European XFEL DAQ and DM computing Technical Design Report – 2009 public version.

Editors: S.Esenov, K.Wrona and C.Youngman

12.10.2009

This document describes the proposed computing model for the initial phase of operation of the European XFEL (EuXFEL) facility. In recognition of the overlap and interdependence of online data acquisition (DAQ) and offline data management (DM), both online and offline handling will be reviewed. The DAQ scope covers the control and readout systems in the photon beam lines, experimental hutches, laser and computing rooms in the experimental hall, and interaction with the electron machine control system. It does not include office computing requirements. The DM scope covers offline systems downstream of the DAQ required for secure on-site data storage and analysis, and remote access to the data. Specific DAQ and DM hardware and software solutions and the IT services required, e.g. networking, access control, etc., by both are described. The scale, manpower and financial cost of the system based on the current understanding of EuXFEL operation are estimated. The understanding of EuXFEL operation will evolve continuously and will require modification of this note. **This version of the DAQ and DM computing TDR is for public use and Appendix B, which contains cost, manpower and time line estimates, has been removed.**

1 How this note is organized

The content of individual chapters and appendices is summarized below. EuXFEL users who do not have time to read the entire document should read Chapter 9.

Chapter 1, How this note is organized
> This chapter!

Chapter 2, Overview of EuXFEL environment
> This chapter gives a brief overview of EuXFEL machine parameters, naming conventions for tunnels, and the anticipated beam to experiment class association.

Chapter 3, Control and readout devices
> This chapter provides an overview of the different photon beam line detector and instruments types that have to be controlled and readout. The data sizes described are used in the experiment requirements chapter.

Chapter 4, Experiment requirements
> This chapter contains showcase studies of experiment DAQ and DM requirements. The yearly anticipated data volumes and estimated instantaneous bandwidths are used in the total cost estimates derived in Appendix B.

Chapter 5, Data acquisition architecture
> This chapter defines the architecture to be used by all detectors and instruments in the DAQ. The 2D pixel detector DAQ and control system is described as a showcase, but it is representative of all detector classes discussed in chapter 3.

Chapter 6, DAQ software
> This chapter outlines the software implementation envisaged for the DAQ systems. The principle thrust of this development is to implement the control layer using standard IDE development tools and recent software developments.

Chapter 7, Data management overview
> This chapter reviews the storage, analysis and software requirements placed on the DM system. These include file formats, software management, and authentication, authorization and accounting schemes. Initial concepts for data storage and analysis are described as well as listing data management policy rules for interacting with the data.

Chapter 8, Data management architecture
> This chapter defines the architecture to be used for data storage and offline analysis which take into account the DM requirements defined in chapter 7.

Chapter 9, Summary
> This chapter contains a summary of the note content which should be useful to EuXFEL users who want to get an idea of what will be provided.

Chapter 10, Glossary
> This chapter contains a list of used acronyms and there meanings.

Chapter 11, Acknowledgements
> This chapter acknowledges those who have contributed to the note.

Chapter 12, References
> This chapter contains a list of references to other information sources.

Appendix A, DAQ and control infrastructure

This chapter contains information plans for network and room usage in the tunnels and experimental hall. It is useful to get a spatial understanding of where various systems will be placed.

Appendix B, EuXFEL DM and DAQ cost, time and manpower estimates

This chapter provides detailed information regarding costing, time lines and manpower estimates. **This appendix will not be made available on the web; it should be requested if required.**

2 Overview of EuXFEL environment

This section reviews those features of EuXFEL which are relevant to the computing model design: photon bunch delivery characteristics, building and tunnel naming conventions, lists of devices, etc.

2.1 Photon beam delivery parameters

The nominal electron bunch timing pattern produced at EuXFEL is shown in Figure 1. The data acquisition systems being designed are required to handle two frequencies, the 5MHz bunch rate and the 10Hz train rate. The number of bunches per train, 3000, is driven by the 600μs RF flat top of the superconducting cavities.
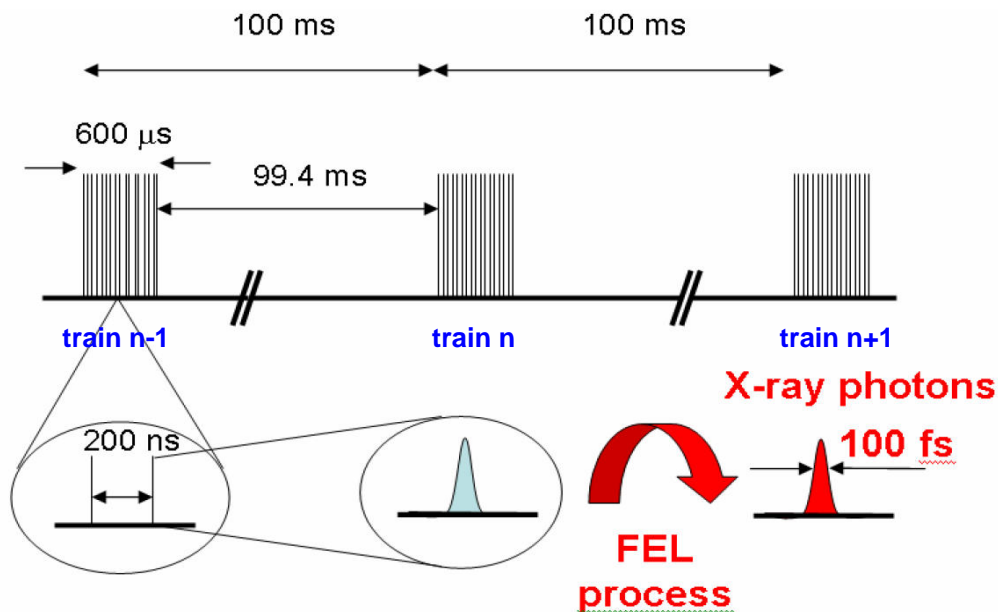
Figure 1 Bunch and train timing at EuXFEL

The beam distribution scheme is shown in Figure 2. The electron bunch train is split into two, by a flattop kicker magnet, and deflected into SASE1 and SASE2 beam lines. A fast kicker is used to dump single bunches during the 20μs flattop field transition period. The two electron trains generate photon bunches in undulators (blue) and are then deflected into additional beam lines with undulators, where the electron bunches are reused to create additional photon bunch beam lines. With the chosen distribution scheme five photon beam lines can be supplied with photon bunches concurrently.
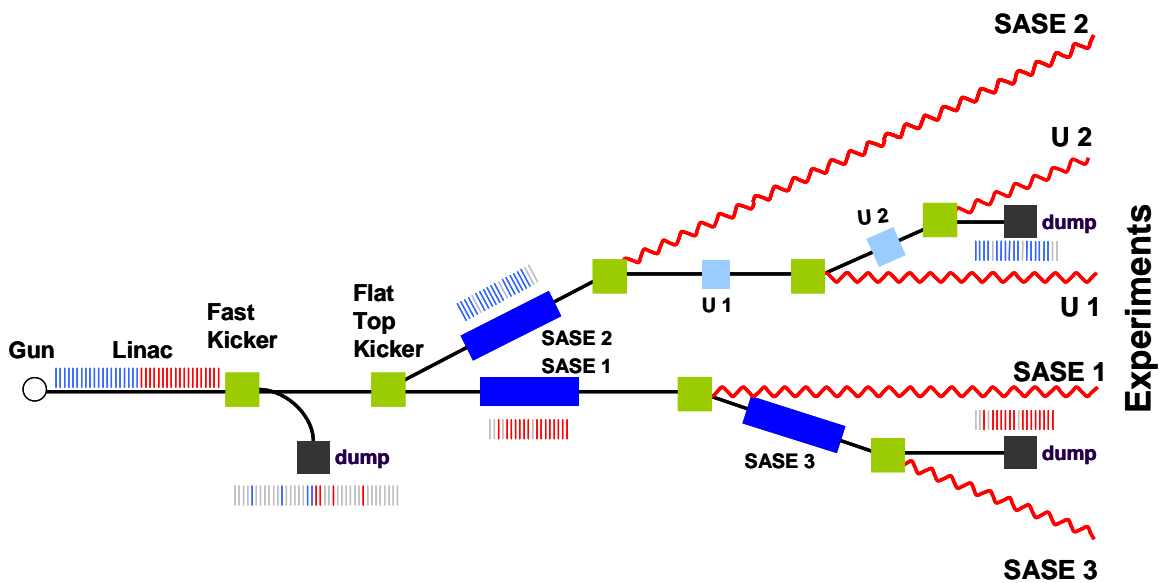
Figure 2 Beam distribution at EuXFEL

The pulse pattern seen by detectors fed by SASE1 and SASE3 are identical, as are those in the SASE2, U1 and U2. The initial train splitting and final electron beam dump absorption limitations mean that ≤1500 photon bunches are seen per beam line. The pattern of filled bunches, see Figure 1, in a train can be modified by firing the fast kicker. Reprogramming the electron gun to produce empty bunches is not likely as this can introduce instabilities into linac operation.

The train repetition rate can be changed within the range 10 thru 30Hz. Increasing the repetition rate may require reducing the beam energy, or reducing the number of bunches per train, or a combination of both. Changing the repetition rate on the fly, between consecutive trains, is not possible. The 5MHz bunch repetition rate cannot be changed, but other rates, e.g. 1MHz, can be produced by not creating or removing bunches.

DAQ relevant EuXFEL parameters are summarized in Table 1.

Parameter	Nominal value	Range
Train rate	10Hz	$10\text{Hz} \leq \text{f} \leq 30\text{Hz}$
Bunch rate	5MHz	fixed
Filled bunches/train/beam line	~1500	≤1500
Beam lines	5	fixed

Table 1 DAQ relevant EuXFEL parameters

2.2 Beam line and experimental hall naming conventions

The tunnel and building naming convention used in this document is shown in Figure 3. Tunnel sections XTD6 thru XTD10 correspond to SASE2 thru SASE3 in Figure 2, respectively. Tunnels and buildings associated with XTD20 and higher relate to a potential upgrade which is not discussed in this document.
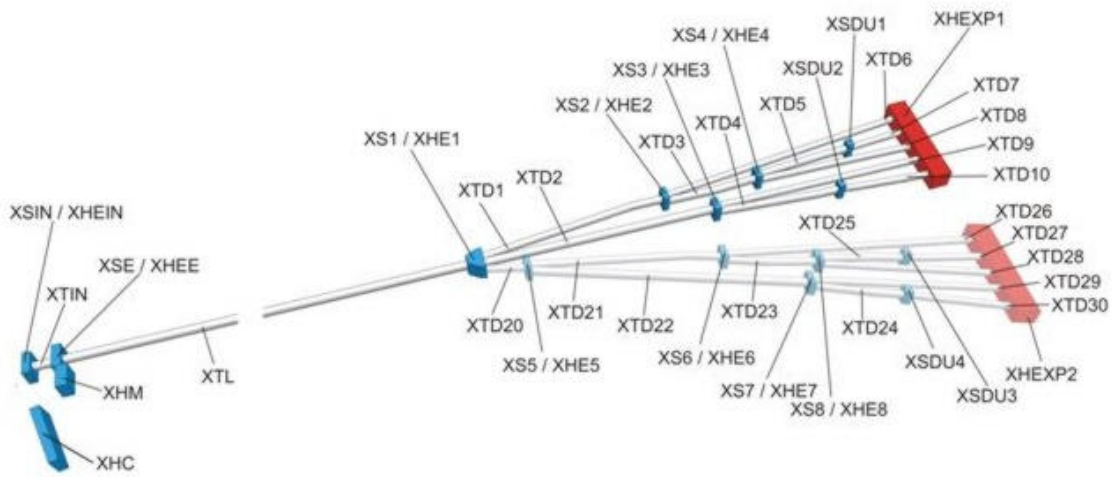
Figure 3 Beam line naming convention

Experiment end stations are located in the underground level (UG01) of building XHEXP1, the floor plan is shown in Figure 4.
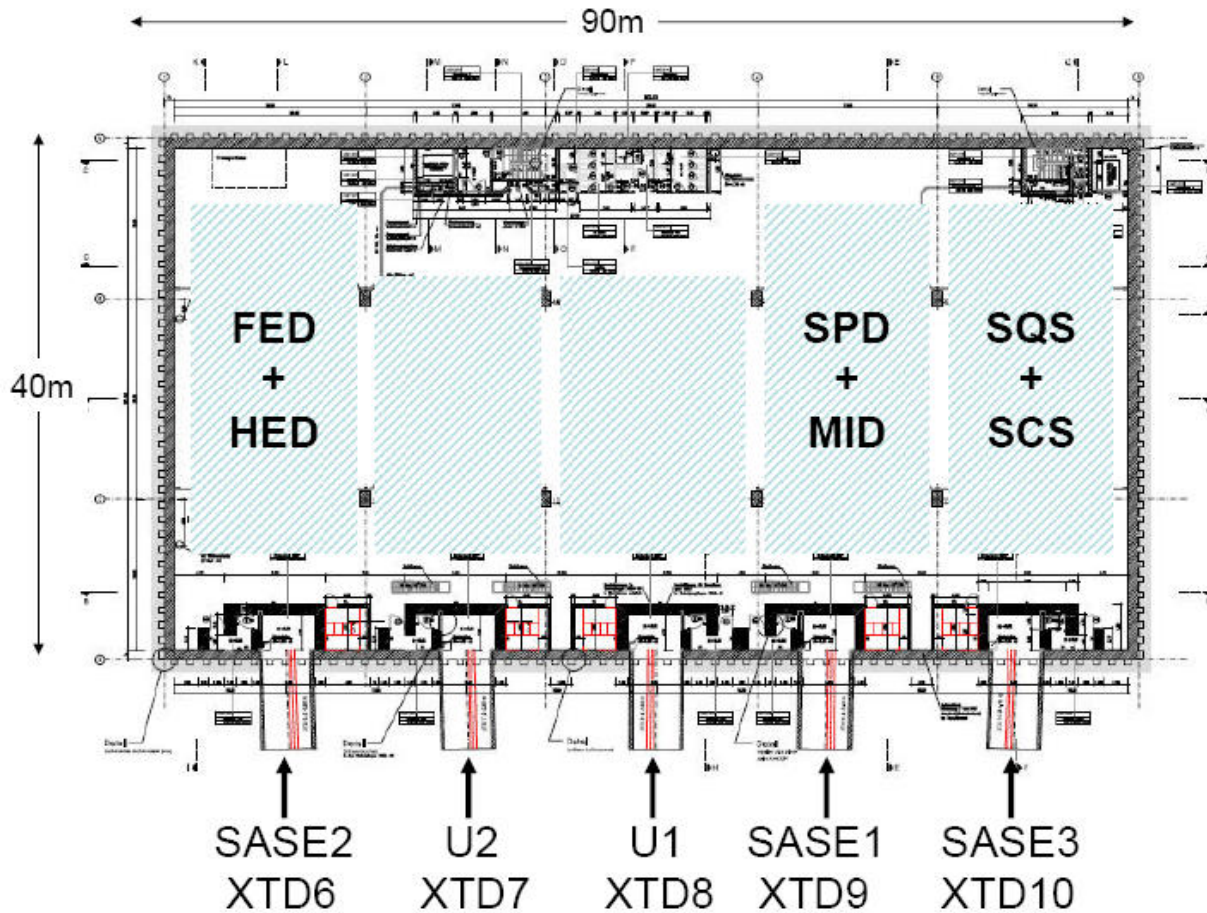
Figure 4 Plan of experiment stations in XHEXP1

2.3 EuXFEL startup time profile

The building and start up time profile for SASE1, SASE2 and SASE3 beam lines is shown in Figure 5.
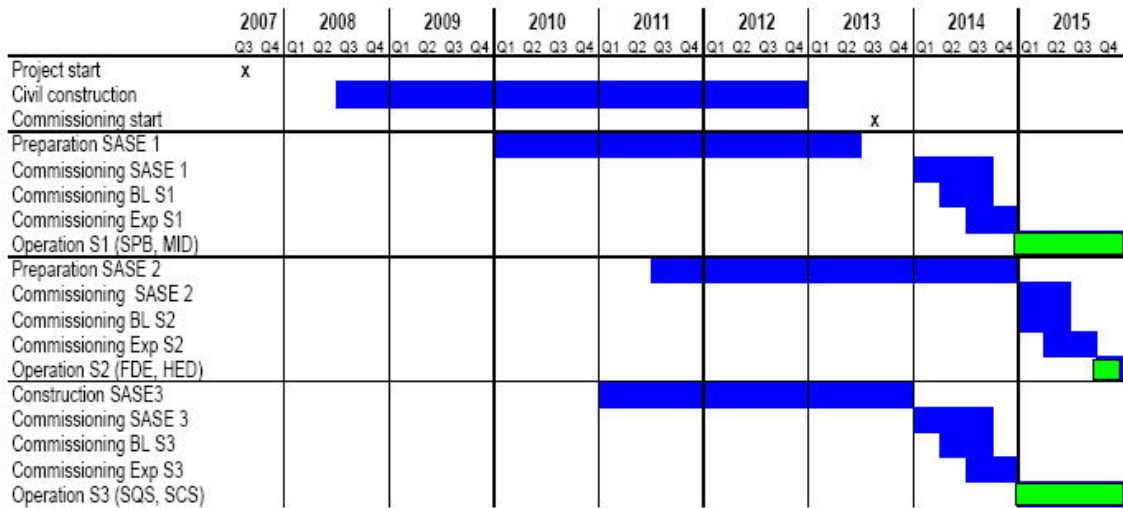
Figure 5 Start up schedule for EuXFEL beam lines

Beam line control system commissioning begins in 2014 with beam operation in 2015. The design and implementation of the control systems for photon diagnostic and detector instruments being built specifically for EuXFEL has already begun and will continue through prototyping and testing phases prior to installation.

3 Control and readout devices

The 5MHz photon bunch delivery rate at EuXFEL requires the design and implementation of new detectors and readout systems. Devices in use at existing synchrotron radiation facilities, such as commercial cameras with single shot readout times of a few milli-secs, will not be able to capture more than one frame per train at EuXFEL. In this section those devices requiring DAQ control and readout are listed.

