The development of smart grid brings great improvement in the efficiency, reliability, and economics to power grid. However, at the same time, the volume and complexity of data in the grid explode. To address this challenge, big data technology is a strong candidate for the analysis and processing of smart grid data. In this article, we propose a big data computing architecture for smart grid analytics, which involves data resources, transmission, storage, and analysis. In order to enable big data computing in smart grid, a communication architecture is then described consisting of four main domains. Key technologies to enable big-data-aware wireless communication for smart grid are investigated. As a case study of the proposed architecture, we introduce a big-data-enabled storage planning scheme based on wireless big data computing. A hybrid approach is adopted for the optimization including GA for storage planning and a game theoretic inner optimization for daily energy scheduling. Simulation results indicate that the proposed storage planning scheme greatly reduces the cost of consumers from a long-term view.

**Abstract**

Smart grid is an innovation of power grid with a high integration of advanced monitoring, sensing, communication, and control technologies in order to provide sustainable, economic, and secure power services to customers. With the rapid development of smart grid, large amounts of smart meters (SMs) and sensors have been deployed with huge coverage. As a result, a large number of multi-sourced heterogeneous smart grid data has been produced. Enormous value can be extracted from these smart grid data which can not only increase the quality of smart grid from the view of utility companies, but also provide better service for different types of customers than traditional power grid. Therefore, the application of big data technology can bring huge benefits for smart grid. For instance, IBM utilized 4 petabytes of climate and environmental history data; a wind turbine location model was designed to estimate suitable installation locations of fans. As a result, the efficiency of wind turbine production has been improved and the service life has been extended. However, big data computing in smart grid brings stringent requirements for wireless communication technologies due to the vast volume, variety, and velocity of smart grid data. Compared to traditional wired communication technology, wireless communication technologies have some unique advantages in deployment, expansion, and cost efficiency. Moreover, with advanced technologies, wireless communications are also characterized by high data rate and reliability. Much research has been conducted on wireless communication technologies applied in smart grid. For example, Yu et al. [2] studied the cognitive radio technology applied in smart grid to achieve optimization of spectrum resource management. Overall, wireless technologies provide promising support for big data computing in smart grid.

Wireless big data computing has wide applications in smart grid, which can be generally categorized into four aspects: customer profiling, demand response, network planning, and pricing. As a case study of the proposed smart grid big data computing architecture, we introduce a residential energy storage planning mechanism in this article. Some work has been done previously on the planning problem of energy storage devices owned by utility companies [3]. However, to the best of our knowledge, there is still little research on the planning of consumer-side energy storage devices. In this article, we focus on the planning of energy storage devices deployed at each consumer's home based on historical energy consumption data. A genetic algorithm (GA)-based approach is introduced to obtain the optimal capability to deploy energy storage devices for each consumer considering long-term total cost. In addition, an energy scheduling scheme is proposed based on game theory for the daily optimization in energy storage planning. As a unique advantage of game theocratic approaches, the Nash equilibrium derived from the proposed energy scheduling game can give all consumers a satisfactory reduction in billing.

This article focuses on the issue of wireless big data computing in smart grid. We first investigate the consistency between the characteristics of big data and smart grid data. To this end, a big data computing architecture is proposed. Four levels are introduced in the architecture: data resource, data transmission, data storage, and data analysis. The data transmission level plays an important role in bridging the other three levels as well as connecting the network with the data level. To address the data transmission in smart grid big data computing, a hierarchical big-data-aware wireless communication architecture is then described. Possible wireless technologies are enumerated for different domains among the architecture. To enable big-data-aware wire-
Big Data Computing Architecture in Smart Grid

With the establishment of a large number of information management systems in smart grid, energy big data is generated ranging from power generation to power transmission, power distribution, and loads, motivating much research into big data for smart grid. Jiang et al. [4] overviewed the related analytic frameworks and the key enabling technologies to achieve a comprehensive understanding of big data application in smart grid. Hu et al. [5] addressed the developments and challenges of big data security in smart grid and gave several taxonomies of energy big data analytics in this field. Overall, big data in smart grid is a very broad area, and the research in this field is still at an early stage. On one hand, the characteristics of big data in smart grid pose new challenges in data computing where the conventional technologies unable to deal with. On the other hand, big data computing infrastructure in smart grid has yet to be unified, and practical difficulties of data acquisition as well as analysis may be encountered.

In this section, we first investigate the characteristics of smart grid data. These features dictate the applicability of big data technologies in smart grid. Subsequently, we propose a big data computing architecture for smart grid, consisting of four main levels: data sources, data transmission, data storage, and data analysis, as shown in Fig. 1. Big-data-aware wireless communication for smart grid at the data transmission level and a big-data-enabled energy storage planning scheme as a case study are described in the following two sections.

