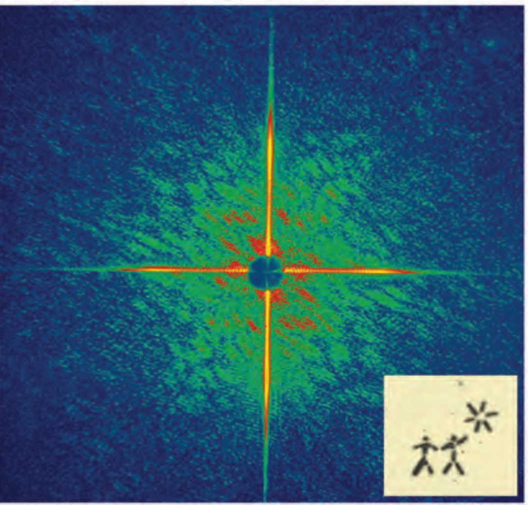
HELMHOLTZ
ASSOCIATION

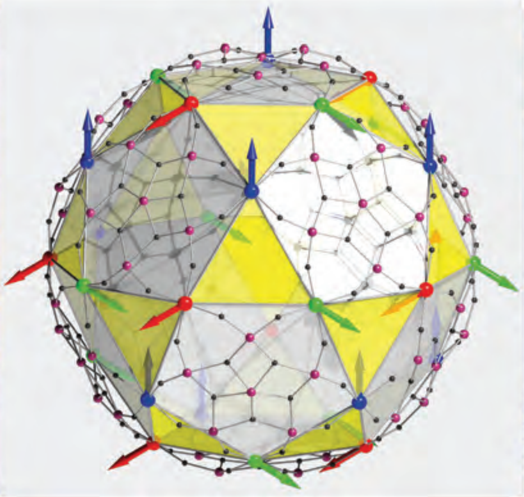
Research Field Structure of Matter

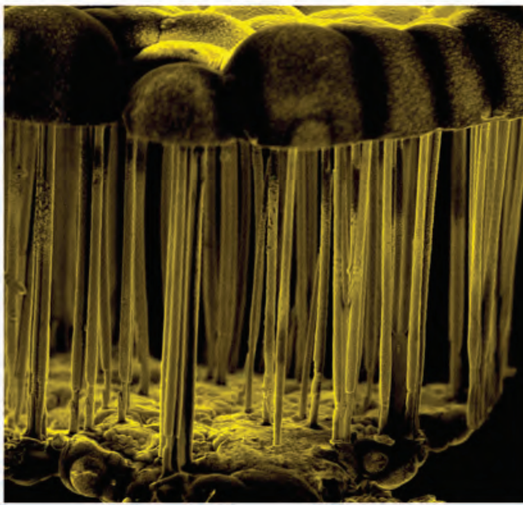
Proposal for the

High Data Rate Processing and Analysis Initiative (HDRI)

within the Helmholtz Research Programme
Research with Photons, Neutrons and Ions (PNI)







of the Helmholtz Centres
Deutsches Elektronen-Synchrotron DESY
Forschungszentrum Jülich
Forschungszentrum Karlsruhe
GKSS-Forschungszentrum Geesthacht
GSI Helmholtzzentrum für Schwerionenforschung
Helmholtz-Zentrum Berlin für Materialien und Energie

2010 – 2014 I Coordinating Centres: GKSS-Forschungszentrum Geesthacht (PNI)
Deutsches Elektronen-Synchrotron DESY (HDRI)

Contents

Executive Summary

Scientific experiments at the PNI facilities dramatically benefit from the opportunities provided by modern, mostly two-dimensional detectors. However, the high data rates generated by these devices pose considerable challenges for the facilities and their users. To address these a High Data Rate Processing and Analysis Initiative (HDRI) is proposed by the PNI facilities for a coordinated and common approach towards developing and providing the appropriate hardware and software tools to cope with high data rates. Within this initiative an online data evaluation system will be established to enable feedback information for fast quality assurance during the course of an experiment, and hence for the most efficient use of valuable experimental time. Further, a standard data format will be established for an easy exchange of data and compatibility with evaluation software, and a data lifetime policy will be established for remote user access to large amounts of data as well as long-term archiving. Another challenging task proposed here is the development of simulation, modelling, and visualization tools enabling experimentalists to obtain scientific results from their samples faster, easier, and more reliably. It is envisaged that this initiative will be pursued in close collaboration with the PNI user community as well as other large-scale facilities for synchrotron radiation, neutron, and ion research outside PNI.

1 Introduction

1.1 Motivation

Experiments at the PNI facilities, especially at third generation synchrotron sources like PETRA III, new high-flux neutron sources like the SNS and ESS, and free electron lasers, dramatically benefit from continuous improvements in source performance like brilliance or flux density, high throughput optics setups, and lately also from developments in detector technology. Especially, with the advent of modern 2D semiconductor detectors like CCDs, pnCCDs or pixel detectors capable of very high frame rates, totally new realms are opened up for high throughput experiments in various scientific fields. Better sampling in both space and time enable experiments of better spatial or q-space quality as well as temporal resolution. Common to all these experiments is the very high data rates that experimentalists have to cope with. Already now experiments can generate data rates up to the order of 10 MB/s for neutrons to several 100 MB/s for X-rays. Recently, the first FEL experiments have produced in excess of 20 TB within one week [1].

