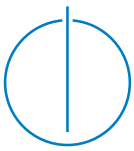
Thomas Kriechbaumer

Methodologies for Distributed Acquisition and Collection of Electrical Energy Data

Technische
Universität
München



Technische Universität München

Fakultät für Informatik

Lehrstuhl für Wirtschaftsinformatik

Methodologies for Distributed Acquisition and Collection of Electrical Energy Data

Thomas Kriechbaumer

Vollständiger Abdruck der von der Fakultät für Informatik der Technische Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr.-Ing. Matthias Althoff

Prüfer der Dissertation:

1. Prof. Dr. Hans-Arno Jacobsen
2. Assoc. Prof. Damian Dalton,
   University College Dublin

Die Dissertation wurde am 29.10.2018 bei der Technische Universität München eingereicht und durch die Fakultät für Informatik am 11.12.2018 angenommen.

It is by logic that we prove, but by intuition that we discover.

– Jules Henri Poincaré

Abstract

One of the breakthrough technologies of the last century is the availability of wide-spread electrical power grids and their pervasive effect on appliances in households, offices, and industry. In recent decades, conserving energy and producing less pollution go hand in hand with renewable energy sources. In order to reduce energy consumption, one has to be aware of consumers and the impact a single appliance can have on the personal electrical energy footprint. Consumption data play a vital role in smart grids and smart homes to understand energy usage characteristics in different environments.

Conventional smart meters offer first insights into consumption patterns, but lack the detailed information about individual appliances and their interaction due to their coarse metering interval. Sampling the voltage and current waveform requires high sampling rates and dedicated data acquisition modules. High-frequency voltage and current data contain valuable information used in power disaggregation and appliance identification tasks – which can help the user to analyze and change the personal consumption profile, or influence operation procedures for industrial machinery and factories.

We develop methodologies for hard- and software architectures for data acquisition, collection, and compression of long-term continuous measurements with a distributed fleet of Internet of Things (IoT) sensors. The integrated approach covers analog signal measurement, embedded systems design, and intelligent data processing in edge- and cloud-computing paradigms. We derive architecture patterns (software blueprints and guidelines) and present two IoT-style data acquisition systems, used for collecting mains (aggregate) and per-appliance waveforms (ground truth), in conjunction with a cloud-based data collection service to provide large-scale storage and analytics capabilities.

Our full-stack architecture patterns have been evaluated by collecting and processing the newly introduced Building-Level Office eNvironment Dataset (BLOND), which contains two long-term measurement series (213 and 50 days) using different components and strategies covered in our design guidelines, a 3-phase mains meter, and up to 90 monitored per-appliance sockets. We provide an in-depth technical validation of the collected data, including sampling rate precision, clock synchronization, and multiple per-file data checks. An appliance log was recorded to label each monitored socket with appliance class, manufacturer, type, and nominal power information.

Zusammenfassung

Eine der wichtigsten Errungenschaften des letzten Jahrhunderts ist die Verfügbarkeit von großflächigen Stromnetzen und deren allgegenwärtige Auswirkung auf elektrische Geräte in Haushalten, Büros, und Fabriken. Die Reduzierung des Stromverbrauchs und der damit verbundenen Umweltverschmutzung sind in den letzten Jahrzehnten direkt an die Einführung von erneuerbaren Energien geknüpft. Um Energieeinsparungen zu erzielen, müssen Kunden und Nutzer über den Energieverbrauch einzelner Geräte informiert werden, mit dem Ziel den persönlichen elektrischen Fußabdruck zu optimieren. Verbrauchsdaten spielen eine wichtige Funktion in intelligenten Stromnetzen und Gebäuden um die Charakteristiken des Energieverbrauchs in verschiedenen Umgebungen zu verstehen.

Übliche intelligente Stromzähler bieten einen ersten Einblick in die Verbrauchsprofile, enthalten aber keine Informationen zu einzelnen Geräten und deren Zusammenspiel aufgrund der groben Messintervalle. Das Messen von Spannung- und Stromwellenformen erfordert hohe Abtastraten und spezialisierte Datenerfassungsmodule. Hochfrequente Messdaten enthalten wertvolle Informationen zur Aufteilung des Energieverbrauchs und Geräteerkennung. Diese Techniken können dem Nutzer bei der Analyse und Anpassung des persönlichen Energieverbrauchsprofils helfen und beeinflussen die Betriebsverfahren von Industrieanlagen und Fabriken.

Wir erarbeiten Methodologien für Hard- und Software Architekturen zur Datenakquise, Sammlung und Kompression von ununterbrochenen Langzeitmessungen mit einer verteilten Flotte von „Internet der Dinge“ (IoT) Sensoren. Der vollintegrierte Ansatz umfasst die analoge Signalmessung, eingebettete Systeme und intelligente Datenverarbeitung in Edge- und Cloud-basierten Paradigmen. Wir definieren Entwurfsmuster (Softwarestruktur und Regelwerk) stellen das Design und zwei IoT-basierten Datenerfassungssystemen vor, welche zur Sammlung von Wellensignalen aus dem Stromnetz (aggregiert) und den Einzelverbrauchern (ground truth) im Zusammenhang mit Cloud-basierten Datensammlungsdiensten verwendet werden um großangelegte Speicherung und Analysefähigkeiten bereitzustellen.

Unsere ganzheitlichen Architekturmuster wurden durch das Sammeln und Verarbeiten von einem neu eingeführten Datensatz evaluiert: Building-Level Office eNvironment Dataset (BLOND), welcher zwei Langzeitmessreihen beinhaltet (213 und 50 Tage), verschiedene

Komponenten und Strategien verwendet, die wir in unseren Designvorgaben spezifiziert haben und ein 3-Phase Stromnetz mit bis zu 90 überwachten Einzelgerätesteckdosen vermisst. Wir erstellen eine tief greifende technische Validierung der gesammelten Daten, inklusive Präzision der Abtastrate, Zeitsynchronisierung und mehreren Datenüberprüfungen pro Datei. Ein Geräteprotokoll wurde erstellt um die Basisinformation wie Geräteklasse, Hersteller, Typ und nominale Leistung der überwachten Steckdose zu erfassen.

Acknowledgments

This dissertation and the pursuit of my doctoral degree took place at the Department of Informatics of the Technische Universät München under the supervision of Prof. Dr. Hans-Arno Jacobsen.

First, I want to thank Prof. Dr. Hans-Arno Jacobsen for accepting me as one of his doctoral candidates, his continuous guidance and support, and providing a fruitful environment that allowed me to conduct my research. He offered great freedom in all aspects of my research and work.

I would like to thank the rest of my thesis committee: Prof. Dr. Damian Dalton for agreeing to be the second examiner, and Prof. Dr.-Ing. Matthias Althoff for being chair of the committee.

I am indebted to my colleagues at the chair and university for their valuable feedback and input. Matthias Kahl, Anwar Ul Haq, and Daniel Jorde for being an instrumental part of our NILM research group. Christoph Doblander for his in-depth expertise and practical advice. Michael Böhmer for his guidance and support in manufacturing various hardware components. Special thanks to Christoph Goebel for starting our NILM research group and acquiring the necessary funding that made my doctoral degree possible. Huge thanks to all my colleagues for not only providing technical input and discussions, but also for the truly great friendship we have formed over the last years.

Last but not least, I want to thank my parents Johann and Renate, as well as my sister Carina for enduring my endless speeches and monologues and for always believing in my dreams and my perseverance.

This research was partially funded by the Alexander von Humboldt Foundation established by the government of the Federal Republic of Germany and was supported by the Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag, and by the German Research Foundation (DFG) and the Technical University of Munich (TUM) within the funding programme Open Access Publishing.

Contents

Abstract **v**

Zusammenfassung **vii**

Acknowledgments **ix**

1 Introduction **1**
1.1 Motivation . . . 3
1.2 Problem Statement . . . 4
1.3 Approach . . . 5
1.3.1 Hardware and Software Architectures . . . 6
1.3.2 Collection and Validation of the BLOND Datasets . . . 6
1.3.3 Waveform Signal and Compression Study . . . 7
1.4 Contributions . . . 8
1.5 Organization . . . 11

2 Background **13**

3 Related Work **17**
3.1 Data Acquisition Systems . . . 17
3.2 Long-Term Continuous Electrical Energy Datasets . . . 19
3.3 Data Compression and Encoding in Datasets . . . 21

4 Mobile Energy Data Acquisition Laboratory **25**
4.1 Hardware Architecture . . . 26
4.1.1 Sensor Board . . . 27

4.1.2 Sampler Board . . . 29
4.2 Software Architecture . . . 30
4.2.1 Microcontroller Firmware . . . 30
4.2.2 Energy DAQ Software . . . 31
4.3 Evaluation . . . 31
4.3.1 High Sampling Rate . . . 32
4.3.2 Resolution and Accuracy . . . 32
4.3.3 Long-Term Recordings . . . 33
4.3.4 Measurement Range . . . 34
4.3.5 Storage Requirements . . . 34

5 Electrical Energy Data Collection Architecture 39
5.1 Design Goals and Requirements . . . 40
5.2 Data Acquisition . . . 42
5.2.1 Analog Data Acquisition . . . 42
5.2.2 Digital Data Acquisition . . . 46
5.3 Data Processing . . . 47
5.3.1 Single-Board Computer Measurement Governor . . . 49
5.3.2 Data Processing Modes . . . 51
5.3.3 Data Collection for Long-Term Continuous Measurements . . . 53
5.3.4 Time Synchronization within a Fleet of DAQ Units . . . 53
5.4 Pull-based Data Processing . . . 56
5.5 Push-based Data Processing . . . 59
5.6 Scalability and Resilience . . . 60
5.7 Evaluation . . . 61
5.7.1 Scalability and Deployment Strategies . . . 65
5.7.2 Scientific Workload: Event Detection . . . 65
5.7.3 Event Detection with Deep Neural Networks . . . 66
5.7.4 Event Detection with k-NN . . . 67
5.8 Discussion . . . 67

6 Building-level Office eNvironment Dataset 69
6.1 Background & Summary . . . 70
6.2 Methods . . . 72
6.2.1 Environment . . . 73

6.2.2 Aggregated Mains Measurements . . . 74
6.2.3 Individual Appliance Measurements . . . 77
6.2.4 Appliance Logs . . . 78
6.2.5 Data Collection . . . 78
6.2.6 Known Issues . . . 80
6.2.7 Code Availability . . . 80
6.3 Data Records . . . 81
6.3.1 BLOND Datasets . . . 81
6.3.2 Appliance Log . . . 82
6.3.3 1-second Data Summary . . . 83
6.4 Technical Validation . . . 83
6.4.1 Data Collection Sanity Checks . . . 84
6.4.2 Sampling Rate Precision . . . 86
6.4.3 Clock Synchronization . . . 86
6.4.4 Per-File Data Checks . . . 88
6.5 Usage Notes . . . 89

7 Waveform Signal Entropy and Compression Study 93
7.1 Evaluated Datasets . . . 96
7.2 Entropy Analysis . . . 98
7.3 Data Representation . . . 100
7.4 Chunk Size Impact . . . 103
7.5 Experimental Results . . . 104
7.5.1 Entropy Analysis . . . 104
7.5.2 Data Representation . . . 105
7.5.3 Chunk Size Impact . . . 108
7.5.4 Summary and Recommendations . . . 109

8 Conclusions 113
8.1 Summary . . . 113
8.2 Future Work . . . 115

List of Figures 121

List of Tables 125

Bibliography **127**

1

Introduction

Electrical energy metering (EEM) has experienced an influx of research activity in recent years due to the shift from mechanical to electronic metering technology. Metering devices used for measuring electrical energy consumption (EEC) and billing consumers are subjected to increased scrutiny over accuracy and reliability. The migration to fully digital EEM is often motivated by potential energy savings and higher comfort levels for occupants. EEC profiles can be generated in smaller time intervals (daily, hourly, by minute), since smart meters allow automated meter readings. Recent studies into the psychological effects of EEM feedback have shown that saving energy and actively managing one's EEC requires frequent feedback over long periods, ideally with an appliance-specific breakdown [1]. However, this requires a significant investment in metering hardware, infrastructure, and reliable communication channels to collect the data from a fleet of smaller meters.

Traditional energy measurement fails to provide support to consumers to make intelligent decisions to save energy, due to the coarse reading intervals (monthly or yearly). Intrusive load monitoring can be used for individual appliance metering by equipping each socket or circuit with a non-intelligent meter. This allows the user to gain insights into the EEC with high granularity, at the cost of additional complexity and costs for hardware, installation, and maintenance. Non-intrusive load monitoring (NILM) attempts to solve this by relying on single-point EEM, ideally utilizing an existing smart meter, to provide

a disaggregated view of the whole-building EEC [2] (power consumption profiles for each appliance). Researchers make use of public datasets to study the characteristics of appliances and to build models representing load profiles and per-appliance usage. This can be beneficial for energy reduction [3, 4], pattern recognition [5, 6, 7, 8], energy demand forecasting [9], and similar fields of study. Modern machine learning approaches (deep neural nets and reinforcement learning) are limited by the available feature space and sampling rate of existing datasets.

In this work, we address the need for methodologies for acquisition and collection of electrical energy data in a distributed fleet of energy sensors. First, we develop a low-cost measurement architecture for high-frequency energy data. This data acquisition unit can be used in a smart meter scenario (aggregated mains circuits) or to collect individual appliance EEC profiles (ground truth per-appliance). Second, we present a distributed edge- and cloud-based data collection and processing architecture, specifically for long-term continuous dataset collection. The software stack utilizes processing power of the measurement units (network edge) to perform data processing tasks, or makes use of cloud-based resources. A large fleet of sensors (and their compute platforms) can be monitored and together form a measurement network with two data collection strategies. Third, we introduce the Building-Level Office eNvironment Dataset (BLOND), a new long-term continuous EEC dataset. The novel feature of the two sub-datasets is the availability of waveform ground truth data for individual appliance EEC. This type of data source was previously unavailable, as existing public datasets only provided low-frequency root-mean-square (RMS) values of voltage and current signals. Fourth, we present a novel study of five whole-building energy datasets with high sampling rates, their signal entropy, and how a well-calibrated measurement can have a significant effect on the overall storage requirements. We show that existing datasets do not fully utilize the available measurement precision, leaving potential accuracy and space savings untapped. We benchmark a comprehensive list of 365 file formats, transparent data transformations, and lossless compression algorithms.

1.1 Motivation

Electronic smart meters are installed by a majority of developed countries, or plans are currently formed to replace conventional electromechanical energy meters [10]. Utility companies want to reduce costs for meter readings and billing, by relying on remote meter reading and telemetry capabilities. The integrated processing and communication devices can be used to acquire readings in shorter intervals, compared to the conventional manual monthly or yearly reading.

Hart [2] first introduced the NILM approach in the 1980s: to measure the aggregate mains voltage and current at the entry point into the building and extract information about individual appliances to help the homeowner or user make informed decisions about their devices. This metering requires higher sampling rates than what was previously available, although recent approaches in this area show promising results [11].

The basic NILM approach did not gain traction due to limited processing power and sensor accuracy of the available technology. In recent years, NILM received significant attention due to the advances in statistics, machine learning, and processing capabilities [6, 12, 13, 14, 15, 16]. Energy saving and smart scheduling of appliance usage are the main promises of NILM-based systems and can be targeted at different stakeholders: consumer, neighborhood, or utility companies. Smart homes and Internet of Things (IoT) further motivated the need for power disaggregation and appliance identification based on non-intrusive EEM [17, 18, 19].

Understanding ones energy consumption can be done through a data analysis of voltage, current, power, and other electrical metrics. NILM, together with electronic smart meters, can further augment a monthly EEC bill by supplying a disaggregated view of the total consumption. Individual appliances can be identified from the live mains signals to directly notify the user or log the activation and runtime period. Isolating appliance transients (start and stop events) can be used for predictive maintenance and fault monitoring [20].

1.2 Problem Statement

Most NILM tasks greatly benefit from high sampling rates of voltage and current signals to capture the sinusoidal waveform. Typical smart meter datasets only report one sample per second (or less), which lacks all high-frequency signal information of the underlying waveform. Although most NILM tasks rely heavily on such waveform measurement data, publicly accessible datasets are still rare and biased towards household environments. A meaningful waveform signal reconstruction requires sampling rates in the Kilo-Hertz range [21]. Recent studies have shown that high sampling rates improve various NILM-related tasks [5, 15]. Differences in power grid systems around the world include: 120 V vs. 230 V, 3-phase vs. 2-phase grids, 50 Hz vs. 60 Hz mains frequencies, and phase shift between the voltage legs. These differences make it difficult to compare and use datasets from Europe and North America for machine learning pipelines, because the underlying waveform data does not match.

Publicly available datasets, such as REDD [22], BLUED [23], and UK-DALE [24], provide voltage and current signals of aggregate mains signals. However, they lack individual appliance signals (ground truth) with similar waveform data, and contain only low-frequency ground truth data (1 sample per second or less). Supervised machine learning requires a fully-labeled ground truth to find patterns in the data during training. Without ground truth waveforms, the exact timestamp of appliance transients and events can only be estimated based on the sampling rate.

The above mentioned datasets were designed for household environments and their typical appliance types, such as fridges, hair dryers, and washing machines. Most modern devices use switched-mode power supplies (SMPS) to transform the mains AC into DC power. These appliances do not show typical pure-resistive characteristics, but contain short pulses of current draw at very high frequencies, which requires high sampling rates to accurately measure the current waveform signal. Office environments mostly contain SMPS-driven appliances, such as computers, battery chargers, monitors, and networking equipment. Collecting the ground truth waveform of such devices requires high sampling rates.

Smart plugs with integrated measurement sensors, as used in existing datasets, can

only supply data at low sampling rates due to their limited processing power and network connectivity. Data acquisition in office environments with multiple rooms at high sampling rates requires a distributed fleet of measurement units to collect voltage and current signals for each appliance, perform local data processing, and access to back-end infrastructure for future analysis.

The sampling rate and number of channels directly correlates to the required storage space in a long-term continuous dataset. Collecting and processing raw data consumes persistent storage, network bandwidth, and CPU resources depending on the bit depth and sampling rate. Lossless compression, well-suited file formats, and encoding schemes can help to reduce the overhead and overall resource requirements to analyze a given dataset. Recent datasets [24, 25] consist of close to 1 million individual files, taking 39 TB of storage space in its compressed form.

1.3 Approach

In this work, we leverage IoT design methodologies and combine them with requirements from energy metering systems. We generalize fully-integrated hardware and software architectures for EEM for multi-circuit aggregated mains and individual appliance ground truth signals. First, we consider the problem of distributed data acquisition in the context of NILM. We describe data collection strategies and transfer methods to reliably measure, process, and store voltage and current waveform data. Second, we apply these procedures to collect a novel long-term continuous dataset with high sampling rates of ground truth waveforms combined with a fully-labeled appliance log for machine learning tasks in power disaggregation and appliance identification. Third, we address the challenge of encoding, storing, and compressing such large-scale datasets based on common design requirements for big data processing systems.

1.3.1 Hardware and Software Architectures

In this approach, we present the physical requirements and the resulting design for a data acquisition system to meter individual appliances (socket-level). The implemented system was built and evaluated: MEDAL, a mobile energy data acquisition laboratory. Our work utilizes an off-the-shelf power strip with a voltage-sensing circuit, current sensors, and a single-board PC as data aggregator. We develop a new architecture and evaluate the system in real-world environments. The self-contained unit for six monitored outlets can achieve up to 50 kSps for all signals simultaneously. A modular design and off-the-shelf components allow us to keep costs low. Equipping a building with our measurement systems is more feasible compared to expensive existing solutions.

We formalized the underlying architecture patterns (design guidelines) and adapted them to fit a electronic smart meter methodology as well as the individual appliance meter use case. We introduced the Circuit-Level Appliance Radar (CLEAR) system [26] based on these principles: a platform to collect voltage and current waveform signals of a 3-phase power grid with up to 250 kSps per channel simultaneously for long-term continuous data collection.

We introduce a fully-integrated hardware and software architecture to collect and process raw measurements. Our system can operate in various modes to collect, store, and stream real-time measurement data. We abstract design rules and strategies to allow for multiple use cases: collection of long-term continuous datasets, real-time appliance monitoring, or ground truth data collection. The core components are designed to be tolerant and self-recovering in case of failures or network congestions.

1.3.2 Collection and Validation of the BLOND Datasets

In this approach, we present the BLOND datasets: BLOND-50 and BLOND-250. The datasets are characterized by a strong focus on office appliance and SMPS-driven power supplies. We outline the measurement environment, data acquisition and collection architecture, and waveform ground truth data stream. The previously introduced hardware and software architecture, including the MEDAL and CLEAR measurement units

are used to collect and store the long-term continuous signal streams – 16 independent measurement units operating as a distributed fleet of sensors.

The collected data was augmented with an appliance log, to gather metadata about the running appliances on each monitored socket. The manufacturer, type, and nominal power of each device was logged and periodically updated to maintain a valid mapping between monitored socket and connected appliance.

The BLOND-50 dataset contains 213 days of continuous data with 50 kSps and 6.4 kSps sampling rates. The BLOND-250 dataset contains 50 days of uninterrupted data with 250 kSps and 50 kSps sampling rates. In total, we recorded and measured 16 appliance classes, 53 different types, and 74 individual appliances.

1.3.3 Waveform Signal and Compression Study

Electrical energy datasets (suitable for NILM tasks) are covering longer time spans, contain data with higher sampling rates, consist of multiple measurement units, and grow in total storage size. While first datasets, such as REDD [22] and BLUED [23] started the era of high-frequency long-term datasets within the NILM community, the signal quality and calibration was not a primary concern until recently. With datasets exceeding tens of terabytes, we found that storage space and data transfer rates become a challenge when processing these datasets.

First, we evaluated five long-term continuous datasets based on the contained signal entropy. We generated the histogram of each signal channel based on the raw measurement values within the dataset. We found that similar measurement setups do not yield the same type of entropy distribution in the published data. The main cause was non-optimal signal calibration and non-matching analog front-ends (voltage probes and current transformers).

Second, we defined a set of requirements for file format and encoding scheme to be used in electrical energy datasets. The primary goal is to reduce the overall dataset size while maintaining an easy-to-use file format and access API. General-purpose datasets

should only be compressed with a lossless algorithm, due to the unknown nature of analytics and feature extraction systems the NILM community might apply to the data. Using a lossy compression masks or softens important characteristics necessary for NILM tasks. We converted and re-encoded the whole datasets in 365 different file formats, data representations, and compression algorithms. The resulting space saving and compression ratio was recorded to compare each combination. We show that with careful selection of file format and encoding scheme, we can reduce the size of some datasets by up to 73%.

1.4 Contributions

The contributions made in this work affect three main topics: new hardware and software architectures for distributed acquisition and collection of electrical energy data, the newly introduced Building-Level Office eNvironment Dataset (BLOND), and a comprehensive waveform signal entropy and compression study of whole-building energy datasets. In the following we present our contributions for each topic:

The main contributions to distributed electrical energy data acquisition systems and collection and processing architectures are:

i. We design the hardware platform for a new data acquisition system, capable of collecting mains electricity data for voltage and current signals. The proposed methodology and architecture was implemented and evaluated in the MEDAL and CLEAR hardware.

ii. We propose a novel data acquisition topology, which provides a fully integrated system from analog measurements up to data processing in a data center. Our systems design combines edge-based processing with cloud-based data collection pipelines.

iii. We extend the topology with real-time data processing at the edge of the proposed measurement network. We evaluate the feasibility of online event detection with deep neural networks at the ground truth data acquisition units.

iv. We conduct extensive evaluations based on the requirements for long-term data collection, as used in the BLOND datasets. We define dataset requirements and

characteristics important to the fields of power disaggregation and appliance identification.

The main contributions of the BLOND datasets are:

i. We published two datasets with electrical energy data: voltage and current waveforms of a 3-phase mains circuit feeding an office building, with individual appliance ground truth measurements.

ii. We provide long-term continuous time series data: BLOND-50 with 213 days; BLOND-250 with 50 days of uninterrupted measurements.

iii. We provide raw waveform data with high sampling rates for both the aggregate and the per-appliance measurements: BLOND-50 uses 50 kSps and 6.4 kSps; and BLOND-250 uses 250 kSps and 50 kSps.

iv. We conduct an extensive technical validation of the collected data: sanity checks while collecting the data, per-file data checks for signal characteristics, sampling rate precision analysis, and clock synchronization verification.

The main contributions to waveform signal entropy and compression study of whole-building energy datasets are:

i. We provide a comprehensive evaluation of electrical energy datasets with regard to their utilized and available measurement precision. We discuss signal calibration and potential space savings that can result from a carefully selected signal conditioning stage.

ii. We compare an extensive list of 365 file formats, transparent data transformations, and lossless compression algorithms. We propose best-practices to reduce the overall dataset size while maintaining an easy-to-use file format and access API.

iii. We analyze the impact of different chunk sizes with regard to the achievable compression ratio and space savings. We show the correlation between compression parameters and the resulting size reduction over the whole data for each of the examined dataset.

Parts of the content and contributions of this work have been published in:

- T. Kriechbaumer, A. U. Haq, M. Kahl, and H.-A. Jacobsen. "MEDAL: A Cost-Effective High-Frequency Energy Data Acquisition System for Electrical Appliances." In: *Proceedings of the 2017 ACM Eighth International Conference on Future Energy Systems*. e-Energy '17. Hong Kong, Hong Kong: ACM, 2017. ISBN: 978-1-4503-5036-5. DOI: 10.1145/3077839.3077844
- T. Kriechbaumer, M. Kahl, D. Jorde, A. U. Haq, and H.-A. Jacobsen. "Large-Scale Data Acquisition Systems Architecture for High-Frequency Electrical Energy Metering." Submitted to ACM Trans. Cyber-Phys. Syst. 2019
- T. Kriechbaumer and H.-A. Jacobsen. "BLOND, a building-level office environment dataset of typical electrical appliances." In: *Scientific Data, an open-access NatureResearch journal* 5.180048 (2018). DOI: 10.1038/sdata.2018.48
- T. Kriechbaumer and H.-A. Jacobsen. *Waveform Signal Entropy and Compression Study of Whole-Building Energy Datasets*. 2018. arXiv: 1810.10887

Related work and additional contributions have been published in:

- M. Kahl, C. Goebel, A. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "NoFaRe: A Non-Intrusive Facility Resource Monitoring System." In: *Energy Informatics*. Vol. 9424. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 59–68. DOI: 10.1007/978-3-319-25876-8_6
- A. U. Haq, T. Kriechbaumer, M. Kahl, and H.-A. Jacobsen. "CLEAR – A Circuit Level Electric Appliance Radar for the Electric Cabinet." In: *2017 IEEE International Conference on Industrial Technology*. ICIT '17. Toronto, Canada, 2017, pp. 1130–1135. ISBN: 978-1-5090-5319-3. DOI: 10.1109/ICIT.2017.7915521
- M. Kahl, T. Kriechbaumer, A. U. Haq, and H.-A. Jacobsen. "Appliance Classification Across Multiple High Frequency Energy Datasets." In: *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm): Smart metering, Demand Response and Dynamic Pricing (SGC2017 Smart Metering)*. 2017. DOI: 10.1109/smartgridcomm.2017.8340664

- M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "A Comprehensive Feature Study for Appliance Recognition on High Frequency Energy Data." In: *Proceedings of the 2017 ACM Eighth International Conference on Future Energy Systems*. e-Energy '17. Hong Kong, Hong Kong: ACM, 2017. ISBN: 978-1-4503-5036-5. DOI: 10.1145/3077839.3077845

- D. Jorde, T. Kriechbaumer, and H.-A. Jacobsen. "Electrical Appliance Classification using Deep Convolutional Neural Networks on High Frequency Current Measurements." In: *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm) (IEEE SmartGridComm'18)*. Aalborg, Denmark, 2018

- M. Kahl, V. Krause, R. Hackenberg, A. U. Haq, A. Horn, H.-A. Jacobsen, T. Kriechbaumer, M. Petzenhauser, M. Shamonin, and A. Udalzow. "Measurement System and Dataset for In-Depth Analysis of Appliance Energy Consumption in Industrial Environment." In: *tm - Technisches Messen* (2018). DOI: 10.1515/teme-2018-0038

1.5 Organization

The rest of this work is organized as follows: Chapter 1 provides an introduction to the problem and motivation for this thesis. Chapter 2 provides background information on fundamentals in power grids, electrical parameters, and measurement values. In Chapter 3, we discuss the related work in the field of NILM and how our work fits into the defined problem.

Chapter 4 presents the data acquisition system MEDAL, which was used to collect ground truth energy consumption data. We define requirements for NILM-focused data collection and integration into a fleet of distributed sensors. The hardware system with analog measurements, data acquisition, and edge-based data processing is described and evaluated for signal integrity and signal quality metrics.

Chapter 5 elaborates on the underlying design methodologies used for the design and

requirements of MEDAL, CLEAR, and the subsequent collection of BLOND. Edge- and cloud-based data processing paradigm are defined and evaluated with two collection strategies utilizing compute resources of the measurement units if available. Monitoring and best-effort workflows are evaluated based on the two BLOND sub-datasets.

Chapter 6 covers the BLOND datasets, its measurement environment, the fleet of MEDAL and CLEAR sensors, and the technical validation of the collected data. We have collected and published two measurement time series: BLOND-50 and BLOND-250 with different sampling rates and time durations. We contribute an office-focused environment with typically SMPS-driven appliances. The most novel feature of the BLOND sub-datasets is their waveform ground truth with high sampling rates (collected with MEDAL units) and the appliance log to label each monitored socket to an appliance class and manufacturer information.

Chapter 7 evaluates a comprehensive list of file formats, compression algorithms and their parameter space. The input data are the five largest long-term continuous datasets, including our BLOND sub-datasets. The results show the best-performing compressors based on the compression ratio and space saving for the whole dataset. A comparison of contained entropy in the raw signal of each dataset was performed to extract reasons for high or low compression ratios, caused by signal conditioning and data formatting.

In Chapter 8, we present unified conclusions about the present topics; specifically about the four main contributions regarding hard- and software architectures, datasets, and their compression.

2

Background

Smart metering of electrical energy consumption creates a constant data stream of not only the measurements, but of metadata, events, and other types of information related to the consumption characteristics of the building under observation (BUO). Meters can be placed at the utility supply point (where the actual meter for billing purposes is placed), or on various locations throughout the BUO, depending on the needs of the customer. Applying the NILM methodology results in a single meter. One can also utilize one meter per floor or wing, to increase granularity. For industrial applications, the facility manager would benefit from an even finer resolution, down to individual assembly lines or manufacturing stages. For residential homes, NILM prevents installing individual meters for each appliance, and therefore reduces costs and complexity. Smart Home technology, such as smart plugs and switchable power outlets, can be used to cost-effectively monitor the electricity consumption of a group of appliances.

Appliance identification provides the user with information about transient events (switch-on and switch-off) and the corresponding appliance class, type, or model. Identification accuracy strongly depends on high sampling rates and features extracted from waveform data [5, 15, 31, 34]. Event detection systems monitor the raw waveform data, and extract a region of interest, which can be used for the classification task [35, 36].

Power disaggregation defines the task of splitting the time series data of a single-point

EEC measurement into the fundamental components (individual appliances). Various techniques have been proposed to generate a per-appliance energy bill or consumption metric [6, 20, 37, 38, 39]. Most of these approaches need to keep track of active (switched on) and inactive (turned off) appliances to correctly attribute portions of the energy to each device.

Given a suitable NILM-capable measurement device, the final metering architecture will consist of a small number of data acquisition systems, and possibly a communication network to aggregate the data streams. However, for the purpose of collecting new datasets, designed for providing valuable source material for research and machine learning approaches, the number of total DAQ units will be vastly higher than any final installation for day-to-day measurements.

Electrical energy has two physical signals that can be measured: voltage and current. Multiple other metrics can be derived solely based on these two signals. Power, energy, harmonics, power factor, and other relevant metrics are defined by voltage and current fluctuations over time. The most commonly used parameters are root-mean-square (RMS) voltage (*Volt*) and current (*Ampere*), together with instantaneous power (1 *Watt* = 1 *Volt* $*$ 1 *Ampere*).

