Pacific Northwest
NATIONAL LABORATORY

PNNL-28724

Incentive-Based Control and Coordination of Distributed Energy Resources

May 2019

A Bhattacharya (PNNL)
J Hansen (PNNL)
K Kalsi (PNNL)
J Lian (PNNL)
SP Nandanoori (PNNL)
HM Reeve (PNNL)
V Adetola (UTRC)
F Lin (UTRC)
T Leichtman (Spirae)
SNG Gourisetti (PNNL)
WJ Hofer (PNNL)
S Kundu (PNNL)
L Marinovici (PNNL)
S Niddodi (PNNL)
D Vrabie (PNNL)
M Chiodo (UTRC)
S Yuan (UTRC)
D Wright (Spirae)

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-28724

Incentive-Based Control and Coordination of Distributed Energy Resources

May 2019

A Bhattacharya (PNNL)
SNG Gourisetti (PNNL)
J Hansen (PNNL)
WJ Hofer (PNNL)
K Kalsi (PNNL)
S Kundu (PNNL)
J Lian (PNNL)
L Marinovici (PNNL)
SP Nandanoori (PNNL)
S Niddodi (PNNL)
HM Reeve (PNNL)
D Vrabie (PNNL)
V Adetola (UTRC)
M Chiodo (UTRC)
F Lin (UTRC)
S Yuan (UTRC)
T Leichtman (Spirae)
D Wright (Spirae)

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Higher penetration of renewable generation will increase the demand for adequate (and cost-effective) controllable resources on the grid that can mitigate and contain contingencies locally before they cause a network-wide collapse. This report details a scalable hierarchical control strategy for coordinating a vast number of heterogeneous distributed energy resources (DERs) in order to provide three ancillary grid services: frequency response, regulation, and ramping reserve. The control architecture consists of 1) a Distribution Reliability Coordinator (DRC), which allocates resource responses based on their declared flexibility, 2) aggregator controllers, which control a collection of DERs, and finally, 3) DER controllers at the individual devices. Test results are provided showing scalable performance on thousands of simulated residential DERs (air conditioners and water heaters) as well as results of a hardware-in-the-loop demonstration of commercial HVAC and microgrid equipment providing real responses in coordination with a grid simulation. Technology commercialization barriers and strategy is also discussed.

Executive Summary

The rapid growth of low-inertia renewable energy resources represents an immense opportunity for the U.S. to minimize its carbon footprint, while presenting a challenge for system operators as traditional “spinning” generation resources are displaced. This transformation requires solutions to robustly and cost-effectively manage dynamic changes on the grid while ensuring quality of service. There is growing recognition that distributed energy resources (DERs – loads, distributed generation, storage, electric vehicles, etc.) represent a great potential to perform this function. This report details a scalable hierarchical control strategy for coordinating thousands of heterogeneous DERs in order to provide three ancillary gird services: 1) frequency response, a rapid (within 2 seconds) service used to maintain the grid at 60 Hz; 2) regulation response, a larger response used to balance on-going differences in supply and demand on the grid; and, 3) ramping reserve, used to bring on capacity to replace lost generation.

The developed control architecture consists of 1) a Distribution Reliability Coordinator (DRC), 2) aggregator controllers, and finally, 3) DER controllers. The DRC coordinates all the flexible loads and distributed energy resources on the distribution system and bids the aggregate flexibility into the wholesale market. The DRC also performs resource allocation (deciding how the flexible devices on the distribution system will be dispatched) to meet the requirements for each of the three ancillary services. The aggregators control a collection of devices that are geographically co-located and provide a platform to aggregate their flexibility and mitigate uncertainty associated with a single flexible load by managing a large aggregation of similar loads (so that any stochastic effects average out). Finally, DER controllers (for example, thermostats for air conditioners, chargers for electric vehicles) adjust the power consumption of individual devices to ensure accurate and reliable delivery of the requested grid services.

Evaluation and validation of this approach was conducted with a mixture of simulation and hardware testing for a range of DERs in residential (air conditioners and water heaters), commercial HVAC (ventilation fans and central chiller plants), and micro-grids (battery energy storage, solar PV etc.). This culminated in DER hardware and simulation models operating in conjunction with grid simulations as part of a large scale federated test-bed demonstration. Simulations of over 10,000 controllable residential devices (a combination of air conditioners, water heaters, batteries, and electric vehicles) demonstrated that this advanced control approach could simultaneously meet all three grid services. Furthermore, hardware-in-the-loop testing was performed with operational micro-grid and commercial building HVAC equipment. The micro-grid equipment (representing PV, battery storage, and back-up generators) demonstrated the ability to meet all frequency response requirements (except ramping time). The commercial building HVAC ventilation equipment successfully provided the regulating response service.

An Industry Advisory Board (IAB) was formed and engaged throughout the project to get feed-back on project risks and potential commercialization barriers. Their input was invaluable in identify key risks stemming from regulatory and market access challenges, financial risks from under-performance, and competitive threats from utility-scale solutions such as integrated storage. A phased commercialization plan was developed that addresses continued risk reduction (value proposition refinement and performance uncertainty assessments), piloting through field demonstrations with key stakeholders, and finally, deployment as an emerging standard for DER control.

Acronyms and Abbreviations

AC	Air Conditioner
ACE	Area Control Error
AHU	Air Handling Unit
ALC	Automated Logic Corporation
API	Application Programming Interface
ARPA-E	Advanced Research Programs Agency-Energy
BAS	Building Automation System
BESS	Battery Energy Storage System
CCSI	Control of Complex Systems Initiative
CDA	Communications Design and Architecture
DER	Distributed Energy Resource
DNP	Distributed Network Protocol
DR	Demand Response
DRC	Distribution Reliability Coordinator
DSO	Distribution System Operator
DSP	Duct Static Pressure
EWH	Electric Water Heater
FNCS	Framework for Network Co-simulation
FOA	Funding Opportunity Announcement
FR	Frequency Regulation
FRC	Frequency Regulation Controller
HPBT	High Performance Building Test-bed
HIL	Hardware-in-the-Loop
HVAC	Heating, Ventilation, and Air Conditioning
IAB	Industrial Advisory Board
ICCP	Inter-Control Center Communications Protocol
IOU	Investor Owned Utilities
ISO	Independent System Operator
NODES	Network Optimized Distributed Energy Systems
OEM	Original Equipment Manufacturer
PID	Proportional Integral Differential
PNNL	Pacific Northwest National Laboratory
QoS	Quality of Service
RMT	Reserve Magnitude Target
RMVT	Reserve Magnitude Variability Tolerance
RTO	Regional Transmission Organizations
SCADA	Supervisory Control and Data Acquisition
SCE	Southern California Edison
T&D	Transmission and Distribution
UTRC	United Technologies Research Center
VAV	Variable Air Volume

VB	Virtual Battery
VFD	Variable Frequency Drive

Acknowledgments

This work was carried out (contract DE-AC02-76RL01830) under the support from the U.S. Department of Energy as part of their ARPA-E NODES program.

Contents

Figures

Tables

1.0 Introduction

1.1 Motivation

Many regions in the United States are seeing a rapid transition to renewable energy resources for gird power generation. For example, on March 5, 2018, solar power generation set record instantaneous percentage contributions in California, contributing over 49% (9.4 GW) of its power supply [2]. Furthermore, Hawaii has committed to meeting 100% of its electrical demand from renewables by 2045 [3]. This transformation requires solutions to robustly and cost-effectively manage dynamic changes on the grid while ensuring quality of service to end-users. There is growing recognition that distributed energy resources (DERs – loads, distributed generation, storage, electric vehicles, etc.) represent a great potential to perform this function. For example, advances in direct digital control of building systems, combined with the increased connectivity of end devices now enable greater participation. However, operators have concerns about their controllability and dependability, especially when not under their direct control. This report presents the work of Pacific Northwest National Laboratory (lead), United Technologies Research Center, Southern California Edison, and Spirae, to develop and test a hierarchical control framework for coordinating the flexibility of a full range of DERs to supply reserves to the electric power grid.

1.2 Objectives of the NODES program

The work presented in this report was funded under the The Network Optimized Distributed Energy Systems (NODES) Program sponsored by the Department of Energy's Advanced Research Programs Agency-Energy (ARPA-E). The goal of the NODES program was to promote the development of *"transformational grid control algorithms and architectures that optimize the usage of flexible load and DERs"* [1]. A key requirement was to reliably manage dynamic changes in the grid by leveraging DERs, while having minimal impact on customer quality of service (for example indoor comfort provided by air conditioning).

1.2.1 Grid Services

The NODES program sought solutions across three categories of ancillary grid services: 1) frequency response reserve, 2) synthetic regulation reserve, and 3) ramping reserve. This project addressed all three service categories. The ability of a device (or population of devices) to meet these grid services was evaluated using a set of key metrics summarized in Table 1 and Figure 1. These metrics are:

- **Initial Response time**: the time required to start responding when a service is requested. This includes communication and control latencies.
- **Ramp time**: the time elapsed between the start of the response and reaching the desired value. This can often be bound by safe-guards on equipment limiting their rate of response.
- **Duration**: the minimum amount of time the service should be provided.
- **Reserve Magnitude Target (RMT)**: the amount of grid service provided as a proportion of the system load.

- **Reserve Magnitude Variability Tolerance (RMVT)**: quantifies the maximum tolerated deviation from the reserve target after the initial ramping interval.
- **Availabililty**: the fraction of time that the DERs are able to provide ancillary services to the grid.

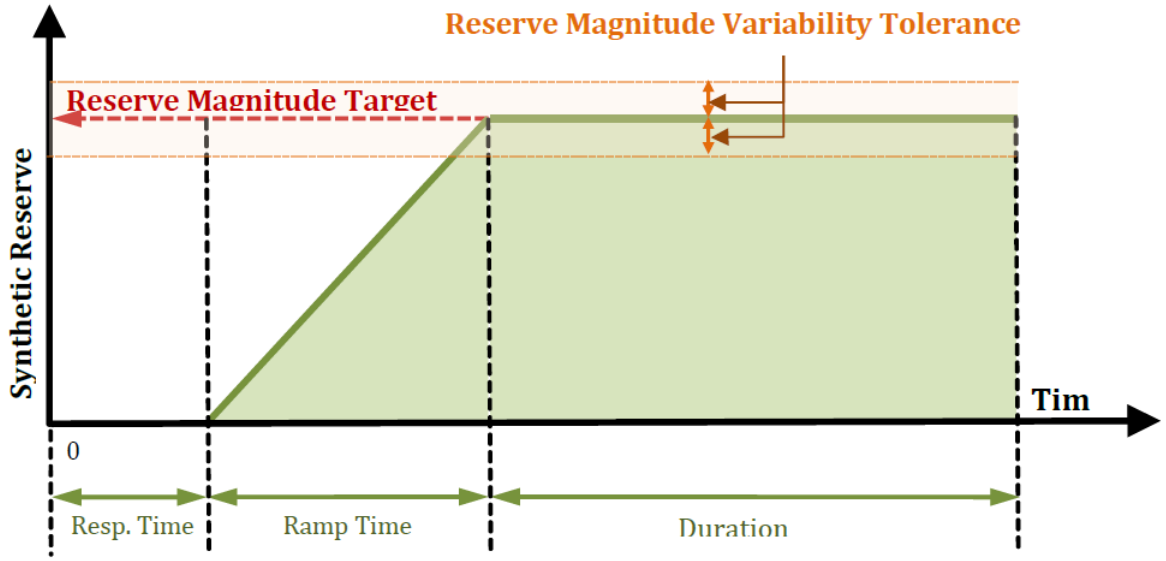


Figure 1. Pictorial description of response time, ramp time, and reserve magnitude variability tolerance (RMVT) [1].

Table 1. Grid Response Requirements [1].

Performance Metric	Frequency Response	Regulating Reserve	Ramping Reserve
Initial Response Time	< 2 seconds	< 5 seconds	< 10 minutes
Reserve Magnitude Target (RMT, % of load)	> 2 %	> 5 %	> 10 %
Reserve Magnitude Variability Tolerance (RMVT)	< +/- 5%	< +/- 5%	< +/- 5%
Ramp Time	< 8 seconds	< 5 Minutes	< 30 minutes
Duration	> 30 seconds	> 30 minutes	> 3 hours
Availability	> 95 %	> 95 %	> 95 %

A brief description of each service and the associated performance criteria is provided in the subsections below.

1.2.1.1 Category I: Frequency Response Reserve

When operating the electric grid the frequency must be maintained within a certain range (60 Hz in North America) to ensure the safe operation of end loads. Achieving this frequency requires that the grid's electrical generation and load be equally matched [4]. An increase in generation (or decrease in load) will result in an increase in grid frequency, while a decrease in generation will result in a decrease in the grid frequency. A sudden loss of electrical generation (for example, from a power-plant unexpectedly going off line) can result in a sudden drop in frequency. If this drop is sufficiently large it can result in other generation sources tripping offline and the cascading failure of the grid. To maintain grid frequency (and prevent systemic grid failure) a certain amount of grid generation is held in reserve to respond to changes in frequency. The amount of required reserve is typically small (several percent) but must be deployed rapidly (within 2 seconds).

1.2.1.2 Category II: Regulation Response

Regulation response serves a similar purpose as frequency response. While the NODES FOA requirements Table 1 allow a slower response the service must be provided for a longer duration. Regulation response is typically provided by dispatchable, quick-ramping generators that can smooth variations within a balancing region between the scheduled power generation and the actual required power generation [1, 5]. Gird operators such as PJM provide a regulation signal (often called an Area Control Error - ACE - signal) that service providers must match. This signal is typically a zero-mean sum signal and compensation is based on how well it is followed.

1.2.1.3 Category III: Ramping Response

Ramping reserve is generation capacity that is bought online in case of loss of generation capacity or unexpected changes in forecast load. The ramping time (< 10 minutes) is less stringent for this service but the delivered magnitude and duration are the largest of all three services.

1.3 Overview of Technical Approach

This project developed a hierarchical control framework consisting of control strategies across multiple time-scales and corresponding to the logical and physical organization of the power system (Figure 2). Each distributed energy resource (DER) that chooses to participate will communicate its ability to provide flexibility and the time scale over which it can provide the service. A distribution reliability coordinator (DRC), which could potentially be seen as a future distribution system operator (DSO), acts as an interface between the DERs and the bulk system, coordinating the resources in an economic and reliable manner. The system coordinator can economically optimize the deployment of the DERs across the different services, taking into account the flexibility of the devices and local system constraints (e.g., capacity). In the case of synthetic frequency response, requiring fast response times, the DERs are armed with active setpoints based on frequency response reserve requirements computed by the DRC, but operate autonomously at faster timescales through self-sensed grid frequency. For regulation services, the devices are armed with active power setpoints based on frequency regulation reserve requirements computed by the DRC, while the control is decentralized, using the area control error (ACE) signal broadcast to the DERs at 4 second intervals. For ramping services, net load will be gradually increased (or decreased) over the interval via setpoints that can be broadcast less frequently (5-15 minutes). This allows the resource to be tailored to current grid conditions, applied in an

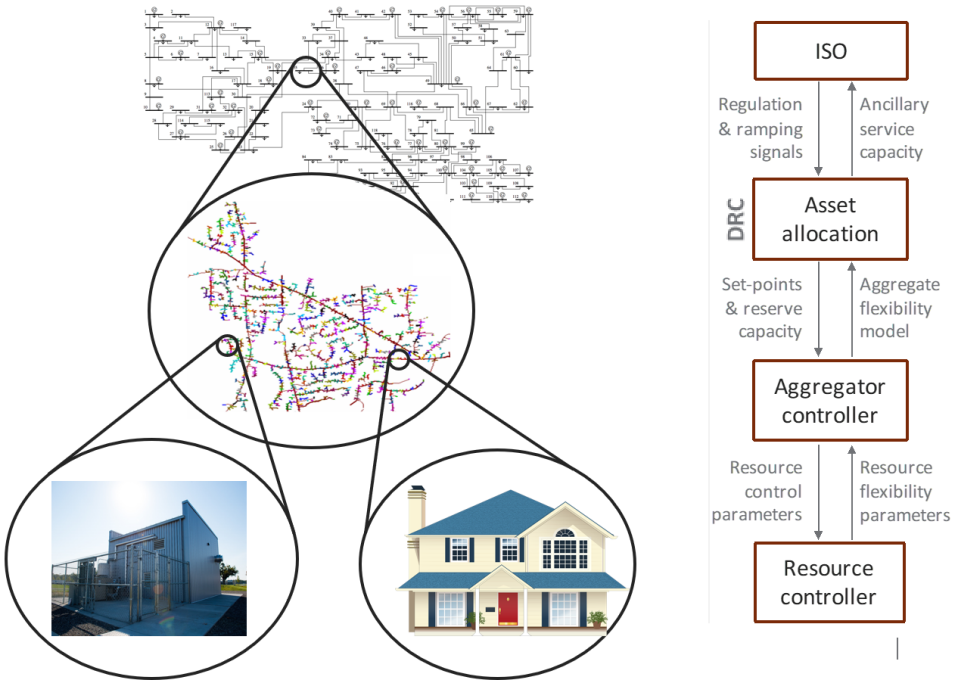


Figure 2. Hierarchical Controls Architecture Overview.

optimal manner at a specific location, and adjusted over time as new DERs are brought online and as required by varying grid operating conditions.

The resulting hierarchical control architecture is shown in Figure 2. The hierarchy corresponds to a logical and physical organization of the power system and consists of 4 layers, each of which is described below:

Independent System Operator (ISO) Layer

At the top of the hierarchy is the ISO (independent system operator). The ISO is responsible for ensuring that the power system functions smoothly, collecting bids from all participants in the transmission level (wholesale market), communicating prices to the market participants and ensuring sufficient reserves for ancillary services. The wholesale market interacts with the traditional generators and distribution reliability coordinators (DRCs).

Distribution Reliability Coordinator (DRC)

The next layer is the distribution reliability coordinators (DRCs). The DRC coordinates all the flexible loads and distributed energy resources on the distribution system and bids the aggregate flexibility into the wholesale market. The DRC also performs resource allocation (deciding how the flexible devices on the distribution system will be dispatched) to meet the requirements for each of the three ancillary services (computed based on the clearing of the wholesale ancillary services market). The DRCs are also responsible for operating the distribution system and any associated retail electricity markets, however this (and the design of incentive mechanisms for engagement of DERs) was outside the scope of this project and warrants further research.

ISO-DRC interface: The distribution system estimates the aggregate flexibility of distributed energy resources (DERs) (flexible loads, distributed solar, distributed storage etc.) and commu-

nicates the same as a bid into the ancillary services (the Category I, II and III services outlined above) markets run by the ISO - this is done once every market period. The ISO takes these bids (and the bids from other market participants), clears the market and comes up with the reserve requirements each DRC must provide for each of the three services. Additional, in real-time, the ISO provides the frequency regulation (Category II) and ramping (Category III) signals to each DRC. For this project the ISO layer is used as boundary conditions for the distribution coordination. That is, the resulting changes in active power and prices at the wholesale level are not modeled and it an area for further research.

Aggregator Controllers

At the third layer are the aggregators - The aggregators control a collection of devices that are geographically co-located and provide a platform to aggregate their flexibility and mitigate uncertainty associated with a single flexible load by managing a large aggregation of similar loads (so that any stochastic effects average out).

DRC-Aggregator interface: The aggregators communicate parameters of an aggregate flexibility model (represented as a virtual battery) to the DRC. The DRC communicates average power setpoint dispatch and reserve margins (for category I and II services) along with participation factors that represent the fraction of the category I and II signals the aggregator must respond to.

Distributed Energy Resource (DER) Control

The fourth layer consists of resource controllers of individual devices (thermostats for air conditioners, chargers for electric vehicles) that determine power consumption of individual devices. The devices we study in this project are *flexible*, ie, they have some flexibility in their power consumption (for example, an AC unit can adjust its power consumption as long as the temperature in the house remains within a certain comfort zone). This flexibility is utilized by the DRC and aggregators to provide grid services by modifying the behavior of the device-level resource controllers.

Aggregator-resource controller interface: The aggregators provide control inputs to the resource controllers of individual devices (thermostats, chargers etc.) to modify the power consumption of devices so as to track the power setpoint (set by the DRC) and provide the category I and II responses required by the DRC. The devices in turn provide information to the aggregators regarding their internal state and other parameters of device level dynamics.

1.3.1 Integrated Federated Co-Simulation Test-Bed

The performance of the resulting system was tested in a co-simulation environment spanning transmission, distribution, ancillary markets, and communication systems. Various classes of actual DERs, (e.g., commercial HVAC systems, battery storage, etc.) and their control systems along with high-fidelity simulations of additional DERs and the electric grid were used to perform hardware-in-the-loop verification of proposed incentive-based control approach.

1.3.1.1 Report Structure

In the following section (Section 2.0) the Distribution Reliability Coordinator (DRC) formulation will be presented. In Section 3.0 the modeling, flexibility estimation and control approach for the

Aggregator Controllers and Residential Distributed Energy Resources (DERs) will be presented along with representative results. Section 4.0 provides details of the development of flexibilty estimators and controls for Commercial Building HVAC DERs. In Section 5.0 the Integrated Federated Co-Simulation test-bed will be described and test-bed demonstration results will be presented in Section 6.0. Finally, the technology commercialization plan is presented in Section 7.0 prior to the conclusion.

2.0 Distribution Reliability Coordinator (DRC) Control and Coordination

The Distribution Reliability Coordinator (DRC)serves as a key interface between the flexible energy resources in the distribution network and the Independent System Operator (ISO) at the transmission level. The primary responsibilities of the DRC are operating the distribution system (with its flexible energy resources) and any associated retail electricity market. It estimates the total available flexibility in the distribution network, based on the flexibility estimates from the aggregator controllers, and bids the flexibility into the wholesale market; and, on market clearance, solves a predictive co-optimization problem to dispatch the flexible assets to meet the requirements of the three ancillary services spanning across a wide range of timescales - seconds (Category I), minutes (Category II), and hours (Category III).

2.1 Flexibility Aggregation

The hierarchical control architecture developed in this work allows the integration of various types of flexible energy resources into the power grid operation. However this requires a uniform way of modeling the flexibility offered by different resources (flexible loads) such that the DRC need not be re-programmed based on the type of resource (e.g. air-conditioners as opposed to electric water-heaters). Such a uniform modeling approach facilitates plug-and-play operation - whereby a resource owner can easily sign up for (or out of) ancillary services as and when desired.

The concept of virtual batteries allows such a uniform modeling of flexibility in various types of energy resources. Certain types of electrical loads are characterized as *energy-driven loads* for which the quality of the end-use service depends on the energy consumption over a duration. Any type of thermostatically-controlled loads (e.g. air-conditioners, electric water-heaters) fall under such categories. For such loads, the thermal energy can be related to the state-of-charge of a physical battery, leading to the conceptualization of the a virtual battery whereby the thermal energy (or the virtual state-of-charge) varies over time based on the power consumption of the load. Moreover, the end-user comfort requirements (quality of service) are modeled as bounds on the virtual energy (state-of-charge). The virtual battery (VB) can be modeled as:

$$x_{t+1} = \lambda_1 x_t + \lambda_2 p_t, \qquad \text{(system dynamics)} \qquad (1a)$$

$$\kappa_t^- \leq x_t \leq \kappa_t^+, \qquad \text{(energy bounds)} \qquad (1b)$$

$$\mu_t^- \leq p_t \leq \mu_t^+, \qquad \text{(power bounds)} \qquad (1c)$$

where

- p_t: additional power consumed over the nominal level (*baseline*) by the VB at time t
- x_t: value of the virtual energy (state-of-charge) state at time t
- λ_1: positive scalar coefficient that determines the self-dissipation of energy
- λ_2: a scalar related to the size of the discrete time-steps
- κ_t^-, κ_t^+: (possibly time-varying) bounds on the energy state
- μ_t^-, μ_t^+: (possibly time-varying) bounds on the power (over nominal)

In the case of flexible loads, the power consumption is treated as being dispatchable (within the bounds specified) and hence a control variable. The objective of the controller is to dispatch the devices such that the end-user quality of service (modeled by the energy bounds) is maintained while providing the ancillary services requested by the system operator. More complex models that include bi-linear terms (product of the energy state and power consumption) may be used if needed (e.g. for commercial HVAC systems).

Identification of VB model parameters for a given collection of DERs is not trivial. Devices typically operate in a nonlinear fashion, with parameters that are mostly unknown (or only partially known). Earlier efforts came up with model-based closed-form expressions for the VB parameters [6, 7], also presented in the examples in Appendix F. Recent and ongoing efforts have been focused on data-driven and optimization-based techniques to identify the *best fit* for the VB model parameters, as found in [8–11].

2.2 Resource Allocation Formulation

The coordinator is responsible for dividing the ancillary service responsibility among the aggregators, as well as for positioning system resources such that they would be able to respond to ancillary service requests in the future over a certain look-ahead period. The ancillary services considered span over three different timescales - Category I or frequency response (seconds), Category II or frequency regulation (minutes), and Category III or ramping (hours). At the start of every allocation period (typically 5-15 minutes), updates are available regarding - 1) the committed (or requested) reserves for the shorter-timescale services (Category I and II); and 2) the requested (target) ramping signal (Category III). The requested amounts could be determined based on a market clearance process that would involve ancillary service bids submitted by the aggregators. However the bidding mechanism and the market clearance process is out of scope for this work. The primary objective of the resource allocation problem is to dispatch the VBs such that their aggregate power profile tracks the target ramping signal (as closely as possible), as well as provision for the faster-timescale reserves - all the while ensuring that the end-user comfort constraints are not being violated.

Next we provide a mathematical description of the main ideas behind the resource allocation problem formulation. Let us use the following notation to denote the ISO requests,

- M_t: the ramping target requested by the ISO (Cat. III),
- R_t: the regulation reserve requested by the ISO (Cat. II),
- E_t: the frequency response reserve requested by the ISO (Cat. I).

The resource allocation problem is designed to share the responsibility among the participating DERs to provide the requested services. As such, we introduce the following variables that denote the power capacity for each group of DERs represented as a virtual battery of each service category:

- m_t^i: the ramping power dispatched to the i-th VB,
- r_t^i: the regulation reserve assigned to the i-th VB,
- e_t^i: the response reserve assigned to the i-th VB.

The ISO requires that the following three conditions are met *as best as possible*

$$\forall t: \quad M_t \approx \sum\nolimits_i m_t^i \qquad \text{(Category III: ramping)} \tag{2a}$$

$$R_t \approx \sum\nolimits_i r_t^i \qquad \text{(Category II: regulation)} \tag{2b}$$

$$E_t \approx \sum\nolimits_i e_t^i \qquad \text{(Category I: response)} \tag{2c}$$

These service requirements are to be met while also ensuring that the end-user comfort constraints are not violated. The bounds on the VB power consumption are modified to accommodate the short-term reserves (Cat. I and II) - i.e. sufficient margin is set aside so that real-time adjustments can be made to provide response and regulation services without violating VB power limits. This is modeled in the following constraints:

$$r_t^i + e_t^i + \mu_t^{i-} \le m_t^i \le \mu_t^{i+} - e_t^i - r_t^i \qquad \text{(VB power bounds)} \tag{3}$$

The VB energy bounds also need to be adjusted. Typically frequency regulation signals are zero-mean, but frequency response events, on the other hand, are uni-directional. As such, in the course of providing frequency response services, the virtual energy state of the VB may increase (or decrease) monotonically and hit the upper (or lower) energy limits. To appropriately account for such event, sufficient margins should also be maintained in the the energy bounds. This is done by adjusting the energy constraints as follows

$$\kappa_t^{i-} + \nu \sum\nolimits_{\tau \le t} e_\tau^i \le x_t^i \le \kappa_t^{i+} - \nu \sum\nolimits_{\tau \le t} e_\tau^i \qquad \text{(VB energy bounds)} \tag{4}$$

where ν is the expected length of a frequency response event (typically 10 s). Finally, the multi-objective problem is constructed that achieves two goals:

1. Performance: meet the service requests as closely as possible, i.e. minimize the errors $\left|M_t - \sum_i m_t^i\right|$, $\left|M_t - \sum_i m_t^i\right|$ and $\left|M_t - \sum_i m_t^i\right|$,
2. Comfort: minimize the end-user discomfort (modeled by $\left|x_t^i\right|$), i.e. operate the VBs as close to their baseline as possible.

As such, the resource allocation is solved via a multi-objective optimization problem as follows:

$$\min \quad w_1 \varepsilon + w_2 \sum\nolimits_{(i,t)} \left|x_t^i\right| \qquad \text{(multi-objective: performance \& comfort)} \tag{5a}$$

$$\text{s.t.} \quad x_{t+1}^i = \lambda_1^i x_t^i + \lambda_2^i m_t^i \qquad \text{(VB dynamics)} \tag{5b}$$

$$\left.\begin{aligned} m_t^i &\le \mu_t^{i+} - e_t^i - r_t^i \\ m_t^i &\ge r_t^i + e_t^i + \mu_t^{i-} \end{aligned}\right\} \qquad \text{(VB power limits)} \tag{5c}$$

$$\left.\begin{aligned} x_t^i &\le \kappa_t^{i+} - \nu \textstyle\sum_{\tau \le t} e_\tau^i \\ x_t^i &\ge \kappa_t^{i-} + \nu \textstyle\sum_{\tau \le t} e_\tau^i \end{aligned}\right\} \qquad \text{(VB energy limits)} \tag{5d}$$

$$\left|E_t - \sum\nolimits_i e_t^i\right| \le \varepsilon \qquad \text{(Cat. I Performance Requirement)} \tag{5e}$$

$$\left|R_t - \sum\nolimits_i r_t^i\right| \le \varepsilon \qquad \text{(Cat. II Performance Requirement)} \tag{5f}$$

$$\left|M_t - \sum\nolimits_i m_t^i\right| \le \varepsilon \qquad \text{(Cat. III Performance Requirement)} \tag{5g}$$

2.3 Testing Grid Service Performance

We present two cases below to illustrate the application of the resource allocation problem. More detailed device specific results will be presented in Sections 3.0 and 4.0, while full integrated test results are presented in Section 6.0. The two cases below utilize electric water heater and air conditioning device models that are described in more detail in Section 3.0 and Appendix F. The simulations were run in MATLAB.

Case 1: We consider five virtual batteries (VBs) with three being groups of electric water-heaters, and two being groups of air-conditioners. Each of the five groups has a population size that is randomly selected between 500 to 1000, and the device parameters are also randomly chosen. VB #2 is assumed to be capable of providing only ramping services, while the other four VBs can provide all three services. The ISO requests are assumed to be:

- Cat. 1 - E_t: $\pm 5\%$ of the aggregated rated power (~ 17.5 MW) of all the devices,
- Cat. 2 - R_t: $\pm 10\%$ of the aggregated rated power of all the devices,
- Cat. 3 - M_t: $+12\%$ of the aggregated rated power of all the devices, with an initial grace period of 30 minutes.

Figure 3 shows how the ramping target is tracked collectively by the 5 VBs over a three hour period. The resource allocation problem is solved every 15 minutes, with updated dispatch for every VB. While the individual ramping dispatch values for each VB changes over time, their sum remains constant throughout the service duration. Figures 4 and 5 illustrate the individual service

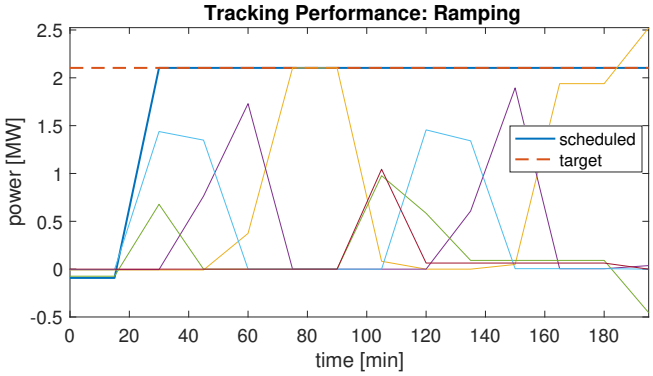


Figure 3. Illustration of a successful dispatch of ramping services across 5 VBs, together achieving the target value of 2.1 MW.

allocations for two of the five batteries - 1) battery #1 which signed up for all three services, and 2) battery #2 which signed up for only ramping services, respectively. Battery #1 is a group of 957 electric water-heaters. The scheduled ramping dispatch, as well as regulation and response reserves are shown on the plots (asymmetric reserve margins are allowed for better flexibility); while the energy reserve margin (for frequency response services), along with the scheduled energy profile and the energy bounds are also shown for battery #1 in Figure 4. Battery #2 is a group of 511 water heaters. Since battery #2 signs up for only ramping services, it does not get assigned any frequency response and regulation reserves. Only the scheduled ramping power and energy profiles are dispatched to battery #2, as shown in Figure 5.

Case 2: Next we consider $> 10,000$ water-heaters and air-conditions split into 20 VBs, each having 500 to 1000 devices. The device parameters are chosen randomly. Frequency response,

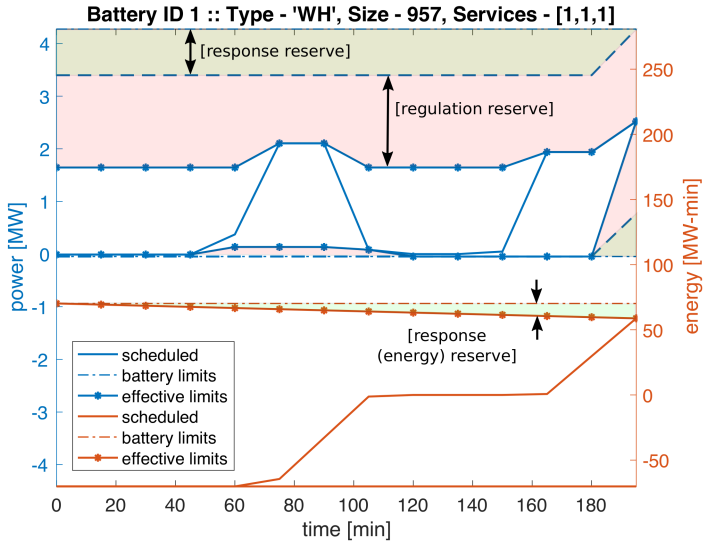


Figure 4. Illustration of the different service allocations on the individual battery #1.

regulation and ramping targets are chosen to be 2%, 5% and 10% of the aggregated rated power of all the devices. In this case, we illustrate the trade-off between the tracking performance and the end-user comfort by varying the relative weights of the two terms in the objective function. Figure 6 shows the comparison of the tracking performance for two values of the relative weights on the end-user comfort in the objective function. Specifically, as a relative weight on the end-user comfort is increased by 10 times (the plot on the right), the tracking performance degrades a little in comparison to the plot on the left. But the tracking error is still maintained within the 5% error bounds.

