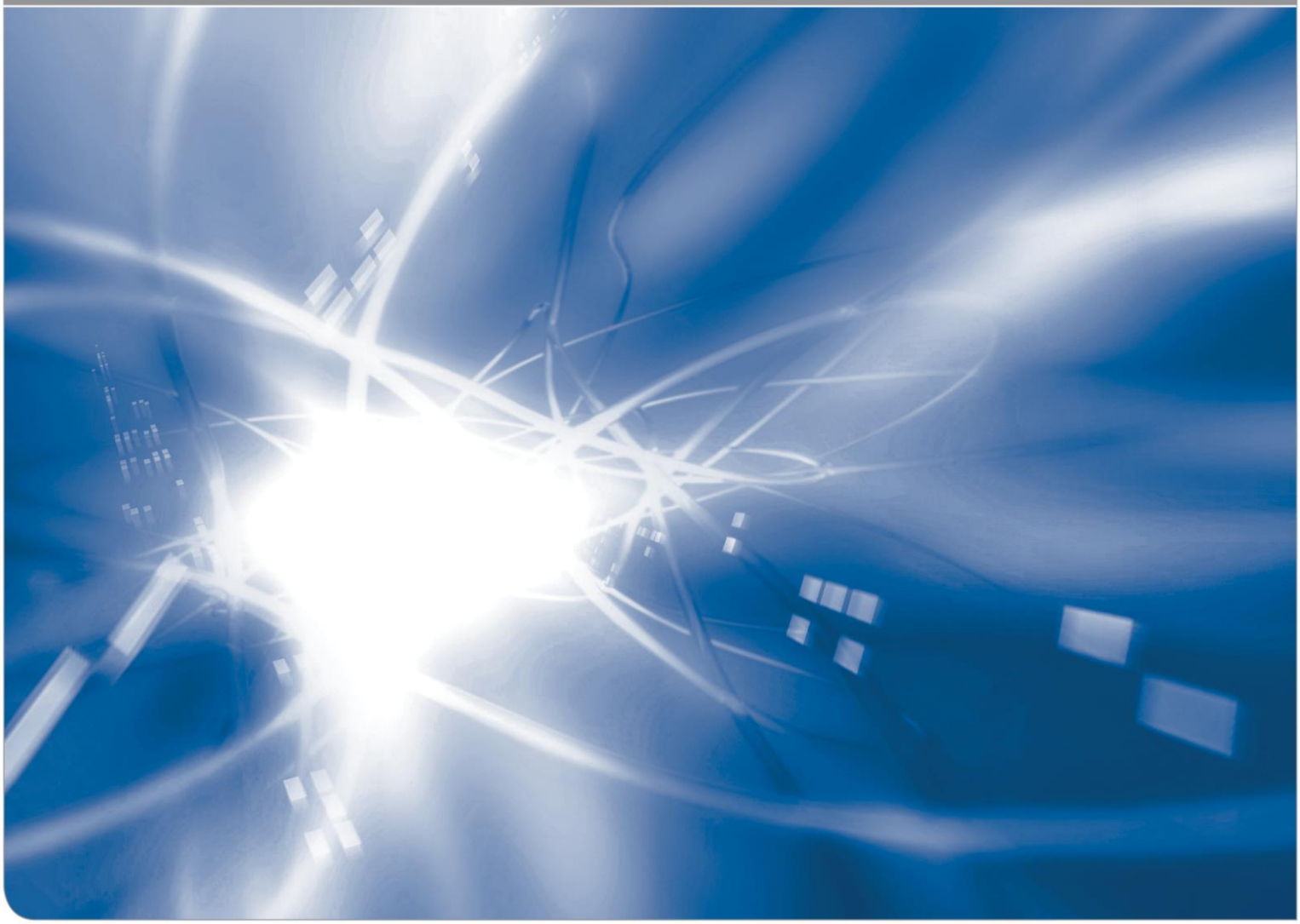
Karlsruhe Institute of Technology

A New Framework for the Analysis of Large Scale Multi-Rate Power Data

by Hüseyin Çakmak[1], Heiko Maass[1], Felix Bach[1], Uwe Kühnapfel[1]

KIT SCIENTIFIC WORKING PAPERS 21



[1] Institut für Angewandte Informatik, Karlsruher Institut für Technologie (KIT)

Institut für Angewandte Informatik
Karlsruher Institut für Technologie
Postfach 3640
76021 Karlsruhe
www.iai.kit.edu

Impressum

Karlsruher Institut für Technologie (KIT)
www.kit.edu

2014

ISSN: 2194-1629

A New Framework for the Analysis of Large Scale Multi-Rate Power Data

ABSTRACT

We face the transformation process of classical power grids towards advanced smart grids, which implicates the merging of information and communication technology with operational technology. Intelligent smart meters and commonly used advanced energy data recorders produce data at a certain update rate, which is not sufficient for the in-depth transient power system analysis and the power grid state estimation. Advanced, high-rate meters provide huge amounts of data but raise issues for the data management. We propose a new framework for multi-rate, large scale data management and analysis, which is able to comprise raw voltage and current data handling at 25 kHz, feature data with a temporal resolution of one second and 10 minutes smart meter data altogether. The introduced high rate power grid data acquisition devices are securely coupled to a big data management system for long term data storage, which forms the data basis for the development of new analysis and simulation methods. The Hadoop based time series analysis is performed locally on the data storage cluster and enables the detection of significant events as voltage sags in large time series. The power grid modeling and simulation tool *e*ASiMoV as part of the framework is capable of handling unified simulation models based on the CIM IEC standard with data interfaces to various simulation software packages. Multiple visualization methods support data exploration for captured, simulated and analyzed energy data. A flexible computing framework enables access to various computing facilities and remote software control for the proposed power grid analysis tools.

Neue Analyseverfahren für umfangreiche Energiedaten mit variabler zeitlicher Auflösung

ZUSAMMENFASSUNG

Im Rahmen der Energiewende vollzieht sich die Transformation der klassischen Stromnetze zu sog. Smart-Grids, bei der neben der Betriebstechnik die Informations- und Kommunikationstechnik eine wesentliche Rolle spielt. Die hierbei eingesetzten Smart-Meter und andere Datenerfassungsgeräte wie etwa PMUs erfassen Verbrauchs- und Betriebsdaten mit einer bestimmten Auflösung, die für eine tiefergehende Stromnetz-Zustandsanalyse meist nicht ausreicht. Aufgrund der kontinuierlichen und großflächigen Energiedatenerfassung fallen zudem riesige Datenmengen an, für die neue Methoden für die Datenspeicherung, das Datenmanagement und die Datenanalyse benötigt werden.

Ein neuartiges System für die Analyse von umfangreichen Zeitreihen mit einer variablen zeitlichen Auflösung von 15 Minuten für Smart-Meter-Daten bis hin zu hochfrequenten Spannungs- und Strom-Messdaten mit einer Datenerfassungsrate von bis zu 25kHz wird vorgestellt. Die neu entwickelten Spannungs- und Stromdatenerfassungsgeräte übertragen die aufgezeichneten hochfrequenten Energiedaten an ein Datenmanagementsystem für große Zeitreihen zur langfristigen Datenarchivierung. Die hochaufgelösten Zeitreihen dienen als Grundlage für die Entwicklung neuartiger, datenintensiver Analyseverfahren zur Identifikation signifikanter Ereignisse und zur Charakterisierung struktureller Datenmerkmale. Die Datenanalyse wird unterstützt durch *e*ASiMoV, einer Software für die Stromnetzmodellierung und -simulation. Geeignete Visualisierungsmethoden dienen zur Datenexploration für erfasste, simulierte und analysierte Energiedaten. Eine neue Softwareumgebung ermöglicht die Nutzung von Hochleistungsrechnern für komplexe Berechnungen und die interaktive Fernsteuerung der *e*ASiMoV-Softwaremodule.

CONTENTS

FIGURES

1. Introduction

The Karlsruhe Institute of Technology (KIT) has a historical leading role in the research of civil nuclear power generation in Germany. In the context of the so-called German energy transition, the Institute established the KIT Energy Center as one of the biggest energy research centers in Europe coordinating multiple energy related projects. Within this scope, the research at the SimLab Energy and at the IAI (Institute for Applied Computer Sciences) focuses on the in-depth power grid analysis.

The integration of renewable energy sources into existing power grids results in fundamental changes in power system functionality and dynamics. Traditional power grids are being replaced by smart grids that combine information and communication technologies (ICT) with operational technology (OT). Smart meters and other power grid data recording devices produce huge amounts of data that need to be handled. As a result, advanced IT infrastructures for big data management, high performance big data analyses and visualization as well as power grid simulation have to be coupled by data exchange and collaborative control offering remote data access and remote software control.

1.1. Related Work

Different research groups have worked on recording, archiving and analyzing energy data combined with power grid modeling and simulation.

Regarding the recording of energy data smart meters seem to be the first choice. These are cheap devices and widely installed in private households. Due to the coarse update rate with a typical frequency of one update every 1 to 15 minutes, they are suited for long-term load forecasting but not for a global fine-grained analysis of power grid dynamics and dependencies. A better alternative are Phasor measurement units (PMUs), these are high speed sensors with the option for synchronous data acquisition and monitoring of the power grid quality by measuring attenuation and phase shift of voltage and current at fundamental frequency. These devices are rarely used in Germany and due to the high costs they are not commonly installed in the distribution grid [1].

An extensive framework development is being pursued by the Tennessee Valley Authority (TVA). They collect and archive PMU data on behalf of the North American Electric Reliability Corporation (NERC) in order to ensure the reliability of the bulk power system in North America [2]. Data from high voltage electric system buses and transmission lines are captured with PMU devices at substation level several thousand times a second, which is then reported for collection and aggregation. Their recorded data contains a GPS time-stamp and the voltage (A, B, C phase in positive, negative, or

zero sequence) magnitude and angle, current (A, B, C phase in positive, negative, or zero sequence) magnitude and angle, frequency, change in frequency over time, digitals and additional status flags. Secure data transfer takes place via LAN-to-LAN VPN tunnels and data is concentrated at several so called “Super Phasor Concentrators” (SPDC). The entirety of the stream, currently involving 19 companies, 10 different manufacturers of PMU devices, and 103 PMUs is then passed to one of three servers running an archiving application, which writes the data to binary files with fixed size on disk. A real-time data stream is forwarded to a server program hosted by TVA, which passes the conditioned data in a standard phasor data protocol (IEEE C37.118-2005) to client visualization tools. A software agent can move archived PMU files into a Hadoop cluster via a FTP interface for long-term storage or alternatively researchers can directly request this data using secure VPN tunnels for further local data processing. The number of PMU devices and the size of the collected PMU data is growing quickly: By the end of 2010 around 40TB of PMU data were collected, the estimated data size for a five years period will be at half a Petabyte.

An analysis environment for low voltage networks is presented in [3]. The framework has modules for metering, data storage, analysis and simulation. The introduced Power Snap Shot method enables for synchronized smart meter data recordings (1-second-rms, P, Q, U), which are used as input for a DIgSILENT/Powerfactory based, automated and remotely controlled load flow simulation. The goal is to optimize simulation models for realism and accuracy, which is restricted by missing specifications of power grid components.

