Synchrotron-light for Experimental Science
And
Applications in the Middle East

Computing Needs of SESAME

Visit Report to Research Centers in Europe

Objective**:**

To Understand the Computing Needs of SESAME Scientific Program

Submitted to: LinkSCEEM Project

20 December 2009

Prepared By:
Mr. Rami Ahmad
Computer Engineer – SESAME

Approved By:
Prof. Hafeez Hoorani
Scientific Director – SESAME

Contents

EXEUTIVE SUMMARY: 4
1. INTRODUCTION:....5
2. VISIT to SLS 6
2.1 Meetings in 6th July 2009....6
2.1.1 Meeting with cSAXS beamline 6
2.1.2 Meeting with MicroXas beamline 7
2.2 Meetings in 7th of July 2009....8
2.2.1 Meeting with PXI and PXII beamlines....8
2.2.2 Meeting with MS beamline 10
2.3 Meetings in 8th July 2009....10
2.3.1 Meeting with SuperXas Beamline....10
2.3.2 Meeting with TOMCAT Beamline 12
2.4 Meetings in 9th of July 2009....13
2.4.1 Meeting with Control and Computing....13
3 VISIT to CERN....16
3.1 Presentations about CERN Computing Infrastructure....16
4 VISIT to ESRF....18
4.1 Meeting with the Head of Computing at ESRF 18
5 VISIT to SOLEIL....21
5.1 Meetings in 15th July 2009....21
5.1.1 Meeting with Control and Computing Group....21
5.1.2 Meeting with CRISTAL beamline....22
5.2 Meetings in 16th July 2009....23
5.2.1 Meeting with PROXIMA beamline 23
5.2.2 Meeting with DIFFABS beamline....25
5.3 Meetings in 17th July 2009....26
5.3.1 Meeting with Computing Group: Network administrator....26
5.3.2 Meeting with SAMBA beamline 27
6 COMMON FINDINGS at SLS and SOLEIL:....28
7 CONCLUSIONS & RECOMMENDATIONS:....29

8 REFERENCES.....31
Appendix 1: VISIT COMPUTING QUESTIONNAIRE.....31
Appendix 2: PERSONS MET 34

EXEUTIVE SUMMARY:

This report presents the outcome of my visit to research centers in Europe for better understanding the computational requirements of SESAME scientific program. The discussions were based on the computing questionnaire I prepared for the visit.

I found that the most intensive computing demanding beamlines are Protein Crystallography, SAXS, and Tomography beamlines and they cannot afford to be on computing resources queue. Interactivity is very important factor in case of beamlines and the online computing resources demands are high, so we have to have local high computing resources at SESAME. Computing power of 11 TFLOPS is estimated to accommodate the online requirements of SESAME complete scientific program (phase-I & phase-II) with any further expansion in number of beamlines which is foreseen in future. The remote computing power allocated by CSTRC is needed to accommodate the beamline users' remote analysis and simulations jobs from their home institute; where the computing power needed by the remote users is equal or can be greater than the online computing power. In addition the remote HPC resources can be used as a scalable or a fail-over HPC to the local HPC facility.

We expect to have a large amount of generated experimental data from the beamlines detectors, few TB per day estimates based on the current cutting-edge detectors technologies. This number is expected to increase in future based on the evolving technologies in beamlines detectors. We need local data storage capacity of 100 TB scalable to 500 TB and an archival system with a capacity of 1.5 PB to keep the experimental data for a period of one to five years. The experimental data is non-regeneratable and precious so we need a disaster recovery local datacenter. For SESAME computing model to work where the HPC is remotely based in the Cyprus Institute, we need a high throughput network connection between SESAME and CyI ranging from 155 Mbps to 1 Gbps to Cyprus to transfer the experimental data efficiently. Beamline local network preparation with 1Gbit/s Ethernet is needed as well, fiber connection with bandwidth 10Gbit/s or Infiniband access is required between HPC cluster and the local data storage.

The needed skills to build and operate the beamlines computing infrastructure are: two system administrators (Linux and Microsoft Windows), one network engineer, one scientific software programmer and four hardware technicians.

1. INTRODUCTION:

This report will address the computing needs of SESAME. It is based upon the requirements collected from several operational synchrotron facilities. One of the main goals of the report is to give a computing model that helps to enhance entire process of collecting vast quantities of data at SESAME from various beamlines highly effective and efficient.

This is a detailed report of my visit to research centers in Switzerland and France; my mission was to analyze the computing infrastructure setup for beamlines in the operational SR sources in Europe for better understanding of the computing requirements of SESAME. We have carefully chosen the light sources to visit where an important consideration was that the storage ring had similar parameters as for SESAME. We targeted only those beamlines where it is well know that the computing and storage requirements are more stringent. ESRF was also chosen in the visit plan to understand the extreme case. The current SESAME computing infrastructure is designed to serve offices and SR accelerator networks only.

The visit included the following research centers SLS, CERN, ESRF, and SOLEIL respectively. The main visits were to SLS and SOLEIL that are SR facilities close to SESAME technical design. I had several meetings at SLS and SOLEIL with representatives from selected X-ray beamlines and from computing groups. At CERN I attended two presentations about CERN grid project, storage and network and at ESRF I had a full day meeting with the head of computing services and I discussed with him the computing setup in ESRF and the computing requirements of a functional SR. All the meetings were based on the computing questionnaire I prepared for the visit; you can find the visit's questionnaire in Appendix 1.

The report concentrates on the computing requirements of the beamlines and may not include some of needed physics technical specifications. The full physics technical specifications can be collected from the publications on the research centers' websites or by contacting the beamlines scientists' contacts in Appendix 2.

The report is ordered according to the meetings place and date; the collected information and the discussed points are broke down under each meeting.

2. VISIT to SLS

2.1 Meetings in 6th July 2009

2.1.1 Meeting with cSAXS beamline

Collected Information:

- End-Station Information:
  - PILATUS 2M: a PSI-developed detector, which operates in single-photon counting mode and thus allows operation without read-out noise, 20bit dynamic range. www.dectris.com
- There are many research interests in this beamline in biology and material science fields.
- The source is Undulator.
- The beamline users before coming to SLS register in DUO beamline users' management tool.
- Available computing resources:
  - 30 TB storage allocated on the file server.
  - 4 core CPU, 16 GB RAM connected to the detector.
  - 8 core CPU, 32 GB RAM. For analysis.
  - 8 core CPU, 64 GB RAM. For analysis.
  - 8 core CPU, 32 GB RAM, Powerful graphic card for visualization.
  - SPEC control machine integrated with EPICS. For information about SPEC http://www.certif.com/ note: SPEC has many macros the cover the users different demands.
  - Media station. Accessing internet, transfer data.
  - Monitoring camera.
- The network connections as the following:
  - 2 GBit/s Ethernet network link between the detector and its server.
  - 1 Gbit/s Ethernet for the beamline network.
- All the computing resources are needed online while the experiment is running.
- No shared resources with other beamlines.
- The experimental data is available online for one month then it is removed from the data storage.
- Remote access to the data and the software application is not provided. But it would be nice if it's available.
- The users of this beamline are not so sensitive about their data.
- The needed software applications:
  - MATLAB. (Data analysis).