Big Data Characteristics of Smart Grid Data

The produced data in smart grid exhibit the characteristics of big data as well as their unique characteristics, that is, volume, velocity, variety, value, and veracity.

Volume: In order to obtain accurate and real-time running status information of devices in smart grid, the number of collection points and the frequency of sampling have increased. The smart grid data volume has become huge, ranging from terabytes to petabytes. The large size of data has brought challenges in data storage, transmission, and analysis.

Velocity: There are two aspects to the requirement for speed of smart grid data processing. On one hand, the real-time data require to be exhibited rapidly in order to truly and completely record the production operation of each detail as well as the complete reflection of the production process. On the other hand, the overall data require to be analyzed within the whole data set in the range of a sampling period in order to support decision making. The smart grid data proposes higher requirements on the processing performance of online data than that of offline data.

Variety: There are many structured, unstructured, and semi-structured data in smart grid, including ordered records, location information, social data sets, tables, and graphs. The unstructured data account for a large proportion of smart grid data and have been growing exponentially by 60 percent [6]. In addition, the frequency of query and processing and performance requirements for these data are not the same either.

Value: The value of smart grid data is directly related to the application environment of data. However, the value density is low. In the case of video, the useful data may be only 1 to 2 s in the continuous monitoring process. The same issue exists in the power transmission equipment state monitoring. There is only a very small amount of abnormal data, but most collected data are normal.

Veracity: Authenticity can be considered as a new attribute of smart grid data for certain requirements of data quality. The authenticity of smart grid data refers to the level of reliability related to the specific type of data. High-quality data achieve an important influence on the effects of data analyzing results. However, even the best data cleaning method cannot remove the inherent unpredictable nature of certain data.
As a basic unit of a smart grid communication system, HAN mainly consists of a variety of smart appliances and a smart meter that collects the real-time power information of smart appliances through wireless communication for energy consumption management.

The correspondence between the characteristics of smart grid data and big data indicates huge numbers of potential applications for big data technologies in smart grid. To this end, a big data computing architecture for smart grid analytics is proposed in the next subsection.

**Big Data Computing Architecture for Smart Grid Analytics**

Considering the generation, transfer, and processing of big data, big data computing is usually summarized into a hierarchical architecture. In this subsection, we propose a big data computing architecture especially for smart grid analytics, and the four levels are introduced as follows.

**Data Source Level:** Smart grid data are generated from different data sources which are longitudinal and/or distributed, that is, distribution and transmission system data, phasor measurement unit (PMU) and advanced metering infrastructure (AMI) data, distributed generation data, intelligence application related data, and so on. These data are distributed in different places and managed by different companies/departments, which belong to different systems, for example, the collection information system (CIS), wide area measurement system (WAMS), distribution management system (DMS), energy management system (EMS), financial management system (FMS), meteorological information system (MIS), geographic information system (GIS), public service, and the internet. These data are not completely independent. There is a complex relationship in the smart grid data that are correlated and influenced by each other. For instance, the meteorological conditions and social economic situation may affect the users’ electricity consumption, which may also affect the electricity market transactions. Power market data can provide the basis for decisions of the relevant public service sector, and the GIS data generated from power enterprise has to use the municipal planning data as a reference. Smart grid big data generated by sensors, infrastructure, and applications are transferred to the network resource database for centralized management, that is, a centralized or cloud data center.

**Data Storage Level:** Smart grid data collected from the data sources contain a lot of meaningless information, which does not need to occupy a large amount of storage space. This brings a new challenge for the coming data computing. Therefore, the smart grid data in a data center may be stored and processed in two modes: stream processing and batch processing [7]. Stream processing means that data are processed immediately when new data arrive and then the demanded results are returned. This model is applicable to the business with high real-time requirements in smart grid, including joint scheduling of power supply and load, online equipment monitoring, and so on. Batch processing means that the data are stored before processing. This model is suitable for the planning of smart grid with low real-time demands but very huge size and complicated business [8].

**Data Analysis Level:** Various analytical methods and tools have been utilized to extract value in different application fields categorized by two directions. One is customer-oriented service, such as demand-side management (DSM). In DSM, customers are categorized based on different characteristics, such as climatic conditions and energy consumption patterns. Based on the category analysis, a total demand response in a certain region or a class of users can be obtained through clustering. As a result, the analysis results can provide fundamentals for the development of demand management response. The other direction is utility-company-oriented service, such as load forecasting [9]. In load forecasting, distributed energy sources and microgrids increase the complexity of load forecasting with wide applications and high prediction accuracy. Thus, a lot of factors need to be considered in load forecasting.

