

Collaborative Load Management with Safety Assurance in Smart Grids

RUI TAN, Nanyang Technological University, Singapore

HOANG HAI NGUYEN, Advanced Digital Sciences Center, Illinois at Singapore

DAVID K. Y. YAU, Singapore University of Technology and Design, Singapore; Advanced Digital Sciences Center, Illinois at Singapore

Load shedding can combat the overload of a power grid 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 overload 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 overload when there are no immediate safety concerns, but (ii) falls back to 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 load shedding.

CCS Concepts: • **General and reference** → **Reliability**; • **Networks** → **Cyber-physical networks**;

General Terms: Performance, Reliability

Additional Key Words and Phrases: Smart grid, demand response, safety assessment, contingency analysis

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 resolve the overload. While this strategy protects the grid against expensive equipment damage or prolonged overload 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

Author's addresses: Rui Tan, N4-02C-85, 50 Nanyang Avenue, Singapore 639798; Hoang Hai Nguyen, Coordinated Science Laboratory, 1308 West Main St., Urbana, IL 61801, USA; David K. Y. Yau, Information Systems Technology and Design, Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372. Part of this work was completed while Rui Tan was with Advanced Digital Sciences Center, Illinois at Singapore. Hoang Hai Nguyen is now with University of Illinois at Urbana-Champaign.

This work was supported in part by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore Agency for Science, Technology and Research (A*STAR), in part by a Start-up Grant at Nanyang Technological University, in part by the National Research Foundation (NRF), Prime Ministers Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-40) and administered by the National Cybersecurity R&D Directorate.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© YYYY ACM. 0000-0000/YYYY/01-ARTA \$15.00

DOI: <http://dx.doi.org/10.1145/0000000.0000000>

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 [Taneja et al. 2012], 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) [Gellings 2009] 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 resolve the overload. By allowing customers to participate in the load management, the impact of overload will likely be milder and more bearable than in the case of the load shedding. 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 be ignored and in fact requires heightened attention. During an overload period, a grid may operate at its stability limit and become unsafe in that its frequency will exceed prescribed safety-critical thresholds following a contingency. An excessive deviation of the system frequency will trigger generators’ self protection mechanisms including disconnection from the grid, leading to widespread blackouts. Both load curtailment and load shedding can prevent the grid from entering an unsafe state, but collaborative load curtailment is a preferred performance feature. However, 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 [Hollnagel 2009]. 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 overload, 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. 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 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 power flow sensors 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. Section 2 reviews background and related work. Section 3 overviews our approach. Section 4 and Section 5 present the IS for safety assessment and the proposed load management framework. Section 6 presents experiment results on a physical 16-bus power system testbed. Section 7 presents simulation results. Section 9 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* [Kundur 1994]. 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 [Kundur 1994]. 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 approaches to restoring the safety.¹ 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) [Sun et al. 2007; Amjady and Majedi 2007; Xu et al. 2012] have been proposed as alternatives. Trained with data generated by offline T-D simulations, these ISes can provide assessment results rapidly. Sun et al. [Sun et al. 2007] build a decision tree to classify a power grid's safety based on data from phasor measurement units. Amjady and Majedi [Amjady and Majedi 2007] 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 [Xu et al. 2012], Xu et al. employ extreme learning machines (ELMs) to construct feedforward neural networks 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 [Sun et al. 2007; Amjady and Majedi 2007] or calculate properties of tolerable contingencies [Xu et al. 2012], this paper

¹In this paper, *load shedding* resolves overload to prevent the grid from entering an unsafe state. When a contingency occurs in an unsafe grid, the conventional *under-frequency load shedding* may be triggered.

constructs ELMs² 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 [Kato and Iwamoto 2002] propose an approach combining generation shift and generator voltage control to extend CCTs. Karapidakis and Hatziargyriou [Karapidakis and Hatziargyriou 2002] propose an online iterative algorithm that calculates a new generation dispatch until a decision tree classifies the grid as safe. Genc et al. [Genc et al. 2010] 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 [Kato and Iwamoto 2002; Karapidakis and Hatziargyriou 2002; Genc et al. 2010] 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 [Xu et al. 2011; Short et al. 2007] explore decentralized demand-side management to improve system performance. Xu et al. [Xu et al. 2011] 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. [Short et al. 2007] 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 [Sha 2001], 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 [Wang et al. 2013]. 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 Section 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 Section 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.

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:

²In this paper, we use “ELM” to refer to the ELM training algorithm and the trained single hidden-layer feedforward neural network interchangeably.

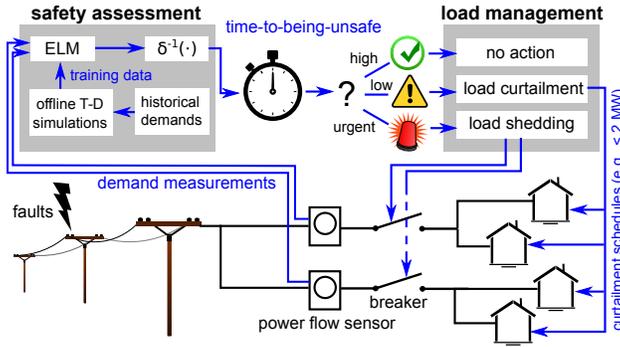


Fig. 1. Overview of our approach (arrows represent data or control flows).

- **A new safety metric:** Based on key observations from extensive T-D simulations and analysis, 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 [Huang et al. 2006] 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 distributed customers. We develop an algorithm for the curtailment scheduling based on a MPC method, which computes a list of suggested *demand ceilings* 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. Hence, DR programs can be introduced incrementally for selected (growing) subsets of the customers.

4. REAL-TIME SAFETY ASSESSMENT

This section defines TTBU and presents the design of ELM to learn TTBU. Lastly, a case study is provided to illustrate the ELM-based real-time safety assessment. Table I provides a summary of notations used throughout the paper, with their description and the sections where they are defined.

4.1. Time-to-Being-Unsafe (TTBU)

We propose the TTBU metric based on key observations from extensive T-D simulations using PowerWorld [PowerWorld 2015]. PowerWorld is a power system simulator that is widely used in the power industry for transmission planning, operation analy-

Table I. Summary of Notation

Symbol	Definition	Section
\mathbf{L}	loading	4.1.2
Δt	time duration	4.1.2
$\delta_i(\Delta t)$	demand ramp-up of bus i	4.1.2
$\delta(\Delta t)$	demand ramp-up of any bus	4.1.2
T	time-to-being-unsafe (TTBU)	4.1.2
P	power-distance-to-being-unsafe	4.1.2
T_W	warning threshold	5.1
T_R	restorative threshold	5.1
T_E	emergent threshold	5.1
k	cycle index	5.1
$d_{i,k}$	actual demand of bus i at beginning of k th cycle	5.2.1
$\hat{d}_{i,k+1}$	predicted demand of bus i at beginning of $(k+1)$ th cycle	5.2.1
$f_i(\cdot)$	demand prediction function for bus i	5.2.1
R	number of most recent demand levels used in prediction	5.2.1
h	prediction horizon	5.2.1
H	optimization horizon	5.2.1
\mathbf{S}_i	curtailment schedule for bus i	5.2.1
$D_{i,h}$	demand ceiling for bus i at prediction horizon h	5.2.1
x_h	demand curtailment	5.2.2
\mathbf{X}	$\mathbf{X} = [x_1, \dots, x_H]$	5.2.2
$\tilde{d}_{i,k+h}$	predicted demand of bus i at time instant $k+h$ after curtailment	5.2.2
\tilde{T}_{k+h}	predicted TTBU	5.2.2
$C(\mathbf{X})$	total cost function	5.2.2
$\sigma(\mathbf{X})$	variation of curtailments	5.2.2
σ_0	maximum allowed curtailment variation	5.2.2
ρ	load shedding percentage	5.3
ξ	commitment	7.1

sis, 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.³ Loading indicates the level of stress imposed on a grid. Intuitively, 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. In the rest of this section, we first present simulation and analytic results on the impact of loading on the generator speed after the occurrence of a transmission line fault (Section 4.1.1). Based on the results, we formally define TTBU (Section 4.1.2).