  - SPEC for data acquisition and experiment control.
  - VGStudio Max for data visualization. http://www.volumegraphics.com/
- 2 TB space is generated for a single experiment during 5 days.
- After the experiment is finished few GBs could be generated after the data analysis.
- Used programming languages: MATLAB, C, and Python.
- The experimental data property:
  - Few millions files
  - The file size range 800 KB – 10 MB
  - CBF “Crystallography Binary File” compression is used. http://www.esrf.eu/computing/Forum/imgCIF/index.html . The new compression method is hdf5 and they are moving to adopt it.

2.1.2 Meeting with MicroXas beamline

Collected Information:

- End-Station Information:
  - Microfocusing (KB-Box).
    80mm after the KB Box (fine-focusing of the x-ray-beam) there is the focal spot.
  - Undulator beam source.
  - 1 micron beam, high resolution.
- Techniques used:
  - Diffraction.
  - Micro chemical images.
  - Micro spectroscopy.
- Research interests: MS, Photon Chemistry.
- Available computing resources:
  - Three control workstations.
  - One development workstation.
  - 4 TB on the file server.
  - Quad core, 4 GB RAM processing server.
  - Media station.
- Users usually discuss the needs of installing their software on the computing resources with the beamline scientists.
- Remote access to the beamline computing resources is provided to beamline scientists not to the users.
- The users are not sensitive about their data as long it is available.
- The needed software applications:
  - EPICS tools to monitor the experiments developed using Python.
  - Data viewer software based on MATLAB.
  - IFEFFIT package.
- No need for computing resources after the experiment is carried out.

- The software development in the beamline is collaboration between users' community and beamline scientists.
- The generated experimental data:
  - Rate: 0.5 TB / week.
  - Size ranges 100 KB – 2 MB.
  - No data compression used.

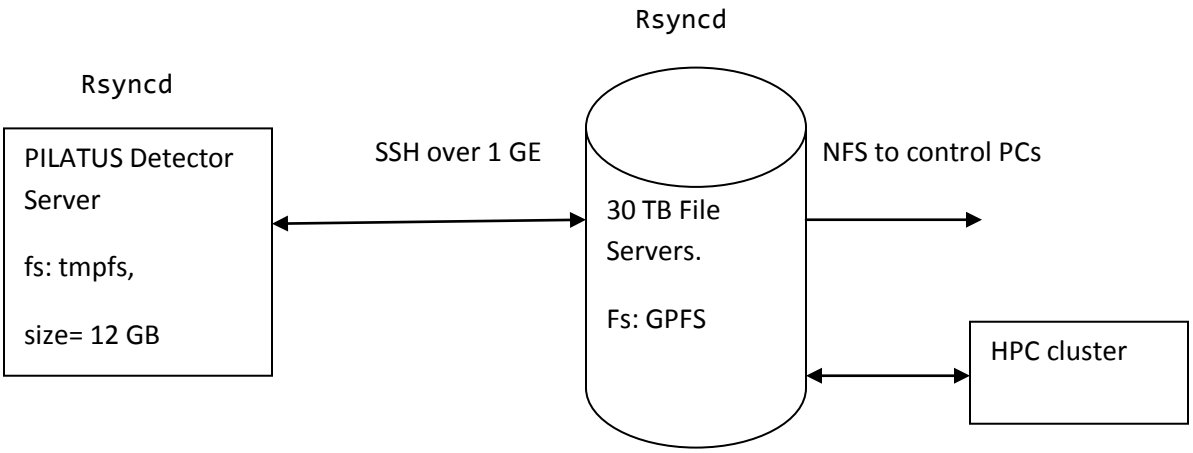
2.2 Meetings in 7th of July 2009

2.2.1 Meeting with PXI and PXII beamlines.

Collected Information:

- End-Station Information:
  - PLITUS 6 M detector.
- The sample change depends on the experiment type, kind of research, and the speed of the detector.
- Provided computing resources:
  - Server that is directly attached to the detector.
  - 8 server nodes used for data reduction and analysis: each 4 core, 8 GB RAM.
  - 30 TB space on the file server.
  - Visualization work station to hard crystallography model: powerful graphic card, 4 GB RAM, and Xeon processor.
  - Normal control PCs: for configuring the experiments, and viewing samples.
  - Monitoring camera.
  - Media station Windows machine.
  - Backup terminal to enable the users manage their experimental data backup.
  - Emergency external disks.
- The server that directly connected to the Pilatus6M detector is part of the detector. It was delivered by the detector's manufacturer Dectris. The connection between Detector and its server is a proprietary hardware using "GigaSTAR" technology; the server hardware is specially chosen to support the GigaSTAR PCI cards. The server specs as follows:
  - 32GB RAM
  - 2x Intel Xeon E5450 @ 3.00GHz CPU
  - 2x 1Gb/s Ethernet Nic onboard
  - Dell Poweredge 1950
  - Linux Operating sytem installed by Dectris, specially tuned towards the Pilatus6M data aquisition task.

- The design of the connection to the storage as in the figure below, rsyncd daemon is used to transfer the data to the file server:



- With the current PILATUS detector at SLS the data can be collected with images at 12.5Hz, which gives about 300MB/s transfer rate. This will change with future Pilatus6M Detector versions.
- SLS computing policy is applied on the beamlines, and the users are not given administrative privileges on the computing resources.
- Remote access to HPC cluster and experimental data is not provided. In case of emergency SLS public FTP is used to transfer the experimental data to the beamline users.
- The experimental data is available for 60 days.
- The generated experimental data is 100 GB/ day in 3 shifts.
- The industrial users are very sensitive about their experimental data, In case of Protein Crystallography at SLS a pharmaceutical company owns one of the beamlines and they bring their IT stuff to monitor the data and transfer it.
- The needed software applications:
  - Clara and Peter: these are in house developed software used to view samples.
  - EPICS clients to setup the experiments.
  - RamDaq to execute the experiments.
  - XDS for analysis.
  - No computing resources needed for offline analysis. The users do the offline analysis from their home institute. Currently it is being discussed at SLS providing a cluster for offline analysis.
  - The programming languages used: Python and PHP to develop the beamline software applications.
- The experimental data description:
  - Number of fram = 16000/day

- Each file size = 6 MB.
- File type= CBF, marccd.

