Karlsruhe Institute of Technology

GPFS Setup at SCC SDM

Uwe Falke

Environment

Karlsruhe Institute for Technology (KIT)

- >9000 staff
- (>5000 scientists)
- >22000 students
- >budget 1Bill EUR

Steinbuch Centre for Computing (SCC)

IT infrastructure
HPC

Data Storage:
- GridKa Storage: LHC Tier 1
- LSDF (Large Scale Data Facility): scientific data incl HPC storage

Some numbers

	GridKa	LSDF	Total
net storage capacity [PiB] ([PB])	43.8 (49.3)	15.2 (17.1)	59.0 (66.4)
# of NSD server nodes	32	14	46
# of NSD client nodes	approx. 65	approx. 850	approx. 900
# of files (inodes)	4.13×10^8	1.453×10^9	1.866×10^9
used storage [PiB] ([PB])	35.8 (40.3)	10.5 (11.8)	46.2 (52.1)
avg. file size [MiB] ([MB])	93.1 (97.6)	7.8 (8.1)	

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

Technology

- Traditional setup:

  x Storage Systems + 2 NSD servers = 1 Storage Unit
- GridKa: IB for RDMA only, Eth GPFS Admin network; LSDF: ipoib
- Storage backend via SAS
- Netapp E5600, E5700, E2800; Seagate EXOS X5U84, (Seagate Corvault, not yet)
- Disk Pooling (Distributed Arrays)
- Data on NL-SAS-HDD
- Metadata on SAS-SSD, (NVMe SSD Shared-Nothing, not yet)

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

Technology

- LSDF: 4 x E5700, 5 x E5600, 10 x Seagate 6575 (EXOS X5U84)
- GridKa: 4 x E5700, 13 x E5600, 2 x E2800
- 2 Separate IB Fabrics in each of LSDF and GridKa
- Server: SuperMicro

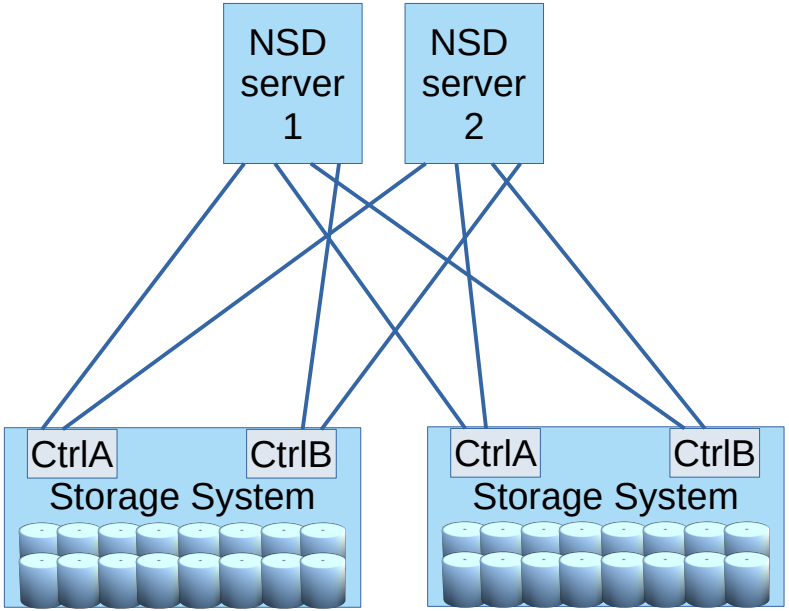
Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

Storage Unit: GridKa I, LSDF I (Netapp E5x00)



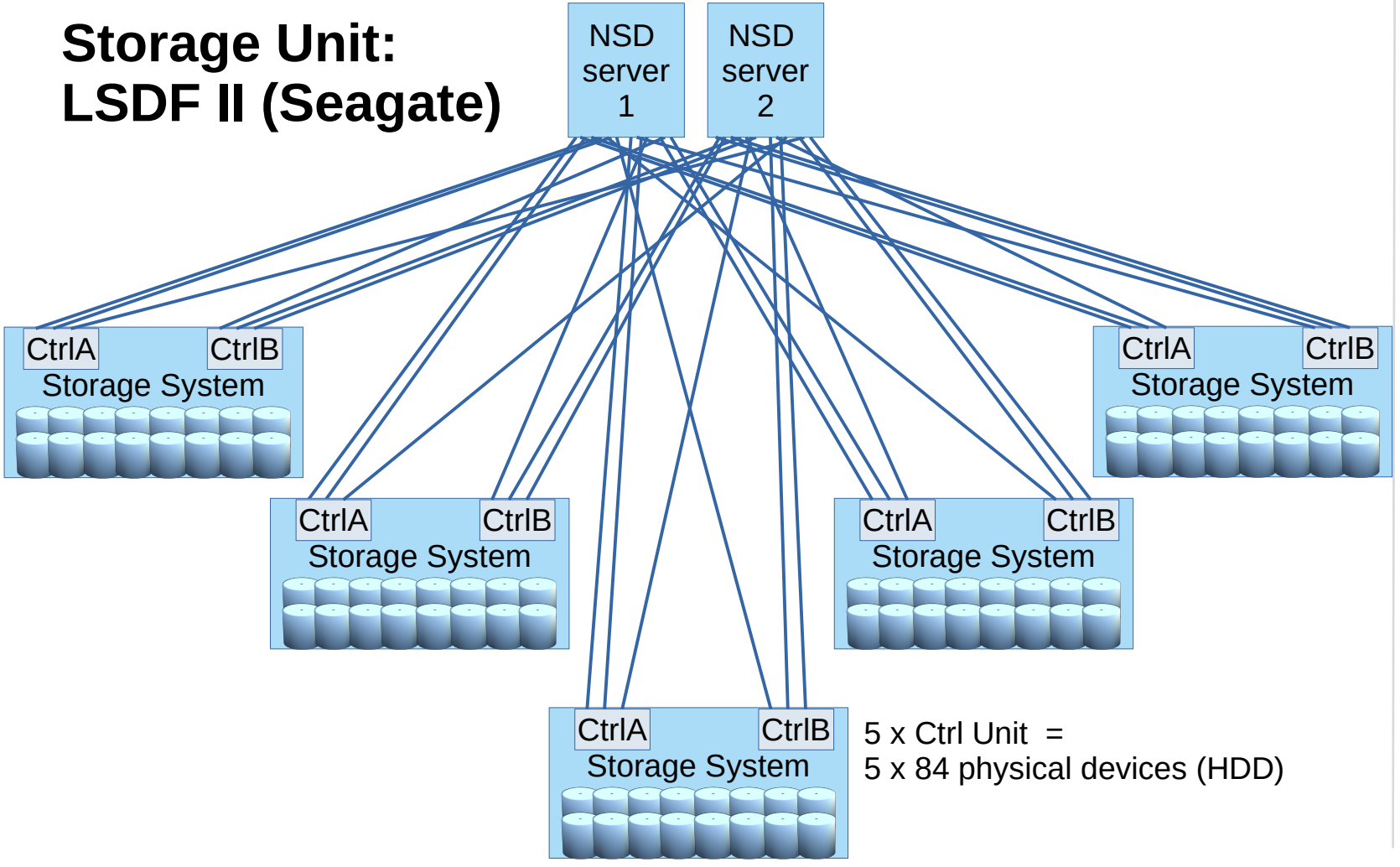
2 x (Ctrl Unit + 4 Expansions) =
2 x 300 physical devices (HDD, partly SSD in LSDF for metadata)