3.1 Beam line optics

Preliminary designs exist for the instrumentation of the EuXFEL photon beam lines. Figure 6 shows the current implementation of SASE1.
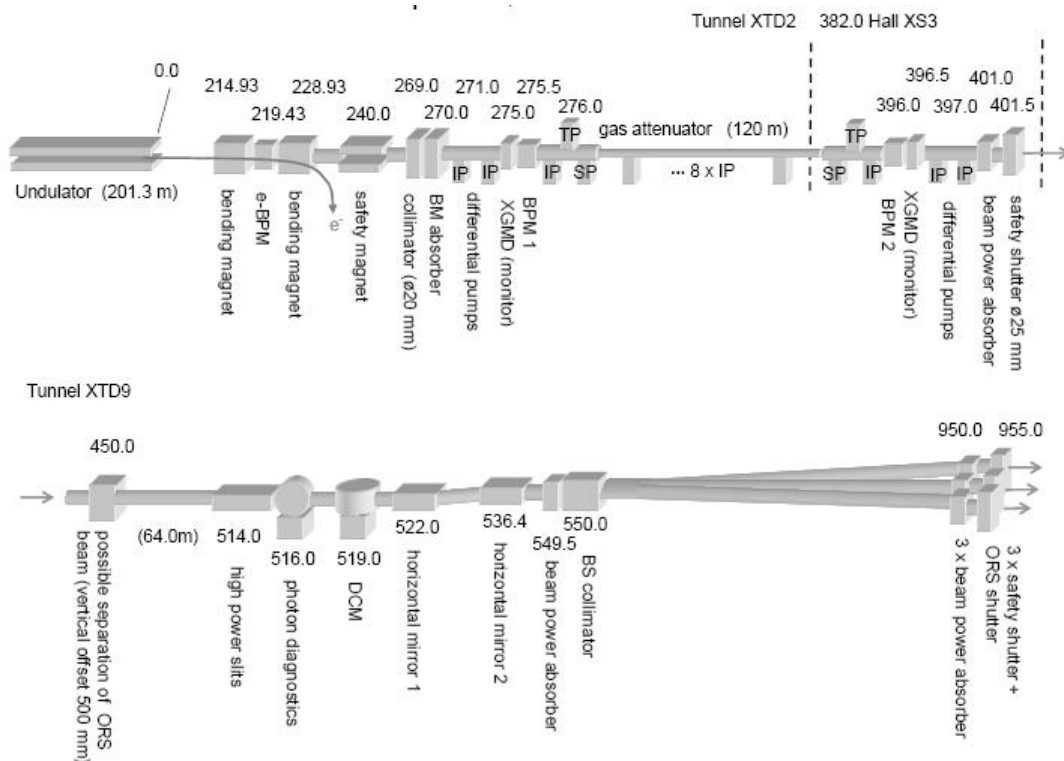
Figure 6 SASE1 beam line components (2008)

The list of the instruments requiring control and readout is given Table 2. A more detailed definition of beam line optics equipment DAQ and control requirements and solutions is being prepared.

Beam line component	Purpose	Control and monitoring requirements	Readout data volume per train (Bytes)
Mirrors	Photon distribution	position	~O(100)
Differential pumps	Vacuum	activation	~O(100)
Shutters	Photon distribution	position	~O(100)
Collimators	Beam collimation	position	~O(100)
Gas attenuators	Intensity control	setting	~O(100)

Table 2 Beam line control and readout components

Optics device DAQ and control systems are typically slow systems and will require train timing synchronization signals and unique markers.

3.2 X-ray photon diagnostics

Photon diagnostic systems are required to assess beam parameters such as intensity, position, and pulse shape for two major purposes: firstly to monitor and control the beam for alignment, steering and FEL optimization and secondly to provide information about the photon beam to users as immediate input to experiments and for later analysis of the acquired data.

The list of photon diagnostic devices expected to be used at the EuXFEL is shown in Table 3. The estimated data rates are summarized in Table 4.

Device	# / BL	Purpose	property	per pulse	train	online
XGMD	2 or 3	Online monitor for beam intensity	intensity	X	full (1500)	X
PIN Diodes	1	Detector for low intensities	intensity	X	full (1500)	-
XBPM	4 (twice x and y)	Online monitor for beam geometry	position profile	X X	full (1500)	X X
Quadrant-BPM	max. 2	Probe with highest spatial resolution	position	X	full (1500)	-
K-Monochromator	1	tune undulator	spectrum	(X)	~ 1 pulse	-
MCP-based detector	1	find and optimize SASE	Basic: tot. energy Visual: image	X X	< 30 pulses ~ 1 pulse	-
Beam Viewer	1 - 3	alignment	image	-	full	-
Photoelectron Spectrometer	1	Measure spectrum and polarization	spectrum	X	full (1500)	X
Wavefront detector	1	Measure spectrum and polarization	spectrum	-	few pulses	X

Table 3 Photon diagnostic devices

The term online is used for non-intrusive devices that can be applied during user runs in parallel with experiments downstream; any intrusive device is understood to be retractable or only temporarily installed; the highest rate assumed is a full pulse train with 1500 pulses at 10 Hz repetition rate.

Device	Control requirements	Readout data volume		
		per pulse	per train	per train per BL
XGMD	Slow controls: HV, gas pressures	few byte (raw data ~10kB ?)	< kB (raw data ??)	< kB (raw data ??)
XBPM	Slow controls: HV, gas pressures	few byte (raw data ~10kB ?)	< kB (raw data ??)	< kB (raw data ??)
Quadrant-BPM	Slow controls: voltage / current	few byte	few byte	few byte
K-Monochromator	Slow controls: HV, motors, camera settings	2 times 12bit 1280x1024 at 10Hz	~ 2 MB	~ 2 MB
MCP-based detector	Slow controls: motors, HV	Basic: 10byte Visual: ~1k x 1k	1 kB 1MB	1 kB 1MB
Beam Viewer	Slow controls: motors, camera settings	VGA type res. 640x480, 8 bit	~ 1 MB	~ 3MB
Photoelectron Spectrometer	Slow controls: motors, HV	t.b.d.	t.b.d.	t.b.d.
Wavefront detector	Slow controls: motors, settings	today 1279 x 1023 pixels at 15 fps		few MB

Table 4 Photon diagnostic control and readout requirements

Photon diagnostic devices require timing signals for synchronization; triggering and unique identification marking of bunches recorded.

3.3 Experiment detectors

EuXFEL detector development has so far concentrated on initiating consortia to develop three 2D-pixel detectors AGIPD, DSCC and LPD. Additional developments including 1D detector and other devices are to be expected.

The DAQ relevant parameters of the proposed 2D-pixel detectors are listed in Table 5. Common implementations have been defined for the timing control and backend readout sub-systems of these detectors.

Parameter (1Mpixel baseline)	2D pixel detector consortia		
	AGIPD	LPD	LSDD
sensor technology	silicon	silicon	silicon - DEPFET
pixel size	200 x 200 µm	500 x 500 µm	200 x 200 µm
sensor topology	tile	tile	tile
single γ sensitive	yes	yes	yes
soft x-ray sensitive	no	no	yes
Max. digitizing rate	5 MHz	5 MHz	5 MHz
gain control	switched 1 fold	3 fold	DEPFET

dynamic range	0 - 5x10**4	0 - 10**5	0 – 10**4
ADC bits	12	12	10
pixel data size	2 bytes	2 bytes	variable ≤2 bytes
pipeline technology	capacitor	capacitor	digital
frame pipeline depth	≤ 400	512	≤1024
Module count	32	32	16
10GE readout links	16	16	16
startup pixel count	1k x 1k	1k x 1k	1k x 1k
startup frame size	2 Mbytes	2 Mbytes	2 Mbytes

Table 5 EuXFEL 2D-pixel detector parameters

The DAQ and control requirements of the 2D detectors are described in Section 5.5.

3.4 Risks

The following risks exist:

- Design of the beam line vacuum infrastructure and associated control system has not yet started. In this note assumptions about the network and rack infrastructure requirements of the vacuum system are derived from discussions with experts.

- It is assumed that the control systems and associated electronics of all beam line systems will be commercial off the shelf products.

- Photon diagnostic and experiment instrument design must conform to the requirements of the DAQ and control group. It is assumed that working turnkey systems will be installed.

3.5 Conclusions

In this note it is implicitly assumed that the proposed 2D-pixel detectors instruments will place the largest demands on the DAQ and control systems.

4 Experiment requirements

The experiments which will initially occupy SASE1, SASE2 and SASE3 beam lines, see Figure 4, are listed in Table 6.

Beam	Class	Experiment
SASE1	SPB	Ultra fast coherent diffraction images of single particles, clusters and bio-molecules, which allow structure determination of single particles: atomic clusters, bio-molecules, viruses, cells, etc.
	MID	Materials imaging and Dynamics, which allow structure determination of nano-devices and dynamics at the nanoscale
SASE2	FDE	Femtosecond Diffraction Experiments, for time resolved investigation of the dynamics of solids, liquids and gases.
	HED	High Energy Density Matter, for investigation of matter under extreme conditions using hard x-ray FEL radiation, e.g. probing dense plasma
SASE2	SQS	Small Quantum Systems, for investigating atoms, ions, molecules and clusters in intense fields and non-linear phenomena.
	SCS	Soft x-ray Coherent Scattering, which measure structure and dynamics of nano-systems and of non-reproducible biological objects using soft x-rays.

Table 6 Experiment classes and their beam line allocations

The DAQ and DM requirements of showcase experiments in the classes in Table 6 were gathered. This information is used in this section to estimate the bandwidths and data volumes which EuXFEL operation will generate.

The estimates derived can only be considered as rough guidelines to the acquisition, archiving and analysis requirements which will exist at EuXFEL. Many unknowns exist: what experiments will actually take place, how often they will be performed, the significance of data rejection and reduction, etc. These problems will decrease when precise detector simulation tools are available and when experience from similar experiments at LCLS and FLASH is known.

4.1 FDE

Femtosecond diffraction experiments on liquid and gas targets typically use continuous flow target injectors, a pump laser to excite the target and a 2D-pixel detector to record the diffraction pattern produced when the x-ray pulse hits the excited target.

The data taking profiles anticipated at EuXFEL for a liquid target experiment is shown in Figure 7. In a five day period, three days are required for experiment setup and the remaining two days are used for data taking. Data taking consists of 30 minutes periods of acquisition during which 1.8 million images are recorded, which corresponds to 100 frames per train at a train delivery rate of 10Hz, followed by 30 minutes of setup time for the next run. Each run contains data from a given target and pump-probe delay.

The total data volume accumulated in the 2 days of acquisition during the 5 day period is 172TB[1]. The instantaneous rate is 1.95GB/s during acquisition, or 0.98GB/s when averaged of the acquisition shift period.

Data rejection or compression is not likely to be significant as the number of camera pixels hit by photons will be large, essentially the entire camera. As the target is continuously supplied all pulses could be used and no target degradation effects will be seen.
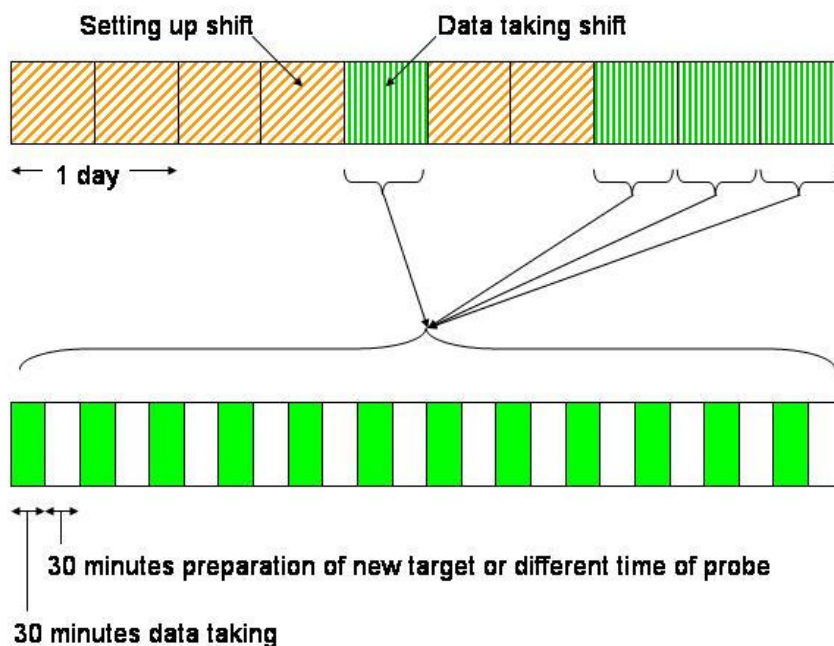
Figure 7 FDE experiment's data taking profile

It is required that data taken be archived at EuXFEL, the raw data can be deleted when the analysis is finished.

The analysis profile anticipated is to create an average image of all data taken during a run and take this data offsite for analysis at the home institute. Depending on whether the averages can be analysed or whether new analysis tools have to be tested, the averages may have to be recalculated 2-4 times. Analysis is completed between 1 month and 12 months of data taking. The algorithms used are in-house developments.

4.2 MID

Material imaging and dynamics experiments use techniques ranging from x-ray absorption spectroscopy to large coherence length imaging using 2D-pixel detectors and laser pumping of targets. Two acquisition modes exist depending on the samples ability to survive without damage the incident photon beam. Radiation hard samples, where the target (e.g. metallic deposits on thin support membranes and crystal generation in fluids at the phase transition boundary) survives many photon pulse trains, and soft samples, where the target (e.g. glasses, polymers, etc.) turns quickly into a plasma

---

[1] 172TB = 48(hrs) x 0.5(run factor) x 3600 x 10(Hz) x 100(frames) x 2(MB)

requiring frequent sample replacement by moving a target support capillary or membrane. In both cases the likelihood of hitting the target is high once alignment has been performed. MID are likely to use 2D-pixel detectors for imaging.

The data taking profile anticipated for EuXFEL for hard and soft targets are shown in Figure 8. The amount of data taken with soft samples is limited due to the overhead of moving to a new target and the probable reduction in the rate of trains and/or pulses per train delivered. The amount of data taken by hard samples could be significantly larger if high q effects are to be seen. Currently it is assumed that 100 frames per train with ~10^4 photons per frame are readout during hard sample data taking, which corresponds to 3.6 million images per hour.
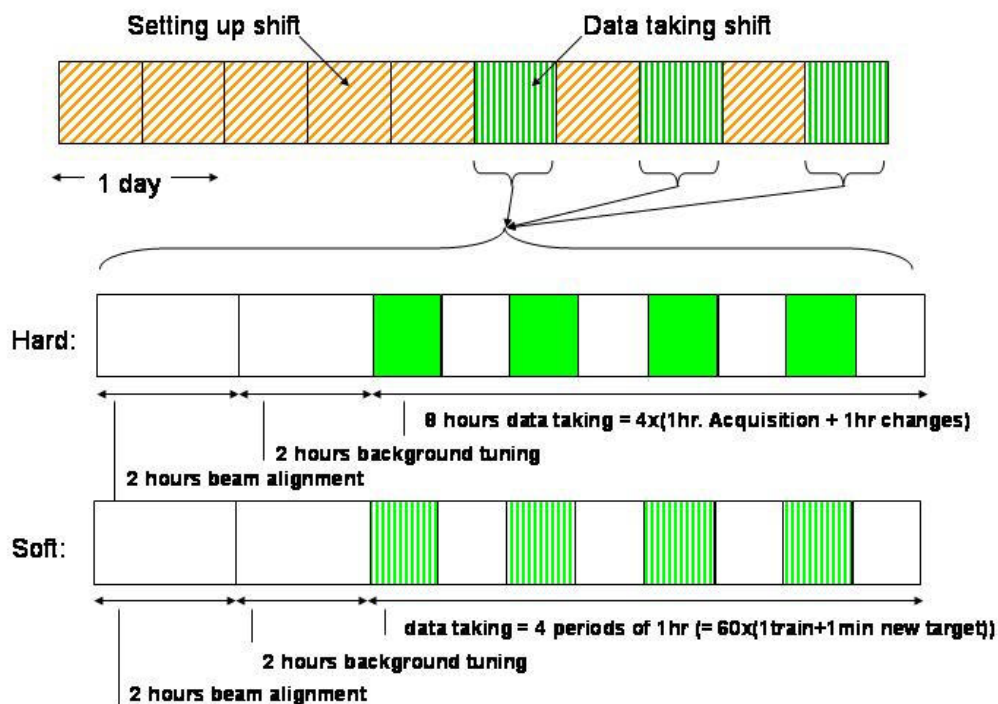
Figure 8 MID experiment's data taking profile

Data compression is likely to be significant for hard samples generating relatively few pixel hits in the detector. Data rejection by selecting frames on the appearance of Bragg peaks during a phase change might be possible.

The total data volume accumulated for a hard target during the 5 day period is 86TB[2], or 28.7TB assuming a 30% compression factor. The instantaneous rate is 1.95GB/s during acquisition, or 0.65GB/s when averaged of the 2 day acquisition period. These bandwidths will drop after compression.

The total data volume accumulated for a soft target during the 5 day period is negligible (~0.1TB) due to the large overhead of moving the target.

It is required that collected data is archived at EuXFEL, the raw data can be deleted when the analysis is finished.

---

[2] 86TB = 12(hrs) x 1.0(run factor) x 3600 x 10(Hz) x 100(frames) x 2(MB)

The analysis profile foresees re-analyses of raw data for six months following data taking. If detector understanding were to improve then a later re-analysis of the data would be useful thus the data should not be deleted early.