2.2.2 Meeting with MS beamline

Collected Information:

- The research interest is structural analysis.
- Provided computing resources:
  - Two normal PCs for control.
  - Regular Windows server.
  - 10 TB on the file server.
  - MAC PC with 2 quad core CPU, 16 GB RAM: used for calculations and analysis.
- The users have no remote access to the computing resources.
- The needed software applications:
  - FullPROF.
  - TOPAS.
  - EXPO.
  - **http://www.ccp14.ac.uk/**
- The used operating systems only MAC and Linux.
- The used programming languages are: Python, C, and FORTRAN.
- Data Description:
  - Number of generated files: 10,000/day.
  - File types: ASCI.
  - Approximate file size=1.5 MB.
  - The generated data size=2 GB/day.
  - Tar and zip used data archiving and compression.

2.3 Meetings in 8th July 2009

2.3.1 Meeting with SuperXas Beamline

Collected Information:

- End-Station Information:
  - 13-Element Ge-Detector.
    - Manufacturer: Canberra. (http://www.canberra.com/)
  - Single-Element Si-Detector (AV450R10-ZR1BE 139)
    - Manufacturer: Ketek (http://www.ketek.net/)
  - X-ray-eye (high-resolution x-ray-eye / normal x-ray-eye)
    - Manufacturer: self made
  - micro-Ion-Chamber

    - Manufacturer: self made
  - Thin and Thick Diode.
    - Manufacturer: Hamamatsu
  - Ion-Chamber.
    - Manufacturer: Oxford Danfysik (http://www.oxford-danfysik.com)
- Beam Source is Pending Magnet.
- The beamline application: design new material based on the electron distribution.
- In surface diffraction PILATUS II novel photon-counting 2-D pixel detector, consisting of 486 x 195 pixels.
- Techniques used in the beamline:
  - XAS: generates 50 KB/hour.
  - Quick XAS: generates 1 MB/min = 100 GB in week at max
  - Fluorescence: 50 KB/hour
- Available Computing resources:
  - Two EPICS control PCs.
  - Laptop for moving motors.
  - Server connected to the detector.
  - Two windows PCs: to connect to internet and data analysis: the Windows PCs generate plots from the generated txt files from the detector. This is used after the experiment is done.
  - 3 TB on the file server.
- The needed software applications:
  - For data scanning:
    - EPICS tools.
    - MEDM. (GUI for Linux)
    - PEP.
    - Python scripts.
  - For analyzing:
    - Origin 7. (Windows and requires license)
    - Gnuplot.
    - Xmgrace.
- The experimental data is deleted manually every couple of months.
- The scientists have remote access but the users don't. The data is sent FTP in case of emergency.
- For analysis normal PCs are needed in this beamline.
- The development of software applications is collaboration between the beamline scientists and the computing group.
- Data description:
  - Number of files = 100 files / day.
  - Average size = 100 KB.
  - Files type txt.
  - Regular compression tools used: tar, gzip.

- Normally the users bring their own data before coming to the light source.

2.3.2 Meeting with TOMCAT Beamline

Collected Information:

- End-Station Information:
  - Xray absorption using CCD detector.
  - The principle is taking many images while rotating the sample. An Image every one millisecond.
- There are two main difficulties in TOMCAT at SLS:
  - Transferring data from the camera to the server.
  - Parallel read write speed on the storage.
- Available computing resources:
  - Control PCs.
  - 28 node linux clusters.
  - File server 15 TB.
  - Detector server = regular Windows machine, C code written to read the data.
  - Visualization workstation, 2 GHz Xeon, 16 GB RAM, Nvidia Quadro FX1700.
  - Normally the users bring their H.D and laptops.
- In the past it was used to provide remote access to some users to the beamline cluster but they faced at SLS access management and coordination problem on the Linux cluster which degraded the performance then after that the remote access to the Linux cluster is stopped.
- The users of this beamline are not sensitive about their data.
- The needed software applications:
  - C & Python parallel code used for homographic reconstructions "3-D".
  - EPICS and SPEC.
  - AVIZO: used for visualization.
  - C code to read the images from the detector.
  - MATLAP for simulation.
  - ImageJ is alternative open source to AVIZO.
- The Cluster mission is to read the generated data and convert it to several 2-D images. AVIZO then combine the images to generate 3-D image.
- The data description:
  - 6000 files / hour.
  - File size 8 MB.
  - Type is tif.
  - Compression 8-bit tif, 16-bit tcfj.

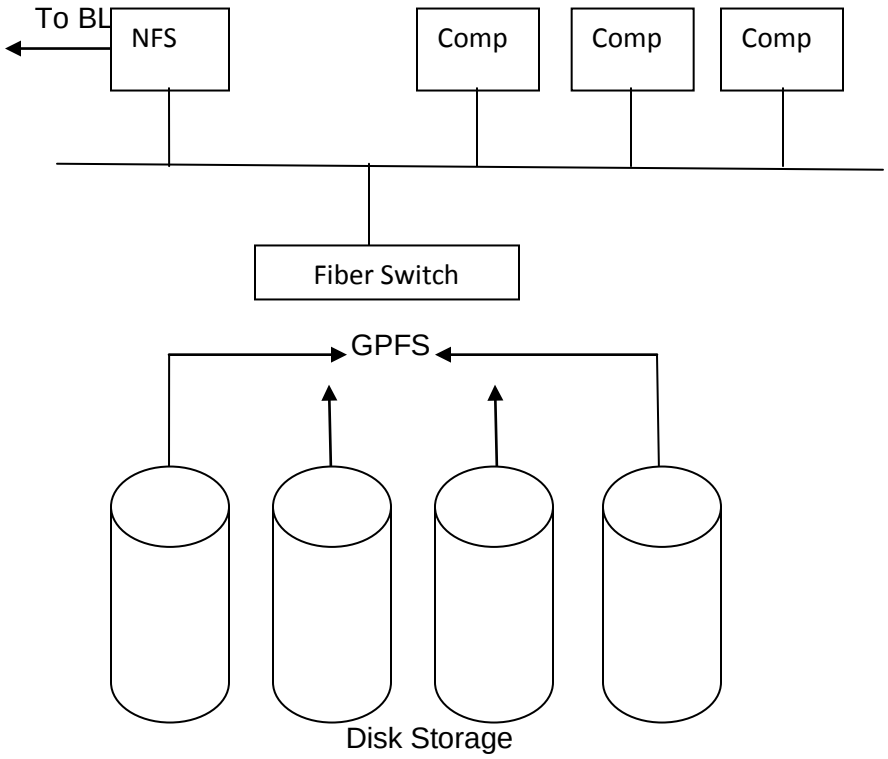
2.4 Meetings in 9th of July 2009

2.4.1 Meeting with Control and Computing

Collected Information:

