UPPSALA
UNIVERSITET

V1.3, March 13, 2017
Author: Jerker Nyberg von Below

SOP – Common service PC File Server

1 Preamble

The service **PC File Server** is produced by BMC-IT and offered to Uppsala University. It is especially suited to departments at BMC that need a place to store a lot of data in a reliable and cheap way, but do not need high availability or high performance.

The service is fully funded by the users, both for hardware and maintenance.

2 Cost

The service costs **1200 SEK/TB/year** when we begin in 2016. The cost is for the full and dedicated storage space - free or in use.

During the startup-phase (years 2016-2019) a contract has to be made for the service several years in advance in order to cover the investment costs. It is our hope that if the service is self-sustained then after 2019 only agreements for one year at a time is needed. If it will be cheaper to produce the service in the future, with larger and cheaper drives, all users will get the same and lower price.

3 Alternative services that you may use instead

Please note that there may be other services that may be better suited for your needs.

- Compared to **Google Storage and Amazon Glacier** (prices from 2016-10-20) the storage is in the same price range. The difference is that the cheap options are charging money for reading back data. The cheapest option for just storing and not reading back data at Google Storage and Amazon Glacier is 0.007 USD/GB/month = **772 SEK/TB/year**. But when reading back all data once the cost will be **4404** or **7488 SEK/TB/year**. There are other options that are either slightly more expensive or also have a fee for reading back data. Please note that you have to follow **Personuppgiftslagen** and other laws and regulations regarding how and where to store your data.
  http://www.datainspektionen.se/lagar-och-regler/personuppgiftslagen/

  Google Storage: https://cloud.google.com/storage/pricing#storage-pricing
  Amazon Glacier: https://aws.amazon.com/glacier/pricing/

- IT-division at University administration has a Hitachi NAS file server service. It costs currently **7000 SEK/TB/year.** Remember that the HNAS has redundant hardware and better availability and has backup via TSM. It is being offered for a

very reasonable price considering what is included. It is well suited for home directories for Windows and similar highly available services.

- UPPMAX has high performance scalable parallel file servers accessible from the shared university HPC (high performance computing) clusters. Storage space can be applied for. It is well suited for storing and working on sequence data. As far as we know the service is free if application is accepted and storage is available. If you need your own storage cluster you should talk to UPPMAX.

- Several organizations (University administration, Rudbeck Laboratory/UAS, Polacksbacken, EBC et al) have their own **campus or department storage** for their own users which may be subsidized or better suited for your requirements, if you are allowed to use it. You may already be paying for it. Please check what your research group, department or campus can do for you.

- The university may be on the way of introducing some kind of **central research storage**. How this will be financed we do not know yet.

4 Technical description

Even though this storage as a service at BMC-IT is starting in 2016, a similar technical setup with simple PC file and backup servers has been used since 2010 at two departments here at BMC.

The servers are put in the BMC computer room with redundant cooling and power.

The data is accessible through Windows SMB/DFS protocols and can be used from Macintosh and Linux computers using SMB. User authentication is handled through USER-AD using the normal university account and password.

Data may be accessed via SMB at *smb://org-sharename.files.uu.se*.

Access to the share is limited at the directory level below the share, normally using a group in the USER-AD that is including the AKKA-group for the research group or department as read-write users. This way the research group leader, when adding new group members, also automatically grants access to the storage. BMC-IT does not want to be involved in the daily adding and removing of user access. Also, we do not want any other ACLs are set on a higher level than the top-level directory.

Data is every night backed up to a secondary backup server using Rsync and snapshots.

The file and backup servers are running CentOS 7 on PC server hardware of a simple kind, for example Supermicro, Dell or HP. No on-site support from the vendor is included.

The service is not intended for use as a home directory file server for Windows computers. The IT-division HNAS file server is a better solution for that. It is not meant

to be a scratch space for computing needing high IOPS or bandwidth either. Use UPPMAX HPC storage for that.

Do not use bidirectional file synchronization connected to the file server. This includes but is not limited to Windows off-line files, Unison file synchronizer and similar software.

Using zip or tar to store a large number of files into a single file (archive) is good because it reduces the number of files in the file server which in turn makes the backups and traversing the file system go faster. Do not however incrementally update the contents of an archive because the snapshots will keep storing new versions of the files. Changing the contents of a file will also in most programs rewrite the full file, unless explicitly told not to. Compression for individual files are in general unnecessary because of the built in compression in the file system.

The drive type will be Seagate Archive SATA drives of 8 TB or larger. These drive have high TB/SEK but low IOPS/TB. Another option was to use enterprise drives with about the same IOPS/SEK, better IOPS/TB but a lot less TB/SEK. We would have preferred the middle class NAS-drives with better IOPS/SEK but slightly lower TB/SEK but have not found a supplier for servers with such drivers.

