



8th TF-Storage meeting
(Thursday-Friday) 3-4 February, 2011
Budapest, Hungary

Table of contents

Table of contents	1
Minutes	1
1. Welcome and announcements.....	1
2. Approval of agenda and minutes.....	2
3. Vendors' presentations	2
4. FileSender project update	3
5. Storage benchmarking Work Item	4
6. National updates.....	5
7. NRENs strategic perspective on storage and cloud	7
8. Next meeting, AOB, and close	7
List of participants	8

Minutes

1. Welcome and announcements

The eighth TERENA Storage Task Force meeting was held on 3-4 February, 2011 hosted by NIIIF/HUNGARNET in Budapest, Hungary. Jan Meijer (UNINETT) welcomed the 31 participants (see the list of participants enclosed) and asked for a roll call. The TERENA secretary was Peter Szegedi.

Jan officially announced that he is renouncing his position as chairman of TF-Storage, effective after the meeting. As the reason of his decision, Jan explained that there is a change in UNINETT's organisation and priorities that puts him at a distance from the type of storage activities typically discussed at TF-Storage, although he is planning to participate in TF-Storage (co-ordinating the FileSender project) in the future.

Jan proposed Maciej Brzezniak (PSNC - the Polish NREN) as the new TF-Storage chairman. Maciej has been a member of the TF-Storage since its inception. He is very experienced with storage and works with the problem space on a daily basis, being part of the Polish national storage for R&E project. Since there was no objection against the new chairman on the tf-storage mailing list, the former chairman, the TERENA Technical Committee, the task force secretary, and the all the attendees of the meeting supported Maciej, he officially become the new chairman of TF-Storage on 4 February, 2011.

2. Approval of agenda and minutes

The meeting agenda was agreed without any changes. The minutes of the last meeting (in Poznan) was also approved. The presentations are available on the TF-Storage website:

<http://www.terena.org/activities/tf-storage/ws10/agenda.html>

3. Vendors' presentations

- Bela Sagi (Fujitsu) introduced the ETERNUS DX storage platform, recently chosen by NIIF as the main storage platform for their HBONE+ project.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110203-fujitsu.pdf>

As Fujitsu understood the call for tender, NIIF partners may apply for storage space as block devices. These could be configured as required and be accessed from partner's side as iSCSI target devices via IP-network and iSCSI protocol. Therefore, it is important to have basic understanding of volume management and iSCSI protocol. Four mid-range storage devices (Fujitsu ETERNUS DX80) serve as a basis for the storage infrastructure. These all have the same 1Tbyte NL-SAS disks to provide a gross total capacity of 120Tbytes each. Two each of four devices are installed at the regional centre of NIIF at the College of Dunaújváros and at the University of West Hungary. All of these storage nodes are connected directly to the HBONE network backbone via Gigabit Ethernet on four ports each.

Fujitsu offered the ETERNUS DX80 platform that is a mid-size box fully redundant against component failures. When power fails, a capacitor maintains power until the cache content is saved to non-volatile memory (flash). The Data Block Guard feature is a 8-byte long check code that is added to every 512-byte data in order to control data integrity both on the drive and in the cache. Data encryption is done by the box itself and not by an application or additional layer. Removed encrypted drives are protected against unauthorized access. Currently there is no feature implemented to easily destroy the encryption key (and make the data unreadable) in case of platform phase out. On the energy saving side Fujitsu applies a mechanism that reduces the rotation of the disks to a minimum speed when they are not touched by the application in a certain period of time. Wide variety of servers and OS are supported.

Concerning platform capacity Bela mentioned that 2TB disks are supported at the moment; 3TB is in the roadmap of the new system. 0.5 PB is the total capacity of the four ETERNUS boxes deployed by NIIF. The SAN can be FC or iSCSI based, there is no iSCSI over 10GE solution yet, will come.