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

Storage Unit: LSDF II (Seagate)



5 x Ctrl Unit =
5 x 84 physical devices (HDD)

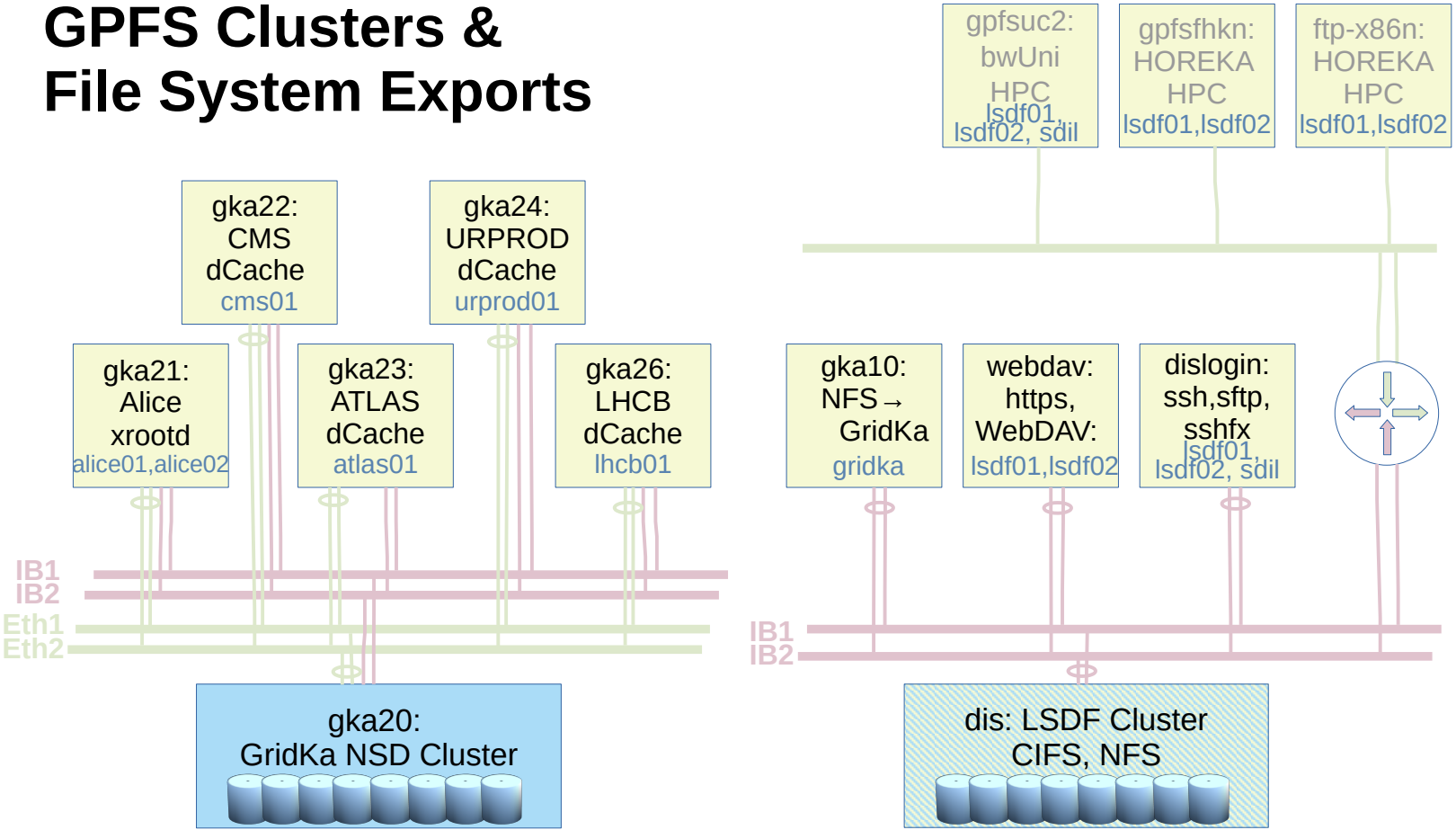
Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS Clusters & File System Exports

