Safety-Assured Collaborative Load Management in Smart Grids

Hoang Hai Nguyen[1] Rui Tan[1] David K. Y. Yau[1,2]
[1]Advanced Digital Sciences Center, Illinois at Singapore
[2]Singapore University of Technology and Design, Singapore
{hoanghai, tanrui, david.yau}@adsc.com.sg

ABSTRACT

When a power grid is overloaded, load shedding is a conventional way to combat the imbalance between supply and demand that may jeopardize the grid's safety. However, disconnected customers may be excessively inconvenienced or even endangered. With the emergence of demand-response based on cyber-enabled smart meters and appliances, customers may participate in solving the imbalance by curtailing their demands collaboratively, such that no single customers will have to bear a disproportionate burden of reduced usage. However, compliance or commitment to curtailment requests by untrusted users is uncertain, which causes an important safety concern. This paper proposes a two-phase load management scheme that (i) gives customers a chance to curtail their demands and correct a grid's undersupply when there are no immediate safety concerns, but (ii) falls back to conventional load shedding to ensure safety once the grid enters a vulnerable state. Extensive simulations based on a 37-bus electrical grid and traces of real electrical load demonstrate the effectiveness of this scheme. In particular, if customers are, as expected, sufficiently committed to the load curtailment, overloads can be resolved in real time by collaborative and graceful usage degradation among them, thereby avoiding unpleasant blackouts in existing practice.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Reliability, availability, and serviceability

1. INTRODUCTION

The conventional power grid has a well understood centralized design. It could be seen as "heavy around the waist," in that all the key control points are situated in the core of the system, which is trusted under careful operator control. For instance, when the grid is overloaded (due to say unexpected loss of generation or a surge in demand driven by extrinsic conditions), some breaker in the core may open, thereby disconnecting a region of customers to shed load and restore the balance between supply and demand. While this strategy protects the grid against expensive equipment damage or prolonged imbalance that could lead to cascading failures and massive blackouts, the disconnected region of customers are nevertheless severely affected in that they will be totally without power for some significant period of time. These "unlucky" customers may suffer significant loss of comfort or money, and their personal safety could be endangered in certain situations.

The emergence of *smart grid* initiatives [21], however, is evolving the architectural design of power grids in the direction of the Internet, whose end systems partake in sophisticated control relative to the core that performs simple well-defined functions. With demand-response (DR) [6] using smart meters and appliances that are capable of automated sensing and control, as well as communicating with the core, the grids of tomorrow are pushing towards a smart *edge* not unlike today's Internet end systems. Such a trend has important potential benefits. For instance, in the previous overload situation, many smart homes/appliances could collaboratively curtail their power consumption to restore the balance between supply and demand. By allowing customers to participate in the load management, the impact of undersupply will likely be milder and more bearable than in the case of the conventional design. This shows that decentralized control at the edge could drive the grid's operation to a *resilient performance* region, in which any needed service degradation is graceful and fair to everyone.

Vis-a-vis improved operation, the safety of smart grids as mission-critical cyber-physical infrastructures cannot

ICCPS'14, April 14–17, 2014, Berlin, Germany.

be ignored and in fact requires heightened attention. Indeed, it is a well known dilemma in system design that by single-mindedly pursuing performance features, no matter how desirable, we may grow the system's complexity unknowingly to the point that its safety and reliability are compromised at high costs [10]. Specifically, with more complex collaborative load management, the trustworthiness of control is significantly weakened because edge users/devices are involved, due to users' erratic behavior and limited/changing commitments, and devices' physical insecurity, variable quality, and possible misconfigurations. For instance, in the collaborative curtailment, unexpected deviations from a prescribed curtailment schedule may cause weakened or delayed responses to undersupply, or even diametrically opposite behaviors to exacerbate the overload, thus leading to extremely expensive failure of the critical infrastructure.

This paper is driven by the key objective to enable desirable performance features such as graceful and collaborative load curtailment in smart grids, while simultaneously assuring their safety at a level no less than that of the well tested conventional design. The fundamental tussle we need to address is that complexity engenders performance features, while the safety of a simple system is typically much stronger than that of a complex one. To resolve the conflict, we let the system run in a collaborative, though untrusted, mode when there are no known immediate safety concerns, but we monitor the collaborative operations continually and proactively to ensure sustained safety. When the monitoring, which is trusted, detects a drift of the system to an impending vulnerable state, the system falls back on a simple and safe, albeit possibly suboptimal, control mechanism in an assured and timely manner (i.e., we are certain to avert unacceptable system failure in time).

We apply the above framework to load management in smart grids to handle overloads. The collaborative operating regime corresponds to DR-driven load curtailment by the edge devices and the fallback mechanism corresponds to the conventional load shedding. The monitoring that triggers the load curtailment and shedding whenever needed is carried out by a real-time and high-fidelity intelligent system (IS) that can assess the grid's safety using sensors such as phasor measurement units (PMUs) distributed in the trusted grid core. In particular, the IS is designed based on a novel safety metric, which we call *time-to-being-unsafe*. The metric measures the minimum remaining time until the grid's possible failure should contingencies occur. It provides foresight to activate the load shedding in time to prevent the grid from entering unsafe regions. It also provides time, before safety becomes emergent, for customers to curtail their demands by following a schedule prescribed by the grid operator using a planning algorithm based on model predictive control (MPC). Extensive simulations based on a 37-bus transmission system and traces of real electrical load demonstrate these features.

The rest of the paper is organized as follows. §2 reviews background and related work. §3 overviews our approach. §4 and §5 present the IS for safety assessment and the proposed load management framework. §6 presents simulation results. §7 concludes.

2. BACKGROUND AND RELATED WORK

2.1 Power Grid Safety Assessment

Safety assessment of a power grid is performed with respect to a *safety criterion* that consists of a *contingency* and a *safety condition* [14]. Take the assessment of generator safety as an example. When a grid is subjected to a fault (e.g., a short circuit on a transmission line), the deviation of speeds of generators in the grid from the nominal value (i.e., 50 or 60 Hz) must be within a range (e.g., 3 Hz) to prevent infrastructural damage [14]. In this assessment, the contingency is usually chosen as a fault located at a line/bus near the generators. The safety condition is the allowed range of generator speeds during the transient period after the contingency. Under a certain *operating condition* that consists of various measurable physical quantities of the grid (e.g., demand at the buses), if the contingency does not cause any violation of the safety condition, the grid is classified as *safe*. Otherwise, it is *unsafe*. For an unsafe grid, *generation shift* and *load shedding* are two conventional approaches to restoring the safety. Specifically, by rescheduling the generation of multiple generators or opening appropriate breakers to disconnect a subset of the loads, these two approaches change the operating condition to meet the safety condition.

2.2 Related Work

Safety assessment can be conducted using time-domain (T-D) simulations. However, they are time-consuming and therefore usually inappropriate for online assessment. Various intelligent systems (ISes) [20, 5, 24] have been proposed as alternatives. Trained with data generated by offline T-D simulations, these ISes can provide assessment results rapidly. Sun et al. [20] build a decision tree to classify a power grid's safety based on PMU measurements. Amjady and Majedi [5] train a neural network to detect future transient instability based on the angle and timing of the first swing of the transient oscillation caused by a fault that occurred. In [24], Xu et al. employ extreme learning machines (ELMs) to estimate the critical clearing time (CCT) of a contingency, i.e., the maximum time duration of the contingency without causing unsafety. Different from these

approaches that classify the grid's safety [20, 5] or calculate properties of tolerable contingencies [24], this paper constructs ELMs to predict the minimum remaining time before the grid becomes unsafe, thereby enabling proactive actions to prevent the unsafety.