4.1.1. Impact of Loading on Post-Fault Generator Speed. We first use a set of T-D simulations for the WSCC 9-bus system [ICSEG 2015] produced by PowerWorld [PowerWorld 2015] to illustrate the effect of loading on grid safety. The topology of this system can be found in [ICSEG 2015]. 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 contin-

³Power 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.

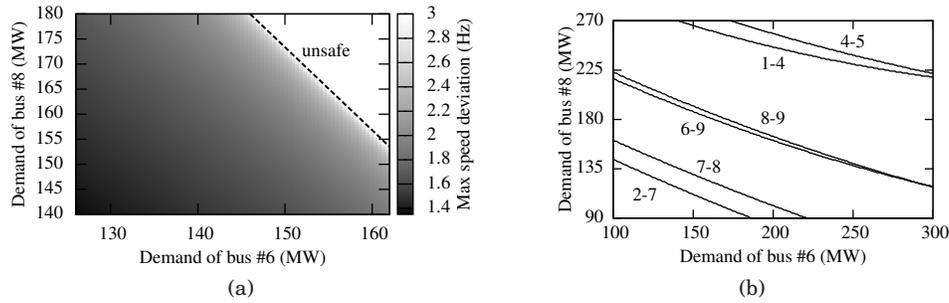


Fig. 2. (a) Maximum generator speed deviation (represented by a gray scale) after a fault vs. the demand of two buses. The white area is the unsafe region with a boundary marked by the dashed curve. (b) Safety boundaries for different contingencies at different transmission lines (labeled by their respective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

gency happens when all the three conductors of the line are shorted together, due to say a lightning strike. 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 bus⁴ 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. In the following, based on certain fault models, we provide an analysis on the impact of a line fault on generator speed to explain the monotonic property observed in the above simulations.

We use the two-bus system in Fig. 3(a) to approximate a multi-bus system when a fault occurs at a line of reactance X_1 . The network (excluding the faulty line) connecting the two terminal buses of the faulty line is simplified as a line of reactance X_2 . For the two buses, E_1 and E_2 denote the electromotive forces, δ_{10} and δ_{20} denote the voltage angles before the fault. For certain fault models, Fig. 3(a) can be further simplified to Fig. 3(b), where X_0 denotes the total reactance between the two buses before the fault and ΔX denotes the change of total reactance during the fault process. For instance, for a balanced 3-phase solid fault with short circuit phase ignored due to its short time period, $\Delta X(t) = \frac{X_2^2}{X_1 + X_2} (u(t) - u(t - \tau))$, where $u(t)$ is the unit step function and τ is the fault clearing time. Before the fault, the power flow from Bus 1 to Bus 2 (denoted by P_0) is given by $P_0 = \frac{E_1 E_2}{X_0} \sin(\delta_{10} - \delta_{20})$. In practice, $\delta_{10} - \delta_{20}$ is mostly within $(-\frac{\pi}{2}, \frac{\pi}{2})$. To simplify discussion, we assume $\delta_{10} - \delta_{20} \in (0, \frac{\pi}{2})$ so that power flows from Bus 1 to Bus 2.

⁴In the rest of this paper, *bus* refers to a *load bus* unless otherwise stated.

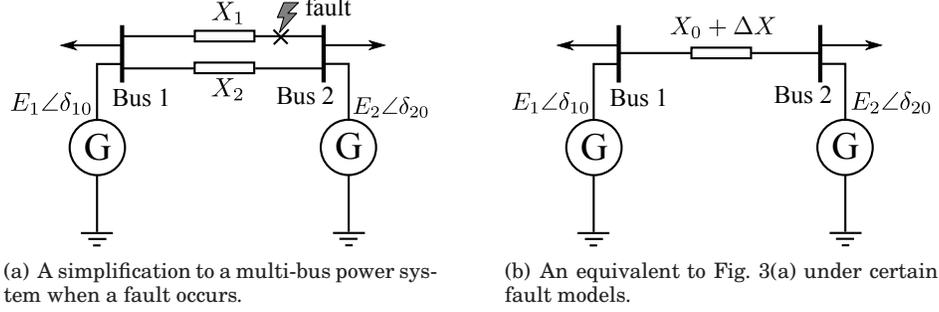


Fig. 3. A system model for analyzing impact of transmission line fault.

Let $\Delta\omega_1$ and $\Delta\omega_2$ denote the speed deviations of the two generators. We analyze the post-fault transients of $\Delta\omega_1$ and $\Delta\omega_2$, based on a classic automatic generation control model [Kundur 1994]. The details of the analysis can be found in Appendix A. The result shows that, in the Laplace domain, the gains of $\mathcal{L}\{\Delta\omega_1\}$ and $\mathcal{L}\{-\Delta\omega_2\}$ with respect to $\mathcal{L}\{\Delta X\}$ are $G_1 = \frac{\mathcal{L}\{\Delta\omega_1\}}{\mathcal{L}\{\Delta X\}} = A_2 \cdot G$ and $G_2 = \frac{\mathcal{L}\{-\Delta\omega_2\}}{\mathcal{L}\{\Delta X\}} = A_1 \cdot G$, where $A_i = D_i + M_i s + \frac{K_i R_i}{s + K_i R_i} T_i(s) \left(\frac{1}{R_i} + \frac{\beta_i}{s} \right)$ for $i = 1, 2$; $G = \frac{s \cdot E_1 E_2 \cdot \sin(\delta_{10} - \delta_{20})}{s \cdot A_1 A_2 X_0^2 + (A_1 + A_2) E_1 E_2 X_0 \cdot \cos(\delta_{10} - \delta_{20})}$; K_i , R_i , β_i , M_i , and D_i are positive constants (their meanings are detailed in Appendix A); and $T_i(s)$ is the transfer function of the turbine. As $(\delta_{10} - \delta_{20}) \in (0, \frac{\pi}{2})$, G_i increases with $(\delta_{10} - \delta_{20})$. Moreover, the pre-fault power flow of the faulty line, i.e., P_0 , increases with $(\delta_{10} - \delta_{20})$. Therefore, the gain G_i increases with P_0 . As a result, the maximum absolute value of the post-fault transient of $\Delta\omega_i(t)$ increases with P_0 .

The above analysis shows that, for certain fault models, the maximum speed deviation of any generator during the post-fault transient process increases with the pre-fault power flow over the faulty line in question. This analytic result explains the empirical monotonic property observed in our T-D simulations. Specifically, as discussed in Section 2.1, grid safety assessment often injects a fault at a line near the generators. Intuitively, the pre-fault power flow over such a line increases with the demand of any bus. Thus, the monotonic property holds under a safety criterion that is based on generator speed deviation. Section 6 further illustrates the monotonic relationship between loading and the grid safety through experiments on a 16-bus physical power system testbed.