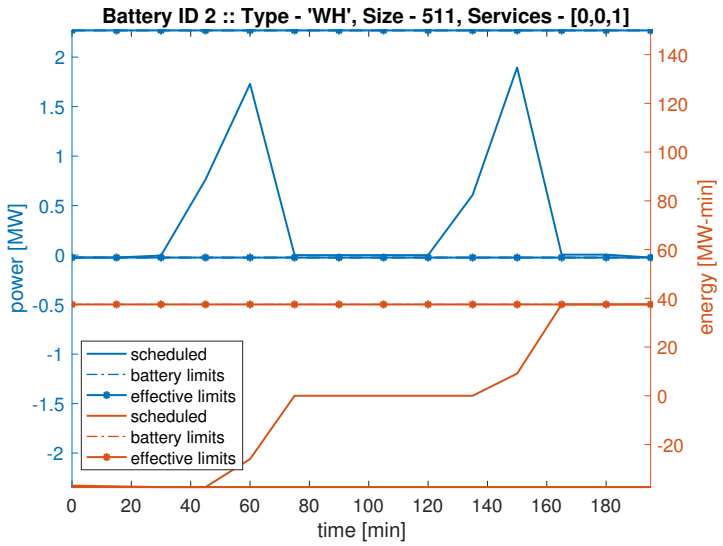


Figure 5. Illustration of the different service allocations on the individual battery #2.

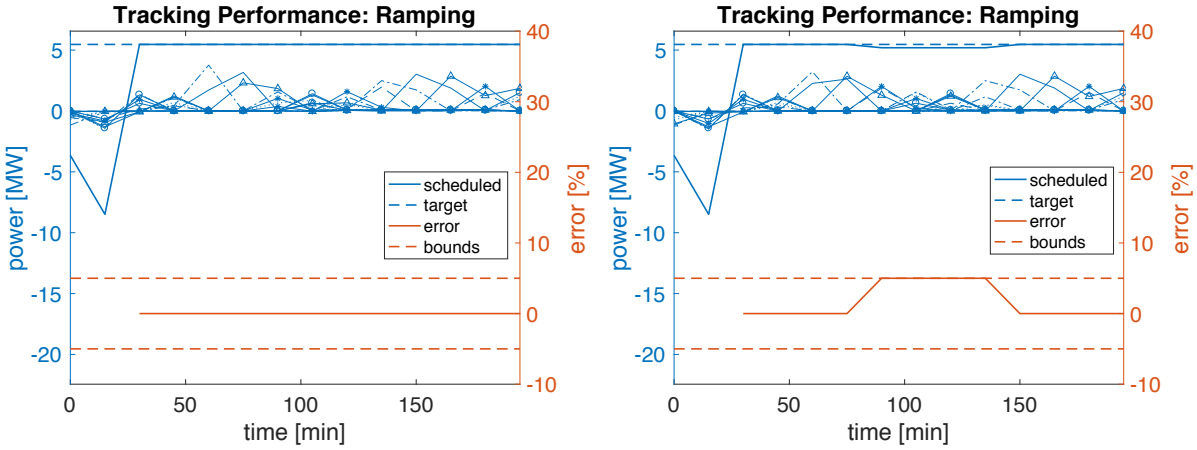


Figure 6. Illustration of the impact on performance when a higher (10x) relative weight is placed on end-user comfort (right) in comparison to the base case (left) - resulting in a slightly higher tracking error.

3.0 Residential DER Control Design

The participation of controllable residential resources (e.g. air-conditioners and electric water-heaters) into the grid ancillary services at the Independent System Operator (ISO) layer is co-ordinated by *aggregators* via the distribution reliability coordinator (DRC). Aggregators serve as the communication and control link between the *resource controller* at the device-level and the DRC. It is the responsibility of the aggregators to characterize the aggregated flexibility of the resources and implement control strategies to coordinate the pool of resources to provide the grid ancillary services (Categories I, II and III) allocated by the DRC, while satisfying end-user quality-of-service (QoS). Typically, the aggregators control a collection of DERs that are geographically co-located and provide a mechanism to mitigate uncertainty associated with a single flexible load by managing a large aggregation of similar loads (so that any stochastic effects average out). The flexibility is represented in the form of a virtual battery (VB), as detailed in Section 2.1, which models the total amount by which the DERs can increase and decrease their power consumption over a given period of time. At the start of each control window (5-15 min), after the DRC communicates the allocated service to the aggregators, each aggregator computes and dispatches control thresholds to individual DERs. Within a single allocation period, the DERs use the dispatched control thresholds to respond to grid signals (e.g. frequency excursions, regulation signals) autonomously, without any intervention from the aggregators.

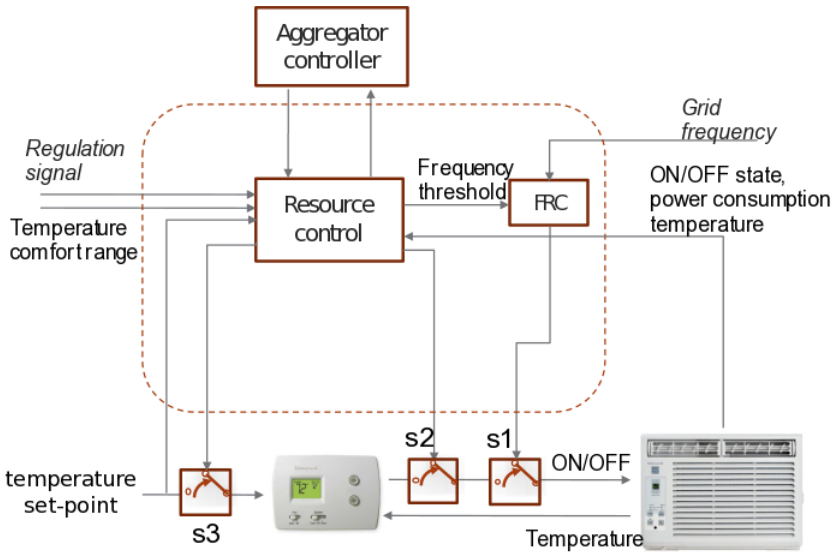


Figure 7. Example of a resource controller integrated with a thermostat-based control loop and equipped with a frequency response controller (FRC).

The *resource controller* acts as the interface between the existing (local) device controller and the load aggregator. It is the responsibility of the resource controller to augment the existing device control to enable grid service response by executing the control thresholds received from the aggregator. Figure 7 briefly depicts the control and communication pathways between an aggregator and a resource controller, using the example of a thermostatic load (e.g. air-conditioner or electric water-heater). In this diagram, the resource controller is augmented with the device-level thermostatic control, which switches the device on/off based on some temperature set-point, while also having a *frequency response controller* module that senses the grid frequency and based on some dispatched frequency threshold (from the aggregator) decides to either switch on

or off the device, only if that forced switching does not violate the local temperature constraints (i.e. the end-user QoS takes precedence over any grid service requests).

The interactions of the aggregators with the DRC and the resource controllers can be summarized as follows:

- DRC-Aggregator interface: Each aggregator chooses a set of ancillary services (Categories I,II,III) to bid into. The aggregators then communicate the flexibility model parameters (in the form of a virtual battery) to the DRC. The DRC communicates the allocated power set-points (for ramping services) and the reserve margins (for frequency response and regulation services) to the aggregators.

- Aggregator-Resource Controller interface: The aggregators provide control inputs to the resource controllers of individual devices (thermostats for air-conditioners and electric water-heaters, chargers for electric vehicles etc.) to modify the power consumption of devices so as to track the power set-point (for ramping) and provide the Category I and II responses, as required by the DRC. The devices in turn provide information to the aggregators regarding their internal state and other parameters of the device-level dynamics.

For a description of the flexibility characterization using a VB model, refer to Section 2.1. In this section, we describe how the aggregator calculates the control thresholds for the DERs to provide the grid services required by the DRC. Before moving on the explaining the aggregator control algorithms, let us briefly describe the dynamical models of the residential thermostatic loads.

Air-conditioner model: The dynamics of an air-conditioning (AC) load is represented by [12]:

$$\dot{T}(t) = -\frac{(T(t) - T_a(t))}{C\,R} - \frac{\eta\, p(t)}{C}\,, \tag{6a}$$

$$p(t^+) = \begin{cases} 0\,, & \text{if } T(t) \le T_{set} - \delta T/2 \\ P\,, & \text{if } T(t) \ge T_{set} + \delta T/2 \\ p(t)\,, & \text{otherwise} \end{cases}\,, \tag{6b}$$

where $T(t)$ is the room temperature; $p(t) \in \{0, P\}$ represent the power draw of the AC; $T_a(t)$ denotes the outside air temperature; and C, R, η are the device parameters representing the house envelopes thermal resistance, thermal capacitance and the load efficiency, respectively. T_{set} is the temperature set-point and δT represents the width of the temperature hysteresis deadband.

Electric water-heater model: The water temperature dynamics of an electric water-heater (EWH) can be modeled using a 'one-mass' thermal model which assumes that the temperature inside the water-tank is spatially uniform (valid when the tank is *nearly* full or *nearly* empty) [13]:

$$\dot{T}_w(t) = -a(t)\,T_w(t) + b(s(t), t)\,, \tag{7}$$

$$\text{where, } a(t) := \frac{1}{C_w}\left(\dot{m}(t)\,C_p + W\right),$$

$$\text{\& } b(s(t), t) := \frac{1}{C_w}\left(s(t)\,P + \dot{m}(t)\,C_p\,T_{in}(t) + W\,T_a(t)\right).$$

T_w denotes the temperature of the water in the tank, and $s(t)$ denotes a switching variable which determines whether the EWH is drawing power ($s(t) = 1$ or 'on') or not ($s(t) = 0$ or 'off'). The

state of the EWH ('on' or 'off') is determined by the switching condition:

$$s(t^+) = \begin{cases} 0\,, & \text{if } T_w(t) \geq T_{set} + \delta T/2 \\ 1\,, & \text{if } T_w(t) \leq T_{set} - \delta T/2 \\ s(t)\,, & \text{otherwise} \end{cases}, \tag{8}$$

where T_{set} is the temperature set-point of the EWH with a deadband width of δT.

3.1 Frequency Response (Category I)

3.1.1 Droop Curve and Frequency Thresholds

Consider an ensemble of N switching loads (e.g. thermostatic loads) which are committed to under-frequency response. As the frequency drops, a fraction of loads are expected to switch off thereby reducing their aggregate power consumption. Typically, any such ensemble of loads will be assigned a specific frequency range over which they need to respond. Thus a typical under-frequency response curve would look like Figure 8, where ω_u and ω_l denote the upper and lower limits of the frequency range assigned to the ensemble, and ω_0 is the nominal frequency (60 Hz). Clearly, $\omega_l < \omega_u \leq \omega_0$ (for under-frequency response). The target frequency response curve is

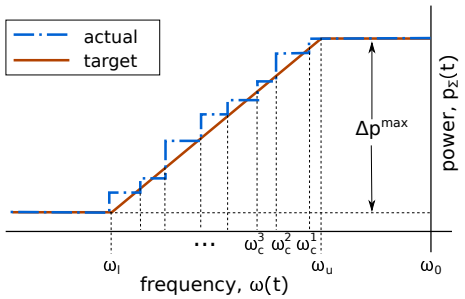


Figure 8. Illustration of a power-frequency response curve.

a smooth line whose slope is determined based on the number (and power consumption) of the devices available to switch their states from 'on' to 'off'. The actual control is implemented by assigning frequency thresholds to each device, such that each device can turn 'off' by monitoring the frequency on its own (see [14–16] for details). An over-frequency response policy can be constructed in a similar way.

Let us identify the DERs in the population with the indices $\{1, 2, \ldots, N\}$. The target response capacity, i.e. the height of the power-frequency response curve in Figure 8 is given by:

$$\Delta p^{\max} := \sum_{i=1}^{N} P^i\,. \tag{9}$$

where P^i denotes the rated power of the DER. For simplicity, let us assume that the corresponding frequency thresholds $\{\omega_c^i(t)\}_{i=1}^N$ are chosen in an ordered way so that:

$$\omega_l \leq \omega_c^N(t) < \cdots < \omega_c^2(t) < \omega_c^1(t) \leq \omega_u. \tag{10}$$

The frequency thresholds are chosen as follows to produce the target response curve in Figure 8:

$$\omega_c^1(t) = \omega_u - \frac{\omega_u - \omega_l}{\Delta p^{\max}} \cdot P^1$$
$$\omega_c^2(t) = \omega_u - \frac{\omega_u - \omega_l}{\Delta p^{\max}} \left(P^1 + P^2\right)$$
$$\vdots$$
$$\omega_c^N(t) = \omega_u - \frac{\omega_u - \omega_l}{\Delta p^{\max}} \left(P^1 + P^2 + \cdots + P^N\right)$$

Each device continuously monitors the grid frequency measurement and compares that to the saved frequency threshold. As soon as the grid frequency falls below the threshold for a particular DER, it switches off provided the temperature limits are not violated (i.e. end-use QoS takes precedence).

The key point to note here is that the values of the frequency thresholds in Equation (10) depend on the availability of the devices to turn 'off' during an under-frequency event (equivalently, turn 'on' for over-frequency event). However, continuous monitoring of the device states in an ensemble has high telemetry requirements, along with potential privacy concerns (for the device owners). A more viable option is to acquire and update the device states information once (at the start of) every fixed control time window, while using that information to estimate the availability of the responsive devices during the control window. This brings us to the problem of quantifying the *fitness* of each of device to provide a certain kind of service (frequency response in this case). The fitness values can be used for the prioritized assignment of the frequency thresholds. For example, the devices that have a high *fitness* value (defined later) will be assigned the thresholds that are closer to the *nominal* frequency, so that when the frequency starts to deviate the 'fittest' devices are called for service first, while the thresholds that are farther away are assigned to the devices with lower fitness values. We propose a metric to quantify the 'fitness' of each controllable device to provide autonomous frequency response.

3.1.2 Fitness and Prioritized Threshold Allocation

Definition 1. *'Fitness' of a device-i for a particular service request is denoted by a scalar $\pi^i \in [0,1]$ which quantifies how likely the device is to successfully complete the service request over some time window.*

Fitness is a qualitative measure of the ability of a controllable load to (successfully) respond to a certain kind of ancillary service request. Evaluation of this metric of *fitness* could depend on several factors, such as the device dynamics, response delays, rate of failure (to respond) and the type of service request. In its simplest form, the fitness measure could be composed of two component metrics - 1) *availability* for response, and 2) *quality* of response.

Definition 2. *'Availability' metric of a device-i for a particular service request, denoted by $\pi^{avail,i} \in [0,1]$, quantifies the probability that the device is available to respond to the particular service request over some time window.*

Definition 3. *'Quality' metric of a device-i for a particular service request, denoted by $\pi^{qual,i} \in [0,1]$, quantifies the probability that the device, when available, completes the service request successfully.*

Consider for example, a plug-in electric vehicle that had signed up for a frequency response service at $t = t_0$ over a time window $[t_0\,,\, t_f]$. Suddenly, at some time $t = t_0 + \Delta t < t_f$ the vehicle

was taken off the charger and driven away. If an event happens any time between $t_0 + \Delta t$ and t_f, the vehicle would be unavailable to respond to it, even though it signed up for it. On the other hand, consider a residential air-conditioner that signs up for frequency response service over some control window. When the event happens, it is ready to respond to it. However, due to delays in control it takes a while to actually switch its mode of operation, thereby delivering a poor quality of service even though it was available to respond to it. The *fitness* metric can be considered as a product of the *availability* and *quality* metrics as follows:

$$\pi^i = \pi^{\text{avail},i} \cdot \pi^{\text{qual},i}. \tag{11}$$

The success of a device in completing a service request (and meeting all the performance metrics) depends on many factors, such as sensors and actuation time-delays, as well as sensors and actuators failures. It is expected that during real-life implementations, the performance of a device to a particular service request could be monitored over time to estimate the *quality* metric, $\pi^{\text{qual},i}$. As an example, in a simple form, the performance degradation due to time-delays can be modeled into the *quality* metric as,

$$\pi^{\text{qual},i} = \exp\left(-\beta\, t_d^i\right) \tag{12}$$

where $\beta > 0$ is an appropriate scaling factor and $t_d^i \geq 0$ is an estimated (possibly over previous such requests) time-delay of the device in responding to the particular service request. A total failure would be captured by the limiting case $t_d^i \to +\infty$ while $t_d^i = 0$ would refer to a success rate of 1. Both the availability and quality metrics can be estimated and updated online by monitoring the device performance in response to similar requests. The fitness values can be maintained online at the aggregator level, and updated at the start of every control window and used in frequency threshold allocation. In the rest of this section, we illustrate how, in the presence of sufficient device-level information, the *availability* metrics can be computed for under- and over-frequency response.

When an under-frequency event happens, the devices that have previously committed for under-frequency response service are requested to turn 'off' one-by-one based on their frequency thresholds. Thus the availability of a device to respond to an under-frequency response can be quantified by the probability that the device is in the 'on' state when the event happens and has meet the device's minimum on-time. It can be argued that the conditional probability distribution of the time of occurrence of a frequency event given the event has occurred within some interval $[t_0, t_f]$, is uniform over the time interval. Then the *availability* metric of a device with respect to an under-frequency response (denoted by the subscript 'resp-') can be obtained as:

$$\begin{aligned}
\pi_{\text{resp-}}^{\text{avail},i} &= \mathbf{Pr}\{\text{device-}i \text{ is 'on' when under-frequency happens}\} \\
&= \int_{t_0}^{t_F} \mathbf{Pr}\{\text{device-}i \text{ is 'on' at time } \tau\} \cdot f_{\text{resp-}}(\tau)\, d\tau \\
&= \int_{t_0}^{t_f} \frac{s^i(\tau)}{t_f - t_0}\, d\tau = \frac{t_{on}^i}{t_f - t_0}
\end{aligned} \tag{13}$$

where, $s^i(\cdot) \in \{0, 1\}$ represents the operational state of the device, taking value 0 in the 'off' state and 1 in the 'on' state; t_{on}^i is the length of time the device spends in the 'on' state during the control window $[t_0, t_f]$; and $f_{\text{resp-}}(\cdot)$ is the uniform conditional probability density function of the time of occurrence of the under-frequency event, given that the event occurs during the interval. Of course, it is not possible to exactly know the 'on' time length t_{on}^i for each device. But with the knowledge of the internal states of the devices, and some forecast of the external conditions, it is possible to estimate $\pi_{\text{resp-}}^{\text{avail},i}$ for each device at the start of each control period.

Using similar arguments, the availability factor for an over-frequency response (denoted by the subscript 'resp+') over an control window $[t_0, t_f]$ could be calculated as:

$$\pi_{\text{resp+}}^{\text{avail},i} = \frac{t_{off}^i}{t_f - t_0} \tag{14}$$

where t_{off}^i is the length of time the device spends in the 'off' state during the control window $[t_0, t_f]$. Note that, under perfect information, the availability metric can be computed exactly using the 'on' time and 'off' time duration. In more realistic scenarios the equations above can be used to estimate certain statistical properties (such as mean, variance) of the random variables t_{on} and t_{off} from any statistical information on the sensor errors and device parameters.

At the start of the control window ($t = t_0$), the *fitness* values of each device in the population are computed (for frequency response). Based on the fitness values, π^i, all the devices in the population are prioritized in an order $\{d_1, d_2, d_3, \ldots, d_N\}$ for consideration of commitment to frequency response, such that:

$$\pi^{d_1} \geq \pi^{d_2} \geq \pi^{d_3} \ldots \geq \pi^{d_N}\,. \tag{15}$$

Accordingly the frequency thresholds are assigned, such that thresholds closer to the nominal frequency are assigned to higher priority ('fitter') devices. Finally, a subset of those devices are selected, based on the priority order, such that their aggregate power rating equals (within some tolerable error, ϵ_p) the target response capacity Δp^{max}, i.e. choose the smallest $m \in \{1, 2, \ldots, N\}$ such that:

$$\left| \Delta p^{\text{max}} - \sum_{i=1}^{m} P^{d_i} \right| \leq \epsilon_p\,. \tag{16}$$

Short Cycling: Short cycling refers to the action of switching a device from ON-to-OFF (or, OFF-to-ON) before switching it back ON (or, OFF) within a very short period of time. Most of the thermostatic loads (air-conditioners, electric water-heaters) have some sort of logic encoded in the device controller that prevents the short cycling. For example, it was reported in [17, 18] that short cycling of the compressor may lead to compressor failure and reduced efficiency. A compressor time delay relay is typically installed to ensure that the compressor spends a certain minimum amount of time ("lockout" time) in the OFF state. Such a lockout effect can be modeled in the calculation of the *availability* metric. During this period, a switching-ON control signal is ignored by the device controller, i.e. a compressor is not considered *available* to execute a switch-ON command until it has already spent the lockout time in the OFF state. Modifying (14) the *availability* metric for the over-frequency response service is adjusted as follows:

- **Case I: The device is OFF and locked-out at $t = t_0$.** Let us denote by δ_{lock} the remaining length of the lockout period. In such a case we have:

$$\pi_{\text{resp+}}^{\text{avail},i} = \frac{\max\left(0, t_{off}^i - \tau_{lock}\right)}{(t_f - t_0)} \tag{17}$$

- **Case II: The device is OFF and past the lockout period at $t = t_0$.** In such a case we simply have:

$$\pi_{\text{resp+}}^{\text{avail},i} = \frac{t_{off}^i}{(t_f - t_0)} \tag{18}$$

- **Case III: The device is ON at $t = t_0$.** Let us denote by T_{lock} the total length of the lockout period. In such a case we have:

$$\pi^{\text{avail},i}_{\text{resp+}} = \frac{\max\left(0, t^i_{off} - T_{lock}\right)}{(t_f - t_0)} \tag{19}$$

Similar adjustments can be made to model ON-state lockout periods, too, when applicable.

3.1.3 Performance

Control performance is evaluated against a metric termed as the *reserve margin variability target* (RMVT) which is expressed as the following:

$$RMVT := \left|1 - \frac{\text{total response provided on request}}{\text{total response requested}}\right|.$$

In order to test the prioritized decentralized frequency response algorithm, a random collection of 1000 ACs and 1000 EWHs was generated[1]. The frequency response range was chosen as:

- 59.7 Hz and 59.995 Hz for under-frequency response, and
- 60.005 Hz and 60.3 Hz for over-frequency response.

Frequency events were created in an IEEE 39-bus network, by injecting disruptions via changing the loads. Figure 9 illustrates two such events. Performance of the frequency response control algorithm was tested against these various frequency events (shifted in time as appropriate).

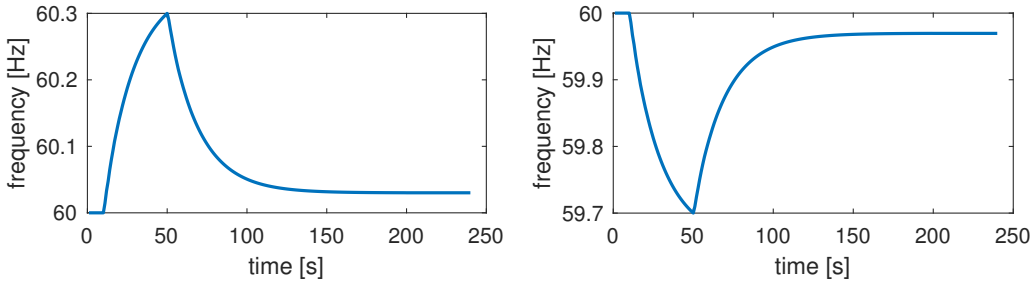
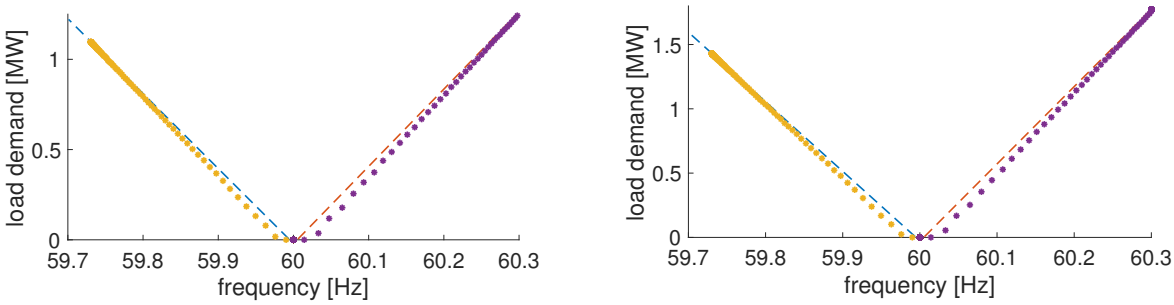


Figure 9. Sample under- and over-frequency events generated in IEEE 39-bus.

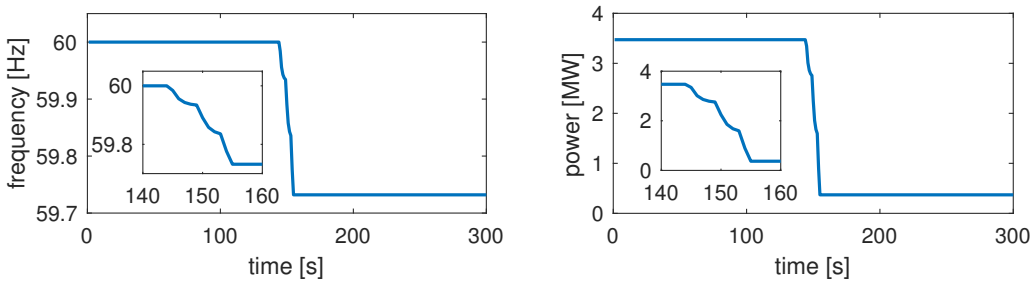
Figure 10 illustrates how the target and achieved frequency response curves typically look like, simulated using 1000 EWHs (Figure 10(a)) and 1000 ACs (Figure 10(b)). The results are obtained for a 5 min control window, with a sampling time of $\Delta t = 1$ s. The achieved response is fairly close to the target response, with some errors being observed for lower frequency deviations. This key observation can be explained by looking at the frequency events in Figure 9. The rate at which the frequency changes is high when the frequency deviation is low (i.e. closer to the nominal value of 60 Hz) and gradually decreases (due to damping effect by the generators) as the frequency deviates further. This suggests that the sampling time should be chosen sufficiently small for faster frequency events.

[1] Parameters for the EWHs were taken from [16], while the parameter values for the ACs are drawn randomly from the following range of values: $P \in [5.5, 6.5]$ kW, $R \in [2, 2.4]\,^o$F/kW, $C \in [3.24, 3.96]$ kW-hr/^{o}F, $T_{set} \in [70, 74]^o$F, $T_a \in [80, 95]^o$F and $\eta = 2.5$,

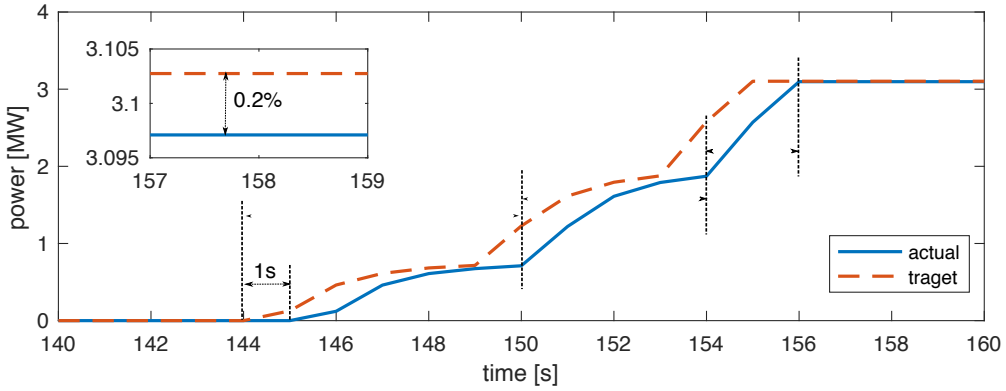


(a) Response of 1000 ACs (b) Response of 1000 EWHs

Figure 10. Target (dashed) and achieved (dotted) frequency response curves



(a) Cascading event (left) and response (right).



(b) Performance metric evaluation.

Figure 11. Performance under a cascading contingency.

Figure 11 shows the response of the group of 1000 ACs and 1000 EWHs in response to a cascading contingency where an initial contingency (created by load drop) leads to two subsequent under-frequency events. Control was chosen as 5 minutes. The net response magnitude achieved is 3.097 MW, which is less than 0.2 % of the target 3.103 MW, i.e. RMVT$\leq 0.2\%$. Note that, the achieved response is always able to track the target response curve within a 1 second delay.

Finally, Monte-Carlo simulations were run to test the algorithm under various different scenarios. With a fixed (but randomly generated) population of 1000 EWHs and 1000 ACs, different scenarios were created by changing the initial operating condition, the length of the control window as well as the time when the frequency event happens. In Table 2 the mean RMVT values for 6 different scenarios are presented. Across all these scenarios, the priority-based algorithm performs uniformly well with the mean RMVT < 0.3 %. However, when not using the priority-based algorithm, the RMVT increases and is sensitive to the length of the control window and when the time of occurrence of the event. Table 4 lists a brief summary of the performance metrics evaluated through the simulation tests. All the metrics specified in the FOA are met satisfactorily.

Table 2. Mean RMVT [%] with priority-based allocation

Control Window \ Event Time	'start'	'middle'	'end'
5 min	0.2078	0.2020	0.2021
15 min	0.2437	0.2602	0.2637

Table 3. Mean RMVT [%] without priority-based allocation

Control Window \ Event Time	'start'	'middle'	'end'
5 min	1.5747	1.4427	1.3613
15 min	1.0400	1.3960	5.7700

3.2 Frequency Regulation (Category II)

3.2.1 Control Design Principle

The objective is to modulate the aggregated power of a population of end-use loads to follow a given power reference signal (for example, a regulation signal broadcast by an ISO or RTO). Generally the regulation signals are meant to balance the small real-time imbalances in generation and load, and are usually calculated as an *area control error* (ACE) by a balancing authority. As such, these signals are typically zero-mean oscillatory signals, and are broadcast every few seconds (e.g. every 4 s for PJM). Note that such a signal, $r(t)$, can always be decomposed in

Table 4. CATEGORY I: Performance Evaluation

Metric	Target	Target Met?	Achieved Performance
Initial Response Time	<2 s	YES	1 s
Reserve Magnitude Target (RMT)	>2 %	YES	>29 %
RMVT	$< \pm 5$ %	YES	<0.5 %
Ramp Time	<8 s	YES	< 7 s
Duration	>30 s	YES	140 s
Availability	>95 %	YES	>95 %
Cascaded Contingency Support	>2 events	YES	3 cascading events

the form of:

$$r(t) = r^{+}(t) + r^{-}(t)\,, \tag{20}$$

where $r^{+}(t)$, referred to here as the *'rising signal'*, is a monotonically increasing component of the signal, while $r^{-}(t)$, referred to here as the *'falling signal'*, is a monotonically decreasing component of the signal. These signals can be constructed from the modulation signal $r(t)$ as:

$$r^{+}(0) = \max\left(r(0)\,,\,0\right) \tag{21a}$$

$$\forall t > 0: \quad r^{+}(t) = \begin{cases} r^{+}(t-) & \text{if } \dot{r}(t) \leq 0 \\ r(t) & \text{otherwise} \end{cases}\,, \tag{21b}$$

$$\text{and} \quad r^{-}(0) = \min\left(r(0)\,,\,0\right) \tag{21c}$$

$$\forall t > 0: \quad r^{-}(t) = \begin{cases} r^{-}(t-) & \text{if } \dot{r}(t) \geq 0 \\ r(t) & \text{otherwise} \end{cases} \tag{21d}$$

As an example, Figure 12 illustrates how a power reference signal can be decomposed into its rising and falling components. Consider an ensemble of switching devices, that have committed to provide a frequency regulation service, by turning ON and OFF as required. The tracking reference (regulation) signal is decomposed into the rising and falling components and fed as two separate tracking signals to the ensemble. Those end-use loads that are currently OFF can be used to deliver $r^{+}(t)$ by turning ON, and those that are currently ON can be used to deliver $r^{-}(t)$ by turning OFF, such that the population of devices as a whole tracks $r(t)$. In the following, a demand-side control with hierarchical decision-making is described for end-use loads to follow the given power reference signal through both slow time-scale coordination (at the aggregator level) and fast time-scale control (at the resource controller level).

At the supervisory-layer, individual end-use loads are managed by the same load aggregator. During real-time operation, the load aggregator needs to receive the power reference signal $r(t)$ from distribution reliability coordinator, determine the rising signal $r^{+}(t)$ and the falling signal $r^{-}(t)$ by decomposing $r(t)$, and broadcast them to individual end-use loads under its authority. On the other hand, the load aggregator has to ensure that the aggregated load response under autonomous threshold-based control can follow the desired power reference signal $r(t)$. This is achieved by coordinating the selection of local thresholds once every allocation (coordination) period, in a method akin to the threshold selection approach adopted for distributed frequency

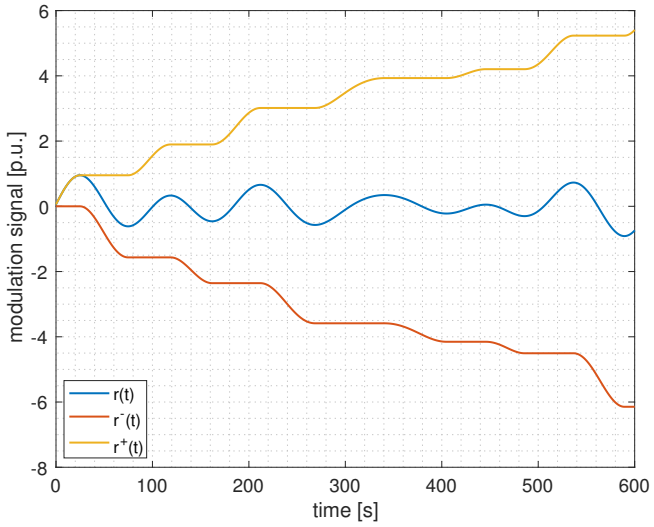


Figure 12. Illustration of decomposing a power reference signal into rising and falling components.

response in [15,16]. The length of each allocation period should be a design parameter depending on the characteristics of end-use loads under consideration, usually chosen to be somewhere between 5 and 15 minutes.