Actions can be taken to correct an unsafe grid. Kato and Iwamoto [13] propose an approach combining generation shift and generator voltage control to extend CCTs. Karapidakis and Hatziargyriou [12] propose an online iterative algorithm that calculates a new generation dispatch until a decision tree classifies the grid as safe. Genc et al. [7] use a decision tree to identify safety regions for generation and load, and employ generation shift and load shedding to restore the grid safety. Different from these approaches [13, 12, 7] that rely on centralized control in the grid's core, our approach allows to cope with potential unsafety in a decentralized manner, through DR involving distributed users.

Recent studies [25, 19] explore decentralized demand-side management to improve system performance. Xu et al. [25] show that dynamic demand technology, which adjusts the power consumption automatically by monitoring the line frequency, can be employed to achieve supply-demand balance. Furthermore, Short et al. [19] show that refrigerators equipped with dynamic demand controllers can replace a certain volume of spinning reserve. However, exploiting the demand side in grid safety monitoring and maintenance has received limited research attention.

The Simplex architecture [18], which consists of a high-performance controller (HPC) and a high-assurance controller, has been proposed to deal with HPC failures and changes of physical dynamics [23]. Both Simplex and our proposed load management feature a mechanism that triggers different manipulation strategies according to the system state. Different from Simplex that applies HPC to pursue *desirable* high performance in normal operation, our approach leverages DR to achieve a resilient solution during selected periods (e.g., peak hours) only, in which the service *degrades*, albeit fairly, for customers. Moreover, in contrast to the *intrinsic* unreliability or untrustworthiness of HPC in Simplex (due to the complexity or unverifiability of software components), the unreliability of load curtailment in our problem is *extrinsic*, i.e., from the customers who can be considered a part of the plant from a control perspective.

3. APPROACH OVERVIEW

Current safety assessment and management approaches such as those reviewed in §2.2 have two major potential issues. First, most of them adopt a centralized management approach (e.g., load shedding) that can be unfair and even hazardous to customers as discussed in §1. Second, many of them adopt a bipolar safety metric, i.e., the grid is classified as either safe or unsafe. Such a metric may suffice for load shedding that can immediately restore the grid's safety on the detection of an unsafe state. However, for generation shift and demand side management that often need significant time to take effect, this bipolar metric can result in unexpected delays in reacting to unsafe states. To address these issues, this paper designs a holistic approach by leveraging machine learning techniques and the increasingly available communication and control capabilities at the edge of power grids.

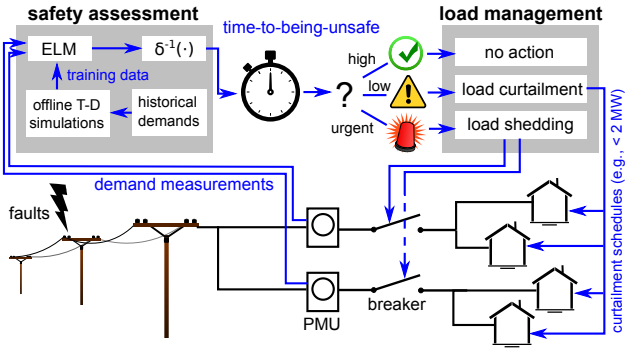

Figure 1: Overview of our approach (arrows represent data or control flows).

As illustrated in Fig. 1, our approach consists of two pillar modules: (i) a real-time *safety assessment* subsystem based on a new safety metric and extreme learning machine (ELM), and (ii) a novel two-phase *load management* subsystem. The contributions of our approach are:

A new safety metric: We propose to adopt the minimum remaining time until the power grid becomes unsafe as the safety metric, which we refer to as *time-to-being-unsafe* (TTBU). With this new metric, we can proactively activate preventive management actions and account for their delays in assuring the grid's safety.

Real-time safety assessment: We develop a real-time safety assessment subsystem based on ELM [11] to estimate the TTBU. We first conduct extensive offline T-D simulations to generate training data for the ELM. At run time, ELM estimates the TTBU according to the current operating conditions. Moreover, it can be repeatedly invoked to assess candidate load management actions and find the best one accordingly.

Safety-assured collaborative load management: We propose a two-phase load management subsystem consisting of a *load curtailment* and a *load shedding* phase. Specifically, when the TTBU drops below a warning threshold, the subsystem enters the load curtailment phase and induces the customers to reduce their consumption collaboratively via DR. This phase decentralizes the safety management by involving dis-

tributed customers. We develop an algorithm for the curtailment scheduling based on a MPC method, which computes a list of suggested *demand ceiling*s for a number of future time periods. However, the curtailment schedules may not be fully realized due to limited customer commitment and/or incorrect edge devices. Once the TTBU drops below an emergent threshold, the subsystem enters the load shedding phase, which immediately disconnects a subset of loads to prevent unacceptable system failures.

Note that our approach allows DR-based load curtailment to run side-by-side load shedding employing existing technologies. Hence, DR programs can be introduced incrementally for selected (growing) subsets of the customers.

4. REAL-TIME SAFETY ASSESSMENT

This section defines the TTBU metric and presents the design of ELM to learn the metric. Lastly, a case study is provided to illustrate the ELM-based real-time safety assessment.

4.1 Time-to-Being-Unsafe

We propose the TTBU metric based on extensive T-D simulations using PowerWorld [4]. PowerWorld is a power system simulator that is widely used in the power industry for transmission planning and operation analysis and visualization. Our simulation results show that the demand vector at all the load buses (which we refer to as *loading*) is the dominating factor for grid safety.[1] Loading indicates the level of stress imposed on a grid. When the grid is under higher stress, it is at a higher risk of being destabilized due to the same contingency. Thus, loading directly relates to grid safety, and we use loading as the operating condition for the safety assessment. An associated advantage of this choice is that, since we aim to protect the grid by load management, using loading for safety assessment can provide immediate decision support to guide the load management actions.

We now use a set of T-D simulations for the IEEE 9-bus system [4] produced by PowerWorld [4] to illustrate the effect of loading on grid safety. This system has 3 generators and 3 load buses (buses #5, #6, and #8) out of totally 9 buses. We evaluate the generator speed deviation against a balanced 3-phase solid fault at the transmission line from bus #5 to #7. This contingency happens when all the three conductors of the line are shorted together, due to say a lightning strike.

[1]Power injections from generators also affect the grid's safety. This paper focuses on non-renewable generation, whose power injections are dependent variables of the economic generation dispatch with demand as a vector of independent variables.

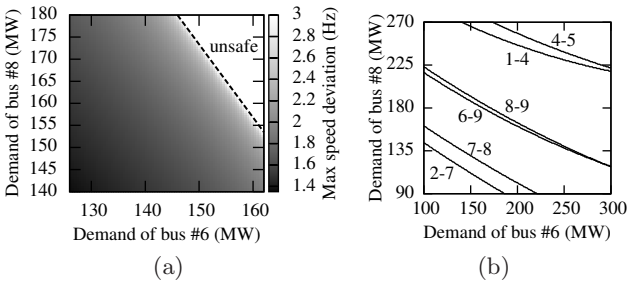

Figure 2: (a) Maximum generator speed deviation (represented by a gray scale) after a fault vs. the demand of two buses. The white area is the unsafe region with a boundary marked by the dashed curve. (b) Safety boundaries for different contingencies at different transmission lines (labeled by their respective terminal bus numbers).

