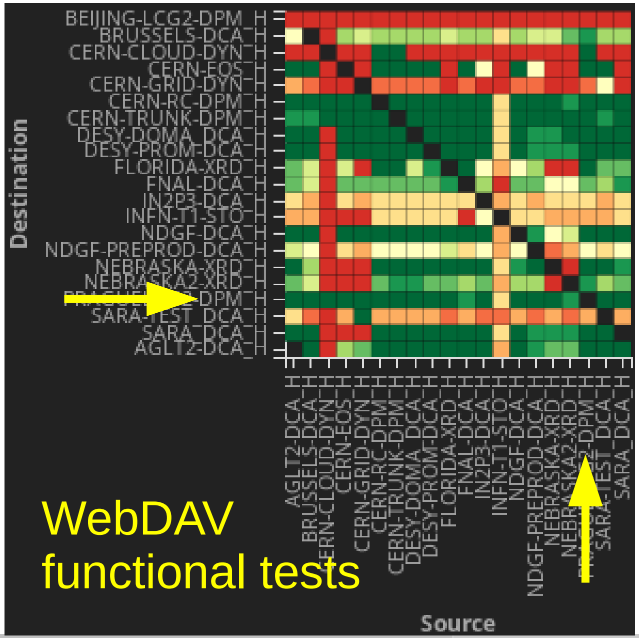
Using legacy **SRM** together **with**
DOME **WebDAV, GridFTP, XrootD**
within one VO should be **avoided**

DOME features – TPC support

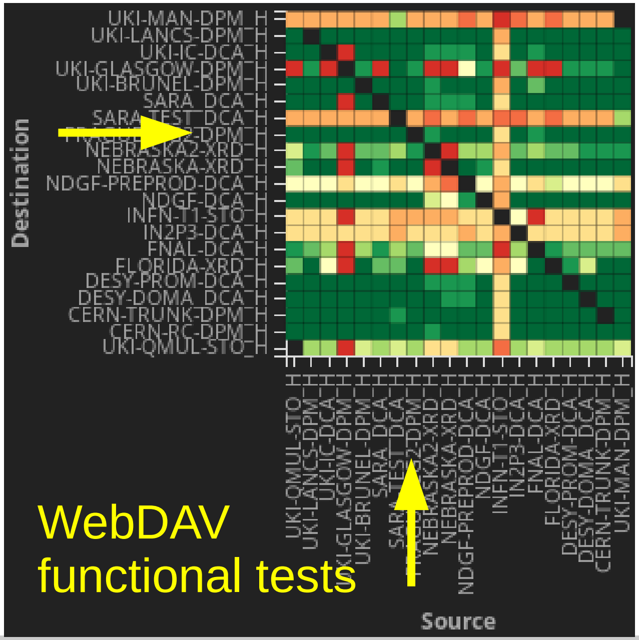
- Good support for non-SRM TPC (WLCG ThirdPartyCopy TF)
  - only supported with DPM DOME configuration
  - X.509 proxy delegation
    - XRootD 4.9, WebDAV with gridsite
  - token based delegation with macaroons
- Testing – functional (smoke-test, matrix), stress
  - details in HOW2019 talk



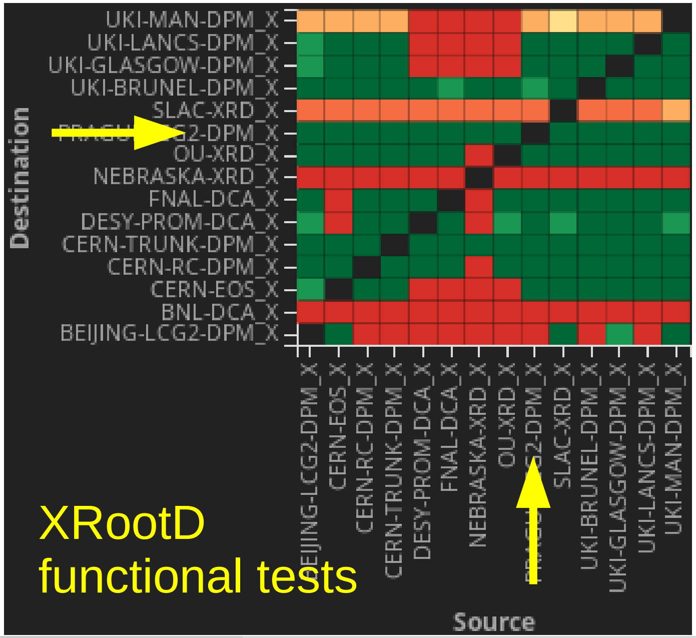
WebDAV
4GB stress
tests (1PB/week)