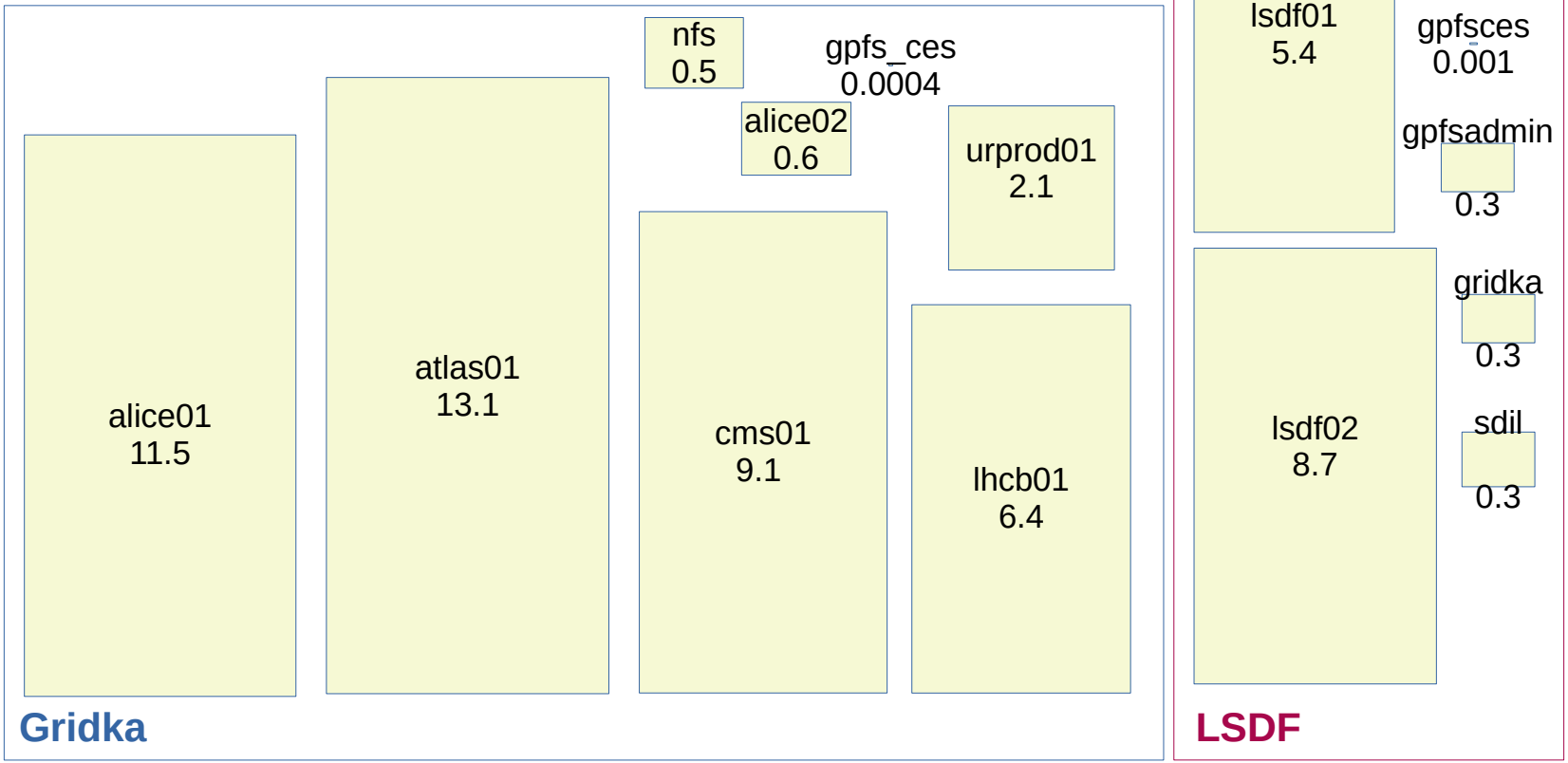


13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS File Systems



13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS File System Configuration (GridKa)

General:

Inode size :	4kiB
Indirect Block Size:	32kiB
File locking semantics	nfsv4
DMAPI	no
Exact mtime mount option	yes

	alice01	alice02	atlas01	cms01	gpfs_ces	GridkaTest	lhcb01	nfs01	urprod01
BlSz (system) [kiB]	256	256	512	512	1024	4096	4096	512	512
BlSz (data) [kiB]	4096	4096	4096	4096	n/a	n/a	n/a	4096	4096
Quotas enforced	u;g;fset	-	-	u;g;fset	-	u;g;fset	u;g;fset	u;g;fset	u;g;fset
block allocation	sc	sc	sc	cl	sc	sc	sc	sc	sc
# NSDs	150	17	162	111	8	8	84	17	33

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS File System Configuration (LSDF)

General:

Inode size :	4kiB
Indirect Block Size:	32kiB
File locking semantics	nfsv4
block allocation	scatter
Quotas enforced	u;g;fset
Exact mtime mount option	yes

	gpfsadmin	gpfsces	gpfstest2	gridka	lsdf01	lsdf02	sdil
BlSz (system) [kiB]	256	512	4096	256	256	256	256
BlSz (other) [kiB]	2048	n/a	n/a	4096	4096	4096	4096
Indirect Block Size [kiB]	32	16	32	32	32	32	32
DMAPI	y	n	y	n	n	n	n
# NSDs	8	2	2	6	108	129	8