A new method for assessing the performance of real time PLC (Power Line Communication) based smart metering systems with the objective of enhanced operation of electricity distribution for low voltage networks is presented in [4]. In the monitoring stage the behavior of the PLC technology and the constant flow of application data are monitored. In the subsequent analysis stage the results are correlated with the physical features of the electricity network for problem identification and location. The underlying technology is developed by the PRIME Alliance (PoweRline Intelligent Metering Evolution), which is established in 2009 and aims to provide an open, extendable, public, and non-proprietary telecommunications architecture for smart grids.

The Open Source Phasor Data Concentrator (openPDC) project, which is administered by the Grid Protection Alliance (GPA) [5], provides a free software system designed to process streaming time series data in real-time. According to the openPDC homepage, captured PMU data is gathered with GPS-time from many hundreds of input sources and sorted by time. The data is provided to user-defined actions as well as to custom outputs for archival. It is scalable and designed to consume all standard PMU input protocols. It provides phasor data transformation and replication without loss, and

supports user configurable output streams, with an extreme low latency. However, performance statistics are logged every 10 seconds only.

Another free Phasor Data Concentrator based on the IEEEC37.118 synchrophasor standard is iPDC [6]. It includes a Database Server for iPDC and PMU simulator modules. The target audience is researchers and students for development and testing of new algorithms and applications related to PMU data. It is released as free software for users without any restriction regarding its usage and modification.

With regard to data analysis classical data mining approaches applied to smart meter data deliver valuable information. However their computational complexity limits the efficient processing of large scale data. A Hidden Markov Model based framework for individual predictability in user consumption at a population level was introduced in [7]. Based on energy consumption data (8 month, 1100 households, 10-min resolution) and local weather data (15-min resolution) this method allows to characterize user consumption and to perform segmentation, similarity clustering and profiling.

The data mining framework presented in [8] enables to process energy data streams with varying granularity for the prediction of energy consumption trends. An interim data summarization component processes incoming data streams, whereas an incremental learning and knowledge accumulation technique enables pattern extraction for voluminous, transient and randomly ordered smart meter data. A fuzzy pattern matching technique is used for data analysis and trend prediction.

Visual data mining in contrast to classical data mining involves the human knowledge, the experience and the intuition into the analysis of big data for discovering patterns, trends and hidden information [9][10]. The selection of appropriate visualization methods for encoding large scale data is challenging. A system for automated anomaly detection for power consumption data is presented in [11]. The introduced ‘pixel based time series visualization’ method supports prediction and clustering for fast anomaly detection. The project Lumberyard [12] provides iSAX indexing for time series data stored in HBase for persistent and scalable index storage. Lumberyard is based on the original iSAX paper [13] and uses the jMotif Java library [14], that provides symbolic aggregate approximation (SAX) [15] and iSAX for time series data enabling outlier detection and the search for so called motifs (often occurring patterns) [16][17]. For the analysis and representation of time series many approaches have been introduced in the past. Representations can be grouped to data-adaptive ones (e.g. Sorted Coefficient, Piecewise Polynomial, Singular Value Decomposition, Symbolic, Trees/Tries) and non-adaptive representations, including Wavelets, Random Mappings, Spectral and Piecewise Aggregate Approximation [13]. Another interesting approach to provide intuitive semantic searches in huge time series databases is described in [18], where so called intelligent icons are introduced.

These are small images representing the content of a file by a color scheme derived from its SAX representation. This allows for unexpected and serendipitous discoveries while working with data and has been integrated into the developed visualization tools in order to provide a great semantic overview of huge amounts of voltage measurement datasets and their similarity.

Within the scope of modeling and simulation for power grids there is a vast number of academic and commercial software packages, that cover all issues regarding the power grid analysis [19]. We are interested in well structured, easily extendable, open source simulation software that supports power flow and transient analysis. From the computing point of view the software package must provide parallel computing (Multicore, MPI, GPGPU) for large scale power grid models and very small simulation time steps. We evaluated commercial software (NEPLAN [20][21], MathWorks SimPower-Systems [22][23], ElektraSoft ELAPLAN [24]) and open source software [25] as LTSPICE [26][27][28], OpenDSS [29][30], PSAT [31][32] and GridLAB-D [33]. In this paper we concentrate on GridLAB-D (see chapter 4.1.), since it is Open-Source and supports multicore computing. It also provides a comprehensive library of power grid components and a text-based interface for hierarchical modeling. The software is developed and improved continuously. A good support for the vital user community exists.

1.2. Requirements for an advanced Power Grid Analysis Framework

Current power grid analysis frameworks rely mainly on smart meter data, which have a single and limited temporal resolution. This is sufficient for power flow analysis, the state of the art control of the distribution grid operation, but not for an in-depth transient power grid analysis. The future framework should support various power grid recording devices at multi-rate temporal resolutions. A low-cost, portable, high rate energy data recording device for the low voltage network is preferred, which enables easy and fast installation at electrical sockets.

The framework should provide secure and reliable data transfer from the recording devices to large scale data storage and should enable long-term storage of the data. The large scale data management must provide services for fast data querying and data retrieval. In addition, it should include the handling of large scale power grid model data and simulation results. Furthermore, power grid related data types from various domains should be supported by the data management, e.g. meteorological data, statistical data, etc.

A further important demand is the coherent processing of data from multi-rate data sources. The data analysis methods must be able to correlate data with varying temporal granularity and from multiple, irregular sources (energy data, weather data, GIS data, and statistical data). Multi-rate and multi-source data should also be supported by the power grid simulation software and the interactive

data visualization in the framework. The used simulation model description should be standardized and adaptable for various power grid simulation types. Suitable data interfaces to various simulation software packages should be supplied. The framework should support data-parallel and data-centric computing for big data analysis and could be supported by advanced remote software control architecture. High performance computing and GPGPU for large scale power grid simulations should be provided.

All tools in the framework should cope with the expected rapidly growing data amount. To our knowledge no such framework exists.

1.3. The KIT Power Grid Analysis Approach

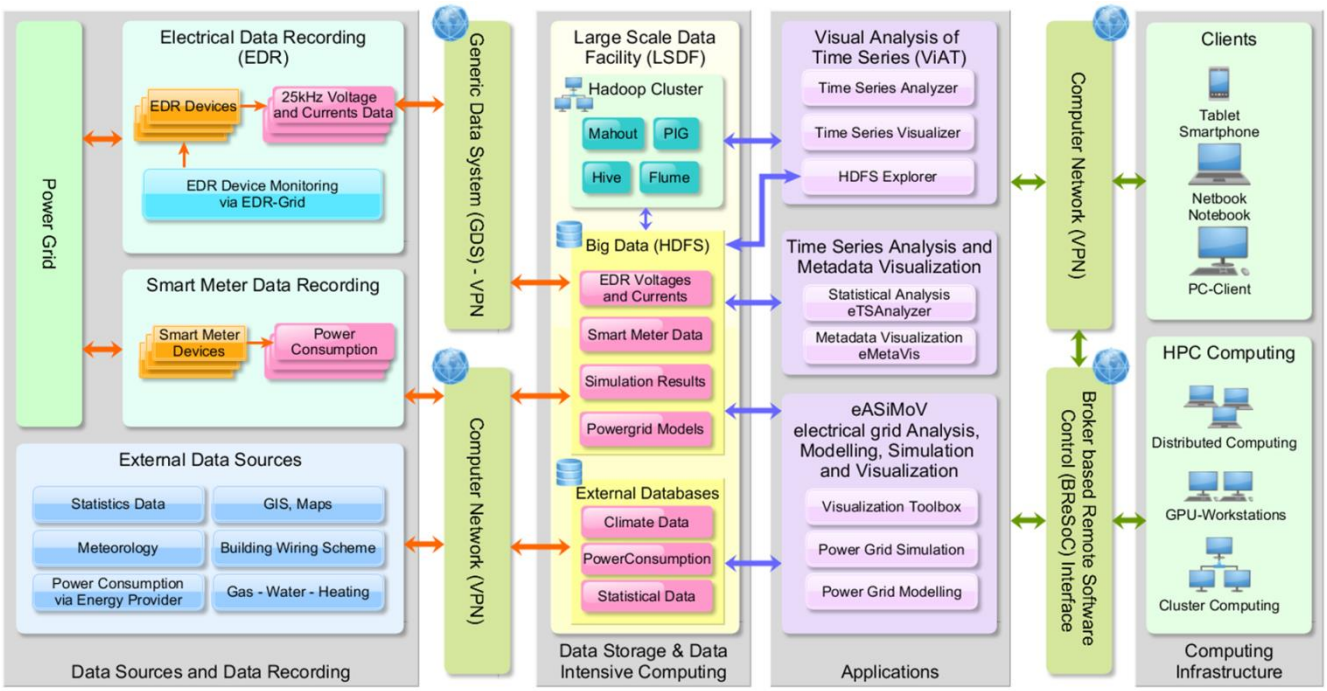
The main goal of the introduced approach is to develop a profound system comprising new analysis methods for the exploration of the power grid system dependencies. Hence, we combine transient power grid data recording, computer aided modeling and simulation techniques and new methods for the big data analysis and interactive metadata visualization. These topics are embedded in a sophisticated computing infrastructure that allows for secure power grid data transfer and remote software control for various computing facilities. As shown in figure 1, the new concept comprises four functional blocks:

- Data recording from the power grid and support for other data sources that relate to smart grid analysis, e.g. meteorology, etc.
- Secure transmission and storage for big data combined with data intensive computing services using the Hadoop file system and Apache Pig.
- Applications for statistical data analysis, modeling of power grids, simulation on high performance computing infrastructures as well as interactive visualization of recorded data, simulation results and metadata gained by data analysis.
- Computing infrastructure comprising the usage of super computers and the support for remote software control.

The first block (f.l.t.r.) in figure 1 describes energy related data sources and methods for the data capturing, which enable the derivation of valuable system state information. Since we assume that aggregated smart meter data is not sufficient when it comes to understanding the complex system dynamics, we capture high rate low-voltage and currents time series at different locations in the island-like network of the KIT Campus North (CN), which is the initial investigation site. The in-house developed Electrical Data Recorder (EDR), which complies with IEC 61000-4-30 and IEC 61000-4-7, is capable of recording three-phase voltage and currents time series at up to 25 kHz synchronously. Time series measurements with EDR devices have been performed since February 2012. Additionally, 534 electrical smart meters are installed at KIT-CN and provide active power

consumption data with an update rate of 10 minutes. Further external data sources can be used to refine the simulation models, e.g. meteorological data for the modeling of PV systems, statistical data for region specific power load curves or GIS data for automated integration of spatial information into the simulation models.

Figure 1. Overview of the KIT power grid analysis framework comprising high rate energy data recording, big data management, power grid modeling, simulation, parallel time series analysis, advanced high performance computing and remote software control.