**Data Transmission Level:** This level builds a bridge among the other three levels and also enables the intra-communication within each level. Confronted with the challenge of big data collection, processing, and analysis, there are enormous requirements for advanced communication technologies. Compared to wired technologies, wireless communications show unique advantages, such as cost efficiency and flexibility. In the next section, wireless technologies are introduced in detail for the implementation of big data computing in smart grid.

**Big-Data-Aware Wireless Communication for Smart Grid**

In this section, we introduce a hierarchical wireless communication architecture for smart grid big data divided into four domains. Some key technologies to enable the big-data-aware wireless communication are then described in detail.

**Hierarchical Wireless Communication Structure For Smart Grid Big Data**

Wireless communication in smart grid can be employed in a wide range of ways [10]. Due to the different size of communication coverage and various utilities, the smart grid can be conventionally divided into three major domains: the home area network (HAN), neighborhood area network (NAN), and wide area network (WAN). In order to support high computing in smart grid, a central unit should be taken into consideration as an extra domain in the proposed hierarchical wireless communication architecture, as shown in Fig. 2.

As a basic unit of a smart grid communication system, the HAN mainly consists of a variety of smart appliances and SMs that collect the real-time power information of smart appliances through wireless communication for energy consumption management [11]. All the collected meter reading data is gathered to generate the total energy consumption information and transmitted between HANs and NANs. Similar concepts to HANs are building area networks (BANs) and industrial area networks (IANS), which are applied to commercial and industrial scenarios, respectively. With relatively small coverage areas and less information exchange, HANs typically need 10–100 kb/s/node bandwidth. Wireless technologies such as Zigbee, WiFi, and cellular networks are widely investigated for HANs.

The NAN is usually defined as a last-mile users’
access network, connecting SMs and other distribution automation collectors to a WAN. Furthermore, the other collectors gathering monitored and controlled information generate a separate network, the field area network (FAN). The FAN has a similar geographical scale as a NAN. Each SM/distribution collector can be viewed as a gateway of NAN/FAN covering a few square kilometer area, and its bandwidth is from 10 to 100 kb/s/node. Cellular technology is usually regarded as a strong candidate for NANs.

A WAN collects information from multiple NAN/FAN and transfers it to a central unit for centralized management. Meanwhile, a WAN also covers all of the transmission and distribution systems. Therefore, a WAN can cover a huge area that is extended to thousands of kilometers. In addition, for a requirement of a large amount of data transmission, a high-speed communication technology needs to be adopted, such as optical fiber communication or cellular communication, which requires 10–100 Mb/s/node bandwidth. For the sake of reliability and latency, utility companies usually adopt a hybrid solution mixing wired and wireless communication technologies for a WAN. In addition to cellular technology similar to a NAN, cognitive radio (CR) is another wireless technology suitable for a WAN.

The central unit consists of a control center and a data center that is a connected server set in an internal connection network. It is established to analyze and process smart grid data and send the control commands to smart grid devices.

**Key Technologies for Big-Data-Aware Wireless Communication**

Once the big-data-aware wireless communication architecture shown in Fig. 2 is designed, some key technologies should be emphasized for improving the wireless communication quality and developing new applications. We describe them as follows.

**Software-Defined Network:** Software-defined networking (SDN) is a revolutionary network architecture designed with the concept of decoupling the control plane and data plane. Due to simplifying the functionalities, the underlying devices such as intelligent electronic devices (IEDs) can make comprehensive packet forwarding decisions, taking into consideration quality of service (QoS), quality of experience (QoE), and application types. The centralized controller is responsible for managing and controlling the underlying hardware to flexibly allocate network resources on demand. To handle the huge amounts of time-critical information in smart grid, an SDN-enabled communication network is investigated, where advanced algorithms for the SDN controller are still needed for the sake of better fault tolerance, real-time capability, and resource efficiency.

**Cloud and Cloudlet:** In smart grid, power should receive real-time management ranging from power generation to power distribution. A cloud-based radio access network (C-RAN) provides a scalable solution for smart grid applications such as condition monitoring. Computing resources including computation, storage, and network are packaged and moved to the cloud so as to centrally optimize the management of massive state data throughout the network [12]. Considering a hybrid mechanism with both central and local data processing, the concept of cloudlets is proposed as a light cloud deployed at base stations (BSs) with limited storage and computation capability [13]. Cloudlets can not only effectively assist the data transmission of delay-sensitive applications but also help with local data processing, which results in reduced real-time traffic.