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS File Systems: More Data

	size / PiB	used / PiB	# files / 10^6	avg f.size/ MiB	# filesets	# snapshots
gpfsadmin	0.31	$2x10^{-2}$	8.5	3.0	14	36
gpfsces	$8x10^{-4}$	$4x10^{-6}$	$4x10^{-3}$	1.1	0	0
gpfstest2	0.07	$3x10^{-6}$	$2x10^{-2}$	0.2	0	0
gridka	0.31	$1x10^{-2}$	9.1	1.0	2	10
lsdf01	5.41	3.96	831.9	5.1	62	351
lsdf02	8.74	6.42	545.3	12.6	12	78
sdil	0.31	0.06	59.9	1.1	3	40
GridkaTest	0.5	0.1	1.1	110.0	1	0
alice01	11.50	10.60	228.5	49.8	0	0
alice02	0.59	$1x10^{-4}$	0.9	0.1	0	0
atlas01	13.12	11.46	33.3	369.2	0	0
cms01	9.10	7.64	22.1	370.4	1	0
gpfs_ces	$4x10^{-2}$	$4x10^{-5}$	$9x10^{-3}$	4.8	0	0
lhcb01	6.39	3.99	5.2	817.7	0	0
nfs01	0.51	0.14	76.7	2.0	9	11
urprod01	2.13	1.80	45.1	42.9	2	0

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS Backup

- Only in LSDF
- Up to now: mmbackup to TSM
- Future: Move to GHI/HPSS (not a native/true backup though), currently underway - Scale Out Backup and Restore (SOBAR)

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS Installation / Updates

- Satellite + Puppet control Linux Kernel, OFED, and GPFS (Beware of unintended changes due to host group inheritance!)
- dCache Pools require downtime to upgrade NSD clients.
- Upgrade of IO cluster theoretically without downtime, however, experience tells that is not going smoothly.
- New nodes join existing Spectrum Scale clusters automatically
- Reinstalled nodes recover Spectrum Scale config automatically
- Nodes join existing CES clusters automatically

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS in Daily Operation

- If it runs smoothly, nobody takes notice :-)
- If it fails, that is often caused by infrastructural issues (connectivity, sudden node reboots)
- Failures not caused outside GPFS are infrequent but the more mysterious in many cases (we do not run GPFS traces permanently ...)
- Usually, GPFS does not harm data even if affected from failures of the environment.
- However: That may happen … GPFS may eat your data ...

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS May Eat Your Data (Recent Issue) ...

- … No, not regularily :-)
- Recent issue : RDMA problem due to a misbehaving OFED (IB HCA driver).
- Detected due to the file consistency checks of dCache
- A number of files found to have an invalid size (all the same, 3968 Byte)
- All these files were written when GPFS saw an assert due to some unexpected return code from RDMA operations and restarted.
- The unexpected code was due to a bug in the OFED driver.
- However, GPFS must be blamed for not throwing an EIO but completing a file write with invalid data (IBM Case TS009847927)

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

Other issues around affecting GPFS

- E5600 Storage Controller lockdown (replacing a defective HDD with a used one)
- Network issues (the usual thing)
- Unexpected Node restarts

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS and Applications (a general remark)

- GPFS has no guaranteed response time - Full Stop
- While GPFS is praised and used for its good performance there may be conditions which ground its operation to a (temporary) halt.
- Any application which is not aware of that might take that for a failure and fail itself - for actually no good reason.
- Reasons for pausing GPFS are most often unexpected connection losses to nodes (either due to network issues or due to instable node operation) - lease wait times apply

Karlsruhe Institute of Technology

13. Jul 2022

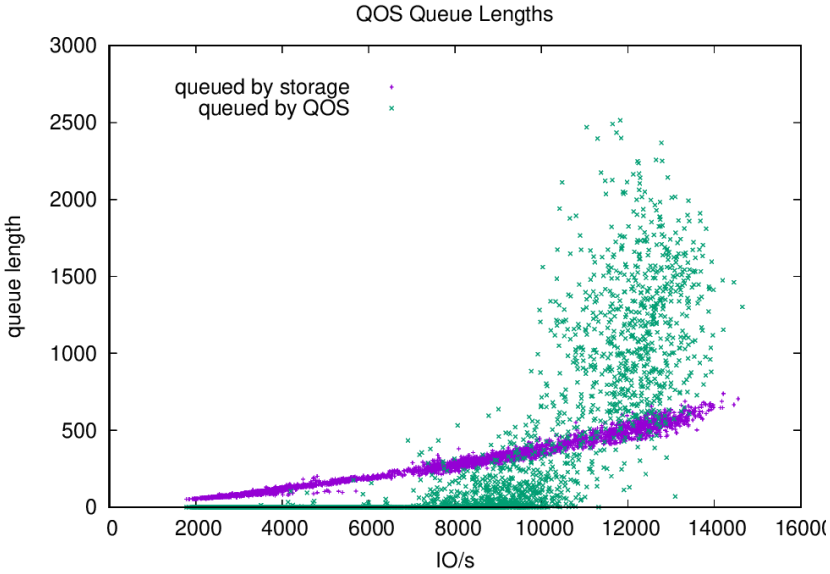
Uwe Falke

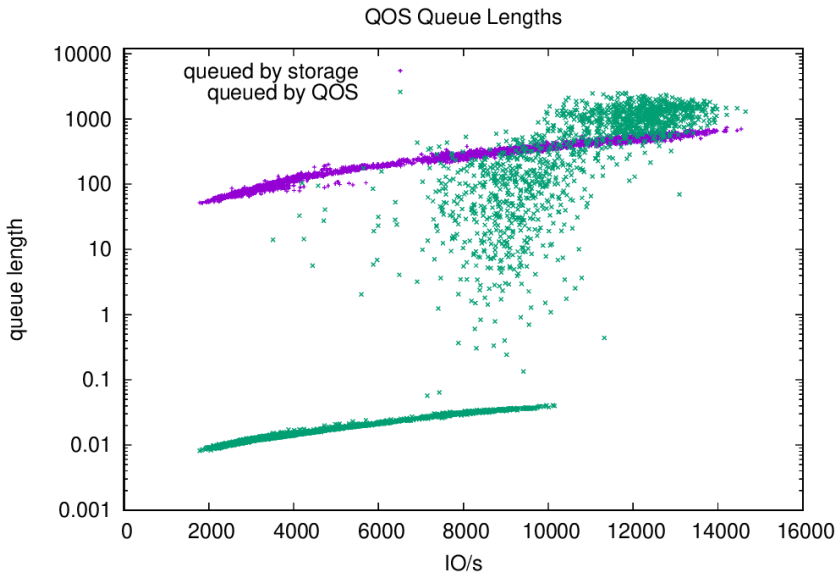
Steinbuch Centre for Computing (SCC)