- Gabor Vitez (Avaxio) talked about Coriad Inc. storage products from a reseller perspective. Coriad is the developer of AoE (ATA over Ethernet) protocol based storage products focusing on exceptional price/performance ratio. ATA over Ethernet (AoE) is an open standards based protocol that allows direct network access to disk drives by client hosts. Using disk storage arrays that support AoE shared storage networks (SAN) can be built that leverage the power

of "Raw" Layer 2 Ethernet.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110203-coraid.pdf>

AoE protocol is designed to be implemented in hardware with very low resource usage. Storage boxes have a SAN-unique ID, similar to an Ethernet MAC address, so the boxes can naturally be connected to Ethernet networks. AoE protocol needs driver to be installed on the client side. Free driver is available on Linux, and a third-party driver for Windows. The administration of the Coriad boxes is easy, only 3 commands are needed under Linux. It provides low SAN complexity due to standard Ethernet usage. There is no strict packet ordering (no TCP applied) although the data integrity is guaranteed by Ethernet checksum. The maximum span is limited by the Ethernet protocol. Ethernet, of course, can be tunnelled over WAN, so a nation-wide solution can easily be built with Coriad. Among the use cases it was mentioned that Coriad is optimal if less than 5TB storage is needed.

From the security point of view it was mentioned that the storage box is not routable so cannot be reached via e.g., the Internet. The transport Ethernet network should use proper Ethernet switched that support e.g., jumbo frames.

At the moment there is no competition on the AoE market, only Coriad is providing that. For testing purposes there is a Linux test platform available for free but its performance is not that good so it is recommended to get a box from a Coriad reseller. Performance numbers are available on the Coriad website.

- Barnabas Nagy (HP) talked about SAS technology that makes the bridge between storage capacity and performance. Serial Attached SCSI (SAS) is a computer bus used to move data to and from computer storage devices such as hard drives and tape drives. SAS depends on a point-to-point serial protocol that replaces the parallel SCSI bus technology and it uses the standard SCSI command set. SAS offers backwards-compatibility with second-generation SATA drives. SATA 3 Gbit/s drives may be connected to SAS backplanes, but SAS drives may not be connected to SATA backplanes.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110204-HP-Storage.pdf>

Barnabas talked about the various cable types. Usually SAS and MiniSAS cables are used but SAS2.1 standard active cables (with a power supply inside) are also available. The standard allows 100m single cable span but in practice 8m cable is the longest one that can be bought. The price of the cables is one third of the fibres' cost. The SAS switch cards are also cheaper than the FC switches. Concerning the bandwidth it was mentioned that 600 MB/sec looks sufficient for now, but in 2-3 years time frame it is planned to be doubled.

4. FileSender project update

Jan Meijer (UNINETT) gave an update on the FileSender project: <http://www.filesender.org/>. The value proposition of FileSender is to provide easy sharing of arbitrarily large files through a

trusted intermediary. The brief history of the collaborative development effort, the open source development community, and the service architecture were briefly explained by Jan.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110203-filesender.pdf>

After 2 years of project lifetime a solid software release 1.0 with good documentation is ready for production. There are known installations at AARNet, ARNES (soon), BELnet, FCCN, HEAnet, SRCE, SURFnet, TERENA, and UNINETT. The project group gained solid understanding on the problem space, established a collaborative cross-organisational community (no reliance on one single NREN), and collected sufficient funding for 2011.

In the FileSender 2.0 roadmap, the highest priority is on the exclusion of Flash (currently needed for the reliable progress bar) and Gears (currently needed for chunking data over 2GB). Everything should be solved under HTML5. Chrome and Firefox is advanced in HTML5 implementation and the preliminary tests show that the data chunking is more efficient so the new release can even be faster.

During the discussions there was an idea to hook up FileSender with a website that checks the originality of the uploaded/sent documents. It was also suggested to support MySQL database. Another idea was to use FileSender as a server image repository for cloud infrastructure. Currently there is no API available on the FileSender side for further integration. All feature requests can be submitted to the project group via its website or mailing list.

5. Storage benchmarking Work Item

Maciej Brzeźniak (PSNC) gave an update on the progress of the TF-Storage "Measuring storage performance" Work Item. Storage benchmarking is a bit of theory and a bit of practice (in the context of tenders). The main motivations behind this work item are: the need for evaluating storage systems taking into account its applications on top, the lack of standard/universal ways of describing system performance, and the verification of vendors' offers in a reliable and efficient manner.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110203-storage-performance.pdf>

Storage benchmarking has a very practical usage in tendering procedures. One of the aims of this work item is to collect participants' experience on this. Most of the cases vendors are selling products via resellers and resellers are chosen by the vendors. It is important to know who is responsible for the appropriate benchmarking. Vendors would appreciate to know the benchmarking procedure used by the clients during the tender evaluations in order to cope with the actual requirements. However, if the benchmarking tools are not open-source, it is hard to use/publish them in public tenders. Most of the cases the vendors test beds and benchmarks are different from the client procedures so the expected performance will never be reached. If the vendor is forced to use new type of testing on their side that may cause extra expenses for the client. Time and money can be saved on both sides if the clients' expectations and evaluation is close to the testing environment and benchmarking used by the vendors. The

aspects of potential benchmarking standards and pre-negotiation with vendors on benchmarking as part of the actual tendering procedure have been discussed.

