A Cost-Benefit Analysis of Data Processing Architectures for the Smart Grid

Akshay Uttama Nambi S. N[†], Matteo Vasirani[¶], R. Venkatesha Prasad[†], Karl Aberer[¶]

[†]Embedded Software Group, Delft University of Technology (TUDelft), The Netherlands
[¶]LSIR, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

(akshay.narashiman,r.r.venkateshaprasad)@tudelft.nl, (matteo.vasirani,karl.aberer)@epfl.ch

ABSTRACT

Information and communication technology (ICT) infrastructure plays an important role to realize the full potential of Smart Grid applications. Smart grids utilize ICT entities to enhance efficiency, reliability and sustainability of power generation and distribution network. Majority of the architectures proposed hitherto focus only on a specific architectural aspect, like communication, storage, processing requirement, etc. Recent studies have shown that lack of knowledge on which architecture best satisfies certain information management requirements has hindered large scale smart grid deployments. In this paper, we investigate the cost-benefit analysis of four data processing architectures for various applications in smart grid. We introduce several key cost indicators to analyze hierarchical data processing architectures for the smart grid. In our evaluation, we consider realistic deployments for both dense and sparse environments. Results reported here are significant for smart grid designers, who can use them to discern the architecture that best fits the system requirements.

Categories and Subject Descriptors

C.0 [**Computer Systems Organization**]: General - System architectures; I.6.3 [**Simulation and Modeling**]: Applications

Keywords

Data Processing, Distributed Information Systems, Smart Grid, Cost-Benefit Analysis

1. INTRODUCTION

Smart Grid (SG) takes advantage of communication and control technologies to integrate the power infrastructure with an information infrastructure [1]. The *power infrastructure* comprises of an interconnected network of power systems that carries electricity from power plants to consumers. The various actors in power infrastructure includes generators, distributors, transformers, circuit breakers, etc. The *information infrastructure* comprises of ICT objects to measure and control power infrastructure and thus, supporting reliable and robust operation of the grid. The information infrastructure supports sensing, computation, control and information exchange capabilities. Actors in smart grid operate autonomously, but need to communicate with other actors to balance energy supply and demand. Smart grid is an ensemble of several applications such as demand response, demand forecast, emergency management, anomaly detection, adaptive pricing, etc. A fundamental building block for all these applications is Advanced Metering Infrastructure (AMI) - a system that measures, collects and analyzes data about energy usage [2].

Wireless Sensor Networks (WSNs) with sensing, computation, communication and control capabilities are widely being deployed to monitor energy consumption of consumers. Bidirectional communication between these devices and utility providers can provide immediate feedback on power usage, power quality, pricing details to the customers. An estimate from a utility provider indicates 22 gigabytes of data being generated every day from its 2 million customers [3]. The overwhelming data generated by smart meters call for developing information management mechanisms for large scale data storage and processing. While there have been deployments of SG (e.g., Grid4EU[1] and SmartWatts[2]) with a few participants, the design of suitable architecture to support envisaged SG applications on a large scale is an important research topic currently [4], [7].

In this paper we provide comprehensive insights about which architecture best satisfy certain information management requirements, such as the accuracy and granularity of collected data, or the privacy level. In particular, (i) we model different data processing architectures (centralized, decentralized, distributed and hybrid) for hierarchical power distribution networks; (ii) we consider realistic SG deployments in both dense (i.e., urban) and sparse (i.e., rural) environments; (iii) we introduce and model several key cost indicators, such as energy consumption, processing power, storage requirements, communication bandwidth, accuracy and privacy; (iv) we provide a detailed cost-benefit analysis of the proposed architectures, which Smart grid designers can use to select the architecture that best fits their requirements.

[1] http://www.grid4eu.eu/ [2] http://www.smartwatts.de/

WiMobCity'14, August 11, 2014, Philadelphia, PA, USA.

http://dx.doi.org/10.1145/2633661.2633665.

The remainder of the paper is organized as follows. Section 2 describes the related literature. Section 3 outlines the different data processing architectures proposed and Section 4 describe various key cost indicators considered for evaluation. Section 5 presents our evaluation setup and simulation parameters considered. We describe the performance results of each architecture and cost-benefit analysis in Section 6 and our conclusions are presented in Section 7.

2. RELATED WORK

The majority of architectures proposed hitherto focus only on a specific architectural aspect, like communication or storage or processing requirement. In recent SG deployments, smart meters collect data at an interval of 5 to 15 min compared to traditional way that only records the meter data once a month [2]. Data values obtained as an average over a 15 minute interval may not be sensitive to realize concepts such as advanced distribution automation, asset management, and appliance energy disaggregation [5]. Thus, smart meters in the near future may well measure values every 30 s, posing a significant challenge in processing and storage of huge amount of data generated.

A secure decentralized data-centric information infrastructure for the SG is proposed in [6]. Kim et al., describe challenges in low latency communication protocols, and security mechanisms for SG. Balancing supply and demand is mapped to constraint satisfaction problem and evaluated using a decentralized hierarchical architecture in [7]. Cloud based SG information management model is proposed in [8] and [9], along with a discussion on key challenges. The focus on cloud computing approaches is to provide adequate resources for the SG. *In contrast with the above works, in this paper we not only propose and analyze several architectures, but we also model important key cost indicators such as energy, communication, storage, processing, accuracy and privacy based on the physical topology of the grid.* Cloud based Demand Response (CDR) architecture using a distributed information infrastructure is proposed in [10]. Scalability aspects of data storage and processing of monthly bills in SG is investigated in [11]. Several data storage mechanisms like centralized relational database, distributed relational database and file systems are compared and evaluated. Similarly, scalability aspects of data communication for AMI application in SG is investigated in [4]. Zhou et al., study how communication cost scales with the number of smart meters and sampling frequency.

A comparison of centralized and distributed monitoring architectures for billing and demand response applications is proposed in [12]. Martinez et al., explore the potential benefits of having distributed architectures compared with centralized ones. In [12], authors evaluate the proposed architectures by considering a fixed number of houses. In comparison, our work improves on the existing literature to provide a comprehensive analysis of various data processing architectures with realistic environments *viz,* urban and rural environments. We consider accuracy and privacy cost indicators to provide a detailed cost-benefit analysis along with other cost indicators like energy, communication, storage, and processing. Thus we provide a holistic approach to model and analyze all key cost indicators in urban and rural environments. The proposed models and cost-benefit analysis are generic and can be applied to any smart grid deployment.