The RMS for a repeating waveform signal S with N samples, is defined as $S_{RMS} = \sqrt{1/N \sum_{i=1}^{N} S[i]^2}$. Such metrics can be given for a single period (sinusoidal mains waveform), or longer segments (1 s, 1 min, 1 h). For longer time frames, the RMS has an "averaging" effect on sub-period deviations from the pure-sine signal. Power grids with alternating current (AC) most commonly use mains voltages of 120 V_{rms} or 230 V_{rms} with a mains frequency of 60 Hz or 50 Hz, respectively. Therefore, the signal waveform peaks at ±169.7 V or ±325.3 V, and repeats itself in a sine-pattern. The fundamental frequency is one of the most important health metrics in a power grid. Even small deviations can cause a total collapse of the delicate supply-demand balance. Harmonics are higher-order signals and typically strongly pronounced at multiples of the mains frequency. The amplitude and distribution of these harmonics also give insights into the state of a power grid. While the voltage follows a sinusoidal pattern, the current can significantly deviate from it, to the point of single bursts of current draw at each voltage peak.

Liang, Ng, Kendall, and Cheng [34] categorized energy data based on the sampling rate: micro level with more than 1 sample per cycle; macro level with less than 1 sample per cycle. Similarly, we define low-frequency in the macro level, and high-frequency in the micro level. Low-frequency measurements typically provide RMS values of voltage, current, or power measurements. High-frequency measurements provide the raw waveform signal for voltage and current. Sampling rate is described as samples per second (Sps) or as frequency in Hertz (Hz). The value unit can optionally be given with an SI-prefix, e.g., 12.5 kSps or 12.5 kHz, 0.0125 MSps or 0.0125 MHz. For low-frequency signals, a coarser unit can be used to define the sampling rate, such as 1 min (equal to $1/60$ Hz), or 1 h (equal to $1/3600$ Hz).

The data acquisition must always happen in the high-frequency domain in order to calculate RMS and other time-averaged metrics, such as phase shift, harmonics, or distortion. Therefore, all electrical energy measurement systems capture the signal waveform at a specific sampling rate, and then either report the raw values (high data bandwidth), or perform the necessary calculations for a predefined list of metrics in an embedded system before reporting these low-frequency values.

For data collection strategies (long-term continuous datasets), a "capture everything" approach means that every measurement will be forwarded to a storage system for long-term persistence. Smart meters and intelligent measurement systems aimed at consumer-level consumption reporting do not require long-term storage facilities, which means in-memory processing can be used to generate high-level metrics (RMS values or total energy consumption per day). The type of metric defines the sampling rate requirement. According to the Nyquist–Shannon sampling theorem [21], the sampling rate needs to be twice the signal frequency one wants to reconstruct. Measuring and reporting the n^{th}-order harmonic of the 50/60 Hz mains frequency requires a sampling rate of at least $2n \cdot 50\,Hz$. While this is the theoretical minimum sampling rate, actual implementations typically use a 5-8× higher sampling rate. The measurement bandwidth, as commonly defined for oscilloscopes, characterizes the required sampling rate to measure an analog sine wave with an amplitude error of 3% [40]. Recording multiple channels with high sampling rates requires oscilloscopes or specialized data acquisition systems as presented in [26, 27, 33, 41]. Most signals are composed of multiple complex waveforms and require a higher sampling rate to cover the entire desired frequency domain [42].

Compressed sensing is a technique to reconstruct a signal by utilizing prior knowledge about signal characteristics. While the Nyquist-Shannon limit is commonly known and refers to a fixed sampling rate, Candès, Romberg, and Tao [43] proved that even fewer samples are sufficient, assuming the signal's sparsity is known. The newly introduced technique achieves a signal reconstruction with a sparse (variable) sampling rate. This optimization in data acquisition can be used for consumer-level intelligent measurement systems, but is not applicable in general-purpose DAQ workflows used for scientific datasets, because the underlying prior knowledge cannot be guaranteed.

3

Related Work

In this chapter, we present related work in the field of electrical energy metering and data acquisition systems used for collection and processing of long-term continuous datasets. We then cover these datasets, their requirements, design guidelines, and measurement environments. NILM-related tasks, such as power disaggregation and appliance identification, are analyzed and examined with regard to the available data and metadata in the publicly available datasets. Finally, we list the related work in the area of data compression for waveform signals, low-frequency measurements, and appliance transient events.

3.1 Data Acquisition Systems

Research in the field of NILM is primarily based on public datasets to analyze, evaluate, and compare new algorithms and approaches [22, 23, 24, 44, 45, 46]. Multiple research groups have created and published datasets, which can be categorized into low- and high-frequency measurements [37, 47]. In recent years, high-frequency measurements have become more popular in NILM, due to technological advances in the field of electronics, while still requiring specialized and expensive hardware (costing thousands of Euros).

Correct identification of appliance transients and load signatures is a key step for power disaggregation and appliance identification [15, 34]. Different appliance types can have discriminating features hidden in the electricity waveforms. The extraction of these features typically requires high sampling rates (Kilo-Hertz range) [14]. High-frequency voltage and current data carry a vast amount of information depending on the appliance type. Such information has been collected and visualized to compare different appliance types on various feature metrics [44, 45, 46].

A high-quality dataset for household energy consumption was provided in [24]. The mains signal was provided with a sampling rate of 16 kHz. Appliance-level data (ground truth) are available in 6 s intervals, which is a significantly lower rate not suitable for reconstructing the mains waveforms. This poses problems in the correct matching of appliance transients (violation of the switch continuity principle) [12], a common concern with many existing datasets [8, 22, 23].

Sensing voltage and current with a custom circuit board allows for a compact recording device and provides configurable parameters for rated current and mains frequency [18]. However, this approach still requires an external analog-to-digital converter and is only capable of monitoring a single appliance. Combining multiple monitored power outlets into a single unit would improve cost efficiency, reduce complexity of analog-to-digital conversion, and allow for easier data collection and storage.

Controlling the turn-on and off time instants based on the voltage zero-crossings was introduced as measurement device in [41]. Using triacs to control the phase angle at which an appliance transient occurs allows for a detailed introspection into angle-correlated power consumption measurements. However, this system is neither portable (no integrated processing unit) nor capable of recording data for a prolonged amount of time (months to years).

There are two common current sensing techniques: using a shunt resistor to measure the voltage drop (with a current transformer), or using the Hall-effect of a current carrying conductor. Both approaches require precise knowledge about the measurement architecture. A new approach allows the measurement of current using non-contact field measurements without the need for precise sensor placement [48]. This has the benefit of being non-intrusive, since the circuit under measurement is uninterrupted and the

sensor can be placed on the outside of the power cable.

Lights, refrigerators, printers, and other consumer electronics are not strict resistive loads and generate different patterns. Using an oscilloscope, it is possible to visualize these patterns and analyze the data digitally [49]. However, most oscilloscopes are only capable of displaying values with 8- or 10-bit resolutions for a low number of parallel signals. Oscilloscopes have limited processing power and storage capacities, which is a key requirement for long-term continuous recordings.

Household energy consumption data are an important source for smart grid solutions. Using new forecasting algorithms and disaggregated energy data, these predictions can be of great value to grid operators and expansion plans. Generating energy demand forecasts for numerous households (thousands) is feasible on commodity hardware [17] and can be accurate to a 15 min time window [9].

3.2 Long-Term Continuous Electrical Energy Datasets

Existing electrical energy consumption datasets have been targeted at use cases involving load forecasting, demand response, and non-intrusive load monitoring. For these purposes, the energy research community shifted to high sampling rates and long-term continuous measurement series in recent years. The following datasets have inspired the design presented in this work and the resulting *Building-Level Office eNvironment Dataset* (BLOND [25]): *The Reference Energy Disaggregation Data Set* (REDD [22]), *Building-Level fUlly-labeled dataset for Electricity Disaggregation* (BLUED [23]), *UK Domestic Appliance-Level Electricity dataset* (UK-DALE [24]). Energy disaggregation and load forecasting have been researched extensively with datasets provided by energy utilities and their Smart Meter measurement architectures for remote meter reading. The resulting data is exclusively available with low sampling rates, typically in the range of one aggregate measurement per minute or hour. Such data does not match the requirements and use cases of most non-intrusive load monitoring and appliance identification tasks, which work best with the raw waveform signal [5, 34].

REDD contains a total of 119 days of data with a sampling rate of 15 kHz for one voltage

and two current signals in multiple residential buildings. The ground truth data was collected at 0.5 Hz and 1 Hz intervals depending on the circuit and appliance type. The raw high-frequency data was reduced with a lossy compression before publishing the dataset. Although the authors claim to present "signal waveforms", the lossy data reduction does not allow a meaningful waveform reconstruction for any given timeframe. The measurement hardware consists of smart power strips for plug-level data, and a NI-9239 data acquisition modules connected to a laptop to collect waveform data of three signals (voltage and two current waveforms). The current signal was generated with a current transformer by TED (split-core clamp-style) and a TA041 voltage probe by Pico. The individual components are well-characterized by the vendors, however, the REDD measurement architecture lacks proper signal calibration and range levels, leading to lost precision.

BLUED contains a whole week with a sampling rate of 12 kHz for one voltage and two current signals in a single-family home. The mains measurements were collected with current transformers (split-core clamp-style) from TED and a Pico TA041 voltage probe connected to a NI-6251 data acquisition module. The ground truth was collected at roughly 1 Hz (RMS values) with plug-level FireFly sensors (smart plugs). The wireless communication was centralized at a gateway device which aggregated all smart plug data and added a NTP-based timestamp. All data was stored on a local computer. The event labeling and signal synchronization was done by visual inspection.

UK-DALE [24] contains 655 days of data with a sampling rate of 16 kHz for a single mains circuit with voltage and current measurements. The measurement hardware consists of a USB sound card attached to an Atom-based general-purpose computer. The analog data acquisition is handled by a simple circuit utilizing an AC-AC step-down transformer with a voltage divider and signal diodes to generate a voltage signal. The current signal is generated by a current transformer and an attached shunt resistor protected by the same signal diode arrangement. The resulting analog signals are AUX-compatible voltage levels and are fed directly into a USB sound card (a single stereo channel). The diodes in both signal paths cause a clipping of the waveform to protect the AUX input. Unfortunately, if the mains voltage exceeds the calibrated signal, the measurements will be skewed, causing erroneous voltage readings. The UK-DALE measurement setup does not provide any potentiometers or other forms of hardware calibration. This results in a degraded

measurement performance, which can be observed in some parts of the UK-DALE dataset. The digital data acquisition is handled by a general-purpose computer running a Python-based recording application. The USB sound card provides OS-level APIs to read audio data from the stereo input channel (left and right for voltage and current data). The DAQ pipeline uses isochronous transfers on the USB interface, which do not provide error detection, error correction, or reliable transfers (packet loss detection). If the employed computer is busy with non-DAQ tasks, the sound card (or the USB subsystem) can silently drop data, resulting in sudden signal jumps (non-continuous sinusoidal waveform measurements). This behavior is mostly visible in the voltage data of the dataset. The published dataset contains 16 kHz data which was downsampled from 44.1 kHz, although the sound card is capable of recording in 96 kHz, however, the authors experienced even more data loss at this sampling rate due to the limited computational resources of the host computer. The ground truth data was collected in 6 s intervals with EcoManager smart plugs.

3.3 Data Compression and Encoding in Datasets with High Sampling Rates

Low-frequency energy data can benefit greatly from compression when applied to smart meter data, as multiple recent works have shown [50, 51, 52, 53]. Electricity smart meters can be a source of high data volume with measurement intervals of 1 s, 60 s, 15 min, or higher. Possible transmission and storage savings due to lossless compression have been evaluated in [52]. While the achievable compression ratio increased with smaller sampling intervals, the benefits of compression vanish quickly above 15 min intervals. Various encodings (ASCII- and binary-based) have been evaluated for such low-frequency measurements, and in most cases, a binary encoding greatly outperforms an ASCII-based encoding. The need for smart data compression was discussed in [54], which further motivates in-depth research in this area. The main focus of the authors was smart meter data with low temporal resolution from 10,000 meters or more. Various compression techniques were presented and a fast-streaming differential compression algorithm was evaluated: removing steady-state power measurements ($t_{i+1} - t_i = 0$) can save on average 62% of required storage space.

High-frequency energy data offers a significantly larger potential for lossless compression, due to the inherent repeating waveform signal. Tariq, Arshad, and Nabeel [55] utilized general-purpose compressors, such as LZMA and bzip2, and achieved good compression ratios on some datasets. Applying differential compression and omitting timestamps can yield size reductions of up to 98% on smart grid data, however, these results are not comparable as there is no generalized uniform data source. The presented results use a single data channel and an ASCII-based data representation as a baseline for their comparison, which contains an inherent encoding overhead. The SURF file format [56] was designed to store NILM datasets and provide an API to create and modify such files. The internal structure is based on wave-audio and augments it with new types of metadata chunks. To the best of our knowledge, the SURF file format didn't gain any traction due to its lack of support in common scientific computing frameworks. The recently published EMD-DF file format [57], by the same authors, relies on the same wave-audio encoding, while extending it with more metadata and annotations. Neither SURF nor EMD-DF provides any built-in support for compression. The power grid community defined the PQDIF [58] (for power quality and quantity monitoring) and COMTRADE [59] (for transient data in power systems) file formats. Both specifications outline a structured view of numerical data in the context of energy measurements. Raw measurements are augmented with precomputed evaluations (statistical metrics), which can cause a significant overhead in required storage space. While PQDIF supports a simple LZ compression, COMTRADE does not offer such capabilities. To the best of our knowledge, these file formats never gained traction outside the power grid operations community.

Lossy compression can achieve multiple magnitudes higher compression ratios than lossless, with minimal loss of accuracy for certain use cases [53]. Using piecewise polynomial regression, the authors achieved good compression ratios on three existing smart grid scenarios. The compressed parametrical representation was stored in a relational database system. However, this approach only applies if the use case and expected data transformation is known before applying a lossy data reduction. A 2-dimensional representation for power quality data was proposed in [36] and [60], which then could be used to employ compression approaches from image processing and other related fields. While both approaches can be categorized as lossy compression due to their numerical approximation using wavelets or trigonometric functions, they require a

specialized encoder and decoder which is not readily available in scientific computing frameworks.

The NilmDB project [13] provides a generalized user interface to access, query, and analyze large time-series datasets in the context of power quality diagnostics and NILM. A distributed architecture and a custom storage format were employed to work efficiently with "big data". The underlying data persistence is organized hierarchically in the filesystem and utilizes tree-based structures to reduce storage overhead. This internal data representation is capable of handling multiple streams and non-uniform data rates but lacks support for data compression or more efficient coding schemes. NILMTK [61], an open-source NILM toolkit, provides an evaluation workbench for power disaggregation and uses the HDF5 [62] file format with a custom metadata structure. Most available public datasets require a specialized converter to import them into a NILMTK-usable file format. While the documentation states that a zlib data compression is applied, some converters currently use bzip2 or Blosc [63].

4

Mobile Energy Data Acquisition Laboratory

Electrical appliances are pervasive around the world. Due to an increased number of household and office appliances, it is infeasible to attach energy meters to each device. Lately, non-intrusive load monitoring (NILM) has gained wide popularity due to its single point sensing and user-friendly deployment. Appliance-level load profiles allow the user to make intelligent decisions in order to reduce the overall energy usage.

Usually, smart plugs are utilized to collect appliance-level measurements as ground truth for larger datasets [17, 22, 23, 24]. Unfortunately, most of these low-cost smart plugs support only low-frequency sampling rates and limited measurement capabilities [38, 64]. This makes it difficult to reconstruct the actual current and voltage waveforms, which can improve tracking of appliance on/off events with higher precision [15].

We proposes a new data acquisition system (DAQ) for energy monitoring and ground truth collection: MEDAL – a Mobile Energy Data Acquisition Laboratory. Our system can capture data continuously, while extracting start-up and switch-off transients for feature extraction, appliance identification, and fault detection. The design is based on an off-the-shelf power strip, a single-board PC for data collection, a voltage-sensing circuit, and current sensors to measure mains electricity signals of up to six appliances. The

proposed hardware was implemented and actual devices were manufactured to evaluate the design based on the the requirements in Table 4.0.1.

Table 4.0.1: Requirements for energy data acquistion hardware.

R1 high sampling rates	**R6** synchronized world clock
R2 long-term recordings	**R7** precise time-stamping
R3 common file format	**R8** high resolution and accuracy
R4 data compression	**R9** retain measurement data
R5 low price per appliance	**R10** resources for data processing

This chapter is structured as follows: We introduce a new hardware architecture in Section 4.1, and propose the software system in Section 4.2. Finally, we present results and design evaluations in Section 4.3.

4.1 Hardware Architecture

MEDAL is designed as mobile cyber-physical system: hardware, software, and algorithms working together in a single platform. The data collection is deeply integrated with an analytics pipeline, either on the system itself, or on a remote server for post-processing (public or private cloud). The measurement system is compact enough to be mobile, while offering access to multiple communication networks. Two sensor types are integrated to fully capture energy consumption data of appliances. This section gives an in-depth description of individual architectural components and outlines the reasoning for specific design decisions. The collection of new datasets can be managed using a fleet of MEDAL units together with a central component to orchestrate large-scale energy measurements. The proposed high-level architecture can be seen in Figure 4.1.1.

The design of physical components is based on an off-the-shelf six-port power strip. The power strip is used as foundation and center piece, with all components designed around it. The current design and all prototypes make use of the *Schuko* socket/plug system used in most EU countries.

The architecture for the DAQ system distinguishes between two main components,

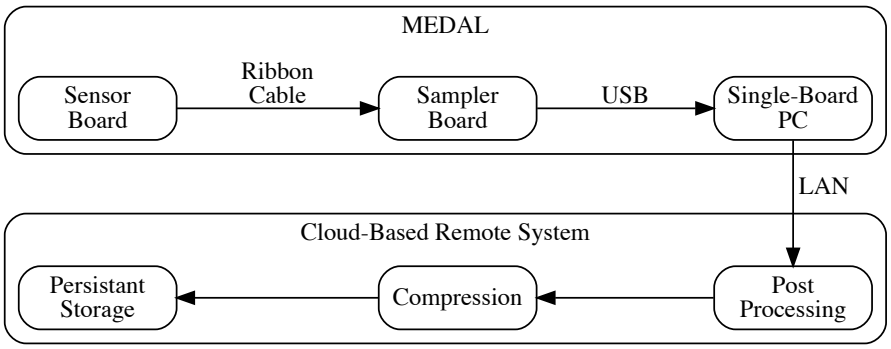


Figure 4.1.1: The architecture of the measurement unit shows the data flow through multiple stages.

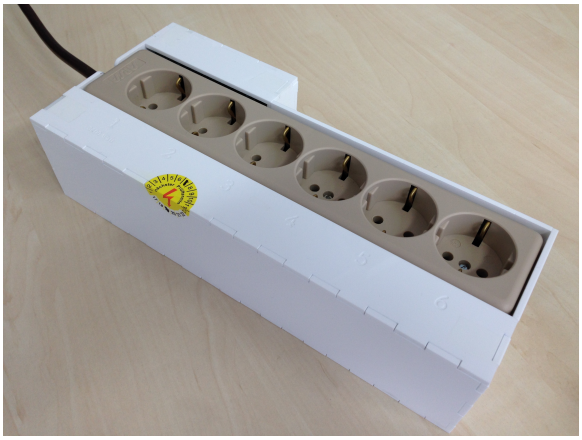
which are designed to fit together into a single assembly: sensor board and sampler board. These are custom-made PCBs with analog signals and digital control circuitry. The data acquisition is handled by a simple hardware module attached via USB. Voltage and current signals are generated with an independent circuit board to isolate high- and low-voltage domains. The main control unit is a single-board PC connected to the local network via ethernet or WiFi.

The six-port power strip is implanted within a custom laser-cut enclosure (white acrylic glass sheets) to protect and shield the electronics and to meet the physical safety requirements. A fully assembled MEDAL unit is compact, portable, and easily installable (**R5**, Figure 4.1.2a). In order to interconnect with the current sensors on the sensor board, the power bus bars are rewired and routed through the sensor, see Figure 4.1.2b.

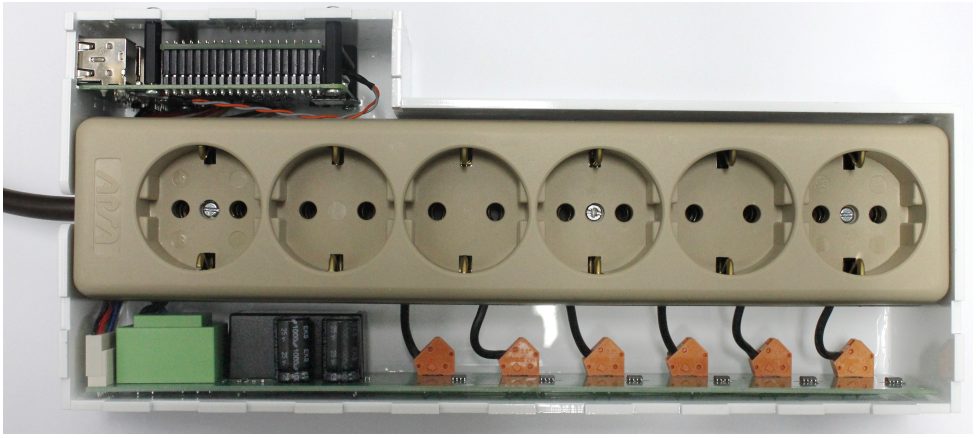
4.1.1 Sensor Board

The sensor board is the only high-voltage component in the system. It contains current sensors and circuitry for voltage sensing. The external power is used for a common 5 VDC power rail used by all sensors, ICs, and the single-board PC.

The voltage signal is generated by an AC-AC transformer acting as galvanic isolator and step-down converter. It is necessary to add a DC-offset to produce a strictly positive signal for the unipolar ADCs. The final signal has a mean of 2.5 V, with peaks at 1.27 V and 3.39 V.



(a) The MEDAL system used to collect ground truth appliance energy consumption data. The laser-cut acrylic enclosure contains a power strip and two boards to measure the voltage and current of each connected appliance.



(b) Without the protective covers, the exposed sensor board at the bottom, and the single-board PC with sampler board at the top-left.

Figure 4.1.2: The physical components of a MEDAL measurement unit.

Six independent current signals (one for each socket) are generated by Hall effect-based sensors. The system can be equipped with ICs calibrated to different sensitivity settings, allowing sensing of currents from 5, 20, and 30 A_{peak}. The selected sensor chip (Allegro ACS712), yields 2.5 V if no current is flowing, and ranges from 0.5 V to 4.5 V.

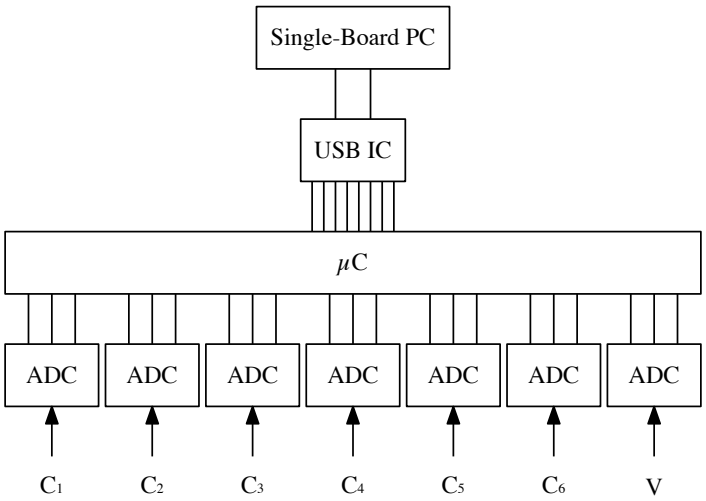


Figure 4.1.3: The data flow on the sensor board. All ADCs are connected via a single data bus to the microcontroller.

4.1.2 Sampler Board

The sampler board converts the analog sensor output into digital data and transfers them via USB to the single-board PC. The main components are: seven MCP3201 chips for analog-to-digital conversion (12-bit, unipolar, 0 V to 4.095 V), an ATmega324PA microcontroller for the control logic, an FTDI232H for USB data transfers, and a Raspberry Pi 3 for processing, buffering, and network connectivity, see Figure 4.1.3.

The ADCs treat each signal independently, while the digital output of all converters are grouped into an 8-bit bus directly attached to the microcontroller (7 data bits + padding). The ADC data packets are then transferred to the USB connectivity chip. Recorded packets are received via USB bulk transfers in the Linux kernel.

Multiple MEDAL units can be time-synchronized using the Network Time Protocol (NTP), which provides accuracy within a few milliseconds [65, 66]. Since MEDAL also supports a stand-alone recording mode without access to an accurate clock master, the sampler board contains a dedicated real-time clock (PCF2123) with a super-capacitor acting as backup battery (**R6**).

4.2 Software Architecture

Raw measurements are controlled by a microcontroller, which contains a static firmware (see Section 4.2.1). The sampler board is connected via a single USB socket and can be controlled directly with a specialized software package (see Section 4.2.2).

Based on the recommendations of the NILMTK project [61], we support HDF5 as storage format (**R3**), which is widely used and supports dataset-based compression (**R4**). Depending on the optional scientific workload, MEDAL is capable of converting the data into HDF5 and applying a simple compression scheme before storing the data file (**R10**). Each file carries a list of attributes: timestamp (UNIX time with microsecond precision), sampling rate (Hertz), a measurement unit identifier (UUID), and a sequence number (chunked data files). Every signal stream (voltage and currents) is stored as an individual dataset with corresponding attributes: calibration factor and removed DC-offset.

4.2.1 Microcontroller Firmware

The microcontroller runs a continuous loop that reads values from the ADCs and forwards data to the USB interface chip. Since all ADCs are controlled via the same I/O bus, the data conversion for all channels is triggered at the same time. This corresponds to a *single-shot* ADC with multiple channels. This task could be managed by the single-board PC; however, since a generic Linux system can fire system interrupts at any time, the required precision can not be guaranteed.

The main loop runs on a interrupt timer based on the main clock frequency with a pre-scaler and a precise crystal oscillator (14.7456 MHz). The sampling rate can be set to any integer value between 225 and 50,000. A slight offset might occur, depending on the pre-scaler and the oscillator frequency and can be determined as follows:

$$t = \lfloor 14745600/f - 1 \rceil$$

$$f_{actual} = 14745600/(t + 1)$$

All ADCs are read in parallel, while reading in the 12 bits sequentially. This means the byte layout is not usable straightaway, and requires a simple bit-wise transformation. This allows us to save additional shift registers or other external data latching to transpose the input vector.

4.2.2 Energy DAQ Software

The single-board PC runs a generic Linux distribution, allowing us to make use of common command and control patterns via USB bulk transfers. All software components run as system service. Configuration can be performed with command line switches or environment variables.

After the initialization phase, a queue of USB bulk transfer requests is filled up. If the USB interface chip on the sampler board has a full buffer, the data is sent to the Linux kernel, which then triggers a callback in the DAQ software. The callback contains a buffer with the sample data. The raw data packet must be reformatted before it can be written into a file. The maximum file size (chunk) can be configured depending on the expected post-processing pipeline and sampling rate. Aligning the file size to 15, 30, or 60 minutes at a specific sampling rate, is commonly preferred.

Operating a MEDAL unit in a stand-alone mode persists all files directly to a mass storage device (**R9**), thus allowing the user to simply download them when needed. This procedure is possible, since all required hard- and software initializations are performed as soon as the unit receives mains power.

4.3 Evaluation

The above described MEDAL system is evaluated based on the defined requirements. Multiple requirements are evaluated with experimental data collections to test the accuracy and overall performance.

We built 20 MEDAL units (designed for the power grid in Europe, 230 V/50 Hz) and have been collecting energy data from more than 50 appliances over the last several months. To accommodate various appliances (different levels of power draw), we configured each MEDAL with a mix of ACS712-05B and ACS712-30A current sensors.

4.3.1 High Sampling Rate

Capturing the signal waveform with very high temporal resolution (**R1**) is necessary to measure harmonics and other high frequency characteristics [15]. The MEDAL system fully supports sampling rates from 225 Hz up to 50 kHz on all channels without data loss, gaps, or buffer overflows. This allows us to capture a 50 Hz mains signal with up to 1000 samples per mains cycle.

4.3.2 Resolution and Accuracy

We chose the MCP3201 as ADC, a successive-approximation ADC offering full 12-bit resolution and a high enough bandwidth to fulfill our target sampling rate of 50 kHz due to its wide availability and simple integration.

The 12-bit vertical resolution yields a measurement step sizes of 5.4 mA_{peak} (ACS712-05B), 15.15 mA_{peak} (ACS712-30A), and 0.25 V for the mains voltage. This gives us a minimal detectable power difference of 0.88 W — less than 0.11% of the 815 W maximum measurement range (**R8**).

Measuring a 5 VDC switched-mode power supply (SMPS) with a 5 W resistive load approaches the limits of MEDAL, see Figure 4.3.1. The signal trace was collected with 50 kHz, resulting in a sample timing precision of 20 µs (**R7**). Initially, no load is attached to the power socket, and the data show the expected low-level channel noise. The sharp transient at 500 ms is caused by internal capacitors being charged up right after plugging in the device. At 1500 ms the SMPS has completed initialization and begins to power the resistive load. The noise energy (16 mA_{rms}) is near the actual load (67 mA_{rms}) compared to the maximum range of 3.5 A_{rms}.

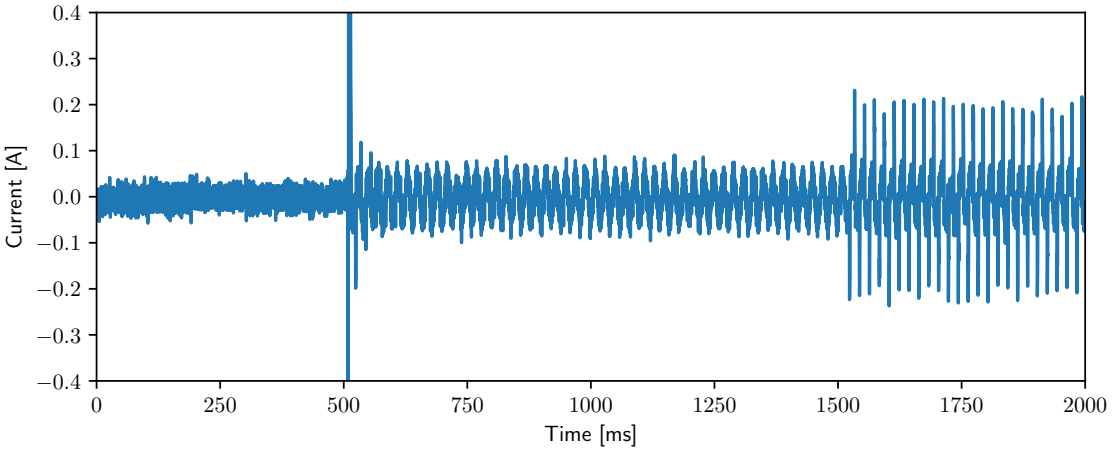


Figure 4.3.1: SMPS with a 5 W resistive load. A sharp transient is triggered by charging up the internal capacitors of the SMPS.

4.3.3 Long-Term Recordings

MEDAL is designed for continuous measurements of energy data without interruptions (**R2**). Given the data are collected and transmitted to a remote storage facility, MEDAL can record indefinitely. We have 15 fully operational MEDAL units in a real-world environment recording data continuously for over 200 days without interruption or data loss (**R9**).

A MEDAL unit was used to record in a typical office environment, with a laptop and monitor connected to the monitored sockets. The mains voltage and frequency changes over a 24-hour period can be seen in Figure 4.3.2a. The current consumption of two appliances in the same time window is depicted in Figures 4.3.2b and 4.3.2c. The recording for the depicted time period consists of 96 files, each 15 min long, with an average file size of 30.2 MB. Root-Mean-Square equivalents of voltage and current are computed over 60 s slices. The laptop is a MacBook Pro 13″ with a 60 W power adapter. The monitor is a 27″ Dell UP2716D with a nominal power of 45 W.

Table 4.3.1: Storage requirements for different sampling rates.

Sampling Rate	CPU Usage	Raw Data for 10 min	HDF5	Ratio
50,000	10.4%	401.0 MB	156.3 MB	38.99%
16,000	3.3%	128.2 MB	48.8 MB	38.07%
4,000	0.9%	32.1 MB	12.3 MB	38.34%
250	0.3%	2.0 MB	0.8 MB	39.91%

4.3.4 Measurement Range

A MEDAL unit can be assembled for different mains environments: 120 or 230/240 V, 50 or 60 Hz mains frequency. In addition, every socket can be equipped with a different current sensor rated for 5, 20, or 30 A_{peak}. This allows us to measure a wide variety of devices: from a small kitchen tool, desktop PCs, to heavy industrial appliances.

Measuring high harmonics allows us to track and monitor the power quality. The measured current and its computed spectrum of a rotary multi-tool can be seen in Figure 4.3.3. A discriminating feature for this appliance is the strong magnitude at 3600 Hz to 3800 Hz likely caused by the brushed motor (commutator bars).

