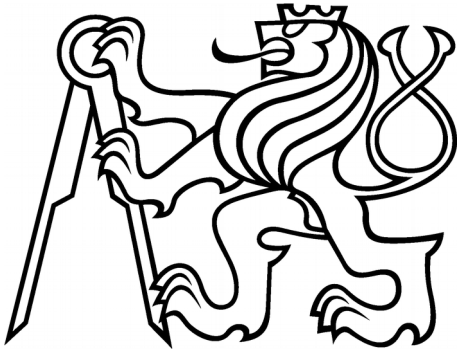


DPM @ Prague

Petr Vokáč

DPM Workshop 2019

13th June 2019

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

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Prague DPMs

- prague_cesnet_lcg2 (CESNET)
  - dpm1.egee.cesnet.cz (DOME enabled, legacy supported)
  - DPM 1.10, dmlite 1.10.4
  - Auger (60TB), Belle (75TB)
- praguelcg2 (FZU – Institue of Physics of the Czech Academy of Sciences)
  - dpmhead.farm.particle.cz
    - testbed (headnode, disknodes, dbnode)
    - various versions & features enabled
  - golias100.farm.particle.cz (DOME enabled, legacy supported)
    - DPM 1.12.0, dmlite 1.12.1 (20190315), lcgdm-dav 0.22.0, dpm-dsi 1.9.15, xrootd 4.9.1
      - just a base release with additional patches (bugfixes, features)
    - ATLAS (2.7PB), DUNE (300TB), Auger (250TB), dteam (10TB)
  - other
    - XRootD (1.6PB for ALICE), NFS (200TB for local users)

Hardware

- Headnode – kvm (6 cores E5-2650v4, 16GB RAM, 100GB disk!)
- DB machine – 2x Intel E5-2630v4, 128GB RAM, 1TB SSD RAID10)
- Disknodes (controllers – Areca 1680, MageRAID 2208 or 3108)

disknode	year	cpu / mem [GB]	os	disk [TB]	net [Gb]
dpmpool1	2011 (2019 hdd)	2x E5620 / 48	CentOS7	455	10
dpmpool9	2011 (2015 hdd)	2x E5620 / 48	CentOS7	210	10
dpmpool11	2011	2x E5620 / 12	SLC6	125	10
dpmpool13	2012	2x E5620 / 24	SLC6	180	10
dpmpool14	2012	2x E5620 / 24	SLC6	310	10
dpmpool15	2012	2x E5620 / 24	SLC6	75	10
dpmpool16	2013	2x E5-2620v2 / 48	SLC6	435	10
dpmpool17	2013	2x E5-2620v2 / 48	SLC6	435	10
dpmpool18	2014	2x E5-2620v2 / 48	SLC6	420	10
dpmpool19	2014	2x E5-2620v2 / 48	SLC6	420	10
dpmpool20	2018	2x Silver 4108 / 96	SLC6	435	4x10
dpmpool21	2018	2x Silver 4108 / 96	SLC6	435	4x10
dpmpool22	2018	2x Silver 4108 / 96	CentOS7	435	4x10
dpmpool23	2018	2x Silver 4108 / 96	CentOS7	435	4x10
dpmpool24	2018	2x Silver 4108 / 96	CentOS7	435	4x10
sum				**5240**	**300**

Configuration

- pools
  - augerpool1 – Auger
  - heppool1 – all experiments except Auger
    - sharing pool → more servers / filesystems → improved performance
  - fragilepool (QOS)
    - old / new machines with uncertain reliability
    - data that we can loose at any time without troubles (Rucio secondary)
- ~ 100 filesystems on 15 servers
- DOME mode enabled, but legacy still configured for SRM access
  - trying to reduce SRM usage not just because legacy EOL
  - using recommended configuration for ATLAS
    - GridFTP redirection enabled
    - XRootD checksums enabled
    - TPC delegation enabled