Andy Valley (Cisco Systems) has volunteered to give a presentation about benchmarking from a vendor's perspective.

ACTION 1 on Peter Szegedi (TERENA) to follow up with Andy Valley (Cisco Systems) on the potential benchmarking talk by the next coming TF-Storage meeting.

Maciej mentioned that the latest results of the work item, including the best practices collected so far, are available on the old TF-Storage Wiki space:

https://wiki.terena.org/index.php/TF_Storage_Benchmarking_Howto

This information is quite old (may be obsolete) so revision is definitely needed! Peter Szegedi (TERENA) noted that the old Wiki content will be migrated to the new TERENA Confluence Wiki platform, anyway.

ACTION 2 on Peter Szegedi (TERENA) to migrate the old TF-Storage Wiki content on benchmarking to the new TERENA Confluence Wiki space:

<https://confluence.terena.org/display/Storage/Measuring+storage+performance>

6. National updates

- Janos Mohacsi (NIIF), on behalf of the hosting NREN, gave a presentation about the HBONE+ project that has been awarded with 15M EUR by the national government.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110203-HBONEplus.pdf>

Janos talked about the network infrastructure, services, and applications. Recently optical dark fibres have been obtained in Hungary and DWDM system has been deployed on top. Some nodes have storage platforms other have supercomputing sites (the distribution is partly caused by power limitations of the sites) although the supercomputing facility looks like one solid platform from the user's perspective. Storage was part of the HBONE+ tender that has resulted a storage system with 0.5 PB total capacity distributed on four major sites of the network.

- Szabolcs Szekelyi and Tamas Kazinczy(NIIF) gave a more detailed overview on the NIIF/HUNGARNET current storage and cloud developments and deployments in the context of the HBONE+ project.

Four Fujitsu-Siemens Eternus DX80s have been installed on four major sites with 480TB total raw capacity. The storage service provided is a low-level "block device" service based on iSCSI. The platform provides iSCSI interface access on 4x1GE links each. It is intended for HBONE user community to do data staging, digital and video archives and backup, as well as to support higher level (cloud) services.

More information about the storage developments is available here:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110204-niif-2.pdf>

The cloud service development activity follows the Infrastructure as a Service principle including features such as: self-service private virtual network management, live migration of VMs between sites, integrated storage management, and console access to VMs.

Further details on the cloud related activity can be found here:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110204-niif-1.pdf>

Szabolcs mentioned that the current cloud platform is lacking proper image management function. API is provided that is a home-grown restful API for all operations (creating virtual machines, etc). NIIF is heading towards a production cloud service although the resource management and accounting issues are not yet fully clarified.

- Rogier Spoor (SURFnet) talked about DRBD WAN replication in combination with VMotion. SURFnet, with contracted partners, has recently studied end-user P2P storage and cloud storage solutions and various technologies for Wide Area Distributed Storage.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110204-SURFnet.pdf>

The Distributed Replicated Block Device (DRBD) installation has been tested by VMotion between Amsterdam and Nijmegen. The potential use cases for the live migration of VMs could be: maintenance, load balancing, or new service deployment at one site. The tests resulted that 4ms roundtrip over 200km distance can be done with DRBD. Synchronous replication is possible and VMs can easily be moved.

The plans for 2011 include a cloud storage solution for 4K streaming, the continuation of FileSender project contribution, a storage service addition in SURFconext, as well as IaaS business case and tech design.

- David Corney (STFC - Rutherford Appleton Lab) gave an update on RAL's storage related activities. There is an LHC Tier1 site at RAL. RAL has some experience with disk failures. Actually, it is hard to convince the storage system supplier that there is a structural problem with the disks provided by a third-party vendor. RAL conclusion was that for core businesses it is worth to buy expensive disks but for commodity services it is fine to have cheaper disks and replace them more often. It depends on the use case.