GPFS QOS

alice01 QOS limits: pool=data,other=70000Iops,maintenance/all_local=inf

data from mmlsqos





13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS and Storage

- Traditional: Controllers in between mmfsd and storage device, additional lag in data transfer
- GPFS GNR (as in IBM ESS, Lenovo DSS): - Distributed Arrays (“Declustered“, DA), direct storage access, but limited to certain HW, relatively short support periods (at least for ESS)
- ECE (Spectrum Scale Erasure Code Edition, Mestor): similar to but more flexible than GNR, BYOH ! → to be considered.
- FPO / Shared Nothing ?

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

NVMe Storage (Shared Nothing)

- NSDs built on node-local storage, here: NVMe devices
- considered for Metadata Storage
- Only safe with 3(4?)-way replication, but -M=2 cannot be increased for an existing file system → FS recreation / data migration is required.
- No redundancy for NSD access: each restart of GPFS on an NVMe NSD node in normal operation requires restarting the disks

Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

Summary

- GPFS forms the base of SDM data storage
- Several Applications / Protocols reside on top
- GPFS is generally reliable
- Applications may benefit from observing GPFS particuliarities
- GPFS perfomance has never been an issue (provided the backing storage sustains it)
- Traditional Storage Setup in Use, new ways to explore (Shared Nothing, ECE)
- IB RDMA (LSDF, GridKa) and IPoIB (LSDF) run generally well (the recent RDMA issue nonwithstanding)

Karlsruhe Institute of Technology

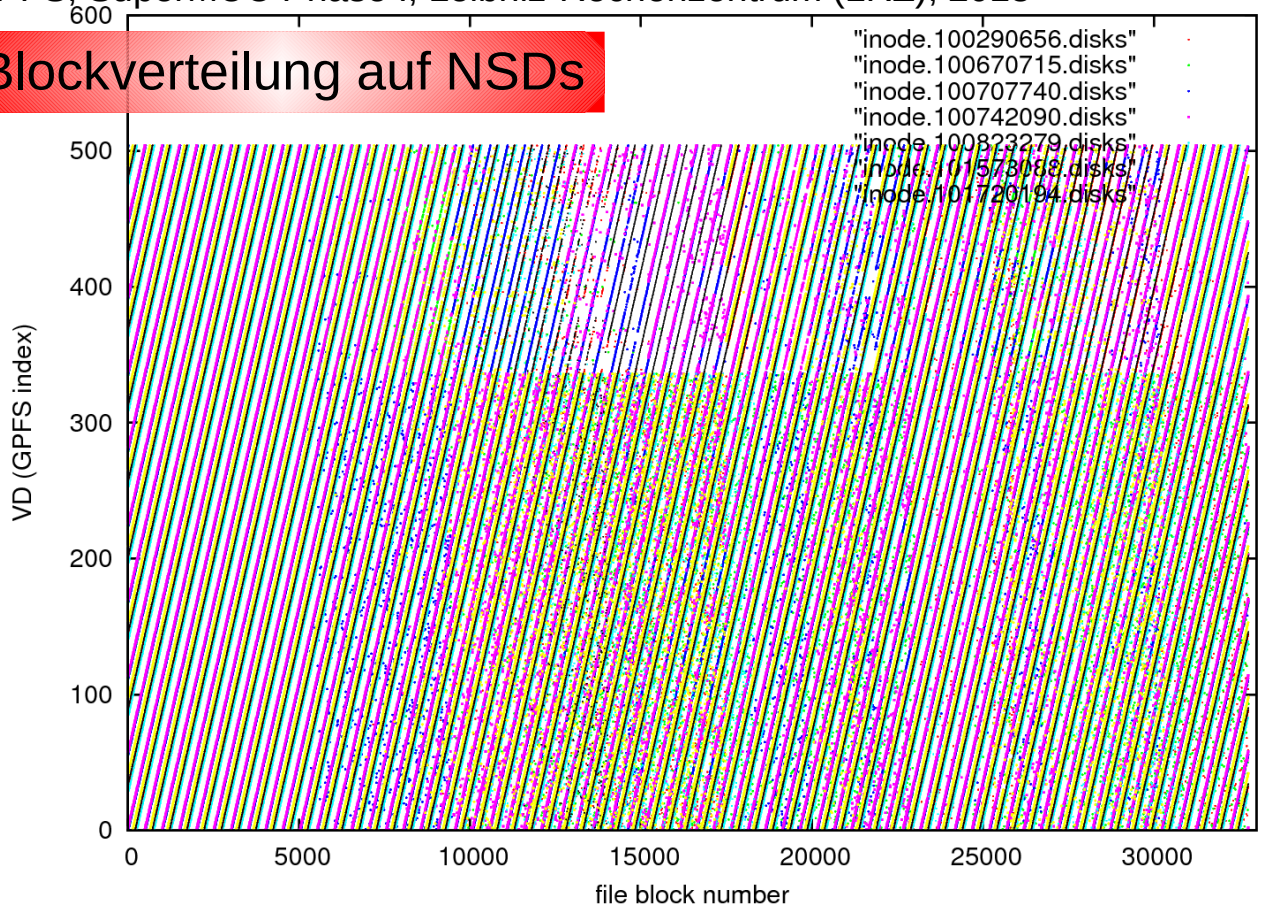
13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