4.3 SPB

Large area 2D-detectors will be used to image the diffraction patterns obtained by scattering spatially coherent x-ray pulses from various SPB targets. The data taking profiles for gas, aligned gas and droplet target injection and conditioning systems have been estimated. Independent of target type it is assumed that a measurement period of 9 days is required, of which the first three days are used for experiment preparation. Due to the relatively low efficiency of hitting the target, gas and aligned gas data taking profiles are characterized by long periods of data taking separated by relatively short target change periods, see Figure 9.
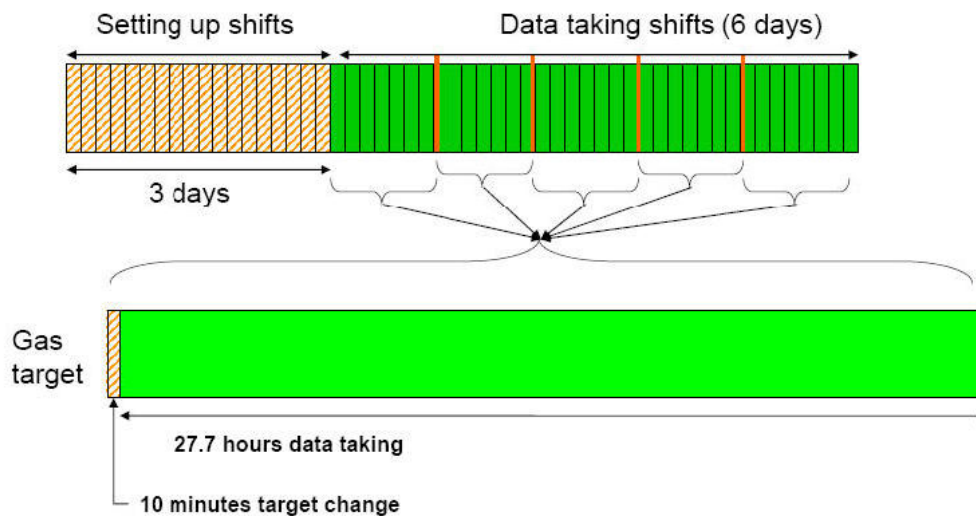
Figure 9 Data taking profile for SPB gaseous targets

For droplet targets the efficiency for hitting the target increases so that the data taking period per target is relatively short, whereas the change of target injector preparation time increases, see Figure 10.

17
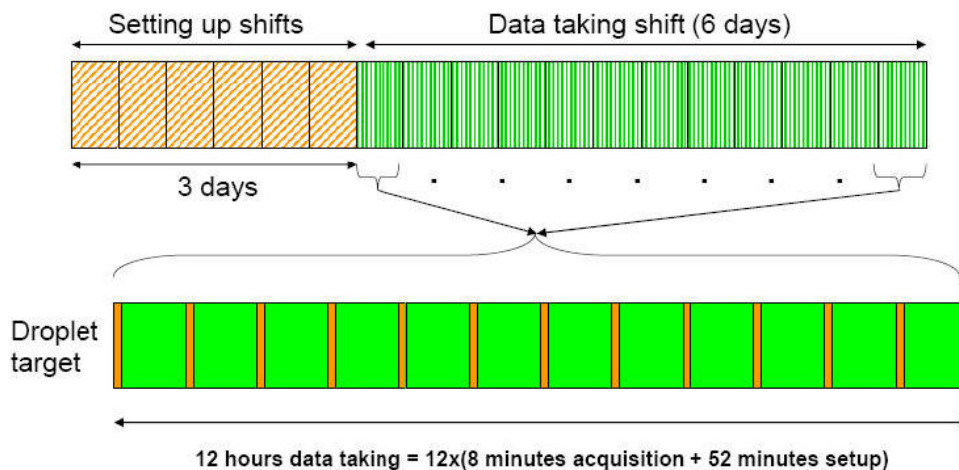
Figure 10 Data taking profile for SPB droplet targets

Details of the numbers used to derive the above data taking profiles are shown in Table 7.

As the multiplicity of photons generated per image is low for gaseous targets, ideal data compression has been assumed as has a noise contribution equal to the signal size. During the 6 days of acquisition the resulting total data volumes generated are 200GB to 200TB, 2TB, and 0.24TB to 1.2PB, for gas, aligned gas and droplet targets, respectively. By assuming that all images acquired for gaseous targets where the target was not hit can be rejected by online vetoing or algorithm based rejection the resulting total data volumes generated go down to 0.2GB to 0.2TB, 0.002TB, and 0.24 to 24TB, for gas, aligned gas and droplet targets, respectively.

Breakdown item	Gas	gas	droplet	Droplet	aligned gas
Injection rate	0.1 Hz	10 Hz	100 kHz	5 MHz	100 Hz
Hit frames per train	0.01	1	20	1000	1
No hit frames per train	999.99	999	980	0	999
Hit frame pixel multiplicity	10 to 10 k	10 to 10 k	100k to 1M	100k to 1M	1 to 100
Average train data size	40kB to 40MB	40kB to 40MB	~2GB	~2GB	4kB to 400kB
Frames required for 1 run analysis	1M	1M	100k	100k	10M
Run time acquisition	115 days	27.7 hrs	8.4 min	10 sec	11.6 days

Target swap time	10 min	10 min	1 hr	1 hr	1 day
Data volume 9 day shift w/o background frame rejection	0.2 to 200TB	0.2 to 200TB	1.2PB	0.24TB	0.02 to 2TB
Data volume 9 day shift with background frame rejection	0.2 to 200GB	0.2 to 200GB	24TB	0.24TB	0.02 to 2GB
220 day annual volume	0.25 to 250TB	5 to 5000TB	30PB	6TB	0.25 to 25TB
220 day volume with background frame rejection	0.00025 to 0.25TB	0.005 to 5TB	600TB	6TB	0.00025 to 0.025TB

Table 7 SPB data taking profile details (1000 frame trains assumed)

Data analysis will require using sophisticated algorithms for image reconstruction. The work on these algorithms has started and initial reconstruction results demonstrate their feasibility. The preliminary CPU requirements are quite challenging. The reconstruction of the signal from 10^6 ideal (almost background free) diffraction patterns requires 64 CPU cores running in parallel for 24 hours. The required CPU time may increase if realistic background is included.

4.4 HED

High energy density experiments typically investigate plasmas, material states or shock wave dynamics, and require the target to be pumped with high power optical laser or, perhaps, auto correlated x-ray beams.

High energy density experiments are currently expected to be relatively undemanding in terms of readout bandwidths and data volumes generated. This is due to the anticipated low rate of laser pumping and the requirement of injecting, or moving to new target material between single shots. A crude estimate of the requirements for such an experiment can be derived by assuming that one pulse of data per train is acquired with 50% efficiency to account for setting up etc. If one additionally assumes a 2MB data payload generated by a streak camera this would generate an instantaneous bandwidth off the detector head of 20MB/s and an annual 220 day data volume of ~400TB.

4.5 SQS

The SQS end station located on the SASE3 beam line is dedicated to the experiments on neutral gas phase targets ranging from atoms and small molecules to large bio-molecules, clusters and nano-particles and experiments with ionic targets ranging from atomic ions to complex molecules.

In order to optimize the use of EuXFEL beam time for each experimental group, a fixed and versatile end station is planned. The experimental chamber will be equipped with various detectors allowing for both photons and particles detection. It is planned to install at least the following detectors:

- a high resolution time-of-flight spectrometer

- ▪ a VMI (velocity map imaging) analyzer
- ▪ a Thomson parabola
- ▪ a COLTRIMS set-up
- ▪ an advanced 2D-photon pixel detector

A repetition rate as high as possible is desirable, but the actual amount of collected data is going to be reduced in cases where the coincidence detection scheme is applied or when the pump-probe experiment technique utilizing additional non x-ray lasers is used. For example if 100kHz laser is used for exciting a target then only max. 60 bunches in the x-ray bunch train can be synchronized to the pumping laser frequency.

4.6 SCS

The Spectroscopy and Coherent Scattering (SCS) instrument is intended for the investigation of atomic and electronic structure as well as of the dynamics of soft and hard matter, biological species and magnetic materials.

Some of the experiment show cases presented in the earlier sections like magnetic materials or biological molecules can also be applied here as they differ primarily in the utilization of the x-ray photons with lower energy as supplied by SASE3 beam line. Therefore it is safe to assume that the data volume for SASE3 experiments (SCS+SQS) should be expected to be comparable with data volume for SASE1 experiment.

4.7 Expected data bandwidths and volumes

The bandwidth of data transferred through and the integrated volume of data produced are estimated in this section.

4.7.1 Instrument limitations

The implementation of the Train Builder 2D pixel detector readout electronics, see Section 5.5, currently limit the number of frames readout per train to 512 with full data payload of 2MB per 1Mpixel detector. The lengths of the storage pipeline in the three detectors additionally limit the number of frames which can be acquired to ≤512.

4.7.2 Running time and inefficiencies

The DAQ efficiency is assumed to be 90%, the experiment hardware efficiency to be 70%, and EuXFEL beam delivery efficiency is 70%. These efficiencies are derived from experience at other facilities.

The annual beam time, in days per beam line, used in the estimations is shown in Table 8. It is assumed that EuXFEL beam line is dedicated to experiments for 220 days when fully commissioned and debugged. The linear turn on profile shown is arbitrary and the turn on of SASE2 is assumed to be accelerated by experience gained with other SASE beam lines.

Beam line	2014	2015	2016+
SASE1	55	110	220

SASE2	0	110	220
SASE3	55	110	220
U1	0	0	55
U2	0	0	55

Table 8 Annual EuXFEL beam time in days

4.7.3 Estimated bandwidths and volumes

The total data volume passing through the DAQ is the sum of beam line component, photon diagnostic and experiment data rate. The numbers derived in this section are based entirely on the detector data volumes described in the experiment requirements section above. Data generated by the photon diagnostic and beam line systems are not included as they will contribute significantly less data.

The annual data volumes and the instantaneous data rates off the detector head are summarized in Table 9. Note that the detector limitations and the run time inefficiencies have been folded into the results. Additionally where ranges of results exist for a given experiment type the larger number has always been used.

Experiment type	Max instantaneous Bandwidth (GB/s) off the detector	Expected data volume per year (compressed)	Expected data volume per year (not compressed)
FDE	1.95GB/s	3.4 PB	3.4 PB
MID hard	1.95GB/s	0.6 PB	1.7 PB
MID soft	-	2 TB	2 TB
SPB gas	10GB/s	1 PB	54 PB
SPB gas aligned	10GB/s	7.6 TB	54 PB
SPB droplet	10GB/s	6. 6PB	6.6 PB
HED	20MB/s	0.4 PB	0.4 PB

Table 9 Experiment data bandwidths and volumes

The data volumes generated by EuXFEL during the turn on period, see Table 8 are shown in Table 10. The largest range results from each experiment class are used to estimate the total volume per beam line assuming that each class uses half of the available beam time, there are two experiment classes per beam line. Note that the SASE3 data volumes are assumed to equal those of SASE1.

Beam line	2014	2015	2016+
SASE1 (SPB+MID)	0.6/10	1.3 / 20	2.8 / 39

SASE2 (FDE+HED)	-	1.9 / 1.9	3.8 / 3.8
SASE3 (SQS+SCS)	0.6/10	1.3 / 20	2.8 / 39
U1	-	-	?
U2	-	-	?
Total	**1.2 / 20**	**4.5 / 31.9**	**8.8 / 81.8**

Table 10 Estimates of PB data volumes per beam line during EuXFEL turn on top of

4.7.4 Risks

The experiment requirements derived in this section should only be considered as rough guidelines as many uncertainties exist. Known risks and their potential effects are itemized below.

- Data taking profile – the total number of frames which must be acquired to ensure that a meaningful analysis can be performed on the data can usually be defined. It is, however, almost impossible to precisely predict the number of good frames acquired per train as target inefficiencies, non auto-correlating pumping rates, target survival times in the beam, and target replacement schemes are currently uncertain. Developments in target replacement systems and non auto-correlating pumping rates could lead to higher rates at EuXFEL than those described above.

- Detector limitations – the data volumes derived are generated assuming use of the three 1st generation 2D-pixel detectors and their readout systems. The DAQ performances of the 2014 baseline detector designs are similar; 1k x 1k pixels (2MB frames), 200-500μm pixels, with ≤512 frames per train readout. A move to larger detectors, 2k x 2k = 4 fold increase, and an increase in the frames acquired per train, 512 to 1024 (DSSC), should be expected.

- Frame rejection and compression – in the numbers derived for the SPB showcase best case lossless compression has been assumed, which significantly decreases the data volumes generated, this assumption may be incorrect. Additionally it is has been assumed that noise only frames cannot easily be rejected by a trigger, e.g. TOF, or by an algorithm, both assumptions may be incorrect.

- Analysis profile – the SPB showcase was the only experiment class which delivered information concerning analysis requirements, but even these are preliminary.

4.7.5 Conclusions

In spite of the uncertainties associated with the experiment requirements the following conclusions can be drawn.

- Bandwidth - the DAQ readout must be capable of sinking the largest data bandwidth generated by the cameras used. Currently this is 10GB/s for a 1Mpxl detector acquiring 512 frames per train. The readout must also be scalable to allow for larger detectors, say 4Mpxl rather than 1Mpxl, and handle a potential doubling of the frames acquired per train.

- Rejection and compression – the DAQ architecture design must be flexible enough to allow rejection and compression at the earliest possible stage, once these factors are understood.

- Data volume – the annual storage requirements for commissioned EuXFEL running are ~10PB. The implementation of the data management system must be capable of handling this size, and be seamlessly scalable in the range 5 through 100PB.

- Rollout – The 10PB EuXFEL storage size can be considered as a sensible starting point. Information from LCLS experiment running and improved detector simulations should allow an improved estimate of this number, and the conclusions should be reviewed in ~2011. Should an increase in storage size beyond 10PB be required due to successful EuXFEL operation; this would require a second funding phase.

- All experiment classes are likely to use the 2D cameras being developed for EuXFEL.

5 Data acquisition architecture

The data handling design is driven by the following concepts:

- handling the large instantaneous data bandwidths generated by detectors
- use of standard IP network protocols and commercial hardware as early as possible
- data processing with the aim of reducing data storage and transfer requirements
- data buffering to decoupling online data taking and offline storage and analysis tasks
- scalability in terms of the number of instruments controlled and data volume produced
- flexibility allowing the use of new technologies

The architecture solution chosen to satisfy these concepts is shown in Figure 11. A detailed schematic of the architecture is shown in Figure 16. Implementing a solution with multi horizontal layers and well defined interfaces between layers allows development work to proceed independently within the different layers and foresees the possibility of swapping entire layers should more suitable solutions be developed. Defining vertical slices for each detector is natural and allows partitioning which improves data security and interoperability.
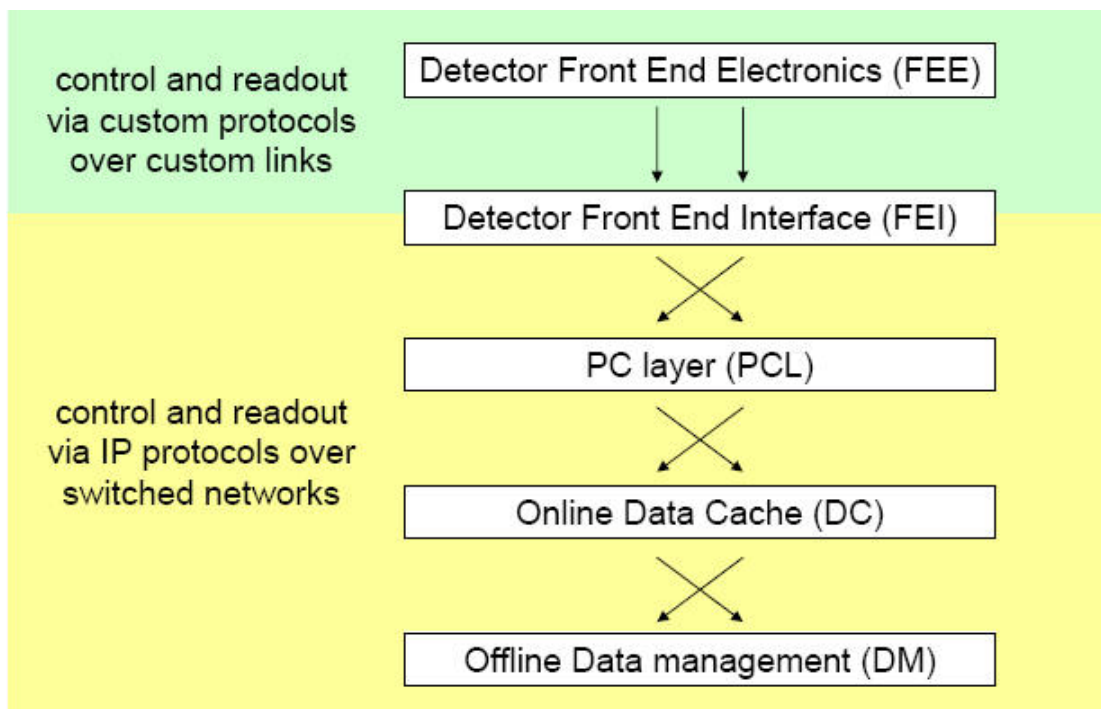
Figure 11 EuXFEL DAQ architecture

5.1 DAQ architecture layers

The implementations of the different layers are described briefly in this section. The proposed AGIPD 2D pixel detector is described as a show case implementation in Section 5.5.

5.1.1 Front End Electronics (FEE)

Front end electronics modules are detector specific units used to control and readout sensors which are required to:

- Interface to a downstream front end interface

In simple detectors, e.g. 1D, the front end electronics and interface may be combined on a single board.

5.1.2 Front End Interfaces (FEI)

Front end interface modules are required to:

- interface to the FEE using a reliable custom or IP protocol
- interface to the machine timing system and, if necessary, the machine protection system
- distribute power and other slow control services to the FEE
- sequence the operation of the FEE using timing system signals
- be controlled and configured via IP
- interface to the downstream readout PCL using a reliable IP protocol
- produce train specific frame ordered data which is sent to the down stream PCL

In the simplest implementation detectors consist of an FEE and FEI, an interface to the machine timing system, an IP output link to the PCL for sending data and for sending and receiving control (and monitoring) information.

The common backend solutions of the 2D-pixel detectors designed for EuXFEL currently split the readout and control functionality as described in Section 5.5.

A common protocol and software implementation of the control functionality should be aimed for.

5.1.3 PC layer (PCL)

The PC layer will be implemented as a PC farm and is required to:

- receive train data originating from the FEI layer
- perform initial file formatting (see below)
- process received data for monitoring purposes
- perform data rejection and compression (R&C)

- send data passing R&C, or a summary place holder for rejected data, to the data cache

For a single 1Mpixel 2D-pixel detector the number of PCL hosts required to simply pass data on to the data cache is ~10. Depending on the data input load, monitoring requirements, and rejection and compression algorithm requirements, the number of machines can be increased to provide sufficient processing time.