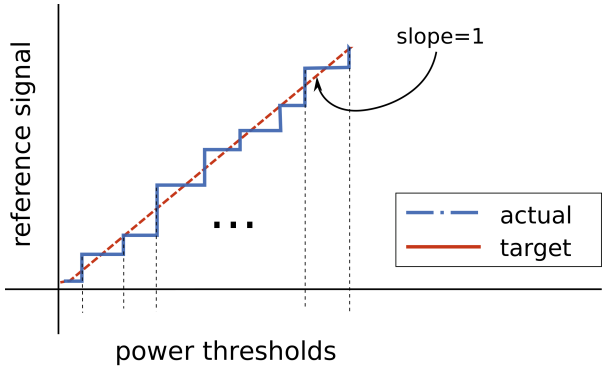


Figure 13. Illustration of selection of power threshold values for reference tracking.

A summary of the threshold allocation for reference tracking is shown in Figure 13. The essential idea is that, as the value of the rising component of the reference signal increases over time, loads are switched ON sequentially, such that the increase in load due to forced switching equals (within tolerable error) the value of the rising signal (note the unity slope of the target response curve in Figure 13). The same concept applies to falling signals as well, in which case the drop in demand due to forced switching (OFF) equals to the absolute value of the falling signal.

At the resource control layer, individual end-use loads are equipped with a threshold-based controller as shown in Figure 7, which operates on top of the original local controllers. When the device is not signed up for any service it simply follows the natural switching logic. As soon as the device is signed up for a service, the resource controller gets activated. If the device gets picked for a up-regulation service, the resource controller monitors the broadcast rising signal

$r^+(t)$ and as soon as the value of the rising signal increases beyond the preset threshold for the device, the device is turned ON. However, it is to be noted here that the end-use preference always take precedence over supervisory control command, i.e. if the local constraints dictate the opposite the device will ignore the supervisory control instruction. Similarly, devices that are signed up for down-regulation are also controlled by the resource controller in a decentralized manner.

Note that the values of the power thresholds depend on the availability of the devices to turn ON or OFF based on reference signal values. However, continuous monitoring of the device states in an ensemble has high telemetry requirements, along with potential privacy concerns (for the device owners). A more viable option is to acquire and update the device states information once (at the start of) every fixed control time window, while using that information to estimate the availability of the responsive devices during the control window. This brings us to the problem of assigning a *fitness* value to each of the devices, quantifying how suitable the device is to provide a certain kind of service. The fitness values can be used for selection and ordering of devices that should be supplied power thresholds for reference tracking. For example, the devices that have high *fitness* value (defined later) will be assigned the thresholds that are closer to zero while the thresholds that are farther away are assigned to the devices with lower fitness values (so that they will be called for only in the extreme cases).

3.2.2 Fitness and Prioritized Threshold Allocation

Fitness metrics are defined for the frequency regulation service in a similar fashion as in Section 3.1.2. Once the fitness values of each device is computed (for regulation service), the order $\{d_1, d_2, \ldots, d_N\}$ in which the devices are prioritized for the service commitment is decided based on their fitness, such that:

$$\pi^{d_1} \geq \pi^{d_2} \geq \cdots \geq \pi^{d_N} . \tag{22}$$

Accordingly the frequency thresholds (for primary frequency response) are assigned, such that thresholds closer to zero are assigned to higher priority ('fitter') devices. Finally, a subset of those devices are selected, based on the priority order, such that their aggregate power rating equals (within tolerable error, ϵ_p) the target response capacity $r^{\max}$, i.e.:

$$r^{\max} \leq P^{d_1} + P^{d_2} + \cdots + P^{d_m} \leq r^{\max} + \epsilon_p , \text{ where } m \leq N . \tag{23}$$

3.2.3 Performance

In order to test the prioritized decentralized frequency regulation algorithm, a random collection of 1200 ACs and 1200 EWHs was generated. Regulations signals from PJM was used to test the algorithm. Starting from a random initial conditions, the ensemble was simulated to estimate the peak power consumption over a 30 minute window. Then the PJM regulation signal (normalized between -1 and $+1$) was scaled appropriately to fluctuate between $> \pm 5\%$ of the peak power value. The signal decomposition technique described above was adopted to generate a rising and a falling signal from the regulation signal. Using a 5 minute allocation period, the devices from each group (of ACs and WHs) were selected to track the *rising* and *falling* signals, for up- and down-regulation, respectively. We used 1 second as the simulation and controller time-step.

Figure 14 shows the total power consumption of the population of devices (1200 ACs and 1200 WHs), tracking a regulation signal from PJM, which is scaled to $\pm 10\%$ of the peak power value (satisfying the NODES FOA metric in Table 1). Moreover, looking at the tracking of the

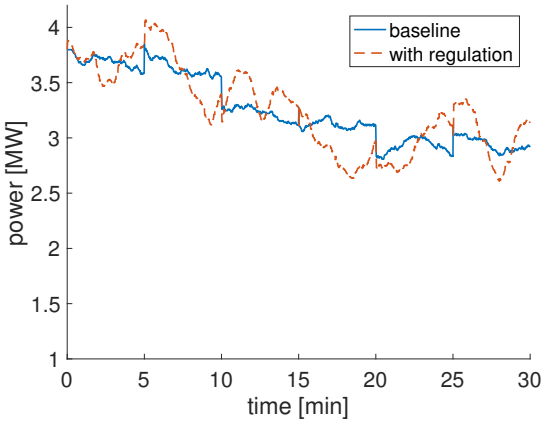


Figure 14. Ensemble of 1200 ACs and 1200 WHs are coordinated to track regulation signal at $\pm 10\%$ of peak power, shown on top of their baseline power consumption.

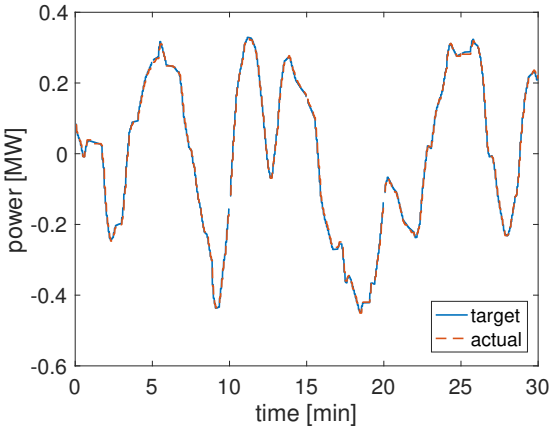


Figure 15. Close tracking of PJM regulation signal, at $\pm 10\%$ of peak power, by an ensemble of 1200 ACs and 1200 WHs over a 30 minutes period.

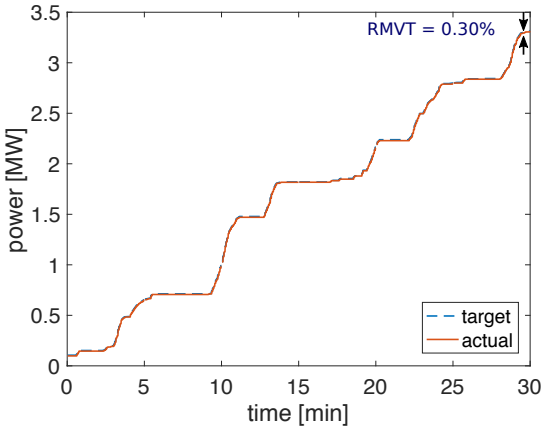


Figure 16. Close tracking of rising regulation signal component over a 30 minutes period, with a total reserve magnitude very close to the peak power.

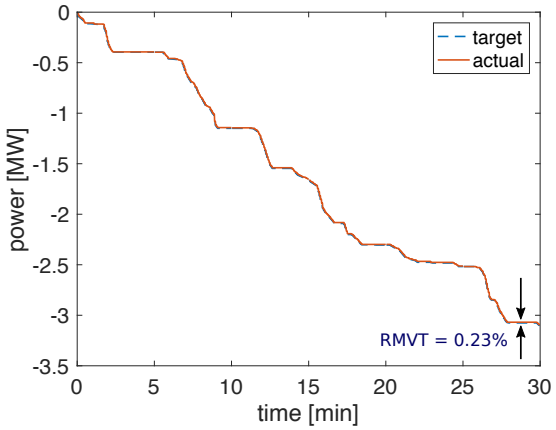


Figure 17. Close tracking of falling regulation signal component over a 30 minutes period, with a total reserve magnitude very close to the peak power.

regulation signal (about the baseline power) in Figure 15, we observe that the initial response time and the ramp time are both very small. In fact those are less than the simulation/control time-step, which is 1 second. Because the devices are being controlled via switching ON/OFF, the response and ramp times are comfortably within the FOA-specified target values. In Figures 16 and 17, we show the tracking performance with respect to the decomposed rising and falling reference signals. Specifically, we highlight how the RMVT metric is met within the target value. Finally, our application of the fitness-values based prioritized device selection approach allows the aggregator to find the best suited devices for regulation services every allocation period, thereby ensuring >95% availability of resource for regulation over a 30 minute window.

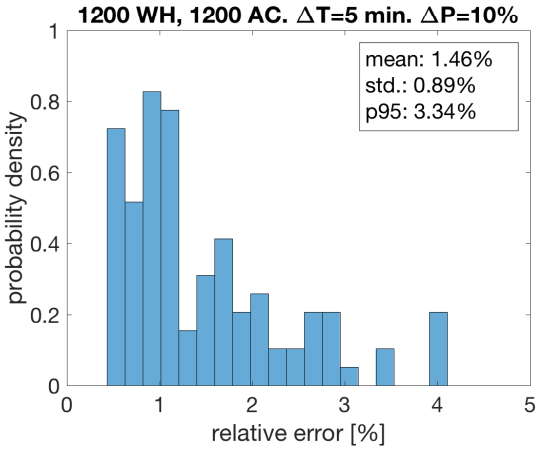


Figure 18. RMVT statistics from a Monte-Carlo run of 50 instances with a population of 1200 WHs and 1200 ACs, under varying initial conditions.

Next we evaluate the performance of the regulation algorithm under different populations, operating conditions, duration of allocation period, and level of service commitment (RMT). Figure 18 shows the plot of RMVT statistics obtained from a Monte-Carlo run of 50 instances of regulation tracking by a (fixed) population of 1200 ACs and 1200 WHs, under varying initial conditions. Allocation period is chosen as 5 min and the regulation signal is scaled to $\pm 10\%$ of the peak power. Each instance gives two RMVT values corresponding to the *rising* and *falling* regulation

profiles. Thus a total 100 RMVT values are obtained from 50 instances. The mean, standard deviation and the 95-percentile value of the RMVT values are found to be 1.5%, 0.9% and 3.3%. Then the similar Monté-Carlo studies are performed for different allocation period, different RMT values (as a percentage of peak power), as well as for a different population of water-heaters and air-conditioners. The results are summarized below in Table 5. It is seen that, in general, the performance improves with shorter allocation period and degrades with larger RMT values. Finally, Table 6 lists a summary of the performance metrics evaluated through the simulation tests. All the metric specified in the FOA are successfully met.

Table 5. Statistical Analysis of RMVT under Various Scenarios

Population	Allocation Period	RMT [%]	RMVT [%]		
			mean	standard deviation	95-percentile
1200 WH 1200 AC	5 min	10	1.46	0.89	3.34
		20	1.66	2.18	6.35
	10 min	10	1.57	2.05	5.89
1200 WH 800 AC	5 min	10	1.80	1.02	3.23
	10 min	10	2.40	3.29	9.89

Table 6. CATEGORY 2 Performance Evaluation

Metric	Target	Target Met?	Achieved Performance
Initial Response Time	$<$5 s	YES	1 s
Reserve Magnitude Target (RMT)	$>$5 %	YES	$>$10 %
RMVT	$< \pm$5 %	YES	$<$0.5 %
Ramp Time	$<$5 min	YES	< 1 s
Duration	$>$30 min	YES	30 min
Availability	$>$95 %	YES	$>$95 %

3.3 Ramping (Category III)

Ramping service is to be provided over a 3 hours period. There have been very few results in the literature focusing on load control design for synthetic ramping reserve. Note that, the delivery of ramping reserve can be viewed as a power reference tracking problem similar to the delivery of regulation reserve except that the service duration is much longer. Therefore, the control approach presented above for delivering regulation reserve can be used as well for delivering ramping reserve. However, it can only deliver a small amount of ramping reserve for a very short period before the flexibility of the population is exhausted due to the imposed temperature deadband around temperature setpoints. Hence, in this project, we propose to control the temperature setpoints of thermostatically controlled loads such as air conditioners and water heaters in order to deliver a synthetic ramping reserve of large magnitude.

3.3.1 Control Design

The proposed ramping control has the structure of hierarchical decentralized control. It consists of two layers, a device layer and a supervisory layer, where the device layer consists of the thermostatically controlled loads (TCLs). At the beginning of every coordination cycle (for example, five minutes), individual TCLs submit to the coordinator their demand flexibility by taking into account local constraints and objectives. The coordinator in the supervisory layer collects all the demand flexibility and then determines the coordination signal, which is referred to as the ramping index, to achieve the desired aggregated power by coordinating this group of TCLs. This ramping index is then broadcast back to the device layer as the coordination signal. After receiving the ramping index, individual TCLs determine local control inputs, i.e., temperature setpoints, independently for the current coordination cycle. In the following, both device-level control and supervisory-level coordination are described in detail by using a residential air conditioner (AC) as an illustrative example.

The indoor air temperature setpoint (T_{set}) of each AC is used as the local control inputs. It is updated inside the thermostat every coordination period in response to the ramping index received from the coordinator. The ramping index actually describes the need of delivering ramping reserve for the coming coordination period. When it is positive, it implies that there is a need of decreasing the aggregated power consumption of TCLs, and vice versa. The magnitude of ramping index denotes the extent of such a need. The mapping from the ramping index (λ^*) to the new temperature setpoint (T_{set}) is specified by the local control response curve as shown in Figure 19, which is determined by several parameters.

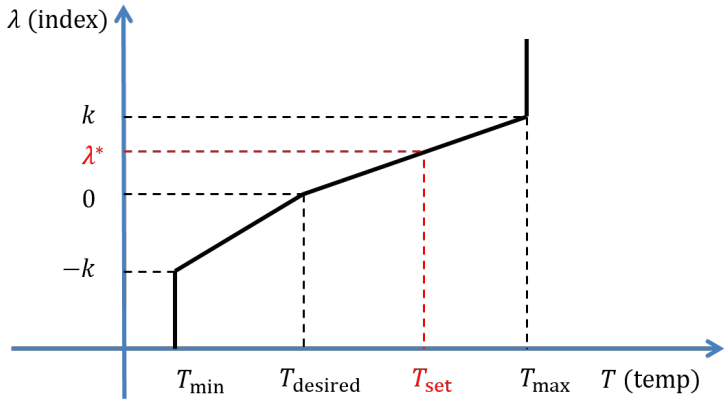


Figure 19. Illustration of local control response curve of a residential AC.

The parameters T_{desired}, T_{min}, and T_{max} are directly specified by users, where T_{desired} is the desired indoor air temperature setpoint, and T_{min} and T_{max} are the lower and upper bounds of the acceptable indoor air temperature setpoint. The parameter $k > 0$ is abstracted from the flexibility of the household owner in providing ramping reserve, which can be either time-invariant or time-varying. When k is very large, the response curve becomes an almost vertical line at T_{desired}. This implies that the household owner is very sensitive to the indoor air temperature, and would like to maintain the indoor air temperature setpoint at T_{desired} unless there is a huge need of ramping reserve. When k very small and close to zero, the response curve becomes an almost horizontal line at 0. This implies that the household owner is very flexible, and is willing to sacrifice indoor comfort to provide ramping reserve whenever there is a need.

In fact, for any given indoor air temperature setpoint, the energy consumption for the coming coordination cycle can be calculated based on the dynamical model of the AC and the current

indoor air temperature. On the other hand, for any given ramping index, the temperature setpoint can be determined from the local control response curve as shown in Figure 19. By considering these two relationships in a composite way, the relationship between the ramping index and the energy consumption can be obtained as shown in Figure 20. This relationship describes the demand flexibility of the household owner for the coming coordination cycle. At the beginning of each coordination period, individual TCLs need to submit their demand flexibility to the coordinator. After receiving all the information, the coordination can determine the aggregated demand flexibility. Then determine the desired ramping index λ^* by finding the intersection between the aggregated demand curve and the desired ramping power as shown in Figure 21. This desired ramping index will be broadcast back to individual TCLs as a coordination signal for them to adjust local temperature setpoints.

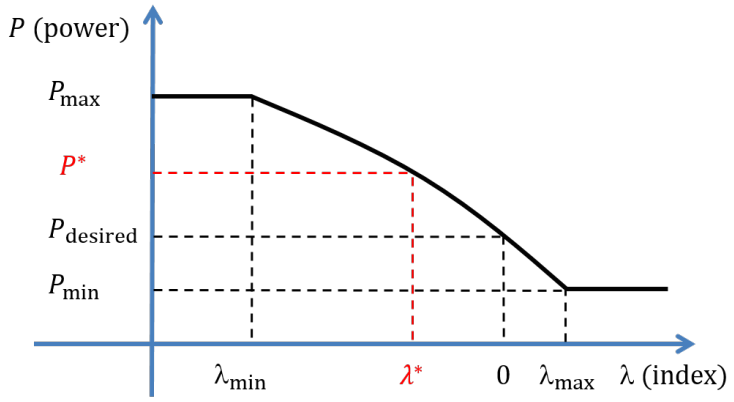


Figure 20. Illustration of demand flexibility of a residential AC.

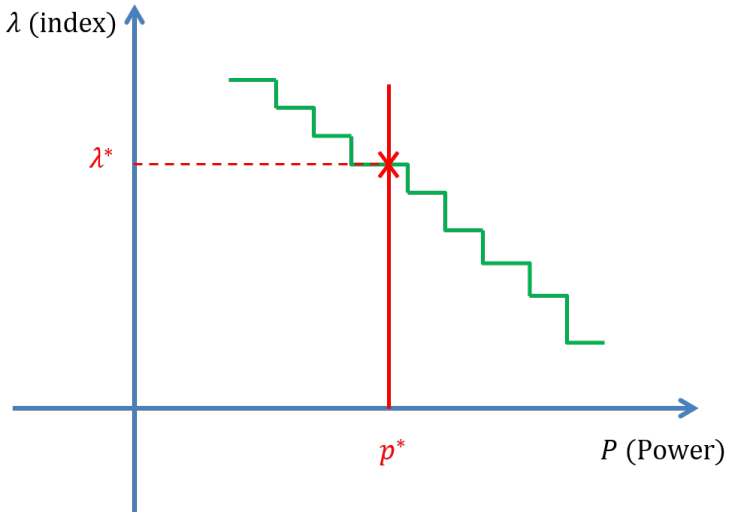


Figure 21. Illustration of determination of desired ramping index.

3.3.2 Performance

Performance metrics for *Category 3: Synthetic Ramping Reserves* from the FOA are listed in Table 1. A random ensemble of 1200 ACs and 1200 EWHs was generated. Starting from a random initial conditions, the ensemble was first simulated with outdoor air temperature as shown in Figure 22, which was measured at Columbus, OH on August 16th, 2009. Two ramping events are considered in order to support high PV penetration levels (the "duck curve"). The first event occurs from 9 am to 1 pm requesting ramp-up reserve at 30% RMT for total 4 hours. The other event occurs from 4 pm to 8 pm requesting ramp-down reserve at 30% RMT for total 4 hours. The recovery period is from 1 pm to 4 pm (total 3 hours), during which, the aggregated power of the load population is increased by another 25%. The coordination period is selected to be 5 minutes with half a minute as the simulation time step. In Figure 23(a), the five-minute average of the aggregated power of the load population under ramping control is compared to the case when it is not. Figure 23(b) shows the ramping control performance by closely examining the tracking of ramping signals, where half-an-hour average power is used. We can see that the response time, as in the case with regulation, is only restricted by the simulation time-step (30 s in this case), and is well within the target value of the metric. As shown in Figure 23, the target values related to the metrics of recovery time, ramp time and RMVT are also satisfied. In order to verify the availability metric, the same testing is performed for the week including August 16th, 2009. The profile of the outside air temperature is shown in Figure 24, where August 16th was the hottest day. The control performance by considering the same ramping events for the other days are shown in Figure 25–Figure 30, respectively. We can see that the proposed control can ensure 100% availability. Table 7 summarized the performance.

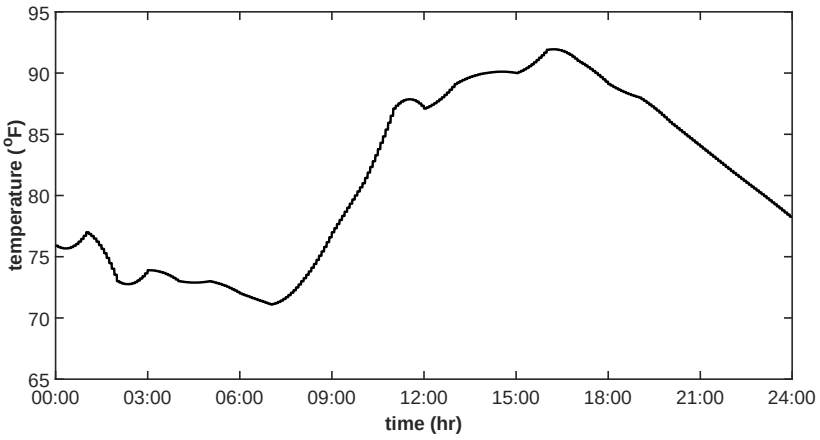
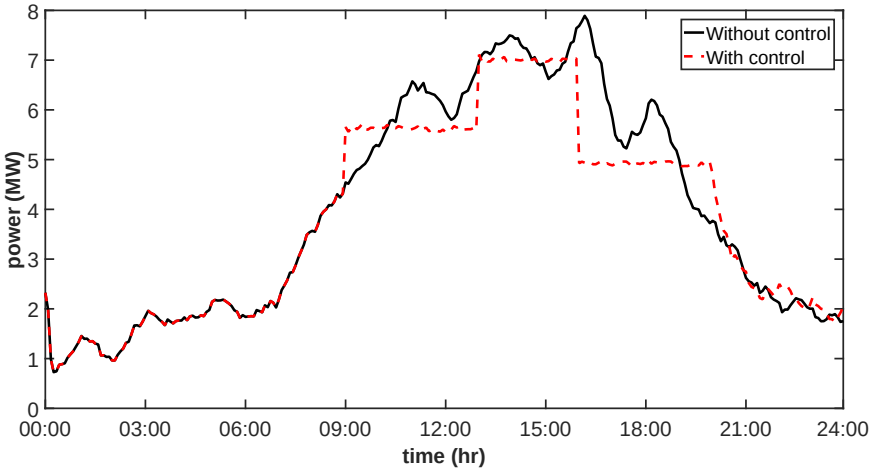
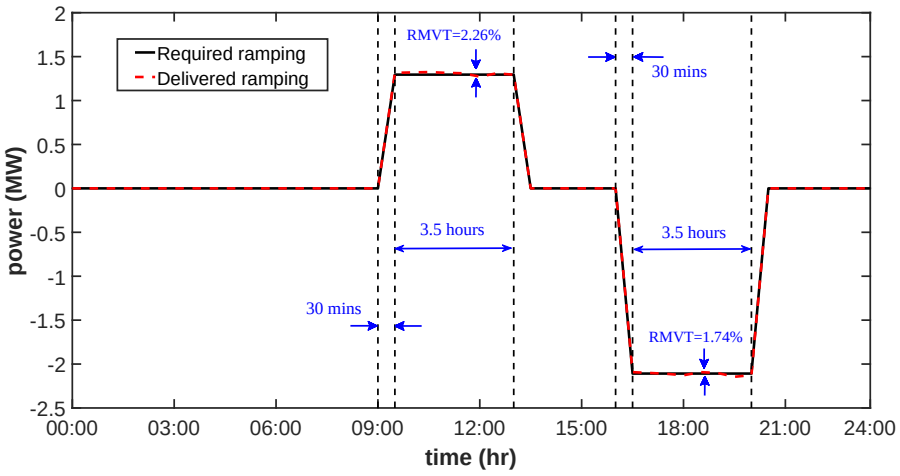


Figure 22. Exterior air temperature on August 16th, 2009.



(a) Aggregated power



(b) Tracking performance

Figure 23. Performance of an ensemble of 1200 ACs and 1200 WHs in supporting two back-to-back ramping events that request $\pm 30\%$ RMVT with a 3 hours recovery time in-between on August 16th, 2009.

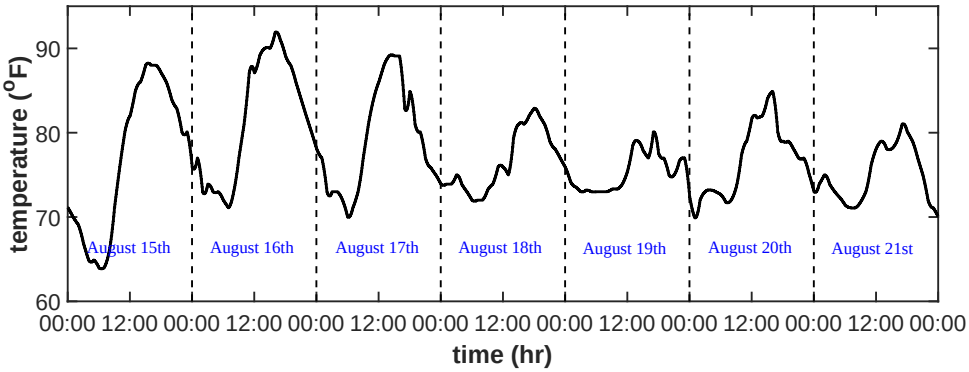
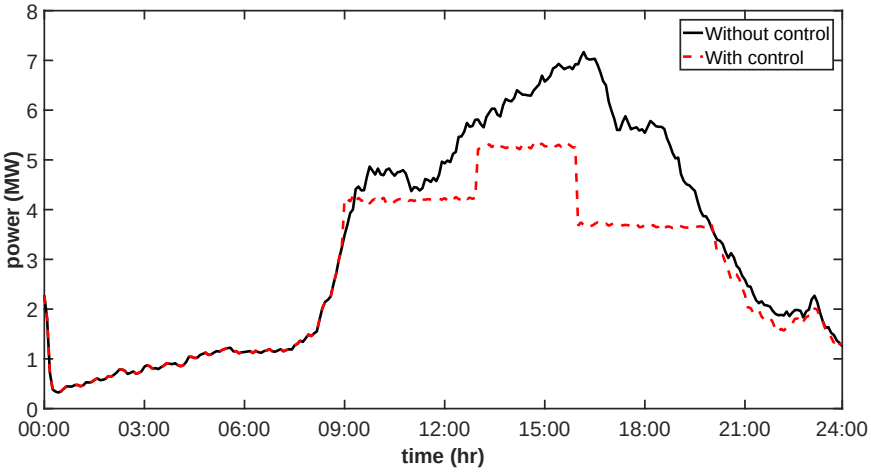


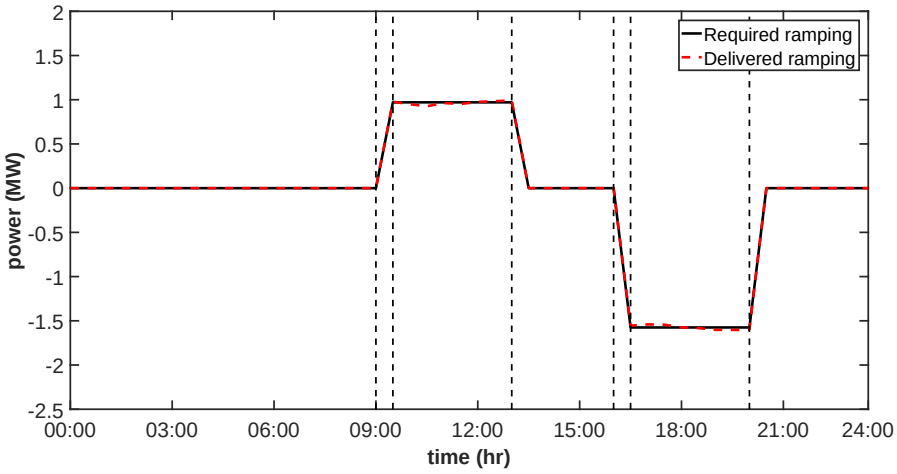
Figure 24. Outdoor air temperature from August 15th, 2009 to August 21st, 2009.

Table 7. CATEGORY 3 Performance Evaluation

Metric	Target	Target Met?	Achieved Performance
Initial Response Time	$<$10 min	YES	30 seconds (limited by sampling time)
RMT	$>$10 %	YES	tested at 30 %
RMVT	$< \pm$5 %	YES	$<$5 % (half-an-hour average)
Ramp Time	$<$30 min	YES	smoothly ramps up within 30 minutes
Duration	$>$3 hr	YES	successfully tracks for 4 hours
Availability	$>$95 %	YES	$>$95 %
Recovery Time	$<$4 hr	YES	recovery time tested is 3 hours

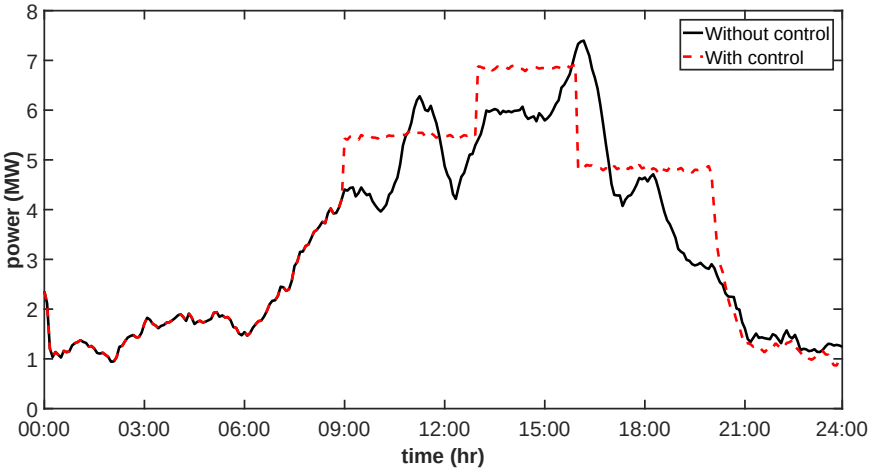


(a) Aggregated power

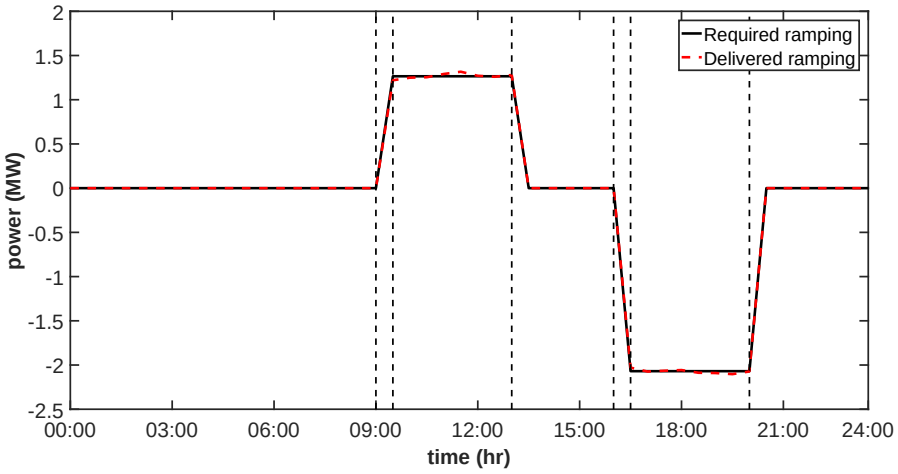


(b) Tracking performance

Figure 25. Performance of an ensemble of 1200 ACs and 1200 WHs in supporting two back-to-back ramping events that request $\pm 30\%$ RMVT with a 3 hours recovery time in-between on August 15th, 2009.

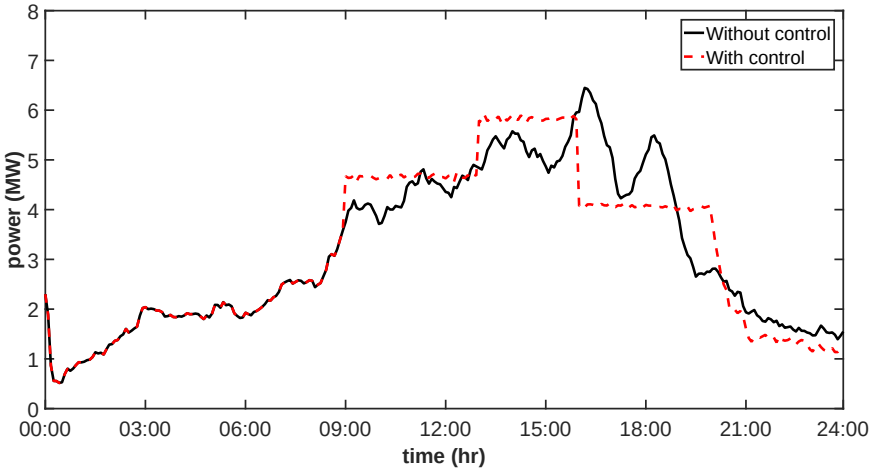


(a) Aggregated power

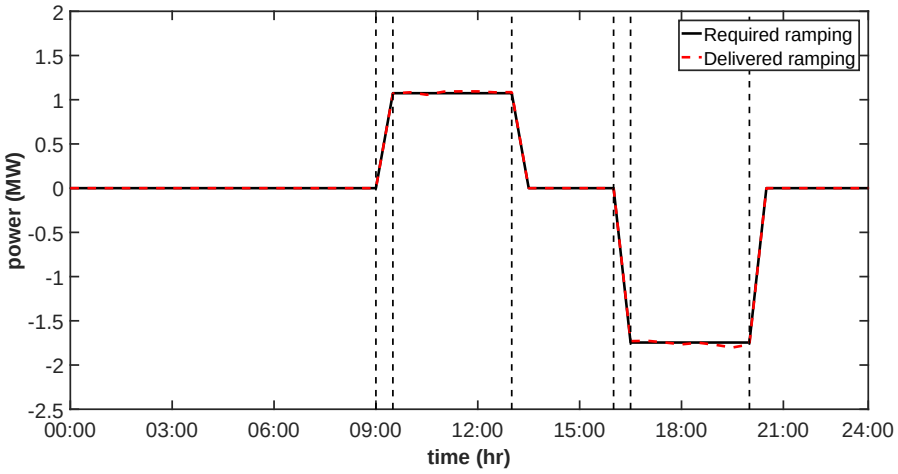


(b) Tracking performance

Figure 26. Performance of an ensemble of 1200 ACs and 1200 WHs in supporting two back-to-back ramping events that request $\pm 30\%$ RMVT with a 3 hours recovery time in-between on August 17th, 2009.