4.1.2. Definition of TTBU. Based on the monotonic property, we define TTBU. Let the vector $\mathbf{L} \in \mathbb{R}_{\geq 0}^m$ denote the loading of a system with m buses. For any time instant t_0 , we define $\delta_{t_0}(\Delta t) = [\delta_{1,t_0}(\Delta t), \delta_{2,t_0}(\Delta t), \dots, \delta_{m,t_0}(\Delta t)]$, where $\delta_{i,t_0}(\Delta t)$ is the ramp-up in demand of bus i at time $t_0 + \Delta t$. We conservatively 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 $\mathbf{L} + \delta_{t_0}(\Delta t)$ is safe if $\Delta t < T$, and it is unsafe otherwise. Under the conservative assumption that $\delta_{i,t_0}(\Delta t)$ is positive and increasing, T will not be overestimated. 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 Section 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 Section 4.2. Although this simplification 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 Section 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) [Huang et al. 2006] 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 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 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 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 L . Second, given L , we may run T-D simulations to determine P . By enumerating L in all the records, we generate a large number of data pairs $\langle L, P \rangle$. 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 [Pereira et al. 2004]. 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 L as measured by trusted metering devices such as power flow sensors 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. Section 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 independent 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.

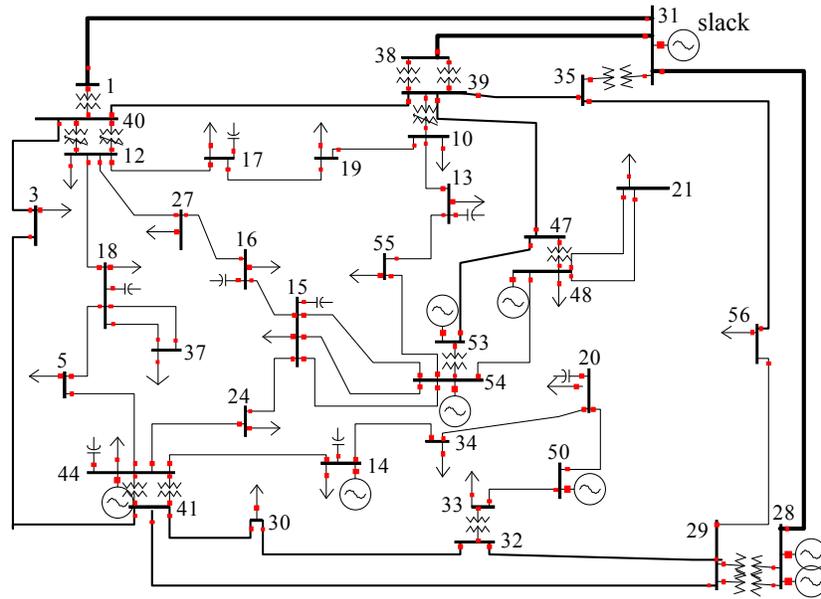


Fig. 4. One-line diagram of the system for case study.

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) [NYISO 2015a]. 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. 4. 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 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 Section 4.1 to learn $\delta_i(\Delta t)$ and $\delta(\Delta t)$. Specifically, the value of $\delta_i(\Delta t)$ is the maximum ramp-up of demand of bus i after a time period of Δt starting from any time point in the load data trace of bus i . The $\delta(\Delta t)$ upper envelopes all $\delta_i(\Delta t)$. Fig. 5(a) 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

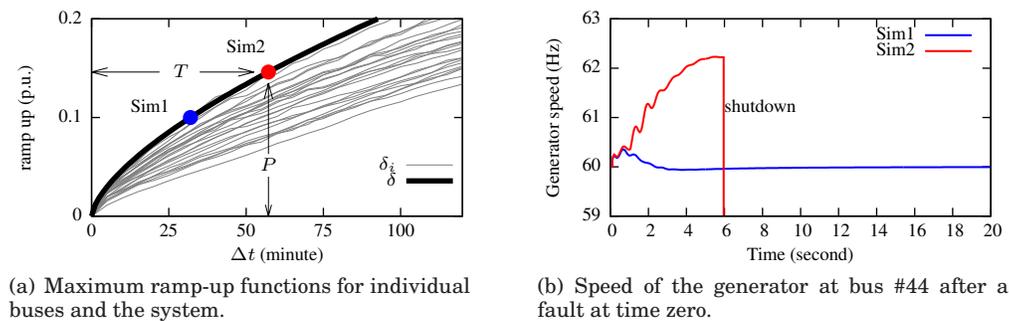


Fig. 5. Ramp-up functions and post-fault generator speed.

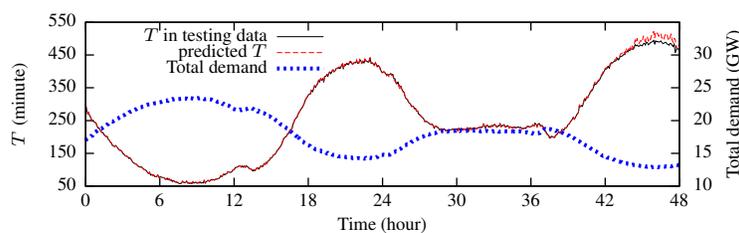


Fig. 6. T and the total demand in two days.

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 Section 4.2 to generate training/testing data for the ELM using PowerWorld. For instance, Fig. 5(b) 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. 5(a). 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. 5(a). By leveraging the monotonic property of grid safety discussed in Section 4.1, we use binary search to speed up the process of finding P given a loading 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 [Guang-Bin Huang 2015] 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 10^5 x 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 Section 5.1. Then, Section 5.2 and Section 5.3 detail the load curtailment and shedding approaches, respectively. Lastly, Section 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 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 ceilings* 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 [Roozbehani et al. 2012a], 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 [Genc et al. 2010; Nguyen et al. 2015] can be applied for the load shedding phase. In this paper, we adopt a hysteresis-based load shedding approach [Nguyen et al. 2015], 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 Section 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 [Roozbehani et al. 2012a], 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 k th 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, an MPC-based algorithm for curtailment scheduling (cf. Section 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 Section 3, the buses may reduce their demand but in this scenario they still exceed the scheduled ceilings. As a result, T keeps decreasing and drops below T_E at $k = 3$. At this point, the grid is in an emergent condition and the hysteresis-based load shedding [Nguyen et al. 2015] is activated to prevent the grid from becoming unsafe. Under this hysteresis-based approach [Nguyen et al. 2015], 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

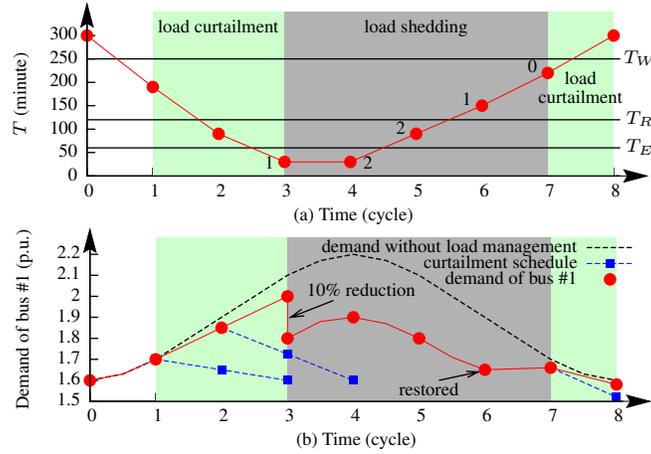


Fig. 7. Illustration of the proposed two-phase load management during peak hours: The digit labels in Subfigure (a) give the numbers of shed buses. Subfigure (b) shows the actual demand and suggested demand ceilings.