5.1.4 Data cache (DC)

The initial design foresaw a data cache as a temporary storage location to be used to decouple data taking and offline tasks and to provide sufficient storage when the primary offline archive was unavailable. In the architecture described here the cache becomes an active part of the system as it will be used to further process data before sending to the data archive.

The data cache is required to:

- receive data originating from the PCL
- store the data in the cache for ~2 days
- allow further R&C processing
- send accepted (positive commit) data to the offline archive

The maximum size of the data cache per beam line per 1Mpxl detector is (2MB frames, 512 frames/train, 2 day continuous operation) is 1.8PB. The data taking profiles described in Section 4 indicate that ≤500TB of storage is required for day one operation.

5.1.5 Offline data management

Offline data management is described in Section 8.

5.2 Equipment location and networks

The physical location and network connectivity of the different layers are described in detail in Section 0.

5.3 Timing information

A TCA crate format Timing Receiver board is being designed by the machine control group. This will be used by instruments to synchronize to the photon delivery.

To operate at other light sources FEIs must interface to both the local timing system and the control and readout interfaces. This appears to be relatively straightforward for FLASH and Spring8. At LCLS this is more difficult due to underlying hardware implementation of the control and readout handling.

5.4 Machine protection system

A TCA crate format machine protection board is being designed by the machine control group. This will be used by instrument control systems to interface with the machine protection systems.

5.5 Showcase 2D pixel detector DAQ implementation

Common implementations have been defined for the control and backend readout sub-systems of the 2D-pixel detectors and this section contains a description of the DAQ architecture implementation for the AGIPD 2D-pixel detector.
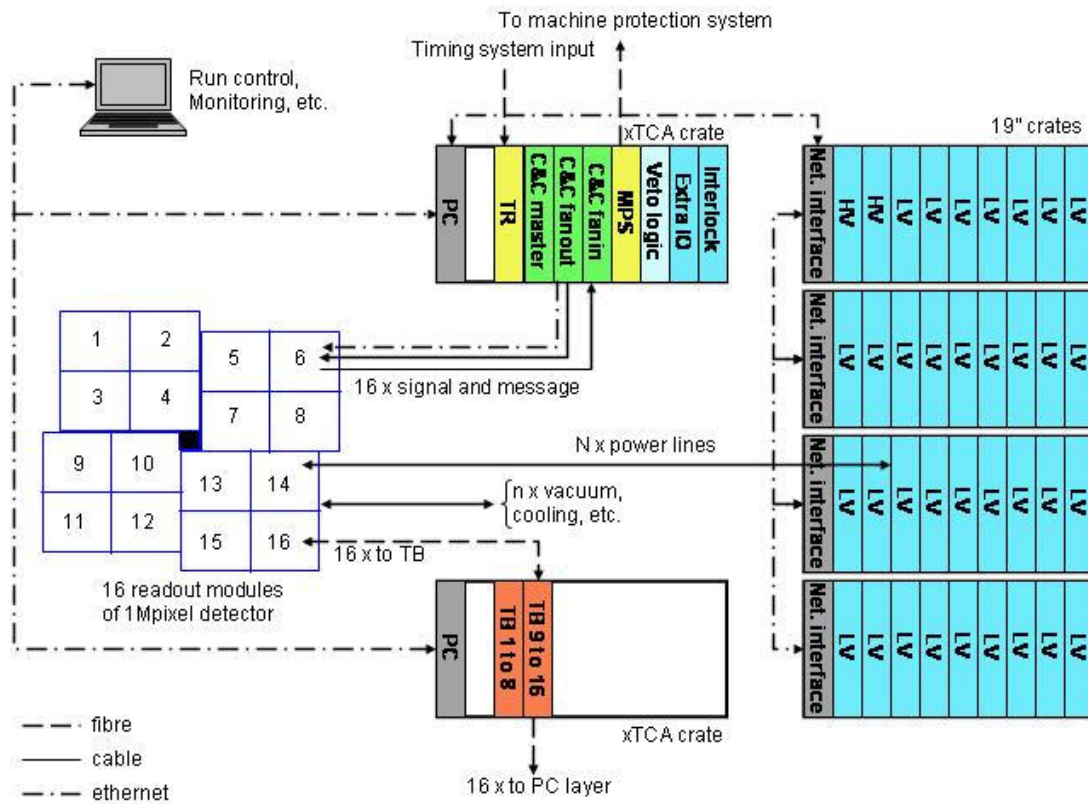
Figure 12 Proposed AGIPD 2D-pixel detector control and readout architecture

The DAQ architecture for AGIPD is shown in Figure 12; the key elements are:

- a xTCA crate (upper) holding control hardware (timing, sequencing, machine protection, slow control, etc.) and Single Board Computer (SBC)
- a xTCA crate (lower) holding FEI hardware and SBC
- 16 FEE 10Gbit/s readout links
- a run control host for run and train configuration

The sequencing and synchronization control, or clock and control (C&C), system is implemented using AMC electronics boards located in a xTCA crate. A C&C master is interfaced through the backplane to a Timing Receiver (TR) board [1], which provides signals and clocks required to synchronize the detector DAQ to the photon bunch delivery time structure. Fast signals, start train, PLL clock, etc., are distributed via cables to the detector specific front end modules (FEE) by fanout boards. Slower information, train bunch pattern, etc., is distributed via LAN using the crate SBC. Control signals generated by the FEE modules are interfaced to the C&C master using fanin boards. The C&C system is being designed and implemented by UCL [2].

The backend readout of the detectors is organized into 16 10Gbit/s SFP+ fibre connections per Mpixel of sensor, which maps well to the 16 or 32 FEE module multiplicities of the detectors. The readout links connect to the train builder board, see below, which performs frame and train building. The Train Builder performs readout data handling for the EuXFEL 2D-pixel detectors (AGIPD, DSSC or LPD).

The baseline requirements of the Train Builder system are listed below:

- The TB design uses 10GE links over SFP+ fibre transceiver input and output modules. The link hardware is fully bi-directional.

- Input data is received from multiple FEE modules, each module has one link.

- The input data transfer protocol will be UDP without retries. The use of other protocols PGP, Aurora, etc. is not excluded

- A maximum of 2 bytes per pixel and 512 picture frames per train will be handled.

- The detector to TB fibre distance is ≤30m.

- The TB builds all the detector data from frames of a given train into a single frame ordered block, and sends this block to the PCL for data processing and analysis.

- The output protocol is TCP or a fully reliable UDP equivalent. The TB to PC separation is ≤300m (i.e. multi mode fibre)

- Network package loss should be minimal; the data encoding chosen should result in the smallest number of picture frames being affected if a packet is lost.

- The nominal EuXFEL bunch train repetition rate is 10Hz, special runs with increased rates up to 30Hz can only be handled at the TB by reducing the maximum number of frames input.

- The option of providing on board processing (data reduction and compression algorithms) exists.

- xTCA crate and form factor standards will be used for the implementation.

The baseline implementation on a standard size ATCA board is shown in Figure 13. The central feature of the architecture is an analogue cross-point switch, which is protocol agnostic and operated as a simple barrel shifter, which is used to build the detector fragments into contiguous ordered events before outputting to the PCL.

Given the components and technologies which are available on the prototype detector timescales it has been decided to target as a baseline a system which would build events from ½Mpixel detectors each

with 8 I/O channels. The two ½Mpixel boards required per Mpixel detector can be connected directly to double NIC PCs in the PCL. The manufacturer's roadmaps for the relevant technologies are however encouraging and it may be possible to achieve 1 MPixel builder systems on longer timescales

FEE outputs arrive on 8 x 10 Gbps SFP+ optical transceivers located on a standard ATCA Rear Transition Module (RTM), allowing different optical receiver components to be used. A first stage of FPGAs receive incoming data from pairs of input channels and may perform initial data ordering (this could be customised logic provided by each detector). The data from a number of bunch trains must be buffered in large external memories before feeding to the switch. The fragments are sent across the cross-point switch, under the control of one of the FPGAs, in a predefined pattern of transfers which results in fully assembled events being fed to a number of processor units also implemented on the baseline board by a second stage of FPGAs where the data is buffered before being sent to the outputs on the front panel.
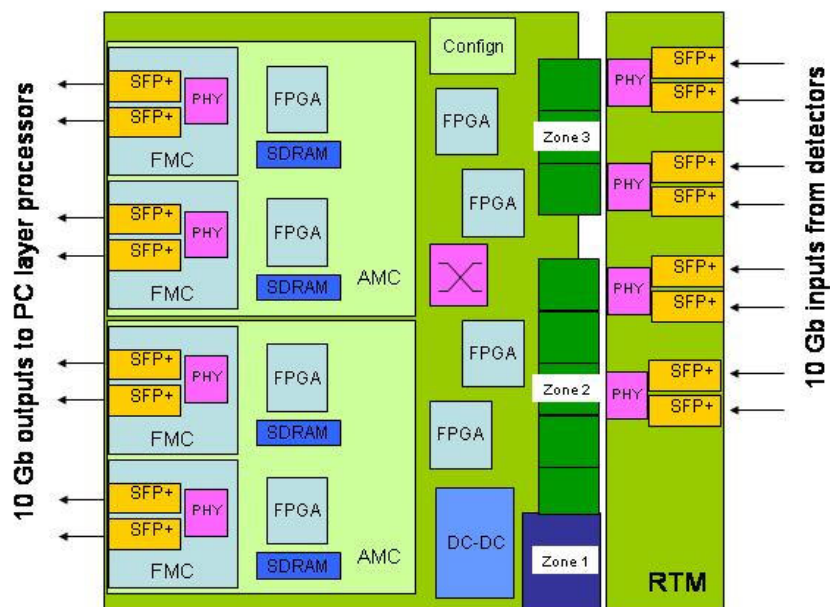
Figure 13 Schematic of Train Builder

The processing units are located at the front panel which leaves open the option of locating them on mezzanine cards which would permit them to be upgraded. The actual processing units employed would depend on the application. FPGAs as a baseline can be used for basic re-formatting tasks and 10 GbE outputs to optical links. But more advanced processing nodes could be envisaged ranging from conventional multi-core RISC processors and graphics processors to Field Programmable Computing Arrays. The outputs of these units, corresponding to data from ½ detector but complete bunch trains, are sent via 8 x 10 Gbps SFP+ optical transceivers to the PC farm for final processing and analysis. The TB readout system is being designed and implemented by STFC [3].

The control and readout sub-systems can be extended, by replication, to fit larger detectors. A 4Mpxl detector installation would require four times more C&C fanouts, TB boards and TB to PCL link connections than a 1Mpxl detector.

Run control and configuration is performed using the run control PC interaction with the SBCs of the C&C and TB sub-systems, and the PCL control system.

5.6 Scalability and resource sharing

The architecture described should be sufficient for the detector devices foreseen, but it should be assumed that larger detectors will be introduced. A 4 Mpixel detector can be produced by replication of 1 Mpixel detectors, and it is foreseen that both C&C and TB can be scaled in the same way.

At the PCL one can scale up the number of PCs and increase the number of NICs (network interface cards) per PC for increased 2D size. The latter has the advantage that the train building, which is no longer complete at the TB output, can be completed after the TB by ensuring that all partial trains from each TB board are sent to the same PCL node.

The DC can be increased in size, but it should be assumed that improvements in R&C will make this less important. The ability to share PCL and DC resources between experiments will also be useful in reducing the size of the final implementation.

5.7 Effect of modifying the bunch delivery time structure

Switching from multi-bunch 10-30 Hz train delivery to continuous (~20 kHz) single bunch operations has a number of consequences:

- The interface to the machine timing system should not require modification as the functionality required for continuous delivery is a subset multi-bunch train delivery.

- Beam line optics systems are not affected as their control and readout timing requirements are loose, ~few seconds.

- The FEE and FEI layers of experiment instruments being built for multi-bunch train operation, AGIPD, LPD and DSSC, may require modification. A design aim of the C&C sequencing for these detectors was to allow operation at LCLS (120 Hz) and other continuous delivery light sources. In this mode of operation the FEE electronics buffers digitized data in memory and when sufficient data is accumulated sends it to the FEI layer which allows the existing train builder implementation to be kept. This design should be acceptable at the significantly higher rates, but requires discussion and confirmation with the detector FEE groups.

- The FEE and FEI layers of photon diagnostic detectors would require modifications similar to those of the experiment detectors.

- In multi-bunch train operation a design feature of the DAQ and DM data handling was the guaranteed that bunches were ordered within a train, where as train ordering was guaranteed only at the macroscopic level. The diagnostic and experiment instrument modifications described above are consistent if an arbitrary consecutive sequence of bunches is considered to be a train.

Any decision to modify the bunch delivery time structure needs to be made as early as possible.

5.8 Risks

A number of risks are present:

- The development of R&C techniques will be critical in determining the final size of the implementation.

- The development of FEE and FEI (TB for 2D detectors) hardware is challenging due to the 5MHz/10Hz bunch delivery structure at EuXFEL.

- As described above, changing the bunch delivery structure will potentially require a redesign of diagnostic and detector instrument FEE and FEI hardware.

5.9 Conclusions

It is anticipated that the architecture proposed, with modifications where necessary, will be the one in use at EuXFEL.

6 DAQ software

The goal is to provide control and readout software which satisfies both users and software developers.

The EuXFEL machine control system will be DOOCS [4] based. Our starting point is not to use DOOCS and instead use the time before EuXFEL operation to look for different solutions using open source software where possible. A later return to DOOCS is not excluded.

6.1 Concepts

The current concept is that a detector or set of detectors form a unit which must be controlled and readout as a single entity called a Readout and Control Unit (RCU). Instruments in an experimental hutch or an intensity monitor are examples of such RCUs, as are beam line optics systems associated with a tunnel and photon diagnostics systems.

Clearly RCUs have to exchange information with other RCUs and external control software systems like DOOCS. This will be performed using gateways which interface the differing protocols used.

Note that because DOOCS has already interfaces to other control systems (TINE, TANGO, EPICS etc.) an appropriate gateway allows us immediate access to these systems. TANGO and EPICS are often used to control and monitor beam line optics systems.

6.2 Tools

Developments in the field of Java Standard Edition, Enterprise Edition, JavaFX scripting language, etc offers many solutions that map directly into a control and monitoring system, these include:

- support for database access
- support for transaction handling
- support for security features
- support for messaging systems
- support for building Rich Internet Applications (RIA)
- RIA supported GUI development with CSS look-and-feel control
- simplification of deployment procedures
- streamlined web and non-web client application development
- standard deployed application launch mechanisms

These tools will be used in the development of control and monitoring software. Additional tools which will also be used (see also section 8.9) are:

- GlassFish application server [5]
- NetBeans Integrated Development Environment (IDE) [6]
- SubVersioNing (SVN) code repository [7]
- Web server for documentation

Application servers offer a number of DAQ useful features: standardized deployment of web and non-web applications, client containers supporting security and transaction management, light weight application client containers, messaging services, database services, etc. Using an IDE to develop code offloads tedious and repetitive tasks from the developer whilst providing standard building blocks for coding.

All code developed will be checked into a SVN repository. Documentation will be provided via the work package web server. The software management guidelines described in Section 7.3 and 8.9.2 will be followed.

The steady stream of new open source software packages is extremely challenging to software design and must be followed. This will require changes in the tools and software packages used and requires that APIs be carefully designed.

6.3 Showcase RCU implementation

A schematic showing the software blocks required to implement the showcase AGIPD RCU hardware implementation, see Figure 12, is shown in Figure 14.

The Application Client Container (ACC) is an important feature of the implementation. The ACC allows clients (RCO, RCM, etc.) to be started in a lightweight container (transaction and security handling is missing) whilst keeping the significant benefits of a full EJB container (annotation dependency injection, web startability, etc.).

Messaging functionality is provided by a Java Message Services (JMS) [8] Topic. A JMS topic implements the publish-and-subscribe model, clients define which messages to receive by specifying an appropriate selector clause in SQL 92 syntax.

Database access is facilitated via the Java Persistence API (JPA) which maps table rows in relational databases to entity objects required during CRUD requests. Entity classes are automatically generated from the network database implementation by the IDE.

The implementation foresees a dedicated database for each RCU, which guarantees that it can be used in a standalone test environment. Each RCU DB will require synchronizing to a central database when running at EuXFEL.
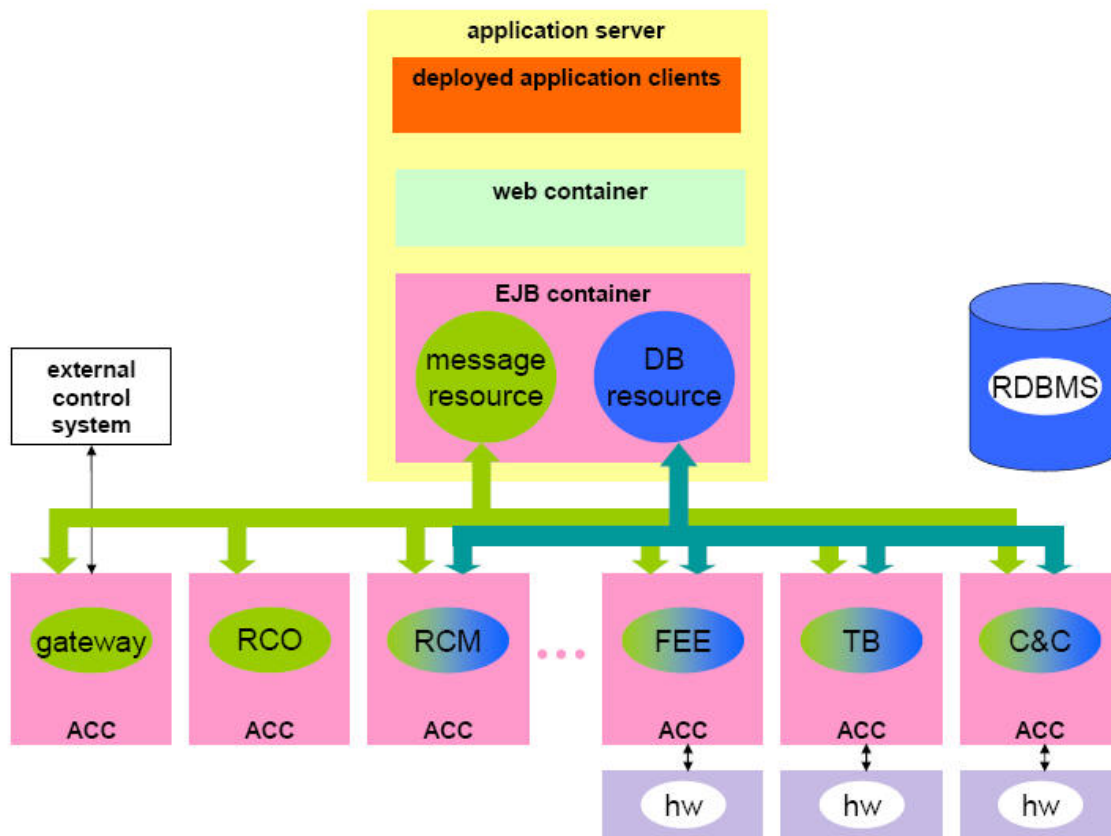
Figure 14 Conceptual implementation of AGIPD DAQ and control software

Before running an RCU required applications (RCM, FEE, etc.) must be deployed to the applications server and the run configuration be defined in an RCU specific database (RDBMS). The RCU can then be run following step definitions shown in a simplified form below:

- the operator client (RCO) selects the run configuration key from configured runs in the database
- required client applications are started on target hosts by a script or application initiated by RCO
- clients participating in message transfer connect to the message hub
- RCO using RCM navigates clients through the state transition definitions of the run

It is anticipated that most control code will be written in Java, but C, C++ and scripting languages will also be used. Test have been performed which demonstrate JMS messaging with C and C++.