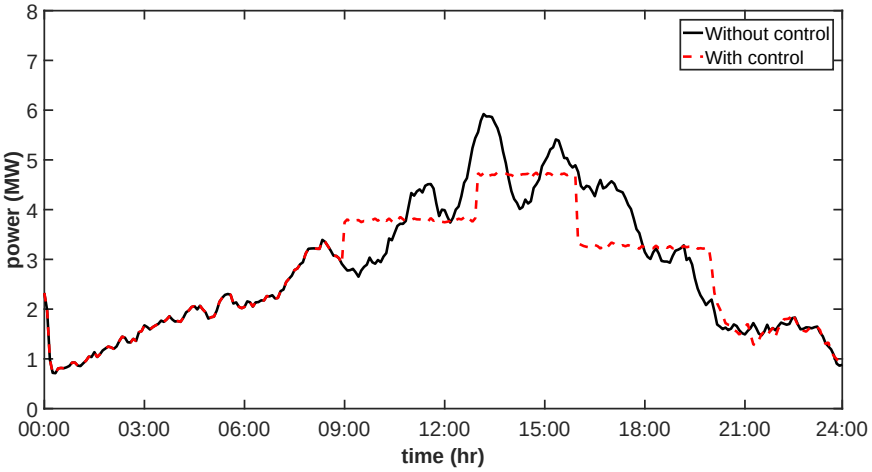


(a) Aggregated power

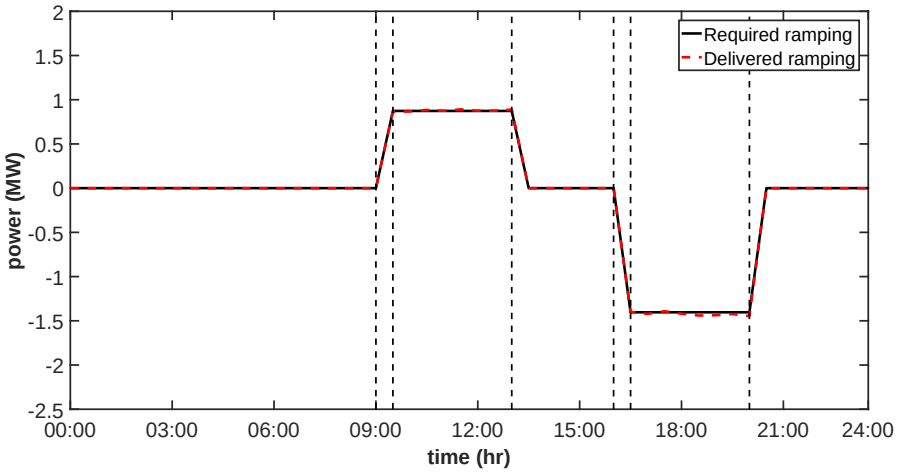


(b) Tracking performance

Figure 27. Performance of an ensemble of 1200 ACs and 1200 WHs in supporting two back-to-back ramping events that request $\pm 30\%$ RMVT with a 3 hours recovery time in-between on August 18th, 2009.

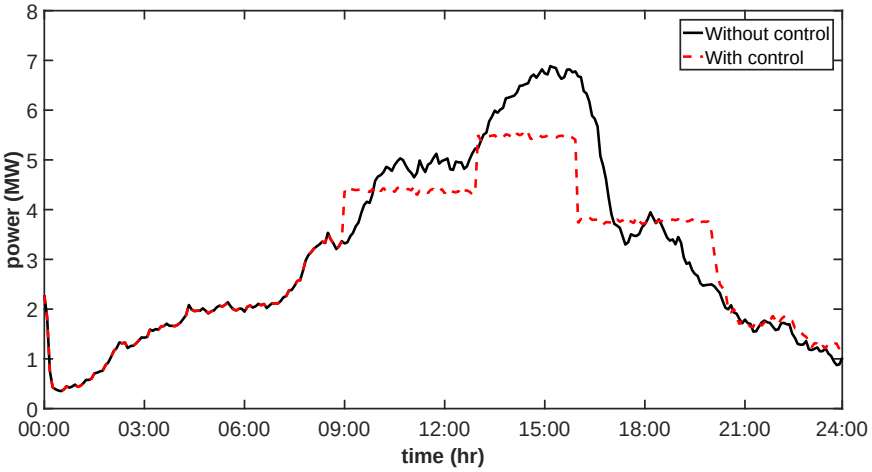


(a) Aggregated power

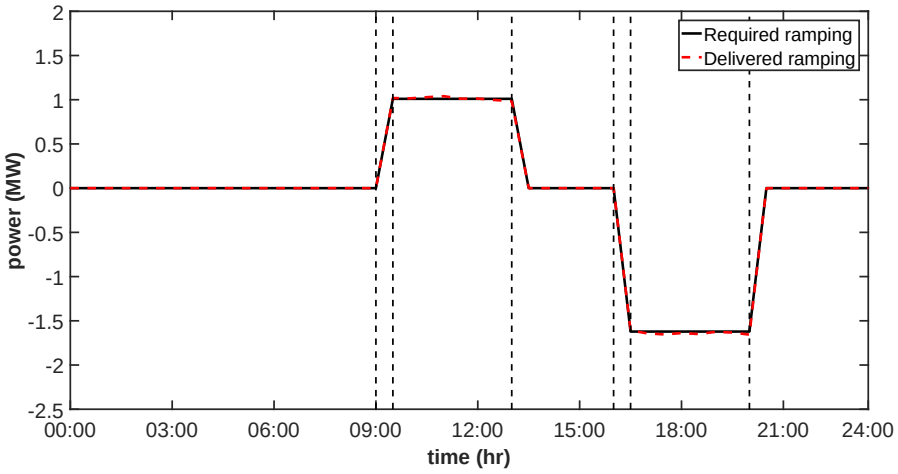


(b) Tracking performance

Figure 28. Performance of an ensemble of 1200 ACs and 1200 WHs in supporting two back-to-back ramping events that request $\pm 30\%$ RMVT with a 3 hours recovery time in-between on August 19th, 2009.

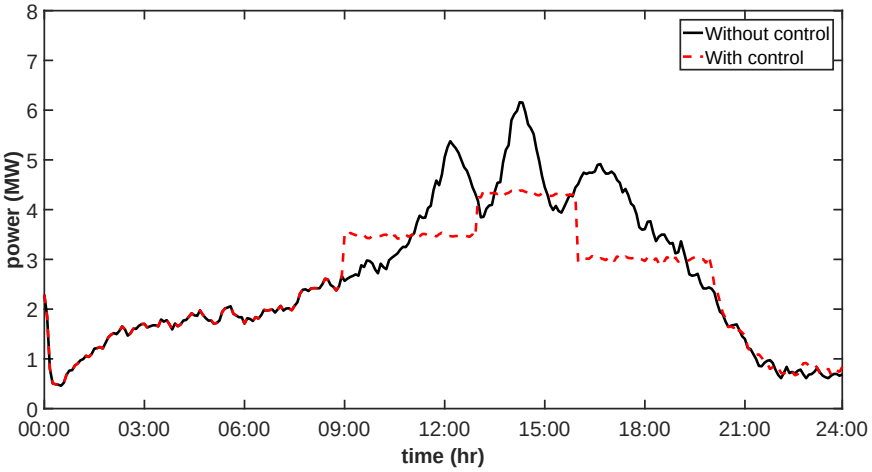


(a) Aggregated power

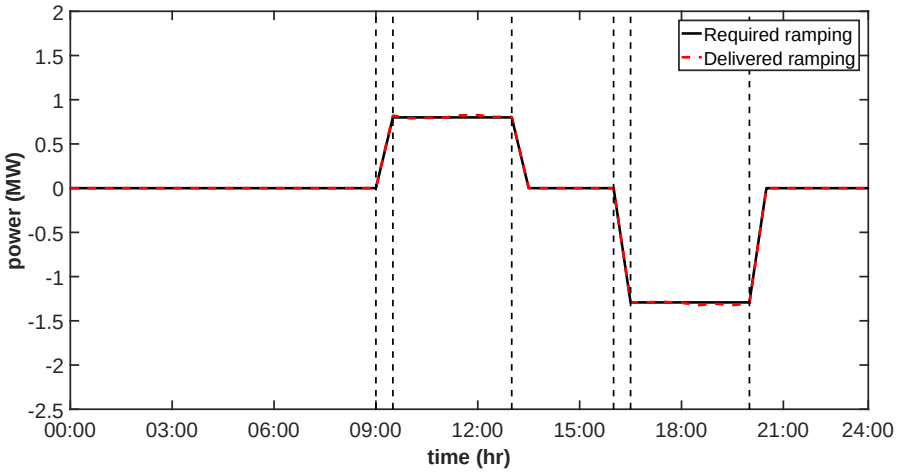


(b) Tracking performance

Figure 29. Performance of an ensemble of 1200 ACs and 1200 WHs in supporting two back-to-back ramping events that request $\pm 30\%$ RMVT with a 3 hours recovery time in-between on August 20th, 2009.



(a) Aggregated power



(b) Tracking performance

Figure 30. Performance of an ensemble of 1200 ACs and 1200 WHs in supporting two back-to-back ramping events that request $\pm 30\%$ RMVT with a 3 hours recovery time in-between on August 21th, 2009.

4.0 Commercial HVAC Characterization, Control Design, and Results

This section presents methods for estimating the demand flexibility of commercial buildings' Heating Ventilation and Air Conditioning (HVAC) system and the device-level control architecture to enable the execution of committed grid reserves while ensuring quality of service.

4.1 Flexibility Characterization

In this subsection, we summarize the work performed to 1) qualify HVAC ventilation fan to provide frequency regulation (FR) and ramping services based on defined metrics for response and ramp time, 2) quantify the magnitude and frequency bandwidth of the FR service it can accurately provide without impacting the building occupants' comfort.

4.1.1 Description of Reference System

UTRC's high performance building test-bed (HPBT), a medium-sized commercial office building with two Air Handling Units (AHUs) and 40 Variable Air Volume (VAV) terminal units was used for the experimental testing. Each of the AHU's is connected to multiple VAV boxes (with reheat coils). The building is operated with a building management system from Automated Logic Corporation (ALC) and can be monitored through its WebCTRL interface. The building has a total of 4 fans (2 supply and 2 return fans). The experimental results presented are for the supply fan that serves 15 interior zones, mostly conference rooms and multi-occupant offices. The main assumption is that the chilled water loop is decoupled from the ventilation fans power consumption due to the small and fast fan power variation during frequency regulation control.

4.1.2 Flexibility Qualification

The qualification test includes procedures that vary the real building fan power consumption by 1) commanding the fan on and off, 2) changing the fan speed directly and 3) changing the duct static pressure (DSP) set-point input to the fan speed controller. The functional testing for AHU fan response to ON/OFF commands turn the fan ON or OFF and accurately capture the time the fan is commanded, when the fan power turns non-zero or starts to decrease and the time the power profile reaches steady state. The fan speed-to-power functional test reduces or increases the fan speed from a steady-state condition by increments of 25% and accurately captures the instant the fan speed changes were commanded, the instant the fan power value changed, the fan speed feedback and the power profiles. Similarly, the static pressure set-point-to-power experiment reduces or increases the DSP from a steady-state condition by increments of 0.25 in H20 and accurately captures the set-point commands, the static pressure feedback and the fan power response. Figure 31 shows a sample incremental step changes in the static pressure set-points, fan speed and fan power consumption.

The experimental results are summarized in Table 8. The AHU fan qualifies to provide frequency regulation and ramping services based on the response and ramp time metrics in Table 1. The three control modes (ON/OFF, direct AHU speed modulation, and closed-loop control of DSP) all resulted in satisfactory dynamics. The On-to-Off control mode has the fastest response and ramp time (1 and 5 seconds respectively), while the DSP control has the slowest response time of 4-6 seconds and ramp time of 38-64 seconds. In the next section, we will show that the ventilation fan also meet the reserve capacity and operational duration metrics.

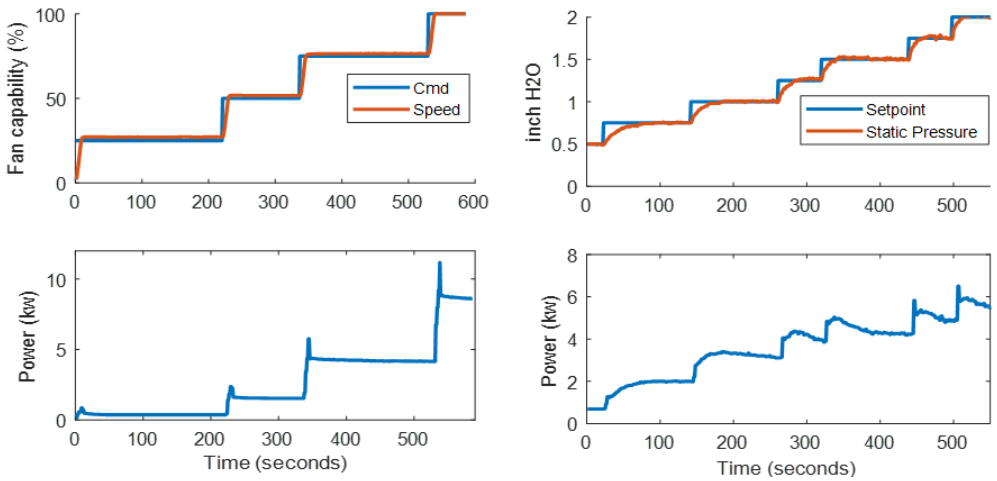


Figure 31. Fan power response to AHU fan speed (left) and DSP set-point (right)

Table 8. Summary of AHU fan qualification results

FR Metric	Requirement	Control mode			
		Off-On	On-Off	Fan speed	DSP Control
Response time (sec)	5	2	1	2	4 -6
Ramp time (sec)	300	32	5	8-12	38-64

4.1.3 Flexibility Quantification

The FR service corrects short-term imbalance on the power grid and the FR reference signal is usually high in frequency and energy neutral on average. Noting that such high frequency change in the fan power consumption will have minimal impact on the indoor environment due to commercial buildings' large thermal mass and inertia, it is sufficient to characterize the building HVAC flexibility for FR in terms of power magnitude and rate limits. To this end, we describe the procedure we used to determine the reserve magnitude and limits on the frequency range for which the building ventilation fan is most effective.

The frequency bands are limited by the equipment operational constraints and the indoor climate requirements. Low frequency variation in the fan speed and hence fan power could result in significant temperature variations, and the existing zone climate controls would react and reject the temperature deviation, resulting in degraded quality of the ancillary service. Similarly, the fan capability to track a "very" high frequency signal could be limited by the equipment operational constraints such as the fan motor ramp-up and ramp-down rate limits that are enforced by the baseline controls to prevent the equipment damage.

The fan controller uses the building duct static pressure (DSP) sensor to modulate the supply fan motor speed, and maintain the static pressure at a desired set-point. As the space temperature deviates from set-point, the zone controller responds by adjusting the damper position to increase or decrease airflow. Varying the zone damper position causes the pressure inside the supply duct to change, causing the controller to adjust the fan speed (and hence supply airflow). The static pressure set-point could be a constant or varied using "Trim & Response" logic, which resets the set-point based on zone damper position. The operation of the return air is synchronized with the supply fan and runs at 90% of the supply fan speed. For frequency regulation,

the fan power consumption can be continuously modulated by varying the fan speed directly or changing the static pressure set-point to the fan controller. The latter is selected for consistency with the baseline controller implementation and to minimize potential reliability issues. Unlike the direct fan speed control, this architecture ensures that the local fan motor constraints will always be satisfied.

The objective of the qualification test is to determine the reserve magnitude and frequency range at which fan power can provide frequency regulation without impacting the building zones thermal comfort. A sequential iterative scheme was used for the experiment design. The set-point input is a sinusoidal signal ($DSP_{sp}^{nom} \pm A \sin wt$), where the perturbation A is selected to respect the allowable DSP when the building is operating:

$$A_{max} = \min(DSP_{sp}^{max} - DSP_{sp}^{nom},\ DSP_{sp}^{nom} - DSP_{sp}^{min}) \tag{24}$$

where $A \in [0.25, A_{max}]$, and the frequency w lies within the band:

$$w \in [1/(10min), 1/(30sec)] \tag{25}$$

For this experiment, the nominal DSP set-point was 1.75 in H_2O and nominal fan power consumption was 3.93kW.

The generated frequency response data from DSP set-point to AHU fan power consumption and zones temperature variation are shown in Figure 32. The indoor climate quality was maintained during the experiment, with the temperature within $\pm 1^oF$ of the set-point. The comfort range, determined by the baseline controls, differs from zone to zone but they are all at least $\pm 1^oF$. The frequency response plot of the experimental data is provided in Figure 33. The fan power can provide up to 1.7 kW (18.9% of its rated power and 43% of its nominal power) of frequency regulation reserve during operational hours without impacting the indoor climate or baseline controls. Since the reserve capacity varies with the frequency of the DSP set-point input, the frequency limits are selected considering controls requirement in addition to the discussed metrics. In particular, the frequency range where the system has large gain is preferred to minimize the control action (fan motor speed command) needed to achieve the regulation goal. The summary of the fan flexibility is provided in Table 9.

Table 9. AHU fan frequency regulation capability

Metric	Requirement	AHU fan
Frequency range (Hz)	NA	[0.0055 , 0.022] Hz ([1/(3min) , 1/(45sec)])
Amplitude (Hz)	NA	±1.7kW
Reserve magnitude (% load)	>5%	18.9% of rated power, (43% of nominal power)
Ramp time	<5 minutes	11.25 – 45 sec

4.2 Frequency Regulation for Building Air-side HVAC System

The commercial building frequency regulation control development and testing is focused on the air-side electricity consumer - the ventilation fans in the air handling units.

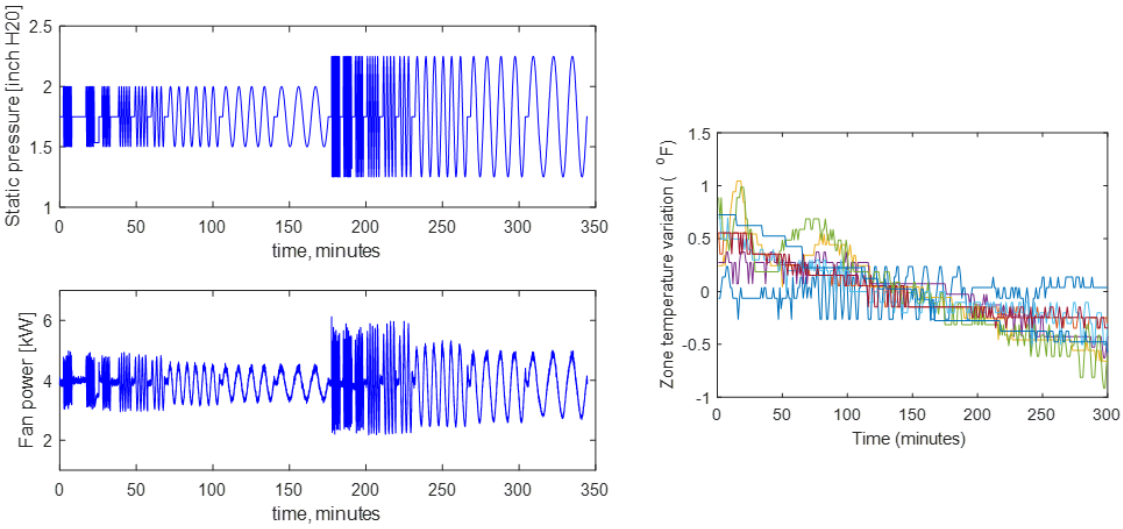


Figure 32. Left: Fan power response (bottom) to sinusoidal signals in static pressure set-point (top). Right: Variability in zone temperature during the frequency regulation experiment

4.2.1 Control Architecture

In this subsection, we describe the design of the frequency regulation controller (FRC) whose bandwidth includes the frequency range identified in the quantification test. The proposed frequency regulation control changes the command to the AHU fan motor speed (and hence power consumption) by indirectly modifying the duct static pressure (DSP) set-point. This architecture is selected for its ease of implementation. It only requires a simple software add-on to the existing HVAC control system; the local fan speed and zone climate control loops remain the same.

The FRC architecture is shown in Figure 34. The output of the controller (ΔDSP spt) is added to the nominal set-point before it is passed to the existing baseline fan control logic. The FRC varies the fan power consumption so that the deviation from its baseline power profile (ΔPower) tracks the reference signal $\Delta Power^{grid}$. The resource controller has two main blocks: nominal power estimator and the reference tracking controller.

Nominal Power Estimator: This block estimates the power the resource would have consumed under a normal operating condition, without providing ancillary service. The nominal power can be estimated through machine learning, low pass filter [19] or model-based constrained optimization [20]. In this study, the nominal estimator is a DSP-to-power data-driven model, identified as a second order transfer function based on the frequency response data generated in Subsection 4.1.3.

Reference tracking controller: The tracking controller computes the required perturbation to the nominal DSP such that ΔPower (= Measured power – estimated nominal power) follows the reference grid power ($\Delta Power^{grid}$). The controller is a classic feedback controller that maximizes the tracking performance within the identified frequency range of interest ($f \in [0.0055, 0.022] H_z$) while minimizing the actuator effort.

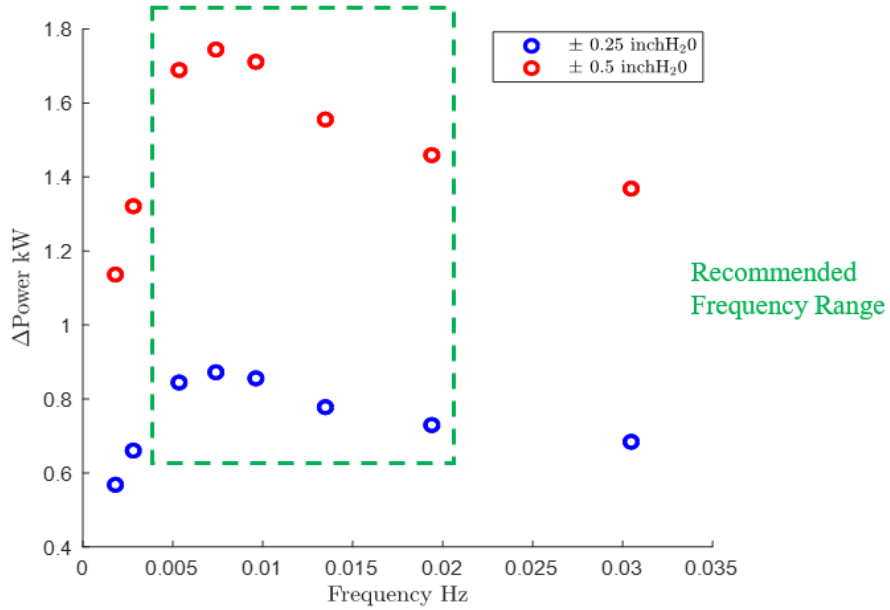


Figure 33. Frequency response plot of fan power response to sinusoidal changes in static pressure

4.2.2 Simulation Results

The frequency regulation controller was implemented in Simulink and communicates with the UTRC building BAS through the WebCTRL server that is connected to the ALC controllers via BACnet over IP network. The set-up is shown in Figure 35. Prior to experimental testing, the FRC controller was tested in a simulation environment with almost perfect tracking.

PJM standard test regulation signals (RegD) were used to assess the FRC performance. The signal ramps up and down fairly quickly and is almost energy neutral (on average) over a period of time. Each experiment was conducted over a forty minute period to meet the test duration for both the NODES program (>30 minutes) and PJM qualification requirements. Figure 8 shows the controller performance while tracking three RegD signals. The reference signals were verified to lie within the HVAC flexible frequency range (otherwise they would need to be filtered) and were scaled to match the available frequency regulation power magnitude.

The controller performance was quantified using NODES metrics, PJM performance metrics based on the formula provided in their manual ([21], pages 54-56) and the tracking error metric used in [19]:

$$r_R = \frac{\frac{1}{N}\sqrt{\sum_{j=1}^{N}\left(\Delta Power^{grid}(j) - \Delta Power(j)\right)^2}}{\max(\text{abs}(\Delta Power^{grid}))} \tag{26}$$

The results show that the building HVAC FRC is consistently able to provide satisfactory grid support. The evaluation against NODES metrics is summarized in Table 10. We note that 1) the response times are less than 4 seconds, meeting the 5 second requirement, 2) the reserve magnitudes range from 29% to 44%, exceeding the 5% requirement, and 3) the ramping times are less than 60 seconds, well within the 5 minute requirement. While the 5% target for the Reserve Magnitude Variability Tolerance (RMVT) was met in a simulation environment, a degraded performance was obtained in the experiments partly due to the measurement noise and communication delay between the local computer used for the FRC and the building controls

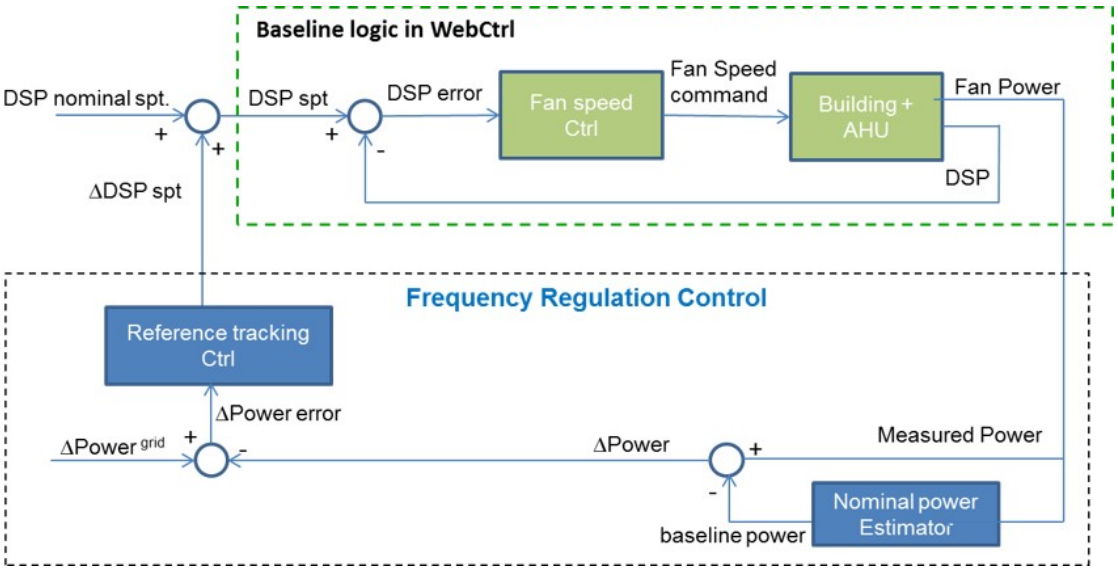


Figure 34. Frequency regulation control architecture

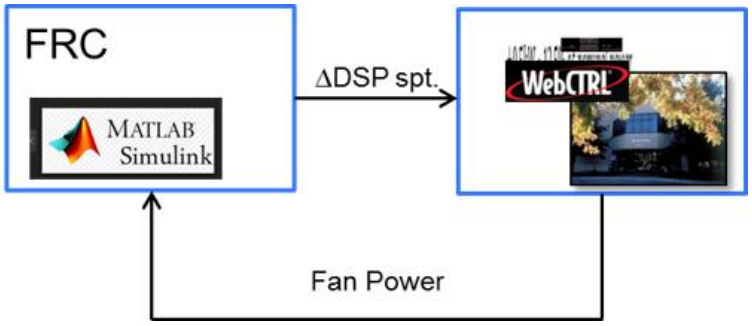


Figure 35. Experimental test set-up

Figure 35. The RMVT is calculated as the maximum variability between the regulation signal and actual power response over a five-minute window. The RMVT score is improved by programming the FRC as EIKON Logic in WebCTRL and using data from the existing control system (See Subsection 4.2.3). The PJM scores are provided in Table 11. The composite scores (0.89, 0.96, and 0.78) exceed the PJM threshold of 0.75. The tracking errors using eqn. 26 are 0.10, 0.11 and 0.08 for regD signals 1 to 3 respectively. The results compare favorably with the 0.19 experimental value obtained in [19].

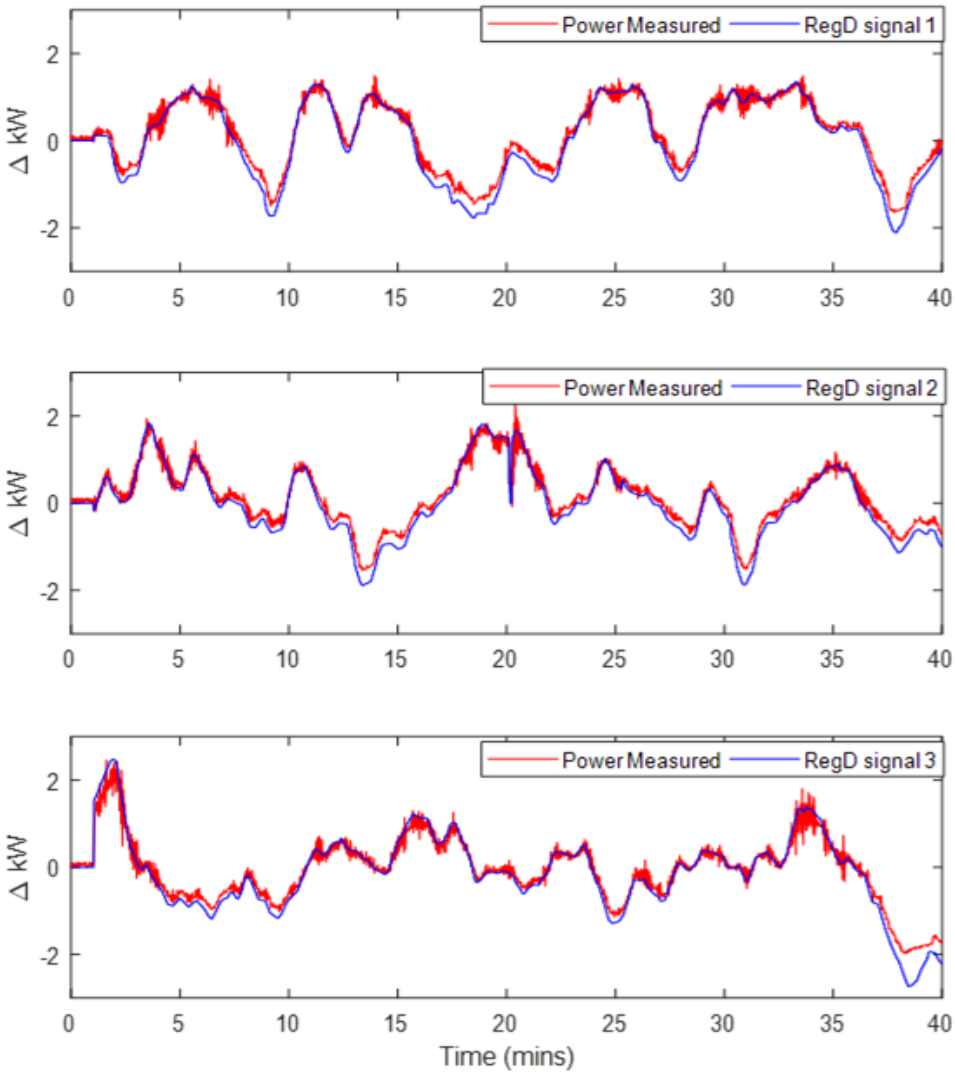


Figure 36. Experiment results for frequency regulation controller with three RegD reference signals

4.2.3 Frequency Regulation Control - Refinement and Implementation

Based on the learnings from the above tests, the FRC reference tracking control was re-tuned and a filter was implemented to attenuate the fan power sensor measurement noise. The regulating control logic was also implemented in ALC Eikon Logic code on the ventilation fan controllers to reduce communication latency. These changes were incorporated as part of the large scale federated test-bed results and are presented in Section 6.3.2. In short, the experimental testing against RegD PJM signal of the refined controller meets all the the NODES targets.

Table 10. FR Experimental results against NODES performance metrics

Metric	Target	Target Met	RegD signal 1	RegD signal 2	RegD signal 3
Initial Response Time	< 5 sec	YES	< 4 sec	< 4 sec	< 4 sec
Reserve Mag. Target	$> 5\%$	YES	29%	40%	44%
RMVT*	$\pm 5\%$	NO	$< 16\%$	$< 18\%$	$< 15\%$
Ramp time	$<$5 mins	YES	< 1 min	< 1 min	< 1 min
Duration	$>$30 mins	YES	$>$30 mins	$>$30 mins	$>$30 mins
Availability	95%	YES	$> 95\%$	$> 95\%$	$> 95\%$

* Reserve Magnitude Variability Tolerance

Table 11. FR Experimental results against PJM performance metrics

Test signal	Correlation score	Delay score	Precision score	Composite (mean) score
Reg D1	0.98	0.99	0.78	0.92
Reg D1	0.97	0.99	0.71	0.89
Reg D1	0.94	0.99	0.73	0.89

4.3 Ramping Service for Building Air-side System and Chiller Plant

The ramping controller leverages the buildings thermal storage capacity through the supply and return fans in the air-handling units and the chiller plant.