3. DATA PROCESSING ARCHITECTURES

The current topology of the power distribution network is arranged according to the voltage [13]. The distribution network is organized into multiple subgrids and consequently forming a hierarchical topology. In this paper, our architectural model adopts hierarchical topology of the power distribution networks. The key elements of our architectural model are the following.

Home Area Nodes (HANs)- are devices interconnected with the smart meter at the consumer premises. HAN receives energy consumption information from all appliances in the household and can employ mechanisms or receive information to match supply and demand at the household level.

Neighborhood Area Nodes (NANs)- act as an intermediate node between consumers and utility providers, and it serves a small geographical area, i.e., a neighborhood consisting of several houses. NAN receives information from the households within the neighborhood. Multiple NANs are deployed to cover utility's territory.

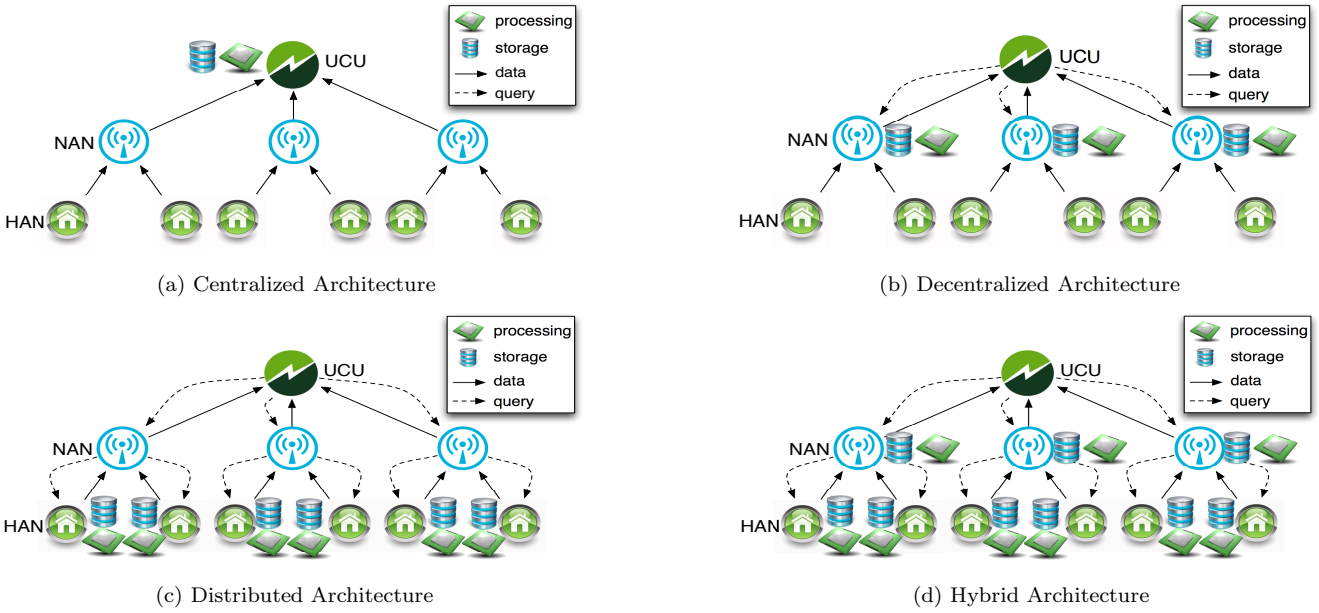
Utility Control Unit (UCU)- are the central control entity of utility providers. This node is responsible for billing, maintaining data, determining electricity price and carrying out demand response. UCU acts as the root node in our architectural model.

Design Choices

In our architectural model, the HANs at each household periodically senses energy consumption and transmits to the respective NAN. NANs act as the intermediate node between HANs and UCU. Communication between HANs and NANs is based on $sub-1GHz$ transceivers which are best suited for both indoor and outdoor environments [14]. The interconnection between NANs and UCU is based on IEEE 802.16 (WiMAX) which supports a maximum data rate up to 1 *Gbps*. The communication choice in this work is supported by some of the recent works [4], [13]. It should be noted that, the analysis in this paper can be further applied to any communication technology.

Data aggregation at the nodes can minimize the overall data communicated and also help in preserving sensitive information of customers. Three major data aggregation mechanisms considered in literature are:
a) *Time-wise aggregation*: where consecutive time-stamped energy consumption readings are aggregated to reduce the granularity of the data collected.
b) *Value-wise aggregation*: where similar energy consumption readings are bucketed to obtain discrete energy readings and thus reducing number of readings.
c) *Consumer-wise aggregation*: where the energy consumption values of several individual customers are aggregated into one time series to obfuscate the consumption of each individual customer.

In this work, we consider only time-wise data aggregation with different granularity. UCU can acquire information from the households by initiating a query and nodes can respond to the query depending on their roles. Based on the data aggregation, processing and storage capabilities of HANs, NANs and UCU, different architectures are proposed. By default, all nodes can send and receive a message, which is the minimum capability assumed at each node. The storage and processing icons in the Fig. 1, shows the addi-



(a) Centralized Architecture

(b) Decentralized Architecture

(c) Distributed Architecture

(d) Hybrid Architecture

Figure 1: Data processing architectures for the smart grids.

tional capability available at each node depending on the architecture.

3.1 Centralized Architecture

In centralized architectures as illustrated in Fig. 1a, only the UCU has data processing and storage capability. HANs periodically sense and transmit the energy consumption values to the respective NANs. NANs act as relays and forward it to the UCU. UCU has all the information and is responsible for processing and storage of the data. Thus, the information flow is uni-directional from HANs to UCU via NANs. No data aggregation is applied in centralized architecture.

3.2 Decentralized Architecture

In this architecture, only NANs have data processing and storage capabilities as shown in Fig. 1b. HANs transmit data periodically to the respective NAN similar to the centralized architecture. Instead of forwarding the data, NAN stores and processes this data locally. In decentralized architectures, since all data is available at the NAN, fine grained data aggregation is possible. NANs can aggregate hourly energy consumption and report to UCU. UCU generates queries to retrieve information from the NANs only when required. Thus, NANs act as central entities in this architecture.

3.3 Distributed Architecture

In distributed architectures, all HANs have data processing and storage capabilities. HANs periodically sense and store the energy consumption values locally. UCU initiates a query to fetch data, which is forwarded to the NAN and in turn to the HANs. HANs process the query and send the reply to UCU via NAN. Thus, making the architecture completely distributed as illustrated in Fig. 1c. HANs are assumed to have sufficient data storage and processing capability and communicate only upon reception of a query.