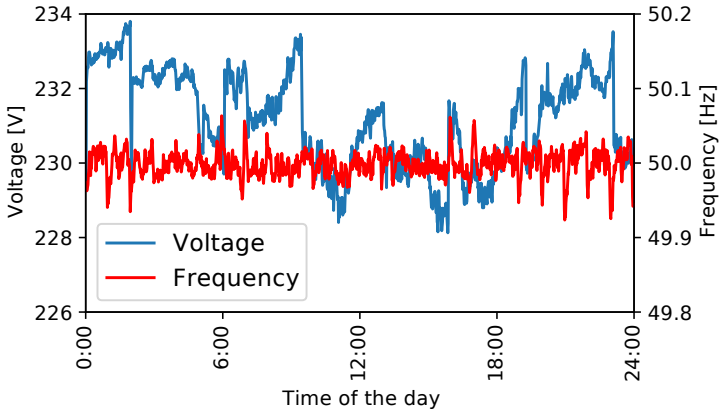
4.3.5 Storage Requirements

Common oscilloscopes and data loggers have to find a trade off between sampling rate, resolution, and storage capacity [49]. MEDAL tries to overcome this limitation by providing fast data acquisition for multiple channels while also capturing and uploading the data stream continuously.

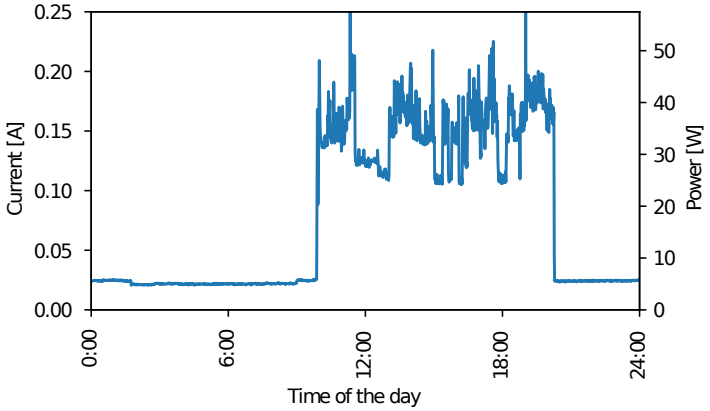
The comparison of five experimental measurement runs with a MEDAL unit and different sampling rates can be seen in Table 4.3.1. The CPU usage clearly shows that the system is running underutilized during normal recording operations and is therefore capable of performing additional scientific analysis in near real-time within the unit (**R10**). Of four available cores, only one is dedicated at all times to the DAQ task. Data conversion and compression is performed in bursts and frees up the CPU as soon as it finishes (at least

two full cores).

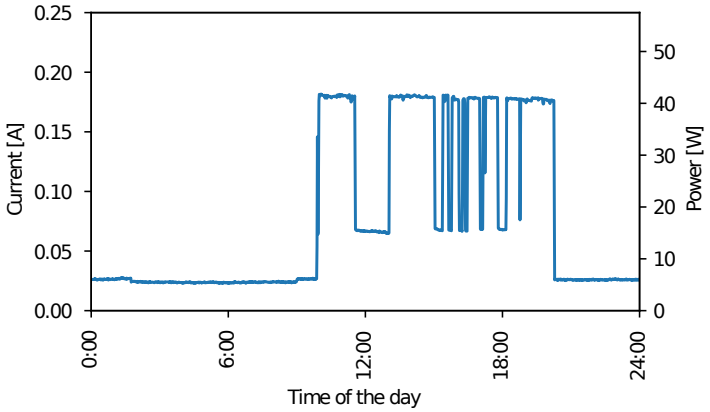
HDF5 provides a well-documented format specification and a good compression by using dataset filters (data manipulation within a file) (**R3**). The HDF5 data in Table 4.3.1 uses the built-in *gzip* and *Fletcher* filters to compress the data (**R4**). More advanced compression techniques might further improve the overall ratio and storage savings.



(a) Voltage and frequency changes.

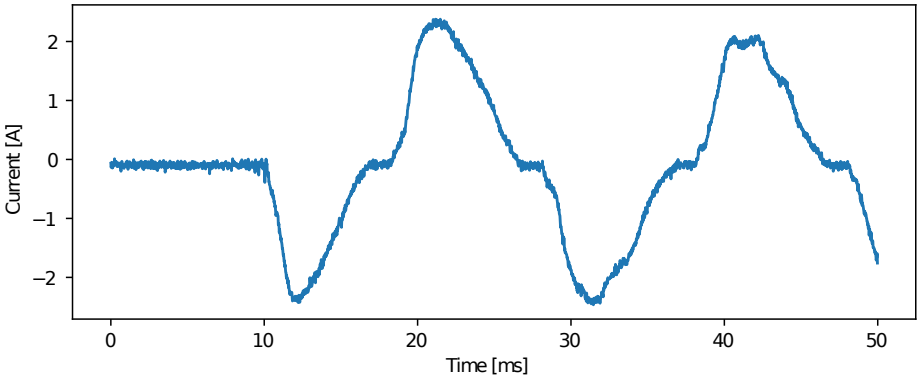


(b) Current consumption of a 13'' laptop.

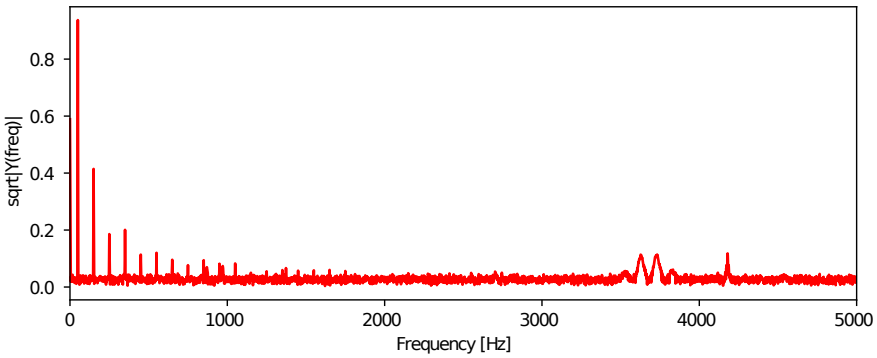


(c) Current consumption of a 27'' monitor.

Figure 4.3.2: A 24-hour recording of two appliances with MEDAL in an office environment.



(a) Rotary multi-tool startup with max. sampling rate. The motor speed control introduces ripples throughout the transients.



(b) Spectrum of a 1s steady-state signal (50 full mains periods). Strong harmonics are visible at 3600 Hz to 3800 Hz.

Figure 4.3.3: Waveform and spectral characteristics of a rotary multi-tool (50 Hz power grid)

5

Electrical Energy Data Collection Architecture

In this chapter, we develop methodologies for electrical energy data collection with purpose-built data acquisition systems, processing pipelines, and collection strategies. The software stack and data flow is closely aligned with the design requirements of CLEAR [26] and MEDAL [27], targeting long-term continuous data collection without data loss or interruptions. Compute resources at the edge of the measurement network are fully integrated and supplement cloud-based infrastructure depending on the operation mode and collection strategy. The BLOND datasets where collected with the methodologies presented in this chapter.

Compared to other datasets with high sampling rates, the novel features of BLOND are twofold: simultaneous 3-phase voltage and current measurements (previous datasets only collected data from a single phase in a 3-phase grid, or a simplified data collection in a 2-phase grid) and a comprehensive ground truth measurement with high sampling rates (previous datasets only collected ground truth data as RMS values for 1 s or 6 s intervals). These unique characteristics of BLOND require a well-conceived data acquisition architecture.

The DAQ systems and BLOND were specifically designed for long-term data collection.

Strong focus was given to resilience against data loss throughout the DAQ data flow chain. From the earliest stage where digital samples are first available, the design guidelines include buffers and timing schedules to guarantee a full data capture without loosing samples. As a final verification step, each file is checked against a rule-based requirements list.

This chapter is structured as follows: We introduce design goals and requirements for long-term energy datasets in Section 5.1. We propose architectural patterns for data acquisition in Section 5.2 and data processing techniques in Section 5.3. Section 5.4 introduces a pull-based data processing scheme, while Section 5.5 describes a push-based strategy for the same task pipeline. Scalability and resilience in the context of long-term continuous data collection is discussed in Section 5.6. Finally, we present an architecture evaluation in Section 5.7 and a discussion of the methodologies in Section 5.8.

5.1 Design Goals and Requirements

Long-term continuous data collection of EEC data with mains waveforms started with the REDD [22] and BLUED [23] datasets. These datasets contain multiple days of uninterrupted data with high sampling rates. Measurement infrastructure for long-term continuous datasets, such as BLOND [25] and UK-DALE [24] are designed for multiple months of uninterrupted measurement series. The design goals and requirements on hardware, software, and infrastructure must be oriented towards never-ending recordings.

We define a set of goals and requirements to achieve long-term continuous data collection with fully-integrated hardware, software, and infrastructure components:

DG1 – Flexible Signal Composition: EEC data collection, with individual appliance meters and multi-phase mains meters, requires a heterogeneous number of signal streams (multiple voltage and current channels) per DAQ system. CLEAR was designed to meter a 3-phase power grid, resulting in 6 simultaneously measurement channels. MEDAL was aimed at a cost-effective ground truth (individual appliance) data collection by providing six monitored power sockets (current channels) and a

shared voltage channel. The data collection pipeline must be capable of processing such data streams independent of the number of signals or their type.

DG2 – Network Connectivity: EEC measurements in office or industrial environment can be physically distributed over multiple rooms and floors within a monitored building. Similarly, collecting data from entire neighborhoods and campus areas requires a distributed fleet of DAQ systems that form a measurement network to collect and process data files. In larger areas, multiple data centers (collection endpoints) can be used to provide additional fail-over domains with automatic load balancing. Local communication links (WiFi or Ethernet network) need to be available to reach a central data collection system. Network channels and compute resources need to be managed to provide optimal uninterrupted service for data collection and processing.

DG3 – Data Quality: While data corruption might be unavoidable, it must be detected as early as possible. Corruption can occur at multiple levels, making files either unreadable, or scrambling individual measurement values. Both types need to be considered during the publishing process of a dataset. The lack of data, due to an outage or malfunction of the DAQ system, needs to be reported as well. In the case of public scientific datasets, the accompanying documentation should list missing, corrupted, or truncated data files. Kelly and Knottenbelt [24] included a dedicated plot with the release of the UK-DALE dataset, which documents up- and downtimes of each measurement system (clearly marked regions of inactive data collection). In addition to the above mentioned sudden signal jumps and discontinuous signals, the dataset also contains 5 partially corrupt files which are not publicly documented. Similarly, BLUED [23] also contains 2 partially corrupt and undocumented files. For the creation of BLOND [25], we provided a comprehensive technical evaluation of all data files and fully documented the data coverage.

DG4 – Collection Reliability: Every outage, interruption, or power surge can result in data loss or corruption. High availability (HA) infrastructure design principles are well-known from data center operations and similar fields. In the case of DAQ systems, one typically has to make trade-offs between redundancy and costs for additional hardware, networking equipment, and storage buffers (similar to the principle of locality). File formats, storage system, and I/O access should also follow general HA principles: error detection or correction on a filesystem level,

fail-over architecture for data retrieval systems, continuous off-site backups, and decoupling data buffers at each communication layer. HA and reliability can be described with the "principles of nines" – the percentage of time a service is fully available in a fixed window (a year, month, or day). *Four nines point five* or 99.995% allows less than 2.2 min per month (on average) of interrupted service (downtime or unavailability).

DG5 – Online Reconfiguration: Data collection networks are evolving over time: adding and removing new measurement systems, adapting the collection strategy, or continuous deployment of new software components must be possible without downtime or service interruption. This allows us to evolve the code base and implement new features while collecting data.

DG6 – Compute Locality: Edge-based compute resources should be used to reduce the necessary infrastructure in the back-end (data centers). Each DAQ system should be capable of performing local data processing tasks, including scientific workloads, such as event detection and appliance identification in real time.

5.2 Data Acquisition

5.2.1 Analog Data Acquisition

The measurement architecture developed for BLOND was implemented in CLEAR and MEDAL and consists of the following main components: an analog-to-digital converter (ADC), an embedded controller, a USB bridge, and a single-board computer (SBC) for the final data processing, see main boxes in Figure 5.2.1. The data flow starts at the signal sensors to condition the measurement signals to be compatible with the ADC input range. The ADC input channels are connected to one analog signal source and can be sampled simultaneously. The digital values (data) are collected by an embedded controller and then delivered to an SBC via a USB connection.

ADCs report an electric potential difference (voltage) as numerical value. The amplitude resolution is defined by the bit depth and the safe voltage input range, yielding 2^n bits

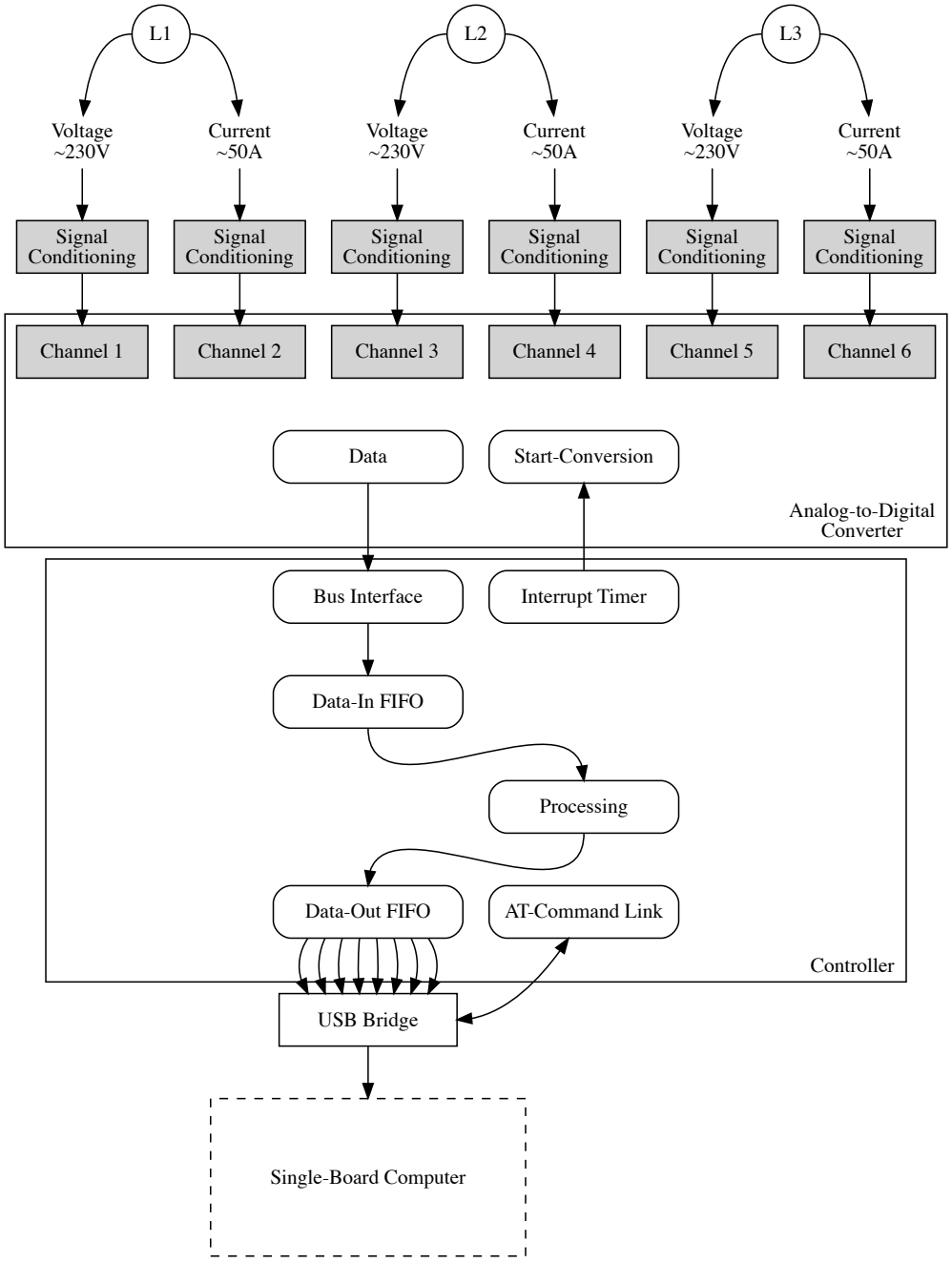


Figure 5.2.1: The data acquisition flow of analog signals from a 3-phase power grid (L1, L2, L3 circuits) with voltage and current sensors into the ADC where the digital values are collected by the controller. The embedded system operates as independent entity and communicates with the SBC via AT-commands.

individually distinguishable integer values, with common resolutions in the range of 12-, 16-, or 24-bit. The signal-to-noise ratio (SNR) of an ADC describes the strength of the desired signal compared to (random) background noise. SNR is typically defined in decibels and uses a logarithmic scale. The performance characteristics of the ADC alone is only of minor importance, because it is strongly affected by the signal conditioning and external circuitry. In order to get a reliably SNR metric, the full DAQ system should be characterized end-to-end.

A unipolar ADC can only report positive voltages: $[0 \dots V_{max}]$, whereas a bipolar ADC spreads the available value range symmetrically around 0: $[-V_{max}/2 \dots 0 \dots + V_{max}/2]$. Common input voltage ranges are up to $\pm 2\ V$, $\pm 5\ V$, or $\pm 10\ V$. The signal gain and expected SNR depend on the input voltage range and need to be considered when designing and selecting components.

Measuring electrical signals typically requires a signal conditioning step, which transforms the raw signal into a suitable (voltage) range, which is compatible with the DAQ hardware. Most systems operate on a low-voltage level (less than 48 V), which could be damaged by directly feeding in mains signals (above 100 V). In the context of this work, we consider the starting point of data acquisition to be the physical sensor for each signal type.

Mains voltage, typically 120 V or 230 V, needs to be conditioned (step-down) to a safe ADC input range. This can be achieved with a simple voltage transformer, or with active differential measurement probes. An additional requirement for voltage sensor selection is the available frequency bandwidth, especially high-frequency signals on the mains voltage might be suppressed or eliminated by improper probing.

If the measurement system provides a conductive path via the mains circuit, the analog signal should be galvanically isolated for safety. This breaks the direct current path between mains and low-voltage elements. Means of isolation can be provided by isolation transformers or optocouplers. An ADC-DAC bridge with optoisolators on the digital signals can achieve the same safety rating, but introduces additional signal noise and unwanted quantization artifacts.

Mains current cannot be directly measured by an ADC and needs to be transformed into a

proportional voltage first. There are two common sensor types: the Hall-effect [67] causes a voltage difference due to the magnetic field of current flowing through a conductor, while a burden resistor simply causes a small voltage drop across a high-precision resistor in series with the mains circuit. A current transformer can step-down the current level if the mains current exceeds the sensor rating. Special care must be taken when using such a transformer, due to its high voltage danger during open-circuit faults. Rogowski coils [68] can only measure AC and require an external integrator to generate a signal proportional to the measured current, but they can be used in an open loop configuration and do not require an iron core.

The ADC conversion (sampling) is triggered by an external signal, which also defines the sampling rate. After receiving the start-conversion trigger, the ADC needs a specified time to perform the sampling, before the digital measurement value can be read. In the case of multi-channel ADCs, either one channel needs to be selected while sending the start-conversation signal, or the ADC provides a single-shot functionality where all channels are sampled simultaneously. Voltage and current signals must always be measured simultaneously, otherwise the time delay between the samples causes a phase shift. While such timing offsets can be compensated with post-processing to some degree, it is generally preferred to have identical measurement timestamps for each channel and sample.

The ADC itself relies on a controlling device to handle all digital I/O functionality. An FPGA or micro-controller connects via a logic bus with the data lines of the ADC. The sampling rate is defined by an internal timer interrupt which generates the start-conversion trigger. The measurement data can then be received via a bus protocol, typically SPI, I²C, or similar interfaces. For single-shot measurements, the controller needs to read $\#channels \times \#bits\_resolution/8 \times sampling\_rate$ bytes per second.

The precise timing for each start-conversion trigger is crucial to provide equidistant sampling of the energy signal being measured. Most mathematical approaches for wave-form analysis, including audio-based signal processing, require uniformly sampled data points. If the measurement timing varies, the underlying signal cannot be reconstructed because of the unknown timing offset. This technique is commonly known as Pulse-Code Modulation (PCM) and describes the uniformly-distributed sampling of analog signals

into digital values. The signal strength (amplitude) is represented by a linear mapping into the available bit depth. Measurements occur in a fixed interval after each other.

5.2.2 Digital Data Acquisition

The controlling device is the first entity to handle digital samples of the measured electricity signals. Depending on the task at hand, various data processing tasks can be implemented at this stage. However, certain real-time restrictions still apply. The controller must respond to the start-conversation trigger signal and retrieve the samples from the ADC. The time between polling the ADC can be used for data processing and other management or communication tasks. A real-time operating system (RTOS) can enforce these deadlines and prohibit data loss due to unavailable compute resources [69]. Depending on the task, a straight-forward bare-metal implementation can be sufficient to retrieve the samples and forward them to a higher processing tier, while buffering data into larger chunks. An FPGA architecture can be beneficial, due to its multiple parallel processing capabilities. Micro-controllers (or micro-processor) typically only provide a single synchronized execution path.

CLEAR and MEDAL employ a similar architecture after the analog data acquisition (**DG1**). The controller in CLEAR is a Lattice XO2 7000-HC FPGA, whereas MEDAL uses a Microchip ATmega324PA micro-controller. The underlying architecture in both systems forwards the collected samples to a USB bridge device (FTDI FT232H), which is then polled by a Linux-based SBC. Direct forwarding of data imposes an elementary real-time restriction: both communication channels to the ADC and the USB bridge must be serviced before the next start-conversion trigger is scheduled. The packet layout is determined by the bit depth of the ADC and the number of channels. CLEAR uses 6 channels with 16-bit each and a 2-byte counter for loss detection, resulting in 14-byte long packets for each sampling. MEDAL uses 7 channels with 12-bit each and the same 2-byte counter, also resulting in 14-byte long packets (**DG1**). The counter is a simple *unsigned short* variable, monotonically increasing with every new start-conversion trigger, and overflows after 65536 packets and restarts at 0, which the *unsigned arithmetic* already handles without further intervention. The SBC can use the counter to detect data loss if the latest counter value is not consecutive with the previous packet.

We chose USB as communication protocol between controller and the SBC due to its versatility and high availability in dedicated bridge devices. Other bus interfaces to provide bi-directional data transfer between systems include Ethernet, CAN bus-based protocols, SPI, or I²C. While some of these protocols might already be available on the employed integrated circuits, the achievable data rate, timing restrictions, and framing complexity need to be considered. Transferring data between an embedded system (RTOS) and a general-purpose computing device (Linux on an SBC) poses problems due to asynchronous data handling. Buffers and decoupling processes are required to prevent data loss or blocking bus access (congestion).

The USB bridge (an FTDI FT232H configured as *FT245 style asynchronous FIFO interface*) provides a signal if the internal send- or receive-buffers are full. The controller checks this signal before writing a new packet, or waits until the USB bridge accepts new write operations. While waiting, a new start-conversion trigger might occur, and another data packet becomes ready, therefore, all outgoing data packets are buffered in a FIFO queue in the controller's internal memory. While a short buffering can be handled, the available memory limits the FIFO queue size and therefore the maximum duration of USB congestion, after which data loss occurs (**DG3**).

5.3 Data Processing

The USB bridge, a fully self-contained integrated circuit, provides a full interface to the USB protocol standard, which offers multiple transfer types to send or request data chunks between the host (SBC) and slave device (USB bridge). The transfer type needs to match the data rates and HA requirements defined by the DAQ specification. Data is sent to the USB bridge via a simple bus interface. Chunks are buffered into larger packets to minimize transport protocol overhead.

USB supports multiple transfer types, which need to be requested and implemented by the host device and driver software [70]. Interrupt transfers are not suitable for DAQ due to its packet size limitation. Control transfers can be used for configuration and command communication, but are unsuitable for measurement data. Isochronous and

bulk transfers are designed for large data packets and fast transmission speeds.

For isochronous transfers, the host requests a new communication pipe and the slave device then periodically streams new data through it. This transfer type is commonly used for audio, video, or other time-based multimedia data streams. While isochronous pipes match the high-level task description of energy measurement DAQ systems, some restrictions and caveats apply to this type of USB transfer mode. Bandwidth and latency is guaranteed within a pipe, but also take precedent over data integrity. This means no error correction or transfer retries are performed. If the non-real-time host is busy with other tasks, the isochronous pipe might discard or stop sending new data.

For bulk transfers, the host needs to register a new bulk transfer request with the USB subsystem of the operating system (Linux kernel). Once the slave is ready and signals that it can fulfill the request, all available data is transferred. After the request is serviced, a new request can be processed. In the case of continuous measurement streams, multiple active requests can be registered, to prevent a queue starvation of the USB subsystem. The controller and the USB bridge only have limited memory dedicated for measurement data buffering, therefore, the bulk transfer queue must always contain active requests to prevent bus congestion (packet or data loss). The number of active requests depends on the sampling rate, chunk size, and buffer lengths. CLEAR and MEDAL units can be configured to ensure reliable data transmission. During the BLOND-50 and BLOND-250 data collection series, we used a request queue length of 2048 (out of a maximal 4096). In USB high-speed mode, a single bulk transfer can contain up to 512 bytes, with two bytes being reserved for error checking. With this architecture, we can transfer 36 full samples per USB bulk transfer using 14-byte packets from the controller through the bridge to the SBC. CLEAR, with 6 channels and 16-bit resolution, can sustain a constant throughput of 250,000 Sps or 3,500,000 B/s without any data loss. MEDAL, with 7 channels and 12-bit resolution can achieve 50,000 Sps or 700,000 B/s without any data loss. The limiting component in both systems is the maximum sampling rate per channel of the ADC.

With these data rates and a queue of 2048 bulk transfers, the USB subsystem is busy for at least 300 ms (CLEAR) and 1500 ms (MEDAL). The Linux kernel uses a preemption

latency of 6 ms by default[1], which gives the process scheduler enough time to call our measurement process to process the data packets and refill the USB bulk transfer queue.

5.3.1 Single-Board Computer Measurement Governor

The Linux-based SBC needs to make the transition from embedded real-time into general-purpose (non-real-time) processing. Some timing and memory restriction can be relaxed, compared to the ADC control. We implemented a dedicated measurement governor that communicates with the real-time components over USB bulk transfers and processes the received data, see Figure 5.3.1. This allows us to decouple the responsibilities of data acquisition and data processing.

The governor is implemented in C to make use of low-level control of memory and process scheduling. It provides multiple configuration options, including sampling rate, file size, and USB bulk transfer queue settings. These options can be passed as command-line arguments, environment variables, or can be read from a human-readable file. Default values are supplied in the source code and can be adapted to fit the ADC and controller specifications. The core communication is handled by a combination of *libusb* [71] and *libftdi* [72] driver frameworks. Monitoring and auto-reset capabilities are provided by a systemd service [73] integrated with the underlying operating system.

The command communication link between the governor and the embedded controller is based on the structure of AT-commands (inspired by the Hayes command set [74]), which can be used to configure the sampling rate and start or stop the data acquisition (start-conversion interrupt timer). A few additional commands have been implemented for debugging purposes to retrieve the firmware version and status registers. After starting the data collection process, the governor starts filling the bulk transfer queue with the configured number of requests and hands them off to the USB subsystem. The kernel now takes care of the direct memory access (DMA) for the USB subsystem and provides a callback once a request is fulfilled. The governor sits idle until receiving the callback, in which it copies the data packets from the completed bulk transfer into a queue for

[1] Linux v4.4.21, `http://elixir.free-electrons.com/linux/v4.4.21/source/kernel/sched/fair.c#L50`

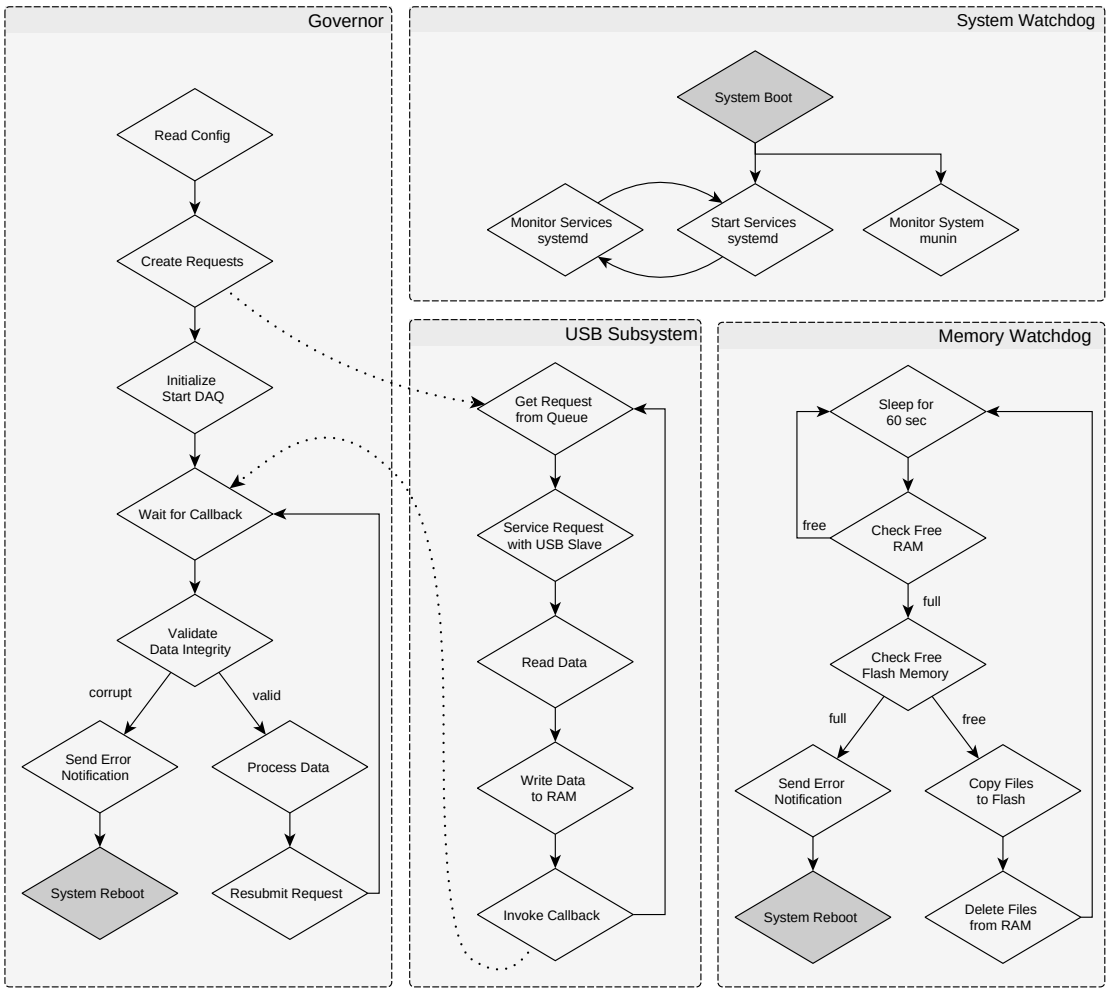


Figure 5.3.1: The software architecture on a single-board computer with the Measurement Governor and subsystems. Each service is managed and controlled by *systemd* and communicates via shared memory regions.

asynchronous processing. The completed bulk transfer can be reset and resubmitted to prevent queue starvation. This process continuous indefinitely until a user-issued stop command is received.

Once the measurement data is available in the SBC main memory and the bulk request is resubmitted, the remaining CPU time can be used for data processing, analytics, or sending the data over the network. The governor can include processing instructions, which are performed on every new data chunk, or it can write the data into a file for external processing. A multi-core SBC, such as a Raspberry Pi 3 (quad-core Broadcom BCM2837) and LattePanda (quad-core Intel Cherry Trail Atom X5), offer sufficient computational power for local processing. USB bulk transfer callbacks, processing of data, and network access can be modeled as individual tasks. The governor, with the help of the Linux kernel, can pin certain tasks to a dedicated CPU core and assign a scheduling class and priority. This can help to prevent queue starvation and other issues caused by blocked or unavailable resources.

The governor provides two endpoints for measurement data: storing all raw data into files, and streaming calibrated measurement samples for all monitored channels into a user-definable executable. The governor is started with an executable filename, which is automatically initialized before data collection begins. The executable receives metadata from the governor via command line arguments. The data is streamed into a named-FIFO pipe managed by the Linux kernel. The pipe provides a time-decoupling and can buffer multiple seconds if the analytics pipeline is busy.