Availability

- 2018

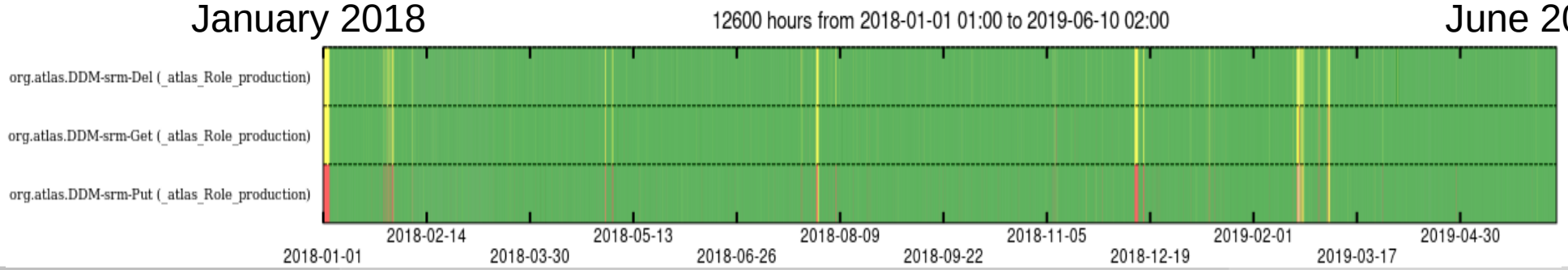
ATLAS SAM monitoring

  - 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

DPM Workshop 2018
Prague, May 30 – June 1

- 2019
  - January 30 – new core switch
  - February 10 – DPM DOME upgrade to 1.11.1 (thread deadlocks)
  - February 17 – enabled GridFTP redirection, problem with one diskserver
  - March 5 – DPM DOME upgrade to 1.12
  - Jun 1 – IPv6 firewall incorrectly configured for new diskserver

January 2018



June 2019

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

Performance markers

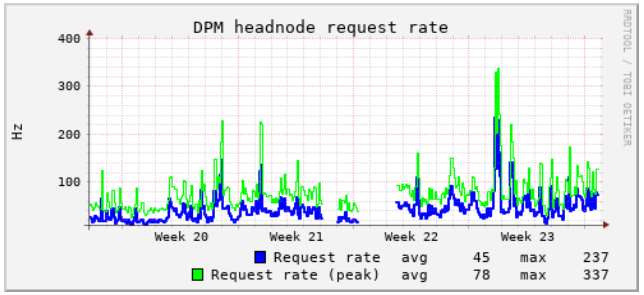
- Since 1.12 dmlite.log provides interal performance markers
- Analyzed data for last month
  - DPM headnode (golias100.farm.particle.cz)
  - DPM disknode (dpmpool22.farm.particle.cz)