3.4 Hybrid Architecture

In hybrid architectures, HANs and NANs both have data processing and storage capabilities as shown in Fig. 1d. Hybrid architectures are extension of distributed architectures, where HANs not only sense and store but also transmit aggregated energy values to NANs. For instance, HANs can sense and store energy values periodically and at the end of the day send an aggregate energy consumption reading to the NAN. The data aggregation granularity may vary depending upon the applications considered.

4. COST INDICATORS

Architectures proposed in the previous section are characterized based on the availability of data storage and processing capabilities at the node, hence monetary costs of a node needs to be modeled in architecture evaluation. Data aggregation is employed by architectures to reduce the amount of data communicated as well as increasing the privacy of the customer. However, data aggregation results in decreased accuracy, since the UCU might need to disaggregate energy readings that have been aggregated over time by the HAN or the NAN. The trade-off between accuracy and privacy cost as a function of data aggregation granularity provides a key insight in the design of architectures. Our cost-benefit analysis hence considers monetary cost, accuracy and privacy as the key cost indicators to evaluate the performance of proposed architectures.

4.1 Monetary cost

Monetary cost C_M is the cost (in $) to deploy and operate the nodes in the architecture (HANs, NANs, UCU).

$$C_M = C_D + C_O. \quad (1)$$

The deployment cost C_D is a one-time cost that accounts for the deployment of storage, processing and communication

capacity.

$$C_D = C_S + C_P + C_T, \tag{2}$$

where C_S is the cost of storage, C_P is the cost of the processing units, and C_T is the cost of the transceivers.

The operational cost C_O is the cost incurred for the operation of the entire network for one month period.

$$C_O = E_{\text{total}} \cdot f_E, \tag{3}$$

where E_{total} is the average energy (in Joules) required by all nodes to be operational for a period of one month, and f_E is the price of energy (in \$/Joule), which is assumed to be constant.

Apart from these factors, the deployment and operational cost may include other factors such as cooling, sensors, peripherals and maintenance, which are not considered in our cost modeling. The components of C_D and C_O are described in detail in the following sections.

4.1.1 Energy consumption

The energy required for the operation of the entire network (expressed in Joules, J) includes the various activities the nodes can perform, such as reading from and writing into the storage, communicating, processing, etc. Energy consumption is calculated for the duration of one month. The energy consumed by a HAN is given by,

$$E_{\text{HAN}} = E_t^{H\to N} + E_{r/w}^{H} + E_p^{H}, \tag{4}$$

where $E_t^{H\to N}$ is the energy consumed for communication, $E_{r/w}^{H}$ is the energy consumed for reading from and writing into the storage, and E_p^{H} is the energy consumed for processing.

The energy consumption for communication is,

$$E_t^{H\to N} = e_{tx}^{H\to N}\ell_{tx}^{H\to N} + e_{rx}^{H\to N}\ell_{rx}^{H\to N}, \tag{5}$$

where $e_{tx}^{H\to N}$ ($e_{rx}^{H\to N}$) is the energy required for transmission (reception) of one byte of information between HAN and NAN, and $\ell_{tx}^{H\to N}$ ($\ell_{rx}^{H\to N}$) is the length in bytes of the messages that have been transmitted (received) by the HAN in one month. The energy consumption due to storage is defined as,

$$E_{r/w}^{H} = e_r^H \ell_r^H + e_w^H \ell_w^H, \tag{6}$$

where e_r^H (e_w^H) is the energy required to read (write) one byte of information, and ℓ_r^H (ℓ_w^H) is the length in bytes of the messages that have been read from (written into) the storage in one month. Finally, the energy consumption of processing is defined as,

$$E_p^H = e_p^H n_p^H, \tag{7}$$

where e_p^H represents the energy required for processing a byte of information at HAN, and n_p^H is the number of processed bytes.

Similarly, energy consumption for a NAN is given by,

$$E_{\text{NAN}} = E_t^{H\to N} + E_t^{N\to U} + E_{r/w}^{N} + E_p^{N}, \tag{8}$$

where $E_t^{H\to N}$ is the energy consumed for communication between HANs and NANs, $E_t^{N\to U}$ is the energy consumed for communication between NANs and UCU, $E_{r/w}^{N}$ is the energy consumed for reading from and writing into the storage, and E_p^{N} is the energy consumed for processing. These terms are defined as,

$$\begin{aligned} E_t^{H\to N} &= e_{tx}^{H\to N}\ell_{tx}^{H\to N} + e_{rx}^{H\to N}\ell_{rx}^{H\to N} \\ E_t^{N\to U} &= e_{tx}^{N\to U}\ell_{tx}^{N\to U} + e_{rx}^{N\to U}\ell_{rx}^{N\to U} \\ E_{r/w}^{N} &= e_r^N \ell_r^N + e_w^N \ell_w^N \\ E_p^{N} &= e_p^N n_p^N. \end{aligned} \tag{9}$$

Finally, for the UCU we have,

$$E_{\text{UCU}} = E_t^{N\to U} + E_{r/w}^{U} + E_p^{U}. \tag{10}$$

The terms $E_t^{N\to U}$ (energy consumption for communication), $E_{r/w}^{U}$ (energy consumption for storage reading/writing) and E_p^{U} (energy consumption for processing) are defined as,

$$\begin{aligned} E_t^{N\to U} &= e_{tx}^{N\to U}\ell_{tx}^{N\to U} + e_{rx}^{N\to U}\ell_{rx}^{N\to U} \\ E_{r/w}^{U} &= e_r^U \ell_r^U + e_w^U \ell_w^U \\ E_p^{U} &= e_p^U n_p^U. \end{aligned} \tag{11}$$

Thus, the total energy consumption for the entire network in a month is,

$$E_{\text{total}} = E_{\text{UCU}} + \sum_{i\in\mathcal{N}} E_{\text{NAN}}(i) + \sum_{j\in\mathcal{M}} E_{\text{HAN}}(j), \tag{12}$$

where $\mathcal{N}$ is the set of NANs and $\mathcal{M}$ is the set of HANs in the network.