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 [Haber et al. 2012]. Since normally bus-level demand is predictable due to its intrinsic strong autocorrelation [Roозbehani et al. 2012a], as well as the predictability of extrinsic factors like weather, MPC is a suitable tool for planning the demand ceilings. In the following, Section 5.2.1 describes the models for demand prediction and curtailment, and Section 5.2.2 develops the proposed MPC-based algorithm.

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 [NYISO 2015b]. 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 $d_{i,k}, d_{i,k-1}, \dots, d_{i,k-R+1}$ 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

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

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.,

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

where h is the *prediction horizon*. In the evaluation in this paper, we adopt a linear autoregressive (AR) model [Roозbehani et al. 2012a], i.e., $\hat{d}_{i,k+1} = \sum_{j=0}^{R-1} a_j d_{i,k-j} + a_R$,

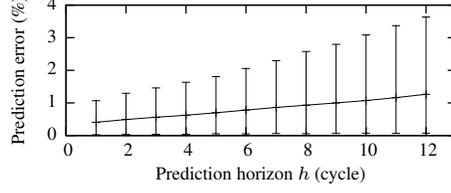


Fig. 8. Relative prediction errors: error bar represents 90% confidence interval; cycle length is 5 minutes; $R = 24$.

which can predict demand accurately. We use one third of the loading data created in Section 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 .

Curtailement Model. The operator of each bus i maintains a *curtailement schedule*, which is a first-in-first-out (FIFO) queue consisting of H demand ceilings. It is formally represented by $S_i = [D_{i,1}, D_{i,2}, \dots, D_{i,H}]$, 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 curtailement scheduling (cf. Section 5.2.2).

At the beginning of each cycle, there is a short *curtailement scheduling session*. During this session, the system operator computes the curtailement schedules and communicates with each bus i to update all the elements in 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 curtailement. 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 curtailement scheduling sessions. The bus can use these estimates to prepare for the curtailements in the subsequent cycles.) Then, the bus duplicates the newest element $D_{i,H}$ and pushes it onto S_i . Although all the elements in S_i will be updated in the next curtailement scheduling session, this self-duplication step simplifies the system design when the load management subsystem switches from the load curtailement phase to the load shedding phase, which will be discussed in Section 5.4.2. The above curtailement model has clear and simple semantics to system operators and customers. Thus, it simplifies the design of supporting DR devices and fosters adoption.

5.2.2. MPC-Based Curtailement Scheduling. This section develops an algorithm to determine the demand ceilings in the curtailement schedules of the buses. Executed in each curtailement scheduling session, the algorithm aims to maintain T at around T_W , assuming that all the buses follow the curtailement schedules exactly. In Section 7, we will evaluate extensively the impact of customer commitment to the curtailement schedules on the performance of the scheduling algorithm. Suppose we need to compute the curtailement schedules for the k th to $(k + H - 1)$ th cycles at time instant k . We denote $x_h \in \mathbb{R}_{\geq 0}$ the *demand curtailement* in p.u. for any bus in the $(k + h)$ th cycle, and define $\mathbf{X} = [x_1, \dots, x_H] \in \mathbb{R}_{\geq 0}^H$. Imposing the same per-unit curtailement on all the buses achieves max-min fairness [Hahne 1991]. Note that since different buses may have different base powers (cf. Section 4.3), the demand curtailements in watts, which are translated from the x_h , vary across the buses. Let $\tilde{d}_{i,k+h}$ denote the predicted demand of bus i at time instant $k + h$ provided that the bus curtailes x_h p.u., i.e.,

$$\tilde{d}_{i,k+h} = f_i(\tilde{d}_{i,k+h-1}, \dots, \tilde{d}_{i,k+1}, d_{i,k}, \dots, d_{i,k+h-R}, \Theta) - x_h. \quad (2)$$

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

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

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:

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

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.

The *curtail scheduling problem* is formulated as follows:

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

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} . As we impose the same per-unit curtailment on all the buses, the complexity of the brute-force search is independent of the number of buses. The evaluation in Section 7 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) [Wah et al. 2007], 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 . Since the search algorithm invokes the single hidden-layer neural network trained by ELM for each candidate solution \mathbf{X} , it has a polynomial time complexity with respect to the input dimension (i.e., the number of buses) due to neural network's polynomial time complexity. Thus, the search algorithm can scale well with the number of buses. It is worth noting that, 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 $\tilde{d}_{i,k+h}$ given by Eq. (2) with $x_h = x_h^*$. As discussed in Section 5.2.1, only $D_{i,1}$ will be applied for

the current cycle and other ceilings in 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 [Haber et al. 2012], 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 [Kundur 1994] to cope with unsafety detected by online contingency analyses such as those discussed in Section 2.2. Integrated with the MPC-based load curtailment in Section 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 [Nguyen et al. 2015] designed to exploit the foresight of TTBU. We now describe briefly its principles, while the details can be found in [Nguyen et al. 2015]. Under this approach, the two thresholds T_E and T_R (cf. Section 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 Section 7, 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., S_i). However, by following the self-duplication step described in Section 5.2.1, each bus i will keep popping the oldest element and pushing a duplicate of the newest element onto 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.

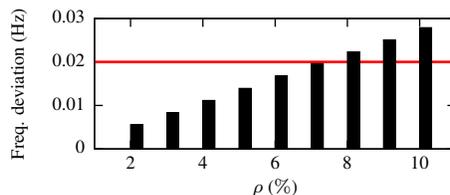


Fig. 9. Speed deviation of the generator at bus #44 after load shedding at bus #21 (red line represents a required upper bound of frequency deviation).

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.

5.4.3. Transients Caused by Load Curtailment and Shedding. A change of load in the load curtailment phase or the load shedding phase will cause transient deviations of the system frequency from the nominal value. To reduce the impact of load curtailment on the system frequency, the system operator can randomize the time instants of sending the curtailment schedules to the buses, within the time window of the curtailment scheduling session. The buses will therefore be unlikely to execute the curtailments simultaneously. Moreover, the system operator may impose an additional constraint of $x_h < x_0$ for the curtailment scheduling problem in Section 5.2.2, where x_0 is a curtailment upper bound that ensures acceptable transient frequency deviations under the curtailments. The setting of x_0 can be obtained via off-line T-D simulations. To reduce the impact of load shedding on the system frequency, as discussed in Section 5.3, the percentage of load to be shed in a cycle should be controlled to avoid an unacceptable system frequency deviation. The system operator can run T-D simulations to determine the setting of this percentage. For instance, Fig. 9 shows the maximum deviation of a generator’s speed from its nominal value versus ρ for the system shown in Fig. 4. If the system operator requires the frequency deviation caused by the load shedding in each cycle to not exceed 0.02 Hz, the setting of ρ should be no larger than 7%. We note that there is no explicit transient physical impact of the transition from one phase to another, because the phases represent operation modes only. Only actual changes of load during curtailment or shedding will cause transient frequency deviations as physical effects.