WebDAV
functional tests



WebDAV
functional tests



XRootD
functional tests

DOME performance

threads→	1	2	4	8	16	32	64	128
read								
SRM	0.6	1.2	2.5	4.6	7.0	8.1	7.9	3.8
gsiftp	2.4	4.7	9.2	17.7	23.2	24.3	24.5	24.8
xrootd	38.77	57.6	81.9	90.3	104.8	111.0	114.3	102.1
webdav	127.4	270.5	513.16	898.5	1236.6	1390.1	1284.2	1304.5
write								
SRM	0.6	1.2	2.3	4.4	7.3	8.1	8.5	3.7
gsiftp	0.6	1.2	2.5	5.1	9.8	19.5	37.3	61.5
xrootd	11.2	27.2	50.6	70.8	76.9	79.3	78.4	81.1
webdav	16.5	34.7	68.0	123.4	169.3	171.4	167.8	175.6
stat								
SRM	2.8	5.4	10.3	20.1	21.9	23.6	24.4	6.7
gsiftp	32.9	64.3	110.6	200.9	220.8	243.1	255.6	263.2
xrootd	84.8	168.8	282.2	454.7	669.1	766.9	913.3	971.8
webdav	206.9	361.0	767.8	1272.8	1762.1	1746.1	1483.3	1535.7
delete								
SRM	1.6	3.3	6.3	12.6	20.5	24.4	24.9	6.7
gsiftp	21.2	42.3	82.8	140.6	153.5	187.5	190.8	205.3
xrootd	38.2	51.7	58.8	65.4	66.6	67.3	69.1	68.9
webdav	51.3	103.1	178.8	311.7	456.8	521.0	519.0	549.0

Data transfers encryption

- In future all transfers will be probably encrypted
  - HTTPS is necessary for TPC
  - XRootD will come with data encryption soon
- Server CPU has build-in support for encryption – AES-NI
  - usually 1 encryption unit per physical core
  - 5Gb/s with single HTTPS connection on low-end modern CPU
    - 16 cores saturate easily 40Gb from mem
    - real file transfers limited by disks
  - 1Gb/s on our oldest storage servers
    - can become quite busy with 10Gb

CPU	openssl	HTTPS one	HTTPS mem	HTTPS disk
2x8core Intel Silver 4108	279.8Gb	4.2Gb	40Gb on 40Gb	30.0Gb disk lim.
2x6core Intel E5-2620	77.7Gb	2.3Gb	10Gb on 10Gb	N/A
2x4core Intel E5620	8.6Gb	0.9Gb	N/A	N/A