Its clearing time is set to be 5 cycles (i.e., 83 ms for this 60 Hz system). For ease of illustration, we fix the demand of the load bus #5 at 200 MW and vary the demand of the two other load buses. The maximum deviation of the three generators' speeds from the nominal value (60 Hz) after the contingency, as a function of the varied demand, is plotted in Fig. 2(a). Under the default setting of this system, a generator will be shut down to prevent infrastructural damage if its speed deviation exceeds 3 Hz. Thus, the system with a generator speed deviation of more than 3 Hz is considered unsafe. From Fig. 2(a), we can clearly observe a *monotonic* relationship between the maximum speed deviation and the demand. That is, as long as the demand increments of the two load buses are non-negative, the speed deviation will increase. Moreover, there is a cut-off boundary between the safe and unsafe regions. Fig. 2(b) shows the boundaries for the balanced 3-phase solid faults that occur at different transmission lines with the same clearing time.

The above monotonic property can also be observed for other types of contingencies such as ground fault. Specifically, for a certain safety criterion, starting from a safe state, if the demand at each bus[2] keeps increasing, at some point the grid will become unsafe. Once the grid is unsafe it will remain unsafe unless the demand at some bus(es) is reduced. Formally, let the vector $\mathbf{L} \in \mathbb{R}^m_{\geq 0}$ denote the loading of a system with m buses. For any time instant t_0, we define $\boldsymbol{\delta}_{t_0}(\Delta t) = [\delta_{1,t_0}(\Delta t), \delta_{2,t_0}(\Delta t), \ldots, \delta_{m,t_0}(\Delta t)]$, where $\delta_{i,t_0}(\Delta t)$ is the ramp-up in demand of bus i at time $t_0 + \Delta t$. We assume that $\delta_{i,t_0}(\Delta t)$ is a positive and strictly increasing function of time duration Δt. There exists a factor T such that the grid at time $t_0 + \Delta t$ with loading $\mathbf{L} + \boldsymbol{\delta}_{t_0}(\Delta t)$

[2]In the rest of this paper, we use *bus* to refer to a *load bus* unless otherwise stated.

is safe if $\Delta t < T$, and it is unsafe otherwise. If $\delta_{i,t_0}(\Delta t)$ is the *maximum* ramp-up in demand at each bus i, then T is the *minimum* remaining time before the grid with loading $\mathbf{L}$ at time t_0 may become unsafe. We call T the *time-to-being-unsafe* (TTBU). Note that T is defined with respect to a certain safety criterion. In §4.2, we will discuss how to address multiple safety criteria, such as multiple contingencies in Fig. 2(b).

In this paper, for simplicity, we assume that each bus at any time instant t_0 has the same maximum ramp-up function, i.e., $\delta_{i,t_0}(\Delta t) = \delta(\Delta t)$, $\forall i \in [1, m]$, $\forall t_0$. Under this simplification, we define the *power-distance-to-being-unsafe* as $P = \delta(T)$, which is the maximum amount of additional power that each bus can draw without causing unsafety. Thus, as long as P is known, T can be obtained as $T = \delta^{-1}(P)$. This intermediate metric P will be used to design the ELM in §4.2. Although the simplifying assumption of identical maximum ramp-up may lead to a loss of accuracy in estimating T, it will not affect the safety assurance if $\delta(\Delta t)$ is appropriately chosen. For instance, we can choose the upper envelope of all the actual ramp-up functions as $\delta(\Delta t)$, i.e., $\delta(\Delta t) = \max_{\forall i \in [1,m], \forall t_0} \delta_{i,t_0}(\Delta t)$. The resulting T will be a conservative estimate for the TTBU. As discussed in §4.2, the safety assessment subsystem can be readily extended to admit per-bus ramp-up functions, albeit with additional overhead.

4.2 Design of ELM

We use extreme learning machine (ELM) [11] to learn the proposed safety metric. ELM is a single hidden layer feedforward neural network with a training algorithm much faster than conventional gradient-based learning algorithms. The design of ELM for our problem boils down to the selection of its input/output parameters and internal configurations. We choose loading $\mathbf{L}$ and power-distance-to-being-unsafe P as the input and output of the ELM, respectively. At run time, given the current loading, the safety assessment subsystem calculates T based on the ELM's output P by $T = \delta^{-1}(P)$. The advantage of using the intermediate metric P as the ELM output is that it isolates the demand behavior (i.e., $\delta(\Delta t)$) from the ELM. Thus, the ELM need not be re-trained if $\delta(\Delta t)$ changes.

We use a synthetic data set, of historical records of $\mathbf{L}$ from an operator's database, and T-D simulation results to train the ELM. First, the maximum ramp-up function $\delta(\Delta t)$ can be easily learned from the records of $\mathbf{L}$. Second, given $\mathbf{L}$, we may run T-D simulations to determine P. By enumerating $\mathbf{L}$ in all the records, we generate a large number of data pairs $\langle \mathbf{L}, P \rangle$. In this paper, we assume that the system model for driving the T-D simulations is accurate. In practice, system operators often improve the model continually based on real measurements for past contingencies [16]. The ELM trained using the generated data pairs will preserve the realistic physical dynamics provided by the high-fidelity offline T-D simulations. The internal configuration for the ELM, such as the number of hidden neurons, is usually determined by an iterative trial-and-error procedure to minimize some error metric (e.g., sum of squared errors of estimation) based on the training data. At run time, the ELM can quickly compute P according to $\mathbf{L}$ as measured by trusted metering devices such as PMUs at the buses. The high-speed computation is highly desirable since it allows us to iteratively look for the best control actions in real time, where the ELM is invoked repeatedly to evaluate candidate actions (cf. §5.2).

We now discuss two practical issues. First, if there is a planned change to the grid (e.g., adding a transmission line or generator), the ELM needs to be re-trained using new T-D simulations based on the new system model. Second, the discussions so far are for a single safety criterion consisting of a contingency and a safety condition. The system operator may want to handle multiple credible safety concerns, e.g., multiple contingencies at different locations. This issue can be addressed by using multiple ELMs, in which each ELM is trained to address one safety criterion. At run time, the minimum of the power-distances-to-being-unsafe estimated by all the ELMs is used to calculate T, which is the minimum time to an unsafe state caused by *any* credible contingency. As power grid design is often incrementally improved to fix known outstanding vulnerabilities to credible contingencies, it is unlikely that T will be confined to explaining a small set of contingencies.

4.3 A Case Study

In this section, we use a case study to illustrate the learning processes for $\delta(\Delta t)$ and the ELM. It is based on real loading data published by the New York Independent System Operator (NYISO) [2]. The measurements were taken every 5 minutes over two months (June and July 2012) for 11 regions. As a detailed model for the NYISO grid is unavailable, in this case study, we use an example system model from PowerWorld as shown in Fig. 3. The model consists of 9 generators and 25 load buses out of totally 37 buses. It includes detailed physical properties of the grid, such as generator configurations and line capacities. However, because there are only 11 regions (buses) in the original NYISO data set, we create an adapted loading data set as follows. The demand of each bus in the new data set is the total demand of two randomly picked regions in the NYISO data set. For each bus, the minimum demand over the two months is selected as the *base power* and each demand record is normalized using this base power. Thus, different buses have different base power numbers, and

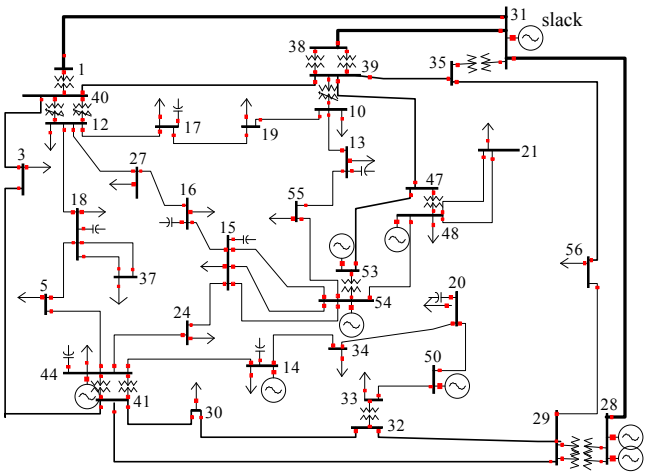

Figure 3: One-line diagram of the system for case study.