6. TESTBED EXPERIMENTS

We conduct experiments on a three-phase 16-bus 400V power system testbed to validate the monotonic relationship between the post-fault generator speed deviation and the loading in Section 4.1.1. The 16 buses are connected to form a ring topology as shown in Fig. 10. Each bus is monitored by a smart meter. A variable load is connected to Bus 4 in the system. Its power consumption can be tuned manually using a knob. A Marelli MJB160XA4 13.5kVA synchronous generator is driven by a motor (which simulates a turbine) and connected to Bus 10 in the system. The input power of the motor is supplied by a Yaskawa V1000 current vector drive (CVD), which communicates with a remote computer. Power engineering researchers have implemented the generation control system discussed in Appendix A using LabVIEW on the computer. Specifically, the LabVIEW program retrieves frequency measurements from the smart meters and then sends commands to the CVD to increase or decrease the power output of the motor that simulates a turbine, both using the Modbus/TCP protocol. We note

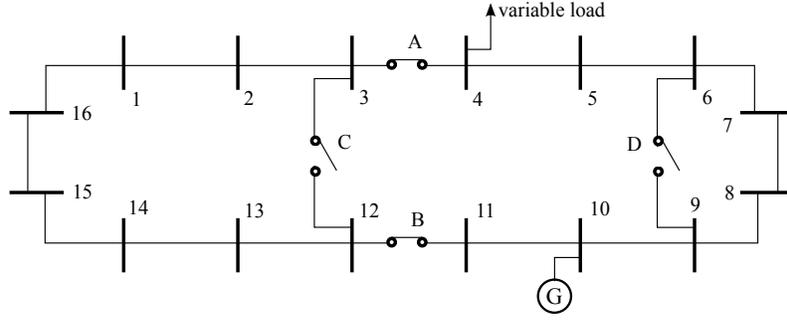


Fig. 10. A 16-bus 400V power system testbed. A 13.5kVA synchronous generator is connected to Bus 10, a variable load is connected to Bus 4, circuit breakers A and B are closed, and circuit breakers C and D are open. Circuit breaker A is manually open to simulate a line loss fault.

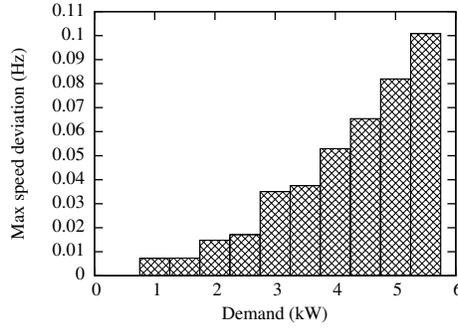


Fig. 11. Maximum generator speed deviation after a line loss fault vs. the demand of the load.

that, in addition to the variable load and the generator connected to Bus 4 and Bus 10, various power system devices are connected to other buses. These devices are not used in our experiments.

In our experiments, we create different loadings by setting the load's power consumption at discrete values from 1.0 kW to 5.5 kW at 0.5 kW intervals. Under each loading, we simulate a line loss fault by manually opening the circuit breaker A shown in Fig. 10 to break the ring structure of the power system. Fig. 11 shows the measured maximum transient generator speed after opening the breaker under different loading settings. We can clearly see the monotonic relationship as observed in our T-D simulations and analysis in Section 4.1.1.

7. TRACE-DRIVEN SIMULATIONS

7.1. Simulation Methodology and Settings

The simulations are based on the 37-bus systems shown in Fig. 4 and the synthetic loading data traces described in Section 4.3. The contingency event and safety condition are the same as in Section 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)

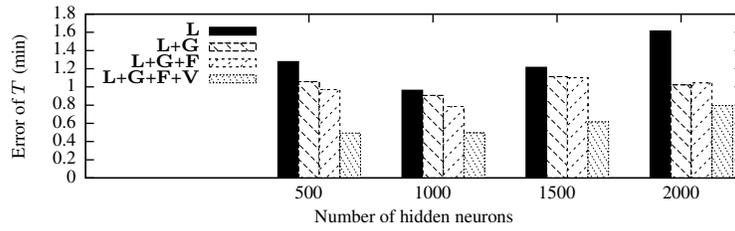


Fig. 12. Estimation error for T under various settings of ELM (L: loading; G: generator outputs; F: branch power flows; V: bus voltages).

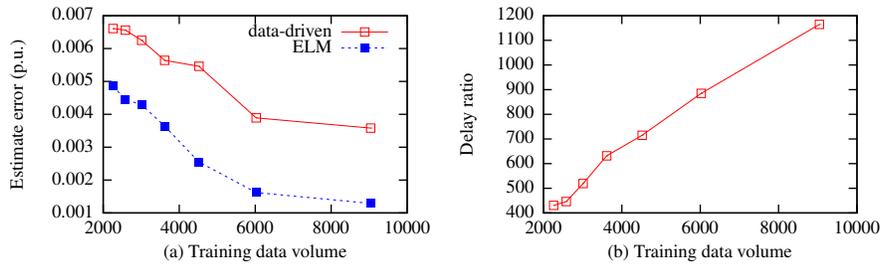


Fig. 13. Estimation errors and delays of ELM-based and data-driven safety assessments.

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 $R = 24$.

7.2. Simulation Results

7.2.1. Effectiveness of ELM-Based Safety Assessment. This set of simulations evaluate the performance of ELM under various settings. Fig. 12 plots the average estimation error for T when the ELM is trained with different combinations of measurable physical quantities of the grid including loading, generator outputs, branch power flows, and bus voltages. We can see that, providing additional inputs (besides loading) to the ELM can improve slightly the estimation accuracy for T (by at most 0.8 minutes). However, doing so will incur additional computation for the MPC-based curtailment scheduling. Specifically, for each candidate solution \mathbf{X} , the additional inputs to the ELM (e.g., power flows) will need to be computed from the predicted loading. The resulting complexity and computation delay may outweigh the benefits of slightly more accurate estimation. Fig. 12 shows that the best setting for the number of hidden neurons is

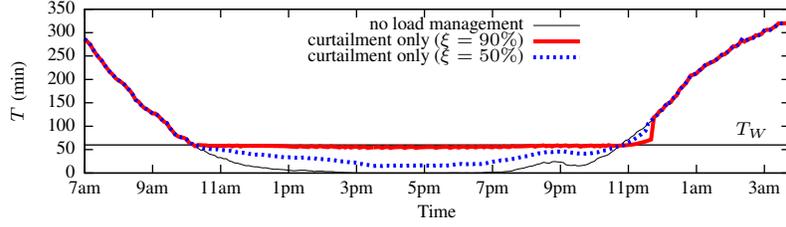
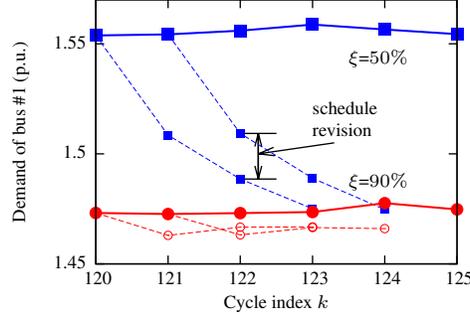
Fig. 14. Peak hours of Aug 02, 2012 ($H = 3$).

Fig. 15. Scheduled demand ceilings and actual demand.

1000 for the ELM trained with loading only. This setting is adopted in the following simulations.

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 (L, P) . At run time, it searches the data pair with an L closest to the input L in terms of Euclidean distance and outputs the P of the pair. Fig. 13 compares the estimation errors and delays of the ELM-based and baseline approaches. Consistent with intuition, their estimation errors shown in Fig. 13(a) decrease with the training data volume. Our ELM-based approach is more accurate. Fig. 13(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.