Summary

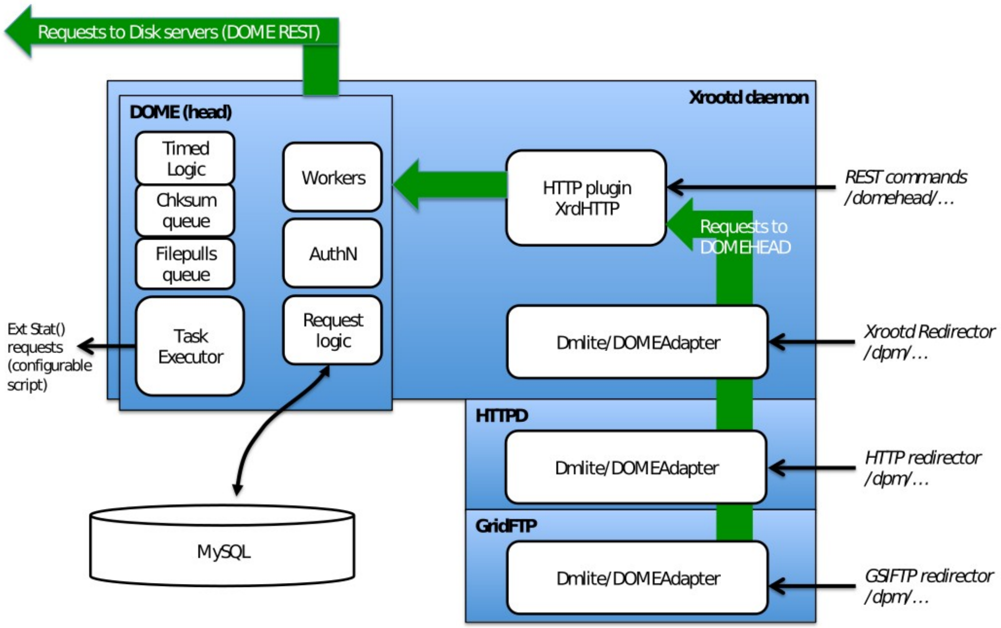
- Active support of legacy DPM ends by June 1st 2019
  - see recommendation provide by WLCG DPM Upgrade Task Force
- New features only in DPM DOME
  - Third Party Copy for non-GridFTP protocols
  - WLCG Storage Resource Reporting
  - substantially improved (metadata) performance
- Missing support for SRM protocol in DOME DPM
  - WLCG experiments can be configured not to use SRM
  - SRM still available through DOME to legacy DPM adapter
  - SRM and DOME transfers in one VO/spacetoken not recommended
- DPM DOME configuration in production at few pioneering sites
  - no critical issue for normal operation
  - few remaining problems (with known workarounds being addressed)
  - reported well described issues gets quickly fixed