Even in such times of continuous growth of computing and data storage capabilities, these volumes of data require dedicated treatment, which extend to the duration of their whole lifetime. For a most efficient use of the experimental time at PNI facilities a fast first evaluation of the measured data is mandatory to guide the experiments. Depending on the experimental techniques this might be as simple as the determination of suitable quality indictors like the internal R-value in single crystal crystallography, the visualization of the data, or in slightly more complicated cases the inspection of some reconstructed information like in tomography. In complex experiments comparison with simulations or model calculations are often necessary. This first data reduction and evaluation step has to be fast enough to allow for an almost real time decision on whether an experiment was successful or not. It is this first step, which is one of the main aims of this collaborative initiative. However, additional efforts are necessary to make the measured data available to the users in an appropriate and safe way until the analysis is finished, and to archive and make data that resulted in publications publicly available.

Since many of these tasks are similar at each of the PNI facilities, this common initiative will try to exploit the synergy of common developments where possible within PNI but also within the European and international synchrotron radiation, neutron, and ion beam communities. This includes facilities as well as user groups from universities and non-university institutions. To enable collaborations, the participating institutes have to agree on certain standards for storage of raw and meta data. The files should contain a full and self-consistent description of the data to allow for later evaluation. Standardized file formats are a precondition for exchanging data and analysis tools among institutes. Hardware standardization can help to keep the programming effort of specialized software at a sustainable level. Data modelling and analysis challenges presented by the large and complex datasets envisaged in this HDRI are formidable. In order to make the task of evaluating data at all manageable, and further to extend their use to the user community as a whole, it is vital that tools are developed which allow ready access to the visualizing of datasets, computationally intensive modelling of the physical behaviour of systems, and full simulations involving instrumental resolution.

1.2 Definitions

Before going into more detail, we shall define the necessary vocabulary to aid the following discussion, which constitutes work package 1 (WP 1) of this initiative. In Figure 1 a simplified scheme for the entire data flow scheme is given.

By the term ‘data acquisition system’ (DAQ) we refer to the part of the data chain which registers a physical quantity (e.g. a count), and provides this information in digital form to an interface within a frontend computer (very left part of Fig.1). HDRI will not generally deal with issues concerning DAQ systems since these are highly specific to individual experimental setups. Instead, HDRI will start with the assumption that measured physical quantities are already in digital form on an interface connected to the frontend computer or in its RAM. In general, data should then be written in a standardized form onto a mass storage system where they can be accessed for immediate evaluation on site (left and middle part on Fig. 1). The data will also stay available, for a given time, for the users for remote evaluation i.e. users will be able to access their data on the mass storage systems of the facilities. Suitable backup strategies will ensure data safety against hardware failure (middle part on Fig. 1). Data leading to publications will be archived for long-term preservation and could be made available in the frame of an open data policy (right part on Fig. 1). The term 'real time' data evaluation in the context of this initiative denotes the evaluation or visualization of the data during the course of the experiment for immediate feedback on the experiment itself.