- The Linux clusters in each beamline are not MPI enabled.
- The central beamline data storage is Xyratex and they are moving to Sun storage.
- The total file server storage is 120 TB RAID 6 shared between the beamlines.
- Each beamline has a separate subnet.
- Currently the offline data analysis is done from the users' home institutes. It is discussed to provide a remote dedicated cluster for offline data analysis and simulation.
- The beamline local network is 1 GE.
- The experimental data is not backed up because it's difficult to manage it and the backup solutions are costly. The risk is put on the users and they should backup the generated data immediately on their external storage.
- GPFS is the adopted parallel file system on the beamline file server storage.
- SLS is moving to adopt hdf5 format on the experimental data in order to store huge number of files, provide fast access to the files, and high compression capability. Important note: some of the beamlines face problems that their applications are not aware of hdf5 so that they cannot deal with hdf5 experimental data.
- The connections to the storage:
  - 1 Gbit/s for the low computing demanding beamlines
  - 5 Gbit/s for the high computing demanding beamlines.
- The problem that TOMCAT beamline has (Parallel read write speed on the storage) it is due that the controller on the disk that is attached to the detector server is malfunction, so that the data is stored outside and this causes the slowness problem.

- The connection to the file server is described in the following diagram:



- Each beamline subnet has its own firewall and different firewall rules based on the beamline requirements.
- The compute nodes access the data on the central storage directly via 4Gb/s Fiber channel SAN / SCSI. Each compute node can reach up to 400MB/s read or write. The cumulated bandwidth for all compute nodes and file server is limited by the Raid storage systems' raid controllers, in a best-case benchmark setup the transfer rate can get close to 1GByte/s total.
- Later 2009 an Infiniband based storage network will be setup. The direct SAN attachment there are some configurations possible with GPFS; other file systems may suggest/require a different setup.
  - DDN (Data Direct Networks) S2A 9900 storage array with direct Infiniband connection is option for the IB network. Another option would be LSI 7900 HPC. Both are large systems which scale to several hundred's of disks and TB in addition they are in use at large HPC and Grid sites. For SLS demands the S2A 9900 is a very large system.
  - Alternatively with both GPFS and Lustre an IB connection can be setup via FC to NSD servers (GPFS) or OSD servers (Lustre) which then serve the storage via Infiniband (or 10GbE) to the compute nodes and

other servers. The benefit is that you can keep the FC/SAN network small, with NSD/OSD server and you can even go without FC switches and just direct-attach. Up to about 64-96 4Gb/s FC ports FC did scale well at SLS site,

  - To grow further, SLS would need to consolidate to some larger (=more expensive) switches, also to get faster SLS would need to move to 8Gb/s, which still is quite expensive.

- In the central storage the disks are 500GB and 1TB SATA disks for the data and a few small fast 15K FC/SAS disks for GPFS metadata.

3 VISIT to CERN

3.1 Presentations about CERN Computing Infrastructure.

Date: 10th July 2009:

- 1st presentation about storage and network at CERN. Duration: 1 hr
- 2nd presentation about CERN Grid and OpenLab project. Duration: 1hr

Collected Information:

- CERN data and computing grid provides the power of 100,000 PCs and 20 PB data storage.
- The computing infrastructure at CERN is classified according to the following tiers architecture:
  - Teir-0: accelerator main computing center.
  - Teir-1: online data analysis
  - Teir-2: simulation.
- Teir-0 computing infrastructure is located in CERN computing centre.
- Teir-1 and Teir-2 are distributed across the countries that participate in the grid project.
- The connection bandwidth between Teir-0 and Teir-1 is 10 GBit/s.
- Glight software is the grid middleware that manages the computing and data grid at CERN.
- The file system used in glight is XROOTD; GPFS can be used in a storage element node in the grid.
- The data storage is managed by software developed at CERN called CASTOR “CERN Advanced STORage manager” used to store physics production files and user files. Files can be stored, listed, retrieved and accessed in CASTOR using command line tools or applications built on top of the different data transfer protocols like RFIO (Remote File IO), ROOT libraries, GridFTP and XROOTD.
- EGEE project is close to finish the project that is responsible for glight development. EGI “European Grid Initiative” EC funded project will adopt glight and take the responsibility to build a grid infrastructure based on glight to serve the European research centers.
- Regarding integrating our computing model with Cyprus we must:
  - Figure out how much data we will transfer.
  - Resources availability.
  - Reliability.
  - Integrating with NRENs is essential in research centers.

- I took a tour in CERN computer center and I found they use many different technologies in servers, tapes, and storage and they have heterogeneous systems (Linux, Unix, Windows) and Windows is mainly used to provide local service.
- At CERN they need high throughput computing system not high performance computing system. Grid is the best design for high throughput computing system.

4 VISIT to ESRF

4.1 Meeting with the Head of Computing at ESRF

Date: 13th of July 2009

Collected Information:

- For managing the beamline users' accounts and their available computing resources there are two adopted systems in synchrotrons which are SMIS (ESRF) and DUO (SLS).
- DUO and SMIT systems are responsible for account creation, disk space allocation and permission setup on the external users' storage zone.
- The data storage for each beamline at ESRF are classified as follows:
  - In house experimental data area 0-8 TB.
  - External experimental data area.
  - Archive area.
- High bandwidth required between some kind of detectors and the storage. In ESRF it is 10 GBit/s.
- MPI and processes schedulers cluster is needed in SR scientific computing work. Generally MPI is needed for simulation software applications and job scheduler needed for analysis software applications.
- ESRF has initiated feasibility study about grid computing to study if it is compliant with SR needs.
- ESRF will bring by end of 2009 a central IBM blade cluster coupled with Infiniband link each node has 4 core CPU and 8 GB RAM. The central cluster has 10 GBit/s connections to the storage.
- The current existed central cluster (NICE) provides 150 CPU.
- The pilot grid project provides 314 CPU.
- Currently a number of SRs in Europe are working on a project called PANDATA to standardize the data policy, data format, analysis software, and authentication. The proposal has been submitted to EC. The proposed experimental data format is hdf5.
- The experimental data retention policy in ESRF is 30 days + 6 months on backup for the visitors. 1 to 3 years + 6 months for in house research.
- The central Storage is NetApp NAS 500 TB GX OS file server. 10 GE is provided to the storage.
- 60% of the NAS (270 TB) is used by actual experimental data (This is a rather low-tide figure, as summertime is typically a low-activity period.). The remaining space is mostly used by general-purpose data (home directories, software) or specific filesystems (source code, databases...).