The second block in figure 1 comprises the data storage and data intensive computing. This data storage facility and the introduced methods are not limited to specific power grid measurement devices. Any kind of data recording device, which produces time series data can be integrated into the proposed system by means of data transfer, data storage and data analysis. In this paper we concentrate on data captured by EDR devices and smart meter data. The large amount of captured power grid data is stored at the Large Scale Data Facility (LSDF), which is maintained by the Steinbuch Centre for Computing (SCC-KIT). The secure data transfer is conducted over Virtual Private Network (VPN). Together with the data transfer monitoring system, which is based on the EDR-Grid infrastructure, the data transmission network allows for a secure, correct and reliable transfer between the EDR devices and the LSDF. We store the entire data derived from the smart meters and from other external databases for modeling and simulation purposes in the LSDF. Besides being used as a pure data storage facility, the Hadoop cluster, as part of the LSDF enables data intensive scientific computing. Further, we developed semantic data analysis methods that

enable searching for significant events in the recorded power grid data as power grid outages, frequency drops, etc. The implemented Apache Pig based scripts for parallel data analysis are incorporated into the in-house developed software ViAT (*Vi*sual *A*nalysis of large *T*ime series data).

A further major software development project at SimLab Energy is *e*ASiMoV, which incorporates analysis, modeling, simulation and visualization methods for power grids (see 'Applications' block in figure 1). Based on the CIM-IEC 61970 specification, we enable the interactive GIS-based modeling of power grids. Interfaces and converters to power grid simulation software enable us to perform power flow simulations for the distribution grid. Currently we support GridLAB-D and aim at providing model data converters to Matlab-SimPowerSystems and other commercial simulators. Additionally, we pursue the in-house development of a simulation kernel, which allows the transient power flow simulation for radial distribution feeders using the Backward-Forward Sweep method. The integrated visualization toolbox provides interactive display and exploration for the simulation results as well as for the recorded EDR and smart meter data via direct access to the LSDF storage. The basic statistical analysis method provided by the software tool *e*TSAnalyzer (*e*nergy *T*ime *S*eries *Analyzer*) generates metadata that is hierarchically structured in the time domain. The interactive data exploration of the hierarchical metadata with *e*MetaVis (energy *Meta*data *Vi*sualizer) allows the detection of missing, faulty or significantly anomalous data.

All software tools can be executed locally, or over the network using the new broker based remote software execution and control environment BReSoC (*B*roker based *Re*mote *So*ftware *C*ontrol), which connects the 'Applications' and 'Computing Infrastructure' blocks in figure 1 and aims to support various computing technologies (cluster computing, distributed computing, etc.). A typical application scenario is the use of thin clients like smartphones or tablets to securely access the power grid data from the LSDF, in order to perform semantic and statistical analysis on the Hadoop cluster and to interactively navigate within the data visualization of the analysis results. A further application scenario is the remote modeling and simulation of highly complex power grids using thin clients.

This paper is structured according to the four functional blocks presented in figure 1. Chapter 2 introduces the new energy data recording device and the data types gathered from various sensors. Chapter 3 presents a unique storage facility for big data together with new methods for parallel semantic data analysis based on Hadoop. In chapter 4 we give an overview of the software tools for modeling, simulation, basic statistical energy data analysis and visualization. In the last chapter we introduce the new computing framework. Results are presented directly in each chapter.

2. Power Grid Data Recording at the KIT Campus North

The initial investigation site is the Karlsruhe Institute of Technology (KIT) Campus North (CN), which is located approximately 10 km north of Karlsruhe, Germany. The electrical power supply is provided by one incoming power transmission line with 110 kV from the local energy supplier. Additionally, a 2 MW block heating station contributes to the electrical power generation. The power consumption of the KIT Campus North has an average load of 20 MW. The base load is about 10MW, the peak loads are about 22MW.

2.1. High Rate Energy Data Recording

We developed the so-called EDR device (Electrical Data Recorder) for recording high rate, low-voltage and current time series [34]. It consists of an 8-channel DAQ board, a GPS-unit and a computation unit with internet access. The maximum capturing rate is 25 kHz for three-phase voltage and the fourth channel is used for precise synchronization with a GPS pulse-per-second signal. Since February 2012 we have been conducting continuous voltage time series measurements in this island-like electrical network. The captured time series raw data is stored locally in the XML data format together with metadata on the recording time, location, measurement device id and further properties. Each XML file holds one minute of recorded data, thus it contains 60 blocks of metadata and recorded voltage and currents raw data, which are Base64 encoded from 16 bit raw binary data. Based on this raw data, the computation unit performs a pre-extraction of features and characteristics according to EN50160 [35]. These are the fundamental frequency, maximum and minimum voltage, 1st to 16th harmonics proportion, THD (total harmonic distortion), the absolute phase angle (the angle of the voltage sine wave at the beginning of a second), effective voltage RMS (Root Mean Square), ARV (Average Rectified Value) and others. The feature data is accumulated for a whole day and saved locally as a separate CSV file for each phase. The CSV file has 86,400 lines with a resolution of one second per line comprising of a time stamp and 29 feature values. The file names provide information about the type of data (Raw, Feature), the recording device id, the channel, date and a time stamp. Currently the amount of the recorded data is about 8.36 GiB a day for one EDR device using 12.8 kHz sampling rate. The feature data file for one phase is about 24 MiB per day per EDR device for each phase.

2.2. KIT Smart Meter Data

Besides smart meters for gas and water, 534 electrical smart meters are installed at the KIT-CN and maintained by the department KIT-TID (KIT-Technical Infrastructure and Services). The power consumption is recorded and registered at a server’s central database. For research purposes, the

KIT-TID is providing the power consumption data of KIT-CN via a network drive. The state of all electrical smart meters is flushed into a CSV file every 10 minutes. Thus a file contains a date and time stamp, the id of the smart meter, the absolute smart meter reading of the power consumption and the unit [kWh]. The filenames consist of the date and the time stamp; all data files are saved in one folder. The network drive holds 52,416 data files (2,127.51 MiB) for one year, the network storage service started in December 2012.

3. Large Scale Data Storage and Data Intensive Computing

High-rate power grid data recording devices produce huge amounts of data, which need to be stored, managed and analyzed. The new concept allows for secure and reliable data transfer and storage mechanisms as well as data-centric and data-intensive analysis methods in the framework. It enables to process any type of power grid time series data, independent of the data source and update rate.

In this paper we focus on data captured by the KIT EDR devices and smart meter data. The EDR devices use the local hard disk of their computation units (notebooks) to cache the recorded data for a limited time period. After aggregation we transfer the data via VPN to the LSDF and thus enable for storage, retrieval and data intensive computing for big data.

3.1. The LSDF Data Storage and Computing Facility

The Large Scale Data Facility (LSDF) is maintained at KIT, Steinbuch Centre for Computing (SCC) [36]. The EDR devices are transmitting the recorded data to the LSDF via the internet using a SOAP-based web service of the so called Generic Data Services (GDS) [34]. The data is stored in the LSDF's HDFS (Hadoop File System) storage with a rather flat file organization: each day is represented by a folder named by date, which contains the raw and feature data files. The data retrieval is currently available via the web browser interface to the LSDF and the in-house developed tools utilizing the Apache Hadoop Java API.

Beside the usage as a pure data storage facility, the LSDF also enables data intensive scientific computing with the Map-Reduce paradigm using Hadoop [37]. The Hadoop cluster as part of the LSDF provides access via two name nodes, which are identical to the data nodes except of the increased 96 GB of RAM and a 10 GE network connection. The cluster consists of 58 data nodes with 464 physical cores, each node has two Intel Xeon CPU E5520 (2.27GHz, 4 cores), 36 GB of RAM, 2 TB of disk, 1 GE network connection, OS Scientific Linux 5.5 and Linux kernel 2.6.18. The Cloudera Hadoop distribution CDH3u5 is currently installed, providing different interfaces. We use the Apache

Pig API [38] for preprocessing EDR data, semantic analysis, and similarity patterns searches in the recorded data and for correlation analysis between low-voltage feature and smart meter data.

3.2. Data-Intensive Computing Methods for Time Series Data Analysis

The main issue that arises when processing huge time series data with a standard PC architecture is the decrease of processing speed whenever the size of the data to be processed exceeds the size of the RAM. Then the next slower storage layer must be used, which is the hard disk. The same problem arises as soon as the storage limit of the hard drive is reached: Data must then be fetched from the archive servers over the Ethernet/internet, which is again slower than reading from the local disk.

Since measured time series are not stationary concerning a one-day-dataset, we need to preprocess or analyze data of many measuring devices for time ranges such as months or even years regularly. This may involve data sizes up to hundreds of terabyte, which would not only be slow on a standard PC but just not possible at all with the current maximal hard disk sizes of ~4 terabyte. An alternative is data processing with Hadoop, which focuses on large scale data intensive computing and can in theory handle unlimited amounts of data by distributing on a scalable number of cluster computers. The approach of Hadoop and similar projects utilizing the Map-Reduce paradigm is to store the data distributed and redundant on many computing nodes and bring the computing to the data. In a previous study [39] we could show that the processing on Hadoop performs very well applied to the recorded time series data and that it outperforms classical multicore parallel processing on the utilized workstation (see chapter 5.2. for the workstation specifications) if the data size grows larger than about 6.2 GiB.

To reduce complexity and development time, we do not directly use the Java Map-Reduce mechanism, but utilize Apache Pig, which is a high-level platform for creating Map-Reduce programs that can be run in environments like Hadoop. It includes a special language called Pig Latin, which abstracts Map-Reduce programming into a high level notation, similar to that of SQL for RDBMS systems. We extended Pig Latin using custom UDFs (*U*ser *D*efined *F*unctions) written in Java and embedded Pig Latin for maximum flexibility for the in-house developed big data exploration and visual analysis software for large time series (ViAT).

3.3. Data Abstraction and Metadata Generation

The Hadoop based computing environment enables for fast large scale data analysis. However, processing may still take some minutes, depending on the size of processed data. We need additional tools for real-time local interaction with the data like interactive visualizations, semantic searches,

finding and clustering similar datasets, pattern mining, histograms etc. So, several tools were developed that may be divided in two categories:

- Category-1: Local tools operating on data fetched from the archive cluster into local RAM.
- Category-2: Tools for data analysis and metadata generation in the Hadoop environment.

Category-1 operates on the original EDR feature data, but the maximum processible data size and therefore the largest time ranges are limited by the client's RAM. For category-2 the data is preprocessed on the cluster first in order to extract descriptive, structural and administrative metadata, which is then visualized or further analyzed on the metadata level.