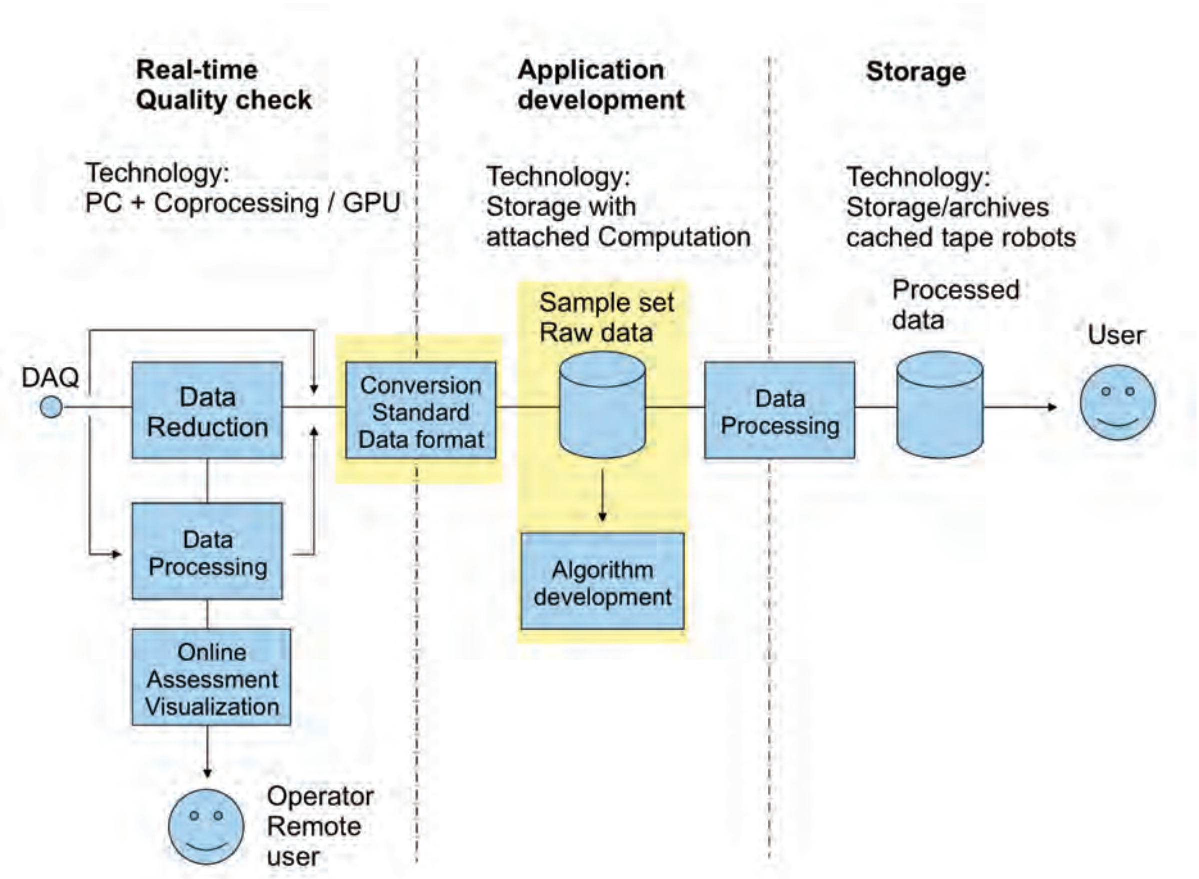

Real-time Quality check
Technology: PC + Coprocessing / GPU
Application development
Technology: Storage with attached Computation
Storage
Technology: Storage/archives cached tape robots
DAQ
Data Reduction
Data Processing
Online Assessment Visualization
Operator Remote user
Conversion Standard Data format
Sample set Raw data
Algorithm development
Data Processing
Processed data
User

Figure 1: Simplified scheme of data flow at PNI facilities.

1.3 Scope of the Proposed Activity

The limited resources available for this initiative will be insufficient to solve all possible data handling, evaluation, and analysis problems at PNI facilities during the present PoF II period (2010 - 2014). The aim of this initiative is rather to start with those experimental techniques where data rates are highest, and focus on where a coordinated effort will be to the benefit of the largest user communities. These include techniques like crystallography, micro tomography, small angle scattering, imaging, time correlation spectroscopy, and fluorescence mapping. Based on the standardizations and software building blocks established during this first phase, further techniques will be included into this initiative. Where possible the standards to be applied in this initiative will be established in close collaboration with European and international facilities, as well as the general communities working in the same field. The proponents of this initiative agree that this initiative should not be a project of limited lifetime but rather a continuous effort beyond the present PoF II period.

2 Data Management

Coordinators: T. Kracht, R. Gehrke (DESY), A. Föhlisch (HZB)
Contributors: D. Herrndörfer, J. Hoffmann (HZB),
W. Drube, G. Wellenreuther (DESY)

This chapter deals with all questions concerning data formats, data handling, remote access, and data archiving. These are the main prerequisites for any common data handling and evaluation strategy.

2.1 Standardization and Data Formats

A common standard for data formats is a prerequisite for efficient data handling and evaluation. Here, the aim is to both facilitate access to data that is independent of specifics of instrument or facility for those researchers having “ownership”, and to facilitate evaluation and analysis by standardising this access. We require the following from the data format, to:

1. provide a complete set of parameters that describe the experimental setup and all the measured data; the format should in turn be self-descriptive,
2. be structured, flexible, extensible, and platform independent,
3. be highly efficient in terms of access speed,
4. implement suitable compression mechanisms, and
5. be readily editable with journaling of changes.

The creation of a standard data format involves the specification of a data model, the definition of the contents, and their implementation.

The data model is an abstract description of the structure of the data, and specifies its organisation. It is expected that the model will feature a hierarchical organization of the data, i.e. the data will be represented in a tree structure with named nodes, named datasets, and attributes of the datasets. The data model must be capable of accommodating all setups and experimental strategies. Flexibility, extensibility, self documentation, and access speed are key design considerations.

The content definition specifies the labels for nodes and elements, plus a detailed description of how these are to be used and interpreted. Catalogues of content definitions for particular experimental applications and techniques (scattering, imaging, spectroscopy, crystallography, …) will be made available, and be straightforwardly tailored for specific experiments. Here, the aim is to avoid the use of proprietary formats, and also to provide enough information to fully determine the experimental setup used just from the data file. In our opinion, this will significantly facilitate the creation of new data evaluation software as well as the creation of interfaces between existing software and the data files.

The implementation of the data model should be based on existing and well-established formats, such as the Hierarchical Data Format HDF5 [2]. An attempt to create a more specific data format for neutron and SR X-ray data is NeXuS [3], which supports the use of HDF5. Both formats are also endorsed by other international laboratories (e.g. ILL, ISIS, ESRF, ANSTO, SOLEIL, DIAMOND, and PSI) and are possible candidates to serve as a basis for the HDRI data format. If necessary for the implementation, the NeXuS structures can be extended.