- Half of beamline data is used by in-house research. This data kept as long as users want (a reminder about old data is sent out periodically to help users cleaning up).
- The other half is used by visitors. This data is kept typically one month after the stay of external users, in order to let them analyze and transfer the data back to their home institute.
- Backup is done on daily bases and the backup retention period is at least 6 months. (Around 9 months in most cases, depending on technical constraints.) Currently, they have 800 TB backups for in-house data and 700 TB for visitors.
- NDMP is not used to backup the NAS storage due to data restoration constraints: they usually perform rather frequent restorations of selected files/directories (typically due to user errors), and NDMP used to be very inefficient in this task as evaluations at ESRF has been made a few years ago about this issue.
- The archiving system has the following components:
  - Six backup servers use a total of 38 LTO tape drives (10 LTO-3 and 28 LTO-4)
  - Two servers have each a dedicated L700 tape library (700 slots). Both tape libraries are almost full.
  - The other servers are sharing a single L8500 tape library (8500 slots configuration, 5000 slots licensed at the moment).
- Beamline data backup represent a total of over 1300 tapes (the volume is still growing due to retention periods). 500 additional tapes are needed in the near future.
- To protect data against potential disasters, the backup server and its tape library is systematically located in the other building (off-site backup) than the NAS. Private interconnections with 10 Gbit/s between the NAS storage and its backup server(s).
- GPU is important for visualization and it saves electricity and it is faster to handle the computing graphics problems. GPU is important in Tomography beamline.
- SSL gateway is used to authenticate the remote users to connect to the central cluster to perform analysis and simulation.
- The experimental Meta data is very important to track the data. Hdf5 will enhance keeping track to the experiments meta data. It is suggested that to hide the data you can hide the meta data, this is important issue in case we want to store the data remotely for instance in Cyprus and we can store the meta data in house in case we have users sensitive about the data.
- TANGO control system is adopted at ESRF.
- There is tens of software applications installed on the central clusters and need a periodic maintenance.

- It is difficult to make distinction between the online and offline computing requirements in SR and it depends on each beamline.
- Archiving the data takes huge amount of storage and this can be easily stored outside SESAME.
- Software parallelization is needed and used in Tomography analysis applications.
- At ESRF each beamline has a subnet 160.103.net.host
- There is no need to Windows HPC.
- At ESRF they have their own Linux distro, the reason behind that is to control the release versioning and to bundle their distro with the required packages.
- There is dedicated directory service based on OpenLDAP used for beamline authentication.
- We need continues budget for IT to renew technologies every 4 years. And our IT resources will depend on how much money we will spend on detectors technologies.
- In ESRF they have two IT departments, one is under the scientific sector and other responsible for managing all the services and there is collaboration between the two teams. There are pros and cons in hierarchy approaches to have one IT department or two IT departments but the key is we have to manage it well and both approaches can success.
- ESRF has 600 employees and 7000/year visitors. And the IT staff in ESRF organized as the following:
  - 20 engineers for beamline: managing beamline control, users support internal servers. ESRF has 40 beamlines
  - 10 scientific programmers.
  - 10 controls programmers: responsible for beamline and accelerator.
  - 10 MIS: database programmers, SMIS development, local management software.
  - 10 desktop support engineers for ESRF staff.
  - 10 center of IT: network, email, web, storage, backup, HPC, etc...)

5 VISIT to SOLEIL

5.1 Meetings in 15th July 2009

5.1.1 Meeting with Control and Computing Group.

- **1st session with head of group and the control team.**

Collected Information:

- SOLEIL central cluster"Fusa" which serves the beamlines HPC current demands provides 120 Xeon CPU.
- The experimental data storage format is transformed to Nexus data format that is used to standardize the data. The benefits of using Nexus format:
  - Ability to store any kind of data.
  - Self-describing format in order to be able to maintain the meta data.
  - Efficiency, which provides high compression and access speed to the data.
  - Adopted in other facilities: DIAMOND, APS, LLB, ISIS, SNS, PSI, ESRF
- For remote access to the experimental data TWIST application is used to access the data via FTPS and HTTP. TWIST is compliant with Nexus.
- DUO is the adopted beamline users' management system.
- TANGO is the adopted control system.
- In each beamline there is 2 servers used for TANGO.
- The application GlobalSCEEN is used to develop applications to talk with TANGO devices.
- Data Acquisition software used is SALSA connected to the scanning device to monitor the work on the beamlines components.
- PASSERELLE software application is used to develop GUI acquisition sequences from the TANGO device.

- **The 2nd session with storage and servers administrator at SOLEIL.**

Collected Information:

- Central HPC clusters are used for scientific computing of beamlines work, below are the summary:
  - 63 nodes, 6GB RAM, internal 1 Gbit/s connection, AMD Opteron CPU, Linux OSCAR using NEC "OSCAR PRO". It provides total power 1 TFLOPS, total RAM 496 GB.
  - 22 nodes, total RAM 464 GB, Xeon CPU 1.5 TF.

  - New proposed cluster: 112 nodes, 8 core Intel NEHALEM CPU, 36 GB RAM each. With Master server SMP 72 CPU 144 GB RAM. This system provides 11 TFLOPS.
  - The central clusters support OpenMPI and Torque/Mavi(PBS) process scheduler.
  - Luster distributed speed filesystem is used.
  - Important note: The memory access and I/O access are important factors we should take care of when designing our Linux cluster, at SOLEIL the 1st cluster with 1 TFLOPS has a better performance from their 2nd cluster with 1.5 TFLOPS because AMD Opteron has a faster memory access than Intel Xeon.
- The adopted Linux distribution is Red Hat.
- There is no need for Windows HPC.
- OpenLDAP is the directory service in the machine and beamlines networks. Microsoft AD is used for the office network only.
- For data DR plan ACTIVE CIRCLE is used.
- The data center resources are divided into two datacenters in two separate buildings.
- The data storage design is as the following scheme: each beamline has its own data storage server inside the beamline local subnet network and it is considered the entrance point to the distributed file system and these internal storage servers are connected to central storage SAN which is 300 TB total separated 150 TB in each building. The archival and backup tape Library size is 2.7 PB spanned over the two buildings.
- Skills needed in the teams: linux/unix administration, storage management, networking, and HPC skills.
- Regarding the life time, the hardware should be setup before 6 months from the commissioning to be ready and take an enough time for testing.
- Points to be taken into account while integrating with Cyprus:
  - Network access. How much the data will be transferred?
  - Guarantee the resources availability on demand and have a full management on the HPC resources.
  - What are the alternative HPC systems?
  - Maintaining the scientists' software applications on the remote HPC in Cyprus.