4.3.1 Description of Reference System

The simulator testbed used for the flexibility estimation and controls assessment is a medium-sized DOE reference office building. It consists of high-fidelity Modelica models of integrated building, HVAC system and baseline controls. The building has 15 zones and three floors with a total area of 54,000 ft^2. The HVAC model consists of a chiller plant and three air-handling units (AHU) and 15 VAV boxes on the air side. The chiller plant has two 80-ton chillers and two cooling towers. The chilled-water system has two loops – a constant-flow primary loop and a variable-flow secondary loop. Each floor is served by one AHU and the temperature of each zone is controlled by a PI controller with a VAV box. The building model captures the thermal dynamics of the building envelope and indoor air by modeling the physics of all the walls, windows, ceilings, floors and indoor air in detail. Daily schedules of internal heat gains from lighting, equipment and occupancy are defined and TMY3 weather data serves as an input to the model. To capture the short-term mechanical and momentum dynamics of the HVAC system, that may impact its performance for frequency regulation, experimental data from a real building functional test was used to calibrate the dynamic time response of the models. In particular, the following nonlinearities were added to the HVAC equipment models that affect their responsiveness. a) The travel time of chilled water going back and forth between chiller plant and AHUs is included. The travel time varies. Give that total length of water pipes is fixed, the travel time is inversely proportional to water flow speed, which is equal to the chilled water volumetric flow rate divided by cross-section area of supply or return pipe. b) The response time of the supply-air flow rate

in AHUs. The response time is due to the fact that the fan VFD has a rate limit on its output, which is 5%/sec up and 100%/sec down (no limit on reducing fan speed). The response delay of supply-air flow is equal to the VFD rate limit $\times$ nominal air volumetric flow of an AHU. c) The response time of supply-air flow rate in VAV boxes. The response time is due to the fact that a damper motor takes time to open and close. This response time is equal to damper travel time multiplied by the maximal design flow of a VAV box.

4.3.2 Control Oriented Model Formulation and Evaluation

The main control oriented models used for the building flexibility estimation and controls development are described below.

Zone temperature dynamics: To ensure scalability and also capture the longer time prediction requirement of the grid ramping service, a 3rd-order nonlinear auto-regressive model (27) with online load estimation is proposed for the zone temperature dynamics [22]. In this model, $\dot{m}_i$ is the mass flow rate of air in to zone i, T_{sa_i} is the discharge air temperature, T_{oa} is the measured ambient temperature and Q_i is the estimated heat load. For each zone, the constant parameters a_j are obtained from functional tests with data generated from a detailed Modelica model of a DOE reference office building. However, the value of Q_i is updated at every sampling instant (15 minutes) via moving horizon estimation to account for variability in internal heat gains and ambient conditions.

$$\begin{aligned} T_i(k+1) = a_1 T_i(k) + a_2 T_i(k-1) + a_3 T_i(k-2) + a_4 \dot{m}_i (T_{sa_i} - T_i(k)) \\ + a_5 T_{oa}(k) + a_6 T_{oa}(k-1) + Q_i \end{aligned} \tag{27}$$

The accuracy of the model is demonstrated in Figure 37. The model is tested with data representative of a typical summer day in Miami from 2 to 5 pm. Note that the measured ambient temperature, T_{oa}, and the estimated heat load Q_i obtained at 2 pm is held constant during the three-hour prediction. The result shows that the model captures the zones temperature evolution of all the zones with less than 0.2^oC of error for the first 15 minutes and within 1^oC error for the three hour prediction. The largest excursions are observed in the larger corner zones on the first floor.

AHU Fan Electrical Power Consumption: Each AHU supply mass air flow rate is computed as the sum of supply air flow to each zone VAV plus constant airflow leakage (28), and the fan power is modeled with a 3rd order polynomial function of the supplied airflow (29). The correlation parameters were trained with simulated data, and the evaluation result is shown in Figure 38 with root mean square error of 0.0249. The return and supply fans are linked, with the AHU return fan motor typically running at approximately 90% of the supply fan speed.

$$\dot{m} = c_0 + c_1 \sum_{i}^{N_z} \dot{m}_i \tag{28}$$

$$P_{fan} = b_0 + b_1 \dot{m} + b_2 \dot{m}^2 + b_3 \dot{m}^3 \tag{29}$$

Chiller Plant Electrical Power Consumption: A scalable data-driven approach was used for modeling the chiller plant power consumption. Rather than developing individual models for the chillers, pumps and cooling tower fans, the approach predicts the total chiller plant power consumption (P_{cp}) as a function of the chilled water supply temperature (T_{chws}), summation of the AHUs coil position (ϑ_{coil}), number of chillers running (N_{ch}) and outside air wet bulb temperature $T_{oa_{wb}}$ (30). Different supervised learning approaches including neural network and nonlinear regression models were tested. It was concluded that a quadratic function of the listed four features meets the

modeling requirement. This approach ensures the scalability of the proposed flexibility estimation framework to diverse chiller plant configurations and simplifies the resulting optimization problems. Combined functional test and normal operation data were used to train the model. The expected trend with changing chilled water supply temperature and water flow rate was also verified. Figure 39 shows an example of the model validation with operational data and the accuracy is provided in Table 12. The comparison of the simplified models with the high-fidelity simulation shows that the mean absolute error is within 10% for power consumption.

$$P_{cp} = f\left(T_{chws}, N_{ch}, T_{oa_{wb}}, \sum_{j}^{N_{ahu}} \vartheta_{coil_j}\right) \tag{30}$$

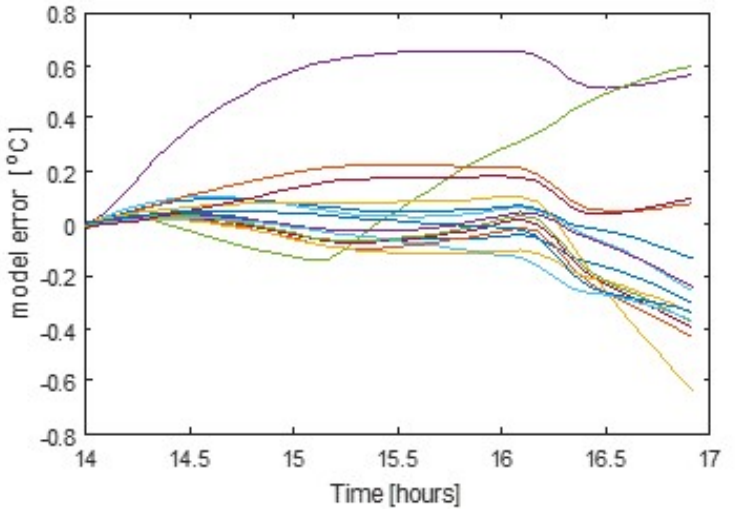


Figure 37. Zone temperature open loop prediction error over three hours with a fixed T_{oa} and Q_i

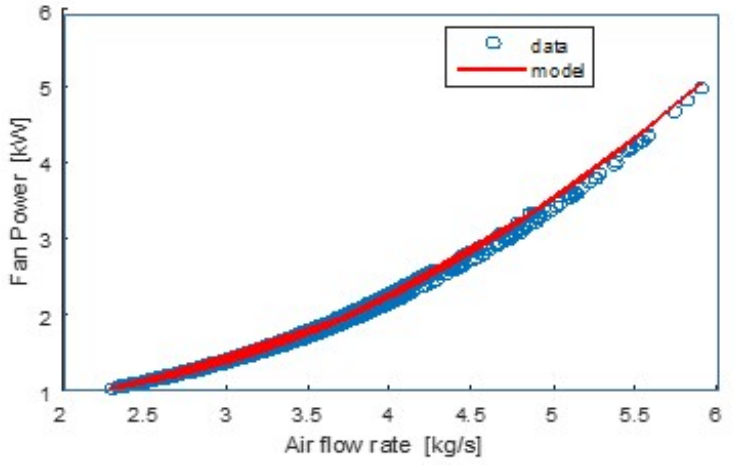


Figure 38. Fan power validation – cubic correlation to airflow

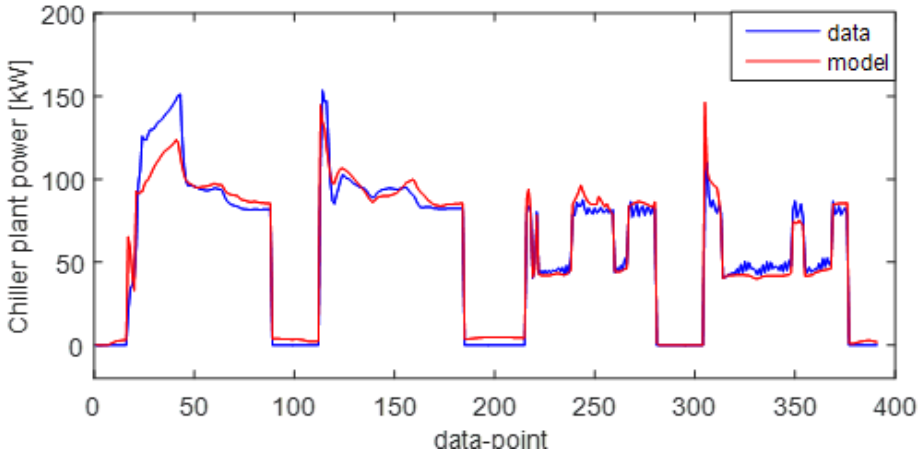


Figure 39. Chiller-plant power model validation

Table 12. HVAC Power model validation

	Root mean square error $\frac{1}{N}\sqrt{\sum_{j=1}^{N}\left(y_j-\hat{y}_j\right)^2}$	Mean absolute percent error $\frac{100}{N}\sum_{j}^{N}\left\|\frac{y_j-\hat{y}_j}{y_j}\right\|$
AHU fan power (kW)	0.025	0.96%
Chiller plant power (kW)	10.2	8.4%

c

4.3.3 Flexibility estimation

The upward and downward flexibilities of the building HVAC system from its baseline operation is represented as a virtual battery model:

$$X_{k+1} = (1 - aT_s)X_k + P_k T_s \tag{31}$$

$$\underline{P}_k \le P_k \le \overline{P}_k, \quad \underline{X}_k \le X_k \le \overline{X}_k \quad \forall k \in \mathcal{K} \tag{32}$$

where $\mathcal{K} = \{1, 2, \ldots, N-1\}$, N is the prediction horizon, T_s is the sampling time, a is the building discharge rate, P_k is the power draw, X_k is the energy state (capacity). $\underline{P}_k/\overline{P}_k$ and $\underline{X}_k/\overline{X}_k$ are the lower/upper power and capacity limits, respectively. To determine the equivalent parameters and constraints for the complex and dynamically interacting HVAC load, our approach combines data-driven and physics based-modeling to learn the building's thermal characteristics and the HVAC equipment efficiencies. Then we used an optimization-based approach to translate the building thermal flexibility to the virtual battery model. The building flexibility is predicted over a time horizon and re-computed at predefined time intervals to adapt to system operating conditions, weather, and heat gains.

Since the building flexibility is defined in terms of the acceptable increase or decrease in the nominal HVAC power consumption, the first step is to determine the baseline power consumption profile. The state-of-the-art building automation system controls the zone temperature to a specified reference set-point T_r or within a comfort bound using decentralized PID controllers and trim-and-response logic. For this work, the baseline power is derived as the total electric power consumed by the AHU fans and the chiller plant to maintain the zone temperature at its set-point

T_r, while satisfying equipment and operational constraints. Note that the baseline power can also be estimated through machine learning techniques if sufficient data is available. Mathematically, the baseline power profile (P_k^{base}) is obtained by solving the following optimization problem:

$$\begin{aligned} \min_{u_{z,k}^i, u_k^i} & \sum_{k\in\mathcal{K}} \sum_{i=1}^{N_z} \left(T_k^i - T_r^i\right)^2 \\ \text{s.t.} \;\; & T_{k+1}^i = f\left(T_k^i, u_{z,k}^i, d_k^i\right) \\ & T_{k+1}^i \in \mathcal{T}, \quad u_{z,k}^i \in \mathcal{U}_z, \quad u_k^i \in \mathcal{U} \end{aligned} \tag{33}$$

where u_z contains the zone level control inputs (mass air flow) and u contains the AHU and chiller level control inputs (e.g., AHU discharge air temperature and chilled water supply temperature). N_z is the number of thermal zones in the building, $\mathcal{T}$ denotes the feasible temperature set determined by the comfort preferences of building occupants. Similarly, $\mathcal{U}_z$ and $\mathcal{U}$ are the feasible sets of control inputs determined from ventilation requirements and the building HVAC operational constraints.

The flexibility of power consumption is derived from the minimum and the maximum power consumed by the HVAC system, while satisfying comfort, equipment and operational constraints. This problem can be expressed as:

$$\begin{aligned} \text{max or } \min_{u_{z,k}^i, u_k^i} & \sum_{k\in\mathcal{K}} P_k \\ \text{s.t.} \;\; & T_{k+1}^i = f\left(T_k^i, u_{z,k}^i, d_k^i\right)) \\ & T_{k+1}^i \in \mathcal{T}, \quad u_{z,k}^i \in \mathcal{U}_z, \quad u_k^i \in \mathcal{U}. \end{aligned} \tag{34}$$

The minimum and maximum power flexibility is then computed as the difference between the optimal objective (P^{min}, P^{max}) and the baseline power consumption as follows

$$\underline{P}_k = P_k^{min} - P_k^{base}, \quad \overline{P}_k = P_k^{max} - P_k^{base}. \tag{35}$$

The battery charge or discharge rates are determined via system identification from active functional tests in which we use data segments during HVAC units transitioning between on/off modes. The rate is computed from the dominant time constant ($\tau = 1/a$) of each zone and then aggregated at the building level. This test is conducted at night time or other unoccupied period to ensure minimal variability in the building internal and exogenous loads. The whole building discharge rate is then calculated as the average across all zones, if the zone sizes and the thermal capacity are relatively close. Otherwise, it is approximated as the weighted average of the zones discharge rates.

$$a = \sum_{i=1}^{N_z} a_i \quad \text{or} \quad a = \frac{\sum_{i=1}^{N_z} a_i M_i}{\sum_{i=1}^{N_z} M_i} \tag{36}$$

where M_i is the maximum airflow rate to each zone.

The lower and upper limits of the energy capacity determine how much energy can be stored or dissipated from the building during a time window. The energy limits are calculated from (31), using the dissipation rate obtained from (36) and power limits from (35).

4.3.4 Flexibility Model Results

The nonlinear optimization model is developed in AMPL (a mathematical modeling language). IPOPT, an efficient open source solver for nonlinear programming, is used to solve the optimization problem. The number of operating chillers is modeled as a function of the building cooling demand, and approximated by a smooth logistic function.

Figure 40 shows a simulation result for the HVAC flexibilities when the zone temperature is constrained to $T^i \in [21\ 25]^oC$, the baseline temperature reference is set to the midpoint value 23^oC and the mean outdoor temperature is 31^oC. On average, the simulated building at the selected load conditions allows approximately 23 kW and 40kW of upward and downward flexibilities over the three hour prediction window. It can be observed from the decomposed power flexibilities (Figure 40B) that the AHU fans consumes about 20% of the total HVAC system power consumption and provides similar ratio in downward flexibility. Chillers are a much larger consumer of electricity, and hence a source of much greater ancillary service. The calculated minimum and maximum power in this example mostly correspond to the limits imposed by the VAV airflow and zone temperature constraints (comfort requirements). To further extend the capability of the building for grid support, there is need for technologies, such as occupancy sensing, that could be leveraged to further relax the minimum airflow constraint without impacting the buildings occupants' comfort. The building discharge rate for the simulated operating condition is estimated as 4.25/hour.

4.4 Ramping Control

This section describes the ramping reserve control design for a commercial building HVAC system. The objective is to track a given power reference assigned by an aggregator or resource allocation controller. A model predictive control approach is developed to coordinate both the supply and demand-side of the building HVAC system in order to track a prescribed load target that corresponds to the the received grid reserve signal. The efficacy of the flexibility estimation and control solutions were demonstrated in a simulation environment using the high fidelity Modelica model of a medium sized commercial building described in Subsection 4.3.1

4.4.1 Approach

The main electrical power consumers for the HVAC system are the ventilation fans and chiller plant (chillers, pumps and cooling tower fans). There are two main challenges in developing the control algorithms: First, when multiple chillers are present there is significant change in power consumption when the number of operating chillers change. The baseline system starts an additional chiller when the operating chiller is near its maximum capacity or the difference between the desired chilled water supply temperature T_{chws} set-point and sensor measurement surpasses certain thresholds for a duration of time. To minimize the impact of chiller cycling or staging on the ramping reserve controller, the baseline chiller sequencing control strategy was modified to limit the chiller cycling at intermediate loads by considering the HVAC power tracking error in addition to the metrics used in the baseline controls to trigger chiller addition or subtraction. Second, transport delay: there is a significant delay between the time a set-point change in the air distribution system is propagated to the chiller plant and impact the water side power consumption. To avoid system instability or overshoot due to the delay, a rate-limited zone temperature set-point relaxation is used on the building side, giving the baseline controls the degree of freedom to modify the AHU operation in response to the set-point changes. On

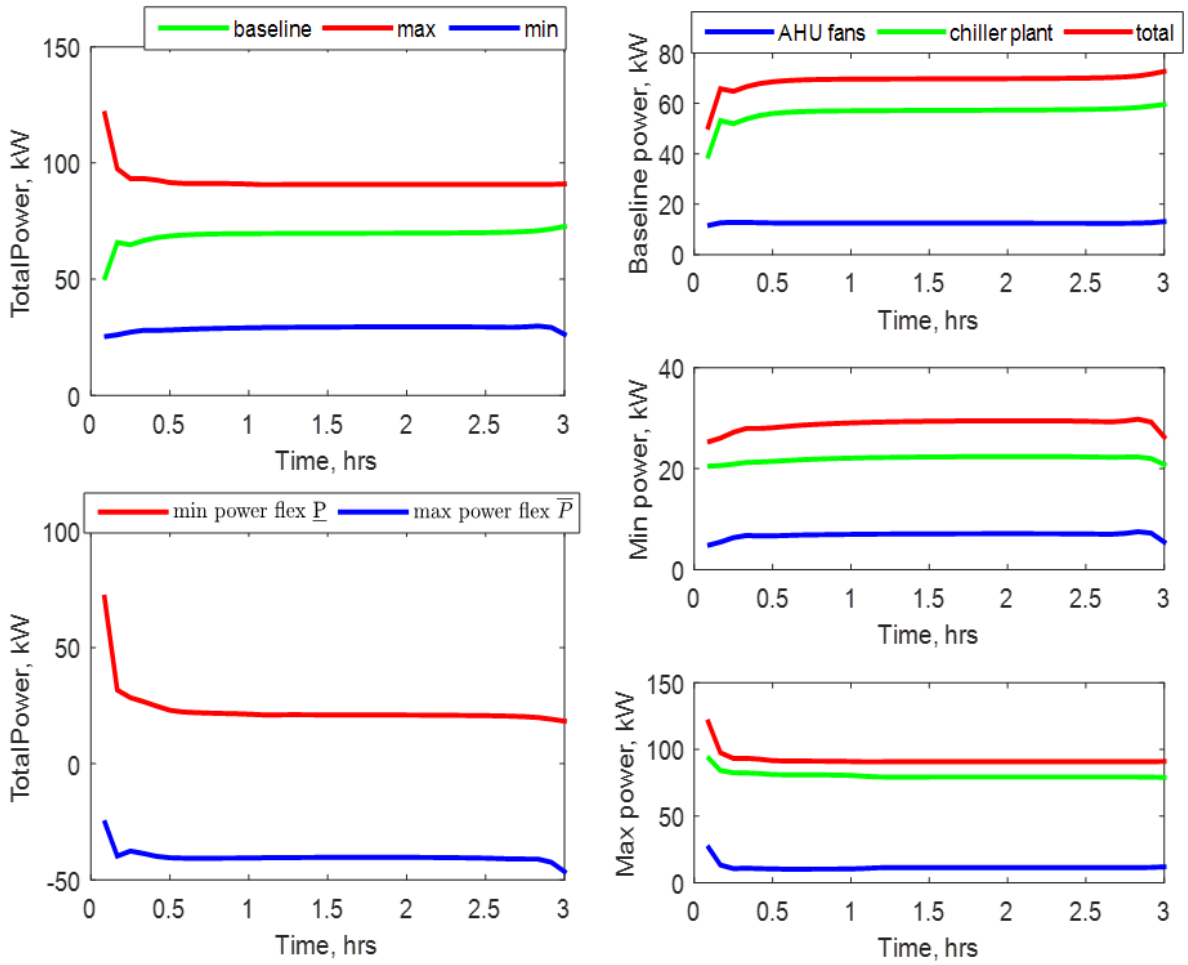


Figure 40. A: HVAC system flexibility, B: Decomposed flexibility.

the water-side, T_{chws} set-point is modified to generate a faster response in the chiller(s) power consumption.

Model predictive control is proposed for the tracking problem. The controller determines the optimal set-points of the chilled water supply temperature (T_{chws}) and the zone set-points (T^{sp}) within a prescribed bounds to minimize the tracking error between the building-level power consumption and the ramping signal received from the aggregator. The optimal control formulation is given below

$$\min_{T^{\text{sp}}_{i,j}, T_{chws}} \quad \sum_{k \in \mathcal{K}} \left(P_{\text{total}}(k) - P^*(k)\right)^2 \tag{37a}$$

$$\text{s.t.} \quad (27),\ (29),\ (30)$$

$$T_{i,j} = f(T^{\text{sp}}_{i,j}) \tag{37b}$$

$$T^{\text{min}}_{i,j} \leq T_{i,j} \leq T^{\text{max}}_{i,j} \tag{37c}$$

$$T^{\text{min}}_{chws} \leq T_{chws} \leq T^{\text{max}}_{chws} \tag{37d}$$

where (37b) is a model of the closed-loop system from the desired temperature set-point to the actual zone temperature and (37c) and (37d) reflect the constraints on the temperature set-points.

4.4.2 Simulation Results

For deployment and evaluation, the optimization based ramping controller is replaced with a simpler control architecture and set-point scheduler learned off-line from the optimal decisions made by the more sophisticated algorithm. The simulation result is shown in Figure 41. The HVAC power consumption is ramped up and down by a significant amount from 8am to 10pm, and the controller yields good tracking performance with 0.2^oC violation in thermal comfort. While the HVAC controller runs two chillers under nominal condition, the ramping controller drops a chiller later in the day at about 6.30pm. Minor undershoot was observed during the transient period as the number of operating chillers changes, overall, the controller performance meets the program metrics (see Table 13).

Table 13. Ramping controller performance

Metric	Target	Target Met	Performance
Initial Response Time	< 10 mins	YES	< 1 min
Reserve Mag. Target	$> 10\%$	YES	Tested at 40%
RMVT*	$\pm 5\%$	YES	$< 5\%$**
Ramp time	<30 mins	YES	< 30 mins
Duration	>3 hours	YES	> 3hours
Availability	95%	YES	$> 95\%$
Comfort violation			$< 0.2C$

* Reserve Magnitude Variability Tolerance
** Excludes ramping transients

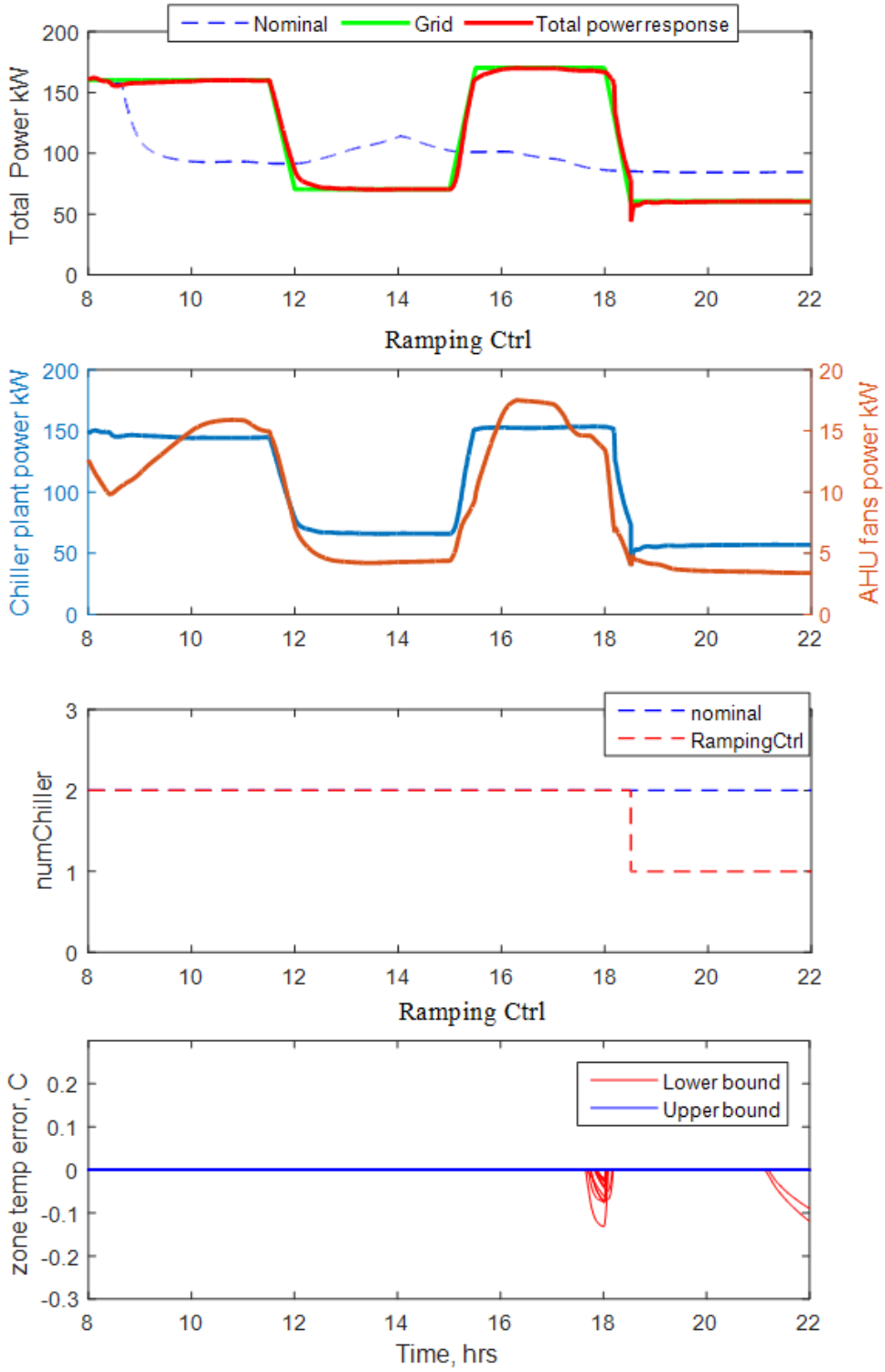


Figure 41. Chiller plant and AHU fans coordinated response to a ramping control reference that overrides typical power usage in a commercial building

5.0 Co-simulation Platform for Large-Scale Testing

This section describes the co-simulation framework developed at PNNL and interface to hardware-in-the-loop (HIL) assets at UTRC and Spirae. The architecture workflow is depicted in Figure 42.

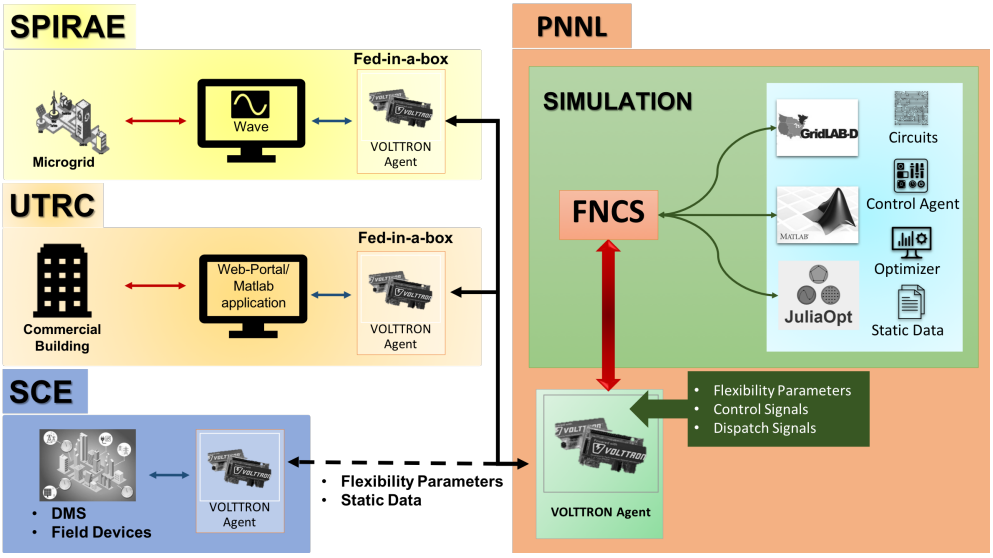


Figure 42. Nodes Experimental Setup and Architecture Diagram

5.1 Simulation Framework

The co-simulation framework is the overarching framework that brings together all tools and simulators using the Framework for Network Co-simulation (FNCS) toolset. FNCS is an open-source co-simulation framework developed by PNNL to integrate multiple simulators across multiple domains, ensuring interoperability across many different commercial and open-sources tools and synchronizing time and information across the different simulators. This allows researchers to explore the interactions of normally stove-piped planning and controls domains such as applications related to building systems and the grid, while developing new control and optimization solutions in tools that they are familiar with. The full simulation framework is depicted in Figure 43.

Each block in the framework describes a type of simulator that will be connected through FNCS. The reader should note that multiple instances of each box could be used for large-scale complex simulations. The framework consists of five groups of simulators. Four of them will be described in this section and the HIL federation will be described in the following section.

ISO Emulator: ISO emulator represents a static data simulator. It contains static recorded data that can be injected into the FNCS stream. This function allows a simplified way of including dynamic parts of the power system, which will not be modelled in detail. In this specific setting, the ISO emulator will provide three signals. The first signal is a capacity reservation signal that will be communicated to the DRC optimizer inside the optimization tool Julia. This signal describes capacity the ISO is procuring from the DRC. The remaining two signals represent a stream for the grid frequency and regulation requirement respectively. These signals will be broadcast directly to individual asserts inside GridLAB-D that participate in the proposed control framework.

Julia: Julia is an open source, high performance, and dynamic programming language. This language is used for building the multi-period power allocation optimization algorithm at the DRC

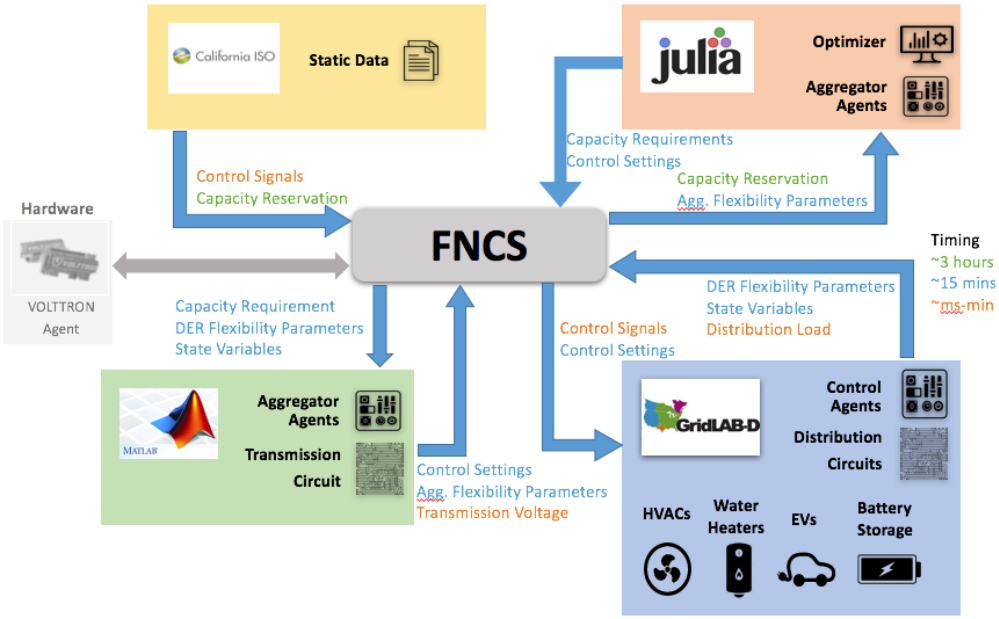


Figure 43. NODES Simulation Framework

level. Julia provides an excellent framework for solving computationally heavy optimization algorithms. The DRC receives the capacity reservation signal from the ISO, along with aggregated flexibility parameters from each individual Aggregator Agent. The DRC solves the resource allocation algorithm and forwards capacity requirements and control settings to each Aggregator Agent.

MATLAB: MATLAB is used in the simulation framework for two features. The first is to incorporate a transmission system solver using MATPOWER. This allows the ability to couple distribution systems with a single transmission system to allow for a more detailed co-simulation that can evaluate control performance in a truly integrated system simulation. Aggregator Agents receive control settings and capacity requirement from the DRC, along with individual DER flexibility parameters and state information from each device in GridLAB-D. This allows the Aggregator Agent to construct the aggregated flexibility parameters and forward them to the DRC. It also forwards control settings to each device inside GridLAB-D.

GridLAB-D: The GridLAB-D simulator is used to model distribution circuits. It also houses dynamic models of all types of controllable devices. This layer receives control signals broadcast from the ISO, along with control settings from the Aggregator Agents. Each device in this layer has a local Control Agent that provides its DER flexibility parameters and state information back to an Aggregator Agent.

The approach above was based on similar simulations frameworks that have shown prove capability for flexibility and scalability. Previous efforts at PNNL, performed under the control of complex systems initiative (CCSI), have shown that this framework can easily support thousands of distribution systems along with hundreds of thousands of controllable devices [23].

5.2 HIL Federation

The second part of the experiment setup are the hardware-in-the-loop (HIL) federation connections to collaborators' hardware assets. By design, the simulated elements are independent from the hardware elements; in other words, the simulations can be run without the HIL, but the ad-

dition of the HIL adds richness and fidelity to the experiments. A conceptual diagram of this connection is shown in Figure 44.