Note that no EJB objects are used in the showcase implementation. Candidates for their use are in web services defined to handle run status and monitoring information stored in the database or buffered online.

6.4 Instrument integration policy and software development path

The DAQ and control software developed must be tested with real hardware in realistic user operating conditions as early as possible. Additionally the interface between the DAQ and control work package (WP76) and other related work packages must be clearly defined. Interfaces and software development paths for the various detector classes are outlined below.

6.4.1 Machine control

Software gateway interfaces are foreseen for bidirectional exchange of information between machine and photon control systems.

As proof of principle a test gateway should be written.

6.4.2 Beam line optics and vacuum systems

As described previously the beam line optics systems are expected to be commercial off the shelf units (COTS) with defined software interfaces, these are usually EPICS. WP76 is responsible for interfacing this system to the machine timing, photon control and machine control systems.

Who buys the interface hardware (crate, SBC…) is not yet defined. In the medium to long term IP based hardware solutions should be used when available.

Exchange of control information between photon and electron beam line systems will be performed using gateway processes.

A provisional design of the vacuum systems (ion pumps, controllers, valves, etc.) will not be available before mid-2010. When available a costing of the system will be possible. Space has been provisionally allocated in tunnel racks , every 95m, for ion-pump HV controllers. It is assumed that WP76 will interface the vacuum system software to photon control system and provide an exchange interface with the electron control system using a gateway.

6.4.3 Photon diagnostics

Photon diagnostic systems are either commercial products or in-house developments with sufficiently sophisticated hardware to handle the 5MHz multi-bunch train structure of EuXFEL. They are considered to be turnkey systems which:

- satisfy and support the timing interface requirements of the timing receiver board provided by WP28

- are compliant with DAQ and control interface requirements for network configuration, monitoring and readout, which are currently being defined.

A turnkey systems is a complete detector and electronics system, which on arrival is integrated with the timing and run control software. Integration with the timing system requires inserting the appropriate timing receiver board and configuring the control and readout software of the device to comply with those at EuXFEL.

Exchange of control information between photon diagnostic and electron control systems will be performed using gateway processes.

6.4.4 Experiments

WP76 is responsible for DAQ and control in the experiment hutches and laser rooms. For the 2D pixel detectors, being developed for use at EuXFEL, WP76 is specifically responsible for backend readout

systems, control and timing interfaces, and control and monitoring software. This model of integration should be applied to additional detector elements (FLASH cameras, 1D detectors,etc.) required.

Exchange of control information between photon detector systems and electron control systems will be performed using gateway processes.

6.4.5 Undulators

WP76 is required to interface to the Global Undulator control system software. The definition of this interface has not yet been determined.

6.5 Risks

The following risks exist:

- The software developed fails to meet the requirements of users.
- Implementation of user specific monitoring and analysis software is time consuming.

6.6 Conclusions

The development outlined above aims at producing an environment using standard tools. This should have significant benefits in reducing the development time required.

7 Data Management overview

The primary goal of the data management (DM) system is to provide a standard set of services (storage, access, analysis…) which allow DAQ and scientific users to perform operations that they require. The DM system must properly deal with various types of data and metadata, define appropriate data formats, provide services for datasets location and efficient transfer, and ensure appropriate data protection policies. In this chapter the scope and specific requirements for the DM system are discussed.

7.1 Types of data

This section describes properties of the different data set types generated.

7.1.1 Raw data

Raw data files contain the set of measured quantities and images associated with the detectors being readout. For each EuXFEL pulse and detector used by the experiment a uniquely identified data record must be created (e.g. a train number and bunch number tag).

The following assumptions are made about the content of raw data files:

- Data records (frames and other information) from the same train will be stored in the same file.
- Data records with the same train number are stored in order of increasing record number.
- Data from consecutive trains from a run may not be stored in the same file. This is due to the finite size of files and DAQ architecture restrictions like writing many concurrent streams of data.
- A run is a period of data taking with the same hardware and software configuration.
- Trains are stored in order of increasing train number in the file.
- Trains and individual data records may be rejected, which means that there may be missing train, or data record numbers. Missing data must be accounted to allow a complete reconstruction of data taking.
- Raw files are immutable; once files are created they cannot be modified.
- Files can be discarded if the data are tagged as not usable.
- As with data records and trains, the lifecycle of a file from creation to deletion must be accounted (logged).

7.1.2 EuXFEL machine data files

The current understanding of photon and electron beam information handling is:

- DAQ and DM are responsible for storing all photon beam line information derived from: the experiments, the photon diagnostic systems, the beam line control systems, etc.

- The DOOCS group is responsible for electron beam (machine) information derived from the beam monitoring system, etc.

A number of problems are associated with this understanding:

- Who store information from the undulator systems which is the interface between bunches and pulses.
- It may be that non of the machine information is of interest to the experiments in which case the photon experiments do not need to store the information. How can this be configured in view of the large number of data types generated by DOOCS?

The following statements can (probably) be made:

- Interfaces between DOOCS and DAQ and DM are needed to exchange information between the two software systems. This would allow both systems to be independent of each other.

7.1.3 Calibration data files

The flexibility of the EuXFEL beam distribution system and detector Front End Electronics (FEE) will allow calibration pulse triggers to be inserted during normal data taking runs. A simple example is taking a pedestal pulse readout type when an empty pulse is configured and delivered. Experience shows that in-run calibration is usually the best way to monitor detector performance over long time periods.

Dedicate data taking runs for calibration purposes are also foreseen and will produce raw data files with the characteristics described in the previous section. These runs are standalone, without beam, where the clock and control system of the FEE provides the required bunch frequency and pulse patterns to be used. At least the following calibration run types may exist.

- Charge injection runs, where a fixed or variable charge (per pixel or pulse) is injected into the pixel readout channel electronics before the pipeline, are used to monitor and tune readout linearity and signal pileup effects.
- Pedestal runs, where no signal input is provided, are used to calculate pixel channel readout noise levels.
- Test pattern runs, where a fixed or variable test patterns (per pixel or pulse) are injected after the ADC, are used to find dropped bits and other errors at the RAM or downstream of it.
- Geometry runs, where x-ray or laser light is directed into the detector, are used to determine the position of the detectors sensors.

7.1.4 Derived data

Results derived from in-run or standalone calibration runs will also require storing as derived data files. This highlights an important feature of data handling at EuXFEL, all information whether raw, calibration, derived or other must be stored in files managed by a common catalogue system. Note that the contents of some files (e.g. the latest derived files) may be transferred to database to ease access by DAQ processes.

7.1.5 Environmental data

Snapshots of conditions prevailing during an experiment run must also be filed. This might contain slow changing information (sensor temperatures, etc.) without an unambiguous train-pulse-number tag or information associated to the train data readout, for example delivered bunch pattern.

7.1.6 Reduced data

If online data rejection is not performed the reduction of data can be performed later using batch mode processing. The reduced data samples can be stored and maintained in the EuXFEL archive or exported to the user home institute for final analysis. The actual scenario for particular dataset may depend on the complexity of the data reduction algorithm and the amount of resources needed to store these intermediate datasets.

7.1.7 Output from user analysis

Users will need to have a possibility of archiving the final results from their analysis in the EuXFEL archive and store any provenance information in metadata catalogues.

7.2 Types of metadata

A primary datasets which consists of all types of data described in the previous section need to be associated with additional information required to provide a full description of the data. This information called metadata is used in all stages of analysis. It describes performed experiments, contains details of the experiment setup, provides information about the collected data, deals with the logical organization of files, defines datasets and collections, stores the history of datasets, describes data quality, or even contain an arbitrary set of user defined attributes or annotations. The different types of metadata expected are described below.

7.2.1 Definition of experiment

Each experiment performed in the facility is associated with an accepted proposal for the scientific study. The relation between particular experiment, the subject of the scientific study and application for the beam time must be preserved.

7.2.2 Experiment run setup

Experiments consist usually of a series of measurements (runs) with predefined conditions. The configuration for every run must be stored. At the end of each run summary information must be created and stored. It may for example contain the reason for run termination, free format description of the experiment and short statistics summarizing the actual conditions during the run.

7.2.3 File related metadata

Data collected by the DAQ system is stored in files and archived in the mass storage system. Each file must be registered in the metadata catalogue with the set of attributes like file location, ownership, size, creation timestamp, and access control list. In case several replicas of a file exists all storage URLs must also be kept.

7.2.4 Datasets and collections

Files can be organized in logical datasets and collections, see Figure 15. A dataset is defined as the group of files with the same origin. Each file may belong only to one dataset. In contrast to a dataset, a collection is defined as a union of files, datasets or other collections. This definition allows one file to be part of many collections. However, the collection may not contain the same file more than once.
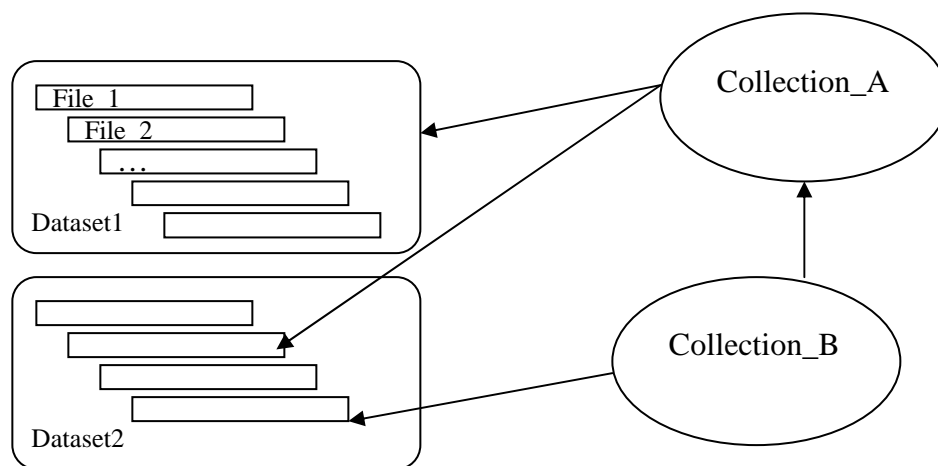
Figure 15 Datasets and collections

Datasets and collections are useful when transferring files using optimized bulk transfer techniques or may ease the definition of input files for analysis jobs. In addition datasets can be used to trace the history of files. Collections are more complex to maintain and standard tools must be provided to allow users simple way of defining collections. Users should be able to exclude some files from one dataset or build a new collection based on previously defined ones using file related metadata attributes.

One should note that the dataset may be treated as an attribute of each file, while collections must be defined as a separate entity.

For each dataset it must be possible to identify its origin. For example, a dataset collected directly from experiment can be uniquely identified by the data type, experiment station and experiment run number. For a derived dataset the origin is another dataset or collection together with full description of the software used for its transformation.

The origin of data is defined for each dataset. Collections do not have this attribute as they may contain several datasets and thus its origin is not defined uniquely.

7.2.5 File content description

Other types of metadata are related to the individual data records or objects within a file. Some of these metadata parameters should be stored in files together with the data itself (encoding, variable type, precision, compression method, definition of data records,…), but other types may require external

catalogue (indices, tags, data quality,…). The later types of metadata may be used at the analysis stage to optimize searches of events of particular interest.

7.3 Software

Software source code, compiled libraries and programs can also be treated as a specific data type and have to be managed through their lifetime. The source code is frequently modified and thus a flexible versioning scheme is mandatory. Software packages usually depend on other software libraries, often forming complex dependency trees. Therefore the software packages must be designed in a way a minimal inter dependency exists between packages and cross dependency is forbidden. For example, if package A depends on B, B depends on C, then package C must not depend either on A nor B.

As in most programming languages (i.e. C, C++) the binary format depends on the compiler suit and the platform architecture where it was built the multi-platform support for libraries and programs is mandatory.

All software needed by DAQ and DM system must be properly managed using a set of standard services and tools.

7.4 Data and metadata format requirements

7.4.1 Files format

Data stored in files has to be properly formatted to ensure easy and long term access to data. A common data model for all types of data should be provided. An appropriate model must be selected based on the following requirements:

- The format must be self describing – this means that even without an external format description the data can be read and understood.

- Data must be encoded in a platform independent way

- Schema evolution – data format and its software will change over a time. It must be possible to read old data with a newer version of software even if the data format is modified.

- The time needed to format and write data to files must be sufficient to sustain design data acquisition rate. The size overhead must be small. Support for data compression would be an advantage.

- Reading performance – both sequential and random access to data in files has to be evaluated.

- Support level for the backend technology has to be reviewed - how big is the user community, what is the user experience, which computing platforms are supported, how well is it documented and maintained

- If needed by analysis the data must be readable after a long time period - this requires a backend technology to be supported for the same time period

- Tools for data content management must be evaluated. These are the data browsers for looking at the content of a file, visualization tools, supported analysis frameworks, etc.

7.4.2 Metadata format

The requirements for metadata formats are not identical to those for data. Metadata information must be easily searchable and in many cases modifications must be allowed for both attribute definition and its content. For example, the data quality attribute may be modified according to the current understanding of data. Therefore, the primary format of metadata must be flexible, extensible and allow for fast query processing. Metadata catalogues are usually implemented using relational databases. However, the snapshots of metadata catalogues should be easily exportable to other formats like XML or PDF.

The standard set of metadata attributes must be defined. Support for extended, user defined attributes should also be provided.

7.5 Authentication, authorization and accounting scheme

Authentication is the process of verifying a claim made by a subject that it should be treated as acting on behalf of a given principal, for example person, computer or service.

Authorization is the process of verifying that the authenticated principal has rights to perform a certain operation. A service may accept or deny processing request coming from authenticated user. Access control list may be used to define authorization.

Accounting is the system of recording, verifying, and reporting the resource usage. It relies on authentication process to associate resource usage to certain principal.

A generic user and service authentication, authorization and accounting scheme is mandatory.

Users must be able to securely authenticate to all services with globally supported methods. Ideally the same method should be used independently of user location, for example when accessing services from local area network while performing an experiment or from user home institute when analyzing data remotely. This means that supported authentication scheme must uniquely identify users and services distributed globally.

The authorization system must allow for a flexible access control list definition for individual users and user groups. Such a scheme must be equally suited for proprietary data and ensure possibility of accessing data from remote sites.

The accounting system must be able to report utilization of various resources and services by users and user groups. These may be allocated storage space size, utilized CPU cycles, or an amount of transferred data in a given time period. Based on that information, authorization system may impose limits and relative priorities between users may be defined.

7.6 Data storage system

Data collected from experiments must be kept on a fast accessible data storage system as long as it is required by analysis or until the data is exported to user home institute. The actual time period when this will be assured depends on the overall data volume, technological limitations and the budget constrains.

A simple model where the data export is performed using external USB disks mounted to a PC or using simple ftp service through the wide area network to the users' home institutes will no longer be valid for the experiments which collect data with very high rates compatible with the EuXFEL pulse pattern. In many cases it will require that the large data sample is pre-processed and reduced close to the data storage system.

The storage system must be capable of accepting the data from the experiments and serve it to clients with minimal latency and adequate bandwidth. A possibility of storing and accessing temporary data files must also be provided to users.

According to the current status of technology the fast data storage systems can be based on magnetic disk systems. By 2014 Storage Class Memory (SCM) technology may be an attractive alternative.

7.7 Data archive

Data archive is defined as a secure, long term supported data storage system. The main role of the archive is to store backup copies of data files and use them for recovery in case the original copy is lost or corrupted. The second role of the archive is to reduce the storage cost for large datasets which are rarely accessed but must be kept for a long time. This can be achieved only if the cost of technology used for implementation of the data archive is lower than for the primary storage system.

All datasets and metadata generated at EuXFEL should be kept in a secure archive, which has a favourable unit cost and proved long archival stability. At the moment tape based technology is the primary candidate for the secure archive but this may change in the future as technologies improve.

Assuming that the archive is based on the tape media two scenarios of data archiving may be foreseen. The first one relies on the backup service performed periodically for each data server. This type of data archiving is suitable for datasets where a small number of files is regularly updated. In that case incremental backup can be easily performed. Large volumes of scientific data have to be treated differently as the simple backup scenario does not give enough flexibility for file management and the performance is not sufficient. Instead, all generated files must be directly stored in the archive and registered in a file catalogue. The catalogue provides global namespace for all files and it is always used when accessing files. The actual disk storage URL (Uniform Resource Locator) of requested file does not have to be known to the user as the catalogue itself maintain the information of all disk copies of the file. This kind of archive architecture also provides the possibility of implementing coherent data access protection scheme.

7.8 Data export

An initial concept of data access assumes two kinds of scenarios depending on the data volume collected by the experiment.

Experimentalists who collect small amounts of data will be capable of transferring the data to their home institute and analyze it there. For that class of users a service must be provided to handle data export from EuXFEL archive in a coordinated manner over the wide area network and by using portable disks. All data and metadata required by analysis must be exportable. This service may also be useful for other experimentalists who deal with much larger data samples but want to use small part of recorded data for testing their analysis software.