5 Data integrity

All data is stored twice. Primarily it is stored on the file server and from there it is synced to the backup server once a day.

The data is stored using RAID6 in groups of four or six drives with two parity drives and two or four data drives. Hardware RAID controller is currently being used but software RAID may be used in the future.

Snapshots are taken on the primary file server every hour. Hourly snapshots are saved for one day. Daily snapshots are saved for a week. Weekly snapshots are saved for a month. Monthly snapshots are saved for a year.

Snapshots are taken on the backup server every night. Read the separate SOP called *Rsync backup to Btrfs snapshot* for how this is done. Daily snapshots are saved for one week. Weekly snapshots are saved for a month. Monthly snapshots are saved for a year.

Remember that the service is as is. We can never be 100% sure that nothing bad will happen but we are, based on our experience over the last couple of years, confident that this solution is for most cases good enough. For higher level of security, when storage of this kind is not good enough, please consider using the IT-division HNAS service. Also consider using the university TSM backup service. You must be aware of the risk of data loss.

6 Disaster recovery and migration of data

If files are accidentally being deleted the snapshots may be accessed by the users themselves.

- In Windows, use the Previous Versions feature in the File Explorer by right clicking on a directory and choosing Properties and then go to the Previous Versions tab.
- For Mac attach to server with Finder at smb://username@ORG-Sharename.files.uu.se/ORG-Sharename/.snapshots
- For both Windows and Mac (and probably Linux as well) command line can be used to enter the hidden directory .snapshots in the root of the share.

The directories are using GMT-timezone in order for Samba to get the right date for Previous versions.

If a single 4 drive RAID6 volume breaks down the drives should be replaced and the data should be read back from backup. If the volume is part of a larger file system (Btrfs multiple device or an extended logical volume with XFS on top of with multiple 4 disk RAID6) the file system has to be recreated even on those that did not broke.

If a server totally breaks down (motherboard failure etc) all drives could be moved over to a second machine and started as is. When the system drives also are moved the second machine will take the identity of the broken machine.

If small parts break (RAM failure) spare parts could be taken from the spare machine to fix the broken one. Replacements are ordered to the spare machine.

If a file system has to be expanded then add a new volume of 4 drive RAID6 and add the device to the file system. Is the server is full, then shut down the file system, move the volume (the 4 drives) to an other server with free drive bays, add the new drives to the same server, and then expand the file system. Export file system in Samba. Move destination of CNAME for the share (*smb://org-suborg-name.files.uu.se*) to the new server.

If the contents of the file server have been corrupted by ransomware on the clients connected to the storage with write access, then all the affected data of the file server has to be restored either from the last uncorrupted snapshot or from the backup. The latest corrupted snapshot has to be removed because else the backups (of the uncorrupted + corrupted files) will soon be too large to fit in the backup server.

7 Performance

Currently each file server can deliver 1 Gbit/s transfers to clients. All the clients simultaneously accessing the server are sharing that bandwidth. It is possible to upgrade to 10 Gbit/s when needed, even though the cheap 10GBASE-TP we are planning to use are still quite new and expensive.

The file server is connected with10 Gbit/s to the backup server.

We do not think file servers with file systems on this kind of hardware are capable of handle more than at most a few Gbit/s per user or file system. Upgrading to 10 Gbit/s may increase the total concurrent bandwidth on that particular server.

The different users (research groups) and file systems are separated to different drives. This means that a single user of a file system will not saturate IOPS on other file systems and drives. It makes it also possible to physically move the file system (the drives) to another server if needed by performance or expansion reasons.

The bandwidth while reading many small files are lower than for large files since the mechanical drives usually have to do a seek before accessing each file. The systems have been running with over 50 million active files and including snapshots over 300 million files. The theoretical limit of Btrfs the limit is 2^64. Restore (using Rsync) even from quite small files is usually running closer to 100 MB/s.

8 Future development

Snapshots could be sent from primary server to backup server instead of using rsync. This would handle renames and duplicates in a more efficient way. But interruptions during transfer of the snapshots is quite hard to handle.

Deduplication is possible in Btrfs. It is using online out of band deduplication. We have ran **duperemove** but only on small data (100GB). We currently do not know the RAM usage for running **duperemove** for large datasets. Deduplication will lead to increased fragmentation. The smaller the block size the bigger is the resulting fragmentation and also space saving. It does not seem like Btrfs has the same drawback as ZFS which become almost useless with not enough memory even when reading from a dedup dataset. Btrfs will happily read deduplicated data in low memory situations, even though duperemove cannot handle the data.

Time machine could perhaps be run directly to the file server. This has been tested technically but has to be packaged as a service in order to be offered. There may be technical problems using the incremental rsync/snapshot backup of the time machine storage.