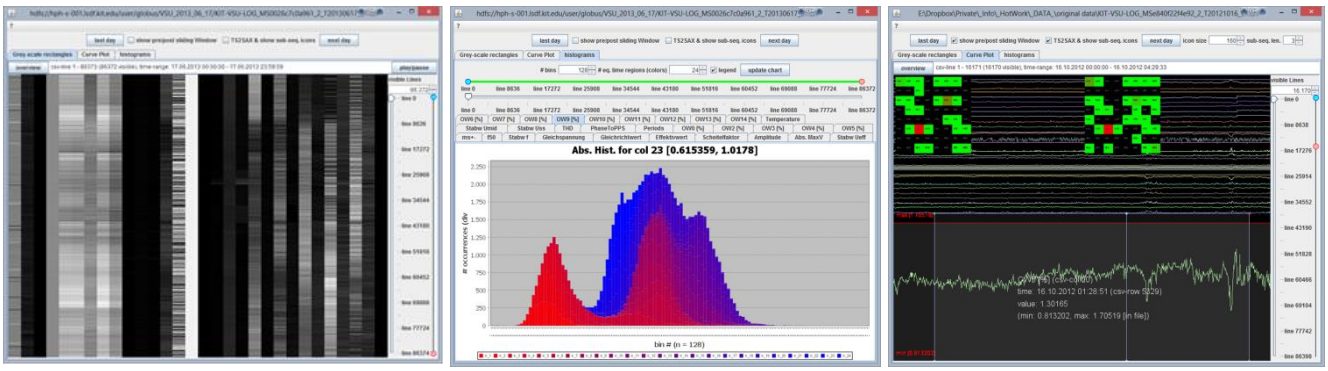
3.3.1. Local Time Series Analysis Tools

The developments in the first category comprise a data exploration tool and a histogram tool. The exploration tool provides two different interactive visualizations for EDR-feature data: the first type is the time-axis oriented plotting of the original EDR-data as curves for each feature. For the second type all feature values are encoded by rectangles with a grey scale color that is computed by linear scaling between minimum and maximum data values (see figure 2-left). All feature values are visualized horizontally side by side along the vertical time-axis as defined in the CSV based feature data sets. Both visualizations are always in sync regarding the current exploration position, visual range and other parameters.

Additional information can be gained with the histogram tool for the feature data sets. For a selected feature we divide the range of values into an arbitrary number of so-called *bins* counting the number of value occurrences. The histogram additionally can be discretized over the time, which gives information on the development of the histogram for one day for a selected feature (see figure 2-middle).

For both visualizations, a conversion into the Symbolic Aggregate Approximation (SAX) representation [15] can be done, enabling the detection of so called *discord* patterns (patterns that are occurring only casually) and *motifs* (often occurring patterns). The approach presented in [16] was also implemented, providing advanced anomaly detection using a sliding window technique that compares an intelligent icon [18] derived from the SAX-representation of a 'history-window' with the intelligent icon built from the SAX-representation of a shorter or equal sized 'future-window'. The window can be interactively resized and moved across the time series data and the two intelligent icons are automatically updated (see figure 2-right). This method allows for quick detection of sudden changes in time series data, which is indicated by the significantly differing color encoding of the intelligent icons for the history and the future window.

Figure 2. New interactive visual analysis tools for time series energy data: normalized color mapping (left), histogram tracing (middle) and SAX representation (right).



3.3.2. Data Analysis in the Pig Environment

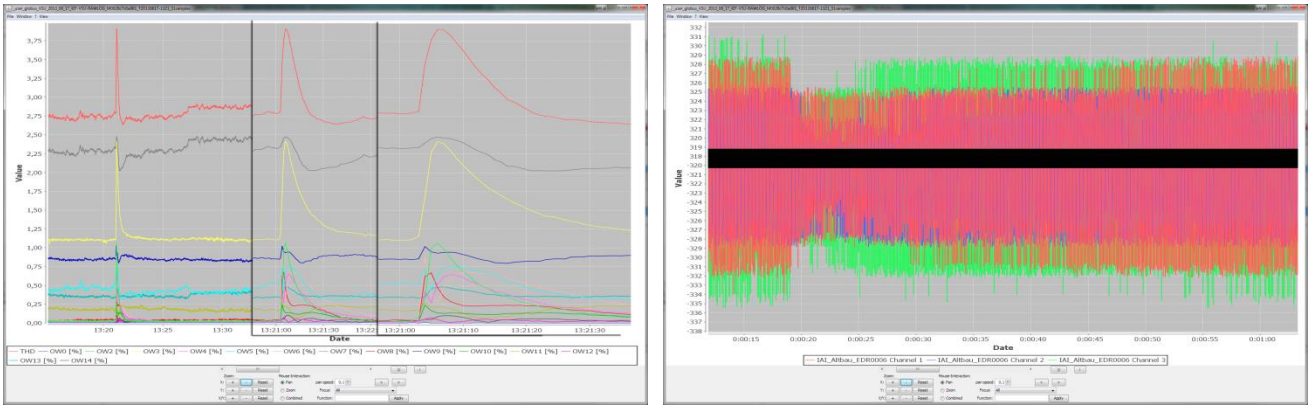
As introduced in [39] we are developing custom Java UDFs (*U*ser *D*efined *F*unctions) for statistical analysis of EDR feature data on the Hadoop cluster. The results of the Pig based statistical analysis are summarized in a hierarchical metadata structure. This metadata will be used for a faster identification and classification of datasets (e.g. for outliers) but also for hierarchical visualization of feature data characteristics using a scalable vector graphics (SVG) based representation, which will be generated on the Hadoop cluster and embedded or linked to the metadata structure. Metadata currently is updated manually, but an automation of this process is in development, so that for each new dataset arriving on the cluster metadata will be generated automatically. The energy metadata will be combined with additional metadata obtained from various data sources, e.g. tweets, RSS news, etc. which will help to correlate real world power grid related events to the measured energy data. In chapter 4.3. the basic statistical time series analysis and hierarchical metadata visualization for EDR feature data and smart meter data is introduced as a multicore implementation, a Hadoop based implementation is in progress.

3.4. Case Study: Detection of a Transmission Network Voltage Sag in the Supply Network

The KIT-TID registered and announced a 30 milliseconds voltage sag in the KIT supply network for the 17th of June 2013 at 2:21 pm UTC+1. The source of the failure was located in the transmission network outside of KIT, the cause was unknown. Based on this information we localized the relevant time series data files on the Hadoop file system. First, we analyzed the feature data available at a rate of one per second and visualized the voltage sag in the low voltage supply network, which is shown in figure 3-left in three zoom levels (f.l.t.r.): we observe peaks in the THD and in the first 14 harmonics and also see the automated voltage recovery. This event could also be identified in the low voltage, 5 kHz sampled raw data shown in figure 3-right. Here we observe an average reduction of the voltage

amplitude at about 4V at 1:21:03.82 pm UTC for phase1 and phase 2, and at 1:21:04.0 pm UTC for phase 3. For an enhanced visualization one data sample was mapped to one millisecond in the diagram in figure 3-right starting at 00:00:00 (time scaling factor 5) and the voltages between -320V and +319V were cropped.

Figure 3. Effects of a voltage sag in the transmission network in the high-rate, low voltage data: the significant changes in the THD and in the first 14 harmonics are shown in three zoom levels (left) and the corresponding raw voltage data with the reduction of the voltage amplitude for all three phases (right).



The benefit of recording and archiving of high-rate raw power grid voltage and currents data is evident: Significant events in the power grid can be provided to research groups for developing and evaluating new analysis methods based on data with high capture rate. The symbolic big data analysis method as introduced in the previous chapter can consider the recorded voltage profile and the derived features characteristics of a known event as input. An encoded SAX representation of this information can serve as a search pattern for massively parallel data mining for similar events. The SAX based data analysis will enable the prediction of outages, frequency changes or voltage sags for real-time captured EDR data streams in the future.

4. Modeling, Simulation, Analysis and Visualization Methods for Power Grids

The *e*ASiMoV (*e*lectrical grid *A*nalysis, *Si*mulation, *Mo*deling and *V*isualization) software was initiated to provide a framework for various software modules comprising modeling and simulation of power grids, the presentation of power grid simulation results, the analysis of simulated and recorded power grid data combined with appropriate visualization methods.

4.1. Power Grid Modeling and Simulation with eASiMoV

Modeling and simulation of the power grid can contribute to understanding the power grids dynamic behavior. We propose a two phase algorithm: In a first phase we conduct model identification with power consumption data gathered from the KIT-CN smart meters and the aggregated voltage data from the EDR devices as input. The goal of this phase is the parameter estimation of the impedances for the power grid components with the known topology. The second phase is the simulation process for the identified model. Here, we will use the voltage measurements as input for the calculation of the currents.

4.1.1. Unified Power Grid Model Description

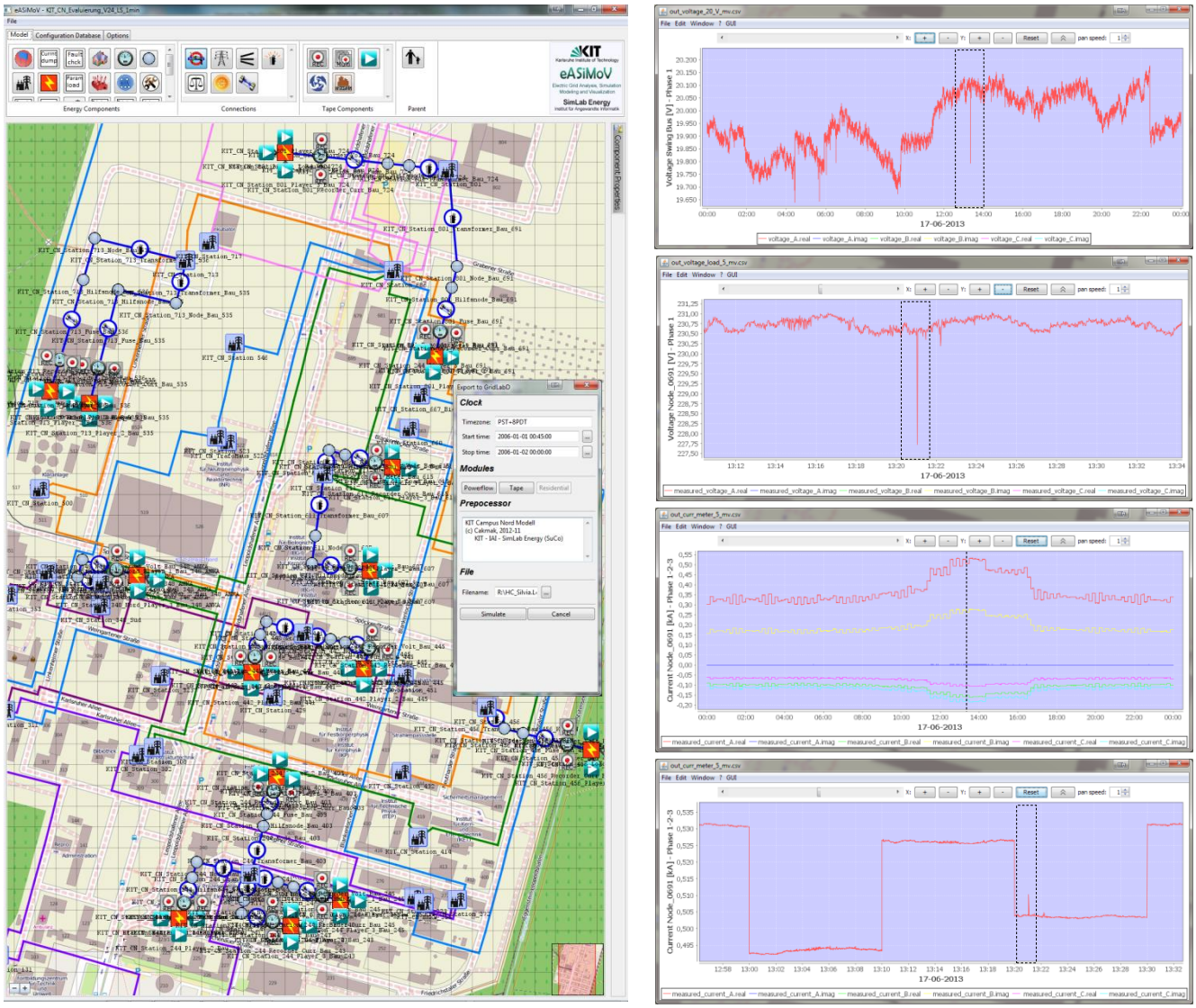
According to the topology of large scale power grids we propose a hierarchical graph-based representation with different levels of detail and additional meta-information. Owing to its simplicity in handling and the huge number of available processing tools, a XML-based graph file format is used. A detailed comparison of available graph description and exchange formats is given in the introduction for GXL [40]. GraphML [41] fulfills the requirements for universality, typing, flexibility, ease of use, scalability, modularity, and extensibility. The possibility for a hierarchical graph description together with sub-graphing enables to model power grids in various levels of detail. The initial specification of power grid components based on the GridLAB-D [33] object description. Since this model description is isolated and has no compatibility to current model description standards, we also support the Common Information Model (CIM) IEC 61970 [42] in *e*ASiMoV. With the distinct object management in *e*ASiMoV we can use the same Java based GUI for modeling with various object representations. For the CIM integration we used the open access CIMTool [43] [44] and the latest available CIM package from the CIM users group. We exported the IEC 61970 package and adapted to the internal class definitions in *e*ASiMoV besides existing GridLAB-D based Java classes for object representation. Current work focuses on creating suitable data interfaces that allow for utilizing and combining the simulation kernels of multiple existing simulation software specialized to different aspects of the power grid.