For the direct implementation a set of Application Programming Interfaces (APIs) is necessary which perform the basic tasks of building the data structures and retrieving the data. Utility programs are required for basic browsing, visualization, and inspection of the data. For both HDF5 and NeXuS generic utilities exist for these purposes. Development of APIs and utilities can thus be based on respective software already provided in the frame of HDF5 and NeXuS.

Special attention must be paid to the implementation of the data format on the data acquisition side. Since the contents of the data structures may differ from experiment to experiment ways are required to define the specific data structure being used. The acquisition system has then to construct the data structures on the basis of this definition or template. The various data sources involved in the experiment deliver the data via a well-defined interface of functions to the data acquisition software. The acquisition software takes the data provided by these functions and constructs the data structure on the basis of the template. Interfacing a new component to the acquisition system involves providing a corresponding function for the data source. In the best case the complete set of data describing the setup and the experiment is generated by the acquisition software fully automatically. This prevents the situation arising that important information is missing from the data. However, it will certainly occur that some meta data may need editing after the acquisition has finished (e.g. comments on the sample, but also some fixed parameters of the setup), therefore, among the utilities, a corresponding editor requires to be provided. Every change made will be tracked and documented in the file (e.g. in an attribute to the corresponding data element) indicating the author and date that the change was made.

The specification of the data model and the collection and specification of the contents in the catalogues of data elements may be completed within half a year after start of the project. In any case, the catalogues can successively be extended, if additional data elements should be identified. For each fundamental experimental area this can be done with the help of users and the experts contributing the evaluation software systems, using a moderated web-based tool (e.g. a Wiki). The final decision on including a new data element into the catalogue is made by a group of experts in the corresponding field. After the definition phase the implementation phase will be started. APIs and utilities have to be constructed and the implementation of the elements for a data acquisition system shall start with case studies at certain instruments.

2.2 Data Access Strategies

Remote data access: The large data sets will initially be stored at the PNI centres and those external users who are not able to transfer all the data to their home institutes must have the possibility to access their data remotely. A web-based portal shall act as a common interface for this access. A survey of established open source software (e.g. Fedora Commons) and of solutions planned or implemented at other institutions (e.g. DIAMOND, EuroFEL) will be performed and a portal prototype will be implemented at DESY. In the envisaged solution the user will have the means to perform keyword-based searching for data, to browse and visualize the contents of large data files, and to transfer specific parts of it without the need to transfer the complete data files.

In the case where users decide to leave their data with the facilities, data will be managed by a suitable caching and tape robot system for a defined period whose length needs to be discussed directly with the user community. A possible option for achieving this is the dCache [4] system developed by DESY and FermiLab.

A long-term goal, which is not part of the first phase of HDRI, is to provide access to experimental data within the frame of a Data-GRID to the users.

Remote Computing: For external users who do not have sufficient infrastructure at their home institutes to handle such huge amounts of data, a platform will be provided for remote data treatment. For this purpose, computing resources and a PNI-wide repository of programs and routines will be established. This will cover the basic needs of data treatment but will also include software needed in specific scientific fields (e.g. tomographic reconstruction). Work on this shall start within the frame of the proposed activity but will have to be carried on beyond that.

Authentication and authorisation: Remote access to data and computing resources requires schemes for authentication and access authorization. It is envisaged to establish a common mechanism for use by all centres participating in HDRI. Based on a detailed evaluation, one of the available solutions for authentication (e.g. open-id, Shibboleth) will be chosed and implemented at the partner sites. The idea behind this effort is for users to have to provide their basic contact information only once within PNI.

2.3 Data Lifetime Management and Archiving

Common policies for data lifetime management and archiving need to be established within the project. These will address the following questions: What kind of data is going to be archived? How long will it be kept? Will published data be made publicly available? etc. According to these policies, hierarchical storage management (HSM) and corresponding archiving facilities will be implemented, and be made available to all project partners. A possible candidate for this implementation is again dCache. In this case DESY already possesses significant expertise and can assist partners in providing a commonly available dCache infrastructure.

3 Real-time Data Processing

Coordinators: M. Münch (GSI), M. Weber (KIT)
Contributors: H. Bräuning, C. Kozhuharov, M. Münch (GSI), C. Jung (HZB),
A. Kopmann, M. Weber (KIT), F. Beckmann (GKSS)

It is common among many PNI experiments that they deliver series of consecutive single measurements (frames) with high frame rates and large amounts of detector and auxiliary data in each frame. Usually, evaluation starts on a frame-by-frame basis, then data obtained from all collected frames is evaluated further, and finally the results have to be visualized and archived. An online data evaluation system that fulfils the requirements of real-time data analysis and visualization must be fast and must provide sufficient flexibility to cover the very different analysis demands of a great variety of experiments.

The development of the different applications of such a system should be driven by the needs of the most data demanding experiments like Small Angle Scattering (time resolved in-situ SAXS, SANS and raster scanning micro/nano-SAXS), neutron time-of-flight experiments, Micro-Tomography, Coherent Diffraction Imaging, X-ray Photon Correlation Spectroscopy, X-ray Ion Coincidence Spectroscopy, micro-fluorescence tomography, and macromolecular crystallography.