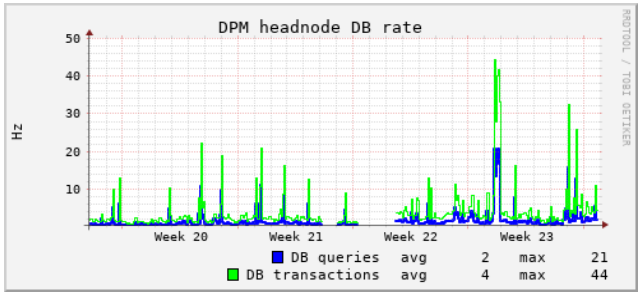
request rate
request peak

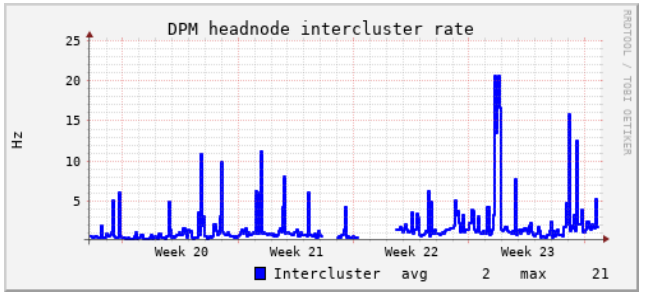
DB queries
DB ransactions

Intercluster

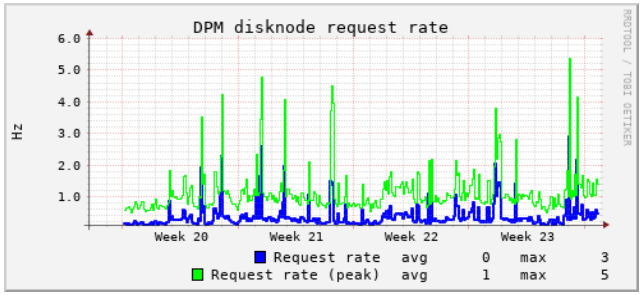
headnode

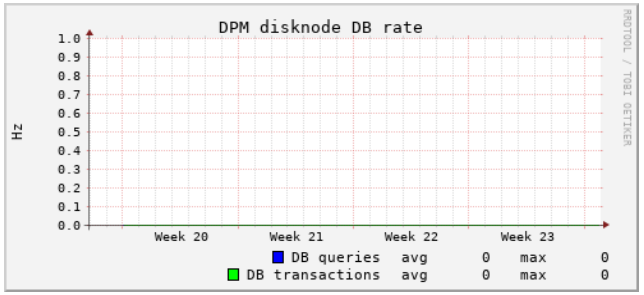


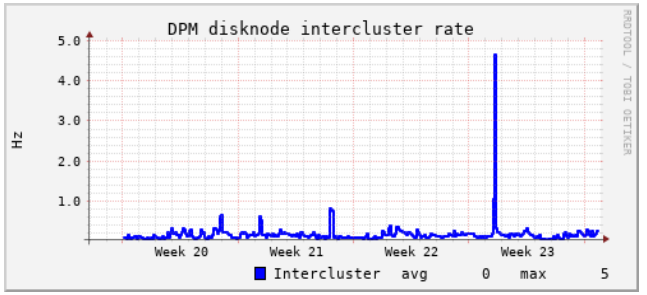




disknode







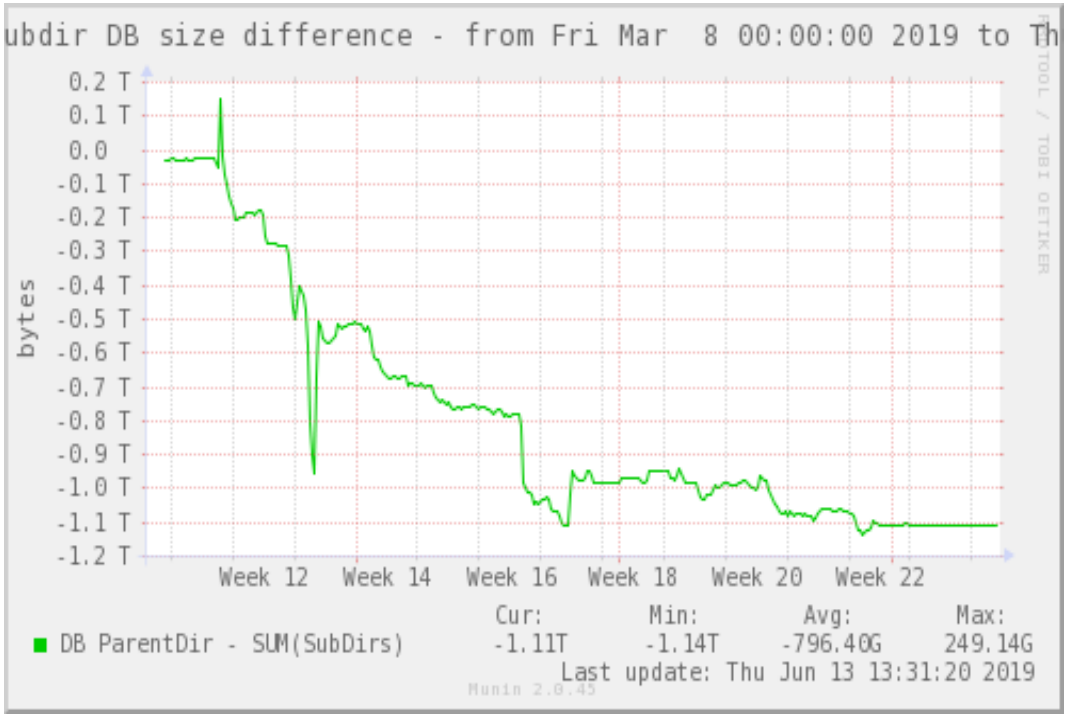
Transfers encryption

- 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

Space counter divergence

- Spacetokens (SRM) vs. quotatoken (DOME)
- 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
  - seems to be fixed with latest DPM update
  - DPM 1.13(?) as part of dpm-dsi
- Still very small chance to diverge with GridFTP
- dmlite-mysql-dirspaces.py
  - can fix ST/QT discrepancy
- Physical disk space issues with SRM transfers
  - fixed in 1.12



GridFTP redirection

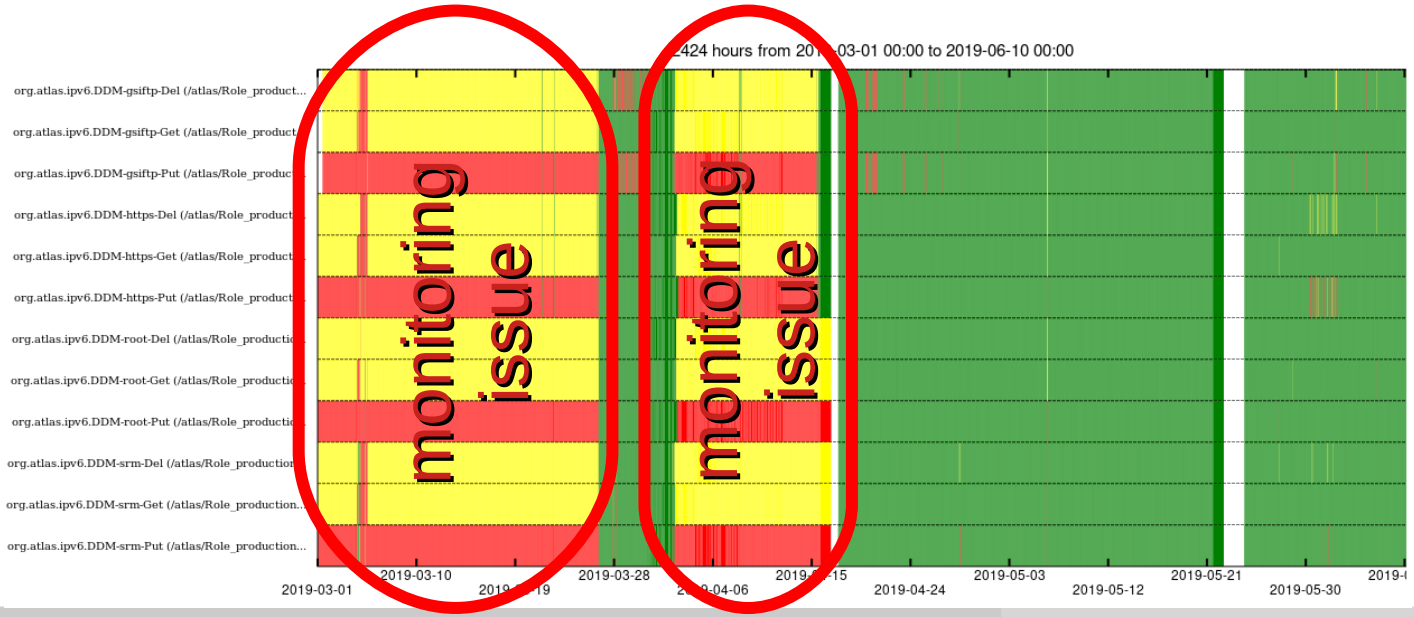
- GridFTP still used exclusively for TPC
- necessary for GridFTP without SRM
  - prevents data tunneling through headnode
  - LCGDM-2748 can cause troubles in case one of diskserver is down
    - workaround – manual removal from gridftp.conf
    - some progres / fixed (?) in dmlite 1.13
  - SRM performance lower ~ 20% with this configuration
    - only metadata operations that mostly affect small file transfers
- it is possible to ***loose data*** with DOME DPM & GridFTP redirection
  - problem in globus protocol
  - transfer is confirmed before DPM gets chance to update metadata
    - FTS / Rucio thinks everything succeeded (update catalog)
    - DPM metadata shows transfer is still in progress
    - in case DPM fails to commit metadata operation → lost data
      - happened for our gsiftp transfers between two pools (lost ~ 300 files)