4.1.2. Case Study: KIT Power Grid Model and Power Flow Simulation with GridLAB-D

As introduced in chapter 3.4. we are able to register the effects of the transmission network voltage sags in the KIT supply network. In a first study we simulated this event utilizing a power grid model of the KIT-CN together with the available smart meter data and the recorded high rate low voltage data. We evaluated the simulation software GridLAB-D 2.2 developed by the U.S. Department of Energy for a three phase unbalanced distribution power flow simulation. With the *e*ASiMoV modeler we created a simplified partial model of the KIT-CN supply network (see figure 4-left). Only seven substations with

14 power lines are modeled and shown in the figure, this is about 1% of the total power grid of the KIT-CN with 34 substations, each with 51 power lines supplying multiple facilities. The underground line installation does not correspond to the reality due to security reasons, which prevents publishing real data.

Figure 4. *e*ASiMoV with the partial simulation model of KIT-CN with seven substations and 14 power lines (left) and GridLAB-D power flow simulation results: 20kV input voltage with the voltage sag and simulated impacts on the low voltage network (right).



The model contains the available grid component specifications. The active power values based on the 10 minutes smart meter power consumption data at KIT-CN and the 5kHz sampled high rate EDR voltage data are used as input for the simulation model. GridLAB-D manages the data synchronization for the multi-rate input time series and performs the calculation of the power flow with

the Newton-Raphson method. The output of the power flow simulation is a collection of time series (currents, voltages, powers), gathered at so-called *recorder objects* in the simulation model.

The goal of the simulation was to determine the value of the voltage sag in the transmission network by parameter variation. For that we compared the simulated average reduction of the voltage amplitudes with the measured 4V voltage sag in the supply network (see chapter 3.4.). The results of the GridLAB-D based power flow simulation are visualized with *e*ASiMoV. Figure 4-right shows from top to bottom the 20kV phase-1 input voltage for the 17th of June 2013 with the determined voltage sag value of 19.793 kV at 1:21:03 pm UTC, the detailed view of the simulated voltage sag in the supply network, the three-phase currents at a selected node for the whole day and a detailed view of the phase-1 current. The load data for the presented node was taken from the smart meter data of a smaller testing plant at KIT.

The main challenge with power grid modeling is to acquire specifications of the installed power grid components for a realistic simulation, which were not entirely available. Another problem is the availability of recorded low voltage data, since currently only a limited number of EDR devices exist. For the medium voltage swing bus we do not have any measured voltage data, thus for the presented simulation the input voltage time series was taken from the closest substation and adapted according to the substation specifications. We also faced several problems with the selected power flow simulator GridLAB-D. It is not possible to feed any number of recorded voltage time series into a Forward-Backward-Sweep or Newton-Raphson based power flow simulation. Recorded voltage data as input is only possible for the swing bus, which is the connection of the 20kV network to the transmission network at KIT.

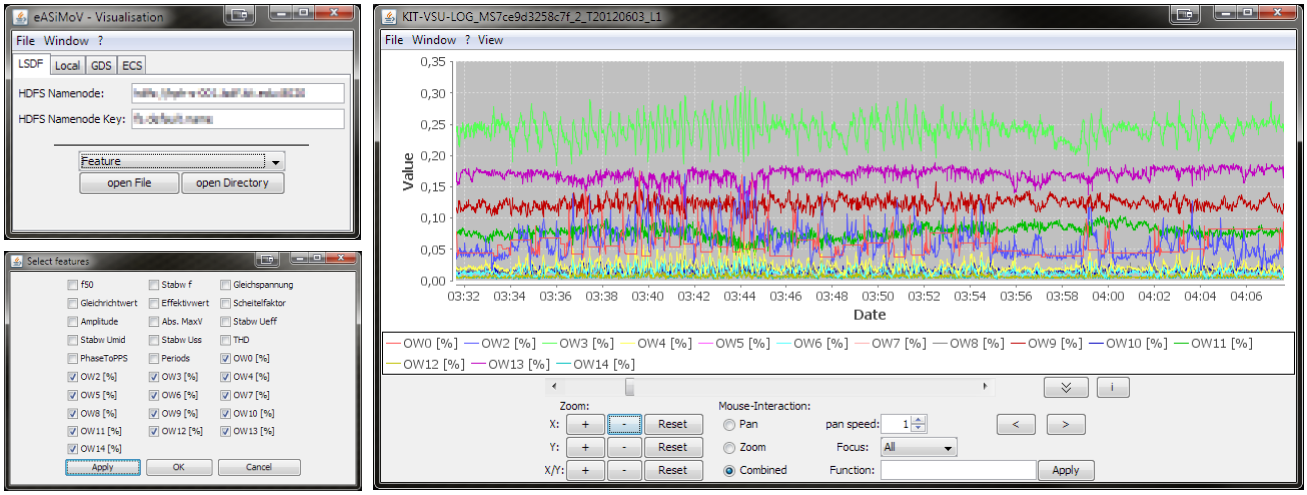
Our experiments showed that a one-second simulation output is not sufficient for transient analysis. Also the limitation of being not able to supply a simulation model with an unlimited number of recorded high rate voltages and currents is a major drawback of GridLAB-D. We also dropped the initial goal of porting the GridLAB-D software to a massive parallel simulation kernel using MPI for computer clusters due to its complexity. Further reasons for this decision were that the software was not running stable, especially in multicore mode, and against the expectations parallel computing was slower than serial calculation. As a consequence, the development of a new power grid simulation kernel *e*SimCore as part of the *e*ASiMoV framework is pursued, which, in its early development phase, supports power flow simulation for radial feeders. Since high performance computing is essential for simulation of large scale power grids with an acceptable time window, *e*SimCore aims to support various electric power simulation topics combined with high performance computing as introduced in [45].

4.2. Interactive Data Exploration of EDR Feature Data with eASiMoV

In addition to the support for the visualization of simulation results as shown in figure 4, the data visualization module of *e*ASiMoV enables interactive data exploration of EDR raw and feature data stored locally or on the LSDF.

For a selected time period the data visualization module creates a list of EDR devices that provide recorded EDR data. After selecting the time scale for visualization, the module renders the selected features using the JFreeChart library. Simple operations for two selected data sets as difference, or the simple detection of outliers based on the median value of a data set are permitted. With a redesign of the GUI interface we support the direct data file access on various storage infrastructures (see figure 5). The tool caches an appropriate time series data range in the memory for smooth transitions during the visualization process. The visualization of EDR raw data is also supported, but does not deliver cognizable visual information.

Figure 5. *e*ASiMoV-Visualization module for direct rendering of EDR raw and feature data sets from the LSDF: connection to the LSDF (top-left), selection of features (bottom-left) and visualization of the harmonics as a subset of the EDR feature data (right).



4.3. Basic Statistical Data Analysis and Metadata based Visualization of Energy Data

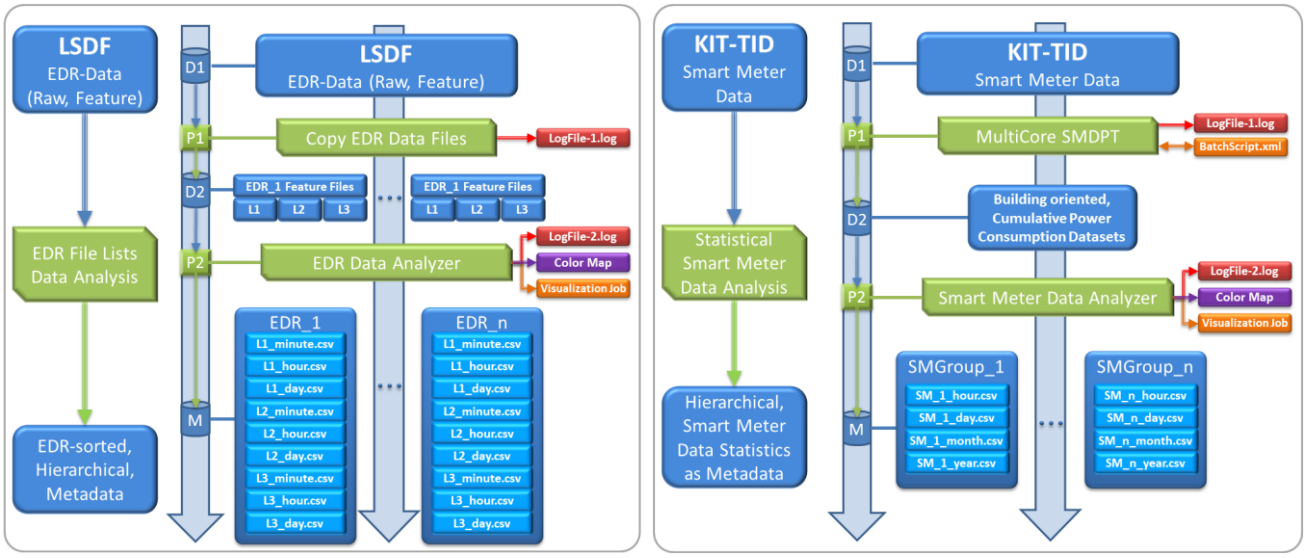
In chapter 3 we introduced current developments for the data analysis operating on the Hadoop cluster enabling the search for similarities, patterns and special properties in the recorded data sets. In addition to this symbolic analysis we provide basic statistical data analysis methods with the goal of extracting hierarchical metadata in the time domain for visualization and assessment. We intend to deliver a compact and informative visual data representation based on metadata for data related services, e.g. data recording, transfer, storage, etc. For a given type of data, we perform a series of hierarchical calculations of selected basic statistical characteristics (minima, maxima, average,

deviation and the median). Based on the analysis results on the highest available level of detail for time series data (e.g. seconds) the results are aggregated to the next coarser representation level (minutes); this is repeatedly done until we have the values of statistical characteristics for the coarsest level (year). The statistical analysis is performed on a local workstation, after gathering the relevant data files from the LSDF. According to the classification in chapter 3.3., this basic statistical analysis is of type category-1: EDR feature data is transferred from the LSDF to a local computer and metadata is generated for local and remote visualization and data exploration. Current work focuses on the upgrading of the analysis methods to category-2 in order to perform data analysis directly on the Hadoop cluster. The *e*ASiMoV statistical analysis software module *e*TSAnalyzer (*e*nergy *T*ime *S*eries *Analyzer*) for energy data has been applied to the following three data modalities:

1. File properties of EDR feature data stored at the LSDF.
2. Content of EDR feature data recorded by the EDR devices.
3. Power consumption data recorded by KIT-CN smart meters.