GPFS, SuperMUC Phase I, Leibniz-Rechenzentrum (LRZ), 2013

Blockverteilung auf NSDs



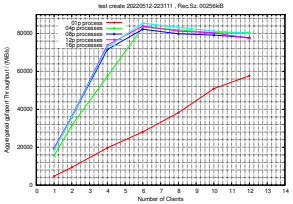
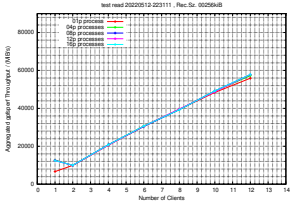
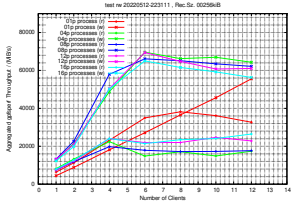
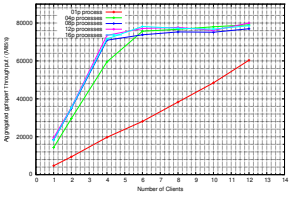
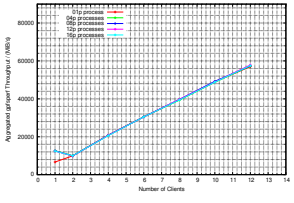
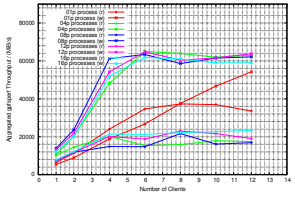
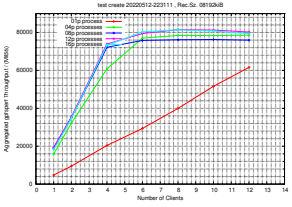
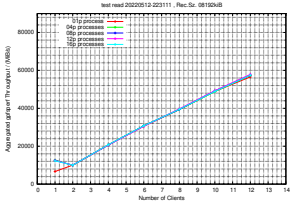
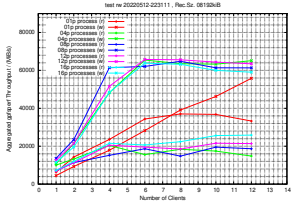
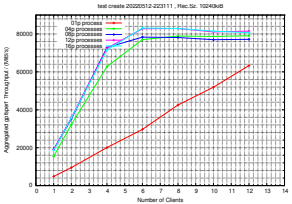
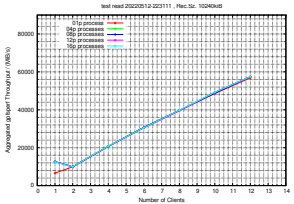
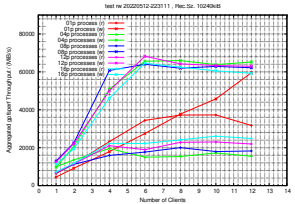
Karlsruhe Institute of Technology