5.3.2 Data Processing Modes

The governor and the data processing pipeline can be configured with two operation modes: STORE and STREAM, see Figure 5.3.2. Both modes can be used independently or simultaneously and are supported on CLEAR and MEDAL units.

STORE provides data collection functionality to collect long-term continuous measurement series, see Section 5.3.3, as used for the two BLOND measurement series. This is the primary operation mode and all decoupling methods are designed to maximize reliability.

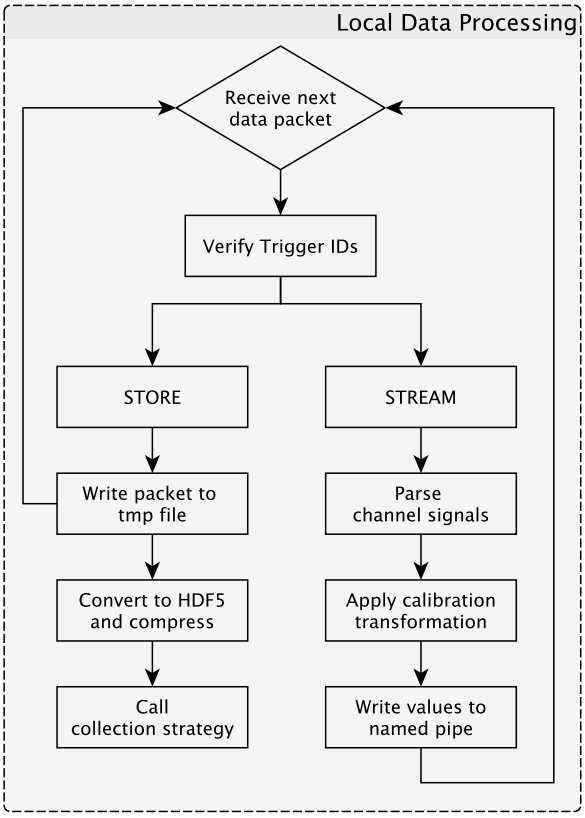


Figure 5.3.2: Governor used in STORE and STREAM data processing modes. Sample use cases are long-term continuous data collection and real-time data analytics tasks.

The data acquisition via the USB subsystem does not stress the CPU significantly, only the file conversion and compression engage the CPU.

STREAM is designed to provide real-time access to measurement data within the DAQ unit for arbitrary analytics task (**DG6**). The signal stream contains ready-to-use calibrated measurement values of voltage and current waveform data with the requested sampling rate for each channel. A uni-directional communication layer between the governor and a child process gets established via named pipes in the operating system. The amount of decoupling buffer can be configured to queue up data if the analytics task is busy. The user is responsible to design and implement the analytics task within the available resources. Excessive use of CPU (all cores for longer periods) or memory (swapping) can cause data loss and system resets. While the DAQ unit "self-recovers", minor data loss can occur.

5.3.3 Data Collection for Long-Term Continuous Measurements

Collecting long-term measurement data with a "capture everything" approach results in multiple continuous data streams. The required bandwidth and storage space per file can be determined based on the bit depth, sampling rate, and number of measurement channels, and optional data compression. Time-series data are commonly stored in equally-sized chunks (files). Each file contains a timestamp and sequence number to find measurement data prior or after the current file.

The governor buffers the collected data chunks in the main memory, before processing them further. The individual processes of the governor share main memory via a memory-mapped filesystem (RAM disk). This provides a simplistic exchange mechanism through the Linux kernel I/O module without the actual overhead of hardware-based I/O. For the two time-series datasets BLOND-50 and BLOND-250 [25], we aggregated the data into larger chunks based on the duration (2 min, 5 min, or 15 min) per measurement unit. The BLOND environment, common to both datasets, utilized 15 MEDAL units for ground truth data collection and one CLEAR system for aggregated building mains measurements, resulting in the periodic creation of 16 new chunks. Once a chunk is complete (reaches its defined maximum file size), the governor hands off the raw binary blob to a processing pipeline and continues with the creation of a new chunk. The blob is augmented with a timestamp, a sequence number, and additional metadata for the processing pipeline.

5.3.4 Time Synchronization within a Fleet of DAQ Units

Collecting measurement data with a fleet of sensors requires some form of time synchronization to re-align the individual time-series data chunks of different measurement units before analyzing them. Each unit generates its own start-conversion trigger based on the internal oscillator that also drives the embedded system. These electronic oscillators are typically precise enough for running micro-controllers and peripheral circuits. However, they can drift when compared to other systems that are not connected. High precision oscillator crystals can achieve an accuracy of less than 5ppm [75], which still drifts by approx. 3 min per year. Temperature and input voltage fluctuations can have a negative

effect as well.

Embedded controllers typically use a main oscillator for the CPU clock frequency and to create timer interrupts. The clock frequency can be derived with an integer factor to generate lower frequency, which are a multiple of the main oscillator frequency. This prescaler value can be chosen to get a finer granularity for the ADC start-conversion timer. The selected sampling rate should match the mains frequency (50 Hz or 60 Hz) to provide a stable number of samples for each mains period. If the sampling rate, oscillator frequency, and prescaler do not match (integer ratio), aliasing effects might occur due to roll-over of individual samples per mains period.

The effect of oscillator drift causes the start-conversion trigger to deviate from the selected sampling rate. The MEDAL and CLEAR measurement units create new data files based on the number of contained samples. Assuming a wall time (the time an observer can monitor on a clock hanging on the wall) of 1 min has elapsed, one unit could have produced 2,999,995 samples, whereas another unit collected 3,000,005 samples. Each data file is combined with a timestamp to indicate the time of the first sample of this file. This time is taken from the operating system clock, which also drifts. In addition to the expected drift, the wall time is also affected by daylight saving time (DST) and leap seconds. While DST can be accounted for by using a Coordinated Universal Time (UTC) timestamp, a leap second needs external input for correction. This correction is relevant for BLOND-50 but not BLOND-250, due to the December 31, 2016, leap second of +1.

For these reasons, an external clock synchronization is required to ensure the individual signals of different measurement units can be aligned and analyzed as a single unified time series. The Network Time Protocol (NTP) [66] is commonly used in IT systems to query, synchronize, and set the time of individual devices with millisecond accuracy [65]. While collecting the BLOND datasets, each measurement unit was synchronized to a stratum-3 time server available on the same Ethernet connection, with periodic updates and corrections. The overall accuracy of the synchronized clock is precise enough to align two independent mains voltage signals based on their 50 Hz or 60 Hz fundamental frequency, with a period length of 20 ms or $16.\dot{6}$ ms.

The Precision Time Protocol (PTP) is an NTP-alternative which offers a higher accuracy in the sub-microsecond range without the need for external GPS receivers at each node

or network segment [76]. The time information is synchronized in a master-slave architecture, where each node (server computers and network devices) in a computer network participates. The clock is distributed hierarchically, therefore each hop needs to support PTP to maintain the precision. This renders PTP unusable for typical Internet-of-Things environments with unknown network devices within the route between a "thing" and the data center. While most enterprise-grade network devices (routers and switches) offer a PTP implementation, the SBC devices in MEDAL and CLEAR lack such support, due to the missing hardware and firmware implementations of the network interface controllers. Although a software-based implementation exists, the promised accuracy of PTP is only achievable with dedicated hardware support.

Clock synchronization is performed during normal runtime of the system. As such, time information is only valid while the system is powered up and operational. During a power outage or reboot, the synchronization is lost and needs to be reacquired. A real-time clock (RTC) is a dedicated hardware clock with a backup battery to maintain an accurate clock even when the system is without external power. Most SBCs lack such RTC modules and backup batteries to save costs. We added a dedicated RTC module and super-capacitor to MEDAL, which maintains the clock information for multiple weeks without external power. The Atom-based SBC used in CLEAR already contains an RTC module and only requires adding an external battery to power it.

When measuring a household, floor, or building, the mains sine wave propagates at a fixed speed from the distribution panel to the individual appliances and measurement units. The speed of electricity can be approximated with the speed of light [77], and therefore the distance between distribution panel and appliances can be disregarded for sampling rates below 1 MHz and copper wire lengths of less than 100 m. For long buildings, such as factories with assembly lines, the wire length can cause a phase delay between the incoming and outgoing current.

5.4 Pull-based Data Processing

The data collection architecture of BLOND-50 takes advantage of the edge computing concept, which performs a majority of the data processing on the local measurement unit themselves. The edge-based fleet of MEDAL and CLEAR measurement units is capable of processing raw measurements, performing basic data verification, and encoding the data into a suitable file format with compression and checksums. Cloud-based components pull the final files from each unit and archive them in a distributed data storage system. The pull-based strategy utilizes a collector, running in the data center, which periodically transfers data files from each unit and performs a few housekeeping tasks, see Figure 5.4.1.

The governor on each measurement system is configured to create a new data file every 15 min for MEDAL and 5 min for CLEAR (in BLOND-50). The local processing pipeline on the measurement unit receives a full data file from the governor and transforms the raw data into HDF5 files. Raw ADC values are encapsulated and split into individual channels before assigning them to HDF5 datasets (a subgroup inside an HDF5 file). This encoding process runs independently (in parallel) of the governor to provide relaxed timing requirements. The resulting HDF5 is stored in the in-memory filesystem, while the raw data chunks are discarded after successful conversion. The HDF5 file format offers built-in compression of all numerical data, which reduces the required bandwidth by 50-60% on average. All these tasks are performed at the edge of the data collection infrastructure, i.e., the CLEAR and MEDAL units (**DG2**).

The governor and the local processing pipeline are designed to use the main memory as primary storage location. Keeping the collected data chunks and their processed HDF5 files in the main memory serves two main purposes: direct RAM access has a high bandwidth and low latency, and it does not cause any I/O operations to the flash memory. Persistent flash memory, such as used in USB thumb drives, SSDs, or SD cards, only provides a limited number of write and erase cycles [78]. Exceeding the vendor-specified write cycles to individual flash memory regions, the data integrity cannot be guaranteed, resulting in possible data loss or corruption. Volatile main memory (SRAM and DRAM), as used in all modern computer architectures, does not suffer from this issue, and can be used for infinite write and erase cycles, but does not retain data after loosing power.

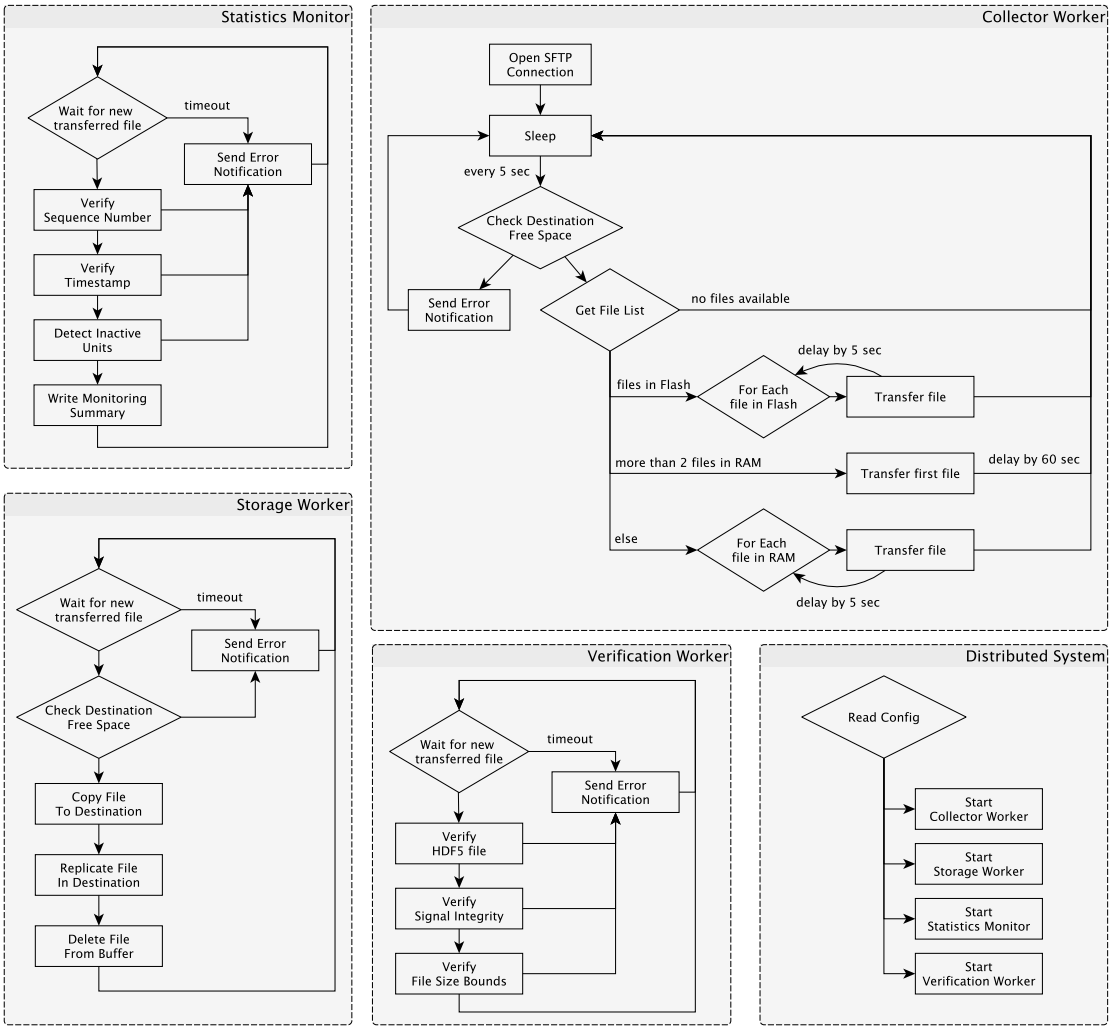


Figure 5.4.1: Pull-based Data Collection and Monitoring used for BLOND-50

A memory watchdog periodically checks the available storage space and swaps older files to the larger persistent flash memory, see Figure 5.3.1. The order in which files can be moved is determined by the pending operations. Raw data files need to be converted to HDF5 and cannot be moved while the conversion is running. HDF5 files cannot be moved while they are being filled with data or transferred over the network.

The pull-based collector periodically checks if files are ready to be transferred and copies HDF5 files one by one, see Figure 5.4.1. The cloud-based collector opens a persistent SSH/SFTP connection to each measurement unit and downloads the list of available files for each storage location (RAM or flash memory). During normal operation, the collector only sees a single file in RAM and transfers it into the data center. If there are multiple files in RAM, or any files in the flash memory, the cloud-based distributed data collection system is not performing in its ideal state. The main objective in such a case is to transfer all files as soon as possible to catch up with the regularly scheduled arrival of new files. The measurement units need to be protected from performance bottlenecks, which means the number of simultaneous file transfers is limited to one, the maximum network throughput is capped, and the collector inserts idle time in between file transfers to allow the non-real-time operating system to perform other tasks (refilling the USB bulk transfer request queue). All these parameters are tested and calibrated to achieve a stable, fault-tolerant, and forgiving data collection system that can self-recover in case of minor outages.

For every file the collector transfers into the data center, three server-side pipelines are triggered: verification, storage, and statistics. Each file is verified with multiple integrity and plausibility checks which could trigger an error notification, as described in [25]. The file is then copied and replicated into a distributed data storage system for long-term storage (archive and backups). All operations touching data and files are designed to "fail safe", i.e., to not loose data or metadata. The full pipeline on the data center can be resurrected in case of outages without loss of state or data. The main goal of BLOND was a long-term uninterrupted data collection, therefore each measurement system and file transfer is logged to generate statistics and error notifications in case a transfer time window passes without any new available files (**DG4**). A human operator was tasked to respond to these notifications within 48 hours. The flash memory in each measurement system can buffer up to 100 hours on CLEAR and 200 hours on MEDAL. The worst case

scenario would be a full flash memory, causing data loss (new data not being stored, or old data being overwritten). Neither BLOND-50 nor BLOND-250 suffered from any such situations. The sufficiently large buffer sizes allow us to deploy new software packages and data center services while collecting data.

5.5 Push-based Data Processing

The BLOND-250 dataset was collected with a 5× higher sampling rate for CLEAR, and 7.8× for each MEDAL unit. The increased data bandwidth for each measurement stream rendered the pull-based strategy unfeasible due to resource limitations of the SBC. While basic data processing tasks can still be serviced, the remaining compute time was not sufficient to convert and compress the data into HDF5 files at the edge of the measurement network. Therefore, we utilized a push-based data processing strategy, which performs all resource-intensive tasks in the cloud, instead of doing them locally.

The governor performs the same tasks as previously, with a configured maximum file length of 2 min for CLEAR and MEDAL units. The data acquisition pipeline immediately continues to write incoming new data to a new file, while the just finished file is handed off to an upload stage, instead of being converted to HDF5. The upload stage uses the same network transfer system of the pull-based strategy, however, the measurement unit is now the initiating endpoint. The file is uploaded into the data center, while handling connection drops with automatic retries and exponential backoffs in case the data center is unavailable. After successfully transmitting the file, it is deleted from the measurement unit.

Due to the increased sampling rate and the missing data compression, the flash memory in each measurement system can buffer up to 10 hours on CLEAR and 13 hours on MEDAL (**DG4**). While this would allow some minor outage of the cloud components to be resolved without data loss, actually transferring the data once the data center becomes available again would be limited by the network bandwidth cap. The human operator tasked with responding to error notifications of the measurement infrastructure becomes a higher priority with the reduced buffer time in a push-based strategy.

The main compute resources used in a push-based strategy are located in the cloud, as the available performance at the edge is not sufficient for data collection and simultaneous compression of completed files. The collector, as depicted in Figure 5.4.1 for the pull-based strategy, is replaced in the push-based strategy by a converter worker, which processes and converts all incoming files to HDF5 and compresses the raw data streams.

5.6 Scalability and Resilience

Main drivers for scalability in large distributed sensor networks are data transfer bandwidth, compute power, and storage requirements. We evaluated the proposed architecture with the collection of two long-term datasets: BLOND-50 and BLOND-250. Both datasets make use of the outlined design principles regarding data acquisition and processing: from the analog signal measurements, into embedded real-time computing devices, SBC edge devices, and finally cloud-based back-end systems for long-term storage. The BLOND datasets use a CLEAR unit as single smart meter for a 3-phase power grid, to collect the whole-building energy consumption. The ground truth energy consumption was collected with 15 MEDAL units as per-appliance energy meters. The datasets consist of 93 current measurement streams and 18 voltage streams.

The measurement environment provides a common Layer-2 IP network, consisting of 1 Gb/s Ethernet LANs, joined via fiber-based uplinks into the data center. Each DAQ unit is connected to the data center and periodically transmits (pull- or push strategy) data files, typically in short bursts as long as data is available. Over the course of collecting a long-term dataset, the time drift of the units will distribute these burst transfers equally. The averaged bandwidth requirement in BLOND-50 to transmit compressed data files was 35 KiB/s for a single MEDAL unit, and 375 KiB/s for CLEAR. The increased sampling rate and the non-compressed data transfers in BLOND-250 generated on average 265 KiB/s for one MEDAL unit, and 1800 KiB/s for CLEAR.

The employed network infrastructure would therefore theoretically accommodate a fleet of over 400 MEDAL units together with one CLEAR unit, all connected via a single Ethernet port. However, in real-world environments, the fleet is distributed over a larger

physical area (spanning floors and rooms), which allows for even larger measurement networks. We therefore conclude that network bandwidth is not a limiting factor for scalability in this context.

While the SBC edge devices with quad-core CPUs do have memory and processing power limitations, the workload and tasks during data collection and processing have been allocated according to the data collection strategies. Scientific data analysis can be implemented at the edge of the sensor network, however, the processing power scales only vertically, based on the availability of SBCs with higher performance density (CPU and RAM resources). The cloud (data center, either local or reachable via the Internet) can provide virtually unlimited resources if required. This allows us to scale horizontally in the data center to accommodate an increasing number of measurement devices without introducing new complexity (**DG2**).

Limitations and real-time requirements have been introduced in Section 5.2.2 and the compliance was validated with an extensive technical validation of all collected data and metadata. We define the main criterion as the resilience of a given data collection network (fleet of sensors communication with the cloud) to power, network, or server outages, as well as deployment of new client and server software packages. A resilient system can cope and recover from any of these issues (within a predefined length or time window).

5.7 Evaluation

The presented architecture and design guidelines were implemented and tested with the creation of the BLOND datasets [25]. The achieved measurement coverage was 99.997%, putting it well above the typical 99.99% cloud provider SLAs [79, 80, 81]. It has to be noted that the uncovered regions are confined to the BLOND-50 sub-dataset, and their primary cause was a direct result of non-conforming operation of the underlying infrastructure: the first event was caused by a corrupt operating system upgrade which lead to a reboot of the measurement unit, and the second event was caused by an improper network filesystem operation where the data was actually collected but erroneously overwritten.

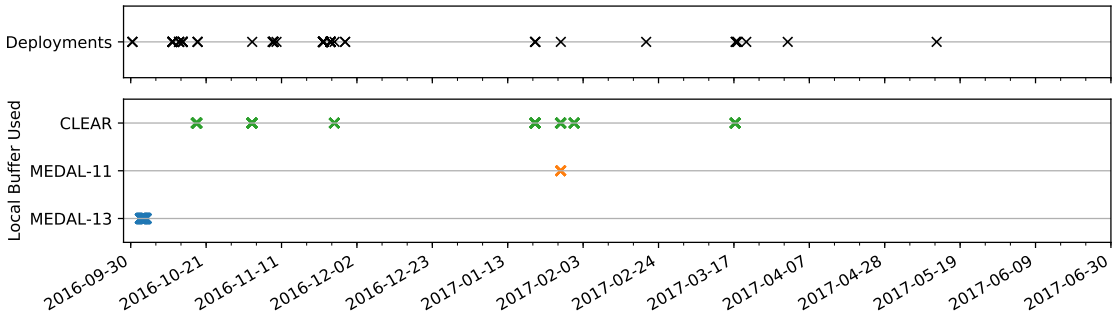


Figure 5.7.1: Analysis of software deployments and local buffering operations at each measurement unit. Local buffering moves data from RAM onto the persistent flash memory. Software services have been continuously improved and deployed without interrupting the ongoing data collection.

During the collection of BLOND-50 and BLOND-250, the cloud-based software services were re-deployed on 70 independent occasions, see Figure 5.7.1. These events where part of a continuous deployment strategy to improve the software quality and feature set. Various components received new functionality and bugfixes during each deployment. Each cloud-based component preserves any intermediary state or operates stateless to handle service restarts gracefully (**DG5**). The decoupling between data acquisition and collection ensured that measurements are continuously collected, even during a network or component outage. Each measurement unit keeps recently collected data files in RAM to save I/O access. If the collector does not retrieve the files, and RAM utilization reaches its limit, older files are moved to the flash memory – which occurred on multiple occasions for CLEAR and two MEDAL units (No. 11 and 13), none of the other measurement units had to make use of this. This safe guard was transparently handled by all components and the system returned to a normal operating state. It can be noted that some of these timestamps correlate with software deployments, which is expected, while others appear to have no obvious trigger.

In the pull-based data collection strategy, each measurement unit performs local data processing to convert the raw data into HDF5 files and to apply internal compression, before the collector can retrieve the files. The edge-based compute power was fully utilized, with one CPU core dedicated to this task (**DG6**). The processing duration for each file was recorded, see Figure 5.7.2. CLEAR shows the fastest task duration, due to its superior CPU, although it had 2.2× more data to process than a MEDAL unit (90 million samples vs. 40.32 million samples per file).

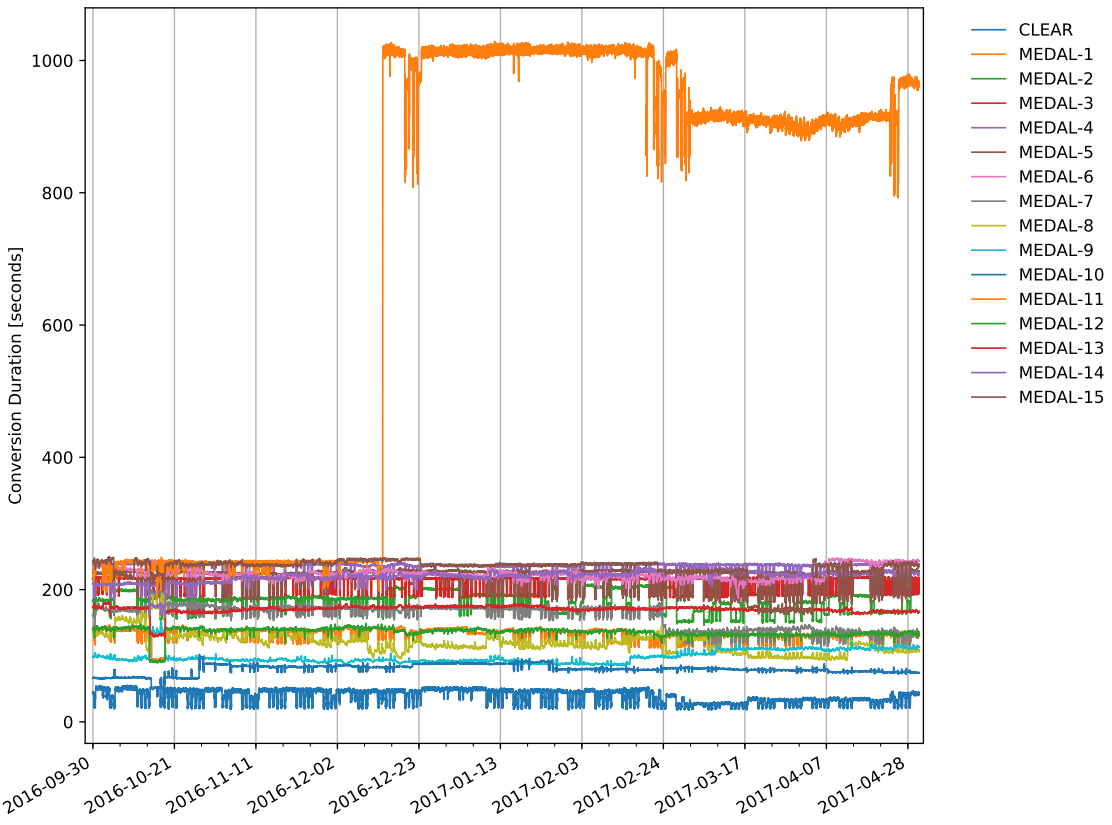


Figure 5.7.2: BLOND-50: Duration of edge-based data conversion and compression for each collected file.

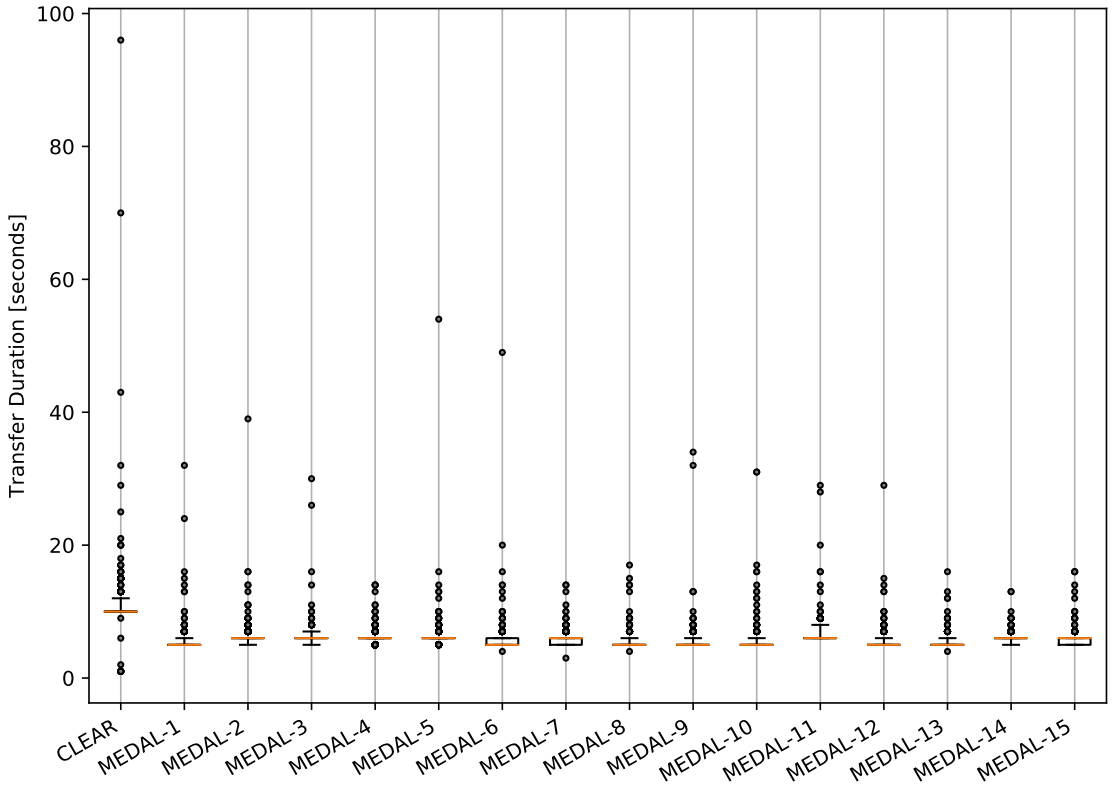


Figure 5.7.3: Distribution of transfer time per file from the measurement unit into the data center (pull-based data collection). The boxplot shows the lower and upper quartile in the box, with a red line for the median value, and whiskers range from 1-99%.

The transfer time per file strongly depends on the overall network utilization and the capped throughput (available bandwidth) for each measurement unit, see Figure 5.7.3. The transfer duration is consistent over the majority of BLOND-50 and BLOND-250, while a few outliers can be attributed to software deployments and local buffering. All MEDAL units show identical distributions and spread. The CLEAR unit generated larger files, and therefore also has higher transfer times. Although one MEDAL unit (No. 11) experienced an out of the ordinary conversion time, see Figure 5.7.2, the transfer time was unaffected. The observed transfer duration only accounts for successful transfer operations, failed and retried connections are not accounted for.

5.7.1 Scalability and Deployment Strategies

The cloud services used while collecting BLOND-50 and BLOND-250 have evolved and were updated frequently to improve usability and monitoring capabilities over the course of the two measurement periods (**DG5**). In total, the individual components were re-deployed 69 times to the data center without any interruption of the measurement time series. On each occasion, the implemented safe guards and buffers performed as designed and decoupled the analog and digital data acquisition from data collection tasks in the data center. The underlying infrastructure (compute and storage servers used to host the data collection services) received regular OS updates without prior knowledge.

5.7.2 Scientific Workload: Event Detection

The proposed architecture consists of a dedicated embedded system for analog and digital data acquisition, a USB-based communication channel to an SBC, and the asynchronous data processing with in-memory buffers. The governor, as well as the pull- and push-based data collection strategies, have been thoroughly evaluated during the creation of the BLOND datasets. One of the main goals of our DAQ architecture is to provide sufficient processing power for scientific workloads (data analysis tasks) at the edge of the sensor fleet network (**DG6**). We provide an easy-to-use programming interface to retrieve measurement data and run experiments, processing pipelines, and machine learning approaches directly on the DAQ system without the need for cloud-based aggregators or additional hardware.

We evaluated the DAQ architecture on MEDAL, due to its reduced processing power (compared to CLEAR). While both systems contain a quad-core CPU, the ARM processor used in MEDAL is less performant than the 64-bit Intel Atom processor in CLEAR. In the context of dataset collection, MEDAL was designed to measure the ground truth electrical energy consumption (per-appliance consumption). Therefore, we implemented an on-device event detection. The use case is a data collection task, combined with a real-time event detection to generate annotated ground truth data (switch on/off events).

5.7.3 Event Detection with Deep Neural Networks

We use an artificial deep neural network to detect appliance switch events in a 12 s time window and continuously evaluate the model prediction. The input data is directly streamed from the governor into a Python-based data pipeline. While the governor is written in the low-level language C, the event detection pipeline (data streaming) is implemented in Python to utilize existing machine learning frameworks, such as Numpy [82], Keras [83], and TensorFlow [84].