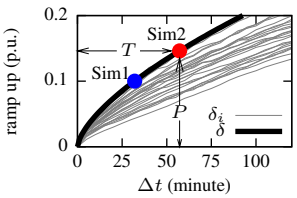

Figure 4: Maximum ramp-up functions for individual buses and the system.

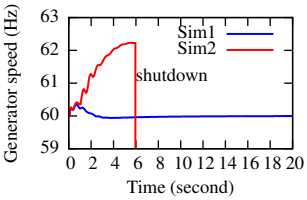

Figure 5: Speed of the generator at bus #44 after a fault occurred at time zero.

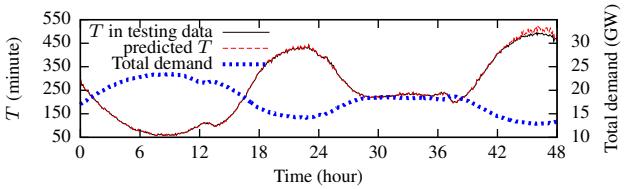

Figure 6: T and the total demand in two days.

the new data set has 18,000 loadings, where each demand record is within the range of $[1.0, 1.7]$ per unit (p.u.). Note that these normalized p.u. values allow us to learn an identical $\delta(\Delta t)$ that is applicable to all the buses. We follow the approach in §4.1 to learn $\delta(\Delta t)$. Fig. 4 plots the maximum ramp-up functions for the individual buses $\delta_i(\Delta t)$, as well as the global system-level ramp-up function $\delta(\Delta t)$.

We evaluate the grid's safety against a contingency of a balanced 3-phase solid fault at the line from bus #31 to bus #38, for a duration of 5 cycles (i.e., 83 ms). As this line is a backbone, the fault may cause non-oscillatory instability, in which all the generator speeds drift away from the nominal value of 60 Hz. Under the default settings for this 37-bus system, if a generator's speed is higher than 62 Hz or lower than 55 Hz for more than two seconds, it will be automatically shut down to prevent damage. Therefore, in this case study, we define the safety condition based on the post-contingency operation status of all the generators. Specifically, under a loading, if all the generators remain in operation 20 seconds after the contingency, the grid is considered safe; otherwise, it is unsafe.

We follow the approach in §4.2 to generate training/testing data for the ELM using PowerWorld. For instance, Fig. 5 shows the speed of a generator in two T-D simulations (labelled by Sim1 and Sim2 respectively) for a loading in the data set with different increments in demand at each bus as illustrated in Fig. 4. In Sim1, the grid is safe. In Sim2, the demand increment is 0.146 p.u. The generator is shut down after six seconds, which violates the safety condition. If the demand increment is smaller than 0.146 p.u., the grid remains safe. Thus, for this loading, $P = 0.146$ p.u., which maps to $T = 57$ min according to $\delta(\Delta t)$ in Fig. 4. By leveraging the monotonic property of grid safety discussed in §4.1, we use binary search to speed up the process of finding P given a loading $\mathbf{L}$. On a workstation computer with an Intel Xeon quadcore CPU at 2.8 GHz, it takes about 15 seconds to process a loading, and hence about three days to complete the data generation. In practice, high-performance computers can be used to speed up this process.

We use one third of the generated data to train the ELM implemented using Python [1] and the other two thirds for testing. The training takes less than one minute. Using ELM to estimate T for a loading takes 0.15 ms only, which is a 10^5x speed-up compared with the binary search based on T-D simulations. Denoting by $\hat{P}$ the power-distance-to-being-unsafe as estimated by the ELM, we define the relative estimation error as $\frac{\hat{P}-P}{P} \times 100\%$. The test shows that all the relative estimation errors fall within the range $[-4\%, 2\%]$, and 99% of them are smaller than 1%. This result confirms the high accuracy of the ELM. Fig. 6 shows the T estimated by the ELM and its corresponding true value in the testing data, as well as the total demand of the 25 buses, in two days. We can see that the ELM can accurately estimate T. Consistent with intuition, T and the total demand exhibit opposite trends over time.

5. TWO-PHASE LOAD MANAGEMENT

This section starts with an overview of the two-phase load management proposed in §5.1. Then, §5.2 and §5.3 detail the load curtailment and shedding approaches, respectively. Lastly, §5.4 discusses several implementation issues.

5.1 Overview of Two-Phase Load Management

The safety assessment and load management are carried out periodically. At the beginning of each *cycle*, the system measures $\mathbf{L}$ and estimates T by ELM. The two phases of the load management are *load curtailment* and

load shedding. According to the T estimated at the beginning of each cycle, the system determines the current phase and applies its corresponding management strategy. In the load curtailment phase, the system curtails the power usage of customers via a DR program. The *curtailment schedule* at each bus is a list of suggested *demand ceiling*s for several subsequent cycles. This paper focuses on scheduling the demand ceilings for buses. A demand ceiling for a bus can be further decomposed into demand ceilings for individual customers connected to the bus and participating in the DR. If the per-bus demand values do not exceed these ceilings, T will be maintained around a predefined level. Note that the demand ceilings could be translated as alternative DR signals including real-time prices [17], as long as a model for bus demand given the DR signal is available. In the load shedding phase, some breaker(s) in the grid's core will open to disconnect some subset of the loading immediately.

The system enters/exits the load curtailment phase by comparing T with a *warning* threshold denoted by T_W. Various existing load shedding approaches [7, 15] can be applied for the load shedding phase. In this paper, we adopt a hysteresis-based load shedding approach [15], which is designed to exploit the predictive safety metric T. This approach uses two thresholds, namely the *restorative* and *emergent* thresholds (denoted by T_R and T_E), which are used respectively as triggering and exiting watermarks for the load shedding. We assume that $T_W > T_R > T_E$. How to set the three thresholds is the subject of §5.4.1. Regarding a cycle, its length should be chosen to achieve a suitable tradeoff between (i) the communication overhead of sending curtailment schedules to a large number of customers, and (ii) prediction accuracy, which typically degrades with longer cycles. Moreover, cycle lengths adopted by existing DR proposals such as real-time pricing [17], which are in low tens of minutes, provide reference settings.