An overview of the data processing chain for the EDR and smart meter data modalities is given in figure 6. For each processing step (P_m) a detailed Log-file is created that holds errors among others. At every processing step also temporary CSV data files (D_{m+1}) are created, serving as an input for the next processing step. The data analysis results are a collection of CSV based, hierarchically organized metadata (M), a calculated color map for visualization and for heuristic data range validity checking as well as a script file for the visualization job. Each analysis chain for the three data types is described in the following subsections.

Figure 6. The data processing chain for the analysis of the content of EDR feature data (left) and the KIT power consumption (right).



For the interactive metadata visualization and exploration, we developed the standalone visualization software *e*MetaVis (*e*nergy *Meta*data *Vi*sualizer), which is highly optimized for fast visualization. As a light-weight tool it is applicable for remote visualization even on portable mobile devices. The hierarchical visualization of metadata based on statistical characteristics in the time domain gives a good overview of the big data quantity and quality by enabling a fast visual detection of missing and deviating data but also of occurring patterns in the time series data. Since errors and outliers are inherited to the next coarser level of detail in the metadata hierarchy, assessment of data quality is possible, i.e. if time series data for a certain time period is within the scope of empirically expected values.

4.3.1. Analysis and Assessment of EDR Data Acquisition and Transfer to the LSDF

Using the Java Hadoop API *e*TSAnalyzer accesses the HDFS, gathers file property information about the stored EDR files (file name, file size and file timestamp) and finally performs an analysis based on these data file properties. The goal is to detect errors during data acquisition and file transfer to the LSDF.

In a first step we gather a detailed file and directory listing information from the HDFS. The filenames of the EDR data sets contain the recording date, the EDR device id and channel-id. Since we are interested in the EDR feature data, in a second step the feature file properties are sorted according to the EDR device and their recording periods. For each data subset a CSV file is created, which contains the date and the file size. During the hierarchical calculation of the selected basic statistical characteristics for the file sizes, we create a Log-file with the number of files with deviating file size from the median file size (ERR_0) and faulty creation date (ERR_1: date in past, ERR_2: date in future), but also the number of errors during the EDR data recording and transmission to the GDS server (ERR_3) as well as the number of errors during the data file transfer from the GDS server to the Hadoop file system (ERR_4). The error recognition is possible, since in case of an occurring error, all subsystems for data transfer will modify the filename by adding a predefined suffix and retry to transfer and save the files.

With this method we analyzed the quality of the EDR data recording, transfer and storage of power grid data on the LSDF in the early phase of the system set up (10/02/2012 – 09/01/2013). The gathering of detailed file and directory information from the LSDF took about 18 seconds for 355 directories containing 955,798 EDR raw files in XML format and 103,198 EDR feature data files in CSV format. The locally performed analysis took about 5.64 seconds for the EDR feature data files on an Intel i5 3.3 GHz CPU in single core mode. We could classify 1,612 EDR data files as plausible, 101,543 files had a significant smaller size than the median of the file sizes (ERR_0), 1,081 files were

dated earlier (ERR_1) and six files were dated in the future (ERR_2). 97,348 files were classified as ERR_3 and 100,211 as ERR_4. The extraordinary high number of file size errors of type ERR_0 correlate to repeated data transfer errors (ERR_3, ERR_4). The data transfer and storage errors classified as ERR_3 and ERR_4 are not disjoint classifiers, a file marked as faulty can be assigned to both error classes and thus counted twice. Detailed analysis of the Log-files delivered valuable information about the remarkable high number of data transfer errors: 97.4% of all errors occurred on the weekend of 30/03/2012. Apparently a maintenance downtime of the LSDF caused a chain reaction, so that the GDS system could not store on the LSDF, and caused the EDR devices to re-send their data repeatedly to the GDS. In consequence of this analysis we revised the data acquisition and the data transfer, which resulted in an increased measurement accuracy and data transfer stability.

4.3.2. EDR Feature Data Analysis and Assessment

We also investigated the quality of the captured EDR feature data. The quality criterion is based on the number corrupt data files, the number of missing data entries and the number of data entries with identical time stamps. We concentrated on the feature datasets, which are derived from the raw data.

As shown in figure 6-left, the tool *e*TSAnalyzer first accesses the EDR feature data files on the HDFS of the LSDF via the Hadoop Java API and copies the relevant data files to the local hard disk. During the copying operation the feature data are sorted according to the filenames and grouped per logical voltage phases (L1, L2 and L3) and per EDR device. In the following step *e*TSAnalyzer processes each EDR feature data file for each phase, and checks for the validity of the data. For each of the 29 features in an EDR feature data file with the resolution of seconds (each line of the original CSV feature data file with total 86,400 lines for one day) the five basic statistical characteristics (minima, maxima, average, deviation and the median) are calculated and the result is aggregated in the next coarser level of detail (minutes). The aggregation is repeated for the hour and day representation.

The tool *e*TSAnalyzer was applied to an early and a recent EDR feature data set in order to investigate the improvement of the EDR data recording quality in the last two years.

The early EDR data set was recorded from 01/06/2012 to 29/06/2012. For the 87 feature data files (29 files for each of the three-phases) with a total size of 1.81 GiB the data transfer from the HDFS to the local workstation took about 27.62 seconds. The total processing time for the analysis was about 114.4 seconds on an Intel i5 CPU with 3.3 GHz in single core mode, the EDR file import and parsing was about 49.1 seconds, and the statistical calculation and file output took about 65.3 seconds. 17 of 29 files for each phase were correct, for each phase three CSV files were corrupt (ERR_0), 95 entries

had the same time stamp (ERR_1), and 79,114 entries in the original CSV files were missing (ERR_2), which was caused by the missing entries for the corrupt data files. The error rate for the total dataset is (3 phases x 79,209 errors) / (86,400 CSV data lines x 87 files) = 3.16%. A detailed inspection of the Log-files indicated technical problems with the data recording for three days in this recording period. The error rate was 42.3 ppm after excluding the corrupt data files from the evaluation.

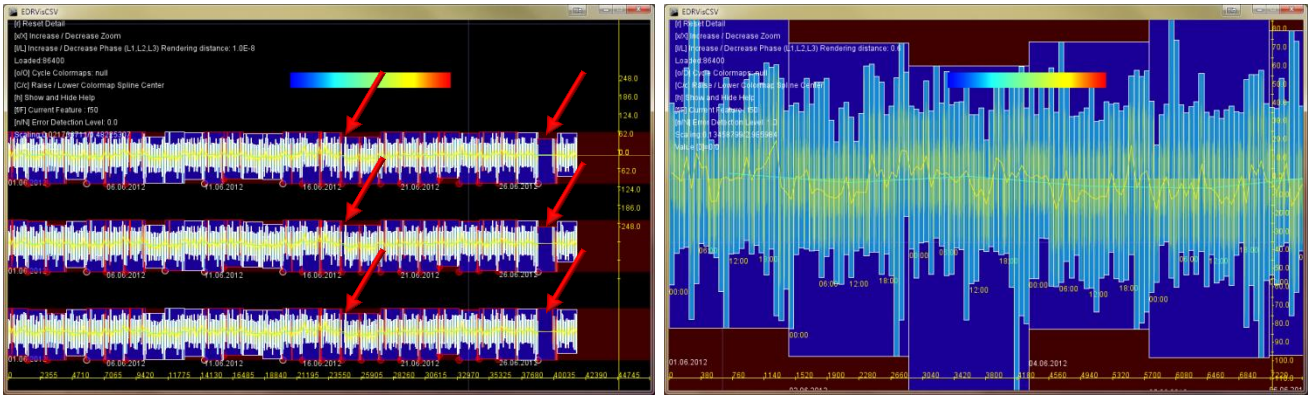
We applied the basic statistical analysis method to the recent EDR feature data set, which was derived from three-phase raw voltage data captured by seven EDR devices from 10/10/2013 to 27/12/2013. The EDR feature data with a total data size of 19.9 GiB are stored in 844 data files with an average file size of 24 MiB. The computation time for the analysis was about 19 minutes for an Intel Xeon 3 GHz CPU in single core mode, 7.5 minutes for a quad core Intel i5 3.3 GHz CPU and 3.1 minutes for a 12 core Intel Xeon 3 GHz CPU. The evaluation of the Log-files enabled the detection of time stamp errors in 65 data files. An in-depth data inspection showed that in total 621 CSV data lines for all phases shared a common time stamp (ERR_1) and thus the same number of data lines for the subsequent seconds were missing (ERR_2). No corrupt data files were found (ERR_0). The error rate for the total dataset was 621 / (86,400 CSV data lines x 844 files) = 8.5 ppm.

Based on the basic statistical analysis of EDR feature data with *e*TSAnalyzer we could show that the quality of the EDR data recording could significantly be improved by factor of 4.98 for the analyzed data sets in the last two years. Thus, the analysis allows for assessment of the EDR data recording quality as well as for classification and exact localization of errors in large time series data.

4.3.3. Visualization of EDR Feature Metadata with *e*MetaVis

The following figure shows *e*MetaVis with the interactive visualization of the metadata, which is a result of the basic statistical data analysis for the early EDR feature data set and which is hierarchically structured in the time domain. The hierarchically organized metadata for the frequency time series is rendered, which is possible for each of the 29 EDR features. Based on the results of the analysis, a data validity range is determined and outliers are visually highlighted with red boxes and red dots. The user may interactively adjust the validity range (dark red background stripe) to determine extreme outliers in the data set. The nested boxes represent years, months, days, etc. from the coarsest to the finest level of detail. The hierarchical data representation indicates missing data blocks for the 18th and the 28th June 2012 (red arrows for each phase). The expansion of the validity range for the data set effects that smaller outliers are classified as correct. An extreme outlier for the 28th June still remains, which apparently is an indication of technical recording problems and the missing data; this requires a further in-depth examination with e.g. the previously introduced semantic data analysis method.