The architecture of the data processing system for real-time data assessment with multi-dimensional pixel detectors is outlined in Figure 1. The processing is split in two domains defined by the data rate that can be handled by today’s PC technology. We propose the development of a high-performance ‘computing system’ based on single PC computer architecture. The internal system bus (PCIe) offers an excellent bandwidth and available high-end graphic processors units (GPUs) can provide the required amount of parallel computing power. The computing cell is fed by digital data from pixel detectors (including analog-to-digital conversion). A significant class of experiments would hugely benefit from processing steps in hardware, in order to shrink data volume before starting the evaluation in the PC. On the other end of the data processing chain (see Figure 1) the data will be stored for later analysis. Real-time data assessment will help to reduce the amount of stored data and at the same time to evaluate their quality.

This chapter focuses on three key aspects for real-time quality checks:

1. Essential data processing with dedicated hardware (main contributor GSI)
2. Real-time data assessment with parallel computing on PCs (main contributor KIT)
3. Analysis methods / Applications (main contributors: DESY, GKSS, HZB)

3.1 Data Processing with Dedicated Hardware

Resolution in one or two spatial dimensions, measurement of energy with high precision, and in particular the introduction of time resolution with high sampling rates drive data rates well beyond the capacity of "commercial off-the-shelf" PC technology.

High-resolution transmission spectrometers for hard X-rays such as the "FOCAL" spectrometer are a major part of the PNI research activities at GSI addressing high-field QED. It strongly relies on 2D position sensitive and energy dispersive Ge(i) micro-strip detectors. To realize the full resolution of the system, pulse shape sampling with a sampling rate of 100 MHz is done, resulting in huge data rates. Diamond detectors used for ion detection as described in 3.3.4 will overcome the rate limitation of 2-dimensional Micro-Channel-Plate detectors and will also produce peak data rates that are not manageable by software.

Pre-processing the raw data of these detectors can be done in an efficient way using programmable hardware. The basic components of the pre-processing hardware are always very similar: an input interface with a fast link to the detector electronics, a processing stage, and the standard interface to commercial hardware. The processing stage may consist of Field Programmable Gate Arrays (FPGA), Digital Signal Processors (DSP) or, more recently, Graphics Processing Units as embedded chips directly on board. Algorithms vary from simple hit detection up to complex waveform analysis or pattern processing. Deployment of programmable hardware allows the changing of algorithms by reprogramming and using the same boards in different areas of PNI.

GSI has developed several generations of FPGA (VULOM [5], FEBEX [6], EXPLODER [7]) and DSP (VUPROM [8]) based pre-processing devices, together with the link and interface electronics for use in nuclear and high energy physics (more information can be found under [9] and [10]).

We propose to first make a survey of available solutions of programmable hardware for signal pre-processing, using the expertise of HGF in nuclear and high-energy physics. Based on the outcome, we would then adopt an existing solution for PNI needs, if available, or join a current development and introduce the requests of PNI.

3.2 Real-Time Data Assessment with Parallel Computing

For online data inspection, evaluation, and control it is desirable to process the raw data as fast as possible. This results in challenging demands for bandwidth and computation power. We propose to address this issue by establishing an appropriate parallel computing infrastructure: a) CPU clusters; b) GPUs.

We propose to develop an extensible set of library functions to meet the requirements of the different application fields. The goal is to develop a stack of software tools, which allow rapid development and deployment of a large range of parallel data processing algorithms. A candidate for a universal implementation language might be OpenCL, a standard for programming parallel architectures [11].

We propose a three layer structure: the first layer implements the core part of the stack which is a set of APIs which would automate the usage of the defined data streams and computing platforms. This shall include: the access to standard data formats, transportation to the processing units, decompression and display of results. The next layer is a collection of parallel primitives for image processing and numerical computation using parallel architectures. The top level of the framework should structure the development process. A task scheduler should distribute the computation tasks between available GPU and CPU cores. The developer is expected to provide only the processing, visualization, and quality check plugins, while everything else is automated by the framework.

In parallel this task will propose suitable parallel computing platforms. The hardware needs to be highly modular to meet the very different resource requirements for the application fields. If the data transfer is the limiting factor a solution within a single standard PC will be preferable. An alternative to CPU clusters is modern graphic processors (GPU), which are being used more and more in scientific computing. Each GPU provides up to 240 cores with in sum 1 TFlops. There have been solutions published with up to 13 GPUs and 12 TFlops in a single PC. For comparison: conventional CPUs like AMDs Opterons reach up to about 4 GFlops/core. Thus a conventional setup with 16 cores and ~60 GFlops total can be accelerated by a factor of 100 or more. The challenge will be to apply the outlined framework for a parallel computing architecture to the task of real-time data assessment.

3.3 Analysis Methods and Applications

In this sub-chapter we give four representative scientific applications illustrating the goals and benefits of HDRI to users. Further examples are given in Appendix A1.

Application	**Current data rates** [GB/h]		**Future data rates** [GB/h]	
	peak	average	Peak	Average
Protein crystallography	500	50	500	200
Coherent diffraction imaging	500	50	4000	400
Tomography	700	50	800	200
Spectroscopy	450	45	18,000	1800
Small angle scattering	1400	140	14,400	4200
Grain mapping	140	80	800	300
In-situ dilatometry	14	12	530	200
In-situ imaging	N/A	N/A	600	110
FOCAL spectrometer	150	150	3000	3000

Table 1: Typical peak and average data rates for different PNI applications. Both current rates and the expect rates in the next five years are given. The average data rates consider the effects of beam time, set-up and sample exchange.