The current signals of all available sockets are transformed into sub-second RMS values. The cumulative sum of these values is then fed into a denoising autoencoder for event classification (event vs. non-event). Our event detection model consists of a 4-layer autoencoder, which uses recurrent layers to encode the input. The event detection algorithm trains the model on non-event data and uses the reconstruction error to find events in the data stream. This error metric increases when unforeseen change points occur in the signal. The model has two encoding and two decoding layers, with the inner most encoding layer being 100-dimensional. Hence, the high-dimensional input signal (current RMS values) is represented in a low-dimensional space and reconstructed again. After detecting signal windows that contain the events, we determine the exact change points by using a simple threshold algorithm. The model has close to 1,600,000 learnable parameters (10 RMS values per second, each 5 mains cycles, over 12 seconds of input data with LSTM cells), making it a comparatively high computational load. The quad-core CPU in MEDAL can process multiple sockets in parallel, making full use of the available computational resources.

The model was trained with the fully-labeled events of phase B in BLUED [23], because of the higher variance in appliances compared to phase A. Accuracy and F-score of the implemented event detector are not relevant for the evaluation of our data acquisition architecture, since the model and its parameters can be adapted easily. The critical metric is computational performance, i.e., if the hard- and software architecture are capable of handling on-device data analysis tasks in near real-time (without significant delays or offline processing).

5.7.4 Event Detection with k-NN

An alternative event detection system was designed and implemented as multivariate supervised binary classification task. We use a k-NN classifier to divide events and non-events based on a 10 s time window. The current signals are streamed from the governor into a model prediction pipeline using Octave [85].

The k-NN model relies on a large number of labeled events and non-events (supervised approach). The temporal position of non-events in the model are implicitly known due to the inverse relationship: all time windows without any labeled events can be used as non-event windows. The cycle-wise cumulative sum of RMS values of each window forms a 500-dimensional feature space. The neighborhood size N and the distance metric define the computational complexity of the model evaluation and can be adapted. It has to be noted that we already achieved high event detection accuracy with $N = 1$ and the `cityblock` metric.

5.8 Discussion

This work presents comprehensive methodologies for electrical energy data collection, data acquisition architectures, and measurement networks. We defined a set of design goals and requirements to support long-term continuous data collection in the context of NILM with uninterrupted waveform data for voltage and current signals.

Our data acquisition and processing pipeline handles a heterogeneous number of signal streams, as implemented in the CLEAR and MEDAL systems, with a unified software control stack. The edge- and cloud-based components are designed to ingest data streams in arbitrary structure (multiple voltage or current signals) (**DG1**).

We collected two measurement series, BLOND-50 and BLOND-250, in an office environment with multiple rooms using a distributed fleet of measurement systems. A central communication network was used to transmit collected data from the DAQ units into a data center for further processing and persistent storage. The measurement network

can be scaled to support a large fleet to cover entire buildings or campus areas without network congestion or data loss. (**DG2**)

Uninterrupted data collection is an important feature for NILM-related datasets to provide gap-less signal streams for power disaggregation and appliance identification tasks. We designed the BLOND measurement infrastructure to high availability requirements and minimized outage and downtime effects. All data files were validated and checked against a comprehensive set of quality metrics to ensure no data corruption. We achieved an overall data availability of 99.997%, which is higher than the standard SLA of cloud infrastructure providers (**DG3** and **DG4**).

The edge-based measurement governor, the cloud-based collector, and processing pipeline are designed to queue incoming data and save all intermediary information to a persistent storage location. We continuously improved our active edge and cloud components and deployed new features and bugfixes without any downtime. All decoupling buffers, queues, and temporary storage locations have been tested and were automatically engaged if the system detected a congestion at one of the pipeline stages (**DG5**).

We provide two data collection strategies and two modes of operation to create a flexible and configurable environment. Pull- and push-based strategies allow us to dynamically use compute resources at the edge or in the cloud and define active endpoints in the communication channels. The STORE and STREAM modes allow us to operate the DAQ systems with two different use case scenarios. STORE can be used to collect long-term continuous datasets, while STREAM can be used for real-time data analytics at the edge of the measurement network with arbitrary scientific workloads for event detection, power disaggregation, or appliance identification tasks (**DG6**).

6

Building-Level Office eNvironment Dataset of Typical Electrical Appliances

Energy metering has gained popularity as conventional meters are replaced by electronic smart meters that promise energy savings and higher comfort levels for occupants. Achieving these goals requires a deep understanding of consumption patterns to reduce the energy footprint: load profile forecasting, power disaggregation, appliance identification, and startup event detection. Publicly available datasets are used to test, verify, and benchmark possible solutions to these problems. For this purpose, we present the BLOND dataset: continuous energy measurements of a typical office environment at high sampling rates with common appliances and load profiles. We provide voltage and current readings for aggregated circuits and matching fully-labeled ground truth data (individual appliance measurements).

This chapter is structured as follows: Section 6.1 gives the background and a summary of the dataset conception. Section 6.2 describes the collection environment and setup, while Section 6.3 covers the data, metadata, and structure of the published data descriptor. Finally, we present a technical validation in Section 6.4.

6.1 Background & Summary

Existing datasets predominately cover household and residential environments [19, 22, 23, 24, 86, 87, 88, 89, 90, 91, 92, 93, 94] due to the cost savings potential for their occupants. Large appliances (space heating, HVAC, washing machines) are being targeted first to achieve an immediate reduction in EEC since households typically contain a manageable number of them. These devices are easier to detect than multiple smaller ones, therefore, most datasets use measurement intervals of 1 sample per second (Sps), 1 minute, or lower. Using sampling rates above 10 kSps is beneficial to the total number and types of distinguishable appliances in a circuit with NILM and appliance identification research questions [15]. The amount of information contained in electricity signals increases steadily with sampling rates ranging up to 1 MHz. Higher sampling rates can capture subtle changes (high frequency ripples), which are useful for appliance identification [5, 15, 18, 49]. Capturing the voltage and current waveforms allows energy disaggregation algorithms such as BOLT [6] to extract patterns directly from the raw measurement data. To the best of our knowledge, only the datasets in [22, 23, 24] provide aggregated sampling rates above 10 kSps. In contrast to the aggregate measurements, the ground truth is only available with low sampling rates, making it difficult to correlate data of individual appliances to the mains EEC with high timing accuracy (see Table 6.1.1).

Table 6.1.1: Overview of long-term energy datasets with high sampling rates. This includes only datasets with long-term recordings of aggregate (above 10 kSps) and per-appliance measurements. In contrast to existing datasets, BLOND also provides ground truth data with a high sampling rate.

Dataset	REDD	BLUED	UK-DALE	BLOND-50	BLOND-250
Aggregate	15 kSps	12 kSps	16 kSps	50 kSps	250 kSps
Ground Truth	1 Sps	1 Sps	0.1$\dot{6}$ Sps	6400 Sps	50 kSps
Circuits / Phases	2	2	1	3	3
Duration	119 days	7 days	655 days	213 days	50 days
Classes	8	9	16	16	16
Appliances	82	43	53	53	53

Office buildings have a large potential for EEC reduction since most office workers are unaware of the energy costs they cause [95]. Modern office environments contain a well-defined set of appliances equipped with switched-mode power supplies (SMPS). Information and Communication Technology (ICT) devices, including computers, moni-

tors, networking equipment, and battery chargers, mostly use direct current (DC) and require a power supply module. Recently, field research and trials have been conducted with buildings offering DC power sockets, removing the need for SMPSs [96, 97]. Recent studies found that SMPSs can have a significant effect on EEM accuracy and can cause deviations of up to 582% when comparing smart meters to conventional meters [98]. This is primarily caused by magnetic interference due to non-linear and fast-switching loads causing distortions in current sensor readings. A significant portion of the reported errors are caused by ripple currents in the frequency range of up to 150 kHz, which is currently not covered by any dataset. The authors found a significant correlation between sensor type and their measurement accuracy. While Rogowski coil-based sensors showed a positive deviation (higher readings), Hall effect-based sensors were found to predominately return negative deviations, compared to conventional electromechanical meters.

In order to study typical office appliances, in particular, ICT devices equipped with SMPSs, in the context of NILM and EEM, we present BLOND: a Building-Level Office eNvironment Dataset. We provide long-term continuous measurements of voltage and current waveforms in a 3-phase power grid of a typical office environment collected in Germany between October 2016 and May 2017. The dataset contains readings for aggregated circuits (smart meter) and the matching fully-labeled ground truth waveform of voltage and current with a high sampling rate for individual appliances. In total, 53 appliance types and 74 appliance instances, grouped into 16 classes, are distributed across 111 recorded channels. All signal traces are precisely timestamped with a globally synchronized clock. The dataset consists of two measurement series in the same environment with different sampling rates. BLOND-50 contains 213 days of continuous readings of all 3 phases (aggregated) at 50 kSps, with ground truth data (individual appliances) at 6.4 kSps. BLOND-250 contains 50 days at 250 kSps (aggregate) and 50 kSps (individual appliances). The setup incorporates one data acquisition system for the aggregated circuits and 15 units to record individual appliances, each capable of measuring up to 6 appliances. We also provide a precomputed 1-second data summary to enable research on data with lower sampling rates.

6.2 Methods

In order to create a new dataset focused on ICT devices equipped with SMPSs, which provides a benefit over existing public datasets, and applicable to NILM-related areas, we define the following requirements and desirable attributes:

High sampling rates are necessary to extract high-frequency features from SMPS and other non-linear loads. Existing datasets (Table 6.1.1) cover the range between 10 to 20 kilosamples per second (kSps), only covering the lower region of the sampling frequency bins described in [15]. New research questions can be posed with higher sampling rates, which could lead to improved accuracy and new types of algorithms.

Ground truth waveforms provide additional information compared to lower sampling rates (e.g., one seconds mean values). Therefore, it is beneficial to collect the per-appliance EEC with sampling rates that are capable of representing the actual mains waveform for voltage and current.

Raw data streams are useful if the desired information cannot easily be extracted during data collection, either because the use case is not known yet or different algorithms and filters might omit import data. This allows us to calibrate and optimize the signal quality for a given task.

Long-term continuous recording results in a gap-less data capture of the entire electrical circuit. Previously recorded datasets contain large gaps where simply no data was recorded or received due to various reasons. While technical systems always have a certain margin of error, integrity and completeness should be a high priority when it comes to high-frequency energy datasets.

Clock synchronization allows for a precise matching between aggregate and ground truth samples. The time-stamping accuracy is a side-effect of high sampling rates. Since most dataset collections happen with a distributed fleet of sensors, maintaining a precise world clock is crucial to the overall timing accuracy. Without proper synchronization, some sensors might drift in time and blur the aggregate-to-ground-truth relation.

6.2.1 Environment

The BLOND dataset was collected at a typical office building in Germany, with the main occupants being academic institutes and their researchers. The measured circuits are part of a single floor with 9 dedicated offices and 160 m^2 of office space with central (non-electric) heating. The average weekday power density over the entire measurement period was 11.7 W/m^2 – which fits into the category of typical office buildings of 9.5 W/m^2 to 13.5 W/m^2 [99]. Throughout the collection of the dataset the population working in the monitored offices varied from 15 to 20 people.

Periods of occupancy are closely aligned with the office work schedule in Germany: Monday to Friday with a majority of occupants being present between 9:00 and 18:00. Weekends show almost no usage of the office spaces and therefore also no electricity consumption. Major public holidays, such as Christmas and New Year, also show minimal presence in the building, as well as personal vacation days taken by occupants individually. This includes business trips, sick days, and other "out-of-office" days. Due to privacy restrictions, no such data were collected.

All occupants perform light-duty office work, utilizing personal computers, monitors, and other electrical appliances typical for this environment. Individuals working in this building spend the majority of their work time at the desk, with certain breaks for meetings or other activities outside their assigned offices. Some occupants are involved in academic work and teaching, giving weekly lectures or attending meetings.

The power system consists of a 50 Hz mains with 3 circuits with a nominal phase shift of 120° (typical 3-phase supply): *L1*, *L2*, and *L3*. Each office room is connected to one or two circuits, with neighboring offices being on alternating circuits (see Figure 6.2.1 and Table 6.2.1). The building is not equipped with electric space heaters or air conditioning. Therefore, the dataset only contains user-operated appliances and base loads.

In order to keep rewiring efforts to a minimum, the existing independent circuits for regular and emergency lighting was excluded from the measurements, and only the user-accessible wall sockets were part of the measurements. The offices are electrically grouped into two sectors, each with an independent 3-phase breaker switch, resulting in

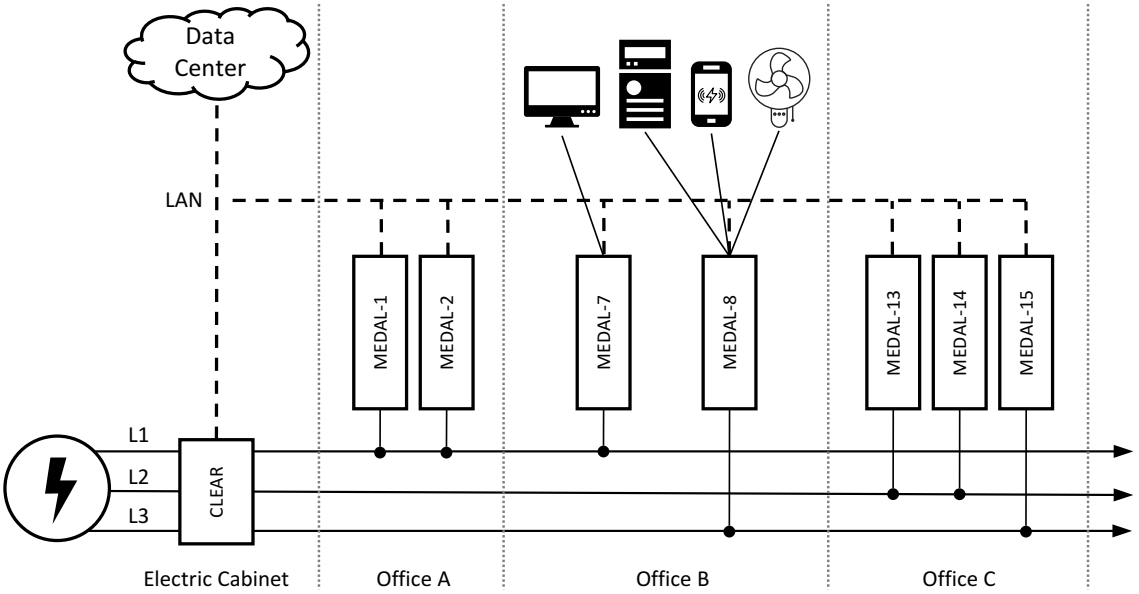


Figure 6.2.1: The measurement architecture of BLOND with physical placement of DAQ systems and connected appliances. A CLEAR unit is used as an EEC meter at the mains input to measure all 3 circuits in the electric cabinet. Multiple MEDAL units are placed in office rooms and connected to different circuits. Each MEDAL can be used to measure up to six appliances simultaneously in a single phase. Only a subset of MEDAL units is depicted; see Table 6.2.1 for a full circuit mapping.

6 circuits. Since the goal of this dataset is to collect aggregated mains EEC, every two circuits per phase are combined for measurement purposes, allowing us to use a 3-phase energy data acquisition system.

6.2.2 Aggregated Mains Measurements

Mains EEM was performed in the distribution board with a CLEAR unit [26], which was designed to meet the BLOND requirements. CLEAR, a circuit-level energy appliance radar, is a specialized data acquisition system capable of measuring voltage and current waveforms with high sampling rates and bit-rate for a 3-phase power grid. The power necessary to operate the sensors and the CLEAR system itself is drawn from a different circuit and not part of the measurement setup.

The CLEAR system (Figure 6.2.2) utilizes three Hall-effect based current sensors, installed in the electric cabinet (Figure 6.2.3), and a measurement box in the adjacent room

Table 6.2.1: Circuit mapping for each measurement system. Associating a MEDAL unit with its aggregated circuit in the CLEAR data is fixed and does not change over time. This corresponds to the wiring of individual office rooms to use one (or more) of three different phases.

CLEAR	MEDAL
L1	1, 2, 3, 7, 12
L2	6, 10, 11, 13, 14
L3	4, 5, 8, 9, 15

Figure 6.2.2: The data acquisition and processing component of the CLEAR system. The laser-cut acrylic enclosure contains three components: sensor board, DAQ board, and a Linux PC.

that contains all electronics and processing units. The electric cabinet and sensors are connected to the measurement box via 2 CAT-6 cables to provide shielded signal transmission and power. The voltage signals are directly tapped off the incoming mains line.

The employed analog-to-digital converter *AD7656A* samples all six channels (3 phases: voltage and current) simultaneously with up to 250 kSps [100]. Each signal channel is converted with 16-bit precision and bipolar value range, allowing for a direct mapping of the AC mains waveform into a digital data stream.

The ADC is controlled by a *Lattice XO2 7000-HC* FPGA to trigger the single-shot and read the data into memory for buffering. The resulting data packets are forwarded to a USB interface chip to allow for direct communication with a single-board PC. The

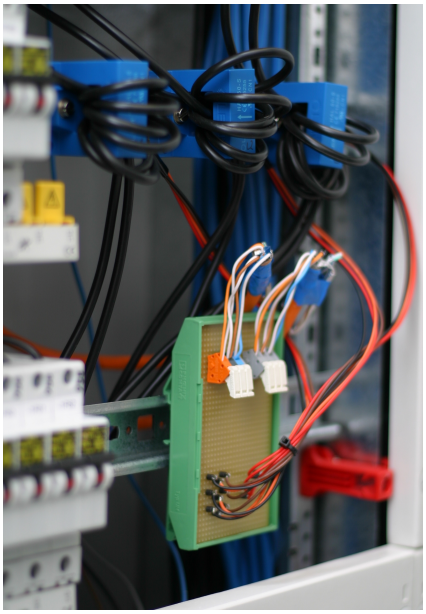


Figure 6.2.3: CLEAR current sensors installed in the electric cabinet. The open-loop Hall-effect sensors employ multiple turns of the mains wiring to increase the usable output signal. A small connection board distributes supply voltage and output signals. All changes and alterations were authorized and conducted by certified personnel.

Linux-based single-board PC receives the data and stores it into files, which then can be sent over the network into the data center for storage.

Each circuit in each room is protected by a 16 A breaker; each mains phase is protected by a 25 A breaker. A preliminary check showed typically less than 16 A per phase of total EEC over the course of a single day. Using *LEM HAL50-S* current sensors, we can utilize 3 primary turns to boost the effective signal bandwidth without exceeding the primary nominal current of 50 A per sensor [101]. The sensors come pre-calibrated and the calibration factor (linear mapping) was computed according to the data sheet.

The voltage signal is generated by an AC-AC transformer, which depends only on the open-circuit voltage and the minimal load during measurements. The calibration factor for the voltage ADC signal was computed by taking multiple RMS readings of a calibrated high-precision voltmeter and mapping it into the ADC signal.

6.2.3 Individual Appliance Measurements

The individual appliance EEM was performed by a fleet of 15 MEDAL units [27] acting as ground truth data for the aggregated mains measurements. MEDAL, a mobile energy data acquisition laboratory, is an off-the-shelf 6-port power strip, augmented with voltage and current sensing infrastructure in a compact and portable enclosure. A single-board PC is used to collect EEC data from the sensing hardware and to run the same software packages as CLEAR. Therefore, the fleet of MEDAL systems and CLEAR behave identically during setup and operation.

Each MEDAL unit measures up to 6 user appliances simultaneously with labeled sockets: #1 to #6. All power sockets in the offices are directly connected to a MEDAL system, used for base load equipment, or rendered unusable to prevent unmonitored appliances from being used. All monitored energy consumption is included in the CLEAR measurements and exactly one MEDAL data stream. MEDAL uses the same voltage sensing circuit and calibration as CLEAR.

All sockets produce an independent current signal with a Hall-effect-based IC from the *Allegro ACS712* family, providing a range of 5 /20 /30 A_{peak} per socket. Due to the expected ICT devices with SMPSs, we chose to configure each MEDAL unit with one high-power socket (up to 3600 W on socket #1), and 5 low-power sockets (up to 815 W, sockets #2 through #6). The maximum safe wattage is properly marked on the enclosure next to the socket. In case the plugged-in appliance exceeds that limit, the signal is limited to the maximum value, while still being electrically safe to operate. The EEC of a MEDAL system is less than 5 W and not measured in the ground truth data.

Most commonly available ADCs that offer simultaneous sampling of all channels can be expensive and are not suitable for a large-scale DAQ system. Therefore, MEDAL uses seven independent single-channel ADCs: *MCP3201* with 12-bit resolution and up to 50 kSps [102]. Precise timing and simultaneous sampling are achieved by using an *ATmega324PA* microcontroller as command & control IC.

6.2.4 Appliance Logs

An office environment with a moving and size-varying population can be an ever-changing setting to collect energy data. A list of observed appliances and their grouping into classes is available in Table 6.2.2. Most of these devices are small and portable, which means they can be moved around, plugged into different sockets, or simply appear and disappear on a daily basis. To prevent the incorrect labeling of appliance ground truth, a mapping between MEDAL sockets and actually plugged in devices was recorded in the appliance log: a spreadsheet containing timestamps, class name, appliance name, nominal power consumption, and socket numbers. The full log for each MEDAL is available in a JSON-based file format and as a spreadsheet file for easy printing and visual inspection. Although the appliance log is mostly based on self-reporting and periodic checks by trained professionals, a certain margin of error cannot be avoided. The curation of this data was carried out to the best of our capabilities and with due skill, care, and diligence.

Monthly checks were conducted to update the appliance log. Occupants were instructed to give notice about changes, so an update can be entered into the appliance log. An in-depth evaluation of the daily EEC was conducted retroactively to further improve the data quality. In cases a mismatch with the actual metered data was found, the appliance log was augmented with additional entries. This was only applicable in cases where a mismatch was deterministically resolvable by either using data from adjacent days, or by questioning the occupant responsible for the MEDAL system. Sockets marked as empty in the appliance log were manually verified by inspecting the daily EEC of the MEDAL system in question. If a mismatch was detected, the log was updated accordingly. Entries in the log dedicate one socket to one specific appliance. This does not include information about being turned on or being plugged in, but only serves as a booking.

6.2.5 Data Collection

BLOND aims for long-term continuous measurements, which requires some fault tolerance in the transmission layer; rendering wireless or mesh-based networks unfit for this task. The building is equipped with spare Ethernet connections in each room, which were used as a reliable transmission network to forward all data into a centralized storage

system. Ethernet, IPv4, TCP, and SSH all provide mechanisms to ensure data integrity and to automatically detect and retransmit faulty data with a very high probability.

BLOND-50 employed a pull-strategy, in which a single central server periodically pulled new data files from each measurement unit and moved them into a distributed storage system. CLEAR and MEDAL convert the raw data into HDF5 files and can buffer data for multiple hours or days if nobody collects new data. The central server only has to move data between systems and also buffers data for up to 24 hours in case the storage system is unavailable. This architecture decouples the various stages to allow for outages and planned maintenance. Buffer sizes and temporary storage devices were chosen carefully to maximize the allowed time before data loss occurs.

BLOND-250 uses a significant higher sampling rate, which renders a pull-strategy unusable due to memory and compute performance limits. Therefore, a push-strategy was used in which each measurement system directly sends raw data files (chunked) to the data center. The files are then converted and moved to the storage system by the server. Due to the higher sampling rate and file sizes, the available buffer time in each stage is also reduced.

CLEAR and MEDAL are built with the same software stack, which enables us to reuse large portions of the collection software and buffering strategies. Each measurement system is capable of buffering multiple gigabytes of raw data to a local storage device (SD-card or USB flash storage) in case of network failures or data center errors. This allows us to survive multiple days of data collection without any transmission capabilities. Upon reestablishing network connectivity, all buffered files are transmitted in bulk at a limited rate to prevent network congestion. Additional actions to further increase fault tolerance were implemented by using "RAM-first" buffering to keep I/O access to a minimum and reduce the risk of memory wear (write endurance of NOR/NAND flash memory). Although the underlying hardware of CLEAR and MEDAL are general-purpose computing devices, some low-level measurement tasks require real-time capabilities, which have been implemented by carefully choosing data structures, in-memory buffer sizes, and I/O access patterns to guarantee error-free data collection.

All networked devices are connected to the same Ethernet and share a synchronized clock via NTP. Two stratum-3 time servers are available on the same layer-2 Ethernet.

The internal system clock is connected to a dedicated real-time clock chip with a backup battery. A daemon process runs in the background to synchronize the system clock continuously; CLEAR uses *systemd-timesyncd* and MEDAL uses *ntpd*.

6.2.6 Known Issues

- Only user-operated appliances are measured as ground truth. Some static appliances (e.g., network switches and wireless access points) that are not user-operated are directly connected to the wall socket and can be considered as base load or static background in the CLEAR measurements (including MEDAL's own energy draw).
- MEDAL uses a unipolar ADC that can cause a slight DC-bias in the signal due to changes in the DC reference voltage. This can easily be accommodated for via proper signal calibration and filtering as part of a preprocessing stage.
- The appliance log was regularly updated and room-to-room checks were conducted. However, there could still be gaps in the log for unknown activity by students bringing their own devices for a short time period without entering the correct details into the appliance log.
- All measurement units were calibrated at the start of the BLOND data collection. Slight deviations in resistor value precision could cause a difference between CLEAR and MEDAL units connected to the same circuit.

6.2.7 Code Availability

We have implemented most of the data collection, technical validation, data processing, and utility tools in Python 3. The individual source files are available under the MIT license in the BLOND repository [103].

Due to the extensive amounts of data, processing is most reasonably done in a distributed and parallelized approach. We provide usage examples that can be scaled up and run in a distributed compute environment.

Software to convert and collect measurement data from a fleet of DAQ units is provided as it was used during the BLOND data collection. All steps in the Technical Validation section can be reproduced with the supplied scripts. The 1-second data summary was created from the raw measurements, and can be fully recreated.

6.3 Data Records

We provide raw voltage and current measurements of multiple circuits and appliances with high sampling rates. Additionally, we derived a data summary by computing various energy-related metrics into 1-second values.

6.3.1 BLOND Datasets

BLOND (Data Citation 1) contains two measurement series with different sampling rates:

- BLOND-50 with 50 kSps (aggregate) and 6.4 kSps (ground truth) over 213 days from September 30, 2016 to April 30, 2017
- BLOND-250 with 250 kSps (aggregate) and 50 kSps (ground truth) over 50 days from May 12, 2017 to June 30, 2017

Raw data and metadata are stored in HDF5 files that can be processed with a variety of open source and commercially available tools. Voltage and current samples of aggregate and ground truth measurements represent the waveform of the underlying electrical signal and are stored as-is from the sensor input. No permanent data cleaning or preprocessing was performed.

Metadata is embedded in each file and accessible as HDF5 attributes, either directly in the file root, or on a specific HDF5-dataset, see Table 6.3.1. Value types are either short integer, floating point, or ASCII-encoded byte strings. Generic information from HDF5 attributes matches to individual parts of the file name.

The structure of each dataset is grouped by date, and unit name into sub-directories: `BLOND-50/2017-03-25/medal-6/` contains 96 files of MEDAL-6 from March 25, 2017. Files of the BLOND-250 dataset can be found in the corresponding directory. This hierarchy is also available in the associated Metadata Record (ISA-Tab). Each file name contains the unit name, date, timestamp of the first sample in the file, a timezone offset, and a sequence number: `medal-6-2017-03-25T17-22-09.499845T-+0100-0016925.hdf5` contains data starting roughly at 17:22 on March 25, 2017, with a timezone offset of +1 hour, and this is the 16925th file in this series of MEDAL-6.

All timestamps and date information are "local time", therefore, special care must be given to the timezone offset during daylight saving time transitions: on 2016-10-30 at 3:00, DST ends (backwards 1h), on 2017-03-26 at 2:00 DST, starts (forward 1h). On December 31, 2016, a leap second was observed, which shifts back all file timestamps by one second.

Since files typically don't start at exactly 0:00 (midnight), the beginning and end of a day can be found in the previous or following file based on the sequence number.

Each measurement unit automatically splits data into chunks while the data acquisition continuous uninterrupted. The size of each chunk (number of samples per file) depends on the sampling rate and type of the unit, see Table 6.3.2. In total, BLOND consists of 945,919 files, amounting to 39 TB.

6.3.2 Appliance Log

The appliance log is available in two file formats: `appliance_log.{json,xlsx}`. Both files contain the same information and can be used interchangeably. The XLSX representation is human-readable and suitable for printing, whereas the JSON data is intended as input for data processing tools.

The JSON file was created from the XLSX data to provide a machine-readable format with the `appliance_log_json_converter.py` script. It contains a list of entries for each MEDAL unit. An entry consists of a timestamp and socket declarations (one for each socket): `class_name`, `appliance_name`, and `power`. Every appliance

instance from the appliance log is summarized in Table 6.2.2.

6.3.3 1-second Data Summary

Each dataset was augmented with a precomputed 1-second data summary: root-mean-square of voltage and current, real power, apparent power, power factor, and mains frequency. The resulting data was stored in one HDF5 file per day per measurement unit, covering all raw data files in each day folder (see `one_second_data_summary.py`). This provides quick and easy access to gain an overview of certain daily characteristics, without the need to download and process thousands of files. The daily data files are accompanied by a corresponding PDF showing selected plots of time series data, e.g., `summary-2017-03-25-medal-6.hdf5` and `summary-2017-03-25-medal-6.pdf`.

6.4 Technical Validation

All raw measurements included in the BLOND datasets are provided as-is, without any post-processing, cleaning, or filtering. This means the raw data must be calibrated and prepared before using the values as input to an evaluation (see the Usage Notes section). During the collection of BLOND, real world effects and noise are captured in the data. The measurement setup (environment) allowed us to have a data coverage of over 99.997% across 16 individual DAQ units during a combined period of 263 days. The missing data amounts to 2.5 hours of uncovered EEC.

An example of the captured waveform of voltage and current signals in a 3-phase power grid with CLEAR can be seen in Figure 6.4.1. A typical load profile (with individual contributions of each measurement system) over the course of multiple hours can be seen in Figure 6.4.2. The static base load was extracted from the day-to-day offset of the total consumption for each circuit. On small time scales (multiple hours), the base load can be assumed constant; day-to-day changes can be accounted for by calibrating the static offset during the night or on weekends (constant load with no occupants). The sum

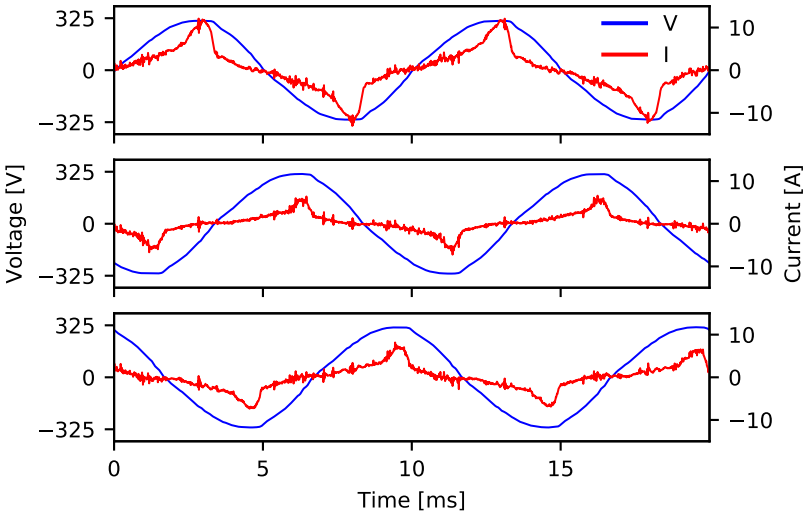


Figure 6.4.1: Waveforms of CLEAR circuits for voltage and current. The 3-phase power grid is characterized by a 120° phase shift between the circuits. The current consumption shows strong SMPS usage with sharp increases at each cycle apex. The voltage shows a typical sinusoidal waveform.

of all MEDAL units matches the measured EEC with CLEAR with reasonable accuracy, however, even small voltage drops or line noise can induce errors.

6.4.1 Data Collection Sanity Checks

While collecting data, each DAQ unit performs sanity checks for each new data chunk. This includes a DAQ continuity and generic transmission error checks. Such errors could be caused by internal queues filling up, full USB transfer queues, or interrupted communication between components. Each chunk contains a sequential identifier that can be validated to match its immediate predecessor and successor. In the case a mismatch is detected, the acquisition stops, reinitializes all components and retries. These identifiers are available in every HDF5 file for offline verification (*trigger ids*). No errors were detected during the collection of the BLOND datasets.

Complementing this low-level check, each newly created HDF5 file gets assigned an increasing sequence number, which marks a continuous uninterrupted series. BLOND-50 and BLOND-250 consist of a single long-term measurement series for each DAQ unit.