Figure 7. Visualization of hierarchical metadata gathered from the basic statistical analysis of EDR feature time series for the three-phase power grid frequency with automatic error recognition. The red arrows show the missing data (left). Zoomed visualization with various levels of detail in the time domain: per day (dark blue), per hour (light blue), per minute (yellow) as well as the daily and hourly averages (right).



4.3.4. KIT-CN Smart Meter Data Analysis

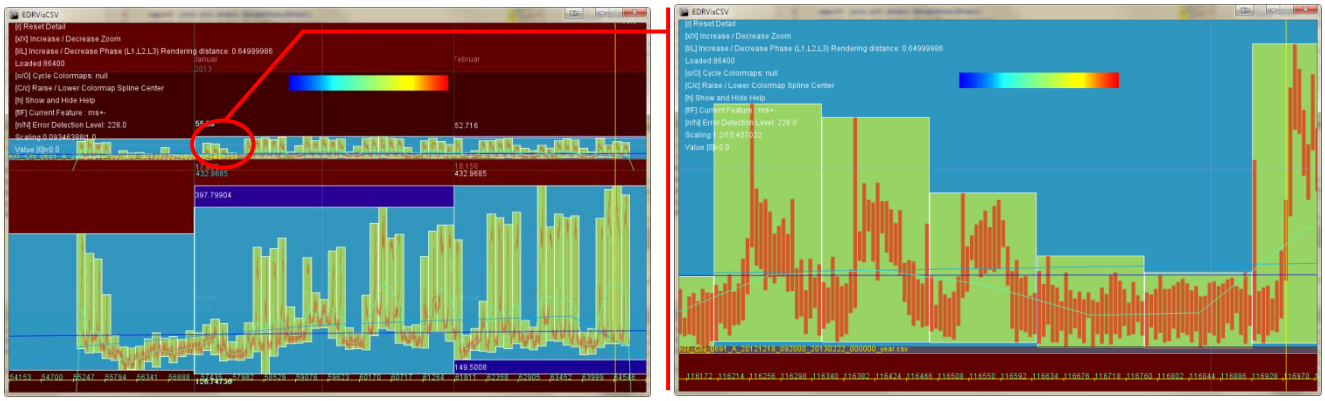
Another domain for the basic statistical data analysis is the power consumption at the KIT-CN. The data is provided by KIT-TID on a network drive, where a new data file is created every 10 minutes holding the absolute value of each of the 534 electrical smart meters. We developed the multicore tool SMDPT (*S*mart *M*eter *D*ata *P*reprocessing *T*ool) to prepare the data for the further basic statistical analysis with *e*TSAnalyzer (see figure 6-right). Based on a user-defined XML specification, it transforms power consumption time series data to flexible facility oriented data for a given time period. All smart meter data for a building are aggregated, and then the power consumption is cumulated for multiple buildings, which are connected to one power supply line of a substation. The facility-oriented, aggregated power consumption data is further used as a new time series for basic statistical data analysis and as an input for the power flow simulation with GridLAB-D. For the analysis, we calculate the five basic statistical characteristics for the power consumption according to the user defined building grouping at the highest level of detail, and we aggregate the data hierarchically over the time units (hour, day, month and year). The analysis further delivers metadata for color mapping and for data validity checking. An automatically generated visualization script simplifies the metadata visualization process in terms of a project file.

4.3.5. Visualization of KIT-CN Smart Meter Metadata with *e*MetaVis

Figure 8-left shows an interactive visualization session of metadata extracted from power consumption data for two buildings at the KIT CN for the period from 18/12/2012 to 21/02/2013. The office building (top) has significant lower power consumption than the testing plant (bottom). Stand-by

phases and work days can clearly be recognized. Even the turning of the year 2012 to 2013 can be identified in the metadata for both buildings. Because of the festive period the power consumption is reduced and only the base load is visible. After zooming into the visualization (see figure 8-right), the average office working hours, with peaks between 8 a.m. - 6 p.m., can purely be identified.

Figure 8. Visualization of power consumption metadata of an office building and of a testing plant (left). Detail view of the first week in 2013 (right). Dark blue boxes show the data range for one year, light blue boxes for each month. Green boxes indicate the days, red boxes the hours.



5. A New Framework for Remote Software Control with HPC Support

The power grid analysis software tools introduced are available as standalone applications. With the embedding of these applications into a general, extendable framework, we enable to access distributed data, to utilize various computing architectures and to remotely execute the applications from any thin client. Although commercial systems are available, the development of an own framework allows maximum flexibility for further research activities.

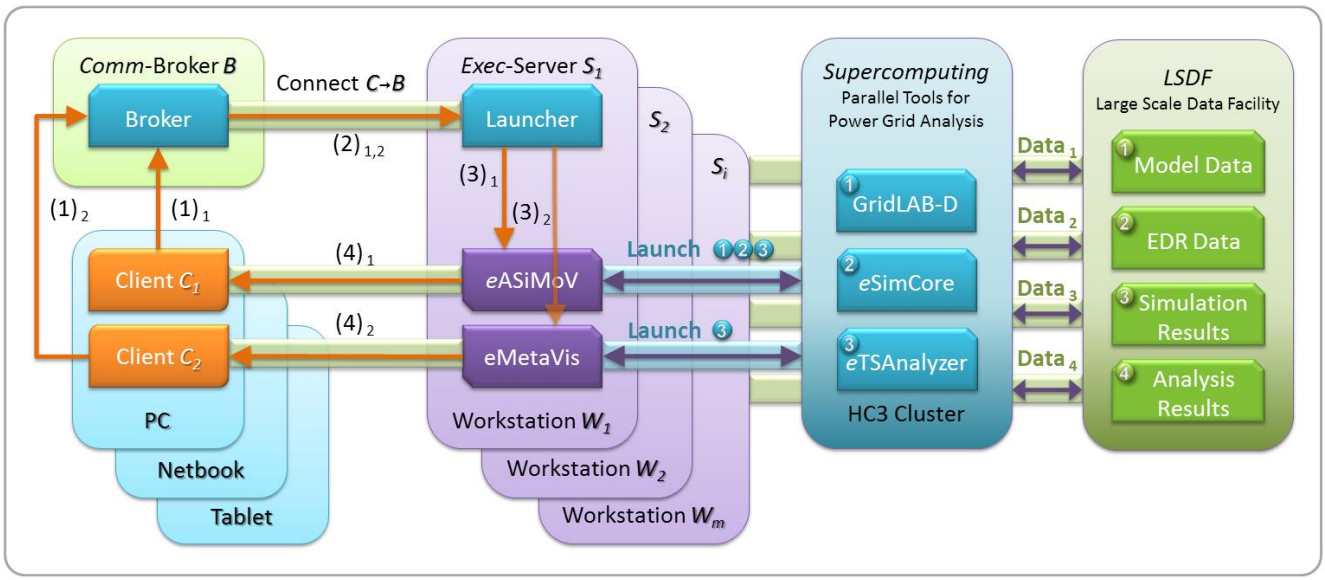
5.1. BReSoC: Broker based Remote Software Control

We initiated the development of BReSoC (*B*roker based *Re*mote *So*ftware *C*ontrol), a computing environment with remote software control support that enables to access data on the LSDF and to use computer clusters as the HC3 (HP XC3000) at the SCC for power grid simulation. We provide in-house developed client software for mobile devices and for Microsoft Windows based PCs.

The concept of BReSoC is shown in figure 9, where the communication path for establishing a direct client-server connection is marked with $(\text{step})_{\text{Client-Id}}$. A client C_1 connects to a dedicated communication broker *B* and requests for a server application $(1)_1$, e.g. *e*ASiMoV for modeling, *e*MetaVis for visualization or *e*TSAnalyzer for basic statistical data analysis, etc. The broker *B*

communicates $(2)_1$ with all available execution servers S_i and selects an available workstation W_m for the client. The remotely executed application, which is launched $(3)_1$ on the workstation, contacts client C_1 and establishes a direct communication and data line $(4)_1$. For this call-back the IP and port number of the client C_1 needs to be passed to the server application along the initial communication path $(1)_1$ - $(2)_1$ - $(3)_1$, which is discarded after the client-server connection is established. The client and the remotely executed application communicate directly, using a simple protocol composed of a header block (control tokens, data descriptive information) and the data block (user input data, compressed data). For the remote software control the client's task is to register any user input, to transfer the interaction to the remote application, to receive real-time image streams from the display of the remote application, and to perform basic 2d pixel write operations into its local canvas for visualization.

Figure 9. BReSoC: A flexible and scalable concept of virtualized software execution, control and remote visualization with support for the computing cluster HC3 and LSDF.



The Windows based applications for modeling (*e*ASiMoV) and visualization (*e*MetaVis) run directly on the workstations, the data analysis software (*e*TSAnalyzer) and the simulation kernels (GridLAB-D, *e*SimCore) can be executed on the workstation, but also be outsourced to the computing cluster HC3 (see figure 9). In a first step we made it possible to link *e*ASiMoV with HC3 in order to carry out simulations with the power grid simulation software GridLAB-D and our power grid simulation kernel *e*SimCore. An *ssh*-based direct connection enables to transfer a power grid simulation model from *e*ASiMoV to the cluster HC3 in order to perform the simulation. The simulation results are being sent back as a compressed archive file. The simulation results are visualized directly in *e*ASiMoV and a screenshot is transferred to the connected client if remote software control mode is activated. The basic statistical analysis software *e*TSAnalyzer for energy data is currently being ported to the

computing cluster. We are also working on the efficient incorporation of the LSDF into the framework, to provide an efficient management of modeling, simulation and analysis data.

5.2. Implementation of BReSoC for Remote Metadata Visualization

We implemented the software for the communication broker and the execution server as well as the client software for the Windows and the Android platforms. Furthermore we extended the software tools for modeling, visualization and analysis by the necessary modules for fast network communication. For testing purposes a low-cost office PC was used as communication broker, a high-end graphics workstation (Intel XEON Quad-Core with 3 GHz, 16 GB RAM, NVidia GeForce GTX 285 and two NVidia Tesla C1060) as execution server, and a netbook as well as a smartphone as the thin clients. For the development of the client software for the latter, the Android SDK was used together with the Android Development Tools (ADT) plugin for the Eclipse IDE. An *ImageView* instance in the Android client software is updated with incoming image streams from *e*MetaVis, the user interaction is captured from the device touch screen interface and transferred to the remotely executed application *e*MetaVis, which in turn applies the interaction to the visualization process and sends new image data. Figure 10 shows the Android client app for the remote software control and visualization of the hierarchical EDR feature metadata. The transferred image has a resolution of 800 x 600 pixels. With the PNG image compression we provide a sufficient good image quality for remote visualization with interactive frame rates about 12 fps.

Figure 10. The Android based client app for remote visualization and control of the *e*MetaVis software, which runs on an execution server in the BReSoC environment.