We now use the example shown in Fig. 7 to illustrate switching between the two phases during peak hours. For simplicity of exposition, the beginning of the kth cycle is referred to as *time instant* k. In Fig. 7, at $k = 0$, as $T > T_W$, the grid is considered safe enough and no management actions are applied. As the loading increases, at $k = 1$, T drops below T_W and the system enters the load curtailment phase. In particular, a MPC-based algorithm for curtailment scheduling (cf. §5.2) computes a *curtailment schedule* for each bus, which aims to maintain T above T_W in a few (two in Fig. 7) subsequent cycles. The square dots in Fig. 7(b) give the demand ceilings for bus #1 in the curtailment schedule. However, as discussed in §3, the buses may reduce their demand but in this scenario they still exceed the scheduled ceilings. As a result, T keeps decreasing

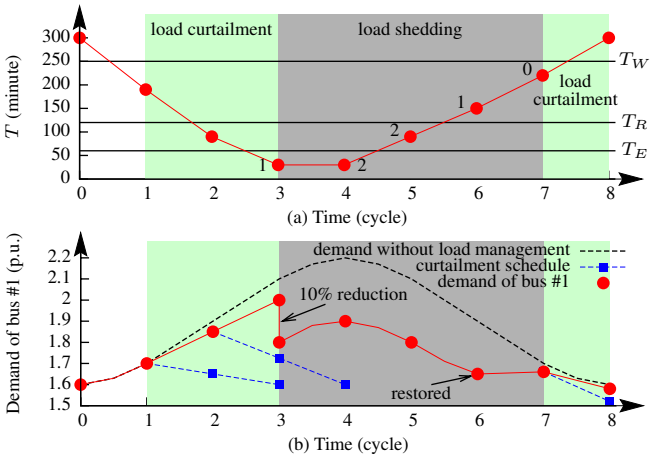

Figure 7: Illustration of the proposed two-phase load management during peak hours: The digit labels in Subfigure (a) give the numbers of shed buses. Subfigure (b) shows the actual demand and suggested demand ceilings.

and drops below T_E at $k = 3$. At this point, the grid is in an emergent condition and the hysteresis-based load shedding [15] is activated to prevent the grid from becoming unsafe. Under this hysteresis-based approach [15], if T is below T_E in a cycle, a certain percentage of load on a newly selected bus is shed. On the other hand, if T is above T_R in a cycle, this approach reconnects the previously shed load at a selected bus. At $k = 7$, all the buses have been relieved from the load shedding. But as T is still under T_W, the system switches to the load curtailment phase. At $k = 8$, $T > T_W$ and no more management actions are needed.

5.2 MPC-Based Load Curtailment

In this section, we formulate the load curtailment problem based on the principles of MPC [8]. Since normally bus-level demand is predictable due to its intrinsic strong autocorrelation [17], as well as the predictability of extrinsic factors like weather, MPC is a suitable tool for planning the demand ceilings. In the following, §5.2.1 describes the models for demand prediction and curtailment, and §5.2.2 develops the proposed MPC-based algorithm for curtailment scheduling.

5.2.1 Preliminaries

Demand Prediction Model. An accurate prediction model is key to the effectiveness of MPC. Various demand prediction models have been applied in practice for generation scheduling and real-time pricing in electricity markets [3]. Due to strong temporal correlation in the demand of a bus in normal operation, most existing prediction models assume that a future demand depends on the most recent past demand.

We adopt an abstract prediction model as follows. Let

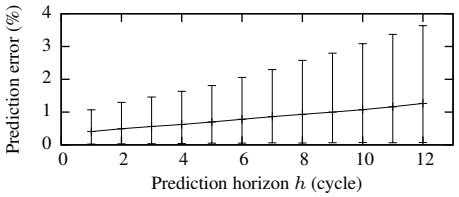

Figure 8: Relative prediction errors: error bar represents 90% confidence interval; cycle length is 5 minutes; $R = 24$.

$d_{i,k}, d_{i,k-1}, \ldots, d_{i,k-R+1}$ denote the R most recent actual demand levels of bus i at time instant k. Its predicted demand in the next cycle, denoted by $\hat{d}_{i,k+1}$, is given by

$$\hat{d}_{i,k+1} = f_i(d_{i,k}, d_{i,k-1}, \ldots, d_{i,k-R+1}, \Theta), \qquad (1)$$

where Θ represents all other time-varying and/or bus-dependent affecting factors such as forecast weather data. This prediction model can be used to generate multi-horizon predictions in a recursive manner, i.e.,

$$\hat{d}_{i,k+h} = f_i(\hat{d}_{i,k+h-1}, \ldots, \hat{d}_{i,k+1}, d_{i,k}, \ldots, d_{i,k+h-R}, \Theta),$$

where h is the *prediction horizon*. In the evaluation in this paper, we adopt a linear autoregressive (AR) model [17], i.e., $\hat{d}_{i,k+1} = \sum_{j=0}^{R-1} a_j d_{i,k-j} + a_R$, which can predict demand accurately. We use one third of the loading data created in §4.3 to learn the coefficients $\{a_j\}_{j=0}^{R}$ and use the rest for testing. Fig. 8 shows the error bars of relative prediction errors for bus #1 versus the prediction horizon h. We can see that the AR model achieves accurate predictions, with relative errors less than 4%. Moreover, consistent with intuition, the prediction error increases with h.

Curtailment Model. The operator of each bus i maintains a *curtailment schedule*, which is a first-in-first-out (FIFO) queue consisting of H demand ceilings. It is formally represented by $\mathbf{S}_i = [D_{i,1}, D_{i,2}, \cdots, D_{i,H}]$, where $D_{i,1}$ and $D_{i,H}$ are the oldest and newest elements, respectively, and H is the *optimization horizon* of the MPC-based curtailment scheduling (cf. §5.2.2).

At the beginning of each cycle, there is a short *curtailment scheduling session*. During this session, the system operator computes the curtailment schedules and communicates with each bus i to update all the elements in $\mathbf{S}_i$. After this session, bus i pops $D_{i,1}$ and uses it as the demand ceiling for the current cycle to guide its curtailment. Ideally, at the end of the current cycle, the demand of bus i is no higher than $D_{i,1}$. (The remaining elements in the schedule, i.e., $D_{i,2}$ to $D_{i,H}$, are estimated demand ceilings for the subsequent $H-1$ cycles, which will be updated by the system operator in the future curtailment scheduling sessions. The bus can use these estimates to prepare for the curtailments in the subsequent cycles.) Then, the bus duplicates the newest element $D_{i,H}$ and pushes it onto $\mathbf{S}_i$. Although all the elements in $\mathbf{S}_i$ will be updated in the next curtailment scheduling session, this self-duplication step simplifies the system design when the load management subsystem switches from the load curtailment phase to the load shedding phase, which will be discussed in §5.4.2. The above curtailment model has clear and simple semantics to system operators and customers. As such, it simplifies the design of supporting DR devices and fosters adoption.

5.2.2 MPC-Based Curtailment Scheduling

This section develops an algorithm to determine the demand ceilings in the curtailment schedules of the buses. Executed in each curtailment scheduling session, the algorithm aims to maintain T at around T_W, assuming that all the buses follow the curtailment schedules exactly. In §6, we will evaluate extensively the impact of customer commitment to the curtailment schedules on the performance of the scheduling algorithm. Suppose we need to compute the curtailment schedules for the kth to $(k+H-1)$th cycles at time instant k. We denote $x_h \in \mathbb{R}_{\geq 0}$ the *demand curtailment* in p.u. for any bus in the $(k+h)$th cycle, and define $\mathbf{X} = [x_1, \ldots, x_H] \in \mathbb{R}^H_{\geq 0}$. Imposing the same per-unit curtailment on all the buses achieves max-min fairness [9]. Note that since different buses may have different base powers (cf. §4.3), the demand curtailments in watts, which are translated from the x_h, vary across the buses. Let $\widetilde{d}_{i,k+h}$ denote the predicted demand of bus i at time instant $k+h$ provided that the bus curtails x_h p.u., i.e.,

$$\widetilde{d}_{i,k+h} = f_i(\widetilde{d}_{i,k+h-1}, \ldots, \widetilde{d}_{i,k+1}, d_{i,k}, \ldots, d_{i,k+h-R}, \Theta) - x_h. \qquad (2)$$