4.1.2 Communication

The communication cost accounts for the data rate (expressed in bits per second, bps) needed to transmit data from a HAN to the UCU through a NAN. Data rate for a HAN is expressed as,

$$T_{\text{HAN}} = \frac{8\ \ell_m^{H\to N}}{t^{H\to N}}, \tag{13}$$

where $\ell_m^{H\to N}$ is the length of the message that has to be transmitted from the HAN to the NAN, and $t^{H\to N}$ is the time period within which a HAN needs to transmit its information to the NAN. Given T_{HAN}, the resulting monetary cost for communication at HAN j is,

$$C_T(j) = T_{\text{HAN}} \cdot f_T(T_{\text{HAN}}), \tag{14}$$

where $f_T(\cdot)$ is a non-linear function that models the cost of bandwidth (expressed in \$/bps). Similarly, data rate for a NAN is expressed as,

$$T_{\text{NAN}} = \frac{8\ \ell_m^{N\to U}}{t^{N\to U}}. \tag{15}$$

The resulting monetary cost for communication at NAN i is,

$$C_T(i) = T_{\text{NAN}} \cdot f_T(T_{\text{NAN}}). \tag{16}$$

Therefore, the total communication cost required for transmission between HANs to NAN and NANs to UCU is expressed as,

$$C_T = \sum_{i\in\mathcal{N}} C_T(i) + \sum_{j\in\mathcal{M}} C_T(j). \tag{17}$$

4.1.3 Storage

The storage cost accounts for the total amount of storage capacity (expressed in bytes) required by the node. The storage cost depends on the sampling interval τ and the

time durationΔ T for which storage is needed. Thus, the storage requirement for a node k is expressed as,

$$S_k = \Delta T \frac{\ell_m}{\tau}, \tag{18}$$

where ℓ_m indicate the length of a message. Depending on the architecture selected and application requirement, ℓ_m and τ varies for each HANs, NANs and UCU. Given S_k, the resulting monetary cost for storage at node k is,

$$C_S(k) = S_k \cdot f_S(S_k), \tag{19}$$

where $f_S(\cdot)$ is a non-linear function that models the cost of storage (expressed in \$/byte).

Thus total storage cost of the network for one month is given as,

$$C_S = C_S(\text{UCU}) + \sum_{i \in \mathcal{N}} C_S(i) + \sum_{j \in \mathcal{M}} C_S(j). \tag{20}$$

4.1.4 Processing

The processing cost accounts for the number of operations (ops) required to respond to a query received at the node. The in-node operations to respond to a query includes mainly arithmetic and relational operations. Processing cost depends on the number of messages to be processed and number of operations to be performed based on the query. The processing cost at node k calculated for one month is expressed as,

$$P_k = \sum_{q \in \mathcal{Q}} n_m \cdot n_q, \tag{21}$$

where $\mathcal{Q}$ is the set of queries generated in the network, which depends on the supported applications, n_m is the number of messages to be processed and n_q represents the number of operations to be performed for query q. These values depend on the architecture selected, the types of query and the node. Given P_k, the resulting monetary cost for processing at node k is,

$$C_P(k) = P_k \cdot f_P(P_k), \tag{22}$$

where $f_P(\cdot)$ is a non-linear function that models the cost of processing (expressed in \$/ops).

Thus total processing cost of the network for one month is given as,

$$C_P = C_P(\text{UCU}) + \sum_{i \in \mathcal{N}} C_P(i) + \sum_{j \in \mathcal{M}} C_P(j). \tag{23}$$

4.2 Accuracy cost

As mentioned before, in order to reduce storage and communication, a possible strategy is to do a time-wise aggregation of consecutive energy readings. In this way, an original time-series X of n readings is reduced to a smaller time-series Y of length m, (where $m < n$) by aggregating each k consecutive values in X into a single value y in Y. However, certain applications may need to restore the original time series X from Y, using a disaggregation algorithm. The restored time-series $\widehat{X}$ may differ from the original one, X.

Accuracy is therefore an important measure of how accurately the original data can be retrieved from aggregated data. We utilize Normalized Root Mean Square Error (NRMSE) as our accuracy cost. Let $x_t \in X$ be the real energy consumption value of a HAN j at time t, and $\widehat{x}_t \in \widehat{X}$ the energy consumption value that has been inferred through the disaggregation algorithm. The NRMSE is expressed as,

$$\text{NRMSE}(j) = \frac{\text{RMSE}(j)}{x_{\max} - x_{\min}}, \tag{24}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum real energy consumption values of X, and

$$\text{RMSE}(j) = \sqrt{\frac{\sum_{t=1}^{n} (x_t - \widehat{x}_t)^2}{n}}. \tag{25}$$

The accuracy cost C_A is therefore defined as the average NRMSE among all the HANs. Formally,

$$C_A = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \text{NRMSE}(j). \tag{26}$$

4.3 Privacy cost

Although the compression of the original time-series X into Y through time-wise aggregation reduces the accuracy of the restored time-series $\widehat{X}$, it also enhances the customer privacy. In fact, Y can be considered as an obfuscated version of X. To quantify the privacy of the aggregation of k consecutive values into a single aggregated value y, we use Shannon entropy [15] associated with the disaggregation of y into k values. In general, higher the entropy, the higher is the customer privacy. The entropy of a system with $\mathcal{S}$ states is expressed as,

$$H(y) = \sum_{s \in \mathcal{S}} -p(s) \cdot log(p(s)), \tag{27}$$

where $p(s)$ is the probability that the system is in state s. In our case, $\mathcal{S}$ is the set of all possible disaggregations, i.e., all the possible ways a value y can be split into k values such that the sum of the k values equal y. The number of possible disaggregations (i.e., the state space size $|\mathcal{S}|$) is called weak integer composition of y into k parts, and it is computed as,

$$|\mathcal{S}| = \binom{y+k-1}{k-1} = \frac{(y+k-1)!}{(k-1)! \cdot y!}.$$

Assuming that each disaggregation of y into k values has the same probability, we can rewrite Eq. (27) as,

$$H(y) = log\left(|\mathcal{S}|\right). \tag{28}$$

The average entropy of the aggregated time-series Y of HAN j is,

$$H(j) = \frac{\sum_{y \in Y} H(y)}{|Y|}. \tag{29}$$

Thus, the privacy cost C_H of a data processing architecture is defined as,

$$C_H = -\frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} H(j). \tag{30}$$

In the next section we present our evaluation setup and simulation parameters considered.