The expected data rates are summarized in Table 1. The typical detector in many of the examples below is a 2k x 2k or 4k x 4k pixel camera with 16 bit resolution. Current frame rates vary from a few Hz to tens of Hz, and for SAXS rates exceed 100 Hz. For most applications the expected future data rates will increase substantially (partially due to the simultaneous use of several 2D detectors with the aforementioned data rates).

Crucial issues are the real-time data reduction, analysis, and quality evaluation so as to be able to decide on the further measurement strategy during the experiment.

3.3.1 (Protein-) Crystallography

Since many years 2D CCD detectors have been used in crystallographic experiments. Finer angular slicing of the data often results in higher signal to noise levels, especially in the case of pixel array detectors. Thus, data sets may contain more than 1800 frames with 32 MB each. For the time of the exposure, data rates in excess of 150 MB/s have to be dealt with. 'Luckily' sample mounting and alignment, even if carried out by a robot, takes place on the scale of several minutes. Regarding expedient evaluation of the measured data, (protein-) crystallography is already in decent shape since several programs are available for integrating and calculating first data quality indicators within minutes, provided that sufficient computing power is available. In this case the most difficult task, namely the implementation of the evaluation software for parallel computing (e.g. SMP), has already been carried out by experts from the community. However, data file formats are mostly proprietary with limited and mostly no automatic means for saving the experiment metadata that need to be recorded, e.g. for a synchrotron radiation experiment. Also, the handling of the huge amount of data is a big issue for this community. Therefore, the structures to be realized by HDRI will certainly streamline data handling processes. Crystallography experiments, therefore, are an ideal test case for the first implementation of the envisaged data flow structures.

3.3.2 Tomography

In recent years micro-tomography using synchrotron radiation (SR) or neutrons has became a well accepted tool for the non-invasive three-dimensional characterization of specimens in the fields of materials science, medicine, biology, and environmental research. Phase contrast and absorption contrast techniques have been developed to investigate millimetre and centimetre sized samples. Due to the steadily increasing requests for the routine application of these methods to series of samples, maximal use of the limited beamtime is required.

We can estimate the data rates using the example of hig*h throughput SRµCT*. Here, the minimum data rate is 2k x 2k, 16 bit, 1 Hz, leading to 64 MB/s, with upgrade to 4k x 4k, 10 Hz possible, leading to >2 GB/s. One must note, however, the following operational demands:

- Continuous operation for 24 h: This includes automatic sample changing, which itself requires sample pre-investigation
- The total amount of data can be estimated as follows:

    - 12 GB / scan (raw data)
    - 32 GB / scan (tomograms)
    - High resolution: 15 min / scan
    - Medium resolution: 1 min / scan
    - Low resolution: 1 second /scan
    - This leads in one day to 1 TB per day raw data in high resolution!
- One must take care of the bottlenecks of online reconstruction during the scan and of high-quality offline reconstruction.

This will become the basis for optimizing the use of the available SR beam time and will increase the throughput of the systems maintaining the dynamic range and contrast in the tomograms. The aim of HDRI is to reduce the time needed for the first possible evaluation of the reconstructed sample from presently more than one hour to about one minute.

3.3.3 Event Recording in Neutron Detection

While neutron sources provide significantly less individual counts than photon sources, neutron experiments will still benefit dramatically from this initiative. With the move to event recording every neutron count is kept - for each event 16 bytes (this is advantageous for stroboscopic experiments and those where sample conditions are continuously changed). Because the count rates scale with the beam intensities, of around 10^6 ns^{-1} for a monochromatic (approx. 50 GB/day) or much higher again for a white beam experiment, it is foreseen that fast data rate methods can be applied to the newest instrumentation such as NEAT2 (HZB) or on ESS and SNS.

The proposed "DiffractomEter for Nonequilibrium States of condensed matter" DENS at the SNS will produce peak count rates of some 40 Tbytes/h (see appendix A1.8).

3.3.4 Ion Detection Using Diamond Detectors

Detector grade carbon vapour deposit (CVD) diamond detectors can be produced either as polycrystalline wafers (up to 5" in diameter) or as single crystals with ≈ 1cm² area and > 1mm thickness. The free market prices for both types of detectors have now become affordable. While both types of detectors have unsurpassed timing properties as well as unsurpassed radiation hardness, a single crystal diamond detector in addition offers a superb energy resolution. The detectors can be stacked, and the contacts on both sides of these detectors can be evaporated in such a way that the detector setups are position-sensitive with a sub-mm spatial resolution. All these properties are an exceptional advantage for ion detection: diamond detectors can be used for a wide range of ions (from protons to uranium), wide range of count rates (from single ion counting to more than 10 GHz), for a wide range of energies (from 100 keV up to GeV), and for a wide range of heavy-ion fluencies (more than 10^{13}/cm²). The superb timing properties (20 ps rise time and slightly longer fall time of the electronic signals) are well suited for tracking and/or coincidence measurements. These extremely fast signals require, on the other hand, very fast sampling with more than 10 GHz and with a minimum of 8 bits (up to 14 for better resolution). Whereas diamond detectors can be operated continuously at very high count rates, the bottleneck remains the electronic read-out, since the total amount of raw data can easily reach 100 TB per day, and this is a formidable challenge for the data acquisition. (Note that in addition, the raw data have to be monitored and selectively filtered during the beam time.)