5.1.2 Meeting with CRISTAL beamline.

Collected Information:

- End-Station Information:
  - Diffractometer 4-Cercles (Newport)
  - Diffractometer 6-Cercles (Newport)
  - Diffractometer 2-Cercles (SMP)

- CRISTAL beamline uses the technique X-ray scattering, and there are no needs for high computing resources in this beamline.
- Available computing resources:
  - One data storage server with capacity 290 GB.
  - Two servers for TANGO control system.
  - Two Terminals for remote access.
  - Three Normal PCs used for instruments server, Windows applications use, and dedicated for the Diffractomètre
- The users are not sensitive about their experimental data and it's not very confidential so they have no problem where to store their data as long it is available.
- The needed software applications:
  - Capoiera: this is java based application to drive the diffractometer.
  - SALSA.
  - Passerelle
  - Mambo: collect the experiment parameters, and extract the data.
  - CrysAlis: this is windows application used for Oxford diffraction to analyze the diffraction image.
  - Bensikin: take snapshots of the parameters.
  - Origin 7: Windows application provides a data analysis and graphing workspace for beamline users.
- They don't use the central cluster.
- The amount of produced data is 10 GB/ 3days.
- Experimental data retention period is one month.
- Number of generated files varies from tens to thousands this depends on the experiment and the instruments used.
- Tango device recorder translates the files from binary, txt, and tif format to nexus format.
- The approximate file size is 1 MB.

5.2 Meetings in 16th July 2009

5.2.1 Meeting with PROXIMA beamline

Collected Information:

- End-Station Information:
  - Currently the CCD detector is used and its proposed to buy PILATUS 6M detector.
  - The beamline is equipped with a large surface area detector (ADSC Q315r)
- There are two ways to collect data in protein crystallography:
  - Rotation on some angle: this is the normal process of data collection and currently it's the only way used at SOLEIL.

  - Shutterless data collection: which is based rotation the crystal sample and taking a snapshot every millisecond. This method can be carried out using PILATUS detector technology at the time being only.
- The computing workflow of a single experiment comprises the following three steps:
  - Setup and control the experiments. Setup monochromator , shutter, focusing, energy, etc…
  - Data Collection: images files are transferred to the cache server using programs written in high level language. The cache server is connected directly to the detector.
  - Data Processing and Analysis:
    - after the diffraction images taken from the detector, data analysis software applications like XDS process these images files to generate txt files contains the " H K L F" representation of the sample.
    - Then a heavy crystallographic atom analysis is run to produce the electron density map. The applications like CCP4 and PHENIX carry out this step.
    - Graphics Visualization applications used to produce the 3D model of the sample.
- Available computing resources:
  - Two control PCs.
  - 4 TB RAID array connected to the cache server and it is NFS shared to the local servers. Note: this storage has a very fast access and it's used for data collection and analysis. The storage server in this beamline that is allocated by the computing group is slow.
  - Two servers called SGE1 and SGI2 that do the data analysis. The servers have 8 processors each. These servers old and they are going to be replaced.
  - Visualization server "Proxima1": Dell XPS Intel i7 processor 940, 6GB Ram, 1 TB disk.
  - Data analysis Server "Proxima2": Dell XPS Intel i7 processor 940, 6GB Ram, 1 TB disk.
  - The current computing resources are not enough and they will replace the 4 TB RAID and the SGIs servers with a new server with 20 TB RAID and 32 cores. This new server will be NFS shared to Proxima1 and Proxima2. Its proposed to buy a new Dell XPS to be Proxima3.
- PROXIMA beamline does not connect to the central cluster because as the scientist reported that they cannot be on the cluster queue and they need these resources immediately on demands.
- The needed software applications:
  - Data integration:
    - MOSFLM.
    - XDS.

  - Data Processing:
    - CCP4.
    - SOLVE.
    - PHENIX.
    - SHARP.
  - Visualization:
    - COOT.
    - PYMOL.
    - RAS MOL.
- TWIST is not used for remote data access because the generated files are not in Nexus format.
- Usually the users are sensitive about their experimental data.
- On the beamline local storage there is no data security. All Data are stored into a disk where all users have read access (with the exception of industrial users). Users' data is hidden just by cryptic data set names. The large generated experimental data and the beamtime schedule is not made public (no one knows when their competitors have beamtime).
- The online computing requirements are high for data analysis and processing.
- The generated data is 85 GB / day after compression it will become 50 GB.
- File size 18 MB
- The files format is binary and txt.
- Number of files is 4,000,000 file.

5.2.2 Meeting with DIFFABS beamline

Collected Information:

- I found that this beamline has no high computing demands, and the meeting was shorten because the beamline scientists were busy with users.
- Techiques used in this beamline:
  - Diffraction.
  - Absorbtion.
  - Fuorecence.
- The available computing resources:
  - Three terminals to control the beamline.
  - Two normal PCs "dual boot" are used by users to access internet and access the experiemental data.
  - Dont use the central cluster.
  - The amout of generated data after each experiement is 1 GB.
- The generated file size is 12 MB.
- Applications used:

- SALSA for data aquision in diffraction teqnique.
- TUMBA used for data aquisition in absorbtion and fourecence teqnique.
- TUMBA saves data in Nexus or ASCI format.
- For analysis the package IFEFFIT used.

5.3 Meetings in 17th July 2009

5.3.1 Meeting with Computing Group: Network administrator

Collected Information:

- The core network has 720Gbit/s bandwidth capacity.
- I was provided with the network map.
- VSS Cisco technology is proposed to be setup between the two datacenter buildings to provide high availability.
- In each beamline there is 3-rd layer switch to provide high availability for each beamline in-case the connection is broken with the core network.
- Each subnet is divided into two vlans: RCL vlan used for control and instruments, REL vlan used for servers and users.
- SOLEIL has 1 Gbit/s link with France NREN “RENATER”.
- The general services servers like LDAP and DHCP are attached to the beamline main router.
- For beamline network they use one class C public IP address sub netted with mask /27.
- In each beamline RCL, one private class C IP network address assigned to RCL VLAN which provides 254 host addresses.
- In each beamline REL network one public class C subnet /27 addresses is assigned.
- REL IPs mainly assigned for beamline internal server and users’ laptops.
- No outbound or inbound traffic is allowed from internet from/to RCLs.
- Tango ports are allowed from the machine network to RCLs.
- NFS and CIFS connections are allowed between REL and RCL in each beamline.
- The following ports are open between REL and main router: DHCP, Squid, LDAP, NFS, samba, NTP, printer, and SSH.
- SFTP and HTTP ports are allowed from internet to access TWIST server.
- The network bandwidth between REL and the central cluster is 1 GBit/s.
- Remote access to SOLEIL network is provided by:
  - VPN to the office network. Using ACS Cisco.
  - SSH relay used on the office network to connect to REL.
- The central HPC cluster is opened for all the remote users if they provide their MAC address for the security on the ssh relay.