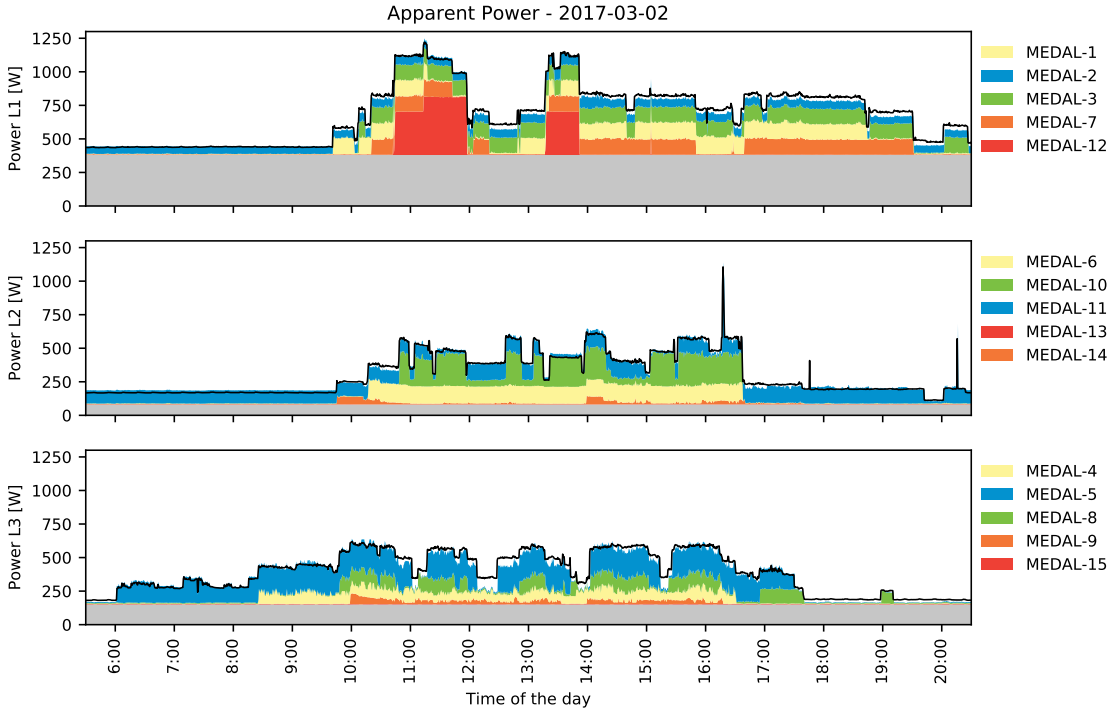


Figure 6.4.2: The load profile over multiple hours of the BLOND environment on 2017-03-02. Each stack plot shows the apparent power consumption of CLEAR (black line) with overlays for the contained MEDAL sub-meters (colored areas). See Table 6.2.1 for the circuit mapping. The individual steps (appliance events) match the overall load profile of the aggregate EEC. Each circuit shows a base load (gray area) which accounts for static background consumers. The 1-second data summary was downsampled to 30 seconds before plotting. Total consumption stays constant in the hours not shown. For visualization purposes, only the sum of all MEDAL sockets was plotted, however, the data contains an appliance-specific breakdown.

Only one interruption was detected: the CLEAR unit in BLOND-50 on 2016-10-18 (last sequence number: 0005172), due to a manual reboot after installing security updates. The gap covers only CLEAR measurements for 2 hours, 19 minutes, and 27 seconds. MEDAL measurements were not affected.

6.4.2 Sampling Rate Precision

Each data acquisition system collects data with a fixed sampling rate. An internal oscillator serves as a precise clock generator to trigger each analog-to-digital conversion. Depending on environmental factors, this process experiences a small unpredictable shift in speed. The actual average sampling rate was calculated based on the timestamps (with NTP precision) of the first and last data file over a 24 hour period (see `average_sampling_rate.py`) since all files contain the same amount of data (samples).

The average sampling rate per day shows an almost constant offset of less than 0.5%, while the actual variations are smaller than 1 Sps over the course of 24 hours, see Figure 6.4.3.

For BLOND-50, CLEAR has a nominal sampling rate of 50 000 Sps, mean of 49952.355, and a standard deviation of 0.057. All MEDAL units combined have a nominal sampling rate of 6400 Sps, mean of 6399.880, and a standard deviation of 0.013.

For BLOND-250, CLEAR has a nominal sampling rate of 250 000 Sps, mean of 248767.169, and a standard deviation of 0.084. All MEDAL units have a nominal sampling rate of 50 000 Sps, mean of 49984.059, and a standard deviation of 0.092.

6.4.3 Clock Synchronization

All timestamps used for marking samples and the beginning of new file chunks are derived from the system clock of the single-board PC in each measurement unit (CLEAR and MEDAL). The NTP precision as reported by `ntpq -c rl` is –20, yielding a theoretical timing accuracy of 0.95 µs. The real-world delay, offset, and jitter values of `ntpq -p`

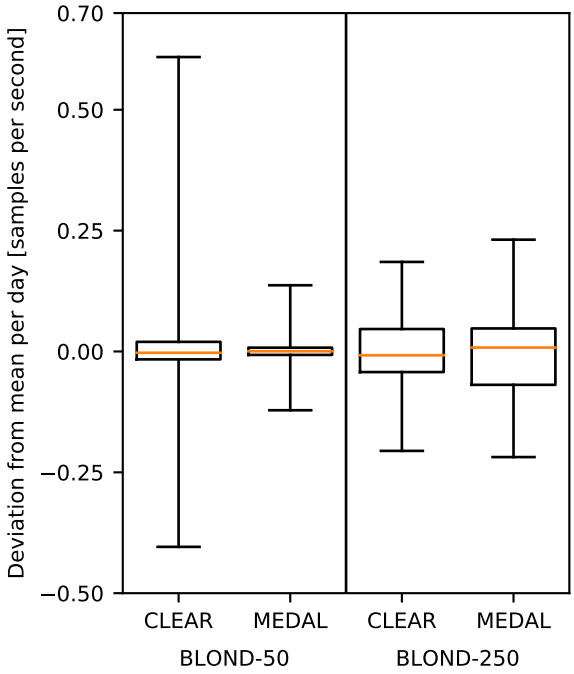


Figure 6.4.3: Boxplot of sampling rate precision per day. This boxplot shows the distribution of average sampling rates and its variation over 213 / 50 days for each measurement system. The whiskers depict the minimum and maximum. The mean was subtracted from each entry to compare only the daily fluctuations. The median is shown as line in the box.

show an average of 1 ms to 2 ms.

The sampled values from the ADC are buffered and transmitted via USB bulk transfers. The timestamp is added on the host device (single-board PC), which could add a short delay between the actual sampling time and the time it gets timestamped. Using USB data packets of 510 bytes containing 36 samples, the resulting average time jitter is 2.81 ms at 6.4 kSps, and 0.72 ms at 250 kSps. The preemption latency for CPU-bound tasks is defined with 6 ms (Linux kernel v4.4.21).

To verify these theoretical values, we used a space heater to generate a visible appliance switch-on event in socket 1 of MEDAL-1 on 2017-06-12 at timestamp 11:10:58. The difference of the sharp transient in the CLEAR to the MEDAL time series data was measured with 6.8 ms, which is within our estimation (see `clock_synchronization.py`). This allows us to synchronize multiple data streams with sub-cycle precision on a 50 Hz mains.

6.4.4 Per-File Data Checks

The correctness of the sampled voltage and current signals was validated by analyzing each data file with 15 individual checks to assert various metrics and raw data streams (see `per_file_data_checks.py`):

Dataset length is the amount of samples per signal in a given file. This is defined by the sampling rate and the file size used for chunking. If the data acquisition is briefly interrupted, stopped, or the file got truncated, the expected length does not match. In BLOND-50, 4 individual files were found that failed this check. The data collection seems to have continued, however, these files were either corrupted during transmission or the storage system failed to persist the data. The files in question have a length of 0 bytes and are not valid HDF5 files. MEDAL 6, 13, and 14 at sequence number 0016123, as well as CLEAR at 0043125 are affected.

Mains frequency is expected to be 50 Hz and should only deviate slightly. The mains frequency was computed using Fast Fourier Transform and selecting the strongest

bin. Erroneous readings would indicate a collapsing power grid or a malfunctioning ADC trigger input (sampling rate). No such errors were found.

Voltage and current root-mean-squared expected values are based on the measurement unit capabilities and can be used as a sanity check to check against unexpected high or low values. Voltage values must be close to 230 V_{RMS}, have an almost zero absolute mean, and the crest factor should be around 1.41. Current values must be below the rated measurement limit of the DAQ unit, have an almost zero absolute mean, and the crest factor must be greater than 1.2. No such errors were found.

Raw voltage value range and bandwidth is defined by the ADC bit-resolution. A 16-bit ADC can yield up to 65536 different measurement values. If the measurement range was calibrated or configured incorrectly, not all values would be used, resulting in degraded accuracy. We checked how many unique values are present in each signal per file and compared the maximum to the minimum, which must be within certain threshold limits. No such errors were found.

Voltage bandwidth is defined by the power grid; for BLOND, we expect a nominal voltage of ± 324 V_{peak}. Including a certain margin of deviation that is allowed during normal operation of the grid, we checked the minimum and maximum voltage values to be within certain threshold limits. No such errors were found.

Flat regions are defined as intervals with identical consecutive values. During initial experiments and prototyping, we detected a malfunction in one of the traces on a PCB. This led to a permanently pulled-low bit on a data bus. The DAQ unit therefore only received the same value over and over again. We checked for flat regions longer than a certain threshold by applying a linear convolution with a filter kernel (length of one mains period) to each signal per file. No such errors were found.

6.5 Usage Notes

The BLOND data files are provided in HDF5 format, which is usable in most scientific computing packages, e.g., Python (h5py/numpy/scipy), MATLAB (h5read), R (rhdf5),

Mathematica (Import), and NILMTK [61]. The metadata (HDF5 attributes) is documented in Table 6.3.1. Each HDF5 dataset was created with the following filters: gzip compression (reduces file size), shuffle (improves compression ratio), and Fletcher (adds checksums to detect data corruption). HDF5 offers a multitude of different filters with potentially better compression, however, we wanted to retain compatibility with most software packages, which typically lack support for 3rd-party filters.

Multiple example use cases for data handling and calibration can be found in the provided source code. We recommend performing a mean-offset normalization for each mains cycle before multiplying the signal with the calibration factor to remove any unwanted DC-biases in MEDAL signals. We deliberately did not clean, back-fill, or strip any of the data. This allows us to retain and extract as much information as possible from seemingly "empty" data (background noise, sampling artifacts, derived data).

Table 6.2.2: List of appliances observed in the BLOND dataset. This list was extracted from the appliance log and contains all devices used in the BLOND environment. A class label was assigned to group similar appliances. The manufacturer, type, and power information was taken from an attached name plate (if available) or the suppliers datasheet.

Class	Manufacturer	Type	Power	Count
Battery Charger	Kraftmax	BC4000 Pro	18 W	1
	DJI	Phantom 3	100 W	1
Daylight	Philips	HF3430	10 W	1
Desktop Computer	*generic*	Intel Xeon E5-1640 v4, NVIDIA TITAN X	1200 W	1
	Dell	OptiPlex 7040	65 W	1
	Dell	OptiPlex 9020	65 W	2
	Dell	T3600	635 W	1
Dev Board	FPGA	Xilinx ML505	30 W	1
	FPGA	Tegra Jetson	90 W	1
	MEDAL	Prototype	5 W	1
Electric Toothbrush	*generic*	inductive charging	5 W	1
Fan	Eurom	VS 16	45 W	1
	VOV	VTS-1641	50 W	1
Kettle	Clatronic	WK3445	2000 W	1
	Severin	WK3364	1800 W	1
Laptop Computer	Apple	MacBook Air 13" Early-2014	45 W	1
	Apple	MacBook Pro 13" Mid-2014	60 W	3
	Apple	MacBook Pro 15" Mid-2014	85 W	2
	ASUS	N750JV	120 W	1
	Dell	E6540	130 W	1
	Dell	XPS13	45 W	1
	Lenovo	Carbon X1	90 W	1
	Lenovo	B560	65 W	1
	Lenovo	L540	90 W	1
	Lenovo	T420	90 W	1
	Lenovo	T450	65 W	1
	Lenovo	T530	90 W	1
	Lenovo	X230 i7	65 W	1
	Lenovo	X230 i5	170 W	1
	Schenker	W502	180 W	1
	Sony	Vaio VGN FW54M	92 W	1
	generic	SMPS, 19V	100 W	1
Monitor	Dell	P2210	22 W	1
	Dell	U2711	133 W	6
	Dell	U2713Hb	130 W	8
	Dell	UP2716D	45 W	2
	Fujitsu-Siemens	P17-1	36 W	1
Multi-Tool	Mannesmann	M92577	135 W	1
Paper Shredder	HSM	Shredstar	250 W	1
Printer	HP	LaserJet Pro 400	425 W	1
Projector	Epson	EB-65950	450 W	1
Screen Motor	Projecta	DC 485	210 W	1
Space Heater	Heller	ASY 1507	1500 W	1
USB Charger	*generic*	single USB power adapter	10 W	2
	inateck	UC2001	15 W	1
	Aukru	BS-522	20 W	1
	Apple	MD836ZM EU	12 W	1
	Apple	MD813ZM EU	5 W	2
	Chromecast	single USB power adapter	10 W	1
	Hama	00091321	10 W	1
	Samsung	Travel	10 W	3
	Sony	single USB power adapter	10 W	1
	Sony Ericsson	EP 800	10 W	1

Table 6.3.1: HDF5 dataset file metadata. Each attribute is accessible via a HDF5-attribute-path. Values are provided in base units (Volt, Ampere, Hertz). Some attributes are only available in the 1-second data summary.

Path	Attribute	Description
/	name	Name of the measurement unit
/	year	Year of the first sample
/	month	Month of the first sample
/	day	Day of the first sample
/	hours	Hours of the first sample
/	minutes	Minutes of the first sample
/	seconds	Seconds of the first sample
/	microseconds	Microseconds of the first sample
/	sequence	Sequence number in this series
/	timezone	Timezone offset (daylight saving time)
/	frequency	Nominal sampling rate in Hz
/	first_trigger_id	Internal trigger number to detect gaps
/	last_trigger_id	Internal trigger number to detect gaps
/<dataset>	calibration_factor	Multiplication factor for calibration
/<dataset>	removed_offset	Removed DC-offset of the signal
/	average_frequency	Average sampling rate over 24h
/	delay_after_midnight	Delay in seconds after 00:00

Table 6.3.2: File chunking and length. BLOND-50 and BLOND-250 use different file sizes to chunk the continuous data stream. The size depends on the available computing resources in each DAQ unit and the configured sampling rate. The final size of the HDF5 only depends on the number of samples and the achievable compression ratio.

Dataset	Unit Type	Sampling Rate	File Length	Samples
BLOND-50	CLEAR	50 kSps	5 min	15,000,000
	MEDAL	6.4 kSps	15 min	5,760,000
BLOND-250	CLEAR	250 kSps	2 min	30,000,000
	MEDAL	50 kSps	2 min	6,000,000

7

Waveform Signal Entropy and Compression Study of Whole-Building Energy Datasets

Home and building automation promise many benefits for the occupants and power utilities. From increased user comfort levels to demand response and lower electricity costs, Smart Homes offer a variety of assistance and informational gains. Internet of Things, a combination of sensors and actuators, can be intelligently controlled based on sensor data or external triggers. Power monitoring and smart metering are a key step to fulfill these promises. The influx of renewable energies and the increased momentum of changes in the power grid and its operations are a main driving factor for further research in this area.

Non-intrusive load monitoring (NILM) can be one solution to identify and disaggregate power consumers (appliances) from a single-point measurement in the building. Utilizing a centralized data acquisition system saves costs for hardware and installation in the electrical circuits under observation. The NILM community heavily relies on long-term measurement data, in the form of public datasets, to craft new algorithms, train models, and evaluate their accuracy on per-appliance energy consumption or appliance identification. In recent years these datasets grew significantly in size and sampling characteristics

(temporal and amplitude resolution). Collecting, distributing, and managing large-scale data storage facilities is an ongoing research topic [104, 105] and strongly depends on the environment and systems architecture.

High sampling rates are particularly interesting for NILM to extract waveform information from voltage and current signals [5]. Early datasets targeted at load disaggregation and appliance identification started with under 2 GiB [22], whereas recently published datasets reach nearly 100 TiB of raw data [25]. Working with such quantities requires specialized storage and processing techniques which can be costly and maintenance-heavy. Optimizing infrastructure costs for storage is part of ongoing research [106, 107].

The data quality requirements typically define a fixed sampling rate and bit-resolution for a static environment. Removing or augmenting measurements might impede further research, therefore no filtering or preprocessing steps are performed before releasing the data.

Data compression techniques can be classified as lossy or lossless [108]. Lossy algorithms allow for some margin of error when encoding the data and typically give a metric for the remaining accuracy or lost precision. For comparison, most audio, image, and video compression algorithms remove information not detectible by a human ear or eye. This allows for a data rate reduction in areas of the signal a user can't detect or has a reduced resolution due to a typical human physiology. Depending on the targeted use case, certain aspects of the input signal are considered unimportant and might be not reconstructable. Encoding only the amplitude and frequency of the signal can lead to vast space savings, assuming phase alignment, harmonics, or other signal characteristics are not required for future analysis. On the contrary, lossless encoding schemes guarantee a 1:1 representation of all measurement data with a reversible data transformation. If the intended use case or audience for a given dataset is not known or is very diverse in their requirements, only lossless compression can be applied to keep all data accessible for future use. Recent works pointed out an imbalance in the amount of research on steady-state versus waveform-based compression of electricity signals [109].

Further consideration must be given to communication bandwidth (transmission to a remote endpoint) and in-memory processing (SIMD computation). The efficient use of

network channels can be a key requirement for real-time monitoring of streaming data. In the case of one-time transfers (or burst transmissions), chunking is used to split large datasets into more manageable (smaller) files. However, choosing a maximum file size depends on the available memory and CPU (instruction set and cache size). Distributing large datasets as a single file creates an unnecessary burden for researchers and required infrastructure.

A suitable file format must be considered for raw data storage, as well as easy access to metadata, such as calibration factors, timestamps, and identifier tags. None of the existing datasets (NILM or related datasets with high sampling rates) share a common file format, chunk size, or signal sampling distribution. This heterogeneity makes it difficult to apply algorithms and evaluation pipelines on more than one dataset. Therefore, researchers working with multiple datasets have to implement custom importer and converter stages, which can be time-consuming and error-prone.

This work provides an in-depth analysis of public whole-building datasets, and gives a comprehensive evaluation of best-practice storage techniques and signal conditioning in the context of energy data collection. The key contributions of this work are:

1. A numerical analysis of signal entropy and measurement calibration of public whole-building energy datasets by evaluating all signal channels with respect to their available resolution and sample distribution over the entire measurement period. The resulting entropy metrics further motivate our contributions and the need for a well-calibrated measurement system.

2. An exhaustive benchmark of storage requirements and potential space savings with a comprehensive collection of 365 file formats, lossless compression techniques, and reversible data transformations. We re-encode and normalize data from all datasets to evaluate the effect of compression. We present the best-performing combinations and their overall space savings. The full ranking can be used to select the optimal file format and compression for offline storage of large long-term energy datasets.

3. A full-scale evaluation of increasingly larger data chunks per file and their final compression ratio. The dependency between input size and achievable compression

ratio is evaluated up to 3072 MiB per file. The results provide an evidence-based guideline for future selection of chunk sizes and possible environmental factors for consideration.

We give an in-depth evaluation of file formats and signal characteristics that directly affect storage, encoding, and compression of such data. Each of the analyzed datasets was created with a dedicated set of requirements, therefore, a single best option does not exist. However, with this study, we want to help the community to better understand the fundamental causes of compression performance in the field of waveform-based whole-building energy datasets. We provide a definition of measurement calibration and its effects on the storage requirements based on signal entropy. Published datasets are self-contained and final, which allows us to prioritize the compression ratio and achievable space saving over other common compression metrics (CPU load, throughput, or latency). We define the achievable space saving and compression ratio as the only criterion when dealing with large (offline) datasets.

This chapter is structured as follows: We describe the evaluated datasets in Section 7.1, which are then used in the experiments for entropy analysis in Sections 7.2, data representation in Section 7.3, and chunk size impact in Section 7.4. Finally, we present experimental results in Section 7.5.

7.1 Evaluated Datasets

While there is a vast pool of smart meter datasets [110], i.e., low sampling rates of measurements every 1 s, 15 min, or 1 h, a majority of the underlying information is already lost (signal waveform) [15]. The raw signals are aggregated into single root-mean-squared voltage and current readings, frequency spectrums, or other metrics accumulated over the last measurement interval. This can be already classified as a type of lossy compression. For some use cases, this data source is sufficient to work with, while other fields require high sampling rates to extract more information from the signals.

All following experiments and evaluations were performed on publicly accessible datasets:

The Reference Energy Disaggregation Data Set (REDD [22]), *Building-Level fUlly-labeled dataset for Electricity Disaggregation* (BLUED [23]), *UK Domestic Appliance-Level Electricity dataset* (UK-DALE [24]), and the *Building-Level Office eNvironment Dataset* (BLOND [25]). We will refer to these datasets by their established acronyms: REDD, BLUED, UK-DALE, and BLOND. Based on the energy dataset survey [110], these are all datasets of long-term continuous measurements with voltage and current waveforms from selected buildings or households. The data acquisition systems and data types are comparable to warrant their use in this context, see Table 7.1.1.

Measurement systems and their analog-to-digital converters (ADC) always output a unit-less integer number, either between $[0, 2^{bits})$ for unipolar ADCs or $[-2^{bits-1}, 2^{bits-1})$ for bipolar ADCs. During setup and calibration, a common factor is determined to convert raw values into a voltage or current reading. Some datasets publish raw values and the corresponding calibration factors, while others publish directly Volt- and Ampere-based readings as float values. Datasets only available as floating-point values are converted back into their original integer representation without loss of precision by reversing the calibration step from the analog-to-digital converter for each channel:

$$measurement_i = ADC_i \cdot calibration_{channel}$$
$$[Volts] = [steps] \cdot [{}^{Volt}/_{steps}]$$
$$[Ampere] = [steps] \cdot [{}^{Ampere}/_{step}]$$

Each of the mentioned datasets was published in a different (compressed) file format and encoding scheme. To allow for comparisons between these datasets, we decompressed, normalized, and re-encoded all data before analyzing them (raw binary encoding).

From REDD, we used the entire available *High Frequency Raw Data*: *house_3* and *house_5*, each with 3 channels: *current_1*, *current_2*, and *voltage*. The custom file format encodes a single channel per file. In total, 1.4 GiB of raw data from 126 files were used.

From BLUED, we used all available waveform data (1 location, 16 sub-datasets) and 3 channels: *current_a*, *current_b*, *voltage*. The CSV-like text files contain voltage and two current channels and a dedicated measurement timestamp. In total, 41.1 GiB of raw data from 6430 files were used.

Table 7.1.1: Overview of evaluated datasets: long-term continuous measurements containing raw voltage and current waveforms.

Dataset	Current Channels	Voltage Channels	Sampling Rate	Values
REDD	2	1	15 kHz	24-bit
BLUED	2	1	12 kHz	16-bit
UK-DALE	1	1	16 kHz	24-bit
BLOND-50	3	3	50 kHz	16-bit
BLOND-250	3	3	250 kHz	16-bit

From UK-DALE, we selected *house_1* from the most recent release (*UK-DALE-2017-16kHz*, the longest continuous recording). The compressed FLAC files contain 2 channels: *current* and *voltage*. In total, 6259.1 GiB of raw data from 19491 files were used.

From BLOND, we selected the aggregated mains data of both sub-datasets: BLOND-50 and BLOND-250. The HDF5 files with gzip compression contain 6 channels: *current{1-3}* and *voltage{1-3}*. In total, 10 246.7 GiB of raw data from 61125 files of BLOND-50, and 11 899.0 GiB of raw data from 35490 files of BLOND-250 were used.

The data acquisition systems (DAQ) of all datasets produce a linear pulse-code modulated (LPCM) stream. The analog signals are sampled in uniform intervals and converted to digital values (Figure 7.2.1). The quantization levels are distributed linearly in a fixed measurement range which requires a signal conditioning step in the DAQ system. ADCs typically cannot directly measure mains voltage and require a step-down converter or measurement probe. Mains current signals need to be converted into a proportional voltage.

7.2 Entropy Analysis

DAQ units provide a way to collect digital values from analog systems. As such, the quality of the data depends strongly on the correct calibration and selection of measurement equipment. Mains electricity signals are typically not compatible with modern digital systems, requiring an indirect measurement through step-down transformers or other

metrics. Mains voltage can vary by up to ±10% during normal operation of the grid [111, 112], making it necessary to design the measurement range with a safety margin. The expected signal, plus any margin for spikes, should be equally distributed on the available ADC resolution range. Leaving large areas of the available value range unused can be prevented by carefully selecting input characteristics and signal conditioning (step-down calibration). A rule of thumb for range calibration is that the expected signal should occupy 80-90%, leaving enough bandwidth for unexpected measurements. Input signals larger than the measurement range get recorded as the minimum/maximum value. Grossly exceeding the rated input signal level could damage the ADC, unless a dedicated signal conditioning and protection is employed.

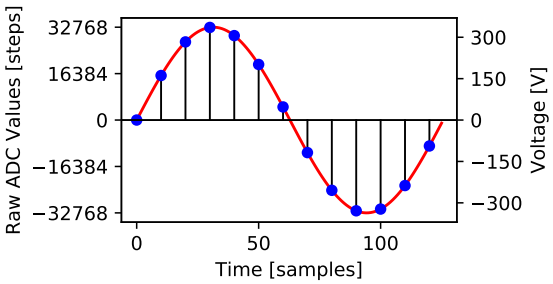


Figure 7.2.1: Linear pulse-code modulation stream of a sinusoidal waveform sampled with a 16-bit ADC. The waveform corresponds to a 230 V mains voltage signal.

We extracted the probability mass function (PMF) of all evaluated datasets for the full bit-range (16- or 24-bit). The value histogram is a structure mapping each possible measurement value (integer) to the number of times this value was recorded. Ideally, the region between the lowest and highest value contains a continuous value range without gaps. However, the quantization level (step size) could cause a mismatch and results in skipped values. We then normalize this histogram to obtain the PMF and compute the signal entropy per channel, which gives an estimation of the actual information contained in the raw data and provides a lower bound for the achievable compression ratio based on the Kolmogorov complexity.

$$X = \left\{ -2^{bits-1}, ..., 0, ..., 2^{bits-1} - 1 \right\}$$
$$hist = histogram(dataset, X)$$
$$f_X = \frac{hist}{\sum_{x \in X} hist[x]}$$

$$\forall x \in X \text{ where } f_X(x) = 0 \; : \; f_X(x) = 1$$
$$H(x) = -\sum_{x \in X} f_X(x) \cdot log_2\left(f_X(x)\right)$$

Each dataset is split into multiple files, making it necessary to merge all histograms into a total result at the end of the computing run. Since all histograms can be combined with a simple summation, the process can be parallelized and computed without any particular order. Computing and merging all histograms is, therefore, best accomplished in a distributed compute cluster with multiple nodes or similar environments.

7.3 Data Representation

Choosing a suitable file format for potentially large datasets involves multiple tradeoffs and decisions, including supported platforms, scientific computing frameworks, metadata, error correction, compression, and chunking. The available choices for data representation can range from CSV data (ASCII-parsable) to binary file formats and custom encoding schemes. From the energy dataset survey and the evaluated datasets, it can be noted, that every dataset uses a different file format, encoding scheme, and optionally compression.

Publishing and distributing large datasets requires storage systems capable of providing long-term archives of scientific measurement data. Lossless compression helps to minimize storage costs and distribution efforts. At the same time, other researchers accessing the data benefit from smaller files and shorter access times to download the data.

Electricity signals (current and voltage) contain a repetitive waveform with some form of distortion depending on the load. In an ideal power grid, the voltage would follow a perfect sinusoidal waveform without any offset or error. This would allow us to accurately predict the next voltage measurement. However, constant fluctuations in the supply and demand cause the signals to deviate. The fact that each signal is primarily continuous (without sudden jumps) can be beneficial to compression algorithms.

A delta encoding scheme only stores the numerical difference of neighboring elements in

a time-series measurement vector. This can be useful for slow-changing signals because the difference of a signal might require less bytes to encode than the absolute value:

$$\forall i \in \{1 \dots n\} \; : \; d_i = v_i - v_{i-1}$$
$$d_0 = v_0$$

We compare the original data representation (format, compression, encoding) of each dataset, reformat them into various file formats, and evaluate their storage saving based on a comprehensive list of lossless compression algorithms. This involves encoding raw data in a more suitable representation to compare their compressed size: $CS = compressed\_size/original\_size * 100\%$, and the resulting space saving: $SS = 100\% - CS$. We define the main goal of reducing the overall required storage space for each dataset, and deliberately do not consider compression or decompression speed. The performance characteristics (throughput and speed) are well known for individual compression techniques [113] and are of minor importance in the case of large static datasets which require only a single compression step before distribution. Performance metrics are important when dealing with repeated compression of raw data, which is not the case for static energy datasets. Repeated decompression is however relevant because researchers might want to read and parse the files over and over again while analyzing them (if in-memory processing is not feasible). As noted in [113], decompression speed and throughput is typically not a performance bottleneck in data analytics tasks.

Building a novel data compression scheme for energy data is counter-productive, since most scientific computing frameworks lack support and the idea suffers from the "not invented here" and "yet another standard" problematic, both common anti-patterns in the field of engineering when developing new solutions, despite existing suitable approaches [22, 53, 56]. Therefore, a key requirement is that each file format must be supported in common scientific computing systems to read (and possibly write) data files.

We selected four format types: raw binary, HDF5 (data model and file format for storing and managing data), Zarr (chunked, compressed, N-dimensional arrays), and audio-based PCM containers.

Raw binary formats provide a baseline for comparison. All samples are encoded as integer

values (16-bit or 24-bit) and are compressed with a general-purpose compressor: zlib/gzip, LZMA, bzip2, and zstd, all with various parameter values. The input for each compressor is either raw-integer or variable-length encoded data (LEB128S [114]), which is serialized either row- or column-based from all channels (interweaving). The LEB128S encoding is additionally evaluated with delta encoding of the input.

The Hierarchical Data Format 5 (HDF5) [62] provides structured metadata and data storage, data transformations, and libraries for most scientific computing frameworks. All data is organized in natively-typed arrays (multi-dimensional matrices) with various filters for data compression, checksumming, and other reversible transformations before storing the data to a file. The API transparently reverses these transformations and compression filters while reading data. HDF5 is popular in the scientific community and used for various big-data-type applications [115, 116, 117, 118]. The public registry for HDF5 filters[1] currently lists 21 data transformations, most of them compression-related. Each HDF5 file is evaluated with and without the shuffle filter, zlib/gzip, lzf, MAFISC [119] with LZMA, szip [120], Bitshuffle [121] with LZ4, zstd, and the full Blosc [63] compression suite, again all with various parameter values.

Zarr [122] organizes all data in a filesystem-like structure, which can be archived as a single zip-archive file or as tree-structure in the filesystem. Each channel is stored as a separate array (data stream) with optional chunk-based compression via zlib/gzip, LZMA, bzip2, or Blosc (with shuffle, Bitshuffle, or no-shuffle filter), again all with various parameter values. Each Zarr file is additionally evaluated with a delta filter to reduce the value range.

Audio-based formats use LPCM-type data encoding (`PCM16` or `PCM24`) with a fixed precision and sampling rate. All channels are encoded into a single container using lossless compression formats: FLAC [123], ALAC [124], and WavPack [125]. These formats do not provide tune-able parameters.

Calibration factors, timestamps, and labels can augment the raw data in a single file while providing a unified API for accessing data and metadata. Raw binary formats lack this type of integrated support and require additional tooling and encoding schemes for

[1] `https://support.hdfgroup.org/services/filters.html`

metadata. Audio-based formats require a container format to store metadata, typically designed for the needs of the music and entertainment industry. Out of these formats, only HDF5 and Zarr provide support for encoding and storing arbitrary metadata objects (complex types or matrices) together with measurement data.

Most audio-based formats support at most 8 signal channels, while general-purpose formats such as HDF5 and Zarr have no restrictions on the total number of channels per file. The sampling rate can also be a limiting factor: FLAC supports at most 655.35 kHz and ALAC only 384 kHz. ADC resolution (bit depth) is mostly bound by existing technological limitations and will not exceed 32-bit in the foreseeable future. While these constraints are within the requirements for all datasets under evaluation, they need to be considered for future dataset collection and the design of measurement systems.