6. Conclusions

Since 2012 we are conducting energy data recording with in-house developed EDR devices at the KIT-CN. These acquired time series data are serving as the basis for further research. The developed tools for big data management allow for long-term storage of high rate, low voltage supply data on the large scale data facility (LSDF). The introduced analysis and visualization software tools for specialized data intensive computing architectures, as the Hadoop system, enable interactive data exploration and the visual detection of patterns and anomalies. A new statistical data analysis tool for multicore architectures was presented, which generates hierarchically structured metadata in the time domain for various data modalities, e.g. EDR feature data. Using this tool we could detect missing and faulty data ranges in the recorded large time series data and demonstrate the quality improvement of the EDR data acquisition in two years. Furthermore, we introduced the GIS-based modeling software *e*ASiMoV with a support for creating interactive power grid models based on the GridLAB-D and the CIM IEC specification. Various visualization tools allow for direct feature data visualization as well as for hierarchical metadata exploration. A new software framework enables the utilization of various computing infrastructures with a remote software control and a remote visualization support for the presented tools.

In the future a widespread power grid data recording with the introduced EDR devices will deliver valuable information for further in-depth data analysis, simulation model parameter identification and system state estimation. The recorded energy data will be made partially accessible for academic research groups for further data mining. Here a demanding challenge will be the data fusion and the correlation analysis of energy data from additional data sources with varying modalities. For an improved fast data access an optimized multi-level storage method on various file systems will be developed: long term and high rate measurement data in raw format will be stored on the GPFS, feature data derived from raw data will be stored on the HDFS of the Hadoop cluster and metadata will be stored on a fast local server. The modeling of very large scale power grids is feasible with automated data processing methods for data gained by crowdsourcing or publicly accessible statistical databases. We will concentrate on the partial automation of the power grid network modeling, transient analysis and power transmission simulation.

A future development could allow researchers to use the introduced system in terms of a web based grid computing service providing interactive modeling, simulation via supercomputing, Hadoop-based big data analysis and simplified data access, retrieval and storage. A customized web interface with reduced and adapted functionality could serve as an electrical information system e.g. for decision makers. We believe that the developed data management and computing infrastructure can easily be adapted to other domains where big data is a central issue.

7. References

1. Depablos, J.; Centeno, V.; Phadke, A.G., Ingram, M. Comparative testing of synchronized phasor measurement units. In *IEEE Power Engineering Society General Meeting* 2004 (1), 948-954.
2. The Smart Grid: Hadoop at the Tennessee Valley Authority (TVA) – Cloudera Developer Blog. http://www.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/ (accessed on 20.12.2013).
3. Stifter, M.; Bletterie, B.; Burnier, D.; Brunner, H.; Abart, A. Analysis environment for low voltage networks. In *Smart Grid Modeling and Simulation (SGMS)*, 2011 IEEE First International Workshop on; IEEE, 2011; pp. 61–66.
4. Sendin, A.; Berganza, I.; Arzuaga, A.; Osorio, X.; Urrutia, I.; Angueira, P. Enhanced Operation of Electricity Distribution Grids Through Smart Metering PLC Network Monitoring, Analysis and Grid Conditioning. *Energies* **2013**, *6*, 539–556.
5. Grid Protection Alliance. http://www.gridprotectionalliance.org (accessed on 20.12.2013).
6. iPDC - Free Phasor Data Concentrator. http://ipdc.codeplex.com (accessed on 20.12.2013).
7. Albert, A.; Rajagopal, R. Smart Meter Driven Segmentation: What Your Consumption Says About You. *Power Systems, IEEE Transactions on* **2013**, *28*, 4019–4030.
8. De Silva, D.; Yu, X.; Alahakoon, D.; Holmes, G. A Data Mining Framework for Electricity Consumption Analysis From Meter Data. *IEEE Transactions on Industrial Informatics* **2011**, *7*, 399–407.
9. Keim, D. A. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on* **2002**, *8*, 1–8.
10. Zhao, K.; Liu, B.; Tirpak, T. M.; Xiao, W. A visual data mining framework for convenient identification of useful knowledge. In *Data Mining, Fifth IEEE International Conference on*; IEEE, 2005; p. 8–pp.
11. Janetzko, H.; Stoffel, F.; Mittelstädt, S.; Keim, D. A. Anomaly detection for visual analytics of power consumption data. *Computers & Graphics* **2014**, *38*, 27–37.
12. Patterson, J. Lumberyard: Time Series Indexing at Scale. *OSCON 2011 - O'Reilly Conferences, July 25 - 29, 2011, Portland, OR*. http://cdn.oreillystatic.com/en/assets/1/event/61/Lumberyard_%20Time%20Series%20Indexing%20at%20Scale%20Presentation%202.pptx (accessed on 20.12.2013).
13. Shieh, J.; Keogh, E. iSAX: disk-aware mining and indexing of massive time series datasets. In *Data Mining and Knowledge Discovery*, 2009, 19(1), 24–57.
14. JMotif – A time series data-mining toolkit based on SAX and TFIDF statistics. http://code.google.com/p/jmotif (accessed on 20.12.2013).

15. Keogh, E.; Lin, J.; Fu, A. Hot sax: Efficiently finding the most unusual time series subsequence. In *Fifth IEEE International Conference on Data Mining*, 2005, 8–pp.
16. Wei, L.; Kumar, N.; Lolla, V. N.; Keogh, E. J.; Lonardi, S.; Ratanamahatana, C. (Ann) Assumption-Free Anomaly Detection in Time Series. In *SSDBM*, 2005, Vol. 5, 237–242.
17. Hao, M. C.; Marwah, M.; Janetzko, H.; Dayal, U.; Keim, D. A.; Patnaik, D.; Ramakrishnan, N.; Sharma, R. K. Visual exploration of frequent patterns in multivariate time series. *Information Visualization* **2012**, *11*, 71–83.
18. Wei, L.; Keogh, E.; Xi, X.; Lonardi, S. Integrating Lite-Weight but Ubiquitous Data Mining into GUI Operating Systems. *J. Univers. Comput. Sci.* 2005, 11, 1820–1834.
19. Power Systems Analysis Software – Open Electrical. http://www.openelectrical.org/wiki/index.php?title=Power_Systems_Analysis_Software (accessed on 20.12.2013).
20. NEPLAN® Power System Analysis and Engineering, Zurich, Erlenbach, Switzerland. http://www.neplan.ch (accessed on 20.12.2013).
21. Bica, D.; Moldovan, C.; Muji, M. Power engineering education using NEPLAN software. In *43rd International Universities Power Engineering Conference (UPEC)*, 2008, 1–3.
22. Electrical Power Systems Simulation – SimPowerSystems – Simulink. http://www.mathworks.com/products/simpower (accessed on 20.12.2013).
23. Foltin, M.; Ernek, M. Model of the Slovak Power System using SimPowerSystems. http://dsp.vscht.cz/konference_matlab/MATLAB07/prispevky/foltin_ernek/foltin_ernek.pdf (accessed 20.12.2013).
24. ElektraSoft – ELAPLAN. http://www.elektrasoft.de/elaplan_sys.html (accessed on 20.12.2013).
25. Moffet, M.; Sirois, F. Review of Open Source Code for Power Grid Simulation Tools for Long Term Parametric Simulation. *CanmetENERGY, TR–2011-137 (RP-TEC) 411-MODSIM*, 2011.
26. Linear Technology – Design Simulation and Device Models: SPICE simulator LTSpice IV. http://www.linear.com/designtools/software/#LTspice (accessed on 20.12.2013).
27. Yang, J.; Li, Z.; Cai, Y.; Zhou, Q. PowerRush: efficient transient simulation for power grid analysis. In *Proceedings of the International Conference on Computer-Aided Design*, 2012, 653-659.
28. Y.-M. Lee; C.-P. Chen. The power grid transient simulation in linear time based on 3-D alternating-direction-implicit method. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 2003, (22), 1545–1550.
29. EPRI – Simulation Tool OpenDSS. http://smartgrid.epri.com/SimulationTool.aspx (accessed on 20.12.2013).

30. Montenegro, D.; Hernandez, M.; Ramos, G. A. Real time OpenDSS framework for distribution systems simulation and analysis. In *Transmission and Distribution: Latin America Conference and Exposition (T&D-LA)*, 2012 Sixth IEEE/PES, 2012, 1–5.
31. Milano, F. PSAT - Matlab-based Power System Analysis Toolbox. http://faraday1.ucd.ie/psat.html (accessed on 20.12.2013).
32. Milano, F. An Open Source Power System Analysis Toolbox. *IEEE Transactions on Power Systems*, 2005, 20(3), 1199–1206.
33. Chassin, D. P.; Schneider, K.; Gerkensmeyer, C. Gridlab-d: An open-source power systems modeling and simulation environment. *Transmission and Distribution Conference and Exposition*, T&D. IEEE/PES, 2008, 1–5.
34. Maass, H.; Çakmak, H. K.; Suess, W.; Quinte, A.; Jakob, W.; Stucky, K. U.; Kuehnapfel, U. Introducing the Electrical Data Recorder as a new capturing device for power grid analysis. In *IEEE International Workshop on Applied Measurements for Power Systems (AMPS)*, Aachen, 2012, 1–6.
35. Voltage characteristics of electricity supplied by public distribution networks; German version EN 50160:2010 + Cor. 2010.
36. Garcia, A. O.; Bourov, S.; Hammad, A.; van Wezel, J.; Neumair, B.; Streit, A.; Hartmann, V.; Jejkal, T.; Neuberger, P.; Stotzka, R. The Large Scale Data Facility: Data Intensive Computing for Scientific Experiments. In *IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW)*, 2011, 1467–1474.
37. White, T. Hadoop: *The Definitive Guide*, 3rd edition; O'Reilly Media, Sebastopol, CA, 2012.
38. Gates, A. *Programming Pig*; O'Reilly Media, Sebastopol, CA, 2011.
39. Bach, F.; Çakmak, H. K.; Maass, H.; Kuehnapfel, U. Power Grid Time Series Data Analysis with Pig on a Hadoop Cluster Compared to Multi Core Systems. In *21th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP 2013)*; IEEE Computer Society: Los Alamitos, CA, USA, 2013; Vol. 0, 208–212.
40. Holt, R. C.; Schürr, A.; Sim, S. E.; Winter, A. GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, Apr. 2006, 60(2), 149–170.
41. Brandes, U.; Eiglsperger, M.; Herman, I.; Himsolt, M.; Marshall, M. S. GraphML Progress Report - Structural Layer Proposal; In *Graph Drawing*, 2002, 501–512.
42. Uslar, M. *The common information model CIM: IEC 61968/61970 and 62325 - a practical introduction to the CIM*, 1st edition, New York: Springer, 2012.
43. Welcome to CIMTool.org. http://wiki.cimtool.org/index.html (accessed on 20.12.2013).
44. Past CIM Releases. http://cimug.ucaiug.org/CIM%20Releases/Forms/AllItems.aspx (accessed on 20.12.2013).

45. Green, R. C.; Wang, L.; Alam, M. High performance computing for electric power systems: Applications and Trends. In *Power and Energy Society General Meeting, 2011 IEEE*; 2011; 1-8.

KIT Scientific Working Papers
ISSN 2194-1629

www.kit.edu