**Crowdsourcing:** Resources can be shared between SMs for the sake of common interests. For instance, under high traffic volume, besides direct data transmission between a BS and SMs, SMs within a certain region can collaborate with others for data dissemination. A similar crowd-enabled data transmission mechanism is introduced for mobile networks in [14] where the transmission channels can be shared with adjacent mobile terminals to avoid network traffic congestion in

![Figure 2. Big-data-aware wireless communication architecture for smart grid.](image-url)
the big data environment. Meanwhile, the incentive mechanism can be adopted for encouraging widespread transmission channel sharing behaviors. Under this circumstance, free transmission channels can be fully utilized and reliable data transmission can be guaranteed.

Cache Control: In smart grid, different applications have different requirements in communications. Although some applications such as load forecasting have less demand in time delay, many other applications such as wide area situational awareness of PMUs need near-real-time data transmission. Distributed caching is an effective way to relieve high real-time traffic volume. Due to limited cache size, cached contents need to be carefully classified, well organized, and promptly updated to maximize the utilization of system bandwidth. Labels are embedded into content objects, marking their properties. According to the labels, cached contents can be well categorized. Contents with types in need of more access counts are continuously cached, while other contents are removed regularly.

Case Study: Big-Data-Enabled Network Planning for Energy Storage

As shown in Fig. 1, network planning in data analysis is an important application of wireless big data computing in smart grid. In this section, we present a residential energy storage planning scheme as a case study of big data computing in smart grids. We consider a system model where users are equipped with smart appliances that are under the charge of SMs. Then each user chooses whether to deploy energy storage devices and the capability of energy storage devices for better energy scheduling and lower cost. The problem could be divided into two sub-problems: an energy scheduling problem to decide the detailed energy consumption pattern of each smart appliance and storage device in each day, and an energy storage planning problem to select a cost-effective storage device for each consumer.

System Model

The power system is considered with one energy source and $N$ consumers. Each consumer is equipped with an SM that can monitor and control the operation of the consumer’s appliances and storage devices according to a given schedule. SMs are connected to the central unit and exchange information to obtain an optimized energy schedule.

We divide one day into $H$ time slots. For instance, assume $H = 24$, and each time slot lasts 1 h. We denote the load of consumer $n$ at time slot $h$ is $l_{nh}$. Thus, the total load at time slot $h$ is

$$L_h = \sum_{n=1}^{N} l_{nh}.$$

We introduce a widely adopted direct load control (DLC) scheme for the billing function, which can be given as a quadratic function $C = K_h L_h^2$, where $K_h$ varies in different time slots.

For each consumer, $A_n$ appliances are deployed. Appliances consist of two classes: non-shiftable appliances and shiftable appliances. Non-shiftable appliances have to operate in a predetermined period of time. However, shiftable appliances can be shifted within a certain range to avoid peak time. The start and end times of this time period are denoted by $t_{start,n,a}$ and $t_{end,n,a}$ and the desired total daily consumption is $E_{n,a}$. Also, $e_{min,n,a}$ and $e_{max,n,a}$ represent the minimum and maximum energy consumption of appliance $a$ during one operating time slot. All these parameters restrict the scheduling of shiftable appliances.

Besides residential appliances, consumers can choose to have storage devices. Energy can be stored in advance and consumed in peak hours. The energy flowing into and out of consumer $n$’s storage device is denoted as $s_{in,n}^h$ and $s_{out,n}^h$. The capability of consumer $n$’s storage device is denoted as $S_{max,n}^h$ and limits the maximum amount of energy stored by consumer $n$. The maximum charging and discharging rate, $s_{max,in,n}$ and $s_{max, out,n}$, restrict the maximum amount of energy that can flow into and out of consumer $n$’s storage device during a time slot.

Energy Storage Planning

We propose a hybrid approach for storage planning, which consists of an outer optimization based on GA and an inner optimization algorithm for energy scheduling introduced in detail in the next subsection.

The GA indicates the storage capability of all consumers’ storage devices. Each energy storage device is assumed to be composed of multiple base units. We denote the base unit as $BU$, and the capability of energy device belonging to a consumer as $1 \leq n \leq N$ is $K_n BU$ with $K_n$ being non-negative integers. A special case is when $K_n = 0$, which means consumer $n$ chooses not to deploy a storage device.