HTTP / apache reload issues

- Apache has to be reloaded (graceful restart)
  - CRL update for HTTPS (some expire in 3 days)
    - normally every 6 hours
    - necessary to support davs – only headnode
    - necessary for TPC on head/disk nodes
  - logrotate – weekly
- Apache can get stuck during graceful restart LCGDM-2699
  - default configuration till dmlite 1.12 not optimal
    - ServerStart = ServerLimit → graceful restart waits
  - slow transfer of big files can take long time
    - old apache process waits
    - sometimes transfers can get completely stuck
      - seems to be fixed in lcgdm-dav 0.22 module
- Restart is fine but kill transfers in progress
- Graceful restart still kills downloads (fixed http 2.4.35 - RHEL8)

IPv6

- Dualstack for years
  - not all protocols were properly configured (not used / monitored)
  - mixed mode with partially enabled IPv6 not supported by DPM
    - client address is part of authentication token used during HTTP redir.
    - GridFTP with redirection doesn't work for IPv4 clients when IPv6 enabled in GFAL (Dirac enforce IPv6 everywhere)
- ATLAS started to monitor IPv6 with SAM tests (March 2019)
  - ATLAS_IPv6_only_GENERAL template
  - still only development version – monitoring issues



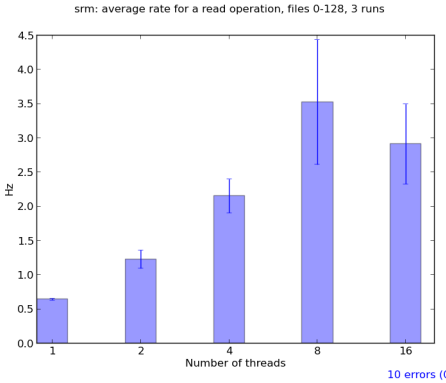
Releases

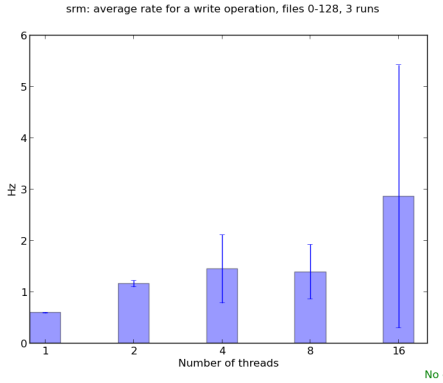
- Slow
  - Most of time I’m running custom build on our production system
  - I would like (important) bugfixes faster in production
  - Also some features should go in production faster
    - cleanup HTTP TPC
    - killing stuck transfers
- Looking forward for 1.13 DPM and dmlite
  - only problem with possibility to lose files with GridFTP redirection
  - problem with GridFTP from IPv4 only clients on dualstack DPM
    - by default enabled by Dirac (hardcoded in sources)
    - it’ll be also enabled by default in next gfal version