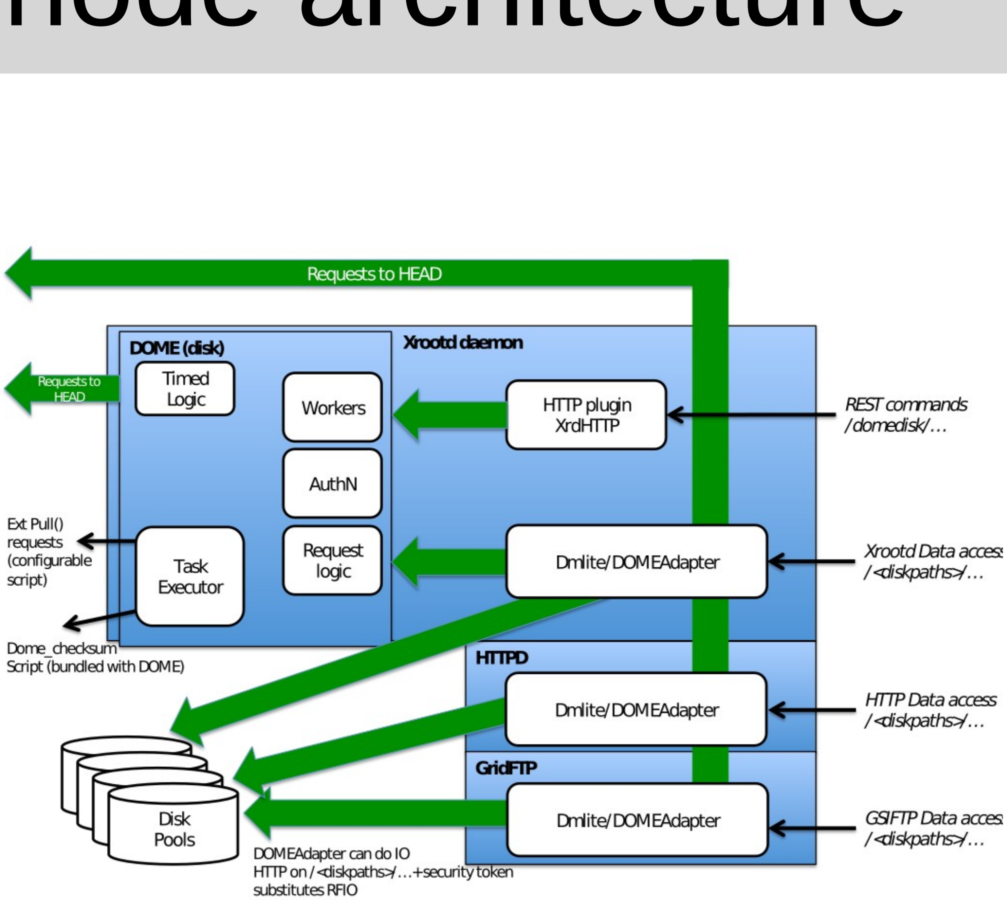
BACKUP

DPM sites > 2PB

Site	Headnode	Size	DPM Version	XRootD protocol	GridFTP
TOKYO-LCG2	lcg-se01.icepp.jp	10559536	DPM/1.9.0-1	0x10030000	GridFTP Server 12.5
ICN-UNAM	tlapiacalli.nucleares.unam.mx	5248671	N/A	N/A	GridFTP Server 11.1
GRIF	node12.datagrid.cea.fr	4554739	DPM/1.10.0-1	0x10030000	GridFTP Server 13.9
UKI-NORTHGRID-MAN-HEP	bohr3226.tier2.hep.manchester.ac.uk	4537796	N/A	0x30000	GridFTP Server 9.1
praguelcg2	golias100.farm.particle.cz	4456037	DPM/1.10.0-1	0x40000	GridFTP Server 13.9
UKI-SCOTGRID-GLASGOW	svr018.gla.scotgrid.ac.uk	3816622	DPM/1.8.10-1	0x10030000	GridFTP Server 12.4
Taiwan-LCG2	f-dpm001.grid.sinica.edu.tw	3269468	DPM/1.8.11-1	N/A	N/A
TR-10-ULAKBIM	torik1.ulakbim.gov.tr	3161019	DPM/1.10.0-1	0x10030000	GridFTP Server 11.3
INDIACMS-TIFR	se01.indiacms.res.in	3107872	DPM/1.9.0-1	0x10030000	N/A
UKI-NORTHGRID-LANCS-HEP	fal-pygrid-30.lancs.ac.uk	3074570	DPM/1.10.0-1	0x10030000	GridFTP Server 13.9
IN2P3-CPPM	marsedpm.in2p3.fr	2642392	DPM/1.9.0-1	0x10030000	GridFTP Server 12.4
INFN-NAPOLI-ATLAS	t2-dpm-01.na.infn.it	2399002	DPM/1.10.0-1	0x10030000	GridFTP Server 13.9
RO-07-NIPNE	tbit00.nipne.ro	2367783	N/A	0x30000	GridFTP Server 7.26
NIKHEF-ELPROD	tbn18.nikhef.nl	2353756	DPM/1.9.0-1	0x10030000	N/A
IN2P3-LAPP	lapp-se01.in2p3.fr	2187150	DPM/1.9.0-1	0x10030000	GridFTP Server 12.5
GRIF	lpnse1.in2p3.fr	2113058	DPM/1.10.0-1	0x10030000	N/A
INFN-FRASCATI	atlasse.lnf.infn.it	2111236	N/A	0x10030000	GridFTP Server 11.8
IN2P3-IRES	sbgse1.in2p3.fr	2032105	DPM/1.10.0-1	0x10030000	GridFTP Server 13.9