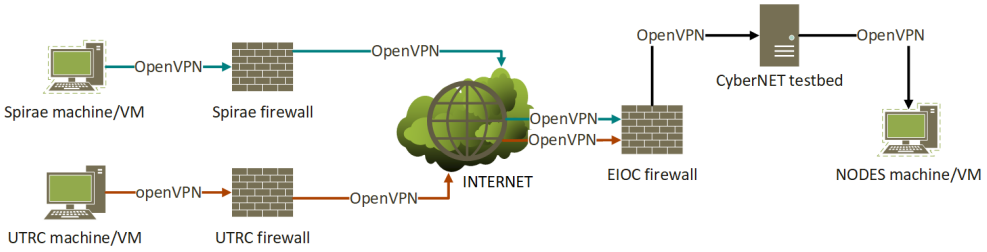


Figure 44. Conceptual overview of the connection from UTRC and Spirae to PNNL (NODES machine)

5.2.1 External Federation

Connecting two systems at two different organizations to exchange data and run integrated experiments is not a trivial task. Each organization's network is behind a firewall that blocks any external traffic into their systems. Hereby, the organization that is receiving the data is referred to as the "host entity" and the organization that is sending the data is referred to as "external entity". This external entity traffic can reach the host entity only when the administrator of the host entity permits the external entity to cross the firewall to let the traffic in. One of the ways to do that is to open a port and give the public facing port address to the external entity. But, doing so could risk the entire network exposure of the host entity to the external entity. From a cybersecurity perspective, an attacker can attack the external entity and potentially gain access to the host's complete network. Therefore, connecting different systems from different organizations poses significant challenges both from cybersecurity perspective.

To facilitate a secure interaction between two different organization, PNNL has designed a system called Fed-in-the-box. Fed-in-a-box is a virtual machine that has pre-scripted OpenSSH scripts that enables it to connect to a client or a server. The user provides the credentials to establish a SSH (layer) tunnel. Although the (two) sites are on different networks, Fed-in-a-box bridges the machines on these two sites (bridged tunnel) and makes it look like they are on same network. See Appendix A for details. By following the defined methodology, instead of exposing the entire network, both organizations would only expose "a" system to each other. This was achieved by opening a secure OpenVPN tunnel and connecting Spirae's and UTRC's systems to the system at PNNL. When these systems are connected through the OpenVPN tunnel, the connection and authentication goes through a CyberNET server and there is complete visibility of all the systems that are coming through this connection. Therefore, this establishes a secure connection and does not violate organizational policies. Fed-in-the-box can be established at both layer-2 (data link layer) and layer-3 (network layer).

The Fed-in-a-box enables coordination between the software systems, simulators and the remote hardware systems. This coordination is implemented in independent VOLTTRON agents installed at remote hardware locations. VOLTTRON [24] is an open-source, secure, extensible, and modular technology that supports a wide range of applications, such as managing end-use loads, increasing building efficiency, integration of distributed variable renewable energy, accessing storage, or improving electric vehicle charging.

After the network channels are successfully established, this setup would enable a seamless communication between the PNNL network port(s) and the remote network hardware systems

located at the collaborators' network. Since both the PNNL network and VOLTTRON already have several security layers ensuring a secure data transmission, the simulators can use all the computational resources required to execute the simulations in a timely fashion. Through the VOLTTRON interface, multiple simulators can transmit power profiles and receive data-points at the same time making these jobs very efficient.

Tying all of this together, the NODES experimental setup depicted in Figure 42 achieves a full-duplex and secure encrypted interface between external hardware assets and the internal (PNNL hosted) simulation testbed. More details about the collaborators testbeds and details about the federated connections is shown in Appendix C.

5.3 Devices under control

For this project, an expansive list of devices were used and actively managed by the designed control system. This includes both simulated assets at PNNL along with physical assets at Spirae and UTRC.

For the simulated portion of the evaluation the device list includes HVAC systems, water heaters (both electric and heat pump), energy storage (both residential and utility scale), and electric vehicles. Each experiment will have different number of devices being controlled. The final experiment will include over 10,000 controllable devices in the simulation. It is furthermore determined that all devices are capable of each service. However, it should be noted that some water heater types can be less suitable for the faster timescale services.

The HIL testing was be performed across the PNNL, UTRC and Spirae test sites. UTRC provided access to commercial building equipment in the form of 42 ventilation dampers and heating coils, and one supply ventilation and return fan. Spirae provided access to the following 42 controllable distributed energy resources: 1 battery inverter, 1 PV system, 2 diesel generator controls, 20 single phase 120 V controllable loads, 8 controllable load feeders, and 10 building controllable loads.

5.3.1 Physical Equipment Capability

It is important to classify what resources are able to adequately provide each category of service. UTRC and Spirae classified what devices are capable of reacting within the response time frame described by the NODES FOA < 2 seconds for frequency, < 5 seconds for regulation, and < 10 minutes for ramping. Table 14 shows the result for equipment at Spirae. Spirae determined that all physical equipment at their disposal can react within the NODES FOA requirements for all three categories.

Similarly, Table 15 shows the result for equipment at UTRC. Not all equipment at UTRC can react within the NODES FOA requirements for all three categories.

Response time for the AHU fresh air damper is typically on the order of $1 - 3$ minutes, which makes them suitable for Category III only. Similarly, the AHU heating coil valve typically responds to set-point changes in supply air temperature in about 60 seconds and the AHU cooling coil valve typically responds to set-point changes in supply air temperature in about 5-7 seconds. The means that both AHU cooling and heating coils are only suitable for Category III. Lastly, ventilation fans can provide service for all three categories. However, for Category I service an ON/OFF command is needed to achieve a typical response rate of 1-2 seconds. The building facility owner may not be willing to provide this flexibility for equipment reliability reasons. Typical response time to set-point changes is about 2-6 seconds for continuous operation.

Table 14. Microgrid (Spirae) device qualification

Device	Category I ($< 2s$)	Category II ($< 5s$)	Category III ($< 10m$)
EV	Yes	Yes	Yes
Battery inverter	Yes	Yes	Yes
PV inverter	Yes	Yes	Yes
Diesel generator	Yes	Yes	Yes
Synchronous generator	Yes	Yes	Yes
Asynchronous generator	Yes	Yes	Yes
Single phase load	Yes	Yes	Yes
Load feeder	Yes	Yes	Yes
Load bank	Yes	Yes	Yes
Building loads	Yes	Yes	Yes

Table 15. Commercial HVAC (UTRC) device qualification

Device	Category I ($< 2s$)	Category II ($< 5s$)	Category III ($< 10m$)
AHU fresh air damper	No	No	Yes
Heating coil	No	No	Yes
Cooling coil	No	No	Yes
Ventilation fan	Yes	Yes	Yes

5.4 Grid System Models and Data Sources

This section describes the grid models that were needed in order to perform the experiments in this project. The models span all the way from wholesale market operations to local distribution system loads

5.4.1 Transmission System Models

Modeling the transmission system, and its power flow constraints, is an important factor in accessing advanced control strategies for providing ancillary services. MATPOWER is the transmission system simulator used in this project to, in real-time, balance the network and dispatch generator portfolio. This model does not capture any of the markets that calculate any of the regulation signals. These signals are attained from collecting historical data from wholesale markets and applying appropriate modifications.

In this project, a modified IEEE 118 Bus system is used. This system is a high-fidelity simulation of the Midwestern U.S. power grid as of December 1962. This system is chosen as it is well studies and can cover the distribution system load required for this project. This system consists of 54 generators, 99 consumers, and 186 transmission lines, connected as shown in Figure 45.

This transmission system case is designed as a peak load study, with a total load of roughly

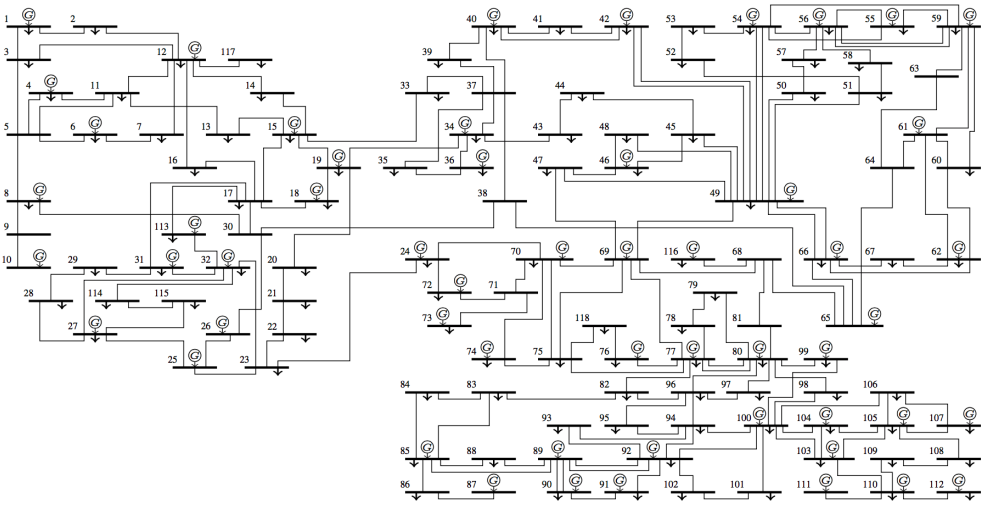


Figure 45. Modified IEEE 118 Bus System

4.2 GW. In order to create a more realistic case the fixed load is scaled by normalized load profiles at each transmission bus. These load profiles are obtained from the distribution system using averages of the historical data. At each bus in the transmission system where the distribution systems are connected, the peak load of this system is subtracted from the fixed load in order to maintain the same peak load level for different numbers of connected distribution systems.

5.4.2 Distribution System Models

Different distribution circuits were simulated in the project using GridLAB-D. These systems ranged from very simple systems that have attractive computational performance, to highly complex and realistic system for detailed control evaluation.

In order to facilitate exploratory testing in this project we developed a set of simple distribution system circuits. They contained the minimum amount of power system components without taking out detailed modeling of the secondary side. This allowed for fast computational performance compared to detailed systems, while still keeping important dynamics of the system intact. Those systems were highly flexible with respect to how many, and what devices are connected to them. That further made them ideal for preliminary testing.

After the preliminary testing the project began to perform a staged approach to use higher fidelity distribution systems. To achieve that, the project first leveraged past PNNL work that developed prototypical distribution circuits for various regions in the continental United States [25]. To develop those prototypical distribution circuits, planning models from distribution utilities across the continental United States were collected, and statistically analyzed to define a subset of representative prototypical feeder circuits that provided an effective approximate of the larger parent population. In that work, five climate regions were used to represent the United States.

In all of the circuits for this project, it is very important to not only model static power flows with simple load models. Instead, simple loads are replaced with physics-based load models including a state-based model of the residential HVAC system, water heater, electric vehicle (EV), and battery storage [26] [27]. These models are described in detail in the data section.

5.4.3 Wholesale Market, Dispatch, and Control Data

Not all aspects of the power grid were modeled in great detail and as such this project relied on recorded data for areas where detail was not needed. Data was also used to calibrate some of the model used as well. This is described further in the following section.

It is important for this project to incorporate realistic wholesale market, dispatch and control data. Since, none of these operations were actively modeled in this project static data was used instead. During control optimization recorded data such as simulated PowerWorld data and PJM regulation data, etc. was used to drive the control system. Data used for each category are as follows.

- Category-I Frequency Response: This was based on PowerWorld simulated data using the WECC planning models. This frequency signal was plugged into the GridLAB-D models as a boundary condition. Note that the PowerWorld model is be tied to any of the simulators but rather the frequency signals are fed into the simulators.

- Category-II Regulation: Historical PJM regulation data (REG-B – faster signal) was used to determine the regulation signal. The regulation signal is a static signal that is known prior to the simulation. It is played into the resource allocation algorithm in the distribution reliability coordinator.

- Category-III Ramping: No historical data was determined to be suitable for this new type of service. Therefore the project elected to have the ramping event be a certain percentage of the peak consumption of controllable devices in order to meet the FOA requirements.

In order to create realistic experiments, each of the models defined in this chapter was calibrated. For the distribution circuits this was done in one of two ways, depending on the models used. For the simplest system, the process of assigning residential houses was determined by calibrating the load models to the peak load of the system. The entire load allocation methodology is described in more detail in [28].

Each HVAC model is driven by outside climate, along with randomized parameters for each house to represent typical house distributions. The population is then calibrated to follow aggregate load shapes from ELCAP data [29]. Similarly, water heater models are driven by water usage schedules, randomized for each device to represent standard water draws (e.g., a shower versus a dishwasher cycle). These schedules are then calibrated to follow aggregate load shapes from ELCAP data (see [29]). EV models are driven by travel data from the National Household Travel Survey (NHTS). The data used throughout this project is from the 2001 survey [30]. This survey contains detailed information about driving patterns and logs of individual trip. This data is used for the EV modeling in GridLAB-D. Lastly, the battery storage is modeled as load following in GridLAB-D with specific settings for each battery storage system.

6.0 Large Scale Simulation, Testing, and Validation

This section describes the scenario used for testing in this project. It also describes in detail the large-scale simulations that were conducted as well as the inclusion of federated hardware experiments performed to quantify the performance of the developed control architecture.

6.1 Large-Scale Simulation Scenario

The evaluation scenario is a co-simulation that includes the simulators according to the framework described in Section 5.0. The scenario includes one transmission that models the modified IEEE 118 system. This system is coupled with 12 distribution systems picked from the prototypical systems, developed at PNNL. These systems have an expansive list of devices that are listed in Table 16 with their respective numbers for the specific circuits used. For the scenario four circuits of each type are used and the total number of devices that can be controlled is listed in Table 16 as well.

Table 16. Distribution system devices

Device	R1-1	R2-2	R2-3	Total
HVACs	750	647	1,182	10,316
Water heaters	1,165	654	1,096	11,660
EVs	159	84	151	1,576
Residential battery storage	84	46	79	836
Utility battery storage	0	1	0	1

With this configuration we ensure that more than 10,000 individual devices will be available for control. The evaluation scenario also includes market signals, as described in Section 5.4.3, frequency data at 1 second intervals, frequency regulation signals at 4-second intervals, and ramping signals every 5 minutes. The additional control elements include resource controllers for every controllable device, an aggregator controller for a specified collection of devices, and a distribution reliability coordinator (DRC) coordinating all distribution circuits.

Market and frequency data is used from the PJM and WECC systems. The WECC system already has a significant proportion of renewable energy resources. For example, on March 5, 2018, 49% of California demand was served by solar) [2]. To meet the requirements of the ARPA-E program, 50% penetration of renewables is needed and the control system must be adaptable to increasing penetrations of primarily solar PV. Distributed PV was added to each circuit to ensure the system was representative of 50% penetration of renewables.

6.2 Large-Scale Response Simulations

This section covers the non-federated experiments where only simulated assets at PNNL are engaged and were actively controlled. For this experiment we are simulating a total of 24 hours of operation. During the time three regulation events happen, one from each category. The details of the events are described below.

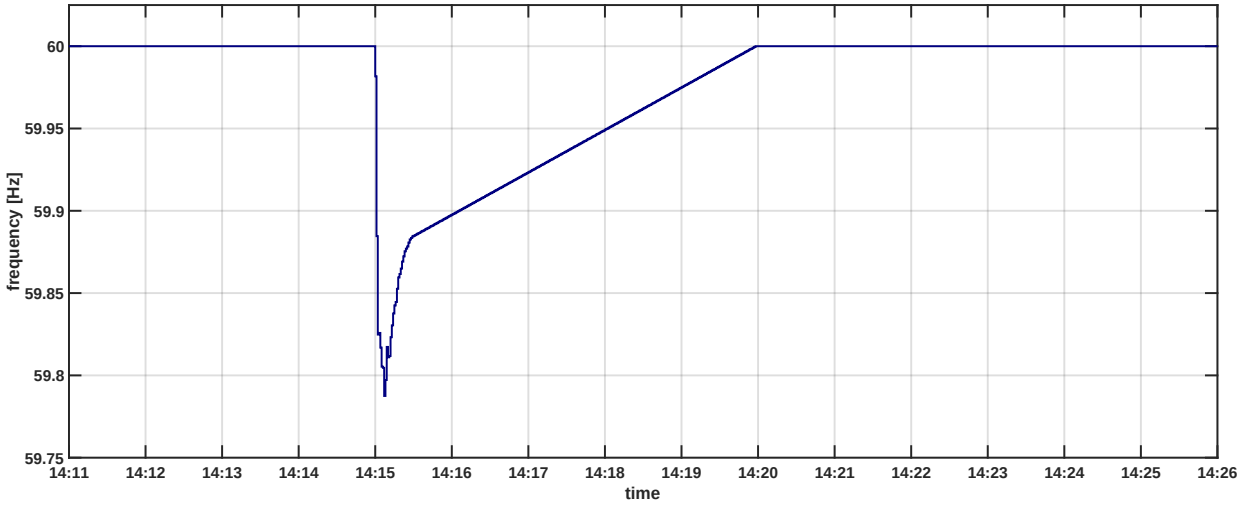


Figure 46. Frequency droop event

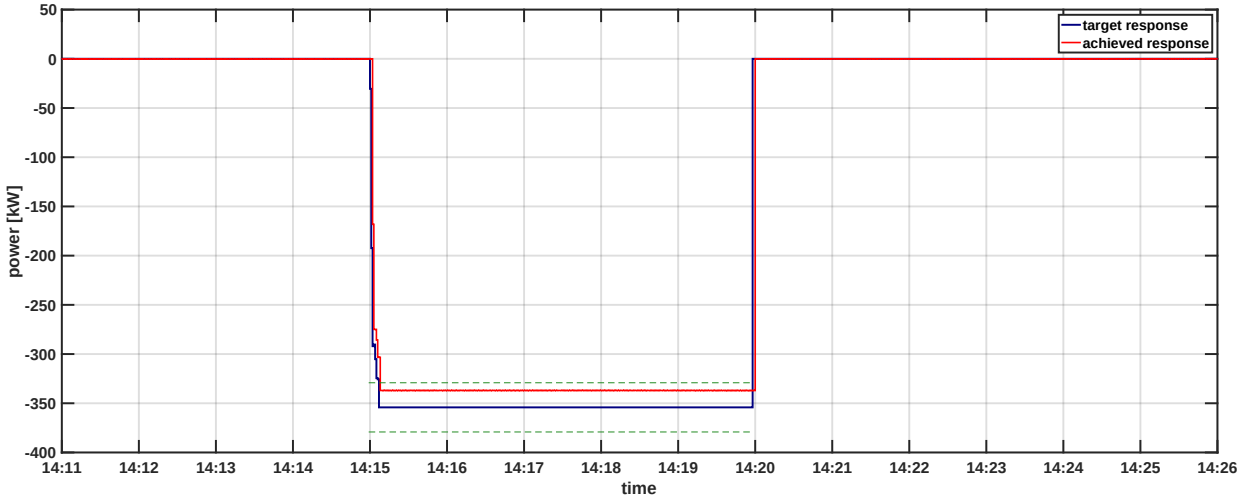


Figure 47. Category I: Frequency droop event response (requirement bounds are shown with green dashed lines)

6.2.1 Category I: Frequency response results

For Category I the ISO is setup to request a total of 500 kW of response capacity to be available from 14:00 to 14:30. During this event the ISO is requesting that this amount of services should be provide over the following frequency ranges:

1. Over-frequency range [60.3Hz, 60.005Hz]
2. Under-frequency range [59.7Hz, 59.995Hz]

For this request the DRC allocates the service request evenly among all aggregators. The allotted amount for this service is 41.6 kW. For this experiment the frequency event is planned to happen at 14:15, and mimics a large generation loss. The event is simulated using PowerWorld and the WECC planning model. The event used is depicted in Figure 46.

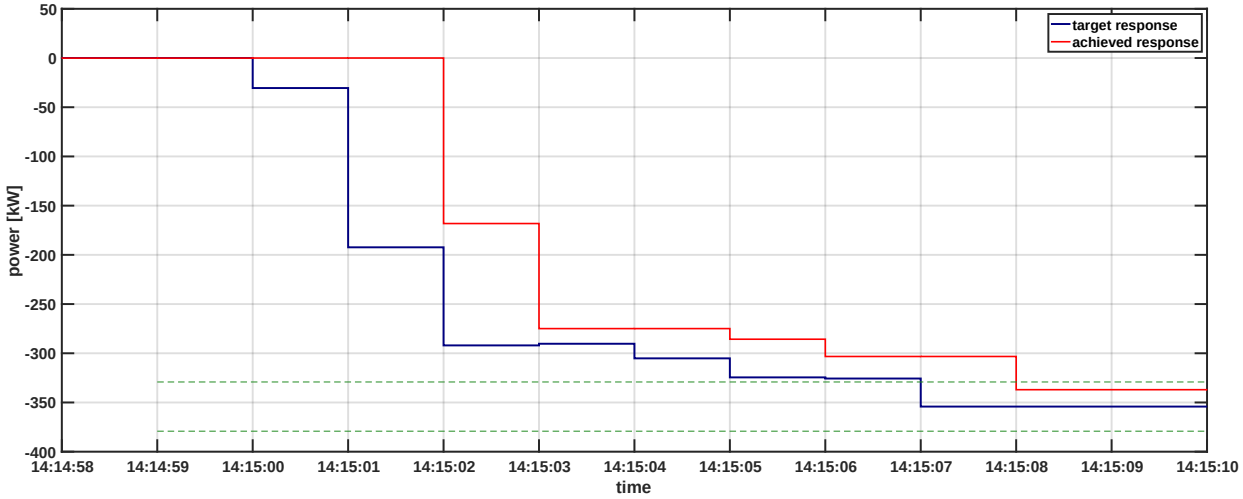


Figure 48. Enlarged view of the Category I frequency response

With the event defined the results from the experiment are shown in Figure 47. In this figure the blue line prescribes the intended response from all of the aggregators. This signal is calculated based on the service capacity request along with the actual frequency event. The red curve is the achieved response calculated based on the devices subscribed for the service. Finally green shows the RMVT error limits. It is clear that the response meets the magnitude and RMVT error limit metrics. Furthermore, Figure 48 shows a zoomed in view of the beginning of the frequency event demonstrating that the initial response time and ramp time metrics are satisfied. All the metrics for Category I are summarized in Table 17.

Table 17. Category I: Frequency response event metrics

Metric	Target	Target Met	Achieved Performance
Initial response time	< 2 seconds	YES	2 seconds
RMT	> 2%	YES	Tested at 2.7%
RMVT	> +/-5%	YES	< 5%
Ramp time	< 8 seconds	YES	2 seconds
Duration	> 30 seconds	YES	Tested at 5 minutes
Availability	> 95%	YES	> 95%

6.2.2 Category II: Regulation results

For Category II the ISO is setup to request a total of 1,000 kW of regulation capacity to be available from 14:00 to 14:30. For this request the DRC allocates evenly among all aggregators. The allotted amount for this service per aggregator is 83.3kW. The regulation event is collected from PJM historical data and depicted in Figure 49.

With the event defined the results from the experiment is shown in Figure 50. In this figure the blue line prescribes the intended response from all of the aggregators. This signal is calculated

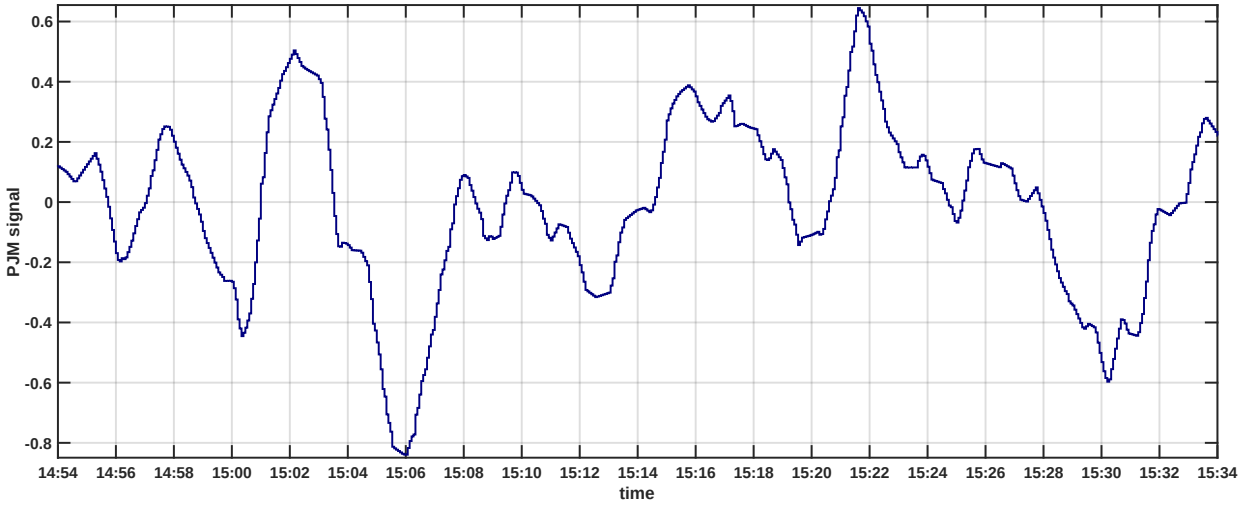


Figure 49. Category II: Regulation event signal

based on the service capacity request along with the actual regulation event. The red curve is the achieved response by calculated based on the devices subscribed for the service.

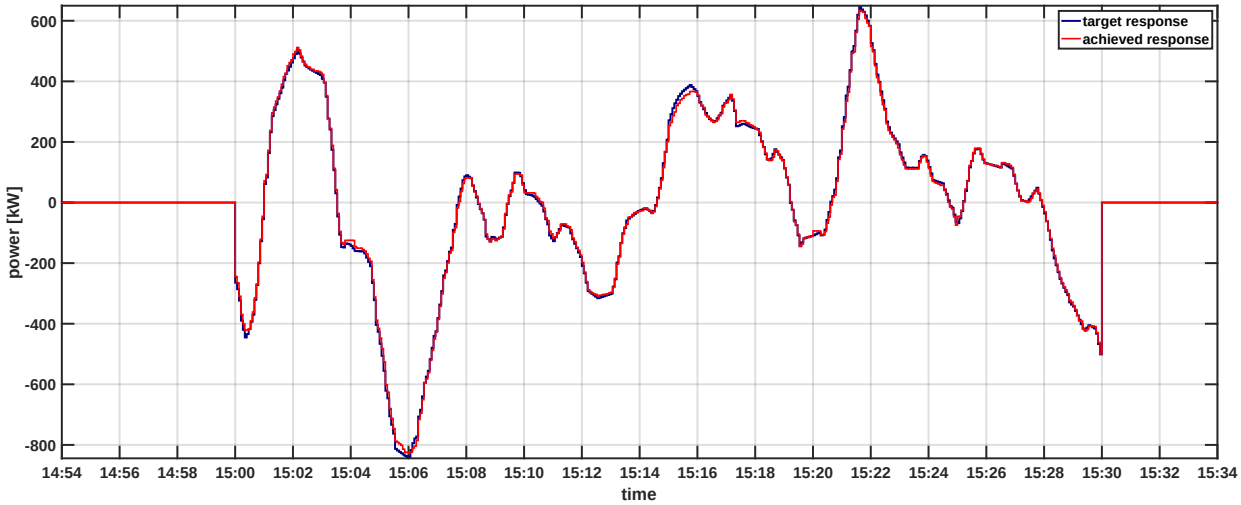


Figure 50. Category II: Regulation event response

It is clear from Figure 50 that we achieve near perfect tracking and are able to met all metrics for this service category. Furthermore, all the metrics for Category II are summarized in Table 18.

6.2.3 Category III: ramping results

For category III the ISO is setup to request a total of 5,000 kW of regulation capacity to be available from 12:00 to 15:00. For this request the DRC allocated request amongst the 12 aggregators (one for each feeder) as shown in Table 19.

The results for the ramping service are shown in Figure 51. In this figure the blue line prescribes the intended response from all of the aggregators. This signal is calculated based on the

Table 18. Category II: Regulation event metrics

Metric	Target	Target Met	Achieved Performance
Initial response time	< 5 seconds	YES	1 seconds
RMT	> 5%	YES	Tested at 5.5%
RMVT	> +/-5%	YES	< 5%
Ramp time	< 8 seconds	YES	1 seconds
Duration	> 30 seconds	YES	Tested at 30 minutes
Availability	> 95%	YES	> 95%

Table 19. Category III: Ramping DRC allocations

Time	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
12:00	186.0	358.3	358.3	186.0	186.0	186.0	-177.5	0.0	0.0	0.0	358.3	358.3
12:05	120.0	234.8	234.8	120.0	120.0	120.0	293.4	228.6	58.2	0.0	234.8	234.8
12:10	120.0	234.8	234.8	120.0	120.0	120.0	145.0	145.0	145.0	145.0	234.8	234.8
⋮												
14:55	120.0	234.8	234.8	120.0	120.0	120.0	145.0	145.0	145.0	145.0	234.8	234.8

service capacity request along with the actual regulation event. The red curve is the achieved response by calculated based on the devices subscribed for the service. Finally green shows the RMVT error limits.

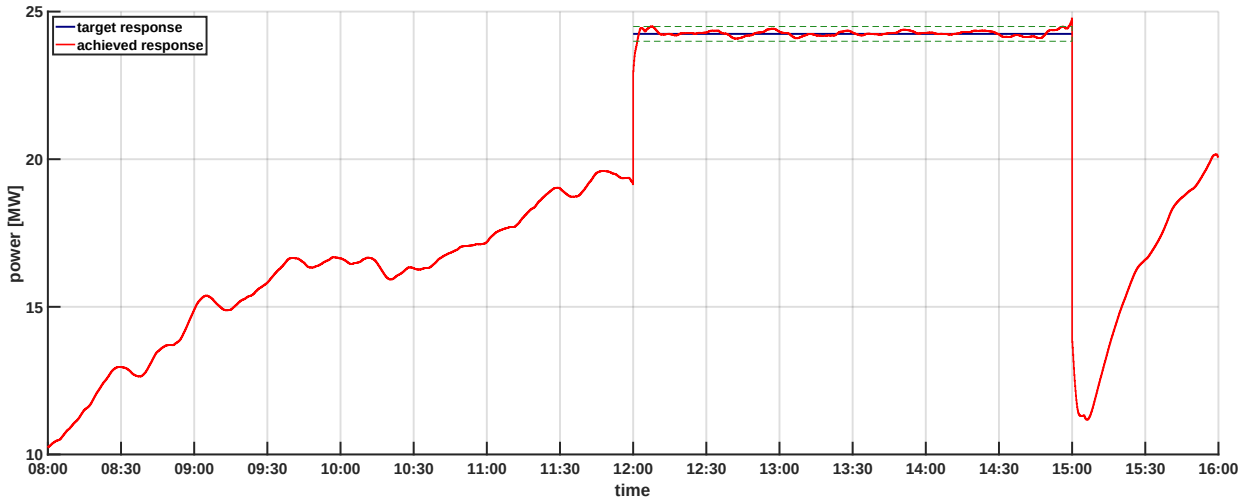


Figure 51. Category III: Ramping event response

It is clear from Figure 51 that we get near perfect tracking and are able to met all metrics for this service category. All the metrics for Category II are summarized in Table 20.

Table 20. Category III: Ramping event metrics

Metric	Target	Target Met	Achieved Performance
Initial response time	< 10 minutes	YES	< 1 minute
RMT	> 10%	YES	Tested at 27.5%
RMVT	> +/-5%	YES	< 5% (10 minute average)
Ramp time	< 30 minutes	YES	< 10 minutes
Duration	> 3 hours	YES	Tested at 3 hours
Availability	> 95%	YES	> 95%
Availability	< 4 hours	YES	> Tested at 3 hours

6.3 Federated Hardware-in-the-Loop Experiments

For the federated joint experiment physical and simulated devices are combined to responds to two events, one from Category I and one from Category II. The experiment is setup to run for one hour and includes a single simulated feeder with 750 controllable devices and physical devices as described in Section 5.3. The feeder selected for this experiment is identical to feeder 1 from the non-federated experiment and in an effort to not duplicate results this section will focus on responses from the physical devices.

In this experiment as mentioned two Categories were evaluated. First, the ISO requests 250kW of Category I Frequency response capacity 10 minutes into experiment and then 100kW of Category II Regulation response capacity 20 minutes into experiment. The following allocations are the outcome of the DRC for the Frequency response event:

- 125kW from PNNL
- 125kW from Spirae

For the Frequency Regulation response event the following it the outcome of the DRC:

- 98.86kW from PNNL
- 1.14kW from UTRC

6.3.1 Category I: Frequency response results

For Category I the ISO is setup to request a total of 250kW of response capacity to be available from minute 10 to minute 40 in the experiment. During this event the ISO is requesting that this amount of service should be provide over the following frequency ranges:

1. Over-frequency range [60.3Hz, 60.005Hz]
2. Under-frequency range [59.8Hz, 59.995Hz]

For this request the DRC allocates evenly among all aggregators. The allotted amount for this service is 125kW. For this experiment the frequency event is planned to happen at 15 minutes into

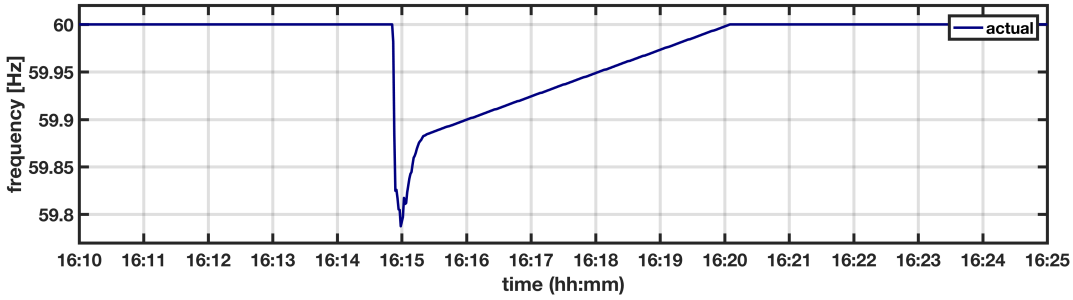


Figure 52. HiL Testing: Frequency droop event used to evaluate the Category I frequency response

the experiment, and mimics a large generation loss. The event is simulated using PowerWorld and the WECC planning model. The event used is depicted in Figure 52.