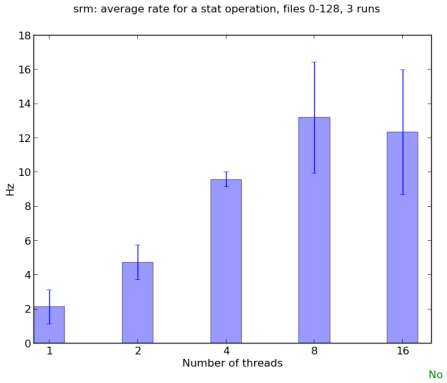
BACKUP

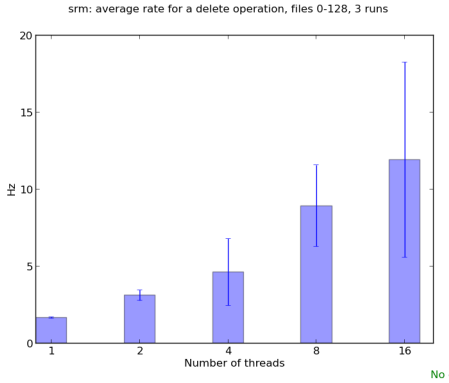
SRM vs. GridFTP performance

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

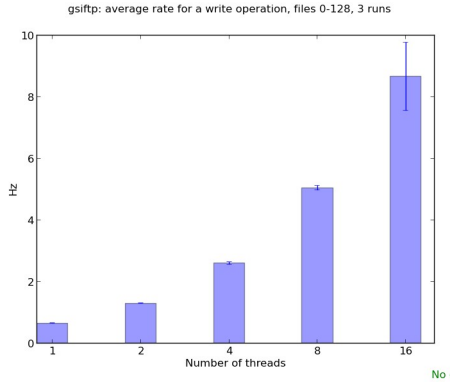


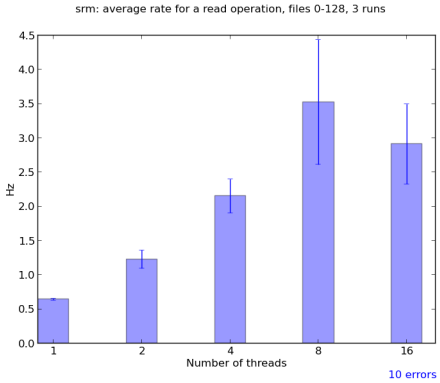