In total, we encoded the evaluated datasets with 365 different data representation formats: 54 raw, 264 HDF5-based, 44 Zarr-based, and 3 audio-based and gathered their per-file compression size as a benchmark. The full analysis was performed in a distributed computing environment and consumed approx. 1,176,000 CPU-core-hours (dual Intel Xeon E5-2630v3 machines with 128 GiB RAM and 10 Gibit Ethernet interfaces).

7.4 Chunk Size Impact

Each dataset is provided in equally-sized files, typically based on measurement duration. Working with a single large file can be cumbersome due to main memory restriction or available local storage space. Assuming a typical desktop computer, with 8 GiB of main memory, is used for processing, a single file from a dataset must be fully loaded into memory before any computation can be done. Depending on the analysis and algorithms, multiple copies might be required for intermediary results and temporary copies. This means the main memory size is an upper bound for the maximum feasible chunk size.

Some file formats and data types support internal chunking or streamed data access, in which data can be read into memory sequentially or random-access. In such environments other factors will limit the usable chunk size, such as file system capabilities, network-attached storage, or other operating system limitations.

The evaluated datasets are distributed with the following chunk sizes of raw data: REDD: 11.4 MiB or 4 min, BLUED: 6.6 MiB or 1.65 min, UK-DALE: 329.2 MiB or 60 min, BLOND-50: 171.7 MiB or 5 min, BLOND-250: 343.3 MiB or 2 min. Measurement duration and file size are not strictly linked, causing a slight variation in file sizes across the entire measurement period of each dataset. Observed real-world time does not affect any of the compression algorithms under test and is therefore omitted. The sampling rate and channel count directly affects the data rate (bytes per time unit) and explains the non-uniform chunk sizes mentioned for each dataset.

We compare the best-performing data representation formats of each dataset from the previous experiment, benchmark them with different chunk sizes, and estimate their effect on the overall compression ratio. For this evaluation, we define the compression ratio as $CR = original\_size/compressed\_size$. The chunk sizes range from 1, 2, 4, 8, 16, 32, 64, 128 MiB, and then continue in steps of 128 MiB up to 3072 MiB. To reduce the required computational effort, we greedily consume data from the first available dataset file, until the predefined chunk limit is fulfilled. The chunk size is determined using the number of samples (across all channels) and their integer byte count (2 or 3 bytes); only full samples for all channels are included in a chunk.

7.5 Experimental Results

7.5.1 Entropy Analysis

Entropy is based on the probability for a given measurement (signal value). The histogram of an entire measurement channel shows the number of times a single measurement value was seen in the dataset (Figure 7.5.2). The plots show the raw measurement bandwidth in ADC value on the x-axis and a logarithmic y-axis for the number of occurrences of each value. The raw ADC values are bipolar and centered on 0: −32768 … 32767 for BLUED, BLOND-50, and BLOND-250; −8388608 … 8388607 for REDD and UK-DALE.

The voltage histogram shows a distinctive sinusoidal distribution (peaks at minimum and maximum values). The current histogram would show a similar distribution if the

power draw is constant (pure-linear or resistive loads), however, multiple levels of current values can be observed, indicating high activity and fluctuations. REDD and BLUED (Figures 7.5.2a and 7.5.2b) show a center-biased distribution, indicating a sub-optimal calibration performance and unused measurement bandwidth. UK-DALE, BLOND-50, and BLOND-250 (Figures 7.5.2c, 7.5.2d, 7.5.2e) show a wide range of highly used values, with the voltage channels utilizing around 90% of the available bandwidth.

REDD and BLUED use only a small percentage of the available range, indicating a low entropy based on the used data type. UK-DALE utilizes a reasonable slice, while BLOND covers almost the entire possible range (Table 7.5.1). Assuming a well-calibrated data acquisition system, the expected percentage should reflect the expected measurement values. Low range usage (REDD, BLUED) leads to lost precision which would have been freely available with the given hardware, whereas high usage (UK-DALE, BLOND) means almost all available measurement precision is reflected in the raw data. Some datasets utilize 100% of the available measurement range, while REDD only uses 5%. A high range utilization does not result in a equally high usage, as the histogram can contain gaps (ADC values with 0 occurrences in the datasets).

7.5.2 Data Representation

The evaluation compares the compressed size (CS, final file size after compression and file format encapsulation in percent of uncompressed size) of 365 data representation formats. For brevity reasons, only the 30 best-performing formats are shown in Figure 7.5.1. Each of the 365 data representation was tested on all datasets. The following evaluation and benchmark uses the raw data from each dataset as described in Section 7.1. In total, raw data with 27.8 TiB was re-encoded 365 times.

HDF5 and Zarr are general-purpose file formats for numerical data with a broad support in scientific computing frameworks. As such, they only support 16-bit and 32-bit integer values, which causes a 1-byte overhead for REDD and UK-DALE. The baseline used for comparison is a raw concatenated byte string with dataset-native data types (16-bit and 24-bit). This allows us to obtain comparable evaluation results, while other published benchmarks compared ASCII-like encodings against binary representations, skewing the

Table 7.5.1: Entropy analysis of whole-building energy datasets with high sampling rates. The amount of unique measurement values for each channel is extracted, which corresponds to a usage percentage over the available measurement resolution. The lowest and highest observed value is used to give determine the observed range.

Dataset	Channel	Values	Usage	Range	H(x)
REDD (24-bit)	current_1	87713	1%	4%	14.3
	current_2	85989	1%	5%	14.9
	voltage	2925155	17%	18%	21.1
BLUED (16-bit)	current_a	5855	9%	10%	7.8
	current_b	7684	12%	13%	9.7
	voltage	11302	17%	18%	13.2
UK-DALE (24-bit)	current	6981612	42%	81%	19.0
	voltage	15135594	90%	100%	23.2
BLOND-50 (16-bit)	current1	51122	78%	100%	12.6
	current2	49355	75%	100%	11.2
	current3	48658	74%	100%	11.3
	voltage1	58396	89%	92%	15.3
	voltage2	57975	88%	91%	15.4
	voltage3	59596	91%	95%	15.4
BLOND-250 (16-bit)	current1	52721	80%	100%	12.4
	current2	51802	79%	100%	10.8
	current3	50989	78%	100%	11.6
	voltage1	58488	89%	91%	15.3
	voltage2	57912	88%	92%	15.4
	voltage3	59742	91%	94%	15.4

results significantly.

Overall, it can be noted that all three audio-based formats performed well, given their inherent targeted nature of compressing waveforms with high temporal resolution. ALAC and FLAC achieved the highest overall CS across all datasets, followed by HDF5+MAFISC and HDF5+zstd, which can overcome the 1-byte overhead. Although the general-purpose compressors and their individual data representation formats were intended to serve as a baseline for comparison of the more advanced schemes (HDF5, Zarr, and audio-based), one can conclude that even plain bzip2 or LZMA compression can achieve comparable compression results. A tradeoff to consider is the lack of metadata and internal structure, which might cause additional data handling overhead as easy-to-use import and parsing tools are not available. Variable-length encoding using LEB128S is a suitable input for the bzip2 and LZMA compressors when combined with a column-based storage format. Delta encoding resulted in comparably good CS in certain combinations.

Some datasets are inherently more compressible than others. This is a result of the entropy analysis and can be observed in the data representation evaluation as well. Compressing BLUED consistently yields smaller file sizes with most compressors than any other dataset. The benchmark shows that higher entropy correlates strongly with higher CS per dataset.

While the majority of tested data representation formats achieves a data reduction, compared to the baseline, some formats are counter-productive and generate a larger output (CS over 100%). This behavior affects most HDF5- and Zarr-based formats, because of the 1-byte overhead (depending on the used compressor).

Choosing the best-performing data representation for each dataset, the following SS can be achieved when applied to all data files as compared against the raw binary encoding: REDD: 48.3% or 0.7 GiB, BLUED: 73.0% or 30.0 GiB, UK-DALE: 40.5% or 2534.1 GiB, BLOND-50: 51.3% or 5252.3 GiB, BLOND-250: 55.4% or 6590.8 GiB. It can be noted that REDD, UK-DALE, and both BLOND datasets perform at around 50-60% of CS, while BLUED shows a significantly smaller CS of below 30% CS, due to it's very low signal entropy (Table 7.5.1). Variable-length encoding (LEB128S) and Delta encoding yield the largest space saving for such types of data (REDD and BLUED).

Two out of the five evaluated datasets (REDD and BLUED) showed the highest space savings with a general-purpose compressor (bzip2) and variable-length encoding. ALAC and HDF5+MAFISC performed best on UK-DALE, BLOND-50, and BLOND-250, given their higher signal entropy and value range utilization.

When comparing the raw space savings against the actually published dataset, which typically is already compressed, we can achieve additional space savings: REDD: 61.2% or 1.1 GiB, BLUED: 96.4% or 295.5 GiB, UK-DALE: -1.3% or −49.1 GiB, BLOND-50: 23.3% or 1519.7 GiB, BLOND-250: 26.0% or 1867.9 GiB. All datasets show space savings, except for UK-DALE, which shows an insignificant increase in the overall dataset size. This means the originally published FLAC files are already compressed to a high extent; this is supported by Figure 7.5.1, showing FLAC among the highest ranking formats in this study. While an absolute space saving of 1.1 GiB for REDD might be insignificant in most use cases (desktop computing and data center), a more compelling reduction in storage space of up to 1867.9 GiB for BLOND-250 can be substantially beneficial.

7.5.3 Chunk Size Impact

The chunk size evaluation (Figure 7.5.3) contains the averaged CR per chunk size for all datasets except REDD, as it only contains 1438.4 MiB of data and was therefore omitted.

The evaluated chunk size range starts with very small chunks, which would not be recommended for large datasets because of the increased handling and container overhead. As such, chunk sizes starting with 128 MiB can be considered as viable storage strategy. The resulting CR ramps up quickly for most formats until it levels off between 32 MiB to 64 MiB. Above this mark, no significant improvement in CR can be achieved by increasing the chunk size. Some file formats even show a slight linear decrease in CR with very large chunk sizes (above approx. 1.5 GiB). ALAC and FLAC compressors show a slight improvement (2-3%) in CR with larger chunk sizes. In most use cases this size reduction comes at a great cost in RAM requirement to process files above 2048 MiB. HDF5 has its own concept of "chunks", used for I/O and the filter pipeline, with a default size of 1 MiB. Internal limitations do not allow for HDF5-chunks larger than 2048 MiB, however, HDF5, in general, can be used for files larger than this limit. The MAFISC filter with LZMA

compression experiences large fluctuations for neighboring chunk size steps and should, therefore, be tuned separately. Overall, increasing the chunk size has a negligible effect on the final compression ratio and only pushes up the RAM requirements for processing.

7.5.4 Summary and Recommendations

The entropy analysis shows a lack of measurement range calibration in some datasets. This results in unutilized precision, that would have been available with the given hardware DAQ units. The used range directly affects the contained entropy, and therefore the achievable compression ratio. A well-calibrated measurement system is a key requirement to achieve the best signal range and resolution.

Choosing a file format for long-term whole-building energy datasets is a crucial component, directly affecting the visibility and accessibility of the data by other researchers. Using an unsupported encoding or requiring specialized tools to read the data is cumbersome and error-prone and should be avoided. We recommend using well-known file formats, such as HDF5 or FLAC, which are widely adopted and provide built-in support for metadata, compression, and error-detection. While ALAC and FLAC already provide internal compression, we recommend the MAFISC or zstd filters for HDF5, due to their superior compression ratio. The serialization orientation (row- or column-based) has only a minor effect.

Large datasets should be split into multiple smaller files to facilitate data handling, reduce transfer speeds and loading times for short amounts of data. We have found that compression algorithms (together with the above-described file formats) yield higher space savings with chunk sizes above 256 MiB to 384 MiB. Small files show a modest compression ratio, while larger files require more transfer bandwidth and time before the data can be analyzed.

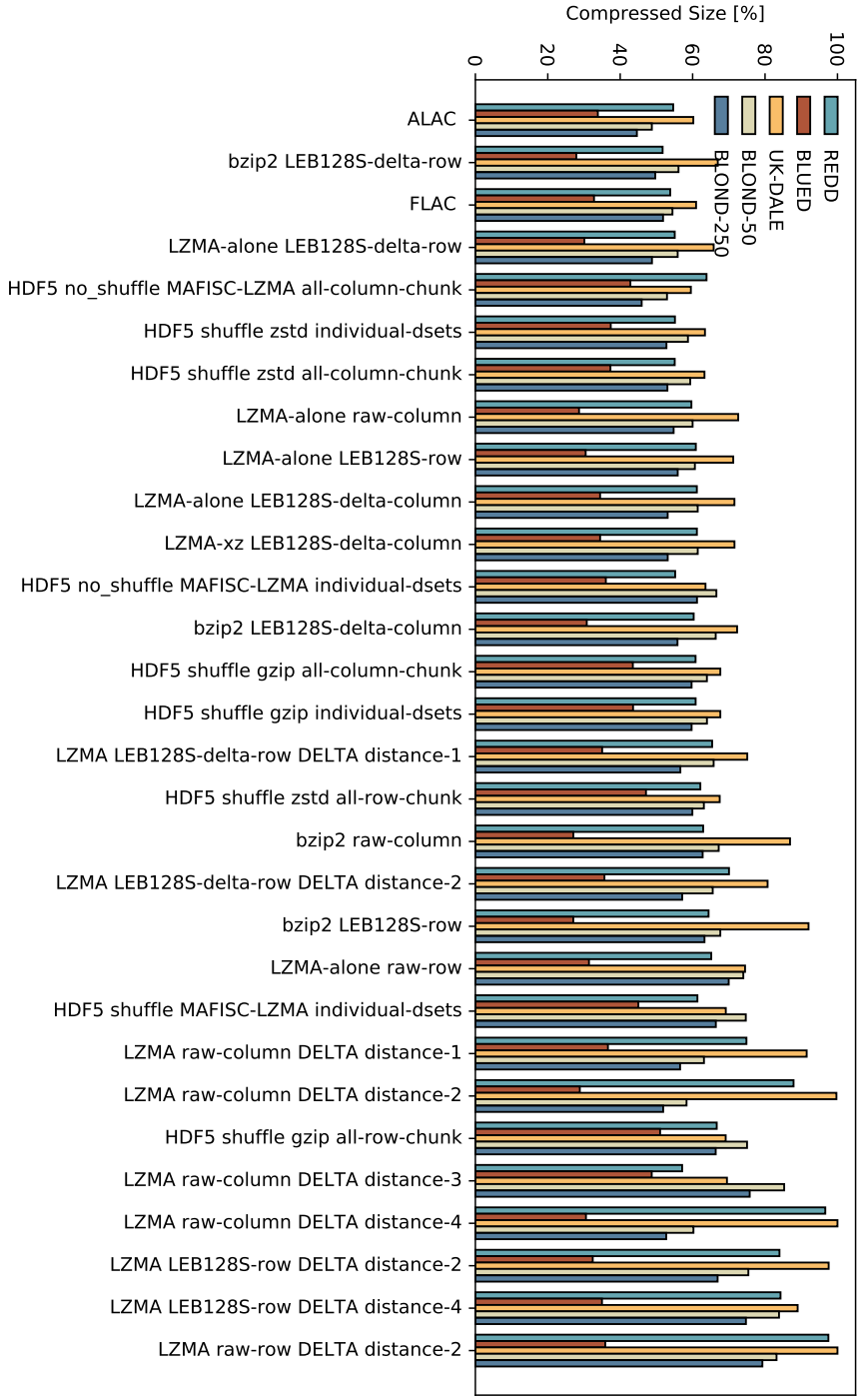
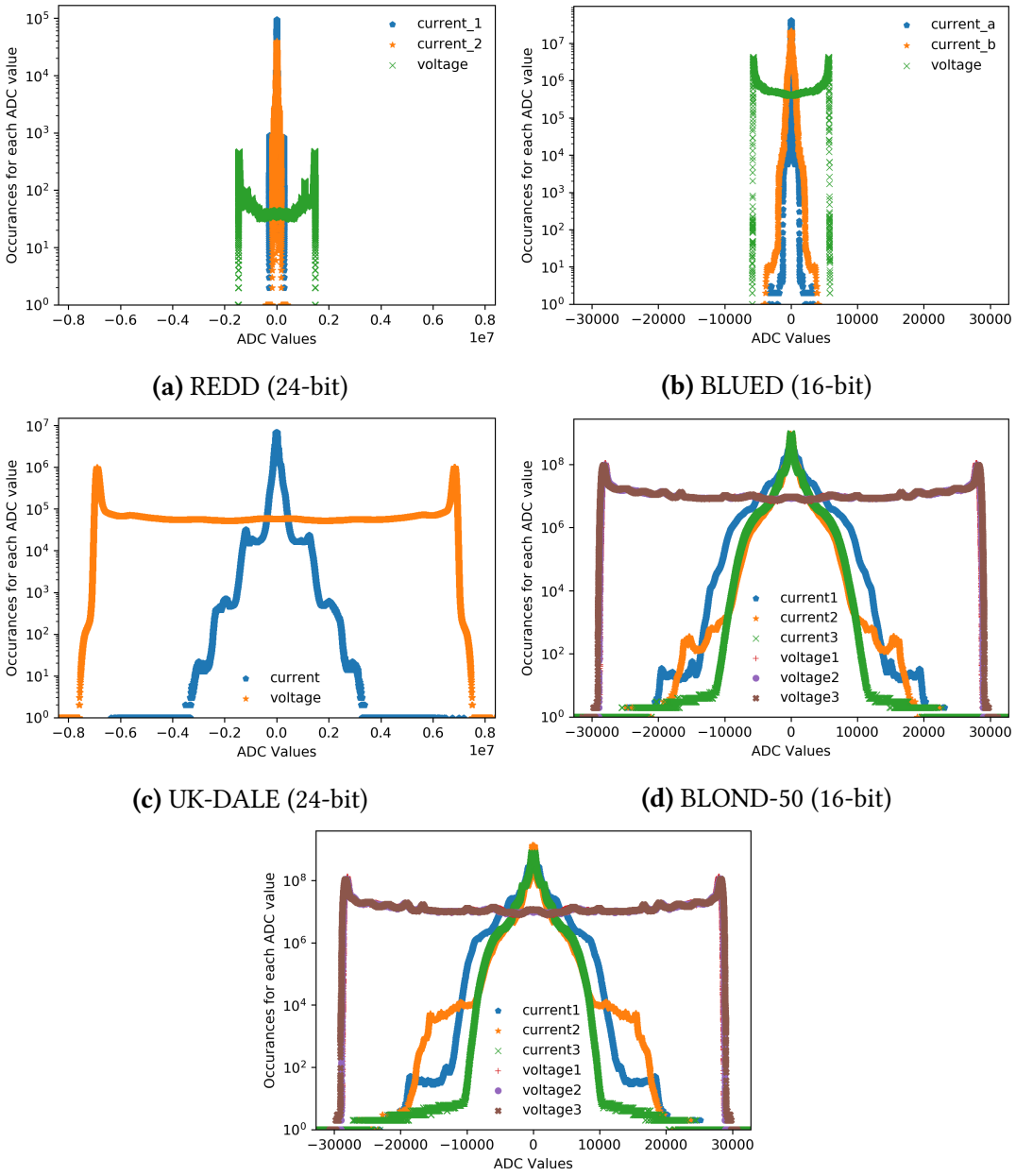


Figure 7.5.1: Compression performance for the top-30 data representation formats and their transformation filters. Each data representation format was applied on a per-file basis to every dataset.



(a) REDD (24-bit)

(b) BLUED (16-bit)

(c) UK-DALE (24-bit)

(d) BLOND-50 (16-bit)

(e) BLOND-250 (16-bit)

Figure 7.5.2: Semi-logarithmic histogram of ADC values for each dataset and channel. Current signals show distinct steps, corresponding to prolonged usage at certain power levels.

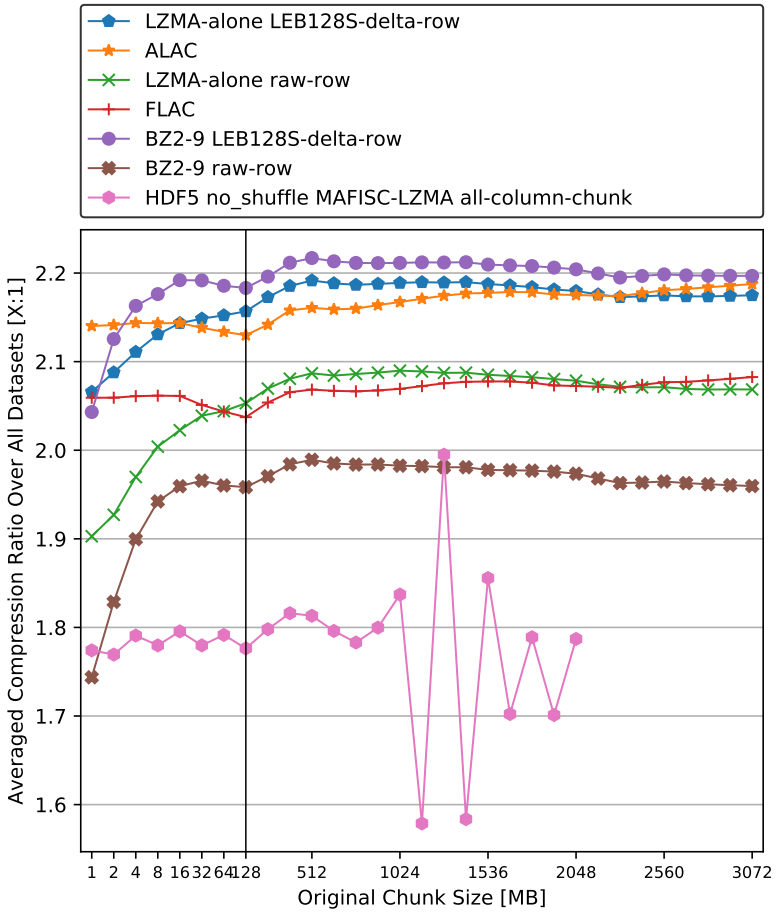


Figure 7.5.3: Chunk size impact of different representations.

8

Conclusions

8.1 Summary

Electrical energy consumption is a key component in the fight against climate change. Understanding the consumption characteristics of appliances and buildings requires significant hardware efforts for metering infrastructure. Office environments are a key energy consumer, where a majority of the daily inhabitants is unaware of their actual energy consumption footprint. SMPS-driven appliances have various operations modes (active, standby, deep-sleep) and can cause unexpected voltage and current spikes in the local power supply.

NILM is a promising approach to provide details energy information without the need for expensive meters and hard-to-maintain data collection systems. Power disaggregation and appliance identification from electrical energy signals, such as voltage and current, are two tasks a NILM system promises. A data-based solution to these tasks requires data acquisition systems and public datasets to model, train, and evaluate machine learning strategies. Collecting ground truth data with high sampling rates provides insights into the waveforms of SMPS-driven appliances, contrary to RMS-based consumption values of existing datasets and DAQ systems.

In this work, we presented full-stack architecture patterns (software blueprints and guidelines) for the distributed data acquisition and collection of electrical energy data. Hardware and software methodologies have been developed and evaluated based on the need for energy datasets with high sampling rates for aggregated and ground truth signals. The resulting datasets BLOND-50 and BLOND-250 are focused on NILM-related tasks, such as power disaggregation and appliance identification, with general-purpose design goals.

We have defined a set of design goals requirements for energy data acquisition systems to be used in NILM and appliance identification tasks. We showed a new system design for high sampling rates and multiple monitored sockets. The proposed design was implemented and a batch of DAQ systems was manufactured for evaluation purposes. The MEDAL architecture proved successful and passed all experimental tests derived from the requirements. We would like to encourage the use of a high-frequency ground truth for future research and new datasets. This could improve data quality for various NILM-related subtasks and similar research fields based on high frequency data.

The analog and digital data acquisition methodologies have been formulated into independent stackable components to be used for long-term continuous data collection, primarily targeted at datasets for power disaggregation and appliance identification with electrical waveform time series data. The proposed concepts were implemented and evaluated by collecting the BLOND datasets: BLOND-50 and BLOND-250 with multiple months of uninterrupted data collection and high sampling rates for aggregated and ground truth signals. All time series data files are validated against an extensive list of metrics and technical rules to ensure the data is valid and usable. The data coverage of 99.997% is significantly higher than any previously published dataset, while also providing novel waveform data for individual appliance EEC. We shared these datasets, including metadata and an appliance log with fully-labeled ground truth data, with the community to extend the foundation and input variety of data-driven machine learning tasks.

We presented a comprehensive entropy analysis of public whole-building energy datasets with waveform signals. Some datasets leave a majority of the available ADC range unused, causing lost precision and accuracy. A well-calibrated measurement system maximizes the achievable precision. Using 365 different data representation formats, we showed that

immense space savings of up to 73% are achievable by choosing a suitable file format and data transformation. Low entropy datasets show higher achievable compression ratios. Audio-based file formats perform considerably well, given the similarities to electricity waveforms. Transparent data transformations are particularly beneficial, such as MAFISC and SHUFFLE-based approaches. The input size shows a mostly stable dependency to the achievable compressed size, with variations of a few percentage points (limited by RAM). Waveform data shows a nearly constant compression ratio, independent of the input chunk size. Splitting large datasets into multiple smaller files is important for data handling, but insignificant in terms of space savings.

8.2 Future Work

Possible future work related to distributed data acquisition and collection includes: *a*) investigating the behavior of the proposed data collection strategies in a multi-building scenario (city neighborhood or campus area) with interconnected data acquisition systems; *b*) supporting alternative data collection topologies with de-/centralized cloud components to increase fault tolerance and HA metrics with multiple active collector and processing services; *c*) adapting the proposed software guidelines to make use of *microservices* and *serverless* architectural patterns [126, 127] to reduce deployment complexity and maintenance periods by splitting the cloud components into fully independent services with a defined communication interface and API versioning schemas to replace and upgrade parts of the data collection system without disturbing an active measurement series.

With the BLOND datasets, we have provided data of typical office appliances with a strong focus on SMPS-driven devices. Future work in the area of characterizing and measuring other types of environments includes: *a*) establishing an extensive appliance taxonomy to build a hierarchical model of electrical consumers and their energy consumption patterns based on different modes of operations, such as motors, heaters, and electronic control modules; *b*) collecting extensive metadata and external sensor information of the environment, such as room occupancy, temperature, humidity, motion, and available WiFi devices, to augment the existing EEC ground truth data; *c*) supplying and measuring

the DC load (low-voltage side) of SMPS-driven appliance (instead of the AC mains load) to revert the masking effect of AC-DC power supplies, increase the resolution, and reduce the noise energy.

Finally, we would propose further research into data representation schemes, encoding, and compression of EEC waveform data: *a*) analyzing the throughput and performance of the best-performing compression schemes and evaluating them in common scenarios, such as NILMTK [61] analytic pipelines or TensorFlow-based machine learning systems which share the requirement for fast and repeated access to measurement data; *b*) designing and evaluating edge-based compression schemes for low-level DAQ systems to compress in real-time before transmitting the data to save bandwidth and reduce the total transferred data size; *c*) raising awareness within the NILM and EEM community about the need for a common data representation and compression layer in publicly available datasets, while also improving the signal measurement range and the associated space-efficient value encoding.

Glossary

AC Alternating Current

ADC Analog-to-Digital Converter

BLOND Building-Level Office eNvironment Dataset

BLUED Building-Level fUlly-labeled dataset for Electricity Disaggregation

CLEAR Circuit-Level Energy Appliance Radar

CPU Central Processing Unit

DAC Digital-to-Analog Converter

DAQ Data Acquisition

DC Direct Current

DMA Direct Memory Access

EEC Electrical Energy Consumption

EEM Electrical Energy Metering

FPGA Field-Programmable Gate Array

HA High Availability

ICT Information and Communication Technology

MEDAL Mobile Energy Data Acquisition Laboratory

NIALM Non-Intrusive Appliance Load Monitoring

NILM Non-Intrusive Load Monitoring

RAM Random-Access Memory

REDD The Reference Energy Disaggregation Data Set

RMS Root-Mean Squared

RTC Real-Time Clock

RTOS Real-Time Operating System

SBC Single-Board Computer

SMPS Switch-Mode Power Supply

UK-DALE UK Domestic Appliance-Level Electricity dataset

USB Universal Serial Bus

List of Figures

4.1.1 The architecture of the measurement unit shows the data flow through multiple stages. 27
4.1.2 The physical components of a MEDAL measurement unit. 28
4.1.3 The data flow on the sensor board. All ADCs are connected via a single data bus to the microcontroller. 29
4.3.1 SMPS with a 5 W resistive load. A sharp transient is triggered by charging up the internal capacitors of the SMPS. 33
4.3.2 A 24-hour recording of two appliances with MEDAL in an office environment. 36
4.3.3 Waveform and spectral characteristics of a rotary multi-tool (50 Hz power grid) 37

5.2.1 The data acquisition flow of analog signals from a 3-phase power grid (L1, L2, L3 circuits) with voltage and current sensors into the ADC where the digital values are collected by the controller. The embedded system operates as independent entity and communicates with the SBC via AT-commands. 43
5.3.1 The software architecture on a single-board computer with the Measurement Governor and subsystems. Each service is managed and controlled by *systemd* and communicates via shared memory regions. 50
5.3.2 Governor used in STORE and STREAM data processing modes. Sample use cases are long-term continuous data collection and real-time data analytics tasks. 52
5.4.1 Pull-based Data Collection and Monitoring used for BLOND-50 . . . 57

5.7.1 Analysis of software deployments and local buffering operations at each measurement unit. Local buffering moves data from RAM onto the persistent flash memory. Software services have been continuously improved and deployed without interrupting the ongoing data collection. 62

5.7.2 BLOND-50: Duration of edge-based data conversion and compression for each collected file. 63

5.7.3 Distribution of transfer time per file from the measurement unit into the data center (pull-based data collection). The boxplot shows the lower and upper quartile in the box, with a red line for the median value, and whiskers range from 1-99%. 64

6.2.1 The measurement architecture of BLOND with physical placement of DAQ systems and connected appliances. A CLEAR unit is used as an EEC meter at the mains input to measure all 3 circuits in the electric cabinet. Multiple MEDAL units are placed in office rooms and connected to different circuits. Each MEDAL can be used to measure up to six appliances simultaneously in a single phase. Only a subset of MEDAL units is depicted; see Table 6.2.1 for a full circuit mapping. 74

6.2.2 The data acquisition and processing component of the CLEAR system. The laser-cut acrylic enclosure contains three components: sensor board, DAQ board, and a Linux PC. 75

6.2.3 CLEAR current sensors installed in the electric cabinet. The open-loop Hall-effect sensors employ multiple turns of the mains wiring to increase the usable output signal. A small connection board distributes supply voltage and output signals. All changes and alterations were authorized and conducted by certified personnel. 76

6.4.1 Waveforms of CLEAR circuits for voltage and current. The 3-phase power grid is characterized by a 120° phase shift between the circuits. The current consumption shows strong SMPS usage with sharp increases at each cycle apex. The voltage shows a typical sinusoidal waveform. 84

6.4.2 The load profile over multiple hours of the BLOND environment on 2017-03-02. Each stack plot shows the apparent power consumption of CLEAR (black line) with overlays for the contained MEDAL submeters (colored areas). See Table 6.2.1 for the circuit mapping. The individual steps (appliance events) match the overall load profile of the aggregate EEC. Each circuit shows a base load (gray area) which accounts for static background consumers. The 1-second data summary was downsampled to 30 seconds before plotting. Total consumption stays constant in the hours not shown. For visualization purposes, only the sum of all MEDAL sockets was plotted, however, the data contains an appliance-specific breakdown. 85
6.4.3 Boxplot of sampling rate precision per day. This boxplot shows the distribution of average sampling rates and its variation over 213 / 50 days for each measurement system. The whiskers depict the minimum and maximum. The mean was subtracted from each entry to compare only the daily fluctuations. The median is shown as line in the box. 87

7.2.1 Linear pulse-code modulation stream of a sinusoidal waveform sampled with a 16-bit ADC. The waveform corresponds to a 230 V mains voltage signal. 99
7.5.1 Compression performance for the top-30 data representation formats and their transformation filters. Each data representation format was applied on a per-file basis to every dataset. 110
7.5.2 Semi-logarithmic histogram of ADC values for each dataset and channel. 111
7.5.3 Chunk size impact of different representations. 112

List of Tables

4.0.1 Requirements for energy data acquistion hardware. 26

4.3.1 Storage requirements for different sampling rates. 34

6.1.1 Overview of long-term energy datasets with high sampling rates. This includes only datasets with long-term recordings of aggregate (above 10 kSps) and per-appliance measurements. In contrast to existing datasets, BLOND also provides ground truth data with a high sampling rate. 70

6.2.1 Circuit mapping for each measurement system. Associating a MEDAL unit with its aggregated circuit in the CLEAR data is fixed and does not change over time. This corresponds to the wiring of individual office rooms to use one (or more) of three different phases. 75

6.2.2 List of appliances observed in the BLOND dataset. This list was extracted from the appliance log and contains all devices used in the BLOND environment. A class label was assigned to group similar appliances. The manufacturer, type, and power information was taken from an attached name plate (if available) or the suppliers datasheet. 91

6.3.1 HDF5 dataset file metadata. Each attribute is accessible via a HDF5-attribute-path. Values are provided in base units (Volt, Ampere, Hertz). Some attributes are only available in the 1-second data summary. 92

6.3.2 File chunking and length. BLOND-50 and BLOND-250 use different file sizes to chunk the continuous data stream. The size depends on the available computing resources in each DAQ unit and the configured sampling rate. The final size of the HDF5 only depends on the number of samples and the achievable compression ratio. . . 92

7.1.1 Overview of evaluated datasets: long-term continuous measurements containing raw voltage and current waveforms. 98
7.5.1 Entropy analysis of whole-building energy datasets with high sampling rates. The amount of unique measurement values for each channel is extracted, which corresponds to a usage percentage over the available measurement resolution. The lowest and highest observed value is used to give determine the observed range. 106

Bibliography

[1] C. Fischer. "Feedback on household electricity consumption: a tool for saving energy?" In: *Energy Efficiency* 1.1 (2008), pp. 79–104. ISSN: 1570-6478. DOI: 10.1007/s12053-008-9009-7.