With the event defined the results from the experiment are shown in Figure 53. In this figure the blue line prescribes the intended response from all of the aggregators. This signal is calculated based on service capacity request along with the actual frequency event. The red curve is the achieved response calculated based on the devices subscribed for the service. Finally green shows the RMVT error limits.

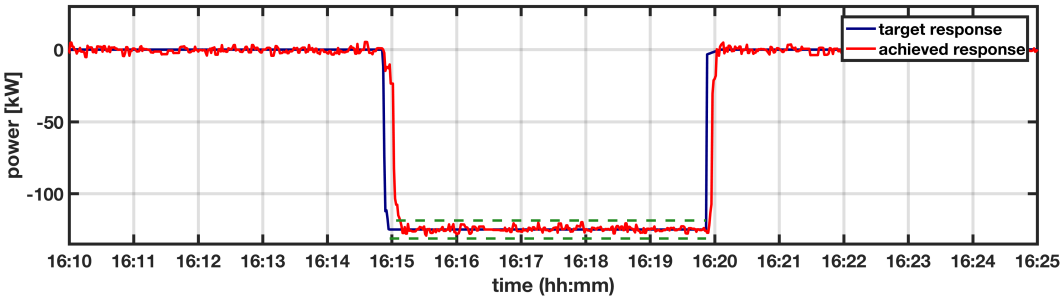


Figure 53. HiL Testing: Catergory I frequency droop response

Figure 53 shows a good overview of the response and that the service magnitude, RMVT, and duration are all met. , However, Figure 54 shows a zoomed in view around the beginning of the frequency event showing that not all the response and ramping requirements are met. The initial response time is met but the ramp time requirement is not met at all times. The requirement is set at less than 8 seconds but for this experiment we have a worst case ramp time of 12 seconds. All the metrics for Category I are summarized in Table 21.

6.3.2 Category II: Regulation reserve results

For category II the ISO is setup to request a total of 100kW of response capacity to be available from minute 20 to minute 50 in the experiment. For this experiment frequency regulating control logic was implemented in ALC Eikon Logic code on the ventilation fan controllers at UTRC. A snapshot of the logic in WebCTRL is shown in Figure 55 and the tracking performance of the controller is provided in Figure 56 and Tables 22 and 23. The experimental testing against RegD PJM signal of the refined controller meets all the the NODES targets. In particular the resulting 2.7% RMVT lies within the $\pm 5\%$ target.

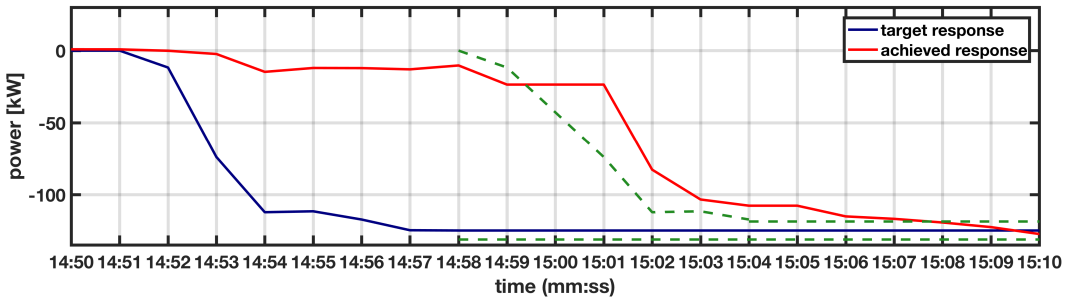


Figure 54. Enlarged view of Figure 53 showing the that the frequency droop response does not meet the 8 second ramp time requirement (shown with a dashed green line)

Table 21. Category I: Frequency Response event metrics

Metric	Target	Target Met	Achieved Performance
Initial response time	< 2 seconds	YES	2 seconds
RMT	> 2%	YES	Tested at 48%
RMVT	> +/-5%	YES	< 5%
Ramp time	< 8 seconds	NO	< 12 seconds
Duration	> 30 seconds	YES	Tested at 5 minutes
Availability	> 95%	YES	> 95%

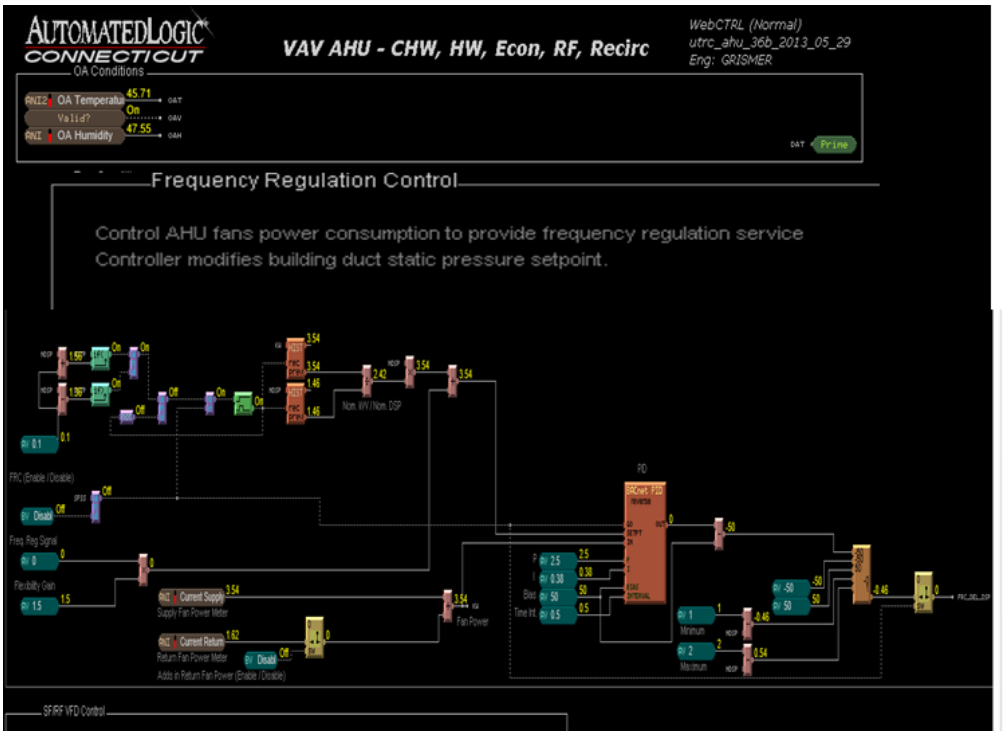


Figure 55. Frequency Regulation Control Implementation in WebCTRL

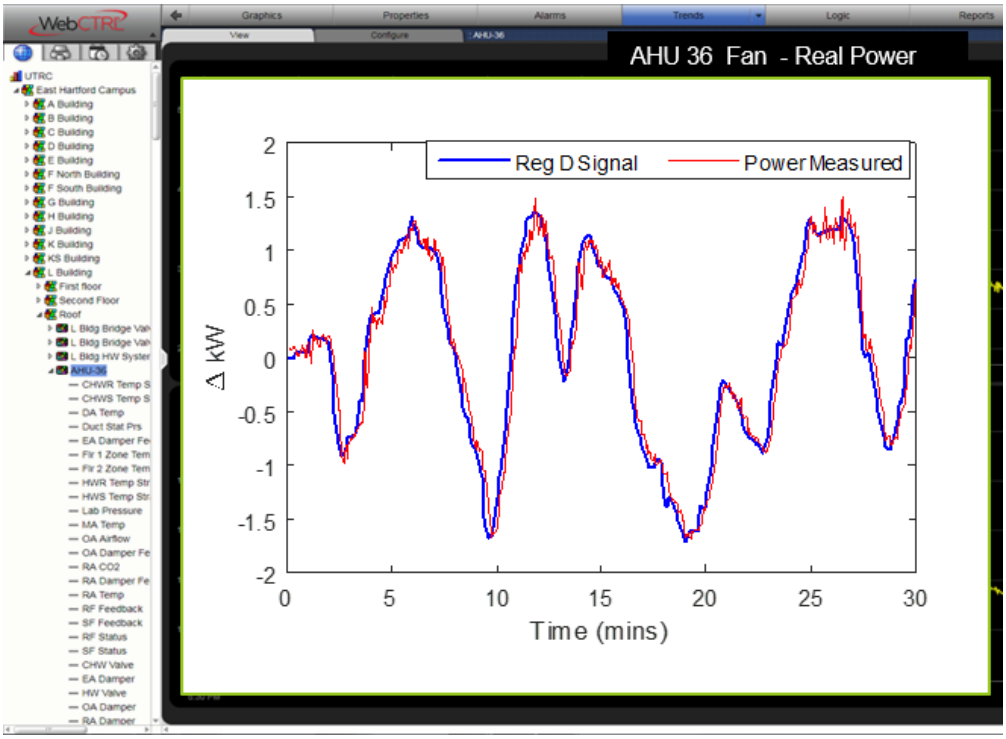


Figure 56. Experiment results for frequency regulation controller with re-tuned Eikon Logic Controller

Table 22. Category II: Frequency regulation experimental results against NODES performance metrics

Metric	Target	Target Met	RegD signal
Initial Response Time	< 5 sec	YES	< 4 sec
Reserve Mag. Target	$> 5\%$	YES	35%
RMVT	$\pm 5\%$	YES	2.7%
Ramp time	$<$5 mins	YES	< 1 min
Duration	$>$30 mins	YES	$>$30 mins
Availability	95%	YES	$> 95\%$

Table 23. Improved Frequency Regulation experimental results against PJM performance metrics

Test signal	Correlation score	Delay score	Precision score	Composite (mean) score
Reg D	0.99	0.99	0.81	0.93

7.0 Technology-to-Market Strategy and Outreach

This section describes how the proposed technology is expected to transition from its current stage of development to ultimate market deployment. This strategy draws from engagement with key industry stakeholders as well as wider dissemination of project results with key communities of practice.

7.1 Industry Engagement and Outreach

7.1.1 Industrial Advisory Board

An Industry Advisory Board (IAB) was formed and engaged throughout the project to get feedback on project risks and potential commercialization barriers. The IAB represented all levels of grid operation including Independent System Operators (ISO) and Regional Transmission Organizations (RTOs), Investor Owned Utilities (IOU), and Original Equipment Manufacturers (OEMs) of DER controllers. The IAB included representation from:

- CAISO (ISO)
- PJM (RTO)
- Exelon (IOU)
- Southern California Edison (IOU)
- Automated Logic Corporation (OEM)
- Spirae (OEM)

7.1.2 Other Engagement and Dissemination

The project partners' presence at trade shows, conferences, and user-group meetings was leveraged to publicize the project including the control algorithms, general methods applied, and test-bed capabilities. Project team members attended and presented at major industry conferences, such as the IEEE PES General Meeting , American Control Conference [31], IEEE Conference of Control Technology and Application [32], Resilience Week [33] and the High Performance Building Conference [11]. Spirae has also be active promoting this area supporting industry webinars and white-papers [34].

As part of these meetings and engagements key items were highlighted: (1) the application of the methods employed; (2) the open-source aspect of the test-bed work; and 3) the performance levels and risk reduction achieved through field demonstrations. This will be essential in developing an ecosystem of stakeholders necessary to ensure that the developed technology provides a persistent technical and economic benefit to the nation.

7.2 Commercialization Strategy

This section details the latest technology trends for grid services, how the NODES technology is differentiated, identified market and commercialization barriers and potential strategies to address these.

7.2.1 Competitive Landscape

There are competitive threats from both utility-scale solutions as well as competing DER approaches:

Utility Scale Approaches: Grid services can also be provided by integrated storage solutions (e.g. battery systems). Utility-scale solar plus storage and dedicated storage solutions have substantially dropped in price and are showing economic benefits to the bulk energy market. For example, Hawaiian Electric Companies has received contracts for multi MW solar plus storage facilities with costs as low as 8 cents/kW-hr [35]. In addition, an early deployment of Tesla battery packs in South Australia has shown their potential for grid services even when not tied to a local generation site. In this case the Hornsdale Power Reserve (HPR), owned and operated by Neoen is the world's largest lithium-ion battery energy storage system, with a discharge capacity of 100 MW and energy storage capacity of 129 MWh. The introduction of the HPR has contributed to removing the need for a 35 MW local (Frequency Control Ancillary Service) FCAS minimum constraint – which was estimated to have added nearly AUD 40 million in Regulation FCAS costs in both 2016 and 2017 [36].

The use of coordinated residential and building equipment DERs is differentiated from utility scale solutions (e.g solar plus storage) in several ways. First, because the approach targets existing devices it offers a substantially lower capital cost, reduced site planning and approval process, and potentially superior proximity to T&D constraints. The use of existing DERs does run the risk of high customer (device) acquisition and enrollment costs on a per kW and kW-hr basis. This can be addressed by ensuring the control technology is implemented on new devices in the factory or by patching equipment in the field (see for example the patching of Hawaii solar inverters [37]). Customer enrollment can be achieved through existing distribution utility customer relationships (potentially in partnership with an aggregator).

Alternative DER Approaches: IAB interviews (with PJM) identified competing approaches at the device level. In the distributed device market Mosaic Power is providing frequency regulation with Grid Interactive Water Heaters (GIWH) in PJM's territory. This is the only active aggregator effort on residential devices for regulating reserve that PJM is aware of. ComEd has a residential thermostat program with Nest (as aggregator) and PJM. SCE's SemiAnnual Report on their Demand Response Emerging Technologies Program [38] also provides a good summary of demand response technologies SCE is evaluating.

An additional emerging risk is that classes of devices will be regulated to have autonomous grid services, for example IEEE std 1547-2018 requiring frequency response as part of the new smart inverter standard. Depending on how prolific these devices become (and how much need there is for grid services in the future smart grid) there may be a significantly diminished need for services from other systems and the associated finacial incentives.

The NODES approach is differentiated from existing DER aggregation approaches in several key ways. First, the use of the virtual battery approach enables device contributions to be abstracted, enabling a heterogeneous mix of devices. Second, the distribution reliability coordinator (DRC) ensures that sufficient flexibility reserves are maintained across the population to ensure that frequency and regulation services can be provided in addition to ramping services. Being able to provide all three grid services from a single population of heterogeneous devices (assuming individual devices are sufficiently responsive) is a differentiating feature. Third, the development of estimators for each DER ensures that their individual flexibility contributions can be committed and delivered without any impact on the quality of service (e.g. comfort, power availability) to the end customer.

7.2.2 Potential Market Barriers and Risks

Performance Reimbursement Risks: Several IAB members (Exelon and Cal-ISO) identified the risk that the quality of the grid services may deteriorate as they are scaled to more devices and implemented in a production environment. This would result in a key risk of diminished compensation if you do not follow the signal closely and in a worst-case scenario if you do not meet an accuracy threshold you get disqualified and must re-enroll. This is an important factor as scaling up the DRC approach may introduce additional uncertainty or latency that may erode performance and therefore compensation. Furthermore, distribution utilities are unlikely to financially penalize retail customers if their devices do not perform as expected. Therefore, since the utilities bare the non-performance risk, they will need to account for margin due to a reduced participation rate or poor device performance. How to successfully estimate and factor in this risk is still an open question.

Split Incentive Risks: A key market adoption risk is that the value accrual from this technology is not aligned with the required investment. For example, it is likely that the required investment (in technology and customer acquisition and enrollment) will occur in large part at the distribution utility level and will need to be approved by Public Utilities Commission. This approval can be challenging for direct load control applications. However, the resulting benefits (and value accrual) may occur at the bulk market level. Based on IAB feedback (Exelon) a clear value proposition will need to be shown for rate-payers to get regulator approval for investment in such a technology by distribution utilities.

Mandated Frequency Response Solutions: Based on IAB feedback (PJM) it is not clear if frequency response services will be procured through a market system or simply mandated from devices (e.g. smart inverters). As such a key recommendation is to focus on regulating and ramping services as the first toe-hold market.

Market Access Risks: Regulations may prohibit the services devices can provide and the pricing available to them. For example, in some markets (Cal-ISO) behind the meter DER's participating in wholesale Demand Response are limited to load curtailment and must be aggregated at the utility facility meter level. (This excludes Frequency Regulation services.) Furthermore, behind-the-meter DERs participating in wholesale markets outside of Demand Response, are subject to retail electricity rates despite bidding into a whole-sale energy market. This can put behind-the-meter approaches at an economic disadvantage.

Additional functionalities (such as M&V and communication) will need to be matured as well. Based on IAB feedback (PJM) there will be a need for a clear Measurement and Verification (M&V) process for distributed resources (especially at the residential level where it is hard to justify the expense of utility grade meters, or interval meters that provide the granularity needed to verify Regulation service, on such small loads). Could leverage or extend PJM's current sampling method [39]. For communication, how behind the meter assets communicate to advanced distribution management systems (DMS) needs to be determined. While M&V and communication approaches are outside the scope of this ARPA-E program PNNL can leverage effort from the Grid Modernization Lab Consortium (GMLC) in these areas.

7.2.3 Commercial Building Deployment

The NODES project technical work was also reviewed with Automated Logic Corporation on multiple occasions. ALC's WebCTRL product includes OpenADR integration to allow for the deployment of application-engineered building control solutions to meet various grid service needs (e.g. load shedding) in various markets. These discussions included representation from ALC's product management and development organizations. Informal discussions were also held with

representatives from several ALC branches.

With some exceptions, building control solutions (e.g. control methodologies, sequence of operations) are developed, delivered, and verified to meet a controls requirement provided by a specifying engineer hired by the building owner or developer. Therefore, broad market adoption of advanced building control solutions requires educating specifying and consulting engineers as to the features and benefits of these new approaches and enabling seamless delivery. There are numerous best practices to achieve this including: dissemination of field demonstration results with early adoption customers and developing and publishing standard implementations to make specifying such solutions easier. This is typically done through industry organizations such as ASHRAE. Even with well documented approaches there can still be ambiguity about the details of such solutions leading to differences and errors in final implementation. This can be addressed by developing and offering standard control block libraries to be utilized for a standardized implementation.

An exemplary example of this is the development of Guideline 36 by ASHRAE. Guideline 36: 'High-Performance Sequences of Operation for HVAC Systems', *"provides uniform sequences of operation for HVAC systems that are intended to…standardized advanced control sequences providing benefits including: reduced engineering time, reduced programming and commissioning time, and a common set of terms to facilitate communication between specifiers, contractors and operators"* [40]. At the end of this initiative BMS vendors such as ALC will have standard library blocks compliant with Guideline 36 that can be used by consulting engineers and installers so these solutions can be readily specified and implemented.

Based on the above discussion and examples key recommendations for accelerating the adoption of advanced grid-service controls in buildings are provided below:

1. Determine the equipment classes, building types, climate zones, and utility incentive structures that will provide the highest return on investment for this technology. This assessment can be done using simulation environments similar to that used by UTRC to develop and evaluate the control technology.

2. Based on the above findings, conduct field demonstrations (with utility stakeholders) at customer sites. Dissemination of the control performance and benefits will be used to educate the wider buildings and utility communities as to the maturity of building interactive-grid technologies.

3. Develop standardized reference implementations of advanced building-to-grid control solutions to simplify the specification and delivery of these technologies. Such a standardized solution can represent the 'base-model' solution allowing individual solution providers to offer improvements above and beyond this.

7.2.4 Intellectual Property Strategy

The product is a collection of algorithms, both at the aggregator level and at the device level, that define the flexibility of the underlying devices and control those devices to deliver the desired flexibility as a grid service while maintaining customer comfort. The platform, including APIs and data requirements, are intended to be openly available. The aggregation algorithms and device-level algorithms are specific and are open for further commercial development.

PNNL has a significant intellectual property position in transactive control, load control, and demand response, due to its leading role in the Pacific Northwest Smart Grid Demo Project, OlyPen Demo Project, AEP gridSMART program, and ongoing project efforts. Multiple companies have licensed IP in this area. Most of the software to be used, such as FNCS and GridLAB-D, are

available under open source licenses, with the latter having a vibrant user community to further drive adoption. PNNL has also generated new IP and data (see two NODES IP filings below). Algorithms and methods regarding distributed level control have been developed by UTRC and PNNL, while those around wholesale markets and bulk grid simulation involve PNNL and SCE. We will also leverage existing software tools that are available via open source licenses, upon review of the obligations under these particular licenses in the context of our wide-scale adoption and deployment strategy as detailed above. A summary of intellectual property generated to date is provided below:

1. PNNL: IPID 30797-E CIP "EXTRACTING MAXIMAL FREQUENCY RESPONSE POTENTIAL IN CONTROLLABLE LOADS" Patent extension.
2. PNNL: IPID 31133-E: Multi-layer Market-based Framework for Seamless Integration of Distributed Energy Resources, first filed 7/14/2017.
3. PNNL: IPID 31275-E: Priority-based Threshold Allocation for Frequency Response, filed as a Continuation-In-Part to IPID 30797-E 2/1/2018. Also filed in Canada.
4. PNNL: IPID 31337-E: New Control Approach for Power Modulation of End-use Loads, filed 3/21/2018. Also filed in Canada.
5. UTRC: 104376US01 (U301933US) - Control System for Advanced Demand response (adopted for filing)

7.2.5 First Markets

We believe that markets in need of new flexibility resources at low costs (to address the rapid integration of renewable generation) and have a large portion of flexible loads (for example a significant proportion of electrical heating or cooling loads) will be more receptive to this technology. California, Ontario/Quebec, TVA, etc. all have historical interests and openness to controllable distributed resources. Aggregators, whether utility or 3rd party, are a natural market segment.

7.3 Proposed Commercialization Plan

PNNL will use the reference platform definition to encourage industry engagement, by openly sharing the process for control and aggregation. The individual algorithms represent commercial products that may be licensed by the individual participants. An instantiation of the platform and its algorithms will be created as part of the Hardware-in-the-Loop (HiL) experiment and incorporated into a commercial partners' existing product line. The individual DER algorithms to estimate and control flexibility will be demonstrated in production software and controllers and delivered to vendors such as Automated Logic Corporation. The team expects the transition of two levels of DER algorithms: 1) a basic implementation that will be openly provided and can serve the basis of a standardized approach that can be specified by consulting engineers and regulators and 2) advanced versions that contain proprietary IP. This will incentivize providers to develop and deliver innovative approaches that can differentiate their offering and improve performance.

7.3.0.1 Phase 1: Risk Reduction

Proposed near-term efforts focus on development, analysis, and evaluation that address key risks and uncertainties. These include:

1. Innovative grid-interactive DER solutions will not be developed by manufacturers and procured by building owners until the suitability criteria and value proposition has been articulated. There is value in extending the resource-level flexibility characterization methodology to use historical data to evaluate the eligibility of a commercial building (and its HVAC system) to provide flexibility and an estimate of the resulting monetary value to the grid.

2. Uncertainty in DER and DRC performance due to communication latency or end-user participation could substantially erode the quality of service provided and resulting payment. The DRC design should be extended to ensure the robustness of the DRC asset allocation optimization to network-level uncertainties and communication failures or latencies. Furthermore, the power flow of the underlying distribution network and the associated constraints regarding line current and nodal voltage should also be systematically considered.

7.3.0.2 Phase 2: Piloting

The risk reduction above must be followed by additional demonstrations and pilots that integrate the developed solution architecture with a utilities physical system. Initially, a demonstration on a portion of a utility's distribution system leading to a larger-scale pilot project would be warranted. A complete solution will require integration of the algorithm to the platform used to operate the Distribution System (e.g. ADMS). Upon a successful pilot, large scale adoption may be possible by integrating the solution to the ADMS and bridging any communications and cybersecurity concerns.

At the building and micro-grid level, prototype DER controllers would need to be adapted with the control development and integration developed in this project and evaluated at key field demonstration sites. These sites would need to cover a representative range of building and equipment types, operating seasons, and grid events. Key success criteria for such a field demonstration would be demonstration of technology maturity and delivery of gird services (and associated monetary compensation) to warrant commercialization.

7.3.0.3 Phase 3: Standardization and Production

After successful demonstration of the value proposition at the pilot scale the final step would be the implementation of any required regulatory changes to compensate DER grid services, development of standards to ensure accelerated delivery of common DER solutions by manufacturers, and integration of DRC capability into the production environment of utilities.

8.0 Conclusions

This report presents a hierarchical control architecture for allocating and controlling Distributed Energy Resources (DERs) to provide ancillary grid services. This approach has a two important and differentiating features. First, it simultaneously addresses all three grid services targeted by the NODES FOA (frequency response, frequency regulation, and ramping). It does this by aggregating flexibility estimates from participating DERs and then optimally allocating that flexibility to ramping, while ensuring sufficient flexibility is reserved to meet the more stochastic needs of frequency response and regulation. This allocation optimization framework, provided by the Distribution Reliability Coordinator (DRC), can be generalized to other grid services - an important characteristic given the diversity of ISO/RTO ancillary service markets across North America and the likelihood for evolution of gird services in the future. The second key feature of this technology is the ability to seamlessly incorporate a heterogeneous mix of DERs. This project demonstrated application to residential water heaters and air conditioning units, commercial HVAC ventilation fans and chiller plants, and microgrid systems (electrical battery storage, back-up generation, electric vehicles and PVs). This was achieved by the use of a standardized virtual battery representation of the flexibility and services each DER could provide. The use of a common interface for DERs to represent themselves to the grid is an important feature of this architecture. The above two features are critical in ensuring a scalable solution. Scalability is important not just in addressing a large number of DERs (the ability to control 10,000s of devices), but also scalability across a mix of existing and emerging DERs, as well as a solution that scales across a range of participating organizations, from the operators of DERs, to aggregators, to grid operators at both the distribution and bulk transmission scales.

Demonstrating and verifying this approach necessitated the implementation and use of a federated test-bed. The test-bed incorporated state-of-the-art modeling environments as well as real-time communication with partner sites (through the VOLTTRON Fed-in-a-box) to enable co-simulation with production DER controllers operating either real hardware or high-fidelity full-building simulations. This capability enabled the demonstration of scalability to >20,000 devices through co-simulation. The test-bed also enabled demonstration of actual DERs at sites across North America providing real-time flexibility estimates and commensurate services in coordination with the DRC. This verified that the approach can be implemented on production hardware and executed in a distributed fashion. The simulation and hardware experiments showed that this DRC control approach can achieve virtually all the FOA metrics. The only current demonstration short-fall was a 4 second delay on micro-grid frequency response due to communication and hardware latencies.

Successful deployment of this approach and similar technologies will require addressing several additional commercialization barriers. First, sufficient incentives are needed for manufacturers to offer grid-service compatible devices and for operators to adopt them. Better understanding the application attributes that have the highest value will accelerate the identification of appropriate first markets. Second, the impact that uncertainty in key variables such as communication latency, residential user participation, and other factors have on the quality of the provided grid services and financial compensation is important. Successful pilot demonstration of the technical capabilities and value proposition would clear the path to the formalization of standardized DER representations of flexibility (such as the virtual battery approach) in conjunction with associations such as ASHRAE and IEEE. These elements (incentives, field demonstration of technical maturity, and industry standards) will be needed for board adoption by suppliers and operators.

References

[1] ARPAE, Network optimized distributed energy systems (nodes) - funding opportunity announcement (2015).
URL https://arpa-e-foa.energy.gov/FileContent.aspx?FileID=afc43091-76da-4d8c-a4b7-ed7e5fc8cbda

[2] Green, Tech, Media, California sets two new solar records (2018).
URL https://www.greentechmedia.com/articles/read/california-sets-two-new-solar-records#gs.47s96t

[3] Utility, Dive, Hawaii far from 100% renewables — but running ahead of schedule, state finds (2018).
URL https://www.utilitydive.com/news/hawaii-far-from-100-renewables-but-running-ahead-of-527171/

[4] Y. G. Rebours, D. S. Kirschen, M. Trotignon, S. Rossignol, A survey of frequency and voltage control ancillary services—part i: Technical features, Power Systems, IEEE Transactions 22 (1) (2007) 350–357. doi:10.1109/TPWRS.2006.888963.
URL http://www2.econ.iastate.edu/tesfatsi/VoltageControlASPart1.Kirschen2007.IEEEas_tech2007.pdf

[5] PJM, Pjm learning center: Regulation market (2019).
URL https://learn.pjm.com/three-priorities/buying-and-selling-energy/ancillary-services-market/regulation-market.aspx

[6] H. Hao, B. Sanandaji, K. Poolla, V. Tyrone, A generalized battery model of a collection of thermostatically controlled loads for providing ancillary service, in: the 51th Annual Allerton Conference on Communication, Control and Computing, 2013.

[7] H. Hao, B. M. Sanandaji, K. Poolla, T. L. Vincent, Aggregate flexibility of thermostatically controlled loads, IEEE Transactions on Power Systems 30 (1) (2015) 189–198.

[8] J. T. Hughes, A. D. Domínguez-García, K. Poolla, Identification of virtual battery models for flexible loads, IEEE Transactions on Power Systems 31 (6) (2016) 4660–4669.

[9] S. P. Nandanoori, I. Chakraborty, T. Ramachandran, S. Kundu, Identification and validation of virtual battery model for heterogeneous devices, IEEE Power & Energy Society General Meeting (arXiv preprint arXiv:1903.01370).

[10] I. Chakraborty, S. P. Nandanoori, S. Kundu, Virtual battery parameter identification using transfer learning based stacked autoencoder, in: 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018, pp. 1269–1274.

[11] V. A. Adetola, F. Lin, H. M. Reeve, Building flexibility estimation and control for grid ancillary services, in: High Performance Buildings Conference, Purdue Herrick Labs, West Lafayette, IN, 2018.

[12] C. Perfumo, E. Kofman, J. H. Braslavsky, J. K. Ward, Load management: Model-based control of aggregate power for populations of thermostatically controlled loads, Energy Conversion and Management 55 (2012) 36–48.

[13] R. Diao, S. Lu, M. Elizondo, E. Mayhorn, Y. Zhang, N. Samaan, Electric water heater modeling and control strategies for demand response, in: 2012 IEEE Power and Energy Society General Meeting, 2012, pp. 1–8. doi:10.1109/PESGM.2012.6345632.

[14] A. Molina-Garcia, F. Bouffard, D. S. Kirschen, Decentralized demand-side contribution to primary frequency control, IEEE Transactions on Power Systems 26 (1) (2011) 411–419.

[15] J. Lian, J. Hansen, L. D. Marinovici, K. Kalsi, Hierarchical decentralized control strategy for demand-side primary frequency response, in: 2016 IEEE Power and Energy Society General Meeting (PESGM), 2016, pp. 1–5. doi:10.1109/PESGM.2016.7741267.

[16] S. Kundu, J. Hansen, J. Lian, K. Kalsi, Assessment of optimal flexibility in ensemble of frequency responsive loads, in IEEE International Conference on Smart Grid Communication (arXiv preprint arXiv:1707.07033).

[17] S. Ilic, C. Bullard, P. Hrnjak, Effect of shorter compressor on/off cycle times on a/c system performance, Tech. rep., Air Conditioning and Refrigeration Center. College of Engineering … (2001).

[18] W. Zhang, J. Lian, C. Chang, K. Kalsi, Aggregated modeling and control of air conditioning loads for demand response, IEEE Transactions on Power Systems 28 (4) (2013) 4655–4664. doi:10.1109/TPWRS.2013.2266121.

[19] Y. Lin, P. Barooah, S. Meyn, T. Middelkoop, Experimental evaluation of frequency regulation from commercial building hvac systems, IEEE Transactions on Smart Grid 6 (2) (2015) 776–783.

[20] J. L. H. Hao, D. Wu, T.Yang, Optimal coordination of building loads and energy storage for power grid and end user services, IEEE Transactions on Smart Grid.

[21] PJM, Pjm manual 12: balancing operations, revision 37 (2017).
URL http://www.pjm.com/~/media/documents/manuals/m12.ashx

[22] S. Bengea, P. Li, S. Sarkar, S. Vichik, V. Adetola, K. Kang, T. Lovett, L. F., K. A., Fault-tolerant optimal control of a building heating, ventilation, and air conditioning system, Science and Technology for the Built Environment 21 (6) (2015) 734–751.

[23] J. Hansen, T. Edgar, J. Daily, D. Wu, Evaluating transactive controls of integrated transmission and distribution systems using the framework for network co-simulation, in: American Control Conference (ACC), 2017, IEEE, 2017, pp. 4010–4017.

[24] Volttron, Onlline documentation (2018).
URL https://volttron.readthedocs.io/en/develop/

[25] K. P. Schneider, Y. Chen, D. P. Chassin, R. Pratt, D. Engel, S. Thompson, Modern grid initiative-distribution taxonomy final report, Tech. rep., Pacific Northwest National Laboratory (2008).

[26] K. P. Schneider, J. C. Fuller, D. P. Chassin, Multi-state load models for distribution system analysis, IEEE Transactions on Power Systems 26 (4) (2011) 2425–2433. doi:10.1109/TPWRS.2011.2132154.

[27] Z. Taylor, K. Gowri, S. Katipamula, GridLAB-D Technical Support Document: Residential End-Use Module Version 1.0, Tech. rep., Pacific Northwest National Laboratory, PNNL-17694 (2008).

[28] J. Fuller, P. N. Kumar, C. Bonebrake, Evaluation of Representative Smart Grid Investment Grant Project Technologies: Demand Respons, Tech. rep., Pacific Northwest National Laboratory (2012).

[29] R. Pratt, C. Conner, E. Richman, K. Ritland, W. Sandusky, M. Tayler, Description of Electric Energy Use in Single-Family Residences in the Pacific Northwest, Tech. rep., Pacific Northwest National Laboratory, Technical Report for Bonneville Power Administration (1989).

[30] US, Department, of, Transportation, National Household Travel Survey (NHTS, Tech. rep., U.S. Department of Transportation, Federal Highway Administration URL: http://nhts.ornl.gov (2001).

[31] F. Lin, V. Adetola, Flexibility characterization of multi-zone buildings via distributed optimization, american control conference, in: American Control Conference, ACC, Milwaukee, WI, 2018.

[32] S. P. Nandanoori, S. Kundu, D. Vrabie, K. Kalsi, J. Lian, Prioritized threshold allocation for distributed frequency response, in: IEEE Conference on Control Technology and Applications (CCTA), IEEE, Copenhagen, 2018, pp. 237–244.