Based on $\{\widetilde{d}_{i,k+h} | i \in [1,m], h \in [1,H]\}$, we can predict the TTBUs for the subsequent H cycles using ELM, which are denoted by $\{\widetilde{T}_{k+h}\}_{h=1}^{H}$. Therefore, these predicted TTBUs depend on $\mathbf{X}$. Moreover, we define the following two quantities. First, as the objective of the curtailment scheduling is to maintain the prediction $\widetilde{T}$ at around T_W, we define the cost function:

$$c(\widetilde{T}|T_W) = |\widetilde{T} - T_W|.$$

Second, rapidly changing curtailments – i.e., large variations in the elements in $\mathbf{X}$ – make it challenging for customers to plan their power consumption in the presence of practical constraints and may thus reduce their commitment to the curtailment schedules. We define an abstract function $\sigma(\mathbf{X})$ to quantify the variation, as follows:

$$\sigma(\mathbf{X}) = \max_{h \in [1,H]} |x_h - x_{h-1}|, \qquad (3)$$

where x_0 represents the implemented curtailment in the $(k-1)$th cycle. Note that other realizations of $\sigma(\mathbf{X})$ such as standard deviation may also be used.

We formulate the following:

Curtailment scheduling problem. *Find* $\mathbf{X} \in \mathbb{R}^H_{\geq 0}$ *to minimize* $C(\mathbf{X})=\sum_{h=1}^{H} c(\widetilde{T}_{k+h}|T_W)$ *subject to* $\sigma(\mathbf{X})\leq\sigma_0$.

In the above formulation, σ_0 is the maximum allowed variation of the curtailments. It can be tuned by the system operator according to empirical past data on how it may impact customer commitment. The problem is a constrained non-linear optimization problem that has to be solved in real time. As $C(\mathbf{X})$ depends on the ELMs, it has no closed-form formulas. Thus, an optimal algorithm of polynomial time complexity is likely unavailable.

Under our realization of $\sigma(\mathbf{X})$ in Eq. (3), the complexity of brute-force search is $\mathcal{O}(N \cdot \lceil \sigma_0/q \rceil^H)$, where N represents the number of ELMs to address multiple safety criteria, and q represents the search granularity for each element in $\mathbf{X}$. The evaluation in §6 shows that a small setting for H suffices. Thus, a brute-force search may still have acceptable delay. For instance, when $H = 4$, it takes 8 to 10 seconds only on a common desktop computer. When a large H is needed, a constrained simulated annealing algorithm (CSA) [22], which can handle cost functions specified by a procedure instead of in closed-form, may be used to obtain near-optimal solutions. Our tests with $H \in [1, 10]$ show that the CSA yields the exact optimal solution and its computation time increases linearly with H. It is worth noting that, since the ELM is invoked for each candidate solution $\mathbf{X}$ in the search algorithm, the high-speed ELM makes it feasible to apply the MPC-based curtailment scheduling in practice. Such predictive scheduling would be infeasible if time-consuming T-D simulations were required.

Our problem can have multiple optimal solutions. In that case, we choose the one with the minimum $\sum_{h=1}^{H} x_h$ such that curtailments are applied only when they are necessary. We denote by $\mathbf{X}^* = \{x_h^*\}_{h=1}^H$ the optimal solution we choose. At time instant k, each ceiling $D_{i,h}$ in the curtailment schedule $\mathbf{S}_i$ is updated with $\widetilde{d}_{i,k+h}$ given by Eq. (2) with $x_h = x_h^*$. As discussed in §5.2.1, only $D_{i,1}$ will be applied for the current cycle and other ceilings in $\mathbf{S}_i$ serve as hints only and they will be updated in subsequent curtailment scheduling sessions before actual application. This is consistent with *receding horizon control* in MPC [8], which is widely adopted to improve system robustness to prediction inaccuracy. This robustness is particularly important in our problem domain, because uncertain and variable customer commitment, as well as potential unreliability of DR devices, may affect the prediction accuracy of Eq. (2) significantly. For instance, underestimated demand ceilings for the subsequent cycles due to limited customer commitment need to be amended in time before application.

5.3 Hysteresis-Based Load Shedding

Load shedding is a well-established technology in power grids [14] to cope with unsafety detected by online contingency analyses such as those discussed in §2.2. Integrated with the MPC-based load curtailment in §5.2, existing load shedding can be used easily as a component in the proposed two-phase management. In our evaluation, we adopt a hysteresis-based load shedding approach [15] designed to exploit the foresight of TTBU. We now describe briefly its principles, while the details can be found in [15]. Under this approach, the two thresholds T_E and T_R (cf. §5.1) serve as the low and high watermarks for the hysteresis. If T is below T_E in a cycle, a certain percentage (denoted by ρ) of load at a newly selected bus is shed; on the other hand, if T is above T_R in a cycle, the previously shed load at a selected bus is reconnected. The load management subsystem exits the load shedding phase if all the shed loads have been restored. The percentage of load to be shed in a cycle is specified by the system operator, subject to a tolerable level of disturbance caused by disconnecting load, as well as other constraints such as service agreements.

5.4 Implementation Considerations

This section discusses two important issues when implementing the two-phase load management.

5.4.1 *Configurations for T_E, T_R, and T_W*

As the load shedding phase is a fallback mechanism, the two thresholds used by the hysteresis-based load shedding approach, T_E and T_R, need to guarantee that T will not reach zero even without the load curtailment phase. Simulations based on past loading data can be conducted to search for the minimum settings of T_E and T_R such that T never goes to zero. With fixed T_E and T_R, T_W should be chosen to minimize the activations of load shedding. However, the effectiveness of load curtailment depends on customers' commitment to the curtailment schedules. Similarly, simulations can be conducted to search for the minimum settings of T_W under different commitment levels, such that T does not drop below T_E. At run time, the system operator can update T_W periodically (e.g., every month) according to the observed commitment level. In §6, we will conduct simulations to illustrate this configuration approach.

5.4.2 *Phase Transits*

We now discuss our design when the load management subsystem switches from load curtailment to load shedding. After the transition, the system operator stops

updating the curtailment schedule for each bus (i.e., $\mathbf{S}_i$). However, by following the self-duplication step described in §5.2.1, each bus i will keep popping the oldest element and pushing a duplicate of the newest element onto $\mathbf{S}_i$ in every cycle. If bus i has *not* been selected to shed load, it will use the popped demand ceiling to guide its curtailment. Otherwise, the loads that belong to bus i but are not disconnected (recall that only $\rho \times 100\%$ of the load is disconnected) are also subject to demand ceilings as part of the popped demand ceiling for the whole bus. When the load management subsystem switches back to the load curtailment phase, the curtailment scheduling is resumed and all the bus curtailment schedules will be updated. If another existing load shedding approach is used to replace the hysteresis-based approach, minor modifications to the above design may be needed for the phase transitions.