[2] G. W. Hart. "Nonintrusive Appliance Load Monitoring." In: *Proceedings of the IEEE* 80.12 (1992), pp. 1870–1891. ISSN: 0018-9219. DOI: 10.1109/5.192069.

[3] D. Kaneda, B. Jacobson, P. Rumsey, and R. Engineers. "Plug load reduction: The next big hurdle for net zero energy building design." In: *ACEEE Summer Study on Energy Efficiency in Buildings*. Vol. 16. 2010, pp. 120–130.

[4] V. L. Chen, M. A. Delmas, and W. J. Kaiser. "Real-time, appliance-level electricity use feedback system: How to engage users?" In: *Energy and Buildings* 70 (2014), pp. 455–462. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2013.11.069.

[5] M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "A Comprehensive Feature Study for Appliance Recognition on High Frequency Energy Data." In: *Proceedings of the 2017 ACM Eighth International Conference on Future Energy Systems*. e-Energy '17. Hong Kong, Hong Kong: ACM, 2017. ISBN: 978-1-4503-5036-5. DOI: 10.1145/3077839.3077845.

[6] H. Lange and M. Bergés. "BOLT: Energy Disaggregation by Online Binary Matrix Factorization of Current Waveforms." In: *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*. BuildSys '16. Palo Alto, CA, USA: ACM, 2016, pp. 11–20. ISBN: 978-1-4503-4264-3. DOI: 10.1145/2993422.2993581.

[7] C. P. Salomon, W. C. Santana, E. L. Bonaldi, et al. "A System for Turbogenerator Predictive Maintenance Based on Electrical Signature Analysis." In: *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*. 2015, pp. 79–84. DOI: 10.1109/I2MTC.2015.7151244.

[8] M. Bergés, E. Goldman, H. S. Matthews, and L. Soibelman. "Learning Systems for Electric Consumption of Buildings." In: *Computing in Civil Engineering*. Vol. 38. Austin, Texas, United States: American Society of Civil Engineers, 2009. DOI: 10.1061/41052(346)1.

[9] H. Ziekow, C. Doblander, C. Goebel, and H.-A. Jacobsen. "Forecasting Household Electricity Demand with Complex Event Processing: Insights from a Prototypical Solution." In: *Proceedings of the Industrial Track of the 13th ACM/IFIP/USENIX International Middleware Conference*. Middleware Industry '13. Beijing, China: ACM, 2013, pp. 1–6. ISBN: 978-1-4503-2550-9. DOI: 10.1145/2541596.2541598.

[10] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and N. Gudi. "Smart meters for power grid — Challenges, issues, advantages and status." In: *2011 IEEE/PES Power Systems Conference and Exposition*. Mar. 2011, pp. 1–7. DOI: 10.1109/PSCE.2011.5772451.

[11] N. Batra, A. Singh, and K. Whitehouse. "Gemello: Creating a Detailed Energy Breakdown from Just the Monthly Electricity Bill." In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 431–440. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939735.

[12] S. Makonin. "Investigating the Switch Continuity Principle Assumed in Non-Intrusive Load Monitoring." In: *IEEE CCECE '16*. Vancouver, Canada: IEEE, 2016, pp. 1–4.

[13] J. Paris, J. S. Donnal, and S. B. Leeb. "NilmDB: The Non-Intrusive Load Monitor Database." In: *IEEE Transactions on Smart Grid* 5.5 (Sept. 2014), pp. 2459–2467. ISSN: 1949-3053. DOI: 10.1109/TSG.2014.2321582.

[14] A. S. Bouhouras, A. N. Milioudis, and D. P. Labridis. "Development of distinct load signatures for higher efficiency of NILM algorithms." In: *Electric Power Systems Research* 117 (2014), pp. 163–171. ISSN: 0378-7796. DOI: 10.1016/j.epsr.2014.08.015.

[15] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert. "Is disaggregation the holy grail of energy efficiency? The case of electricity." In: *Energy Policy* 52 (2013), pp. 213–234. DOI: 10.1016/j.enpol.2012.08.062.

[16] D. He, W. Lin, N. Liu, R. G. Harley, and T. G. Habetler. "Incorporating Non-Intrusive Load Monitoring Into Building Level Demand Response." In: *IEEE Transactions on Smart Grid* 4.4 (Dec. 2013), pp. 1870–1877. ISSN: 1949-3053. DOI: 10.1109/TSG.2013.2258180.

[17] A. Veit, C. Goebel, R. Tidke, C. Doblander, and H.-A. Jacobsen. "Household Electricity Demand Forecasting: Benchmarking State-of-the-art Methods." In: *Proceedings of the 5th International Conference on Future Energy Systems*. e-Energy '14. Cambridge, United Kingdom: ACM, 2014, pp. 233–234. ISBN: 978-1-4503-2819-7. DOI: 10.1145/2602044.2602082. URL: http://doi.acm.org/10.1145/2602044.2602082.

[18] F. Englert, T. Schmitt, S. Kößler, A. Reinhardt, and R. Steinmetz. "How to Auto-configure Your Smart Home? High-resolution Power Measurements to the Rescue." In: *Proceedings of the 2013 ACM Fourth International Conference on Future Energy Systems*. e-Energy '13. Berkeley, California, USA, May 2013, pp. 215–224. ISBN: 978-1-4503-2052-8. DOI: 10.1145/2487166.2487191.

[19] S. Barker, A. Mishra, D. Irwin, et al. "Smart*: An open data set and tools for enabling research in sustainable homes." In: *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*. Beijing, China: ACM, Aug. 2012, pp. 112–118.

[20] P. Shenavar and E. Farjah. "Novel embedded real-time NILM for electric loads disaggregating and diagnostic." In: *EUROCON 2007 - The International Conference on "Computer as a Tool"*. Sept. 2007, pp. 1555–1560. DOI: 10.1109/EURCON.2007.4400282.

[21] C. E. Shannon. "Communication in the Presence of Noise." In: *Proceedings of the IRE* 37.1 (Jan. 1949), pp. 10–21. ISSN: 0096-8390. DOI: 10.1109/JRPROC.1949.232969.

[22] J. Z. Kolter and M. J. Johnson. "REDD: A Public Data Set for Energy Disaggregation Research." In: *SustKDD '11*. Vol. 25. San Diego, California, USA, 2011, pp. 59–62.

[23] K. Anderson, A. Ocneanu, D. Benitez, et al. "BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research." In: *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD)*. Beijing, China: ACM, Aug. 2012, pp. 1–5.

[24] J. Kelly and W. Knottenbelt. "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes." In: *Scientific Data* 2.150007 (Mar. 31, 2015). DOI: 10.1038/sdata.2015.7.

[25] T. Kriechbaumer and H.-A. Jacobsen. "BLOND, a building-level office environment dataset of typical electrical appliances." In: *Scientific Data, an open-access NatureResearch journal* 5.180048 (2018). DOI: 10.1038/sdata.2018.48.

[26] A. U. Haq, T. Kriechbaumer, M. Kahl, and H.-A. Jacobsen. "CLEAR – A Circuit Level Electric Appliance Radar for the Electric Cabinet." In: *2017 IEEE International Conference on Industrial Technology*. ICIT '17. Toronto, Canada, 2017, pp. 1130–1135. ISBN: 978-1-5090-5319-3. DOI: 10.1109/ICIT.2017.7915521.

[27] T. Kriechbaumer, A. U. Haq, M. Kahl, and H.-A. Jacobsen. "MEDAL: A Cost-Effective High-Frequency Energy Data Acquisition System for Electrical Appliances." In: *Proceedings of the 2017 ACM Eighth International Conference on Future Energy Systems*. e-Energy '17. Hong Kong, Hong Kong: ACM, 2017. ISBN: 978-1-4503-5036-5. DOI: 10.1145/3077839.3077844.

[28] T. Kriechbaumer, M. Kahl, D. Jorde, A. U. Haq, and H.-A. Jacobsen. "Large-Scale Data Acquisition Systems Architecture for High-Frequency Electrical Energy Metering." Submitted to ACM Trans. Cyber-Phys. Syst. 2019.

[29] T. Kriechbaumer and H.-A. Jacobsen. *Waveform Signal Entropy and Compression Study of Whole-Building Energy Datasets*. 2018. arXiv: 1810.10887.

[30] M. Kahl, C. Goebel, A. Haq, T. Kriechbaumer, and H.-A. Jacobsen. "NoFaRe: A Non-Intrusive Facility Resource Monitoring System." In: *Energy Informatics*. Vol. 9424. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 59–68. DOI: 10.1007/978-3-319-25876-8_6.

[31] M. Kahl, T. Kriechbaumer, A. U. Haq, and H.-A. Jacobsen. "Appliance Classification Across Multiple High Frequency Energy Datasets." In: *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm): Smart metering, Demand Response and Dynamic Pricing (SGC2017 Smart Metering)*. 2017. DOI: 10.1109/smartgridcomm.2017.8340664.

[32] D. Jorde, T. Kriechbaumer, and H.-A. Jacobsen. "Electrical Appliance Classification using Deep Convolutional Neural Networks on High Frequency Current Measurements." In: *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm) (IEEE SmartGridComm'18)*. Aalborg, Denmark, 2018.

[33] M. Kahl, V. Krause, R. Hackenberg, et al. "Measurement System and Dataset for In-Depth Analysis of Appliance Energy Consumption in Industrial Environment." In: *tm - Technisches Messen* (2018). DOI: 10.1515/teme-2018-0038.

[34] J. Liang, S. K. K. Ng, G. Kendall, and J. W. M. Cheng. "Load Signature Study—Part I: Basic Concept, Structure, and Methodology." In: *IEEE Transactions on Power Delivery* 25.2 (Apr. 2010), pp. 551–560. ISSN: 0885-8977. DOI: 10.1109/TPWRD.2009.2033799.

[35] R. Cox, S. B. Leeb, S. R. Shaw, and L. K. Norford. "Transient event detection for nonintrusive load monitoring and demand side management using voltage distortion." In: *Twenty-First Annual IEEE Applied Power Electronics Conference and Exposition, 2006. APEC '06.* Mar. 2006. DOI: 10.1109/APEC.2006.1620777.

[36] O. N. Gerek and D. G. Ece. "2-D analysis and compression of power-quality event data." In: *IEEE Transactions on Power Delivery* 19.2 (Apr. 2004), pp. 791–798. ISSN: 0885-8977. DOI: 10.1109/TPWRD.2003.823197.

[37] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar. "Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey." In: *Sensors* 12.12 (2012), p. 16838. ISSN: 1424-8220. DOI: 10.3390/s121216838.

[38] PNNL - Pacific Northwest National Laboratory. *Characteristics and Performance of Existing Load Disaggregation Technologies*. 2016. URL: http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-24230.pdf.

[39] L. Mauch and B. Yang. "A new approach for supervised power disaggregation by using a deep recurrent LSTM network." In: *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Dec. 2015, pp. 63–67. DOI: 10.1109/GlobalSIP.2015.7418157.

[40] I. S. Association. *IEEE 1057-2017 - IEEE Standard for Digitizing Waveform Recorders*. May 2017. URL: https://standards.ieee.org/standard/1057-2017.html.

[41] M. N. Meziane, T. Picon, P. Ravier, et al. "A Measurement System for Creating Datasets of On/Off-Controlled Electrical Loads." In: *2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*. June 2016, pp. 1–5. DOI: 10.1109/EEEIC.2016.7555847.

[42] I. Hickman. *Digital Storage Oscilloscopes*. Elsevier Science, 1997. ISBN: 9780080504513.

[43] E. J. Candès, J. K. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements." In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223. DOI: 10.1002/cpa.20124.

[44] J. Gao, S. Giri, E. C. Kara, and M. Bergés. "PLAID: A Public Dataset of High-resolution Electrical Appliance Measurements for Load Identification Research: Demo Abstract." In: *ACM BuildSys '14*. Memphis, Tennessee, 2014, pp. 198–199. ISBN: 978-1-4503-3144-9. DOI: 10.1145/2674061.2675032.

[45] M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen. *WHITED - A Worldwide Household and Industry Transient Energy Data Set*. 2016. URL: http://nilmworkshop.org/2016/proceedings/Poster_ID18.pdf.

[46] T. Picon, M. Nait Meziane, P. Ravier, et al. "COOLL: Controlled On/Off Loads Library, a Public Dataset of High-Sampled Electrical Signals for Appliance Identification." In: *arXiv preprint arXiv:1611.05803 [cs.OH]* (2016).

[47] T. Babaei, H. Abdi, C. P. Lim, and S. Nahavandi. "A study and a directory of energy consumption data sets of buildings." In: *Energy and Buildings* 94 (2015), pp. 91–99. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2015.02.043. URL: http://www.sciencedirect.com/science/ARTICLE/pii/S0378778815001486.

[48] D. Lawrence, J. S. Donnal, and S. Leeb. "Current and Voltage Reconstruction From Non-Contact Field Measurements." In: *IEEE Sensors Journal* 16.15 (Aug. 2016), pp. 6095–6103. ISSN: 1530-437X. DOI: 10.1109/JSEN.2016.2574245.

[49] W. Lee, G. Fung, H. Lam, F. Chan, and M. Lucente. "Exploration on Load Signatures." English. In: *ICEECS '04*. Vol. 2. Japan: EECS, July 2004. eprint: 10.1.1.120.5328.

[50] M. Ringwelski, C. Renner, A. Reinhardt, A. Weigel, and V. Turau. "The Hitchhiker's Guide to choosing the Compression Algorithm for your Smart Meter Data." In: *Energy Conference and Exhibition (ENERGYCON), 2012 IEEE International*. Sept. 2012, pp. 935–940. DOI: 10.1109/EnergyCon.2012.6348285.

[51] A. Unterweger and D. Engel. "Resumable load data compression in smart grids." In: *IEEE Transactions on Smart Grid* 6.2 (2015), pp. 919–929. DOI: 10.1109/TSG.2014.2364686.

[52] A. Unterweger, D. Engel, and M. Ringwelski. "The Effect of Data Granularity on Load Data Compression." In: *Energy Informatics: 4th D-A-CH Conference, EI 2015, Karlsruhe, Germany, November 12-13, 2015, Proceedings*. Cham: Springer International Publishing, 2015, pp. 69–80. ISBN: 978-3-319-25876-8. DOI: 10.1007/978-3-319-25876-8_7.

[53] F. Eichinger, P. Efros, S. Karnouskos, and K. Böhm. "A Time-series Compression Technique and Its Application to the Smart Grid." In: *The VLDB Journal* 24.2 (Apr. 2015), pp. 193–218. ISSN: 1066-8888. DOI: 10.1007/s00778-014-0368-8.

[54] M. Nabeel, F. Javed, and N. Arshad. "Towards Smart Data Compression for Future Energy Management System." In: *Fifth International Conference on Applied Energy*. 2013.

[55] Z. B. Tariq, N. Arshad, and M. Nabeel. "Enhanced LZMA and BZIP2 for improved energy data compression." In: *2015 International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)*. May 2015, pp. 1–8.

[56] L. Pereira, N. Nunes, and M. Bergés. "SURF and SURF-PI: A File Format and API for Non-intrusive Load Monitoring Public Datasets." In: *Proceedings of the 5th International Conference on Future Energy Systems*. e-Energy '14. Cambridge, United Kingdom: ACM, 2014, pp. 225–226. ISBN: 978-1-4503-2819-7. DOI: 10.1145/2602044.2602078.

[57] L. Pereira. "EMD-DF: A Data Model and File Format for Energy Disaggregation Datasets." In: *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*. BuildSys '17. Delft, Netherlands: ACM, 2017, 52:1–52:2. ISBN: 978-1-4503-5544-5. DOI: 10.1145/3137133.3141474.

[58] I. P. Ẽ. Society. *IEEE 1159 - PQDIF: Power Quality and Quantity Data Interchange Format*. Jan. 2018. URL: http://grouper.ieee.org/groups/1159/3/docs.html.

[59] I. S. Association. *COMTRADE: Common format for Transient Data Exchange for power systems*. Jan. 2018. URL: https://standards.ieee.org/findstds/standard/C37.111-2013.html.

[60] A. Qing, Z. Hongtao, H. Zhikun, and C. Zhiwen. "A Compression Approach of Power Quality Monitoring Data Based on Two-dimension DCT." In: *2011 Third International Conference on Measuring Technology and Mechatronics Automation*. Vol. 1. Jan. 2011, pp. 20–24. DOI: 10.1109/ICMTMA.2011.12.

[61] N. Batra, J. Kelly, O. Parson, et al. "NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring." In: *ACM e-Energy '14*. New York, NY, USA: ACM, 2014, pp. 265–276. DOI: 10.1145/2602044.2602051.

[62] M. Folk, G. Heber, Q. Koziol, E. Pourmal, and D. Robinson. "An Overview of the HDF5 Technology Suite and Its Applications." In: *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*. AD '11. Uppsala, Sweden: ACM, 2011, pp. 36–47. ISBN: 978-1-4503-0614-0. DOI: 10.1145/1966895.1966900.

[63] F. Alted. *Blosc: A high performance compressor optimized for binary data*. 2017. URL: http://blosc.org/.

[64] A. U. Haq and H.-A. Jacobsen. "A Step towards Advanced Metering for the Smart Grid: A Survey of Energy Monitors." In: *CoRR* abs/1607.07780 (2016). URL: http://arxiv.org/abs/1607.07780.

[65] C. D. Murta and P. R. Torres-jr. "Characterizing quality of time and topology in a time synchronization network." In: *49th IEEE Global Telecommunications Conference, IEEE GLOBECOM*. 2006.

[66] D. Mills. *Computer Network Time Synchronization: The Network Time Protocol*. CRC Press, 2006. ISBN: 9781420006155. DOI: 10.1201/9781420006155. URL: https://books.google.de/books?id=pdTcJBfnbq8C.

[67] E. H. Hall. "On a New Action of the Magnet on Electric Currents." In: *American Journal of Mathematics* 2.3 (1879), pp. 287–292. ISSN: 00029327, 10806377. URL: http://www.jstor.org/stable/2369245.

[68] W. Rogowski and W. Steinhaus. "Die Messung der magnetischen Spannung." In: *Archiv für Elektrotechnik* 1.4 (Apr. 1912), pp. 141–150. ISSN: 1432-0487. DOI: 10.1007/BF01656479. URL: https://doi.org/10.1007/BF01656479.

[69] A. S. Tanenbaum and H. Bos. *Modern Operating Systems*. 4th. Upper Saddle River, NJ, USA: Prentice Hall Press, 2014. ISBN: 9780133591620.

[70] USB Implementers Forum, Inc. *USB 2.0 Specification*. 2000.

[71] libusb Community. *libusb: A cross-platform user library to access USB devices*. 2018. URL: https://libusb.info.

[72] Intra2net AG. *libFTDI - FTDI USB driver with bitbang mode*. 2018. URL: https://www.intra2net.com/en/developer/libftdi/index.php.

[73] systemd Community. *systemd System and Service Manager*. 2018. URL: https://freedesktop.org/wiki/Software/systemd/.

[74] D. Hayes. *Hayes AT Commands*. 2018. URL: http://home.intekom.com/option/hayesat.htm.

[75] M. Frerking. *Crystal oscillator design and temperature compensation*. Van Nostrand, 1978. ISBN: 9780442224592. URL: https://books.google.de/books?id=ISlTAAAAMAAJ.

[76] I. S. Association. *IEEE 1588-2008 - IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems*. May 2008. URL: https://standards.ieee.org/standard/1588-2008.html.

[77] I. M. Gottlieb. *Practical RF Power Design Techniques*. McGraw-Hill Professional Publishing, 1993. ISBN: 9780070239869.

[78] P. Pavan, R. Bez, P. Olivo, and E. Zanoni. "Flash memory cells-an overview." In: *Proceedings of the IEEE* 85.8 (Aug. 1997), pp. 1248–1271. ISSN: 0018-9219. DOI: 10.1109/5.622505.

[79] I. Amazon Web Services. *Amazon Compute Service Level Agreement*. May 2018. URL: https://aws.amazon.com/compute/sla/.

[80] Google. *Google Compute Engine Service Level Agreement*. May 2018. URL: https://cloud.google.com/compute/sla.

[81] Microsoft. *SLA for Virtual Machines*. May 2018. URL: https://azure.microsoft.com/en-us/support/legal/sla/virtual-machines/v1_8/.

[82] P. F. Dubois, K. Hinsen, and J. Hugunin. "Numerical Python." In: *Computers in Physics* 10.3 (May 1996).

[83] F. Chollet et al. *Keras*. https://keras.io. 2015.

[84] M. Abadi, P. Barham, J. Chen, et al. "TensorFlow: A System for Large-Scale Machine Learning." In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. Savannah, GA: USENIX Association, 2016, pp. 265–283. ISBN: 978-1-931971-33-1. URL: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

[85] J. W. Eaton, D. Bateman, S. Hauberg, and R. Wehbring. *GNU Octave version 4.2.0 manual: a high-level interactive language for numerical computations*. 2016. URL: http://www.gnu.org/software/octave/doc/interpreter.

[86] G. Hébrail and A. Bérard. "Individual household electric power consumption data set." In: *É. d. France, Ed., ed: UCI Machine Learning Repository* (2012).

[87] C. Holcomb. "Pecan Street Inc.: A Test-bed for NILM." In: *International Workshop on Non-Intrusive Load Monitoring*. 2007, pp. 271–288.

[88] A. Monacchi, D. Egarter, W. Elmenreich, S. D'Alessandro, and A. M. Tonello. "GREEND: An energy consumption dataset of households in Italy and Austria." In: *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 2014, pp. 511–516. DOI: 10.1109/SmartGridComm.2014.7007698.

[89] C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, and S. Santini. "The ECO Data Set and the Performance of Non-intrusive Load Monitoring Algorithms." In: *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. BuildSys '14. Memphis, Tennessee: ACM, 2014, pp. 80–89. ISBN: 978-1-4503-3144-9. DOI: 10.1145/2674061.2674064.

[90] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajić. "AMPds: A public dataset for load disaggregation and eco-feedback research." In: *2013 IEEE Electrical Power Energy Conference*. 2013, pp. 1–6. DOI: 10.1109/EPEC.2013.6802949.

[91] S. Makonin, B. Ellert, I. V. Bajić, and F. Popowich. "Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014." In: *Scientific data* 3 (2016).

[92] L. Pereira, F. Quintal, R. Gonçalves, and N. J. Nunes. "SustData: A Public Dataset for ICT4S Electric Energy Research." In: *International Conference on ICT for Sustainability (ICT4S 14)*. Atlantis Press. Stockholm, Sweden: Atlantis Press, Aug. 2014. DOI: 10.2991/ict4s-14.2014.44.

[93] M. Maasoumy, B. Sanandaji, K. Poolla, and A. S. Vincentelli. "BERDS - BERkeley EneRgy Disaggregation Data Set." In: *Proceedings of the Workshop on Big Learning at the Conference on Neural Information Processing Systems (NIPS)*. 2013.

[94] D. Murray, L. Stankovic, and V. Stankovic. "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study." In: *Scientific Data* 4 (2017), p. 160122.

[95] A. Menezes, A. Cripps, R. A. Buswell, J. Wright, and D. Bouchlaghem. "Estimating the energy consumption and power demand of small power equipment in office buildings." In: *Energy and Buildings* 75 (2014), pp. 199–209.

[96] M. Fantauzzi, D. Iannuzzi, M. Pagano, A. Scalfati, and M. Roscia. "Building DC microgrids: Planning of an experimental platform with power hardware in the loop features." In: *2015 International Conference on Renewable Energy Research and Applications (ICRERA)*. 2015, pp. 1507–1512. DOI: 10.1109/ICRERA.2015.7418659.

[97] B. Wunder, J. Kaiser, F. Fersterra, et al. "Energy distribution with DC microgrids in commercial buildings with power electronics." In: *2015 International Symposium on Smart Electric Distribution Systems and Technologies (EDST)*. 2015, pp. 425–430. DOI: 10.1109/SEDST.2015.7315246.

[98] F. Leferink, C. Keyer, and A. Melentjev. "Static energy meter errors caused by conducted electromagnetic interference." In: *IEEE Electromagnetic Compatibility Magazine* 5.4 (2016), pp. 49–55. ISSN: 2162-2264. DOI: 10.1109/MEMC.2016.7866234.

[99] M. Sheppy, L. Gentile-Polese, and S. Gould. *Plug and Process Loads Capacity and Power Requirements Analysis*. Tech. rep. National Renewable Energy Laboratory (NREL), Golden, CO., 2014.

[100] Analog Devices, Inc. *AD7656A Datasheet*. 2017. URL: http://www.analog.com/media/en/technical-documentation/data-sheets/AD7656A.pdf.

[101] LEM Holding SA. *HAL50-S Datasheet*. 2015. URL: http://www.lem.com/docs/products/hal%5C_50%5C_600-s.pdf.

[102] Microchip Technology Inc. *MCP3201 Datasheet*. 2007. URL: http://ww1.microchip.com/downloads/en/DeviceDoc/21290D.pdf.

[103] T. Kriechbaumer. *Evaluation source code for BLOND*. Aug. 2017. DOI: 10.5281/zenodo.838974. URL: https://doi.org/10.5281/zenodo.838974.

[104] D. Yuan, Y. Yang, X. Liu, and J. Chen. "A cost-effective strategy for intermediate data storage in scientific cloud workflow systems." In: *2010 IEEE International Symposium on Parallel Distributed Processing (IPDPS)*. Apr. 2010, pp. 1–12. DOI: 10.1109/IPDPS.2010.5470453.

[105] E. Deelman and A. Chervenak. "Data Management Challenges of Data-Intensive Scientific Workflows." In: *2008 Eighth IEEE International Symposium on Cluster Computing and the Grid (CCGRID)*. May 2008, pp. 687–692. DOI: 10.1109/CCGRID.2008.24.

[106] G. Liu and H. Shen. "Minimum-Cost Cloud Storage Service Across Multiple Cloud Providers." In: *IEEE/ACM Trans. Netw.* 25.4 (Aug. 2017), pp. 2498–2513. ISSN: 1063-6692. DOI: 10.1109/TNET.2017.2693222.

[107] K. P. Puttaswamy, T. Nandagopal, and M. Kodialam. "Frugal Storage for Cloud File Systems." In: *Proceedings of the 7th ACM European Conference on Computer Systems*. EuroSys '12. Bern, Switzerland: ACM, 2012, pp. 71–84. ISBN: 978-1-4503-1223-3. DOI: 10.1145/2168836.2168845.

[108] A. Bookstein and J. A. Storer. "Data compression." In: *Information Processing & Management* 28.6 (1992). Special Issue: Data compression for images and texts, pp. 675–680. ISSN: 0306-4573. DOI: 10.1016/0306-4573(92)90060-D.

[109] J. C. S. de Souza, T. M. L. Assis, and B. C. Pal. "Data Compression in Smart Distribution Systems via Singular Value Decomposition." In: *IEEE Transactions on Smart Grid* 8.1 (Jan. 2017), pp. 275–284. ISSN: 1949-3053. DOI: 10.1109/TSG.2015.2456979.

[110] NILM Community. *NILM datasets*. May 2018. URL: http://wiki.nilm.eu/datasets.html.

[111] European Committee for Electrotechnical Standardization. *CENELEC Harmonisation Document HD 472 S1*. 1989.

[112] American National Standards Institute. *ANSI C84.1-2016: Standard for Electric Power Systems and Equipment—Voltage Ratings (60 Hz)*. 2016. URL: https://www.pge.com/includes/docs/pdfs/mybusiness/customerservice/energystatus/powerquality/voltage_tolerance.pdf.

[113] R. Arnold and T. Bell. "A corpus for the evaluation of lossless compression algorithms." In: *Data Compression Conference, 1997. DCC '97. Proceedings*. 1997, pp. 201–210. DOI: 10.1109/DCC.1997.582019.

[114] Free Standards Group. *DWARF Debugging Information Format Specification Version 3.0*. 2018. URL: http://dwarfstd.org/doc/Dwarf3.pdf.

[115] S. Blanas, K. Wu, S. Byna, B. Dong, and A. Shoshani. "Parallel Data Analysis Directly on Scientific File Formats." In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD '14. Snowbird, Utah, USA: ACM, 2014, pp. 385–396. ISBN: 978-1-4503-2376-5. DOI: 10.1145/2588555.2612185.

[116] L. Gosink, J. Shalf, K. Stockinger, K. Wu, and W. Bethel. "HDF5-FastQuery: Accelerating Complex Queries on HDF Datasets using Fast Bitmap Indices." In: *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*. 2006, pp. 149–158. DOI: 10.1109/SSDBM.2006.27.

[117] M. T. Dougherty, M. J. Folk, E. Zadok, et al. "Unifying Biological Image Formats with HDF5." In: *Commun. ACM* 52.10 (Oct. 2009), pp. 42–47. ISSN: 0001-0782. DOI: 10.1145/1562764.1562781.

[118] S. Sehrish, J. Kowalkowski, M. Paterno, and C. Green. "Python and HPC for High Energy Physics Data Analyses." In: *Proceedings of the 7th Workshop on Python for High-Performance and Scientific Computing*. PyHPC'17. Denver, CO, USA: ACM, 2017, 8:1–8:8. ISBN: 978-1-4503-5124-9. DOI: 10.1145/3149869.3149877.

[119] N. Hübbe and J. Kunkel. "Reducing the HPC-datastorage footprint with MAFISC—Multidimensional Adaptive Filtering Improved Scientific data Compression." In: *Computer Science - Research and Development* 28.2 (2013), pp. 231–239. ISSN: 1865-2042. DOI: 10.1007/s00450-012-0222-4.

[120] HDF Group. *Szip Compression in HDF Products*. 2017. URL: https://support.hdfgroup.org/doc_resource/SZIP/.

[121] K. Masui, M. Amiri, L. Connor, et al. "A compression scheme for radio data in high performance computing." In: *Astronomy and Computing* 12.Supplement C (2015), pp. 181–190. ISSN: 2213-1337. DOI: 10.1016/j.ascom.2015.07.002.

[122] A. Miles. *Zarr: A Python package providing an implementation of chunked, compressed, N-dimensional arrays*. 2018. URL: https://zarr.readthedocs.io/en/latest/.

[123] Xiph.Org Foundation. *FLAC: Free Lossless Audio Codec*. 2018. URL: https://xiph.org/flac/.

[124] Apple Inc. *ALAC: Apple Lossless Audio Codec*. 2018. URL: https://macosforge.github.io/alac/.

[125] D. Bryant. *WavPack: Hybrid Lossless Audio Compression*. 2018. URL: http://www.wavpack.com/.

[126] S. Fowler. *Production-ready Microservices: Building Standardized Systems Across an Engineering Organization*. O'Reilly, 2016. ISBN: 9781491965979.

[127] C. Gurturk. *Building Serverless Architectures*. Packt Publishing, 2017. ISBN: 9781787129191.