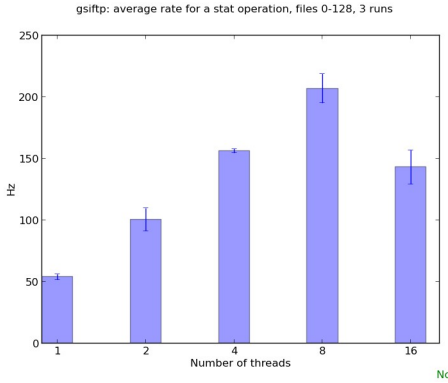


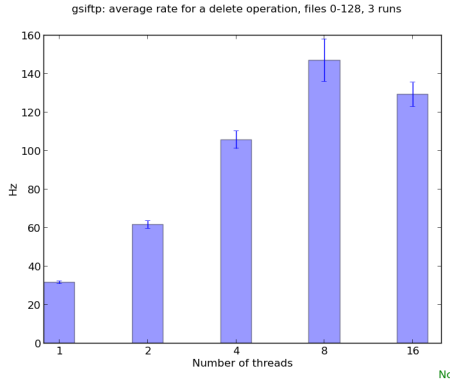


- GridFTP performance – write, read, stat, delete

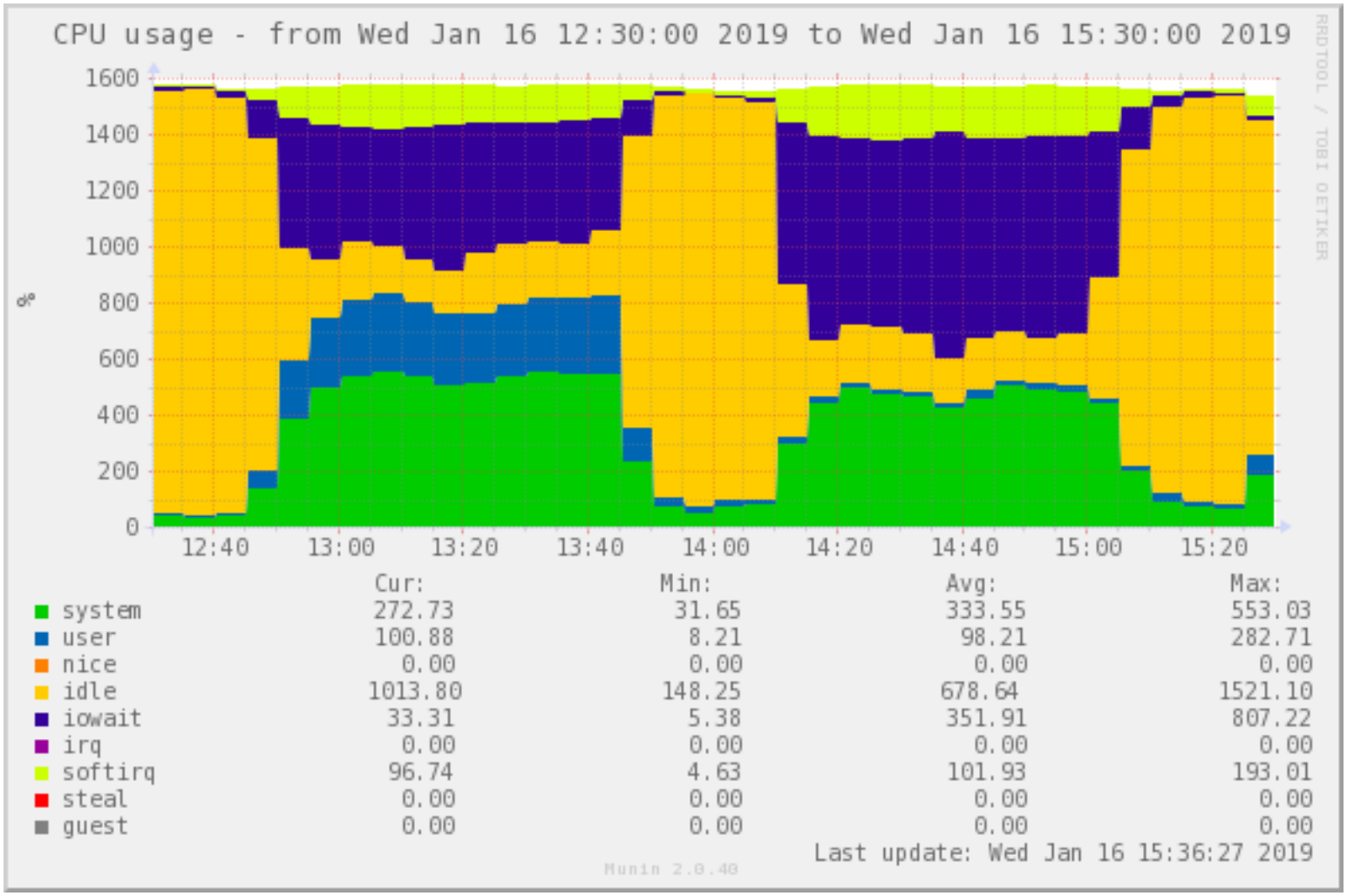








AES-NI cpu utilization



- 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