6. TRACE-DRIVEN SIMULATIONS

6.1 Simulation Methodology and Settings

The simulations are based on the 37-bus systems shown in Fig. 3 and the synthetic loading data traces described in §4.3. The contingency event and safety condition are the same as in §4.3. We use the testing data traces to drive the simulations. When no load management actions are applied, the demand of each bus is set with the testing data. When the system enters the load curtailment phase, we simulate the *desired demand* of each bus as follows. Let $d_{i,k}$ and $\bar{d}_{i,k}$ denote the demand of bus i at time instant k in the simulation and the data traces, respectively. The desired demand of this bus at the next time instant, denoted by $d^*_{i,k+1}$, is set by $d^*_{i,k+1} = \hat{d}_{i,k+1} + \eta \cdot (\bar{d}_{i,k} - d_{i,k})$, where η is a constant and $\hat{d}_{i,k+1}$ is given by Eq. (1) with $f_i(\cdot)$ realized by a trained linear AR model. Note that as the AR model captures the trend of demand only, it alone cannot drive the simulations. Thus, we use the term $\eta \cdot (\bar{d}_{i,k} - d_{i,k})$ in $d^*_{i,k+1}$ to capture the reason for demand increase in the data trace.

The buses curtail their demands with a certain *commitment* $\xi \in (0, 1)$. Specifically, the actual demand of bus i at time instant $k+1$, i.e., $d_{i,k+1}$, is set to an interpolated value $\xi D_{i,1} + (1-\xi) d^*_{i,k+1}$, where $D_{i,1}$ is the demand ceiling in the curtailment schedule for the current cycle. Once bus i is selected to implement load shedding at time instant k_0, the shed demand is $\Delta d_i = \rho \cdot d_{i,k_0}$ and the actual demand is reset by $d_{i,k_0} = d_{i,k_0} - \Delta d_i$. The actual demand at any subsequent cycle under load shedding is set to $d_{i,k} = \bar{d}_{i,k} - (\bar{d}_{i,k_0} - d_{i,k_0})$, such that the simulated actual demand preserves the shape of the real demand trace over time, but is reduced by a constant $(\bar{d}_{i,k_0} - d_{i,k_0})$. Once bus i is restored from the load shedding, its actual demand is reset to $d_{i,k} = d_{i,k} + \Delta d_i$. To simplify the simulations, we set the cycle length to 5 minutes, which is the same as the period of the loading data. Default settings for other parameters are: $T_W = 60$ min, $T_R = 40$ min, $T_E = 20$ min, $H = 2$, $\sigma_0 = 0.01$ p.u., $\eta = 0.5$, $\rho = 5\%$, and $R = 24$.

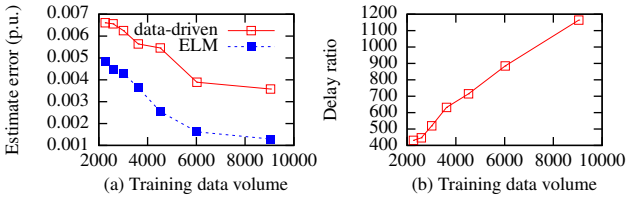

Figure 9: Estimation errors and delays of ELM-based and data-driven safety assessments.

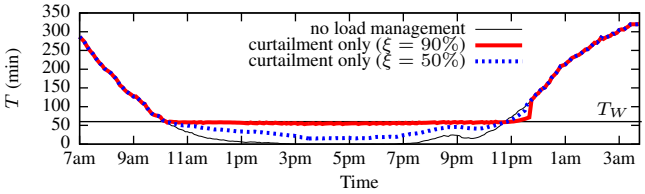

Figure 10: Peak hours of Aug 02, 2012 ($H = 3$).

6.2 Simulation Results

6.2.1 *Effectiveness of ELM-Based Safety Assessment*

We compare our ELM-based approach with a *data-driven* baseline approach. This baseline approach stores the training data set consisting of a large number of data pairs $\langle \mathbf{L}, P \rangle$. At run time, it searches the data pair with an $\mathbf{L}$ closest to the input $\mathbf{L}$ in terms of Euclidean distance and outputs the P of the pair. Fig. 9 compares the estimation errors and delays of the ELM-based and baseline approaches. Consistent with intuition, their estimation errors shown in Fig. 9(a) decrease with the training data volume. Our ELM-based approach is more accurate. Fig. 9(b) shows the ratio of the delays of the baseline and ELM-based approaches, where the delay of ELM is within $[0.10, 0.15]$ milliseconds. The delay of the baseline approach increases with the training data volume, and is much longer than that of ELM.

6.2.2 *Effectiveness of MPC-Based Load Curtailment*

In this set of simulations, we disable the hysteresis-based load shedding and focus on the evaluation of MPC-based load curtailment. Fig. 10 shows the evolution of T during the peak hours of a day. Without any load management, the system is unsafe (i.e., T is around zero) for more than four hours. With load curtailment, if the customer commitment is high (90%), T can be maintained at around T_W. If the customer commitment is low (50%), T keeps below T_W but will not reach zero. Note that the system exits the load curtailment phase if T has been above T_W for a certain time duration (0.5

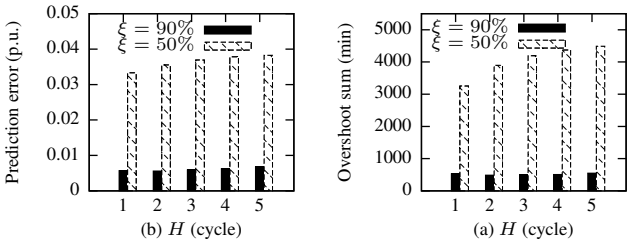

Figure 11: Impact of H on load curtailment.

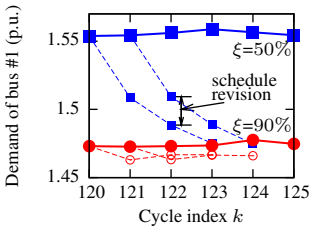

Figure 12: Scheduled demand ceilings and actual demand.

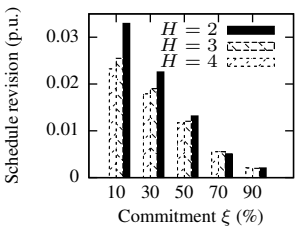

Figure 13: Average schedule revision versus commitment.

hours in Fig. 10).

We also evaluate the impact of H on the system performance by the following two metrics. First, for a certain time instant, we compute the error between the actual demand and the predicted demand given by Eq. (2), averaged over all prediction horizons $h \in [1, H]$. We further average the errors over all the time instants. Fig. 11(a) plots the average prediction error versus H. Second, we refer to the sum of $(T_W - T)$ over time when T is below T_W as *overshoot sum*, which characterizes the effectiveness of load curtailment. Fig. 11(b) plots the overshoot sum versus H. From Figs. 11(a) and 11(b), we can see that both the prediction error and overshoot sum decrease with commitment. When $\xi = 90\%$ and $\xi = 50\%$, both metrics are minimal when $H = 2$ and $H = 1$, respectively. For MPC, a larger H does not necessarily improve the overall performance due to a decreasing prediction accuracy with h [8]. Moreover, customer commitment also significantly affects the prediction accuracy. Nevertheless, we can see that the overall performance of load curtailment is not sensitive to H. Thus, a small setting for H (e.g., 1 to 3) will suffice.

Fig. 12 shows the actual demand of bus #1 under different commitment levels for a few cycles around 5pm in Fig. 10. With a low commitment, the demand of the bus is high, leading to a low T in Fig. 10. Fig. 12 also plots the scheduled demand ceilings computed at $k = 120$ and $k = 121$. For a low commitment, a demand ceiling for a certain cycle needs to be significantly changed due to the reduced prediction accuracy caused by the low commitment. Such a change is referred to as *schedule revision*, which is illustrated in Fig. 12. Fig. 13 shows the average schedule revision versus commitment with different optimization horizons. We can see that the schedule revision decreases with commitment. It implies that, if the customers follow the curtailment schedules better, their schedules change less in return, thereby mitigating the challenges in power consumption planning.