4 Data Analysis, Modelling, and Simulation

Coordinator: T. Brückel (Jülich)

Contributors: M. Monkenbusch, S. Mattauch, A. Ioffe (Jülich), S. Roth (DESY), D. Lott (GKSS), A. Tennant (HZB)

New techniques and detectors yield more and more complex data that need special expertise for treatment and interpretation. In scanning applications rapid scanning over extended and laterally inhomogeneous multi-dimensional gradient samples leads to enormous amounts of two- and higher dimensional data. Parametric studies as a function of thermodynamic parameters (temperature, field, pressure etc), or kinematic studies as a function of time, produce enormous sets of data, which can only be treated efficiently by automated routines. On the other hand developments on the computing side allow for more and more sophisticated evaluation and modelling procedures requiring expert knowledge and eventually access to high performance computing facilities.

The PNI centres see it as a central task, not only to provide state-of-the-art instruments at high intensity sources, but also to supply corresponding software and computing tools. A unified approach to this problem shall support both the expert and non-specialists users. Diffraction in crystallography has already reached a state that may serve here as a successful model of such an enterprise that could be adopted for other scattering methods. For crystallographic data treatment, standardized programme packages exist, covering the whole chain of data treatment from data correction via simulation and analysis to real space visualization (e.g. atomic structure incl. thermal ellipsoids and bond lengths) which are then used by diverse communities such as biology, chemistry, geo-science, materials science, engineering, physics etc.

By supplying appropriate treatment and analysis procedures, we aim to open up state-of-the-art PNI instrumentation and experimental techniques such as to make them accessible to the non-expert users, allowing them to concentrate on the scientific problem in question and not on the technical details of data evaluation.

As a first example for prototyping we choose Grazing Incidence Small Angle Scattering (GISAS), to develop and access the architecture and software tools that in a later stage will enable us to extend the supplied analysis and visualization tool box to other experiments.

Indeed GISAS, we believe, could have a substantial impact in nanoscience for a broad community of users from biology (membranes), chemistry (catalysts), engineering (surface treatment), physics (thin film magnetism) etc. if a similar ease of use can be established. This requires also real space visualization of representative surface-, interface- or domain structures similar to the structure plots in crystallography.

4.1 Proposed Conceptual Design

We intend to standardize and integrate the key data analysis activities of visualization, modelling, and simulation in order to form a toolbox whereby specialized codes and high performance computing can be accessed. The aim is to build an architecture whereby a simplified data analysis environment can be offered. This will utilise the advantages presented by the data management tools and common formats developed in the other work packages. This environment will further link in with the provision of high capability data processing and computing resources of the Helmholtz centres.

The overarching aim is to establish an open source environment where codes can be contributed by user groups and facilities scientists which will then extend the scope of application with time. The basic work to create this environment and provide the key codes necessary will be undertaken within the HDRI. The plan is to pursue three directions simultaneously: (i) the development of the architecture and core programmes, (ii) linking tools to the data sets and computing resources, and (iii) development of the example application - GISAS. It is hoped that in doing so we will deliver a common data analysis platform for scientists, fostering a collaborative approach and bringing together the most powerful computational and visualization resources to aid the user community.