7.2.2. 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. 14 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 hours in Fig. 14).

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. 17(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. 17(b) plots the overshoot sum versus H . From Figs. 17(a) and

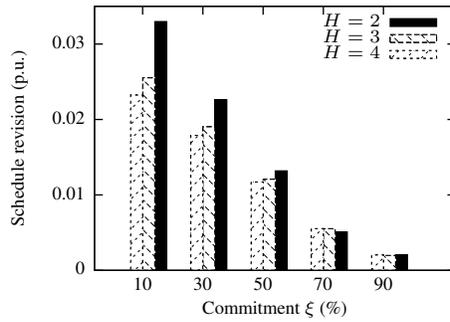
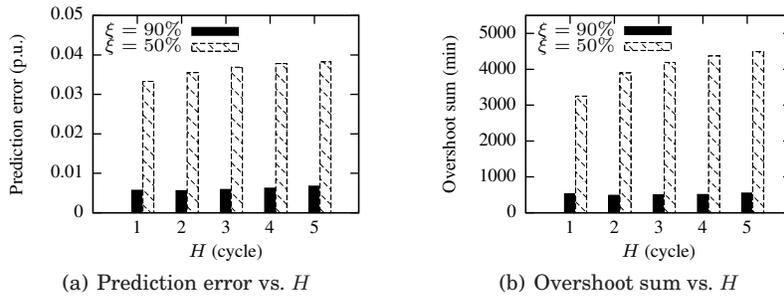


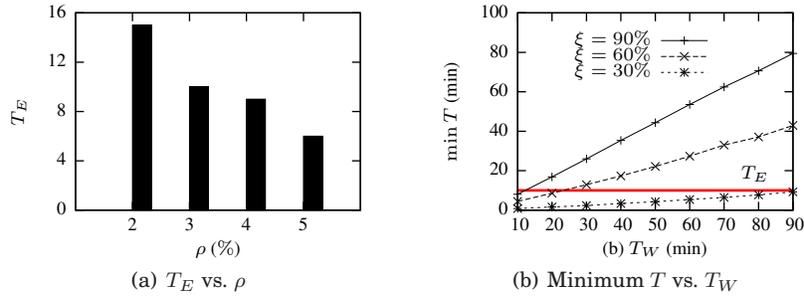
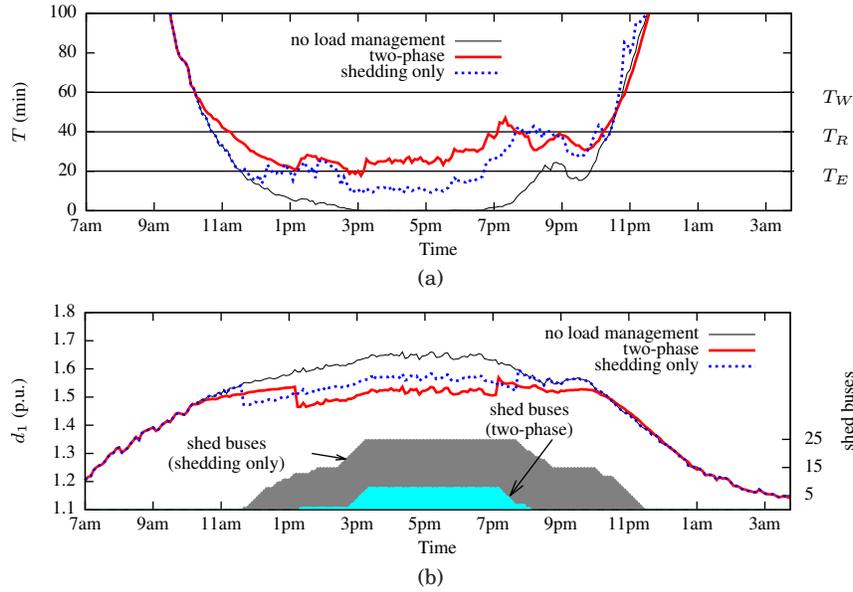
Fig. 16. Average schedule revision versus commitment.

Fig. 17. Impact of H on load curtailment.

17(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 [Haber et al. 2012]. 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. 15 shows the actual demand of bus #1 under different commitment levels for a few cycles around 5pm in Fig. 14. With a low commitment, the demand of the bus is high, leading to a low T in Fig. 14. Fig. 15 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. 15. Fig. 16 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.

7.2.3. Configurations of T_E and T_W . Fig. 18(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. 18(a), T_E must be larger than 10 minutes. Fig. 18(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.

Fig. 18. Configurations of T_E and T_W .Fig. 19. Peak hours of Aug 02, 2012 ($\xi = 30\%$).

7.2.4. 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. 19(a) shows the evolution of T . Fig. 19(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.

8. DISCUSSIONS

This section discusses two issues that have not been addressed in the previous sections and also summarizes the key assumptions made in this paper.

8.1. Open Issues

8.1.1. Customer's incentives to curtail load. The proposed two-phase approach offers opportunities for the customers to participate in the load management by curtailing their

loads, while ensuring the grid's safety by enforcing the load shedding should the customers be uncommitted to the requested curtailments. The risk of totally losing power during the load shedding period may already provide an incentive for the customers to respond to the curtailment requests.

Additionally, there exist various other incentive DR mechanisms that have proved their effectiveness in pilot experiments/deployments. For instance, a pilot experiment [Allcott 2009] conducted from 2003 to 2006 with 693 household consumers in Chicago shows that the household consumers did respond to real-time utility prices. The experiment adopted different methods to inform the households of the real-time prices in its different phases. For instance, a small plastic globe that changed colors in real time on a continuous spectrum from blue to red, indicating low to high prices, was given to each household. The experiment concludes that households can be significantly price elastic and the information technology can significantly increase its price elasticity further (Section 2.4 of [Allcott 2009]). With increasing adoption of smart appliances and home automation systems, the DR is becoming more prevalent and transparent. For instance, in the Load Guard Automatic Price Response Service provided by ComEd [ComEd 2014], a cyber-enabled controller automatically regulates the duty cycling of a central air conditioner based on ComEd's real-time prices. Moreover, some DR programs have become legally required options for consumers, e.g., real-time pricing of electricity in Illinois [Public Act 094-0977 2014]. All these pilot experiments/deployments, as well as current industry/legal trends towards DR, suggest the DR's feasibility and effectiveness.

A detailed quantitative/empirical study of the DR (similar to [Allcott 2009; Roozbehani et al. 2012b]), and translation of the curtailment model in this paper to other DR mechanisms are interesting, but are out of the scope of this paper, whose focus is on the design of the load curtailment scheduling to maintain grid safety. The results of related studies (e.g., [Allcott 2009; Roozbehani et al. 2012b]) can be used to configure relevant parameters of the proposed curtailment scheduling algorithm (e.g., σ_0) or adapt the proposed scheduled demand ceilings to other DR signals. As shown in existing studies [Allcott 2009; Roozbehani et al. 2012b], a customer's response to DR signal has uncertainty. As a challenge to the MPC formulation in Section 5.2.2, this uncertainty may adversely affect the optimality of the curtailment schedule. The resilience of the MPC approach against demand uncertainty is of interest in future work.