Experimentalists who collect large data volumes may not be able to export raw data to external computing site via wide area network, but rather they will analyze collected data close to the archive. If required the results of analysis must be exportable using the same services as with the previous case.

7.9 Computing clusters

Computing clusters close to the data storage system seem to be essential for EuXFEL given the expected data volumes. It is expected that significant part of data analysis will be performed on site. The primary goal of this analysis should be to condense the data volume and make it exportable. In cases where data reduction is not possible the complete analysis would need to be performed on site.

The computing infrastructure must be available on different levels of the DM systems. The CPUs will be used for data formatting and compression, background discrimination, rejection of images without signal, data quality monitoring, and offline data reduction or full analysis.

Computing clusters should be capable of running many instances of the same program (jobs) in parallel, in batch mode. A simple approach uses trivial parallelism techniques where each job processes a different data sample. More advanced techniques are often needed where certain information is exchanged between jobs. This technique requires support for Message Passing Interface (MPI) [9].

The CPU requirements are not well known as most of the algorithms do not exist yet or their developments have just been started. The CPU requirements must be evaluated repeatedly over the next years as the definition of experiments improves, understanding of their efficiencies and background conditions is improved and analysis techniques are developed. The experience gathered by performing analysis of data collected at LCLS should significantly help when estimating the required CPU resources.

7.10 Data Management Policies

The policy for data management system has to be defined and agreed between EuXFEL and its users. The following issues must be addressed:

- What are the general responsibilities of EuXFEL and experimentalists in terms of long term data storage?
- How long will raw data and associated metadata be stored in the archive? An initial assumption is 12 months.
- Who and under which conditions can remove data?
- How long the data derived from user analysis can be stored in the archive?
- What are the limits for storing derived data from user analysis?
- How to define priorities for massive data transfer between different users?
- Rules for space reservation on disks and allocation of the computing power.
- Security rules for connecting users' devices (notebooks, etc) to EuXFEL networks.
- Conditions for running users' programs on EuXFEL computing resources (legal issues, licensing)

Before users are granted access to any computing or storage resources at EuXFEL, they must accept the policy and security rules. These rules must be formally defined and made available to users. Every user must sign them upon formal registration.

7.11 Conclusions

Data management system is expected to deal with all data and metadata produced by photon beam system and experiment instruments. A common data handling model for all experiments should be provided. Secure and long term data storage, efficient data location services, fast file transfers and data export services must be implemented. Taking into account the expected data rates and limitations of the data export bandwidth an appropriate amount of computing power for data reduction or full analysis on site has to be provided. Local and remote access to data and services must be protected using common

authentication and authorization scheme. DM services should be equally suited for experimentalists who deal with large or small data volumes.

8 Data Management Architecture

In this section the proposed architecture and implementation of the DM system are described.

8.1 Architecture

The proposed architecture of data management system is shown on Figure 16. This figure extends the schematic view of DAQ system shown in Figure 11. It shows additional details relevant for DM processing including the flow of data beyond the archive layer.
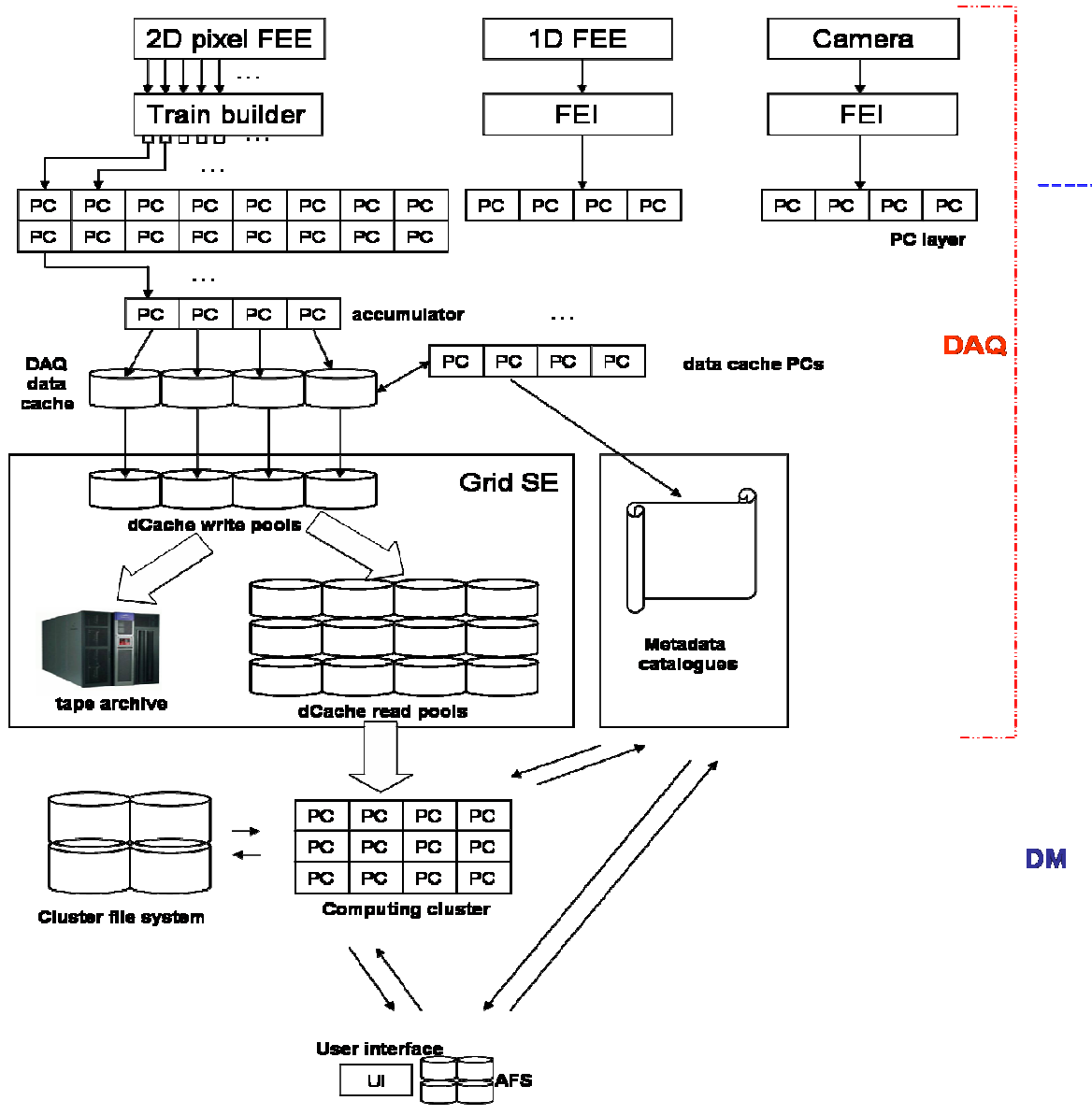

Figure 16 Data Management System Architecture

The following DM components can be identified:

- PC layer

- DAQ data cache

- Data archive – currently assumed to be a tape storage

- dCache write pools – disk based cache in front of tape archive

- dCache read pools – disk based storage system

- Metadata catalogue service

- Cluster file system – used as a temporary storage for user data

- PC cluster – computing farm available for users

- User interface – working group servers for interactive analysis, software development, batch job management, web interface – multiple instances at EuXFEL and at remote locations

- AFS – shared file system for home directories

Data coming from detectors, via *FEI*s (ie: *Train Builder*) will be monitored, formatted and possibly rejected and compressed on the *PC layer*. The number of PCs in this layer can be scaled according to CPU requirements. At least one PC per data channel (FEI output link) is foreseen. If the CPU requirements are more demanding each channel can be served by a larger number of PCs. In this case an additional *accumulator* PCs is inserted with the task of accumulating data from the other channel PCs, formatting the data and storing it in files on *DAQ data cache*. At this point train ordering can be done and the size of the files can be optimized before writing into the data cache. Once data files reach the data cache, experiment operators may perform more advanced data quality monitoring (DQM) if appropriate algorithms exist. If data is found to be of bad quality the files should be discarded. If collected data passes the DQM checks files are transferred to the *data archive* and registered in the *metadata catalogue*. *Data cache PCs* will be used for DQM and as a client for data archiving.

Data from successful experiment runs are transferred to the EuXFEL Computer Centre (CC) using high speed optical links. In front of the tape archive a disk based cache storage system is required *(dCache write pools)* to optimize archiving procedures and at the same time pass the data stream to the fast storage system *(dCache read pools)* where users' can access data for further offline analysis. Separation between write and read pools ensures that there is a minimal interference between online and offline access and secures the data archiving processes from unpredictable user access patterns. Files will remain on the offline disk storage until they are pre-processed *(computing cluster)* and exported out of EuXFEL CC. Reduced data can be exported using reliable Grid services. Once this is done data can be removed from disk storage system while the tape copy may be kept longer until the full analysis is performed and results are published. The time when the data are completely removed from EuXFEL mass storage system (including tape backup) will depend on the overall data size, available resources and the EuXFEL DM policy.

All files stored in the archive will be globally managed using metadata services compatible with Grid middleware. The Grid based authentication scheme can provide the same authentication methods for users independent of the working place. Compatible authorization schemes will ensure that proprietary data will not be publicly accessible until the owner of the data decides to share it with other scientists. Services will be provided to users for easy searches and lookup to the archive content. Computing

clusters may be part of the local Grid infrastructure or can be built as a specialized PC farm with support for advanced parallel processing using MPI. Standard *User Interface* computer nodes will be available for interactive work and a lightweight web based interfaces will be provided for convenience.

The assumptions described above can be summarized as follows:

- A tape archive will be used for long term data storage
- Data from unsuccessful experiment runs should not be stored
- Raw data stored in the archive will remain immutable
- A disk storage system (dCache) will be used as the front end to the tape archive
- Data archive and disk cache system will be part of the Grid infrastructure
- Computing clusters will be used for offline data analysis
- Grid services will be provided for transferring data outside EuXFEL CC
- Data will be protected using Grid authentication and authorization methods
- Metadata services will be used to describe the content of the archive
- User interfaces for data access will be provided

The architecture proposed above provides a solution for the EuXFEL data management. Uncertainty exists with respect to the available technologies and their costs in 2014 when the first data are expected. The architecture has been designed to take into account that certain components may be implemented using newer technology if available. In particular the implementation of the data archive is currently assumed to be based on tape media which currently ensure the best level of security for long term data storage and are cost effective compare to disk based storage. The functional requirement of the data archive is to guarantee an acceptable degree of data security and non tape implementations may become economically (investment and operational) attractive before EuXFEL start up. The technological outlook is described below.

8.2 Technology trends

This section reviews the hardware technologies which can be used in the DM.

8.2.1 Tape archive technology

Figure 17 shows the trends for tape storage capacity and transfer rates for the next few years using as an example LTO technology [10]. Based on these available trends further estimations are made until 2015 covering the first period of EuXFEL operation.
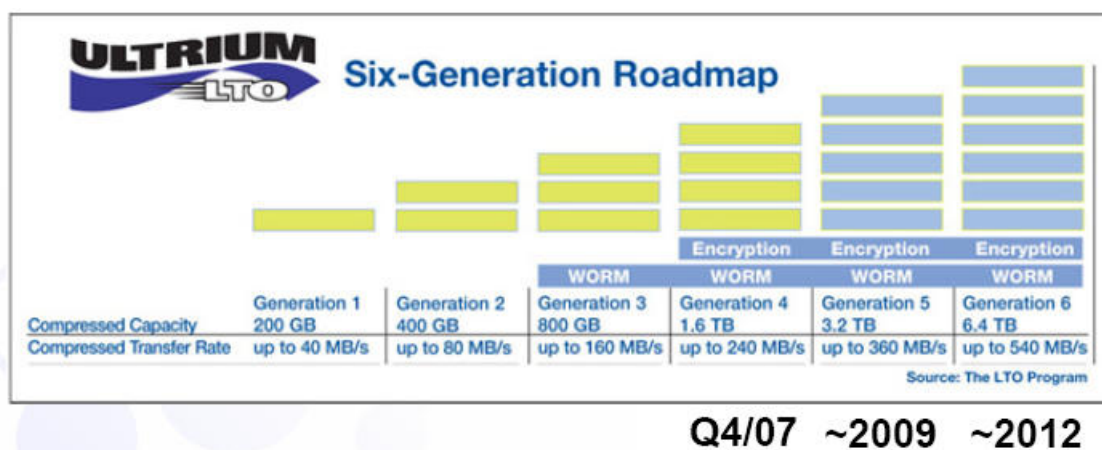
Figure 17 LTO tape capacity and access speeds roadmap

The figure shows compressed capacity and compressed transfer rate. As the native tape compression factor is not known (i.e. due to already compressed images) it is safe to assume that the real tape capacity and maximum transfer rates should be divided by two. This means that in 2012 one should expect 3.2TB tape size and 270MB/s transfer rate per drive. The maximum number of drives per robot is estimated to 64 and number of tapes per robot to 10000. This gives the total capacity of 32PB and nominal transfer rate ~17GB/s per tape robot. One should expect further increase in tape sizes and transfer rates by 2015. The tape capacity should be between 5-6.4 TB and the transfer rate around 400MB/s per drive. It is worth to note that the close to maximum transfer rate per drive can be achieved only if the data are written to or read from tape sequentially with minimal overhead for drive head positioning.

A fully equipped tape robot could handle the overall data volume up to 50-64PB with supported overall transfer rate up to 25GB/s. These numbers are important if the system has to be up scaled to fulfill increased requirements from experiments.

8.2.2 Fast storage system technology

Magnetic disk technology is currently used to implement fast access to the large data volumes. The growth in the disk capacity (density) is estimated to double every 2.5 years which leads to about 4TB per disk capacity in 2014/2015. In contrast to that the input/output rate and the access latency is not expected to follow the density growth resulting in much slower data rate per volume unit. This may lead to problems where large data volumes stored on magnetic disks cannot be easily accessed with the required bandwidth. Especially the random data access patterns used by analysis programs may be problematic and special optimizations may be required according to the concrete data access pattern.

As an alternative to the magnetic disk technology Storage Class Memory (SCM) developments should be followed over the next years. It is assumed that SCM will eventually overtake magnetic disk technology mainly due to a much improved latency and data rates per capacity unit. The main reason for this improvement is that SCM does not rely on moving mechanical components as opposite to the magnetic disks. It is however very difficult to predict the time when this technology is available and corresponding cost per storage unit is affordable for large scale data storage system.

8.2.3 CPU developments

The CPU clock frequency is not likely to be significantly increased over the next years. This is mainly due to the thermal issues. Instead, the increase in computational power will be assured by packing

49

multiple CPU cores on a single chip. Assuming the continued increase in number of cores per chip, there should be 32 cores per chip (socket) available in 2014. As the number of cores per CPU increases and clock frequency stays on the same level the techniques using multi-threading or MPI based code are getting more important. A single thread program will not be able to benefit from the increased computing power unless the full computation task can be easily divided to many independent sub-tasks, each of them analyzing different part of data sample. Another issue which must be carefully watched is related to the memory access, which may get slower if the number of cores increases.

8.3 Data archive implementation

Potential implementation of data archive is described below.

8.3.1 Data storage systems

Currently it is assumed that the tape technology will be initially used for the secure data archive and the disk based storage system will be used for fast storage system implementation.

The DM architecture is however designed in a way that underlying technology can be replaced without affecting the general DM concept. Possible configurations of fast data storage system and data archive are: disk-tape, disk-disk, SCM-tape, or SCM-disk.

8.3.2 Disk cache for data archiving

The dCache [11] system is planned to be used as a front-end to the tape archive. One of the roles of this component is to optimize data archiving process. Deferred writing of files will optimize usage of tape drives. A configurable amount of data will be collected on magnetic disk based system or SCM and then flushed to tape when enough data is available to fill a single tape. The multi pool architecture of dCache system allows input data streams to be directed to different servers which makes the system highly scalable. It is assumed that files will be written at the same time to the tape archive and transferred to offline disk storage (dCache read pools) for further access by users. In this scenario archived files will be used only if data on disks are no longer present, for example when data are needed to be reanalyzed after long time period or in case a disk copy got corrupted. All archived files will be registered in data catalogues allowing easy way to view available samples and searches through the archive. The dCache system must be located close to the archive. File transfer between DAQ data cache and dCache write pools needs to be performed using parallel data streams to achieve required transfer rate.

8.3.3 File format implementation

It is currently assumed that the EuXFEL file formats will be based on the Hierarchical Data Format 5 (HDF5) data model [12].

HDF5 provides general methods for formatting and storing scientific data in files. It is capable of storing complex data objects and associated metadata in platform independent format. The data representation is self-describing in the sense that the format defines all the information necessary to read and reconstruct the original objects of the abstract data model. Software library support is provided on a broad range of computational platforms. High level APIs are available for C, C++, Fortran 90, Java and IDL languages. HDF5 addresses issues like access time and storage space optimizations (compression). IO library provides interface to MPI. A collection of generic tools exists for managing, manipulating, viewing, and analyzing data in HDF5 format. HDF5 based data format is also planed to be used at LCLS.

The actual data record format for all instruments must be designed and implemented in a close cooperation with detector developers and user community. Performance of the software has to be evaluated keeping in mind high data rates.

8.4 Metadata catalogue services

A Chimera [13] catalogue which is part of the dCache software distribution suit will be used as a central namespace provider for the primary EuXFEL archive. Chimera presents all registered files in a view of regular file system and exports it via NFS4 [14] protocol. The only difference compare to the standard file system is that the file transfer must be done using dCache compatible API. All other files attributes like ownership, size, creation timestamp, access control list are compatible with the NFS4 API.

The data management services in the Grid environment require the presents of additional catalogues. Every file on the Grid is identified by its Grid Unique Identifier (GUID). The GUID is automatically generated when a file is registered in the Grid Storage Element (for example in dCache). A user friendly identifier called Logical File Name (LFN) will be assigned to each GUID. The mapping between GUID and LFN is implemented in the LFC File Catalog. This catalogue is also capable of handling multiple replicas of the same file across Grid sites.

An additional catalogue must be used to manage other types of metadata described in section 7.2.