5. EVALUATION SETUP

Applications in SG have different data requirements, which imply different data acquisition queries generated by the UCU. In our evaluation we consider Billing and adaptive pricing (BAP), Demand Response (DR), Demand Forecast

Table 1: Queries generated (E=energy consumption)

Queries generated	BAP	DR	DF	EM
E / 10 seconds		X	X	
E / 30 seconds		X		X
E / 15 minutes		X	X	X
E / hour	X	X		
E / day	X	X	X	X
E / month	X		X	X
(Min, Max, Avg)E / day		X	X	
(Min, Max, Avg)E / month	X			X

Table 2: Energy consumption for different operations. [17]

Operations	Energy consumption
Transmission @sub-1GHz	0.164 mJ/byte
Reception @sub-1GHz	0.08 mJ/byte
Transmission @IEEE 802.16	0.324 mJ/byte
Reception @IEEE 802.16	0.100 mJ/byte
Read from flash	0.09 μJ/byte
Write to flash	0.8 μJ/byte
Processing	0.14 μJ/byte

(DF) and Emergency Management (EM) applications. The requirements of the applications considered for our cost-benefit analysis are described in Table 1.

Billing and adaptive pricing (BAP). In the future, utility providers will be able to bill consumers based on the real-time demand-supply balance. Consumers will also get real-time pricing information in order to alter their energy demand. Thus, queries related to minimum, maximum and average energy consumption, as well as hourly and monthly energy consumption are generated by UCU for this application.

Demand-response (DR). DR strategies are designed to reduce or shift energy consumption from peak periods to off-peak periods. Thus, energy consumption readings at high frequency during peak periods and low frequency consumption readings at off-peak periods are required to envisage DR.

Demand forecast (DF). Demand forecast algorithms can assist utility providers towards efficient distribution of electricity and better planning of resources. Aggregate energy consumption readings and high frequency readings at peak periods can assist in accurate demand forecast.

Emergency Management (EM). Cascading failures and robustness of the grid are some of the challenges that are handled using emergency management strategies. To detect abnormal energy consumption patterns, readings at high frequency are required.

5.1 Environment

In our cost-benefit analysis we consider two environments *viz*, *urban* and *rural*. We define the number of HANs and NANs in an urban and rural setup based on Electric Power Research Institute (EPRI) [16] survey about the NAN, population density and number of households in the USA.

Average total population in an urban environment is around $4.8M$, with a maximum population density of roughly $33.7K$ persons per km^2 and land area of 121 km^2. Thus, an urban environment is composed of $1.6M$ households. To provide adequate coverage to the collection of energy data from the households, 73 NANs operating at sub-1GHz are required [16].

Rural environments with different terrain and population density is considered to have total population of $1.4M$ and land area of 215000 km^2. Thus, rural environment consists of around $476K$ households, with 76 NANs operating at sub-1GHz to provide coverage [16].

5.2 Simulation parameters

In this work, a standard wireless sensor node (WSN) is considered as HAN and its configuration depends on the architecture. Each HAN samples data by default every 5 minutes, which can be programmed based on the requirement or upon reception of the query. Each HAN is associated with a sub-1GHz transceiver to communicate with the NAN. Similarly, NANs are equipped with both sub-1GHz and WiMAX transceivers to communicate with HANs and UCU respectively. Table 2 shows the energy consumption for different operations performed by the HAN.

In this work, *Data* message contains HAN number, time stamp and energy consumption values. The *Query* message includes the HAN number and query number. Similarly, the *Query-reply* message carries the energy consumption value, HAN number and query number. Finally, the *Aggregated data* includes HAN number, aggregation granularity and aggregated energy consumption value. Message size of data, query, query-reply, aggregated messages are considered to be 50, 5, 10 and 10 bytes respectively.

6. PERFORMANCE RESULTS

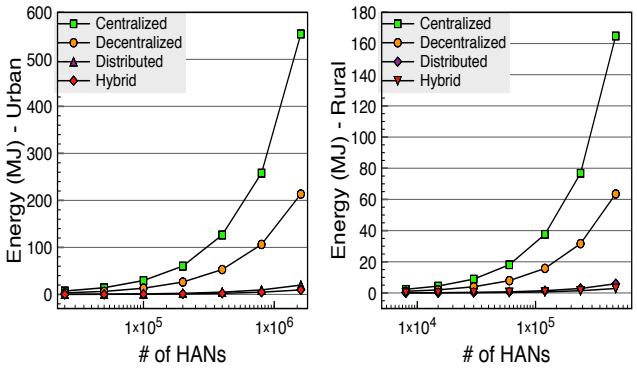
This section describes the performance of each architecture based on the key cost indicators for urban and rural environments. To calculate the key cost indicators, we used the data over a duration of one month in our simulations.

6.1 Energy consumption

Energy consumption cost per architecture for both urban and rural environments[3] is illustrated in Fig. 2a and Fig. 2b. In *urban environments*, it is evident that centralized architecture consumes significant amount of energy compared to other architectures. In centralized architecture, complete data needs to be relayed to the UCU, thus increasing the number of transmissions and the energy required. Distributed and hybrid architectures consume much lower energy compared to centralized and decentralized architectures. The significant energy saving in distributed approaches is due to the reduced number of transmissions. Energy consumption of the hybrid architecture is the lowest compared to all other architectures. This energy saving is achieved by sending aggregated data to NANs as compared to storing data only at HANs, as in distributed architecture.

In general, the total energy consumption increases rapidly as the number of houses increases for centralized and decentralized architectures. In case of distributed and hybrid architectures, the increase in energy consumption is very

[3] In our experimental evaluation we considered two cases: (i) each NAN has the same number of HANs, and (ii) each NAN has a uniformly distributed random number of HANs. We found that there is not much difference in energy consumption between the two cases. Thus, for simplicity we consider equal number of HANs being allocated to each NAN.



(a) Urban (b) Rural

Figure 2: Energy consumption across architectures.

Table 3: Energy consumption distribution for urban environments with $1.6M$ HANs.

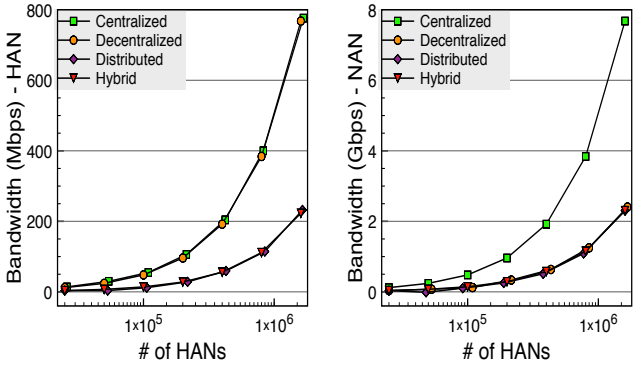
Architectures	Storage	Proc.	Comm.	Total Energy
Centralized	4%	12.6%	83.4%	554.0 MJ
Decentralized	2.3%	16.2%	81.5%	213.5 MJ
Distributed	15.9%	20.4%	63.7%	19.8 MJ
Hybrid	7.1%	3%	89.9%	9.6 MJ

gradual, thus increasing their scalability. Similar trends can be seen for rural environments, as shown in Fig. 2b.