13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

2022-05-12

	create	read	rw
256kiB			
2048kiB			
8192kiB			
10240kiB			

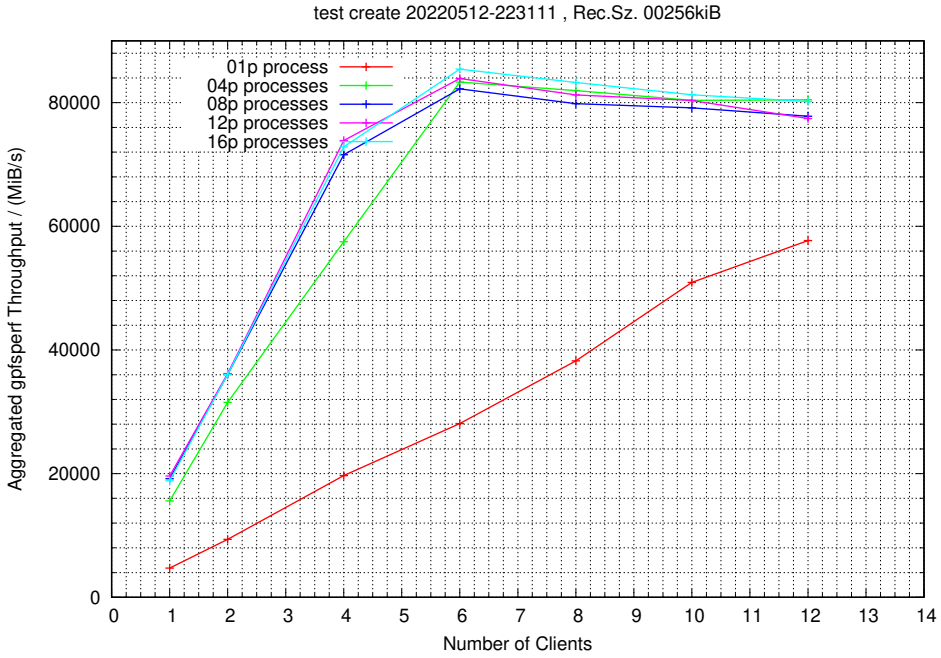
Karlsruhe Institute of Technology

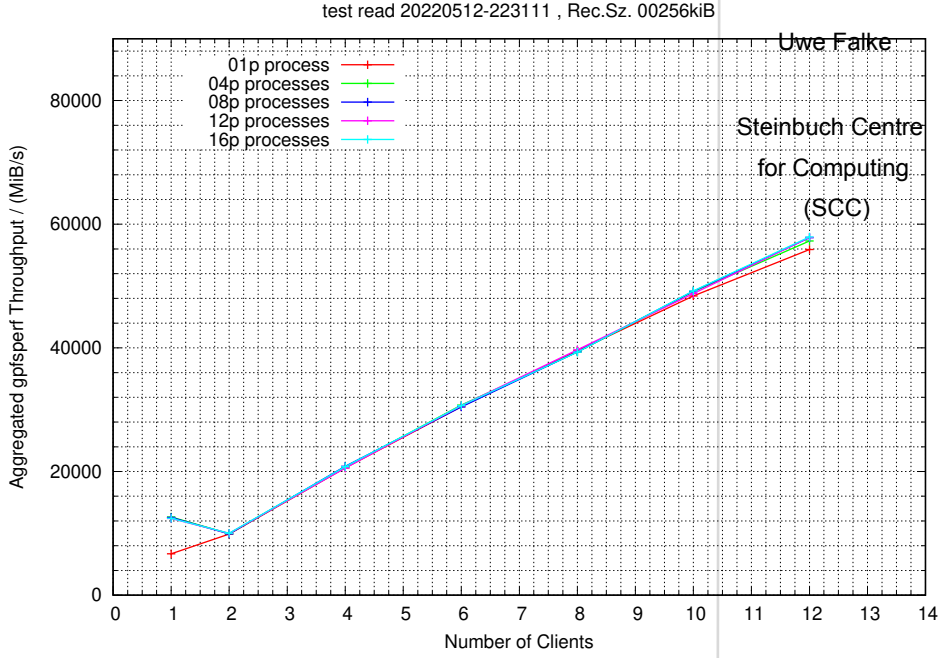
13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

2022-05-12, 256kiB IO Sz, Create (left) + Read (right)

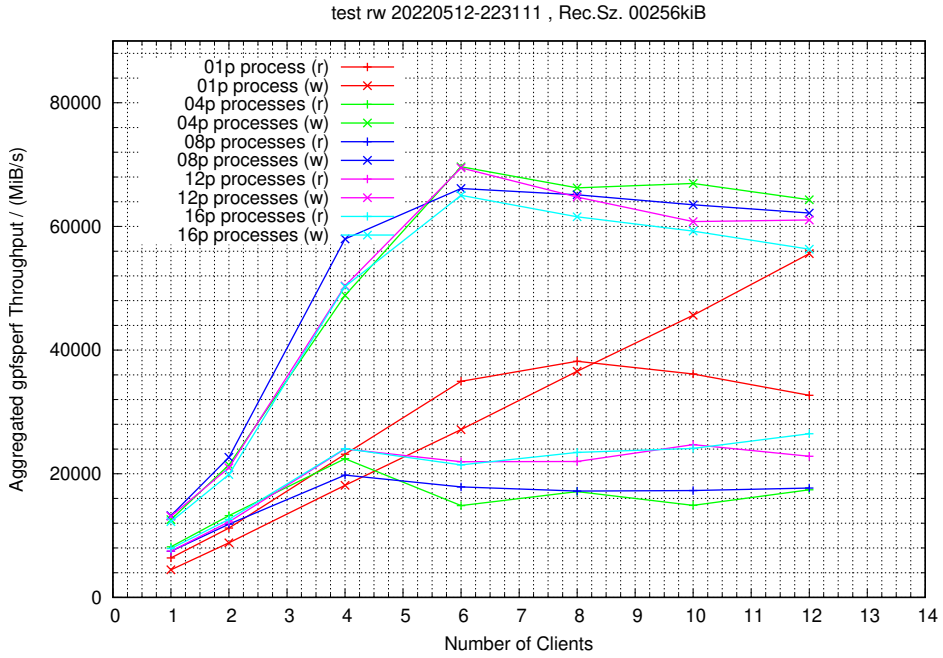




Uwe Falke

Steinbuch Centre for Computing (SCC)

2022-05-12, 256kiB IO Sz, parallel Read-Write (RW)