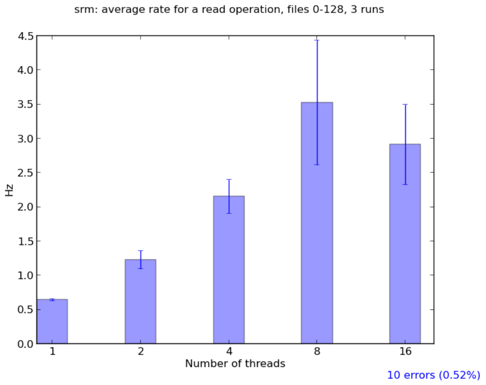
DOME head / disk node architecture

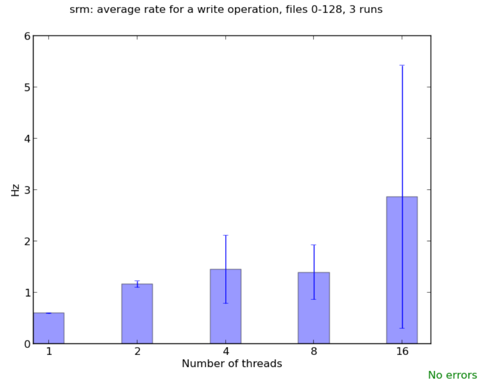


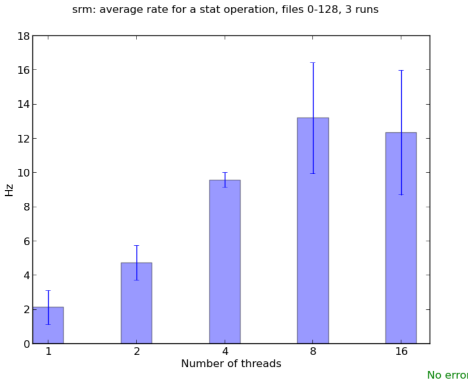

Requests to HEAD
Requests to HEAD
DOME (disk)
Xrootd daemon
Timed Logic
Workers
HTTP plugin XrdHTTP
REST commands /domedisk/…
AuthN
Request logic
Ext Pull() requests (configurable script)
Task Executor
Dmlite/DOMEAdapter
Xrootd Data acces /<diskpaths>/…
Dome_checksum Script (bundled with DOME)
HTTPD
Dmlite/DOMEAdapter
HTTP Data access /<diskpaths>/…
GridFTP
Dmlite/DOMEAdapter
GSIFTP Data acces /<diskpaths>/…
Disk Pools
DOMEAdapter can do IO HTTP on /<diskpaths>/…+security token substitutes RFIO

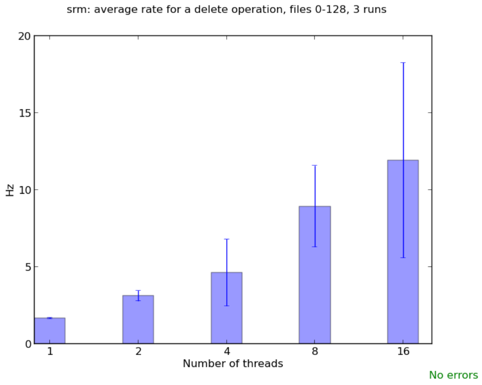
SRM vs. GridFTP performance

- DOME DPM 1.11.1, XRootD 4.9, dpm-dsi 1.9.15 with redirection
- SRM performance – write, read, stat, delete