Table 3 shows energy consumption distribution for different architectures (number of HANs = $1.6M$). We remark that the energy consumption considers only communication, storage and memory operation, although other factors could be considered, such as cooling, lights, etc. It is evident that the most significant energy factor in all the architectures is communication. Hence, reducing communication needs can in turn reduce overall energy consumption, as can be seen in distributed and hybrid architectures.

6.2 Communication

The communication cost as described in Section 4.1.2 is the average data rate required to support the SG applications considered in this work. The time of reference $t^{H\rightarrow N}$ and $t^{N\rightarrow U}$ in Eq. (13) and Eq. (15) are considered to be

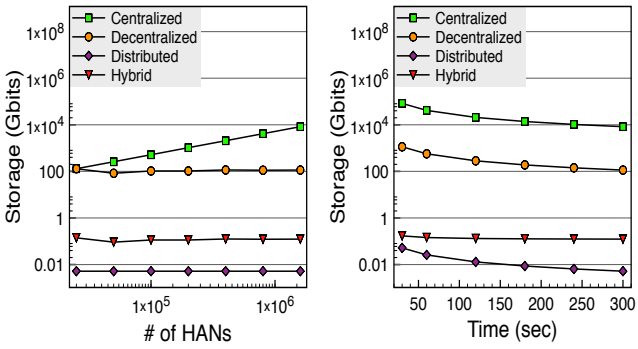


(a) Bandwidth-HAN (b) Bandwidth-NAN

Figure 3: Bandwidth required in urban environment.

Table 4: Bandwidth required for various architectures.

	Urban		Rural	
Architectures	HAN	NAN	HAN	NAN
Centralized	480 bps	11 Mbps	480 bps	3 Mbps
Decentralized	480 bps	32 Mbps	480 bps	1 Mbps
Distributed	144 bps	3 Mbps	144 bps	1 Mbps
Hybrid	138 bps	2 Mbps	138 bps	0.5 Mbps



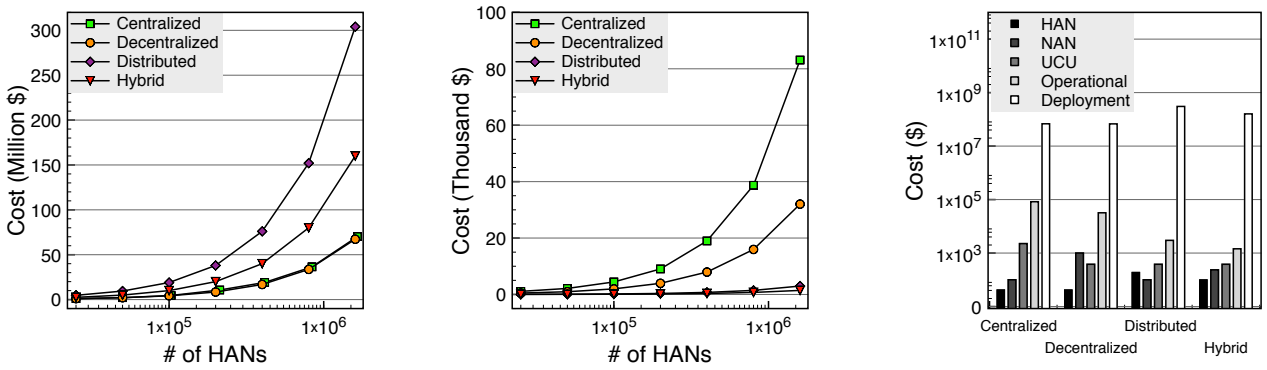
(a) # of HANs (b) Sampling Interval

Figure 4: Required storage in urban environment.

1 s. Table 4 shows the average bandwidth requirement at each HAN and NAN for both urban and rural environments with $1.6M$ HANs and $476K$ HANs respectively. Data rate requirement for a HAN is same, irrespective of the environment, as each HAN transmits the same data based on the architecture selected. However, the data rate required at NANs in urban environment is higher than rural environment, since more HANs are associated with each NAN in an urban environment. It is evident that the bandwidth supported by our communication choice of sub-1GHz and WiMAX can indeed allow all network operations. The bandwidth requirements from HANs to NAN and NANs to UCU in an urban environment are shown in Fig. 3a. The needed bandwidth between HANs and NAN is higher for centralized and decentralized architectures.However, since distributed storage and processing is adopted in distributed and hybrid architectures, the number of transmissions performed at the HAN is reduced. Thus, the bandwidth requirement is significantly reduced in these architectures. Similarly, the average bandwidth requirement between NANs and UCU is shown in Fig. 3b. In general, the bandwidth increases with the number of houses as seen in the Fig. 3. Similar trends with scaled-down bandwidth requirements are observed for rural environments.

6.3 Storage

Storage required by each node for different architectures in urban environment, for duration$\Delta\ T = 1$ year is shown in Fig. 4a. The default sampling interval of 5 minutes is considered to determine the storage cost as described in Eq. (18). As with other costs, storage cost for centralized architecture is the highest compared to other architectures, as all data is stored at one place i.e. the UCU. The storage required by other architectures is much lower than the centralized architecture, with distributed architecture having the low-



(a) Deployment cost (b) Operational cost (c) Cost distribution

Figure 5: Monetary cost details for deployment, operational and each node in urban environment.

Table 5: Processing operations in urban environment.

Architectures	Number of operations
Centralized	9000 M-ops
Decentralized	9000 M-ops
Distributed	0.36 M-ops
Hybrid - NAN	15.75 M-ops
Hybrid - HAN	0.008 M-ops

est. As the number of HANs increase, the required storage increases linearly in centralized architectures. On the other hand, since equal number of HANs are allocated to each NAN, for all the other architectures storage is constant, as seen in the Fig. 4a.

Finally, the required storage as a function of sampling interval is shown in Fig. 4b. Higher sampling intervals indicate less frequent sensing of energy values. Storage cost in general decreases with increase in sampling interval, regardless of the architecture.