Upgrade to 10 Gbit/s 10GBASE-TP on the switch to get faster bandwidth. 10 Gbit/s is available in the server room routers but currently not in the top-of-rack switches.

9 FAQ

Q: Can the storage be expanded?

A: Yes and no. The storage is physically located on dedicated drives in the machines. It is possible to add a new RAID-set and then add that device to the Btrfs file system. I would say that in general each file system will not be able to easily expand since that requires free drive space on the PC file server. It is also possible to use logical volumes and Ext4fs or XFS on top of it. So it is possible to move drives to another server and

them make room for new drives, on the new or on the old server. Large file systems also increase the risk of failing the whole file system. It is preferred to create a new file system on new drives.

Q: How large can the storage be?
A: In theory all storage on a single server can be used in one large file system, but we prefer smaller chunks making restore after file system failure easier. We prefer keeping each file system at 16 or 32 TB.

Q: How small can the storage be?
A: The storage has to be at least 16 TB usable. This is 4 drives with 8 TB in a RAID6 configuration. With 1200 SEK/TB/year and 16 TB storage this would give a yearly cost of 19200 SEK/year.

Q: Are the PC file servers redundant?
A: No! Neither is the network switch. And no service during weekends or vacation either. Some components like fans (several), power supplies (two) and drives (RAID6) are redundant in a single server.

Q: I do no not trust these shenanigans with Linux and Btrfs and Rsync. I want something enterprise-ready I can rely on!
A: Ok! Use the IT-division HNAS file server service instead. But remember that for example Amazon, Google and Oracle are using Linux and Btrfs a lot together with cheap hardware and we do consider them enterprise. Everything is about using the right tool for the job.

Q: Can quota be used?
A: No. Technically yes, but this is a simple service, so in order to keep maintenance to a minimum we skip that.

Q: Can ACLs be used?
A: We set base ACLs for the folders at the top level so that only the persons that should be able to access data can access it.

Q: Can invoices be sent to different receivers for the storage?
A: No. One invoice for each customer.

Q: Who owns the storage?
A: The hardware belongs to BMC-IT. The data is yours.

Q: Can this break down?
A: Yes, it most probably will break down at some point in time. Then you have to wait until the service can be restored. If this happens during vacation time you may have to wait until the end of the vacation.

Q: I want all these extra features.
A: If you develop them and test them maybe we can implement them. But this is supposed to be a simple and cheap service.

Q: Why are you not using Ceph or GlusterFS instead?

A: We have experience with introducing both of these systems at Uppsala University at UPPMAX already in 2007 on the old Hagrid cluster. We have also been running Ceph with block-level storage for virtual machines at BMC-IT for testing a few times since then. We consider them both interesting systems possible to scale up to quite high capacity and performance. CephFS is of particular interest. But they both introduce a high level of complexity that we don't like in this kind of service. If you need that level of scalability, contact UPPMAX who are specialized in this and use their services instead.

10 Ordering

To order the service, send an email to helpdesk@bmc.uu.se containing information about the amount of storage needed, a preferred name for the storage file system and share and contact information for sending invoices.

The delivery time is around two months from ordering to using the storage. If there is already free capacity online then the delivery time is shorter.

11 Current status in interest, ordering and users

Contact person / *Department*; *Section* / Size / Status

1. Mats Pettersson / Leif Andersson
   Department of Medical Biochemistry and Microbiology (IMBIM); Genomics
   Size 96 TB
   Interest 2016-05-10
   Confirmed (100 TB) 2016-06-08
   Name IMB-GenomicsLA1 (32TB)
   Name IMB-GenomicsLA2 (32TB)
   Name IMB-GenomicsLA3 (32TB)
   Delivery 2016-09-09
   Billed at 2017-12-07 period 2016-10-01 – 2016-06-30 for 86400 SEK

2. Carl-Johan Rubin
   Department of Medical Biochemistry and Microbiology (IMBIM); Genomics
   Size 16 TB
   Interest 2016-05-10
   Confirmed (20 TB) 2016-06-15
   Name IMB-GenomicsCJR (16 TB)
   Delivery 2016-09-09
   Billed at 2016-12-07 period 2016-10-01 – 2016-06-30 for 14400 SEK

3. Anna Nilsson / Per Andrén
   Department of Pharmaceutical Biosciences; Medical Mass Spectrometry
   Size 32 TB
   Interest 2016-05-30
   Confirmed 2016-06-15

UPPSALA
UNIVERSITET

Name FBV-MSImaging (32 TB)
Delivery 2016-09-09
Billed at 2016-12-07 period 2016-10-01 – 2016-06-30 for 28800 SEK

4. Lars Bäckström
   Department of Medical Sciences; Molecular Medicine
   Size 32 TB
   Interest 2016-06-08
   Name MOL-EXTBMC (32 TB)
   Delivery 2016-09-09
   Billed at 2016-12-07 period 2016-10-01 – 2016-06-30 for 28800 SEK