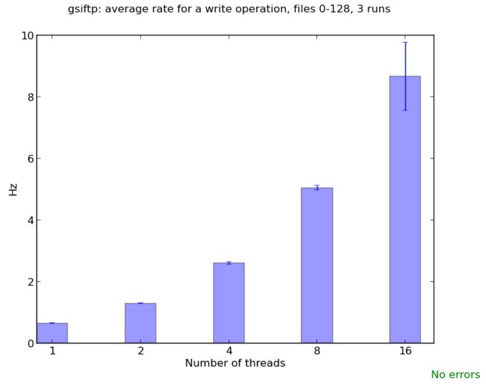


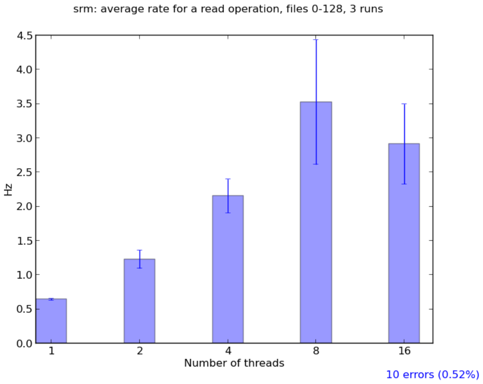


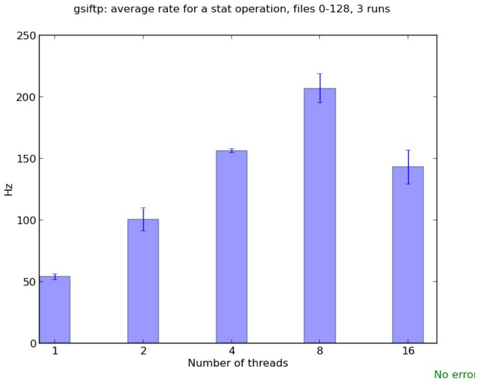


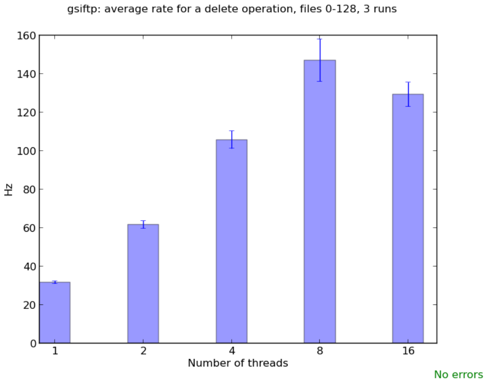


- GridFTP performance – write, read, stat, delete



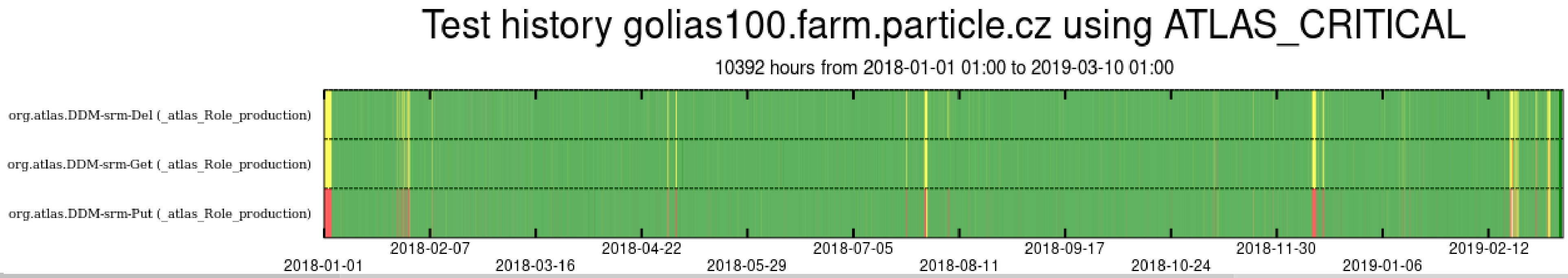






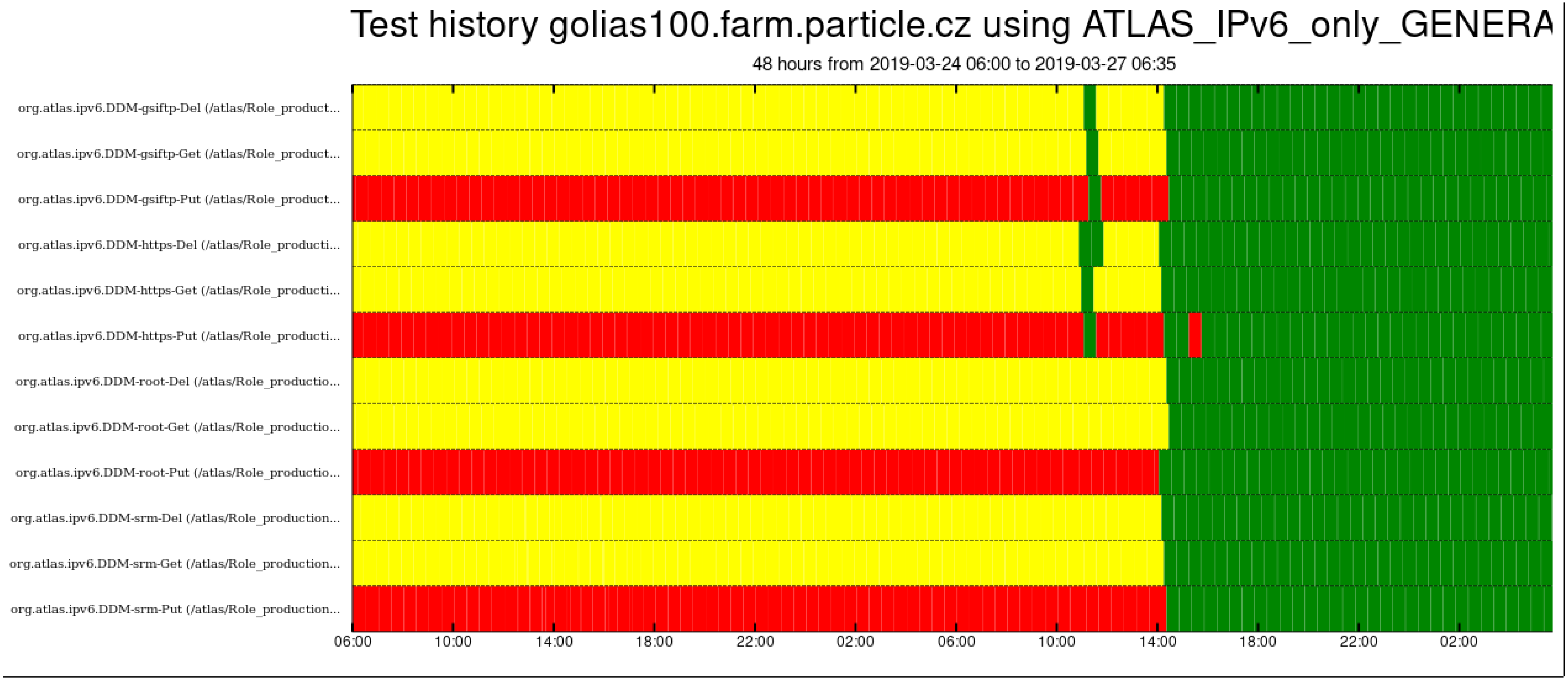
Issues – failing SAM SE tests

- 2018
  - January 1st – UPS / power failure
  - January 26-30 – performance issues (non-local DNS queries)
  - May 11, 21 – troubles with one diskserver
  - June 28 – DPM 1.10.3 upgrade
  - July 23 – DPM reconfigured to DOME mode
  - July 29-31 – enabled/disabled GridFTP redirection
  - August 11 – spacetoken synchronization issues
  - December – troubleshooting core switch / network problems
- 2019
  - January 30 – new core switch
  - February 10 – DPM DOME upgrade to 1.11.1
  - February 17 – enabled GridFTP redirection