6.4 Processing

Processing accounts for the number of operations performed to respond to a query as described in Section 4.1.4. Processing requirements depend on the number of messages the node has to process before replying. For each query, all messages until the reception of query is processed and each query is independent of other queries. Thus the processing requirement depends on when the query is received (in turn number of messages to be processed) and the operations performed. The processing requirement for one month duration in an urban environment with 1.6M HANs is shown in Table 5.

In centralized architectures, since UCU performs all processing, the processing requirements increase with the number of houses. Since, number of HANs per NAN is constant, processing at each NAN in decentralized architectures is merely a constant with increase in number of houses. In a distributed architecture, processing is done in a distributed manner at each HAN, thus reducing the number of operations at each HAN by order of four compared to decentralized architecture. In hybrid architectures, processing effort is distributed at both HANs and NANs and has the least processing cost at each HAN. Similar trends are also observed for rural environments.

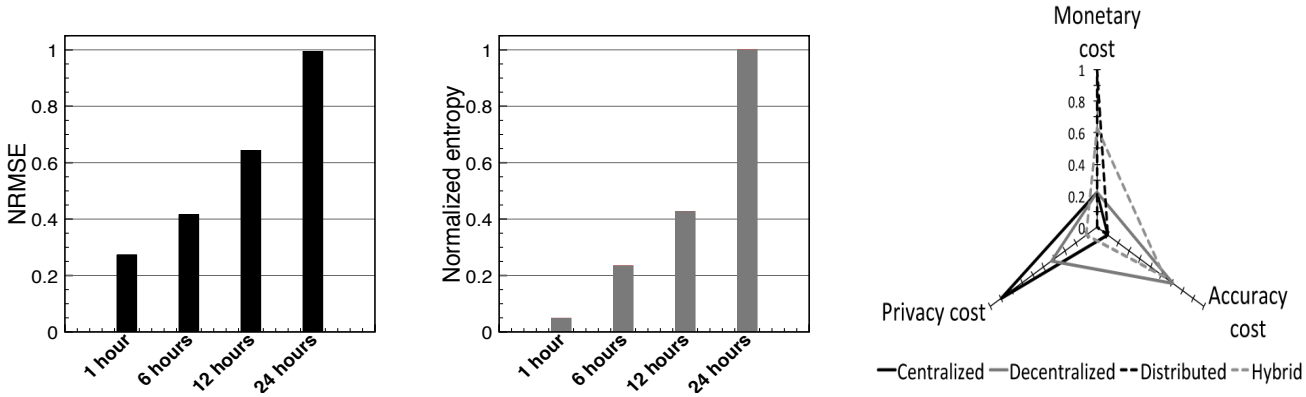
6.5 Cost-Benefit Analysis

Monetary cost. The monetary cost as described in Section 4.1 accounts for deployment and operational costs. Deployment cost is the cost for installing nodes (HANs, NANs, UCU) and varies based on the capabilities provided to each node with respect to processing, storage and communication. Operational cost is calculated based on the energy required to operate all nodes for one month. The cost of electricity (f_E) in Eq. (3) is considered to be 0.194 $/KWh, based on Pacific Gas and Electric Company[4] yearly average electricity costs. Based on the storage, processing and communication requirements obtained in previous Section, appropriate modules and price details are considered from digikey[5]. In general, $f_S(\cdot)$ in Eq. (19) model the cost of storage per byte and is considered to vary from 2$ to 60$ for 256KB to 500GB of storage. Similarly $f_T(\cdot)$ in Eq. (14) models the cost of transceivers and is considered to be 5$ and 10$ for $sub-1GHz$ and WiMAX transceivers respectively. $f_P(\cdot)$ in Eq. (22) model the cost of processing and varies from 5$ (MSP43016xx processor) to 60$ (ARM Cortex-M3 processor).

The deployment cost for each architecture as a function of number of houses is shown in Fig. 5a. Since processing and storage is performed by only UCU in centralized architecture, the total deployment cost is the lowest compared to all other architectures. In decentralized architectures, all the NANs have storage and processing capabilities. The distribution of processing and storage capabilities to the NANs overcome single point failure but follows the same trend in monetary cost as compared to centralized architectures. In distributed architecture, each HAN is equipped with processing and storage capabilities, thus increasing the monetary cost of each node in the network. Due to sheer number of HANs, the total deployment cost of the distributed architecture increases rapidly with number of HANs. The deployment cost in hybrid architectures is lower compared to a distributed architecture, since processing and storage is distributed at both HANs and NANs.

The operational cost for various architectures in urban environment is shown in Fig. 5b. Similar to energy consumption cost (refer Fig. 2), the operational cost is the highest for centralized architectures and lowest for hybrid architectures.

[4] http://www.pge.com/. [5] http://www.digikey.com/



(a) NRMSE (b) Normalized entropy (c) Cost-Benefit analysis

Figure 6: Accuracy, privacy and cost-benefit analysis.

Fig. 5c shows the break-up of monetary cost incurred for each architecture in urban environment with 1.6M HANs. The figure shows the cost in $ for each HAN, NAN, UCU (with storage, processing, communication), as well as the operational and deployment cost across various architectures. As seen in the figure, adding storage and processing features to each HAN increases the deployment cost of distributed and hybrid architectures. However, distributed and hybrid architectures has the least operational cost compared to centralized and decentralized architectures and hence more energy efficient.

Accuracy cost. Centralized and distributed architectures do not employ data aggregation, as data is stored either in UCU or HAN. Since complete data is available at each NAN in a decentralized architecture, low data aggregation granularity of 1 hour is employed. In hybrid architecture, NAN receives data aggregated with granularity every 12 hours from each HAN. In this work, NRMSE is the metric used to determine the accuracy based on data aggregation granularity. To calculate the accuracy cost, we used two weeks of data collected by the REDD initiative [18]. Data aggregation granularity of 1, 6, 12, 24 hours are considered, while the step-size of the time-series to be restored is assumed to be 5 minutes. The focus of this paper is not on accurate disaggregation algorithms. For this reason, we employ a rather simple algorithm that equally split an aggregated reading into the 5 minute buckets of the time-series to be restored. Fig. 6a shows the NRMSE values for various data aggregation granularities. It is evident that, higher the data aggregation granularity, the higher is the NRMSE and thus lower is the accuracy.