As described above there are at least three catalogues needed for proper data management. The Chimera namespace provider and LFC Grid File Catalogue may be used as out of the box services. The metadata catalogue, which deals with datasets, collections, EuXFEL predefined attributes, experimental setup, software repositories, data provenance and user defined schemas has to be implemented. As far as possible the standard software components must be used for that. AMGA [15] metadata catalogue is one of possible choices as the base for development of such metadata schemes. Other option is to base a development on Java EE software tools.

The following requirements are defined for metadata catalogues:

- Metadata catalogues must be based on database technology
- Database services must have very good performance and must be scalable to hundreds of millions of records
- The database structure must allow efficient metadata partitioning
- Online and offline service must be integrated
- Definitions of policies for metadata manipulation are required:
    - ❖ how to insert metadata during data taking
    - ❖ which metadata attributes can be modified and which are immutable
    - ❖ procedures for data quality metadata handling
    - ❖ datasets and collections
- Wide area access to metadata catalogues (web services) must be provided
- Support for a single global authentication and authorization mechanisms compatible with other services used at EuXFEL

- APIs for metadata access for the used languages must be provided (C++, C, Java)

- Compatible formats for database and XML, XML snapshots, conversion tools are needed

- The secure backup of metadata catalogues must be provided

8.5 Authentication and authorization

It is proposed that the storage and computing system for EuXFEL will be integrated with the Grid infrastructure. Authentication and authorization schemes in the Grid environment are well established and standard methods are provided. In this context a user is identified by a personal X.509 certificate. Authorization scheme for access to computing and storage resources is based on the Virtual Organization Membership Service (VOMS). The user must be registered in the appropriate Virtual Organization (VO). Members of a VO can be organized in hierarchical tree of groups with no limitation on its depth. For the purpose of EuXFEL a new VO – xfel.eu – has been defined and will be maintained at DESY. A combination of NFSv4 access control list scheme and VOMS could be used to protect access to proprietary datasets. This scheme would give a possibility to define fine grain access control list. It will be supported by the Grid storage element implementation based on dCache software.

The same authentication method based on personal certificate can be used for non-grid services. In this case however users must also have a record in the local user registry. To allow fine grained authorizations, the grid certificates must be automatically translated into kerberos5 [16] ticket through a gateway mechanism. Access to any resource inside EuXFEL network infrastructure can be fully controlled. In addition the remote access to the raw and derived experimental data will be possible at any time as long as the owner of the data holds a valid certificate and regardless if the person is still registered as an active user on site.

The gateway mechanism may be based on the globus gsi implementation of a gsi-enabled openssh [17]. Such a gateway has recently been installed in DESY at the National Analysis Facility (NAF) of the HGF-Terascale project [18], and has been used by various groups from the HEP community to perform parallel batch analysis outside the standard Grid computing scheme, but fully utilizing the standard Grid middleware. Other possible approach is based on Shibboleth system which can handle transparently short living certificates.

The translation of grid-credentials allows defining data access authorization on a personal, group or any arbitrary level, such that data remain fully protected from un-authorized access.

A detailed case study must be performed to guarantee that this authorization scheme can be used by all services provided to the EuXFEL user community: access to web resources, archived data, metadata and software repositories.

It is expected that by 2014 the certificate based authentication will be more popular and the procedures will be very well established. Since this technology is currently not yet well known in the photon science user community practical procedures will be described and made available to the users.

8.6 Data access methods

8.6.1 Offline disk storage system

Appropriate numbers of disk based storage servers are required for the purpose of data analysis. The amount of required disk storage space depends on the time the average analysis is performed and the overall data rate. The disk storage will be managed by the dCache system. Data can be retrieved using

the gsidcap protocol which is compatible with proposed authentication and authorization scheme. Also support for NFSv4.1 interface is expected on a time scale of 2 years, which should further simplify the access to archived data. Wide area network transfer can be performed using Grid based File Transfer Service (FTS) between Grid sites, if required. Users will be provided with lightweight tools (ie data portal), where the transfer of small datasets can be performed between EuXFEL data storage system and personal computer at any location in the world up to the network limits.

8.6.2 Cluster file system

A cluster file system connected to the computing farm via an InfiniBand [19] switch may be used when the short latency for data access is required. A typical use case for this scenario is running parallel MPI based analysis jobs. The cluster file system may also be used as a temporary space for analysis jobs output.

8.7 Computing clusters

One of the possible implementations of a computing cluster can be based directly on the Grid middleware where the computing resources are shared between various Virtual Organizations according to the predefined policy. This provides a possibility to better utilize computing resources by borrowing CPU cycles from other VOs when needed and giving them back when there is more CPUs than required.

Another possibility is to build a dedicated cluster for EuXFEL users. This may be attractive in case the MPI support is required (exchange of messages between jobs). In that scenario the network connection between computing nodes requires short latency. Typical implementation is based on InfiniBand network. Batch job management software supporting MPI and Grid authentication scheme is required. The idea is to use experience of DESY IT division which has built and maintains similar cluster for NAF users.

At the moment it is not possible to define which implementation will be better suited for EuXFEL analysis patterns as the majority of algorithms are not yet known. An improvement in understanding the computing requirements is needed.

8.8 User Interface

Each user who gets access to any computing resource on site must be registered in the user database. Functional accounts may still be useful but each person with access to it must be registered and posses a private account. It is assumed that the software developed at DESY for handling user accounts will be installed and used by EuXFEL.

Each registered user can be provided with a small amount of disk space for home directory which can be shared between all computers on site and provide consistent setup for user preferences independent of the computer used. It is currently assumed that this space will be implemented using Andrew File System (AFS) [20] in linux type environment. Windows home directories may be also needed.

A set of computers will be provided for interactive login. They will act as a user interface for the data management services. Basic computing services like printing, wireless access, possibility of connecting private notebooks to the network should exist.

The management of xfel.eu VO will be supported by DESY – in particular it includes the management of users' certificates and interactions with Grid Information System.

An easy to use, preferably web based interface will be provided to users who want to analyze data remotely. Web based access to metadata catalogues, interface for file transfer services and software repository is mandatory.

8.9 DM software

As described in previous sections, DM consists of several subsystems. Each of them requires appropriate software to implement and maintain its designed functionality. As much as possible this software should be based on standard, third-party middleware components (databases, data format, standard authentication, etc). They need to be carefully selected and evaluated to ensure that their functionality, performance and maintainability could meet EuXFEL DM requirements.

On top of these middleware software components EuXFEL specific services and tools must be designed and implemented. As an example one can consider the software handling data format. Using HDF5 software which provides tools for data format realization a specific data record format for each type of detector needs to be implemented. Furthermore, different data records (from different detectors) need to be consolidated and mapped to the physical file structure. This requires development of features like network data record transfer, synchronization of data records, compression, IO buffers and selectors. High level interfaces for users must be defined to ensure easy access to data for offline analysis. A typical EuXFEL user should not be exposed to the complexity of this software, but rather be able to concentrate on scientific issues related to the data analysis.

8.9.1 Development path

It is planned to develop as much as possible common data management tools for all EuXFEL instruments. DM software development must be connected to the real scientific use cases through the involvement in data analysis process with selected experiments performed at FLASH and LCLS.

8.9.2 Software repository

Software used by DAQ and DM systems must be centrally maintained in software repository. Such repository should deal not only with the source code management but also provide tools and generic environment for building libraries and programs. In addition any external software component used at EuXFEL must also be maintained.

The central repository is also needed for users' code in case the data provenance is required. In this case an appropriate catalogue must keep information about the data origin, including the application package used to transform the data, software release tag and parameters used to run the program.

In the following a list of requirements for the software repository is given:

- source code maintenance
- authentication and authorizations based on Grid certificates, preferably supporting VOMS services
- wide are network access to software repository
- standard environment for developers
- coordinated and automatic software building procedure (nightly builds) on various operating systems
- support for unit tests

- environment for execution

- flexible software distribution scheme

- definition of support model for multiple operating systems, architectures and compilers

- integration of online and offline software

Initially it is assumed that the source code repository will be based on an open source system SubVersion [7]. Another open source package called cmake [21] will be used to facilitate C and C++ software building, testing, packaging and installation. It is known that cmake supports broad range of operating systems, IDEs and compilers.

For Java SE and Java EE software the NetBeans Integrated Development Environment (IDE) [6] and Ant [22] systems will be used for development and software packaging. An open source application server GlassFish [5] will be used to ease deployment and launching the client applications within internal network environment.

8.10 Risks

The following risks can be identified:

- Technology trends – anticipated trends in technology development may not be correct. This may lead to changes in the implementation of the secure archive and fast storage systems. The complete replacement of the technology may be required or the size must be adjusted according to the available bandwidths and the capacity per storage unit.

- Computing clusters – computing cluster may require different types (InfiniBand, etc.) of hardware depending on the analysis techniques as developed in 2014.

- AAA – if the authentication, authorization and accounting schemes based on personal certificates cannot be applied to the EuXFEL environment (ie. they are too complex for users), they must be replaced by different methods.

- File format – HDF5 has been initially chosen for the data format implementation as it satisfies most of the requirements listed in section 7.4.1 and its probable use at LCLS and other Light Source Facilities. However, the usability and performance of the corresponding software needs detailed evaluation in the context of the high data rates expected at EuXFEL.

.

9 Summary

In this note a description of the DAQ and control and DM systems to be implemented at EuXFEL are described. DAQ and control address issues like configuring and controlling hardware and reading out data from it. DM address data management issues like how long data is stored for, what data format is used, what metadata is provided and how data can be accessed and analyzed.

DAQ and DM requirements and solutions are clearly tightly interrelated which has lead to partitioning the implementation into five layers:

- Front end electronics of the detectors, stepping, etc.

- Front end control interface to the front end electronics

- PC layer cpus for data accumulation, file formatting and monitoring

- Data cache for onsite temporary data storage and analysis

- Archive data storage and analysis clusters

Multi layered architectures with well defined APIs decouples layers should provide flexibility in view of future software and hardware developments.

User interaction with the machine control system, photon beam line optics, photon diagnostic, laser pumps and experiment instruments will be provided by the photon control DAQ system. This will offer the normal suite of web applications and client programs that allow the user to configure, control and monitor an experiment.

Integration of exotic instruments into the photon control system will be allowed and supported provided instruments:

- satisfy and support the timing interface requirements of the timing receiver board provided by WP28

- are compliant with DAQ and control interface requirements for network configuration, monitoring and readout, which are currently being defined.

Online monitoring of images acquired will be performed by computers in the PC layer, immediately following the instruments, or quasi-online by computers accessing the data cache. The analysis programs operate on data files and will be either:

- standard analysis packages providing image pictures, histograms (radial and horizontal projections, etc.) and sums.

- Experiment specific packages provided and the experiment.

All experiment data files are written through the PC layer into the data cache where it is stored for 2-3 days. The data cache provides onsite local storage and analysis processing capabilities; the later should be used for additional quality control before finally committing the data to archive.

The DM realm is entered when data files are transferred from the data cache to the storage archive. Data and metadata files will be stored on disk and archived to secure (e.g. tape) storage. The disk copy will be deleted when not enough disk space is available.

The data access model currently envisaged for EuXFEL is that the bulk of experiment data is analyzed onsite in computing clusters with batch job submission and data organization tools, Grid middleware, developed in the context of e-Infrastructure being used. Lightweight client access allowing small amounts of data to be transferred over WAN to home institutes for analysis tuning are foreseen.

The following policy decisions relating to DM usage are in place:

- the initial size of the data storage system will be 10PB scalable to 100PB.

- data will be archived for 1 year before deletion

- a reasonable amount of archive storage will be provided to experiments

- a reasonable amount of computing power will provided to experiments

Access to DAQ and DM resources used for data taking and offline activities will be restricted to users groups through the coherent authentication and authorization schemes. Personal certificate based authentication scheme will be tailored to the needs of the photon science community, taking into account ease of use and ensuring unique identification of users across the glob. Fine grain authorization scheme will offer the possibility of protecting access to data as defined upon accepting the proposal for experiment.

9.1 User activity summary

The list of user group activities from proposal of an experiment to publication are:

- a proposal is submitted via the EuXFEL proposal portal defining beam line, detector and other requirements

- if the proposal is accepted personal certificates are generated and user's group is defined and an entry in portal is opened for the experiment. The user group is informed if the proposal is rejected

- using the portal entry and interaction with the beam line scientist time is allocated and a schedule for any preparation, in particular, EuXFEL side work like detector integration, is defined.

- the user group arrives for the allocated preparation and data taking period and access the corresponding beam line resources which include a fully functional soft and hardware DAQ and control environment (run control, monitoring PCs, target injection and positioning devices, beam line control, etc.) using their certificates.

- experiment runs are performed and data are cached.

- cached data which should be kept is committed to the archive. The online status of the experiment and data cache can be seen through the beam line control system, cached and archived data via the EuXFEL portal.

- whilst onsite data can be transferred to non EuXFEL storage media (laptop disk) using a web client. Data on EuXFEL supported storage is protected by the authorization mechanisms. If a user makes a copy of the data outside of EuXFEL storage infrastructure the user is entirely responsible for protecting the access to that data.

- access to and analysis of data on EuXFEL storage is performed within the model defined above and uses web and other tools which will be provided for job submission, interactive analysis, etc. Whether data are analysed onsite or offsite depends on factors like: total volume of data, ability to transfer data to the site where analysis is to be performed, etc. Data files which have migrated to secure storage will be migrated back to fast storage transparently.

- once analysis is complete the user group can delete data files from EuXFEL storage. If storage has not been freed after the guaranteed 1 year retention period the data may be deleted if the user group does not respond to explanation requests.

10 Glossary

AGIPD	Adaptive Gain Integrating Pixel Detector
API	Application Programming Interface
BPM	Beam Position Monitor
C&C	Clock and control FEI control development for 2D pixel detectors
CC	Computer Centre (EuXFEL)
DAQ	Data Acquisition and Control
DC	Data Cache storage and computing layer
DM	Data Management
DQM	Data Quality Monitoring
DSSC	DEPMOS Sensor with Signal Compression
FDE	Femtosecond Diffraction Experiments
FEE	Front End Electronics
FEI	Front End Interface
HED	High Energy Density matter
LPD	Large Pixel detector
MID	Materials Imaging Dynamics
MPS	Machine Protection System
PCL	PC readout and computing Layer
R&C	Rejection and Compression
SASE1..3	EuXFEL SASE beam lines 1 thru 3
SCS	Soft x-ray Coherent Scattering
SPB	Single Particles Bio-molecules
SQS	Small Quantum Systems
TB	Train Builder FEI readout development for 2D pixel detectors
TR	Timing receiver board
U1..2	EuXFEL undulator beam lines 1 thru 2
VO	Virtual Organization
VOMS	Virtual Organization Membership Service
XGMD	X-ray Gas Monitor Detector
XHEXP1	EuXFEL experimental hall and office building
XTD	EuXFEL Tunnel Distribution

11 Acknowledgements

Experiments: Ch.Gutt, H.Chapman, H.Ihee and J.Hajdu

2D pixel: J.Coughlan, M.Wing, P.Goettlicher, A.Kugel, M.Zimmer, H.Graafsma, G.Potdevin and I.Sheviakov

IT: V.Guelzow, P.v.d.Reest, K.Woller, K.Ohrenberg, Th.Witt, U.Toeter, M.Gasthuber, B.Lewendel, P.Fuhrmann and F.Schleunzen

Infrastructure: P.Dost, S.Feuker, J.Havlicek, F-R.Ullrich and E.Negodin

Safety: S.Schrader

Photon diagnostics: J.Gruenert and K.Tiedtke

Beam line optics: H.Sinn and A.Trapp

Management: A.Schwarz and Th.Tschentscher

Appendix A   DAQ and control infrastructure

In this section the DAQ and control infrastructure implementation for racks, rooms and networks are described. The lists and related numbers generated during this process have been used when costing the system.

The scope of the discussion is restricted to the following areas:

- photon beam line optics, diagnostics and vacuum systems in the tunnels
- computing services rooms in the experimental hall
- experiment hutch and laser rooms in the experimental hall

Note that the infrastructure required for implementing the undulators and their control systems in the tunnels are not described in the computing TDR.

Note also that the DM data storage and computing infrastructure is not described in this section. It is assumed that DM infrastructure is provided by an offsite service provider such as DESY-IT when implementing the design outlined in Section 8.

A.1   Implementation documentation

Documents containing detailed descriptions of the hardware infrastructure implementation of the tunnel [23] and computing service [24] areas have been prepared. These documents list and describe the currently proposed implementation solutions for a larger infrastructure set (power, cooling, racks, networks, safety, compressed air, air-conditioning, lighting, UPS power, etc.) than is required to understand the DAQ and control system (racks, rooms and networks) implementation. Interested parties should read the documents provided to obtain a full understanding of the infrastructure implementations.

The infrastructure requirements for the experiment hutch and laser room areas are currently being defined. For the purposes of this note it is assumed that the DAQ and control requirements are small, approximately two full height (~220cm) electronics racks per room, and that any additional resources will be provided by hardware located in the computing service rooms.

A.2   Tunnel area infrastructure

It is assumed that all infrastructure consumables, such as water and power, needed by instruments in the tunnel will originate from XHEXP1. This simplifies considerably the implementation described below.

It is intended to use EuXFEL standard [25] electronics half racks mounted under the beam line for installation of all photon diagnostic instruments, optical control elements and vacuum systems in the tunnel sections. It proposed to use the standard EuXFEL tunnel rack currently being specified by the EuXFEL technical coordination group. The standard rack has the following characteristics

- standard 19 inch wide crates
- approximately 130 cm (~28U) high to fit under the beampipe
- 3 racks form a closed rack triplet unit with shared cooling, fire safety etc.
- ~3kW closed cycle water cooling unit per rack triplet
- controller unit (network controlled) per rack triplet

In tunnel sections containing electron beam line rack units will be shielded from radiation. Although currently intended for use with xTCA crates, with front-to-back airflow, the racks should be modifiable for other crate standards, e.g. VME (bottom-to-top), which may be needed by photon beam line or diagnostic systems. The racks will therefore contain electronics crates, patch panels, power supplies, stepping motor controllers, etc.