There is a maximum number of deployed
base units for each consumer, which we denote as $K_{\text{max}}$. This maximum value is influenced by some economic and technology conditions. Thus, the GA is constrained by the value $K_{\text{max}}$, that is, for any $1 \leq n \leq N$, it should satisfy $k_n \leq K_{\text{max}}$. The fitness function indicates the overall cost for the initial installation of storage devices and everyday power consumption.

The specified optimal energy scheduling of consumers for each day during the given period is obtained with the proposed energy scheduling scheme. Then a fitness function is calculated to evaluate the population. In the inner loop, the consumers so that the fitness function is calculated to evaluate the individuals and decide whether to terminate the GA.

The GA terminates when the minimal fitness function value remains constant over an assigned number of generations or the maximum number of iterations is reached. The output of GA indicates the optimal capability of each user’s energy storage device.

**Energy Scheduling**

In this subsection, we use a game theoretic approach to model the energy scheduling problem. In the proposed energy scheduling game, consumers are considered as players who focus on minimizing their own costs. The strategy set of each consumer is the schedule of his/her appliances and storage device, which is restricted by the limitations described before. The cost function of each user is the billing paid by the user, given as

$$C_n = \sum_{h=1}^{H} K_h \left( \sum_{a=1}^{A} p_{a,n} + s_{h,n} \right)^2.$$  

There are multiple Nash equilibria existing in the proposed energy scheduling game and yielding the same cost. We can obtain one optimal schedule using the proximal decomposition algorithm together with the best response algorithm [15].

**Implementation of the Proposed Mechanism**

The storage planning mechanism consists of an outer optimization and an inner optimization. Each consumer’s data can be processed in parallel, and the load and other information is transferred once an optimized solution based on the current circumstance is solved.

In the outer optimization, populations are updated continuously. Once a new population is generated, the inner optimization is called to specify its optimized schedule and costs among the consumers so that the fitness function is calculated to evaluate the population. In the inner optimization, the proposed energy scheduling mechanism has a nested loop which can be further divided into two phases: initialization phase and runtime phase. At first, in the initialization phase of the outer loop, the central unit broadcasts $K_h$ for $h = 1, \ldots, H$. Meanwhile, the initial strategies of all the consumers are randomly chosen. As the process then goes to the initialization of the inner loop, the initial strategies of the regularized game of all the consumers are randomly initialized, and their corresponding loads are announced. Then the runtime phase of the inner loop begins. The optimization problem of each consumer is solved separately, and its new schedule is announced to the others. Consumers’ updated optimization problems are solved constantly once new loads are received until there are no new updates announced. At this point, the regularized game has been solved, and the solution is assigned as the next strategy. Then the process enters the inner loop with the new strategy, and the new regularized game is solved continuously until the strategy converges to the solution of the energy scheduling game.

The total cost across the grid after the proposed storage planning scheme is compared to the cost without storage deployment. The simulation results are shown in Fig. 3. In the case with storage planning, a relatively large amount of cost appears at the very beginning due to the installation of storage devices. Then there are lower daily costs subsequently, which gradually narrow the gap between the total costs in the two cases. After 169 days, the total cost in the case with storage planning becomes lower than in the other case. If we consider a lifetime of 10 years, the case with storage planning would save a total of $24,497$ comparing to the case without storage deployment.

For each single day, the energy scheduling scheme contributes a lot to flattening the peak load in the grid. The load at each time slot during a day is shown in Fig. 4, comparing the case with the proposed energy scheduling mechanism and the origin energy consumption pattern according to consumers’ desires. With the energy scheduling scheme, we can find that the peak load is reduced, while off-peak loads increase and the peak-to-average ratio (PAR) reduces from 4.66 in to 4.34 (i.e., 6.87 percent less). This directly leads to the saving in daily costs shown before.

**Conclusions**

In this article, we have discussed the characteristics of smart grid big data and proposed a big data computing architecture for smart grid analytics. Then the hierarchical big-data-aware wire-
The storage planning mechanism consists of an outer optimization and an inner optimization. Each consumer’s data can be processed in parallel and the load and other information is transferred once an optimized solution based on the current circumstance is solved.

Acknowledgment

This work is supported by NSFC (61572262); NSF of Jiangsu Province (BK20141427); NUPU (NY214097); Open research fund of Key Lab of Broadband Wireless Communication and Sensor Network Technology (NUPT), Ministry of Education (NYKL201507); Qinlan Project of Jiangsu Province; The projects 240079/F20 funded by the Research Council of Norway; the project Security in IoT for Smart Grids, with number 248113/070 part of the IPTPLUS program funded by the Norwegian Research Council, and the Knowledge Foundation (KSS) Sweden. The corresponding authors are Der-Jiunn Deng and Yanfei Sun.

References