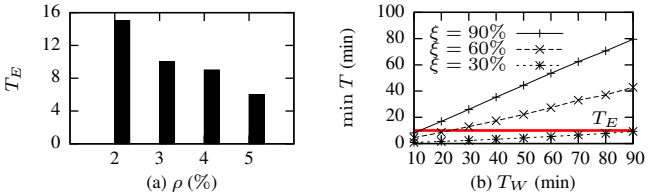

Figure 14: Configurations of T_E and T_W.

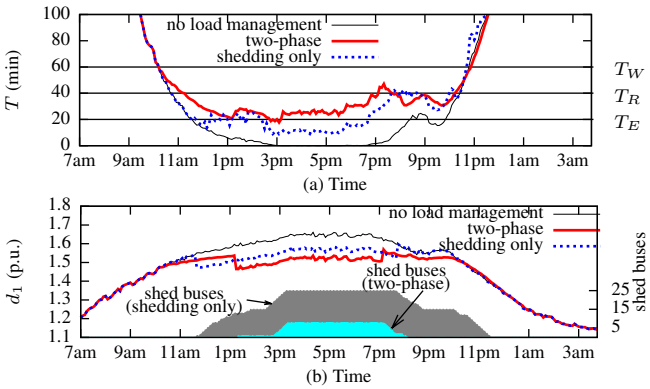

Figure 15: Peak hours of Aug 02, 2012 ($\xi = 30\%$).

6.2.3 *Configurations of T_E and T_W*

Fig. 14(a) shows the minimum T_E that ensures a non-zero T versus ρ. For instance, if ρ is set to 3% to meet a tolerable level of disturbance due to load shedding, from Fig. 14(a), T_E must be larger than 10 minutes. Fig. 14(b) shows the minimum of T in the simulations versus T_W under different commitment levels. If T_E is 10 minutes (represented by the horizontal line), T_W should be set to 90, 20, and 13 minutes if the observed commitment is 30%, 60%, and 90%, respectively.

6.2.4 *Performance of Two-Phase Load Management*

We compare the two-phase load management scheme with a baseline scheme with hysteresis-based load shedding only. Fig. 15(a) shows the evolution of T. Fig. 15(b) shows the demand of bus #1 as well as the number of shed buses. Under the two-phase scheme and the setting $\xi = 30\%$, at most 8 buses need to shed any load. Note that if ξ is higher than 51%, no buses need to shed load. In contrast, for the baseline scheme, all the 25 buses need to shed load from 3pm to 7pm. This result shows that with load curtailment, the chance of load shedding can be reduced.

7. CONCLUSION AND FUTURE WORK

We presented a two-phase smart-grid load management scheme that switches between the load curtail-

ment and shedding phases, in which the swtiching is controlled by a new safety metric called *time-to-being-unsafe*. The scheme allows the grid to correct its undersupply gracefully in the common case, when users participate in the collaborative demand-response as expected. At the same time, it guarantees to avert unacceptable system failure even in the case of erratic customer behavior, by falling back on conventional load shedding in an assured and timely manner where necessary. Simulations based on a 37-bus system and real traces of electrical load demonstrate the features of the proposed design.

For future research, succinct but realistic models of customer commitment to curtailment schedules are interesting. Inclusion of these models in the MPC-based curtailment scheduling may further improve its effectiveness and reduce the need for load shedding. The curtailment scheduling may also be relaxed to admit different per-unit curtailment for different buses, to better account for the heterogeneity of buses in the real world.

Acknowledgments

This research was supported in part by Singapore's Agency for Science, Technology and Research (A⋆STAR), under a research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center.

8. REFERENCES

[1] ELM implementation. `http://bit.ly/115Y3iJ`.

[2] NYISO load. `http://bit.ly/1lwcKLX`.

[3] NYISO load forecast assumptions. `http://bit.ly/15Blkwg`.

[4] Powerworld (version 17). `www.powerworld.com`.

[5] N. Amjady and S. Majedi. Transient stability prediction by a hybrid intelligent system. *IEEE Trans. Power Syst.*, 22(3):1275–1283, 2007.

[6] C. W. Gellings. *The smart grid: enabling energy efficiency and demand response*. The Fairmont Press, Inc., 2009.

[7] I. Genc, R. Diao, V. Vittal, S. Kolluri, and S. Mandal. Decision tree-based preventive and corrective control applications for dynamic security enhancement in power systems. *IEEE Trans. Power Syst.*, 25(3):1611–1619, 2010.

[8] R. Haber, R. Bars, and U. Schmitz. *Predictive Control in Process Engineering-From the Basics to the Applications*. Wiley, 2012.

[9] E. L. Hahne. Round-robin scheduling for max-min fairness in data networks. *IEEE J. Sel. Areas Commun.*, 9(7):1024–1039, 1991.

[10] E. Hollnagel. *The ETTO principle: efficiency-thoroughness trade-off, why things that go right sometimes go wrong*. Ashgate Pub, 2009.

[11] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

[12] E. Karapidakis and N. Hatziargyriou. Online preventive dynamic security of isolated power systems using decision trees. *IEEE Trans. Power Syst.*, 17(2):297–304, 2002.

[13] Y. Kato and S. Iwamoto. Transient stability preventive control for stable operating condition with desired CCT. *IEEE Trans. Power Syst.*, 17(4):1154–1161, 2002.

[14] P. Kundur. *Power system stability and control*. Tata McGraw-Hill Education, 1994.

[15] H. H. Nguyen, R. Tan, and D. K. Y. Yau. Safety-assured collaborative load management in smart grids. Technical report, ADSC, 2014. `http://publish.illinois.edu/cps-security/files/2014/01/load.pdf`.

[16] L. Pereira, D. Kosterev, D. Davies, and S. Patterson. New thermal governor model selection and validation in the wecc. *IEEE Trans. Power Syst.*, 19(1):517–523, 2004.

[17] M. Roozbehani, M. Dahleh, and S. Mitter. Volatility of power grids under real-time pricing. *IEEE Trans. Power Syst.*, 27(4):1926–1940, 2012.

[18] L. Sha. Using simplicity to control complexity. *IEEE Software*, 18(4):20–28, 2001.

[19] J. Short, D. Infield, and L. Freris. Stabilization of grid frequency through dynamic demand control. *IEEE Trans. Power Syst.*, 22(3):1284–1293, 2007.

[20] K. Sun, S. Likhate, V. Vittal, V. Kolluri, and S. Mandal. An online dynamic security assessment scheme using phasor measurements and decision trees. *IEEE Trans. Power Syst.*, 22(4):1935–1943, 2007.

[21] J. Taneja, R. Katz, and D. Culler. Defining CPS challenges in a sustainable electricity grid. In *ICCPS*, 2012.

[22] B. W. Wah, Y. Chen, and T. Wang. Simulated annealing with asymptotic convergence for nonlinear constrained optimization. *J. Global Optimization*, 39(1):1–37, 2007.

[23] X. Wang, N. Hovakimyan, and L. Sha. L1Simplex: fault-tolerant control of cyber-physical systems. In *ICCPS*, 2013.

[24] Y. Xu, Z. Y. Dong, J. H. Zhao, P. Zhang, and K. P. Wong. A reliable intelligent system for real-time dynamic security assessment of power systems. *IEEE Trans. Power Syst.*, 27(3):1253–1263, 2012.

[25] Z. Xu, J. Ostergaard, and M. Togeby. Demand as frequency controlled reserve. *IEEE Trans. Power Syst.*, 26(3):1062–1071, 2011.