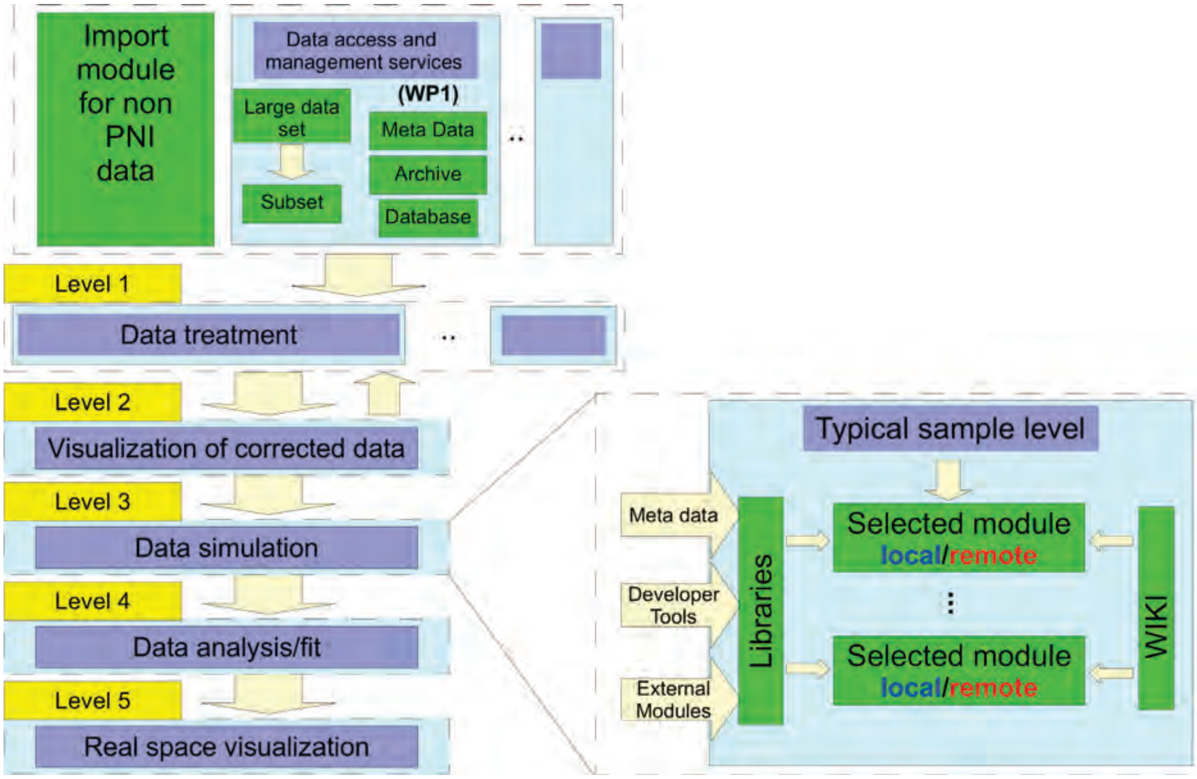


Figure 2: Schematic interdependency of the key components and the work flow of the planned data correction, analysis and visualization framework.

4.2 Key Components

We propose a data analysis management platform that allows free and full interaction with the data handling, modelling, simulation, and visualization resources of the PNI centres supported by core routines on the users own computer. This will be achieved through well-defined API which will allow legacy and external high performance resources to be linked in an effective and timely manner.

Levels Involved in Data Analysis Work Flow

Level 1: Data treatment

The first program block comprises fast, efficient, common, generic libraries, implemented at server or stand-alone level providing the essential initial data treatment steps, e.g. transformation from angles, detector pixel positions, etc. into physical space like (Q_x,Q_y,Q_z,E) including normalization, background and sensitivity corrections.

Level 2: Data visualization

This allows for the interactive presentation of n-dimensional data as projections, cuts or iso-surfaces. For the challenges presented here, recognition algorithms for correlations in the data and automated search strategies for clustering and classifying are of particular significance. This enables the identification of relevant sections of parameter space. This involves constructing common libraries, but the final aim is for user tools tailored for the physical problems.

Level 3: Simulations

This offers access to expert software and CPU-intensive applications. The aim is to perform on-demand and quasi-real time simulations to both obtain first information within constraints on physical parameters as well as input for further more sophisticated analysis. Envisaged capabilities include access to Monte Carlo instrument simulations, DFT and analytical expressions, but also the combination of these to simulate the total scattering response.

Level 4: Modelling/Analysis

Here, the connection with the physical parameters is made. Comparison is made between model and data by the use of statistical analysis, e.g. regression or Bayesian analysis. Tailored plug-in modules from the user community are extremely important. The challenge is to address comprehensive data sets, so simultaneous fitting of multi-technique/-multi-probe data is central. Open source and user-contributed software have successfully been implemented recently, for example in McSTAS/VITESS by the instrument simulation community, and we will follow the best practice here.

Level 5: Real space visualization

Especially for the non-expert user, a visualization of the model in easiest terms is of highest importance, which usually means real space. This step is always an important one in making the connection between structure and functionality. What is required here is to translate the determined system parameters into a readily viewable representation, such as crystal structure or set of vibrational modes. Again, the most convenient and meaningful representation goes hand in hand with the type of system being studied, and we take the same approach as before in writing core libraries linked with tailored tools.

4.3 Prototyping and Scientific Examples

GISAS is both an important topic for modern nano-science and strong in-house competence for the PNI partners. This means that several existing software packages at the proposing centres can be contributed, expediting the efficient development of exemplary software. The main task is to merge the subunits into one user friendly software tool using an open architecture and allowance for continuous further development, and thereby to meet the needs of the non expert users from chemistry, biology, or engineering materials backgrounds. This will cover neutron and X-ray applications equally well and allow for multi-probe analysis. Users of the PNI instruments REFSANS (GKSS), MARIA (JCNS), KWS1 and KWS2 (JCNS), V6 and V18 (HZB), NANO (ANKA), MiNaXS and BW4 (DESY) will profit immediately from such an initiative.

For GISAS, representative 3D images of surfaces, interfaces, or domain structures will be produced depending on the model chosen: nano-particles on surfaces, fractal interface morphologies, magnetic domains etc. Recently, a JCNS workshop on “modelling and data analysis for grazing incidence and off-specular scattering”, enabled an overview of the main challenges for their application to be clearly identified. Identified to be of particular importance was the development of modelling methods especially the Distorted Wave Born Approximation (DWBA), and also the bringing together of the various international efforts in theory and experiment through extended cooperation between HGF centres and external scientists. Indeed, if the advances in modelling and analysis can be brought into a useable package then this would have the immediate impact of disseminating the most advanced techniques directly to the broader community.