8.1.2. Sequential Faults. In Section 4.2, we discussed how to address multiple independent faults. Similarly, we may address multiple contingencies in sequence. A sequence of faults can be treated as a single contingency process. An ELM can be trained based on the T-D simulation data with the sequential contingencies injected. The system operator may identify credible sequential contingencies from, say, cascading failure analysis [Crucitti et al. 2004].

8.2. Summary of Assumptions

This section summarizes the key assumptions made in this paper and discusses their implications on the application of the proposed load management scheme in practice.

- Section 4.1.1 assumes a monotonic relationship between the generators' maximum speed deviation and the demand. This assumption has been empirically verified using simulations and explained analytically in Section 4.1.1. When applying the proposed load management scheme, this assumption can be evaluated via extensive simulations based on the system model of the target power grid.
- Section 4.1.2 assumes that each bus at any time instant has the same maximum ramp-up function. By conservatively choosing the upper envelope of all the actual ramp-up functions as this global ramp-up function, this assumption will hold and the

approach presented in this paper will give a conservative estimate of TTBU. Based on the conservative TTBU estimate, the safety assurance will not be weakened.

- The MPC-based load curtailment scheduling approach in Section 5.2 assumes that the customers will follow the curtailment schedule. As discussed in Section 8.1.1, existing studies have shown effectiveness of DR mechanisms. The load curtailment can be implemented based on these DR mechanisms. Although customers may deviate from their expected responses to DR signals, the load management system can fall back to the load shedding approach to ensure grid safety.

9. CONCLUSION AND FUTURE WORK

We presented a two-phase smart-grid load management scheme that switches between the load curtailment and shedding phases, in which the switching is controlled by a new safety metric called *time-to-being-unsafe*. The scheme allows the grid to correct its overload 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 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. Moreover, how to trade off the cost of load curtailment/shedding (e.g., loss of utility value) with the risk of contingency-induced system failure due to the absence of load curtailment/shedding is an interesting issue for further study.

Acknowledgement

We acknowledge Dr. Hoay Beng Gooi for providing a power system testbed for conducting the experiments described in Section 6.

REFERENCES

- Hunt Allcott. 2009. Real Time Pricing and Electricity Markets. (2009). <http://economics.stanford.edu/files/Allcott3.13.pdf>.
- Nima Amjady and Seyed Farough Majedi. 2007. Transient Stability Prediction by a Hybrid Intelligent System. *IEEE Transactions on Power Systems* 22, 3 (2007), 1275–1283.
- ComEd. 2014. ComEd Residential Real-Time Pricing Program. (2014). <https://rrtp.comed.com/>.
- Paolo Crucitti, Vito Latora, and Massimo Marchiori. 2004. Model for cascading failures in complex networks. *Physical Review E* 69, 4 (2004), 045104.
- Clark W. Gellings. 2009. *The Smart Grid: Enabling Energy Efficiency and Demand Response*. The Fairmont Press, Inc., Georgia, USA.
- Istemihan. Genc, Ruisheng Diao, Vijay Vittal, Sharma Kolluri, and Sujit Mandal. 2010. Decision Tree-Based Preventive and Corrective Control Applications for Dynamic Security Enhancement in Power Systems. *IEEE Transactions on Power Systems* 25, 3 (2010), 1611–1619.
- Guang-Bin Huang. 2015. ELM Implementation. (2015). http://www.ntu.edu.sg/home/egbhuang/elm_codes.html.
- Robert Haber, Ruth Bars, and Ulrich Schmitz. 2012. *Predictive Control in Process Engineering-From the Basics to the Applications*. John Wiley & Sons, New York, USA.
- Ellen L. Hahne. 1991. Round-Robin Scheduling for Max-Min Fairness in Data Networks. *IEEE Journal on Selected Areas in Communications* 9, 7 (1991), 1024–1039.
- Erik Hollnagel. 2009. *The ETTO Principle: Efficiency-Thoroughness Trade-Off, Why Things That Go Right Sometimes Go Wrong*. Ashgate Publishing, Farnham, Surrey, UK.

- Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. 2006. Extreme Learning Machine: Theory and Applications. *Neurocomputing* 70, 1 (2006), 489–501.
- ICSEG. 2015. WSCC 9-Bus System. (2015). <http://publish.illinois.edu/smartergrid/wsc-9-bus-system>.
- E.S. Karapidakis and N.D. Hatziargyriou. 2002. Online Preventive Dynamic Security of Isolated Power Systems using Decision Trees. *IEEE Transactions on Power Systems* 17, 2 (2002), 297–304.
- Yoshinori Kato and Shinichi Iwamoto. 2002. Transient Stability Preventive Control for Stable Operating Condition with Desired CCT. *IEEE Transactions on Power Systems* 17, 4 (2002), 1154–1161.
- Prabha Kundur. 1994. *Power System Stability and Control*. McGraw-Hill, New York, USA.
- Hoang Hai Nguyen, Rui Tan, and David K. Y. Yau. 2015. *Safety-Assured Performance Features in Complex Systems: Smart Grid Collaborative Load Management*. Technical Report. ADSC. <http://publish.illinois.edu/resilient-grid/files/2015/02/master.pdf>.
- NYISO. 2015a. New York Independent System Operator (NYISO) Load Data. (2015). <http://mis.nyiso.com/public/P-58Blist.htm>.
- NYISO. 2015b. New York Independent System Operator (NYISO) Load Forecast Assumptions. (2015). http://www.nyiso.com/public/webdocs/markets.operations/market_data/load_data/FERC890aDALFDesc.pdf.
- Les Pereira, Dmitry Kosterev, Donald Davies, and Shawn Patterson. 2004. New Thermal Governor Model Selection and Validation in the WECC. *IEEE Transactions on Power Systems* 19, 1 (2004), 517–523.
- PowerWorld. 2015. PowerWorld (version 17). (2015). <http://www.powerworld.com>.
- Public Act 094-0977. 2014. (2014). <http://www.ilga.gov/legislation/publicacts/94/PDF/094-0977.pdf>.
- Mardavij Roozbehani, Munther A. Dahleh, and Sanjoy K. Mitter. 2012a. Volatility of power grids under real-time pricing. *IEEE Transactions on Power Systems* 27, 4 (2012), 1926–1940.
- Mardavij Roozbehani, M Ohannessian, D Materassi, and M. A. Dahleh. 2012b. Load-Shifting under Perfect and Partial Information: Models, Robust Policies, and Economic Value. (2012). <https://dahleh.lids.mit.edu/wp-content/uploads/2012/08/2012-Load-ShiftingUnderPerfectAndPartialInformation.pdf>.
- Lui Sha. 2001. Using Simplicity to Control Complexity. *IEEE Software* 18, 4 (2001), 20–28.
- Joe A. Short, David G. Infield, and Leon L. Freris. 2007. Stabilization of Grid Frequency Through Dynamic Demand Control. *IEEE Transactions on Power Systems* 22, 3 (2007), 1284–1293.
- Kai Sun, Siddharth Likhate, Vijay Vittal, V. Sharma Kolluri, and Sujit Mandal. 2007. An Online Dynamic Security Assessment Scheme Using Phasor Measurements and Decision Trees. *IEEE Transactions on Power Systems* 22, 4 (2007), 1935–1943.
- Jay Taneja, Randy Katz, and David Culler. 2012. Defining CPS Challenges in a Sustainable Electricity Grid. In *The 3rd ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*. ACM, Beijing, China, 119–128.
- Benjamin W Wah, Yixin Chen, and Tao Wang. 2007. Simulated Annealing with Asymptotic Convergence for Nonlinear Constrained Optimization. *Journal of Global Optimization* 39, 1 (2007), 1–37.
- Xiaofeng Wang, Naira Hovakimyan, and Lui Sha. 2013. L1Simplex: Fault-Tolerant Control of Cyber-Physical Systems. In *The 4th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*. ACM, Philadelphia, PA, USA, 41–50.
- Yan Xu, Zhao Yang Dong, Jun Hua Zhao, Pei Zhang, and Kit Po Wong. 2012. A Reliable Intelligent System for Real-Time Dynamic Security Assessment of Power Systems. *IEEE Transactions on Power Systems* 27, 3 (2012), 1253–1263.
- Zhao Xu, Jacob Østergaard, and Mikael Tøgeby. 2011. Demand as Frequency Controlled Reserve. *IEEE Transactions on Power Systems* 26, 3 (2011), 1062–1071.