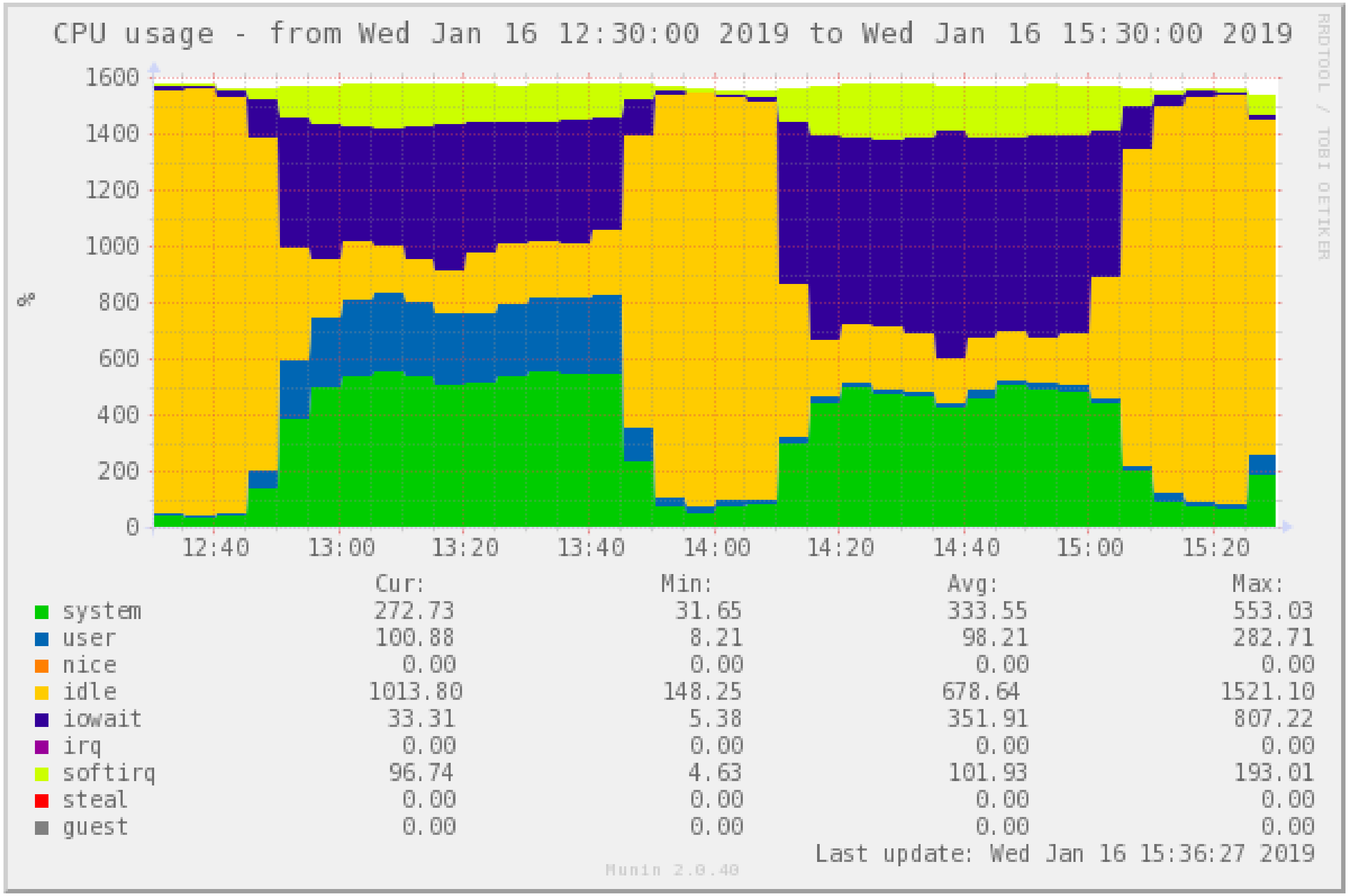


SAM IPv6only storage monitoring

- Monitoring tests available in “dev” branch
  - storage works fine
  - preproduction monitoring issue



AES-NI cpu utilization



CPU usage - from Wed Jan 16 12:30:00 2019 to Wed Jan 16 15:30:00 2019

	Cur:	Min:	Avg:	Max:
🟩 system	272.73	31.65	333.55	553.03
🟦 user	100.88	8.21	98.21	282.71
🟧 nice	0.00	0.00	0.00	0.00
🟨 idle	1013.80	148.25	678.64	1521.10
🟪 iowait	33.31	5.38	351.91	807.22
🟪 irq	0.00	0.00	0.00	0.00
🟩 softirq	96.74	4.63	101.93	193.01
🟥 steal	0.00	0.00	0.00	0.00
⬛ guest	0.00	0.00	0.00	0.00

Last update: Wed Jan 16 15:36:27 2019

- CPU utilization while reading 1GB files from disk and sending them with average speed ~ 30Gb/s encrypted with TLSv1.2,ECDHE-RSA-AES256-GCM-SHA384,2048,256