Privacy cost. This cost factor also depends on the data aggregation granularity. Intuitively, higher the data aggregation granularity, the higher is the entropy and thus lower is the privacy cost. In centralized architectures, all sensitive data of customers is available at UCU, thus making centralized architecture the least privacy-preserving architecture. On the other hand, distributed architectures with distributed storage ensure that customer's sensitive data is stored locally at HAN, making them completely privacy-preserving.

However, privacy cost varies in decentralized and hybrid architectures. As before, we used two weeks of data collected by the REDD initiative [18] to calculate the privacy cost. Fig. 6b shows the average entropy for different data aggregation granularities. The figure shows that the higher the data aggregation granularity, the higher the entropy, therefore the lower the privacy cost (see Eq. (30)).

Cost-Benefit Analysis. The radar plot in Fig. 6c shows the performance of each architecture according to the monetary, accuracy and privacy costs. All cost indicators have been normalized to the [0,1] interval. Clearly, the lower the value of monetary, accuracy and privacy costs, the more desirable is the architecture for the smart grids.

Centralized architectures has low monetary cost, high privacy cost and low accuracy cost, thus making them less privacy-preserving but economically cheaper, since all data is stored and processed at the UCU. Thus, centralized architectures are less scalable and suffer from single point of failure but with low deployment cost.

Decentralized architectures on the other hand have low monetary cost, moderate privacy cost and moderate accuracy cost. The privacy cost reduction is achieved with data aggregation at NANs, which also increases the accuracy cost. Decentralized architecture distributes the processing and storage efforts to the NANs thus achieving moderate privacy and low monetary cost.

Distributed architectures have the highest monetary cost, the lowest privacy and accuracy costs. Distributed architectures are clearly the most privacy-preserving system, since data is stored and processed locally, however this increases the monetary cost of the architecture.

Hybrid architectures have lower monetary cost compared to distributed architectures, with low privacy cost and moderate accuracy cost. Distributing storage and processing at both HANs and NANs reduces the deployment cost, with a loss in accuracy due to data aggregation at NANs. Also, hybrid architectures are more energy efficient but with high deployment cost.

7. CONCLUSIONS

The future Smart Grid - relying upon an extensive ICT infrastructure - will be fundamentally different from the current power distribution systems. We presented different data processing architectures for the smart grids. This paper presents the first step towards understanding and modeling

the key cost indicators for large scale deployment of smart grids. To gain insights, the proposed architectures were evaluated based on energy consumption, processing power, storage requirements, communication bandwidth, accuracy of data collection and privacy. From our cost-benefit analysis, we can conclude that even though centralized architectures perform well in terms of accuracy and deployment cost, they are less scalable and least privacy-preserving. On the other hand, distributed architectures overcome privacy issues with local storage and processing, but with additional deployment cost. Decentralized architectures perform well in terms of accuracy and monetary cost, while hybrid architectures increase the privacy by increasing the deployment cost. Thus, the choice of the architecture could be to have a more energy efficient architecture or highly scalable distributed architecture with high deployment cost or simple less scalable architecture and depends upon the objective of the implementation.

8. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's Seventh Framework Programme (EU FP7) under grant agreement no. 287708 (iCore), 288322 (Wattalyst) and 288021 (EINS).

9. REFERENCES

[1] X. Fang, S. Misra, G. Xue, and D. Yang. Smart grid - the new and improved power grid: A survey. *IEEE Communications Surveys and Tutorials*, 2012.

[2] U.S. Dept. of Energy. Advanced metering infrastructure. February 2008, [Online] https://www.smartgrid.gov/sites/default/files/pdfs/advanced_metering_infrastructure_02-2008.pdf.

[3] M. Shargal and D. Houseman. The big picture of your coming smart grid, 2009, [Online] http://www.smartgridnews.com/artman/publish/commentary/The_Big_Picture_of_Your_Coming_Smart_Grid-529.html

[4] J. Zhou, R. Hu, and Y. Qian. Scalable distributed communication architectures to support advanced metering infrastructure in smart grid. *Parallel and Distributed Systems, IEEE Transactions on*, 2012.

[5] S. Akshay Uttama Nambi, T. G. Papaioannou, D. Chakraborty, and K. Aberer. Sustainable energy consumption monitoring in residential settings. In *INFOCOM, Proceedings*, IEEE, 2013.

[6] Y.-J. Kim, M. Thottan, V. Kolesnikov, and W. Lee. A secure decentralized data-centric information infrastructure for smart grid. *Communications Magazine, IEEE*, 2010.

[7] D. Rech and A. Harth. Towards a decentralised hierarchical architecture for smart grids. In *Proceedings of the Joint EDBT/ICDT Workshops*, USA, 2012. ACM.

[8] S. Rusitschka, K. Eger, and C. Gerdes. Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain. In *IEEE SmartGridComm*, 2010.

[9] X. Fang, D. Yang, and G. Xue. Evolving smart grid information management cloudward: A cloud optimization perspective. *Smart Grid, IEEE Transactions on*, 4(1):111–119, 2013.

[10] H. Kim, Y.-J. Kim, K. Yang, and M. Thottan. Cloud-based demand response for smart grid: Architecture and distributed algorithms. In *IEEE SmartGridComm*, 2011.

[11] M. Arenas-Martinez, S. Herrero-Lopez, A. Sanchez, J. Williams, P. Roth, P. Hofmann, A. Zeier. A comparative study of data storage and processing architectures for the smart grid. In *IEEE SmartGridComm*, 2010.

[12] Martinez R, Ramos, F, Gormus, S et al. A Comparison of Centralized and Distributed Monitoring Architectures in the Smart Grid. In *Systems Journal*, IEEE, Dec. 2013.

[13] A. Aggarwal, S. Kunta, and P. Verma. A proposed communications infrastructure for the smart grid. In *Innovative Smart Grid Technologies (ISGT), 2010*, 2010.

[14] S. Aust, R. Prasad, and I. Niemegeers. Performance evaluation of sub 1 ghz wireless sensor networks for the smart grid. In *IEEE Local Computer Networks*, 2012.

[15] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[16] Electric Power Research Institute *Wireless Field Area Network Spectrum Assessment*, 2010. [Online]

[17] T. Prabhakar, S. N. Akshay, R. Venkatesha Prasad, S. Shilpa, K. Prakruthi, and I. Niemegeers. A distributed smart application for solar powered wsns. In *NETWORKING, LNCS*, Springer, 2012.

[18] J. Z. Kolter and M. J. Johnson. Redd: A public data set for energy disaggregation research. In *proceedings of the SustKDD workshop*, 2011.