A. POST-FAULT GENERATOR SPEED TRANSIENT

Let P_{12} denote the power flow from Bus 1 to Bus 2 in Fig. 3. Let X denote the total reactance between the two buses. Note that X is a variable over time due to the fault. From power flow analysis, we have

$$P_{12} = \frac{E_1 E_2}{X} \sin(\delta_1 - \delta_2). \quad (4)$$

In the rest of this appendix, let Δx denote the change of a quantity x from its pre-fault operating point and \tilde{x} denote the Laplace transform of x . The change of P_{12} (i.e., ΔP_{12}) can be approximated by the linear component of the total derivative of P_{12} at

the pre-fault operating point $(X, \delta_1, \delta_2) = (X_0, \delta_{10}, \delta_{20})$. Specifically,

$$\begin{aligned} \Delta P_{12} &= \left. \frac{\partial P_{12}}{\partial X} \right|_{X=X_0} \cdot \Delta X + \left. \frac{\partial P_{12}}{\partial \delta_1} \right|_{\delta_1=\delta_{10}} \cdot \Delta \delta_1 + \left. \frac{\partial P_{12}}{\partial \delta_2} \right|_{\delta_2=\delta_{20}} \cdot \Delta \delta_2 \\ &= -\frac{E_1 E_2 \sin(\delta_{10} - \delta_{20})}{X_0^2} \cdot \Delta X + \frac{E_1 E_2 \cos(\delta_{10} - \delta_{20})}{X_0} \cdot (\Delta \delta_1 - \Delta \delta_2). \end{aligned} \quad (5)$$

In practice, $\delta_{10} - \delta_{20}$ is mostly within $(-\frac{\pi}{2}, \frac{\pi}{2})$. For simplicity of exposition, we assume that $\delta_{10} - \delta_{20} \in (0, \frac{\pi}{2})$ such that the power flows from the left bus to the right bus before the fault. In time domain, we have $\Delta \delta_i(t) = \int_0^t \Delta \omega_i(t') dt'$, where $i \in \{1, 2\}$. In the Laplace domain, this relationship is expressed by $\widetilde{\Delta \delta}_i = \frac{1}{s} \cdot \widetilde{\Delta \omega}$, where s denotes the complex variable in the Laplace domain. Thus, in the Laplace domain, Eq. (5) can be written as

$$\widetilde{\Delta P_{12}} = -\frac{E_1 E_2 \sin(\delta_{10} - \delta_{20})}{X_0^2} \cdot \widetilde{\Delta X} + \frac{E_1 E_2 \cos(\delta_{10} - \delta_{20})}{s \cdot X_0} \cdot (\widetilde{\Delta \omega_1} - \widetilde{\Delta \omega_2}). \quad (6)$$

We use a widely adopted generator model [Kundur 1994] that addresses the primary speed control with droop characteristics and the secondary speed control. The two layers of speed control ensure that the speed of a generator can converge to its nominal value in the presence of load changes or a non-fatal fault as observed in Fig. 5 in the manuscript. Specifically, the generator dynamics can be described by the following Laplace-domain equations [Kundur 1994]:

$$\widetilde{\Delta p_{Mi}} = -\frac{K_i}{s + K_i R_i} T_i(s) \widetilde{\Delta \omega_i} - \frac{K_i R_i \beta_i}{s(s + K_i R_i)} T_i(s) \widetilde{\Delta \omega_i}, \quad i = 1, 2, \quad (7)$$

$$\widetilde{\Delta p_{Mi}} - \widetilde{\Delta P_{Li}} + (-1)^i \cdot \widetilde{\Delta p_{12}} = D_i \widetilde{\Delta \omega_i} + M_i s \widetilde{\Delta \omega_i}, \quad i = 1, 2, \quad (8)$$

where, for the generator connected to the bus i , P_{Mi} is the input mechanical power, K_i is a constant gain of the primary speed control, R_i is a constant of droop control, $T_i(s)$ is the transfer function of the turbine, β_i is a constant gain of the secondary speed control, and M_i is the inertia constant; for the load connected to the bus i , D_i is the load-damping constant. Note that K_i , R_i , β_i , M_i , and D_i are all positive. Eq. (7) describes the aggregated dynamics of the primary and secondary speed controls. Eq. (8) is from the law of conservation of energy [Kundur 1994].

Our analysis ignores the changes of load during the transient process of the fault that is often much faster than the load change process. By solving Eqs. (6)-(8) with $\widetilde{\Delta P_{Li}} = 0$, we have the gains of $\widetilde{\Delta \omega_1}$ and $-\widetilde{\Delta \omega_2}$ with respect to $\widetilde{\Delta X}$:

$$\begin{aligned} G_1 &= \frac{\widetilde{\Delta \omega_1}}{\widetilde{\Delta X}} = \frac{s \cdot A_2 E_1 E_2 \cdot \sin(\delta_{10} - \delta_{20})}{s \cdot A_1 A_2 X_0^2 + (A_1 + A_2) E_1 E_2 X_0 \cdot \cos(\delta_{10} - \delta_{20})}, \\ G_2 &= \frac{-\widetilde{\Delta \omega_2}}{\widetilde{\Delta X}} = \frac{s \cdot A_1 E_1 E_2 \cdot \sin(\delta_{10} - \delta_{20})}{s \cdot A_1 A_2 X_0^2 + (A_1 + A_2) E_1 E_2 X_0 \cdot \cos(\delta_{10} - \delta_{20})}, \end{aligned}$$

where

$$\begin{aligned} A_1 &= D_1 + M_1 s + \frac{K_1 R_1}{s + K_1 R_1} T_1(s) \left(\frac{1}{R_1} + \frac{\beta_1}{s} \right), \\ A_2 &= D_2 + M_2 s + \frac{K_2 R_2}{s + K_2 R_2} T_2(s) \left(\frac{1}{R_2} + \frac{\beta_2}{s} \right). \end{aligned}$$

We have assumed that $(\delta_{10} - \delta_{20}) \in (0, \frac{\pi}{2})$. Thus, the gain G_i increases with $(\delta_{10} - \delta_{20})$. Moreover, according to Eq. (4), the pre-fault power flow of the faulty line increases with

$(\delta_{10} - \delta_{20})$ when $(\delta_{10} - \delta_{20}) \in (0, \frac{\pi}{2})$. Therefore, the gain G_i increases with the pre-fault power flow of the faulty line. As a result, the maximum absolute value of the post-fault transient of $\Delta\omega_i(t)$ increases with the pre-fault power flow of the faulty line.