5.3.2 Meeting with SAMBA beamline

Collected Information:

- End-Station Information:
  - Ionisation chambers.
  - Multi-elements Germanium Detector.
  - Total electron yield.
- Absorption spectroscopy technique used.
- The experiment computing workflow:
  - Setup the experiment beamline instrument and devices: this is done by applications written in Python.
  - Collecting data: using PASSERELLE.
  - Analysis and Simulation.
  - XANES calculation: this requires high computing resources.
  - EXAFS calculation.
- The heaviest computing part is the analysis and simulation. The central HPC cluster is used for this purpose.
- Normal PCs used for control.
- For remote connection to the beamline experimental stations they use ssh, and NXCONTROL.
- The users in this beamline are not sensitive about their data.
- The needed software applications:
  - For control: Python application “Pysamba”.
  - TANGO device is used to monitor the status and collect data.
- Data analysis and simulation:
  - For EXAFS method: HORAE, CHEOCKEE, GNXAS, VIPER.
  - For XANES method, the Monte Carlo simulation applications: FEFF8, MXAN, CONTINUUM, and FDMNES.
- The beamline scientists rely on them self in writing their SW applications. The scientist mentioned that they need a software engineer to guide them in writing software applications.
- The experimental data produced is 2 GB/ year and its expected to be increased to 10 GB/ year.
- The internal storage file server is 200 GB and 1 TB allocated on the backup server.
- Number of files can be approximated it depends on the experiment.
- The approximate file size of the experimental data is 100 KB.
- The access to the central cluster is not provided from the beamline. Where it's preferred to have the remote access.

6 COMMON FINDINGS at SLS and SOLEIL:

Common Answers from SLS and SOLEIL:

- The computing resources that the users bring with them are mainly their laptop and their external data storage that contains their previous experimental data or it is used to take a backup of their experiments results.
- Regarding the IT policy on the beamline IT resources I found it is complaint with IT computing policy of SR and every beamline has its own subnet and firewall rules.
- The beamline users have minimal administrative privilege on the computing resources.
- Remote access to HPC is not provided to beamlines users at SLS while at ESRF and SOLEIL the remote access is provided. This point is discussed currently at SLS and considered as a hot issue.
- The software applications are developed by the scientists and the computing group. Some of the applications are from the users' community. Few property Windows software applications are used mainly for visualization and few data analysis.
- The Industrial users are very sensitive about their experimental data. Especially the pharmaceutical factories in Protein crystallography beamlines.
- No requirements for Windows HPC facility.
- DUO is the beamline users' management tool.

SLS:

- PILATUS detector is used in a number of beamlines.
- The beamlines that require high computing resources are: cSAXS, PXI, PXII, and TOMCAT.
- There is no shared computing cluster at SLS but each beamline has its own processing servers for analysis.
- The central file server for beamlines is RAID6 120 TB.
- The experimental data is not backed up.
- The beamlines computing clusters are not MPI aware.

SOLEIL:

- PROXIMA is planning to purchase PILATUS detector. But I didn't find PILATUS detector in the other beamlines.
- Central HPC and file server is provided and available for beamlines. The computing resources are distrusted in two separate disaster recovery buildings.
- The Central clusters are MPI enabled.

- PROXIMA and SAMBA beamlines have high computing demands.
- Most of the experimental data is formatted in Nexus format. And the remote access to the data is only provided to Nexus data files.
- High available beamline unit design is implemented: storage servers, control servers, computing cluster and 3rd layer switch in each beamline station.

7 CONCLUSIONS & RECOMMENDATIONS:

1. Interactivity is essential for beamlines so that the online computing demand is high.
2. The computing workflow for beamlines is classified according the following three steps:

    i. Experiment setup: This step is before the experiment start and requires light computing resources and basically it includes control system tools.
    ii. Data acquisition: This step requires high computing resources especially in terms of storage and connection bandwidth to the detector. At this step computing resources must be available online while the experiment is running.
    iii. Data analysis and simulation: This step needs computing power to accomplish the complex calculations. For some beamlines at this step computing resources are needed online and for other beamlines it is possible to wait until the experiment is completed.

3. Not all X-ray beamlines have high computing demands; there are many factors to consider in this regard: detector type, kind of research, the needed algorithms and analysis methods to accomplish the experiments. The most intensive computing demanding beamlines are Protein Crystallography, SAXS, and Tomography beamlines.
4. Most of the beamline users perform their offline analysis from their home institute. This step is not fully clear to me and how much computing resources are needed. For this particular scenario we need to get in touch with the SESAME SR users' community to figure out the needed computing resources.

- **Recommendations**

1. Our computing model and needs will strongly depend on our proposed detectors' technologies and endstations. We should have an adequate computing infrastructure to accommodate the rapidly evolving technologies in beamlines' detectors.
2. We should take care of the all necessary steps in computing and make sure no bottlenecks are there that could negatively affect the whole experiment workflow.
3. In our computing design we should consider the follows:
    i. Central computing resources and early designing. To avoid the problem that SLS is having that they have distributed clusters for each beamline and they think now to have central HPC.
    ii. Our HPC should be MPI enabled and job scheduler aware to cover all scientific software applications needs.
    iii. For the disaster recovery setup, we need to have two datacenters in separate buildings.
    iv. Adoption of hdf5 experimental data format, which is used in ESRF and SOLEIL. SLS is moving to hdf5 files format now. In this context we need to check out the compatibility of EPICS and the scientific applications with hdf5.
    v. Remote access should be provided to the HPC facility and experimental data. The remote access for the data should be provided regardless of the data type to avoid the problem at SOLEIL with TWIST.
    vi. Our computing design should be scalable in order accommodate the detectors technologies in future.
    vii. Experimental data backup is important and should be designed carefully.
4. Points to consider when integrating with Cyprus:
    i. Tier-0: This is in-house located resources, which will cover the three computing steps I mentioned above.
    ii. Tier-1: Used for moving the experimental data backup, secondary HPC to fail over the local HPC, and offline remote users analysis and simulation jobs. Note: this could attract the beamline users by providing a remote HPC facility to run their analysis from their home institutes.
    iii. The connection speed is essential factor and will control our integration design with Cyprus HPC facility.
    iv. Maintaining the installed software applications in the remote HPC.
    v. Resources availability and manageability.
    vi. To solve problems with users that are sensitive about their data we could think of a design that hides the Meta data and store it

locally, or simply we don't not send the experimental data of industrial users remotely and keep the data locally.