A provisional allocation [23] of electronics racks required has been made, see Table 11. Groups of racks associated with optical and diagnostic control systems, which are geographically close to each other, are grouped and identified by a per beam line alpha-numeric code, e.g. A1 and B1. Vacuum system racks appear as singletons placed periodically along the tunnels to satisfy the distributed pumping requirements of the beam pipe and are identified with by a X-numeric code, e.g. X1 and X2. Singleton racks also provide locations for switches serving the ~180m tunnel lengths. Control cables and network connections are routed over a dedicated cable tray on the tunnel wall. The cable tray allows inter-rack, inter-group, group-to-group and group to XHEXP1 connections.

	SASE1			SASE2			SASE3			U1			U2		
	Group	#	m	Group	#	m	Group	#	m	Group	#	m	Group	#	m
electron and photon beam line	X1	1	946	X1	1	856	X1	1	406	X1	1	584	X1	1	224
	X2	1	856	X2	1	766	X1	1	316	X2	1	494	X2	1	224
	X3	1	766	A1	1	763	A1	6	233	A1	1	482	A1	1	145
	A1	1	726	B1	4	734				B1	5	453			
	B1	4	697	X3	1	676				C1	1	431			
	X4	1	676	C1	5	679				X3	1	404			
	X5	1	596												
	C1	5	571												
photon beam line only	D1	1	511	D1	1	608	X3	1	226	D1	5	400	B1	2	140
	X6	1	496	X4	1	586	B1	8	183	E1	2	359	X3	1	114
	E1	8	449	F1	1	512	C1	1	141	H1	3	328	C1	9	104
	F1	1	409	X5	1	496	X4	1	136	X4	1	314	D1	3	88
	X7	1	406	X6	1	406	E1	1	130	X5	1	224	E1	1	63
	X8	1	316	G1	8	450	F1	1	75	F1	1	178	X4	1	44
	G1	1	258	H1	1	408	X5	1	46	X6	1	134	F1	1	3
	X9	1	226	X7	1	316	G1	1	40	X7	1	44			
	X10	1	136	I1	1	258	H1	1	3	G1	2	2			
	X11	1	46	X8	1	226									
	H1	1	6	X9	1	136									
				X10	1	46									
				J1	2	3									

Table 11 Beam line control system rack allocation - for each rack group identifier the number of racks and the groups furthest distance from XHEXP1 is shown.

The current design of the EuXFEL standard tunnel rack foresees three racks mounted in a holder cage with shared power, cooling, safety and other services. The total number of tunnel racks required by the photon beam line systems is currently 59, which corresponds to 30 triple rack units. Triple rack units will be either shielded or unshielded if located in tunnel sections with electron and photon or photon only beam lines, respectively.

A.3   Hutch and laser room infrastructure

EuXFEL when fully commissioned is expected to have a total of ten experiment hutches located on the ground floor (UG01) and a number of laser rooms. It is assumed that all Front End Electronics required to control the instruments and lasers will be located close to the instruments. As much down stream, of the FEI, electronics as possible should be located in the computer service balcony rooms. Maximizing the amount of electronics in the balconies will reduce the infrastructure requirements close to the experiments.

A.4　Computer services room infrastructure

XHEXP1 balcony rooms UG01/28, 29, 30 and 38 will be used to house the computing services for DAQ and control and IT. These rooms, see Figure 18, are ideally positioned with respect to the tunnels, which enter the hall immediately below the balconies, and to the experimental hutches and laser rooms on the hall floor.
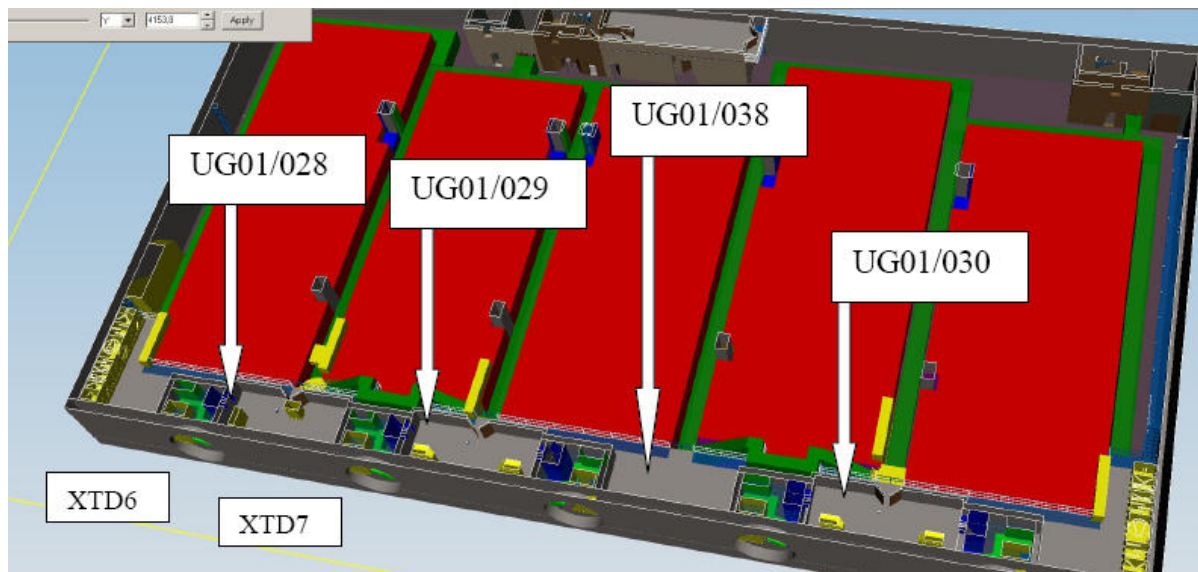
Figure 18 Computing service balcony rooms

All rooms have the same height and almost the same floor size and it is possible to install twelve standard 220cm electronics racks in a single row, as shown in Figure 19. In the final implementation standard computer centre racks will be used which allows using the same costing per rack as assumed for DM rack installations. Concentrating computer services in a small number of geographically close rooms foresees sharing of common infrastructure and improves maintenance and installation operating efficiency.

Figure 19 Balcony room rack footprint

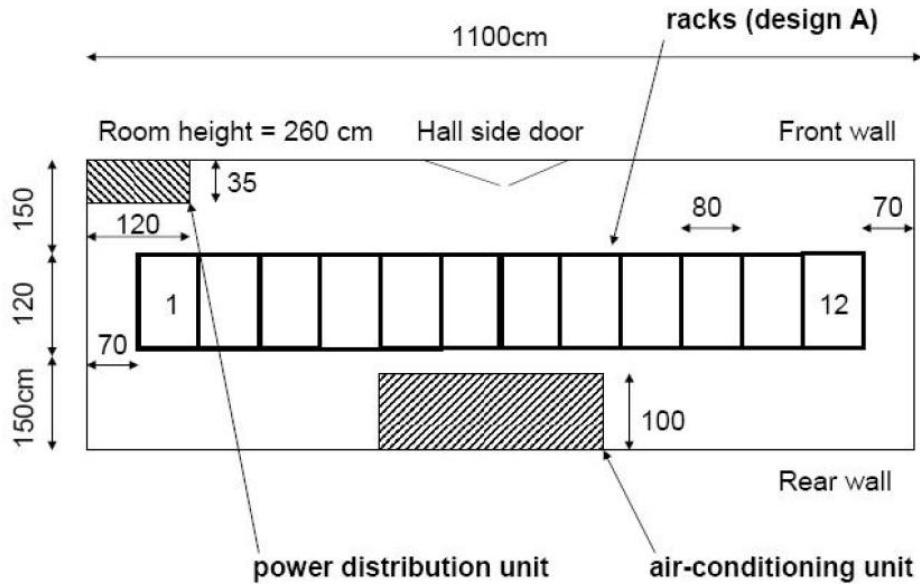The anticipated rack usage per room is shown in Table 12. In principle only five half rooms are required to service the five EuXFEL tunnel, but given the uncertainties associated with the final computing requirements all four rooms should be planned for full use with a final size decision deferred until ~2011.

Rack number	Power (kW)	Usage (owner or architecture layer)
1	~0	Patch panels (IT / DAQ/DM)
2	~2	Network switches (IT / DAQ/DM)
3	8	1PB disk storage (DC)
4	8	1PB disk storage (DC)
5	20	PC layer (PCL)
6	20	CPU (DC)
7	20	CPU (DC)
8	20	PC layer (PCL)
9	8	1PB disk storage (DC)
10	8	1PB disk storage (DC)
11	~2	Network switches (IT / DAQ/DM)
12	~0	Patch panels (IT / DAQ/DM)

Table 12 Computer services rack usage

Each balcony room connects to its nearest tunnels, see Table 13, through labyrinths at either end of the room. Two additional breeches in the hall side balcony room wall are used to interconnect, via a cable tray running above the balcony walkway, the balcony rooms and receive control cables and fibres on ceiling mounted cable trays coming from the hutch and laser room areas and other hall areas requiring IT connections.

UG01 room	Connected to tunnels	
28	XTD6 – SASE1	XTD7 – U2
29	XTD7 – U2	XTD8 – U1
30	XTD8 – U1	XTD9 – SASE1
38	XTD9 – SASE1	XTD10 – SASE3

Table 13 Balcony room tunnel connections

A.5   Network infrastructure

The implementation of the network is a key element in the DAQ architecture implementation. The network design is driven by the following concepts:

- links from all instrument areas (tunnel, hutch, laser room, etc.) terminate in the computing service balcony rooms
- links from other hall area, XHEXP1 office networks, etc. may terminate in the balconies
- termination end points consist of a patch panel and a switch rack
- end points are interconnected
- point-to-point connectivity will be used where necessary, e.g. FEI to PCL.
- Ethernet is used throughout
- link speeds are determined according to requirement
- all DAQ links are physical media and not wireless
- the network will be partitioned and fire walled to ensure correct operation
- network performance and error conditions will be monitored by software
- a ~50 GB/s link will link XHEXP1 with the DESY-IT DM provider

Implementation notes concerning the network are described in the next sections.

A.5.1   Connection endpoints

All DAQ and control networks will be connected to endpoints in the computer service balcony rooms. Each endpoint will consist of a switch rack, containing single or multiple switch enclosures, and a patch panel rack. The patch panel rack decouples the switch installations from incoming and outgoing

fibre and copper link cables. The maximum number of links which can be handled per endpoint is ~500 which will satisfy startup requirements; again a review of the size of the system will be useful in 2011.

A.5.2   Hutch and laser room connections

Hutch and laser rooms will be connected to computer services in the computing rooms by links running from rack mounted patch panels to the computer room endpoints. The route taken by the links, either via ceiling or floor mounted cable trays is not yet fixed, but the expected length of the link cables will not exceed 90m which will allow use of copper Cat 6 connection cables as well as multi-mode fibres.

A.5.3   Tunnel network

The following concepts have driven the network design needed to support sparsely distributed installations in the tunnel:

- Electronics, and other, racks are grouped geographically so that the network connections required can be provided by a single local switch and patch panel installation.

- Connections within rack groups can use copper or fibre connections.

- Any point in the tunnel should be ≤90m, the maximum copper Cat 6 patch cable length, from a switch. This allows copper connections to be installed quickly as needed without requiring the installation of additional hardware.

- Switches are connected to the end-points using SL mode fibre running along the planned tunnel cable trays.

- Switches and patch panels are mounted directly in the instrument racks (A1, X1, etc.).

- By 2014 10Gbit/s copper connections should be standard and switch fibre uplinks will be 40Gbit/s or higher.

- Fibres running along the tunnel cable trays will be required to pass through a chicane between the tunnel and the end-point rack area.

Using the tunnel rack allocation shown in Table 11 a provisional scheme for switch locations in the tunnels has been derived, see Table 14.

SASE1		SASE2		SASE3		U1		U2	
Group	Ports	Group	Ports	Group	Ports	Group	Ports	Group	Ports
X1/946	0/4 X1	X1/856	0/4 X1	X1/406	0/4 X1	X1/584	0/4 X1	X1/224	0/4 X2
X2 856	0/4 X2	X2/766	0/4 A1	X2/316	0/4 X2	X2/494	0/4 X2	X2/224	0/4 X2
X3/766	0/4 A1	A1/763	0/4 A1	A1/233	0/4 B1	A1/482	0/4 X2	A1/145	0/4 X2
A1/726	0/4 A1	B1/734	4/16 A1	X3/226	0/4 B1	B1/453	4/16 X2	B1/140	4/16 X2
B1/697	4/16 A1	X3/676	0/4 D1	B1/183	4/16 B1	C1/431	0/4 E1	X3/114	0/4 E1
X4/676	0/4 X5	C1/679	0/4 D1	C1/141	0/4 B1	X3/404	0/4 E1	C1/104	4/16 E1
X5/596	0/4 X5	D1/608	4/16 D1	X4/136	0/4 B1	D1/400	4/16 E1	D1/88	4/16 E1
C1/571	4/16 X5	X4/586	0/4 D1	E1/130	0/4 B1	E1/359	0/4 E1	E1/63	0/4 E1
D1/511	0/4 E1	F1/512	0/4 G1	F1/75	0/4 X5	H1/328	0/4 E1	X4/44	0/4 E1
X6/496	0/4 E1	X5/496	0/4 G1	X5/46	0/4 X5	X4/314	0/4 E1	F1/3	0/4 E1
E1/449	4/16 E1	X6/406	0/4 G1	G1/40	0/4 X5	X5/224	0/4 X5		
F1/409	0/4 E1	G1/450	4/16 G1	H1/3	0/4 X5	F1/178	0/4 X5		
X7/406	0/4 E1	H1/408	0/4 G1			X6/134	0/4 X7		
X8/316	0/4 G1	X7/316	0/4 I1			X7/44	0/4 X7		
G1/258	0/4 G1	I1/258	0/4 I1			G1/2	0/4 X7		
X9/226	0/4 G1	X8/226	0/4 I1						
X10/136	0/4 X11	X9/136	0/4 X9						
X11/46	0/4 X11	X10/46	0/4 J1						
H1/6	0/4 X11	J1/3	0/4 J1						

Table 14 Tunnel network connections. The group field contains the name of the rack group and the distance (m) from XHEXP1. The port field contains the number of 1/10Gbit/s ports and the rack group containing the switch.

A.5.4   Computer service networks

All PCL and DC storage and processing hardware will be installed in the balcony computer service rooms. As these rooms also contain the network endpoints connections from tunnel and hutch or laser rooms interconnections are easily made. This configuration removes the requirement of installing significant amounts of DAQ hardware in the hutches. It also centralizes infrastructure requirements like cooling, power distribution, etc.

A.5.5   XHEXP1 to DESY-IT network link

Ideally the DM data archive and processing hardware, or as it will be called later the data centre, should be located on the DESY site. It is planned to install 50GB/s of link capacity between XHEXP1 endpoints and the DESY computer centre. This will require ≤20 fibres provided wavelength division multiplexing can be used in 2014. Fibres will be connected to endpoints in computer service rooms 30 and 38, traverse the SASE1 tunnels on the control signal cable trays, enter the under floor fibre channel at XS1, and reach DESY-IT via XSIN.

A.5.6   Wireless networks

The wireless network discussed here carries IP protocols over standard mobile wireless technology like GSM. This should not be confused with the digital wireless used by fire and police services which EuXFEL is required to install in tunnels and XHEXP1 in order to obtain safety certification.

Wireless access should be provided in the hall and tunnels for portable access as it will simplify maintenance and other activity when cutoff from the cabled network. It should be noted that wireless will not be used for DAQ and control.

Test tunnel wireless installations in FLASH and PETRA3 at DESY, and no doubt at other laboratories, will provide information concerning which solutions should be used for signal distribution and protocol standards to be used. It is too early to specify an implementation for EuXFEL tunnels at the moment. Given the twelve month specification to installed period a final decision must not be made before mid-2012. The cost of the installation is likely to be high.

A.6   Risks

The following risks can be identified:

- Details of the infrastructure required were derived during the process of writing this TDR and details may change.

- The Global Undulator control systems interface to the photon control system has not yet been defined.

- The beam line vacuum system rack and network requirements have been derived from discussions with HASYLAB-FS. No detailed planning work has been performed.

- The cost of wireless networks has not been included in the costing supplementary note.

Appendix B EuXFEL DM and DAQ cost, time and manpower estimates

This version of the DAQ and DM computing TDR is for public access. Appendix B, which contains cost, manpower and time line estimates, has been removed.

12 References

1 Timing requirements and a proposal of a timing concept for the Eu-XFEL (WP28 document)

2 2D pixel detector Clock and Control in-kind proposal (UCL)

3 2D pixel detector Train Builder in-kind proposal (STFC)

4 DOOCS, http://doocs.desy.de

5 GlassFish Application server, https://glassfish.dev.java.net

6 NetBeans IDE, http://www.netbeans.org

7 SubVersion (SVN), http://subversion.tigris.org

8 JMS, http://java.sun.com/products/jms

9 MPI http://www.mcs.anl.gov/research/projects/mpi/index.htm

10 LTO technology Homepage, http://www.lto-technology.com

11 dCache Homepage, http://www.dcache.org

12 http://www.hdfgroup.org/HDF5/index.html

13 http://www.dcache.org/manuals/chep04/Chimera-paper.pdf

14 NFS version 4 protocol http://nfsv4.org/

15 AMGA metadata catalogue http://amga.web.cern.ch/amga/

16 The Kerberos Network Authentication Service (version 5), RFC4120,
   http://tools.ietf.org/html/rfc4120

17 GSI-enabled openssh, http://grid.ncsa.uiuc.edu/ssh

18 HGF-Terascale project homepage, http://terascale.desy.de

19 IBTA homepage, http://www.infinibandta.org

20 OpenAFS, http://www.openafs.org

21 cmake homepage, http://www.cmake.org

22 Apache Ant, http://ant.apache.org

23 Infrastructure requirements (rack, cooling, power, network, etc.) for photon beam line
   systems are described in note
   http://edmsdirect.desy.de/edmsdirect/file.jsp?edmsid=1332691 and excel spreadsheet
   http://edmsdirect.desy.de/edmsdirect/file.jsp?edmsid=1332741

24 Infrastructure requirements (rack, cooling, power, network, etc.) for computing services in
   the experimental hall are described in note
   http://edmsdirect.desy.de/edmsdirect/file.jsp?edmsid=1451201

25 Proceeding of the XFEL rack workshop held in DESY March 2009,
   https://indico.desy.de/conferenceDisplay.py?confId=1877