5. Jennifer Meadows
   Department of Medical Biochemistry and Microbiology, Genomics
   Size 64 TB
   Interest 2016-09-07
   Name IMB-GenomicsKLT1 (32 TB)
   Name IMB-GenomicsKLT2 (32 TB)
   Ordered 2016-12-09
   Delivery 2017-03-09
   Billed at 2016-12-09 period 2017-02-01 – 2020-01-31 for 230400 SEK

6. Malin Lagerström
   Department of Neuroscience, Developmental Genetics; Sensory circuits
   Size 16 TB
   Interest 2016-06-20
   Name INV-SC (16 TB)
   Ordered 2016-12-07
   Delivery 2017-03-09
   Billed at 2016-12-09 period 2017-02-01 – 2017-06-30 for 8000 SEK

Equipment:

1. Supermicro with 36 HDD 8 TB – primary
   Ordered 2016-06-15 unpacked 2016-08-22
2. Supermicro with 36 HDD 8 TB – backup
   Ordered 2016-06-15 unpacked 2016-08-22
3. Supermicro with 36 empty slots – cold stand-by
   Ordered 2016-06-15 unpacked 2016-08-22
4. Supermicro with 36 HDD 8 TB – primary
   Ordered 2016-12-09 arrive 2017-01-30 unpack 2017-02-01
5. Supermicro with 36 HDD 8 TB – backup
   Ordered 2016-12-09 arrive 2017-01-30 unpack 2017-02-01

UPPSALA
UNIVERSITET

12 Event log

2016-09-16 14.00 bmc-pcfs1: Service started.
2016-10-01 02.06 bmc-pcfs2: arcmsr hang, bus reset, two offline devices
2016-10-02 09.00 bmc-pcfs2: Recover by reboot server.
2016-10-05 11.00 bmc-pcfs1: Move from server room C6:3 to D11:0. New IP.
2016-10-05 12.00 bmc-pcfs{1,2}: Connect with 10GBASE-TP.
2016-10-06 11:00 bmc-pcfs2: Upgrade to arcmsr v1.30.0X.23-20151225 and reboot
2016-10-07 10:00 bmc-pcfs1: Upgrade to arcmsr v1.30.0X.23-20151225 and reboot
2016-10-07 12.45 bmc-pcfs1: Lost connection to USER-AD. Restart Winbind.
2016-10-12 11.00 bmc-pcfs1: Samba stopped working. Restart Samba.
2016-10-20 12.00 bmc-pcfs1: Start Netatalk for experimental Time Machine service.
2016-10-24 08.40 bmc-pcfs2: rejecting I/O to offline device. Reboot. Start investigation with Southpole.
2016-10-31 11.00 bmc-pcfs1: mlocate filled up /var, blocking login. Remove mlocate.
2016-11-09 06.01 bmc-pcfs2: arcmsr0 abort scsi_cmd again. Contact Southpole support again.
2016-11-09 14.40 bmc-pcfs2: Upgrade to arcmsr v1.30.0X.26-2016-20161104 and reboot.
2016-11-15 16.10 bmc-pcfs1: Upgrade to arcmsr v1.30.0X.26-2016-20161104 and reboot.
2016-11-17 03.30 bmc-pcfs2: arcmsr0 abort scsi_cmd again. Contact Southpole support again.
2016-11-18 13.40 bmc-pcfs2: reboot after testing btrfs balance.
2016-11-30 11.05 bmc-pcfs1: fixed wrong permissions on GenomicsLA3
2016-12-12 17.00 bmc-pcfs1: Domain controller unreachable error. Restart Winbind.
2016-12-13 05.13 bmc-pcfs1: Domain controller unreachable for MMS. Reported 13:40 (11:40).
2016-12-13 14.24 bmc-pcfs1: Restarted Winbind. Domain controller reachable again.
2017-02-06 08.30 bmc-pcfs2: arcmsr0 abort scsi_cmnd again. Reboot.
2017-02-17 08.30 bmc-pcfs2: arcmsr0 abort scsi_cmnd again. Reboot. Contact Southpole.
2017-02-20 08.30 bmc-pcfs2: arcmsr0 abort scsi_cmnd again. Reboot.
2017-02-21 08.30 bmc-pcfs2: arcmsr0 abort scsi_cmnd again. Reboot. Contact Southpole.
2017-02-21 16.00 bmc-pcfs2: Change Areca settings: ncq, tler, timeout & spoweron
2017-03-09 15.00 bmc-pcfs4: Online
2017-03-09 15.00 bmc-pcfs5: Online
2017-03-09 19.30 bmc-pcfs4: arcmsr0: abort sccsi_cmnd:
2017-03-10 08.00 bmc-pcfs4: reboot. Contact Soutpole.