For the better definition of SESAME computing model, we need a clear understanding of the resource allocation at CyI dedicated for SESAME. This implies knowing today the available CPU power, disk storage and manpower, as these resources will be taken into account when determining the overall SESAME computing needs. A software repository related to synchrotron applications should be maintained at CyI that will require at least hiring one dedicated software engineer at CyI.

As CyI will be a service provider for SESAME, a 24/7-complaint centre must be managed at CyI.

There is a talk in Europe today for the creation of a Virtual European Synchrotron Radiation Data Analysis Centre. This is an important development and should be noted by CyI and SESAME.

As the planning of HPC at CyI has a direct bearing on SESAME, we need a well-defined mechanism that the SESAME directorate is directly involved in this planning.

8 REFERENCES

- SLS: Swiss Light Source http://sls.web.psi.ch/view.php/about/index.html, Villigen, Switzerland.
- CERN: European Organization for Nuclear Research www.cern.ch Meyrin, Switzerland.
- ESRF: European Synchrotron Radiation Facility www.esrf.eu Grenoble, France.
- SOLEIL Synchrotron: www.synchrotron-soleil.fr Paris, France.

Appendix 1: VISIT COMPUTING QUESTIONNAIRE

Part I – beamline Scientists:

1. Collect the beamline technical specification.
2. How frequent is the sample change in a single experiment for this beamline experiments?
3. What is the source of the beamline:
    - Magnet.

- Regular.
- Undulator

4. What are the research interests?
5. What is the experiment workflow? How do you book a place in the beamline? What are the computing equipments that the LS provide you? What are the computing equipments that the scientists get with them?
6. What is the policy of using the computing resources in the SR?
7. Do you need remote access to your applications or data? If yes why do you need it? How often do you connect?
8. Are you willing to store you experiment data in safe remote data storage outside the SR? E.g. storing the data in a data grid project.
9. What are the software applications do you need to carry out your experiments?
10. What are the applications and codes used for simulation, data processing, mining and analysis?
11. What are your needs to computing resources while the experiment is running and the needs to computing resources after the experiment finished? For how long can you wait till you get the results from the computing application in each case?
12. What are the SW and HW requirements for the applications? Provide us the applications data sheet if available.
13. Who is responsible for developing the scientists' applications?
14. How much is the total needed data storage?
15. How experimental data is sent / transported to users' home institute?
16. Describe the data generated from the experiments of this beamline?
    - Approximate Number of files?
    - Files type?
    - Files sizes' range?
    - Do you use data compression? If yes what is the compression mechanism?
17. Do you obtain and update data from remote data store? If yes what is the data size and the required bandwidth to the remote data store?

18. Attending a live demo from the beamline scientists about how they use the computing resources during and after the experiments.

Part II – Control & Computing Group:

1. Taking a tour in SLS computing infrastructure.
2. Discuss the beamline network infrastructure design.
3. What are the HPC system types and levels?
4. What is the total memory requirement in each level?
5. What is the amount of submitted jobs to the HPC facility in terms of kjobs/day?
6. What is the online CPU requirement in terms of FLOPS? "while the experiment is running"
7. What is the CPU requirement in terms of FLOPS for the development environment and tools?
8. What is the needed CPU requirement in terms of FOLPS for data processing and analysis after the experiment is carried out?
9. Discuss the needs of multiplatform HPC facility.
10. Discuss the adoption of Linux distribution in SLS. And the needs of having our own Linux distro.
7. What are the applications and codes used for simulation, data processing, mining and analysis?
8. What is the data storage design? And how do you classify the data in your data storage?
9. How much is the required bandwidth between the high performance computers facilities and the data storage?
10. How long is the data retention period for each data class?
11. Do you obtain and update data from remote data store? If yes what is the required bandwidth to the remote data store?
12. How much is the archival data capacity requirement? And what are the backup retention types and periods?
13. Discuss the storage DR and high performance design.
14. Describe the remote access of the beam line users to the beam line station?

15. Do you provide direct remote access to the data? If yes how is it implemented?
16. What is the internal bandwidth needed between the end station and data storage?
17. How do you authenticate experiment users on beamline computers?
18. How do you guarantee data privacy between beamline users?
19. What file services (NFS, SAMBA, FTP, etc) do you provide for beamline users?
20. What are the computing staff requirements? Skills, human power, and on call working hours …etc.
21. What are the staff duties and job description?
22. Discuss the technology life time issue.
23. Discuss our proposed computing model that is integrated with Cyprus Institute HPC.

Appendix 2: PERSONS MET

SLS:

- Department: cSAXS beamline
  - Contact: Oliver Bunk (oliver.bunk@psi.ch)
- Department: MicroXas beamline
  - Contact: Daniel Grolimund daniel.grolimund@psi.ch
- Department: PXI, PXII
  - Contacts:
    - Anuschka Pauluhn [anuscka.pauluhn@psi.ch]
    - Exequiel.Panepucci@psi.ch]
  - Department: MS
    - Contact: Phil Willmott philip.willmott@psi.ch
- Department: SuperXas
  - Contact: Evegueni Kleimenov (evegueni.kleimenov@psi.ch)
- Department: TOMCAT
  - Contact: Rajmund Mokso (rajmund.mokso@psi.ch)
- Department: Control and Computing
  - Contacts:
    - Rene Kapeller (rene.kapeller@psi.ch)
    - Heiner Billich (heiner.billich@psi.ch)

CERN:

- Alberto Pace [Alberto.Pace@cern.ch] conducted the presentation about Network and Storage. Duration 1 hour
- Sverre Jarp [Sverre.Jarp@cern.ch] conducted the presentation about HPC and Grid Computing. Duration 1 hour.

ESRF:

- Contact details:
  - Rudolf Dimper [dimper@esrf.fr]

SOLEIL:

- Department: Control and Computing.
  - Contacts
    - Brighitte Gagey [brigitte.gagey@synchrotron-soleil.fr], Head of Computing
    - MARTINEZ Philippe [philippe.martinez@synchrotron-soleil.fr], HPC and Storage.
    - Contact: GATTONI Pascal [pascal.gattoni@synchrotron-soleil.fr], Network Infrastructure.
- Department: CRISTAL beamline.
  - Contact: Pierre Fertey, pierre.fertey@synchrotron-soleil.fr
- Department: PROXIMA beamline
  - Contact: Andrew Thompson andrew.thompson@synchrotron-soleil.fr
- Department: DIFFABS
  - Contact: REGUER Solenn Solenn.reguer@synchrotron-soleil.fr
- Department: SAMBA
  - Contact: Emiliano FONDA emiliano.fonda@synchrotron-soleil.fr