RAL's experience with tape library is good. Compared to disks it is an order of magnitude better so tape is not dead at all (do note that CESNET's tape library is an example for recent deployment). The robot can fail but mechanical failures (3-4 can happen per year) can be predicted and avoidable. LHC has only one copy at RAL (single copy) but more copies exist around the world.

RAL's plan is to shut down Datastore and use open-source CASTOR instead. However, the migration to CASTOR might be a nightmare. Would be good to point the metadata to the new machine (moving the metadata only) and leave the data on tapes (not moving the data

at all). Currently we are discussing about a couple of PBs that can be doable but the data will ever grow (10PB soon).

- Maciej Brzezniak (PSNC) talked about the Polish national project NDS2 on secure sharing, publishing and exchanging data. With NDS2 PSNC tries to address: more sophisticated backup/archive related user requirements, provide transparent versioning, increased security (encryption) and performance (hardware-aided), and improve overall system scalability.

For more details, see the slides:

<http://www.terena.org/activities/tf-storage/ws10/slides/20110204-nds2.pdf>

7. NRENs strategic perspective on storage and cloud

Peter Szegedi (TERENA) gave an update on the outstanding action item (TF-Storage/TSec(10)064 ACTION 1) on TF-Storage participants' strategic perspective on storage and clouds. Peter has contacted with the TF-Storage participants and asked them to provide any information available related to storage/cloud national strategies. SURFnet, CESNET, NIIF, PSNC and SWITCH have replied and indirect information has been collected from GRNET, HEAnet, UNINETT, JANET and RAL.

Peter presented the draft table of content of a potential white paper collecting national strategy in storage deployments and cloud related developments (if appropriate). The objective would be twofold; on one hand it would be beneficial for the TF-Storage group and the broader TERENA community, on the other hand it can be fed back to the European Commission policy making process in order to assist them taking the right assumptions.

For more details, see the slides:

http://www.terena.org/activities/tf-storage/ws10/slides/20110204_tf-storage_szegedi.pdf

Peter asked for more contribution! The progress on the white paper will be reported on the next coming TF-Storage meeting.

8. Next meeting, AOB, and close

TF-Storage meeting participants agreed to have a brief discussion during the TERENA Networking Conference 2011. Peter reported that a half-day slot is reserved for TF-Storage on Thursday afternoon (on **19 May 2011** in Prague, Czech Republic). The logistic details and the planned meeting agenda will be available on the TNC2011 website:

<https://tnc2011.core.terena.org/>

Moreover, after a quick poll on the mailing list, it turned out that there is sufficient interest in a full task force meeting in June timeframe. GRNET has volunteered to host the TF-Storage meeting in Athens, Greece. The meeting dates are fixed as follows: **16-17 June 2011**. The logistic details and the planned meeting agenda will be available on the TF-Storage website:

<http://www.terena.org/activities/tf-storage/ws11/>

Both registrations are OPEN, please, visit the websites!

The meeting was closed by Maciej Brzezniak (PSNC), the new chairman of the task force. Meeting host, presenters, and participants were thanked for their efforts and contribution to this successful meeting.

List of participants

Andrej Bagon	Arnes
Filiz Bektas	SWITCH
Brian Boyle	HEAnet
Maciej Brzezniak	PSNC Poznan
Sasa Cavara	CARNet
david Corney	STFC - Rutherford Appleton Lab
Miroslav Ivančević	CARNet
Jaroslav Kremenek	CESNET
Martin Kämpf	SWITCH
Faidon Liambotis	Greek Research and Technology Network
Rosend Llurba	NCF
Tamas Maray	NIIF/Hungarnet
Ivan Marton	NIIFI
Jan Meijer	UNINETT
Janos Mohacsi	NIIF/Hungarnet
Zsombor Nagy	NIIFI
Martin Osmundsvåg	UNINETT
Jernej Porenta	Arnes
Branko Radojevic	CARNet
Roger Skjetlein	Uninett
Rogier Spoor	SURFnet
Peter Stefan	NIIFI
Peter Szegedi	TERENA
Szabolcs Székelyi	NIIFI
Béla Sági	Fujitsu Hungary
Andy Vallely	Cisco Systems
Mark van de Sanden	SARA
Mario Vandaele	Belnet
Peter Vercimak	CESNET
Valentin Vidic	CARNet
Gábor Vitéz	Avaxio