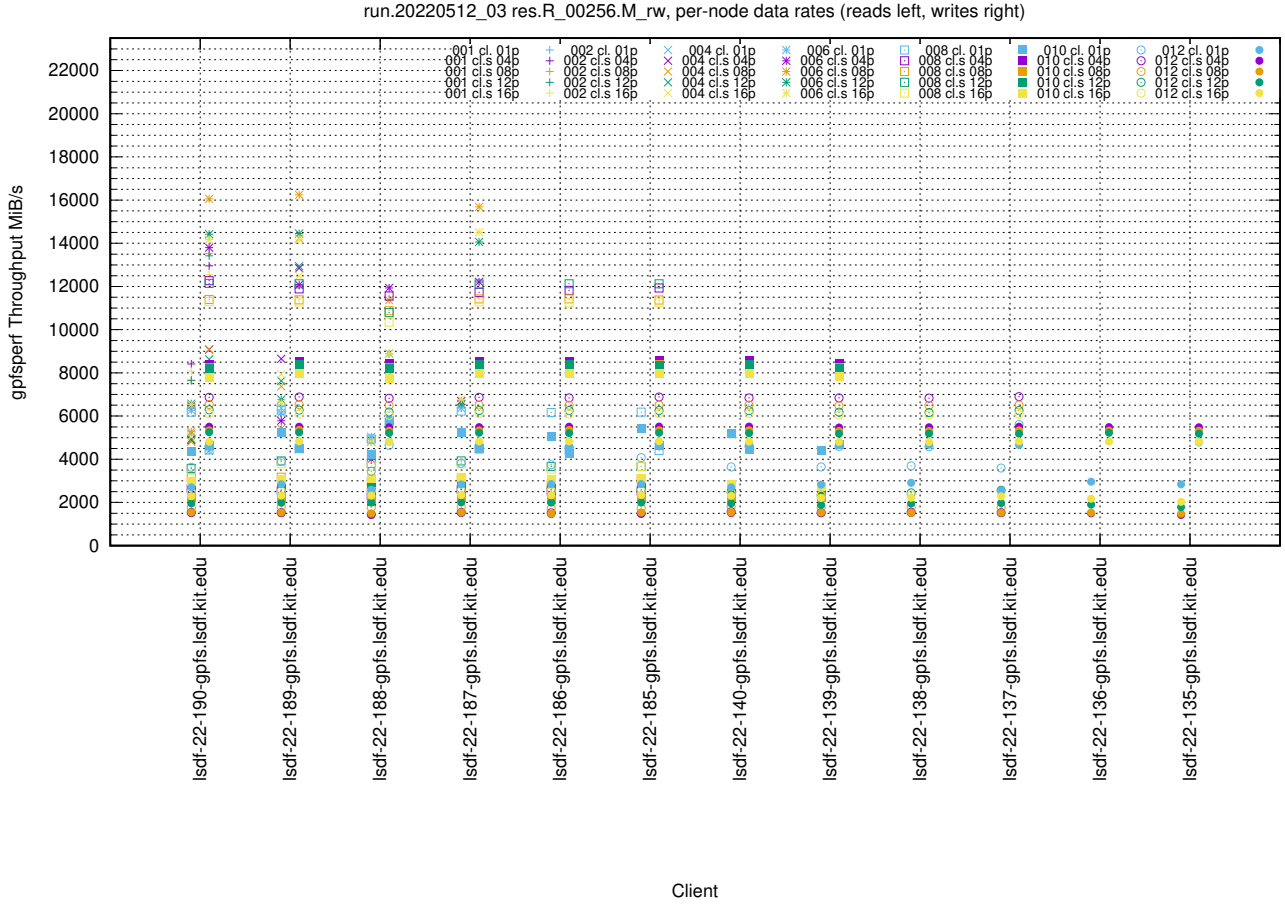


13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

2022-05-12, 256kiB IO Sz, parallel Read-Write (RW)



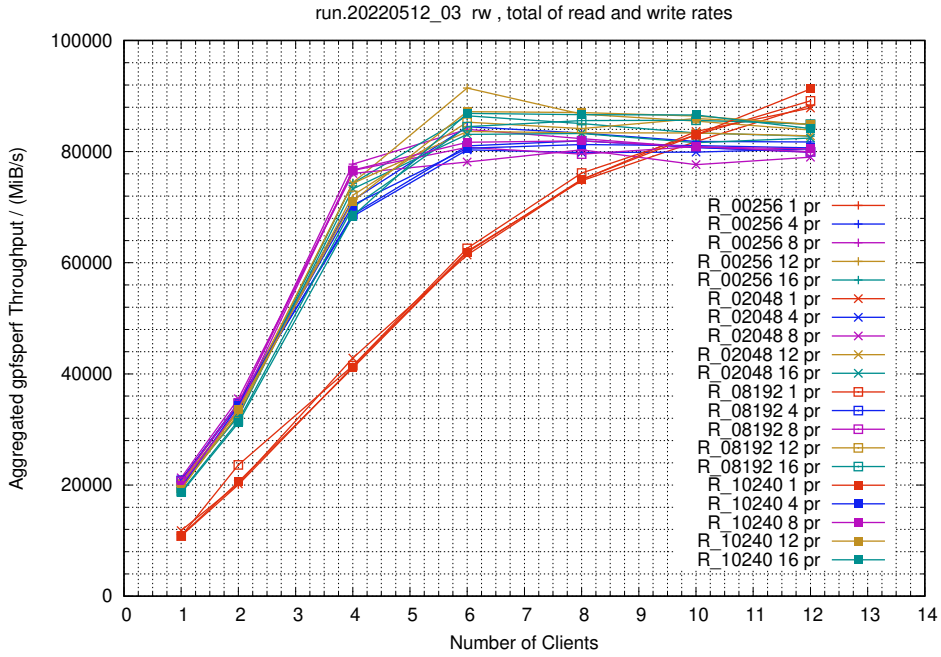
Karlsruhe Institute of Technology

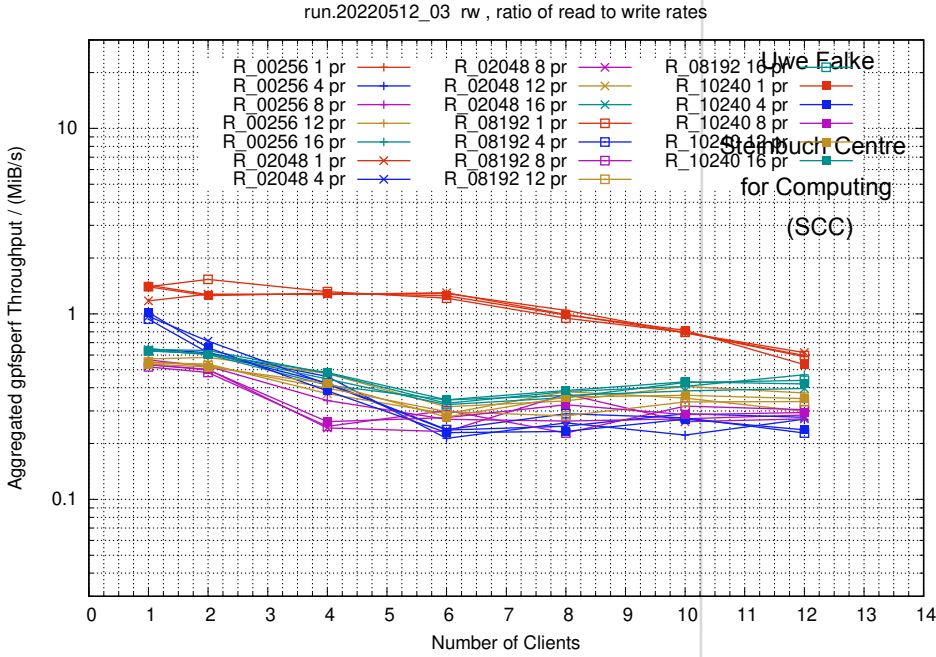
13. Jul 2022

Uwe Falke

Steinbuch Centre for Computing (SCC)

2022-05-12, aggregated R+W rates, R/W ratios





Uwe Falke
Steinbuch Centre
for Computing
(SCC)