[33] S. Gourisetti, J. Hansen, W. Hofer, D. Manz, K. Kalsi, J. Fuller, S. Niddodi, H. Kley, C. Clarke, K. Kang, H. M. Reeve, M. Chiodo, J. Bishopric, A cyber secure communication architecture for multi-site hardware-in-the-loop co-simulation of der control, in: Resilience Week, Denver, CO, 2018.

[34] S. Cherian, P. Asmus, Liberating microgrids (and all der) (2016).
URL https://dta0yqvfnusiq.cloudfront.net/spira17343983/2018/02/Navigant-Spirae-Liberating-Microgrids-and-DERs-wp-5a74f48e1b280.pdf

[35] Energy, Manager, Today, Grid-scale hawaiian solar projects offer record low fuel prices (2019).
URL https://www.energymanagertoday.com/hawaiian-solar-projects-0180962/

[36] Aurecon, Hornsdale power reserve: Year 1 technical and market impact case study (2019).
URL https://www.scribd.com/document/395050069/Aurecon-Hornsdale-Power-Reserve-Impact-Stud
fullscreen&from_embed

[37] I. Spectrum, 800,000 microinverters remotely retrofitted on oahu—in one day (2015).
URL https://spectrum.ieee.org/energywise/green-tech/solar/in-one-day-800000-microinverters-remotely-retrofitted-on-oahu

[38] SCE, Demand response emerging markets and technologies program: Semi-annual report: Q3 – q4 2017 (2018).
URL http://www3.sce.com/sscc/law/dis/dbattach5e.nsf/0/DD5EFFA730C0C91388258263007E1B51/$FILE/A1103001_R1309011-SCE%20EMT%20Semi-Annual%20Q3-Q4%202017%20Cover%20and%20FINAL%20Rprt.pdf

[39] PJM, Pjm manual 19: Load forecasting and analysis: Revision: 33 - attachment c (2018).
URL https://www.pjm.com/-/media/documents/manuals/m19.ashx

[40] ASHRAE, Ashrae guideline 36-2018: High-performance sequences of operation for hvac systems (2018).
URL https://www.ashrae.org/news/esociety/new-guideline-on-standardized-advanced-sequences

[41] Z. Xu, R. Diao, S. Lu, J. Lian, Y. Zhang, Modeling of electric water heaters for demand response: a baseline pde model, IEEE Transactions on Smart Grid 5 (5) (2014) 2203–2210.

Appendix A – Fed-in-a-box

Fed-in-a-box is a virtual machine that has pre-scripted Open-ssh scripts that enables it to connect to a client or a server. The user would give in the credentials to establish ssh (layer) tunnel. Although the (two) sites are on different networks, fed-in-a-box bridges the machines on these two sites (bridged tunnel) and makes it look like they are on same network. This means, fed-in-a-box has two interfaces: one interface faces the (lab) while the other interface faces the world-wide web and the VPN bridge allows a layer 3 communication between the two sites/interfaces i.e., LAN on PNNL can talk to LAN at anywhere on collaborator's network through a secure connection that bridges those two sites and allows shared communications over layer 3 or above. In this case, fed-in-a-box is between the VOLTTRON agent at remote location and the PNNL firewall. Fed-in-the-box architecture comes with several advantages:

1. At an organizational level, connecting the systems or receiving traffic from external networks and entities is restricted. Especially, at PNNL the IT-Cybersecurity office strictly prohibits data routing from an external entity. Through Fed-in-the box approach, the VPN bridge connects selected/defined external devices to a PNNL computer over a private connection. This way, neither of the sites is exposed to each other. Therefore, during the data exchange, both the sites (external entities and PNNL) are secure from a cybersecurity perspective;

2. Multiple tests have been conducted using the Fed-in-the-box approach and using this architecture has been proven to handle hardware-in-the-loop connections;

3. As part of the fed-in-the-box development conducted at PNNL, several cybersecurity aspects and requirements were considered and the systems are closely monitored in a tight isolated network. Since the data is routed through a VPN tunnel, the data is already encrypted during the transfer/exchange;

4. This architecture, in the past, has proven to be efficient and fast in exchanging data at higher sampling frequencies (even at less than sub-second sampling time). This provides great flexibility to the NODES project to perform both fast and slow experiment.

A.1 PNNL to External Collaborator Connection

The HIL co-simulation platform to enable such connection comprises of two main sections: 1) Connections inside PNNL; 2) Connections entering PNNL. It is important to understand how an external collaborator such as Spirae or UTRC will be able to interact with PNNL resources. A conceptual diagram of this connection is shown in Fig. A.1.

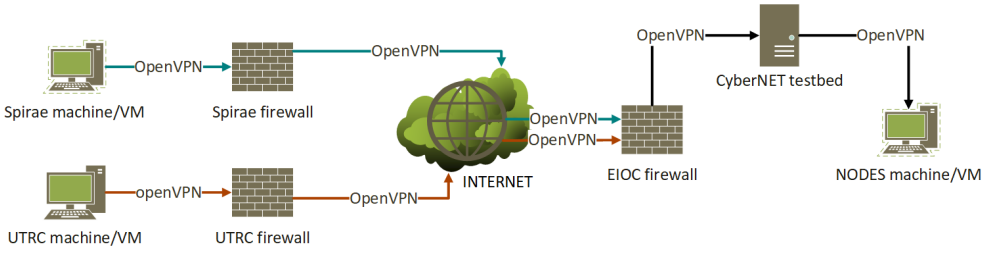


Figure A.1. Conceptual overview of the connection from UTRC and Spirae to PNNL (NODES machine)

As show, all the computers/Virtual Machines (VMs) from the external network that are allowed VPN access would be able to connect to the NODES machine at PNNL (through that VPN tunnel). This federation (fed-in-the-box) setup establishes a bridge between the external machine and the NODES machine on a private network. The traffic is routed via the internet, through the cyberNET server (compute domain at PNNL) to reach the NODES machine. Such a fed-in-the-box architecture can be implemented at both layer-2 (data-link layer) and at layer-3 (network layer). The capability of connecting the external devices on both layer-2 and layer-3 level is important because devices such as smart meters cannot connect on layer-3. Layer-3 connection is the most viable connection when the external device, such as a computer, has two Network Interface Cards (NICs). Through the fed-in-the-box framework both types of connections are possible and can be deployed depending on the need of the external collaborator. A detailed illustration of layer-2 connection is shown in Fig. A.2.

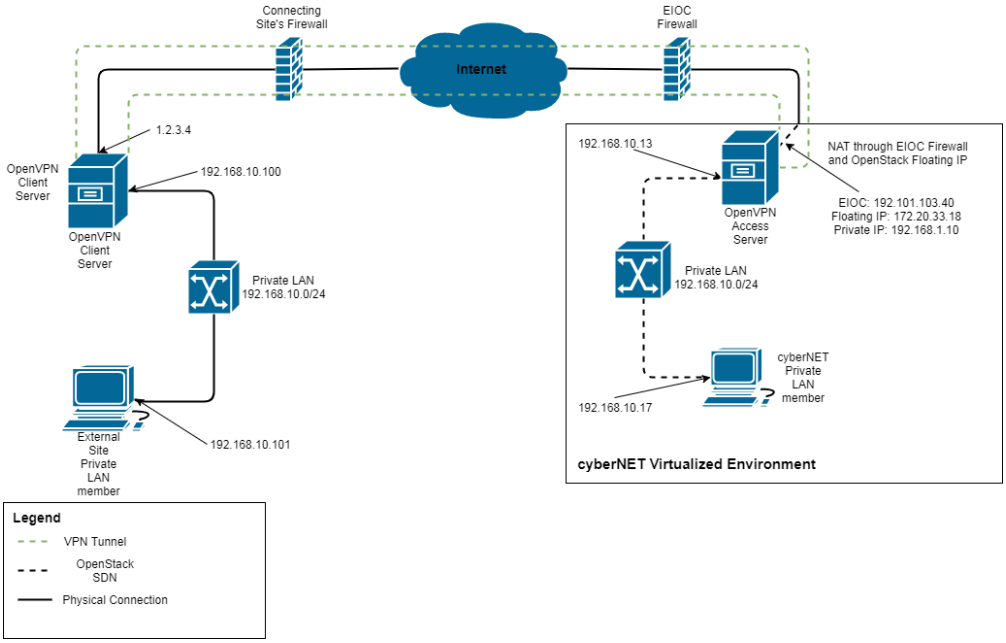


Figure A.2. An illustrative overview of Fed-in-the box connection in the use-case (under testing)

As part of the initial testing VOLTTRON and FNCS were installed on the NODES machine and the two connect the VOLTTRON agent and FNCS to exchange data. This test was performed successfully and later sections of the document is geared to provide detailed overview of those tests. The NODES machine is also established in the CyberNET testbed. CyberNET testbed is located at PNNL that comprises of high performance computing resources. The testbed is located on an independent network isolating itself and the devices connected to it from the PNNL main network. Importing the NODES machine on CyberNET is advantageous because:

1. this would minimize the number of firewall hops for the external traffic to reach the NODES machine;

2. The testbed has connection rules and cybersecurity policies that maintains the integrity of the co-simulation;

3. Federation server is part of the testbed which makes it very efficient to connect the NODES machine. Along with this prototype test to exchange traffic from an external network to the NODES machine have been performed.

These steps have been extensively tested and achieved the seamless dataflow connections successfully. In the prototype tests performed, an external laptop is used to connect to the NODES machine and route the traffic. Three tests were performed and the latency for each test was $10mSec$, $100mSec$, $10mSec$, respectively. Those results strongly indicate that a HIL co-simulation with external devices can be achieved with minimal to no discrepancies. Currently, the team is performing direct connection and data exchange tests with VOLTTRON and FNCS in the loop with the NODES machine (on cyberNET) and a machine/VM on an external network. Currently, the external computers/VMs can be connected through layer-2 or layer-3 connection.

A.1.0.1 UTRC-to-PNNL Connection

In the case of UTRC, BACnet connection is established between UTRC's Building Automation System (BAS) and a commercial building. The BAS is connected to a VOLTTRON agent that is on the same network. The BAS would send the flexibility parameters to their VOLTTRON agent. Those flexibility parameters represent the current state and flexibility of the commercial building. Those parameters are communicated back to the PNNL simulators through a PNNL-networked VOLTTRON agent as illustrated in Fig. A.3. Once the simulators located at PNNL receive the flexibility parameter, they would run the power flow and optimization algorithm to generate a power profile and sends it to UTRC's BAS through the intermediary VOLTTRON agents (see Fig. 42 and Fig. A.3). The dispatched power profile includes Settings and set points, market (sampled at 10 minutes) and control signals (regulation signal sampled at 2 seconds; frequency response signals).

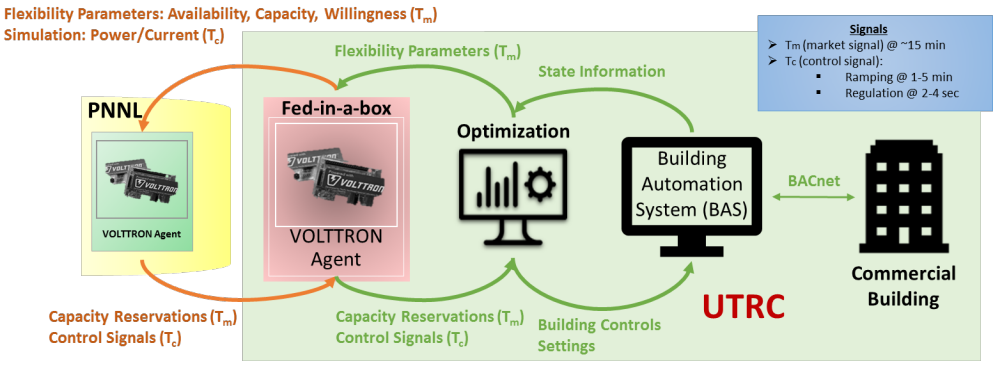


Figure A.3. Full-duplex connection between UTRC and PNNL

A.1.0.2 Spirae-to-PNNL Connection

Similarly, a web protocol connection (formerly this was using OpenFMB, but was down scoped) is established between Spirae's Wave platform and microgrid devices. In this case, there are three independent microgrids connected to three independent Wave platforms. Each of those Wave platforms are connected to independent VOLTTRON agents, located at Spirae's facilities. All the hardware systems and VOLTTRON agents are under the hood of Spirae's internal network. Each individual VOLTTRON agent receives the virtual battery model parameters representing the current state and flexibility of the microgrids. Those flexibility parameters are communicated back to the PNNL simulators through a PNNL-networked VOLTTRON agent as illustrated in Figure 6. Note that only one instance of hardware connections is shown in Fig. A.4 as the remaining two are exact replicas of what is depicted here. Once the simulators located at PNNL receive the

flexibility parameter, they would run the power flow and optimization algorithm to generate a power profile and sends it to Spirae's Wave platform through the intermediary VOLTTRON agents (see Fig. 42 and Fig. A.4). The dispatched power profile includes Settings and setpoints, market and control signals (regulation signal sampled at 2 seconds; ramp signal sampled at 1 minute).

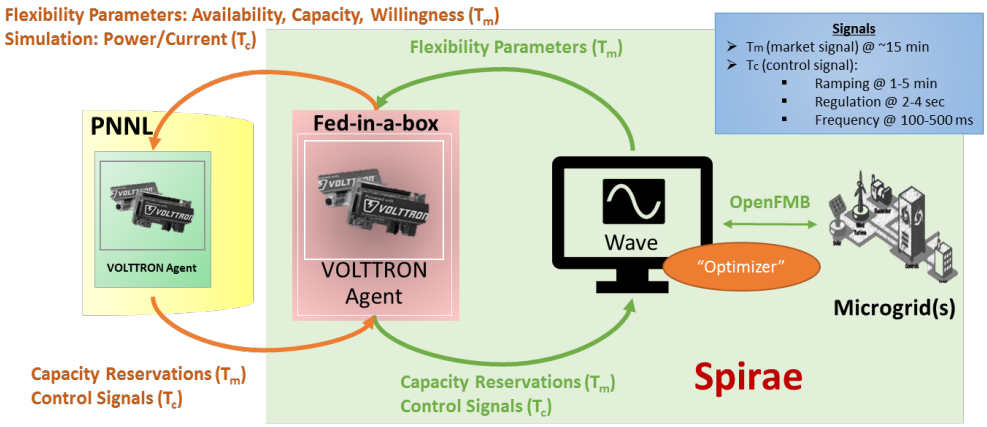


Figure A.4. Full-duplex connection between Spirae and PNNL

A.2 A Brief Overview of VOLTTRON FNCS Bridge

VOLTTRON FNCS Bridge is essentially a FNCS client and a VOLTTRON agent. Through the FNCS bridge, the VOLTTRON agent is connected to simulators such as GridLAB-D to exchange data and run experiments in real time. As shown in the Fig. A.5, at the VOLTTRON agent stage, the flexibility parameters are subscribed as topics. The header is stripped, and the payload is sent to FNCS. The intermediary step between the VOLTTRON agent and FNCS is the FNCS bridge. FNCS bridge forwards the messages from the VOLTTRON message bus to the FNCS message bus and vice versa. FNCS receives the payload and they are subscribed as different topics and eventually the FNCS broker sends to subscriber as it sees the data/message(s).

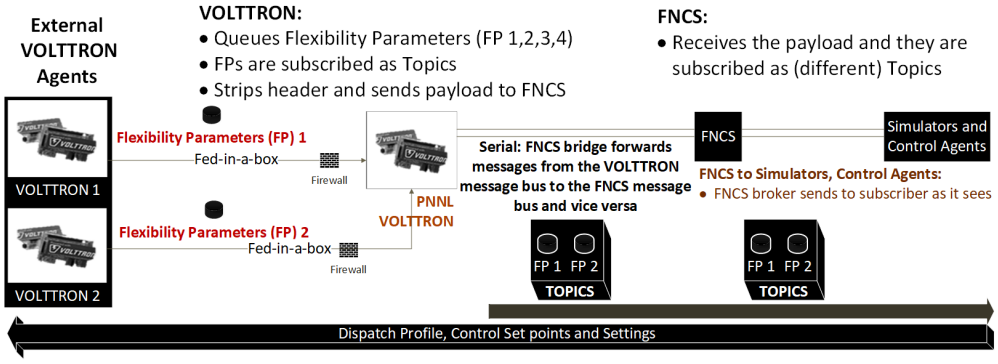


Figure A.5. Architectural overview of VOLTTRON-to-FNCS-to-Simulators

An initial test was performed on an Ubuntu Linux virtual machine to test the FNCS bridge by establishing a VOLTTRON to FNCS connection and exchanging data between them. A pre-created data file of FNCS recognizable messages were transmitted from VOLTTRON agent that is on the same VM as well as the VOLTTRON agent that is located on a different machine/VM.

The test was successful and the data exchange procedure was as expected. This means that the VOLTTRON-FNCS-Simulators (such as GridLAB-D) is successfully established on the NODES machine at PNNL. This setup is ready to receive data/messages from the external VOLTTRON agents (ex: Spirae, UTRC) and would be able to send the dispatch profiles at the desired time intervals. Establishing the FNCS bridge is a fairly straight forward process but it would require appropriate configuration by following a sequence of steps as shown in Fig. A.6 (detailed diagram can be found in Appendix - B).

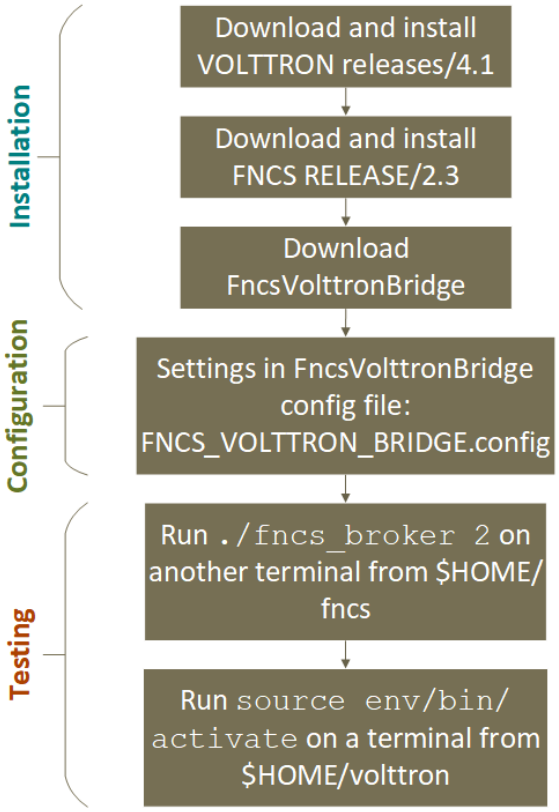


Figure A.6. Sequence of steps to establish VOLTTRON to FNCS connection

A.3 VOLTTRON to FNCS connection metrics

Based on the tests performed, it is evident that the FNCS bridge can handle messages at any speed the VOLTTRON is configured to transmit to FNCS and vice versa. Although, the bridge does not have limitations on transmitting the data, the receiving agent (example: if VOLTTRON is sending the message, FNCS would be the receiving agent) would not be able to see the message until the next timestep. A key aspect to consider prior to deciding the data transmission speed is the sampling time (or sampling frequency) of the control agent or the simulator. For example: if GridLAB-D is connected to FNCS, as long as the data exchange interval (or speed) between the VOLTTRON agent and FNCS is equal to the sampling time of the simulator (in this example: GridLAB-D), the data will be time synchronized. If not, the data will not be time synchronized and may result in undesirable results. Please see Figure 5 for an illustration of the above data flow and time synchronization requirement. Fig. A.7 demonstrates the routing process with an example. As shown, the data packet/message (denoted as M1) sent from VOLTTRON to FNCS

at the very first timestep (denoted as Time Step-1) cannot be seen by FNCS at that timestep. But, M1 is seen at the next timestep (denoted as Time Step-2). A numerical way to explain this example is if M1 is sent to FNCS at $10/10/20178 : 00 : 00$ and M2 is sent to FNCS at $10/10/20178 : 00 : 01$, FNCS sees M1 at $10/10/20178 : 00 : 01$ and M2 at $10/10/20178 : 00 : 02$.

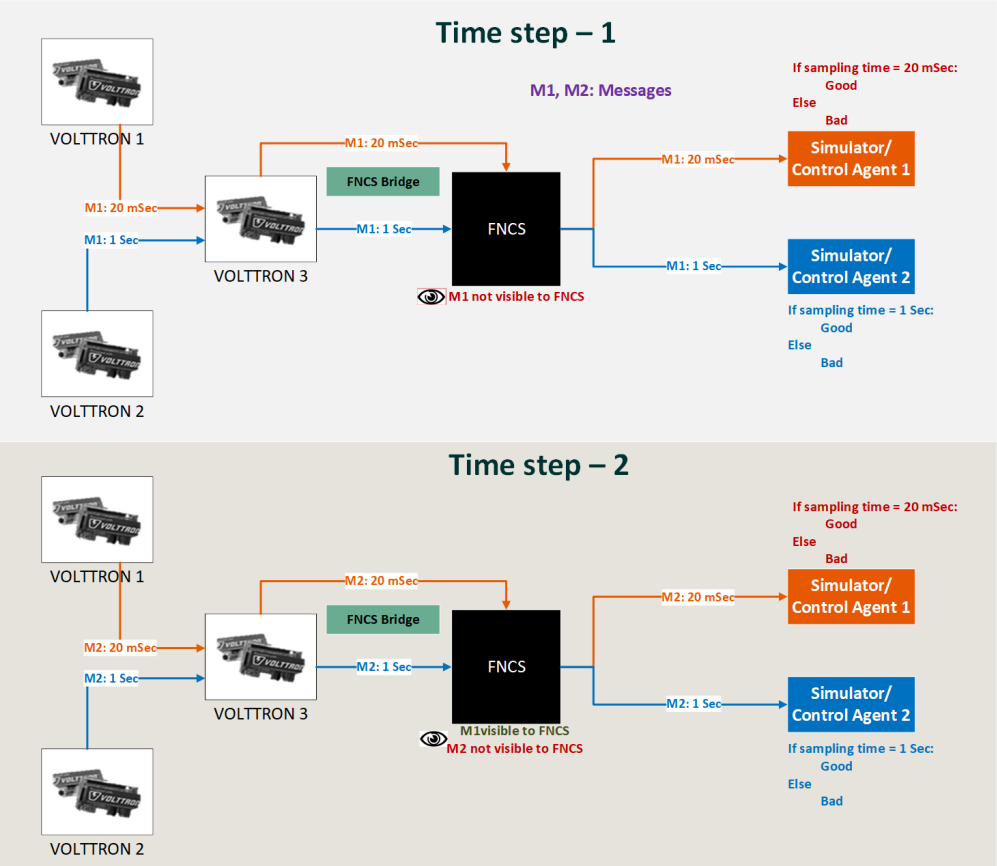


Figure A.7. Sequence of steps to establish VOLTTRON to FNCS connection

VOLTTRON to FNCS test was conducted on Ubuntu virtual machine. The test file used for this stream has 60 entries at 1 second interval. Once the VOLTTRON and FNCS instances are initiated to start the data exchange, each entry in the test file was expected to move from VOLTTRON to FNCS (and vice versa) at the designated time interval (1sec). As expected, the streaming was smooth and the data was exchanged. The data exchange results are shown in Table. A.1. As per the NODES requirement, data packets may be exchange at 5min intervals. Since this stream/exchange test was conducted to transmit the data at 1 sec interval, this proves that the established architecture would work seamlessly for the purposes of this project.

Note that in the above test, the external half where the data packet would be received by VOLTTRON agent from an external source was excluded. Currently, the routing connections between the external devices and the NODES machine are being tested. Once those connections are validated, a VOLTTRON (external) to VOLTTRON (NODES) data exchange pipeline would be established. In order to enable VOLTTRON to VOLTTRON data exchange, a well-tested forward historian would be activated on both the agents and data packets would be sent from both directions. The final test would include data exchange between the external VOLTTRON agent, NODES VOLTTRON agent, FNCS, and the simulators/control agents.

Table A.1. Specifications for physical devices at Spirae

Item	Output
Data Sending Software	VOLTTRON
Data Receiving Software	FNCS
Sampling time (Data transmission interval)	$1Second$
Data Packets sent	$60Packets$
Total Transmission Time	$1min$
Latency/delay	$<= 100mSec$
Operating System	Ubuntu Linux
Host	Virtual machine and a physical machine

Appendix B – VOLTTRON, FNCS, FncsVolttronBridge Installation

This appendix focuses on a detailed walk-through of VOLTTRON, FNCS, and FncsVolttronBridge installation, configuration, and testing. Fig. B.8 is an expanded version of Fig. A.6. It shows a detailed sequence of steps to install VOLTTRON and FNCS. Following Fig. B.8, a detailed set of instructions are defined that were strictly followed to establish the current connection.

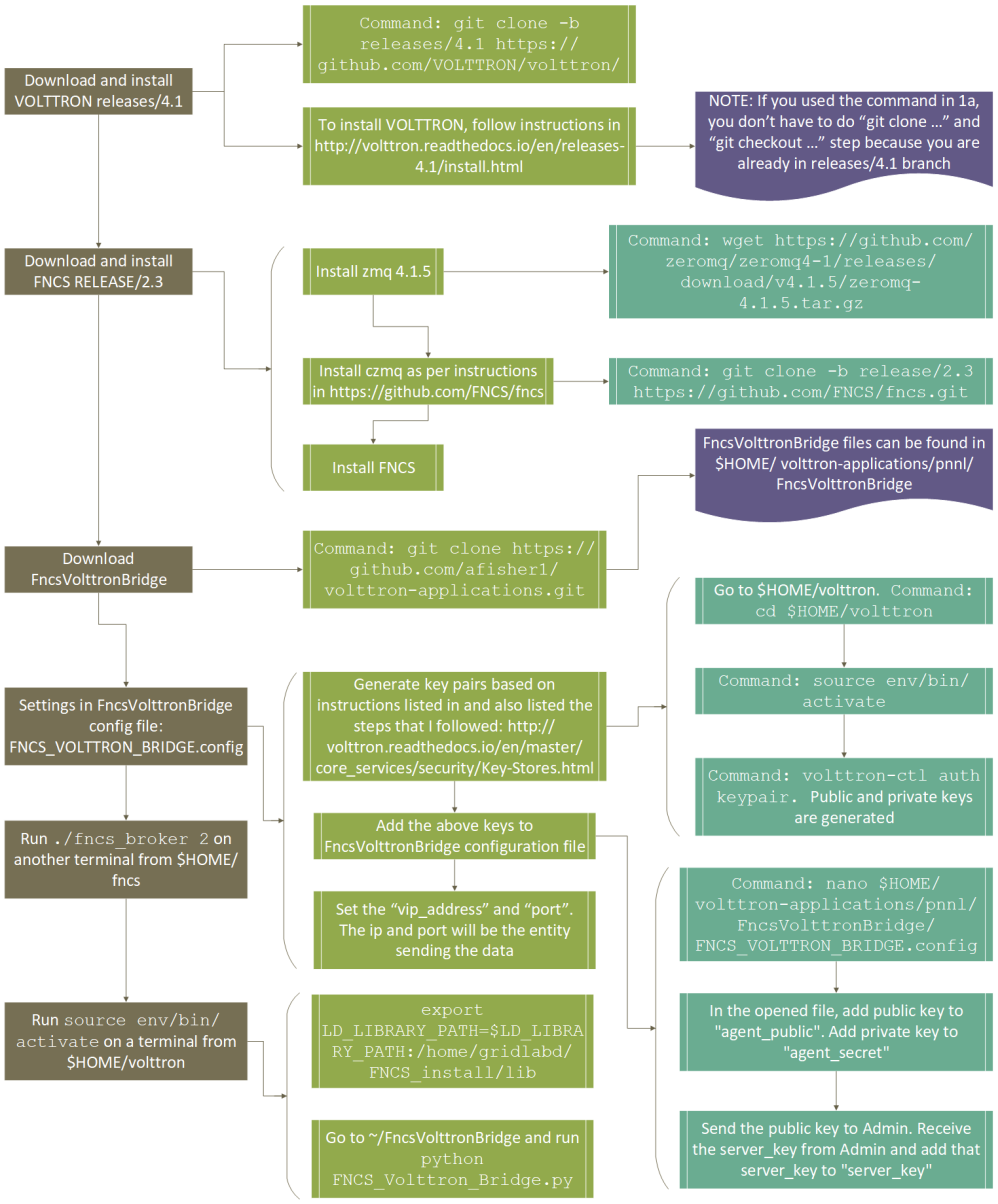


Figure B.8. Detailed sequence of steps to establish VOLTTRON to FNCS connection

1. this would minimize the number of firewall hops for the external traffic to reach the NODES machine;

2. The testbed has connection rules and cybersecurity policies that maintains the integrity of

the co-simulation;

3. Federation server is part of the testbed which makes it very efficient to connect the NODES machine. Along with this prototype test to exchange traffic from an external network to the NODES machine have been performed.

1. Everything below is installed in `$HOME` on 64bit Ubuntu Linux
2. Download and install VOLTTRON releases/4.1:
   a. Command: `git clone -b releases/4.1 https://github.com/VOLTTRON/volttron/`
   b. To install VOLTTRON, follow instructions in http://volttron.readthedocs.io/en/releases-4.1/install.html
      i. **NOTE**: If the command in 1a is used, skip "git clone …" and "git checkout …" steps because, that automatically clones releases/4.1 branch.
3. Download and install FNCS RELEASE/2.3 and associated third party libraries:
   a. First, install zmq and czmq as per instructions in https://github.com/FNCS/fncs
      i. **NOTE**: DO NOT install zmq version specified in the above instructions. For the Fncs-VolttronBridge to work, install zmq 4.1.5. **Command**: `wget https://github.com/zeromq/zeromq4-1`
      ii. Czmq installation version in 2a works fine which is version 3.0.0.
      iii. Go to 3b (below), once zmq and czmq are installed.
   b. **Command**: `git clone -b release/2.3 https://github.com/FNCS/fncs.git`
   c. Install FNCS as per instructions in 3a (above)
4. Download FncsVolttronBridge:
   a. **Command**: `git clone https://github.com/afisher1/volttron-applications.git`
   b. FncsVolttronBridge files can be found in `$HOME/ volttron-applications/pnnl/ FncsVolttronBridge`
5. Settings in FncsVolttronBridge config file: FNCS_VOLTTRON_BRIDGE.config
   a. Generate key pairs based on instructions listed in and also listed the steps that we followed: http://volttron.readthedocs.io/en/master/core_services/security/Key-Stores.html
      i. Go to `$HOME/volttron`. **Command**: `cd $HOME/volttron`
      ii. **Command**: `source env/bin/activate`
      iii. **Command**: `volttron-ctl auth keypair`. You should see public and private keys.
   b. Below steps walks through the process of adding the above keys to FncsVolttronBridge configuration file.
      i. **Command**: `nano $HOME/volttron-applications/pnnl/FncsVolttronBridge/FNCS_VOLTTRON_BRI`
      ii. In the opened file, add public key to "agent_public". Add private key to "agent_secret"
      iii. Then send the public key to the VOLTTRON Admininstrator (email). Receive the server_key from the VOLTTRON Administrator (email) and add that server_key to "server_key"

c. Now, set the “vip_address” and “port”. The ip and port will be the entity sending the data. For example: if UTRC is sending some data over this bridge, the ip and port will be the ip of the UTRC computer and the port is the port UTRC is using to send the data. BUT, THE above is just an example and is not applicable in our case. Because UTRC <—> PNNL is VOLTTRON instance <—> VOLTTRON instance connection

6. (*Copy fncs.py from fncs/python to the bridge folder*)

7. **Test step-1**: Run `./fncs_broker 2` on another terminal from `$HOME/fncs`

8. **Test step-2**: On another terminal, run `./fncs_player_anon 60s fncsplayerout.txt` from wherever the fncs_player_out.txt is located.

9. **Miscellaneous steps: Do this if you are testing the bridge on localhost itself** i.e., Fncs Bridge talking to the VOLTTRON instance on the SAME machine (Run source `env/bin/activate` on a terminal from `$HOME/volttron`).
   a. Generate the server key: `vctl auth serverkey` (do this only after `source env/bin/activate`)
   b. **Might be useful**: `cat /.volttron/config` to see IP and port#
   c. **Might be useful**: `vctl auth` list has all keys. The “Credentials” under user_id: “platform” is the server_key. Make sure you see the public_key (under some user_id other than “platform” and “listeneragent-*”)
      i. **NOTE**: if you want to simply check if the public key is already added, then do this:`vctl auth list | grep <public_key>`
      ii. If public_key is not added, add it by following this:`vctl auth add` and keep clicking enter under “credentials” are asked: add the public key here
   d. `volttron-cfg` to set up the configuration (IP, port, etc.)
   e. (*Start the platform: `volttron -vv`*)

10. **Test step-3**: Run `source env/bin/activate` on a terminal from `$HOME/volttron`
    a. `export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/home/gridlabd/FNCS_install/lib`
    b. Go to `/FncsVolttronBridge` and run `python FNCS_Volttron_Bridge.py`

11. **Miscellaneous steps: To listen (see all exchange log messages):**
    a. Go to /usr/local/volttron/examples/StandAloneListener
    b. **Edit Setting.py**: set correct public, private and server key (Same as in FNCS_VOLTTRON_BRIDGE.conf
    c. Run `python standalonelistener.py`

Appendix C – Collaborator Testbed and federation

The objective of this appendix is to describe the testbeds at UTRC, Spriae, and SCE. In adition, this appendix also provides intricate details about the federated connection between PNNL and the collaborators (UTRC and Spirae).

C.1 UTRC Testbed Integration

The current solution employed to federate with UTRC involves a Layer-2 site-to-site VPN tunnel using an OpenVPN access server and transport layer security (TLS) encryption. The PNNL team set up an OpenVPN access server in the CyberNET testbed environment that is reachable by the internet through Port Network Address Translation (PNAT) on a designated public IP and port.

PNNL's OpenVPN access server is deployed within an OpenStack cloud environment. It is dual-homed. The first interface is attached to a private software-defined subnet that is reachable externally via network address translation (NAT). A second interface is attached to a private software-defined subnet that houses the NODES VM. Using a Linux bridge the VPN TAP interface created by OpenVPN and the second interface on the server are bridged on Layer 2 of the Open Systems Interconnection (OSI) model.

On UTRC's end, a dedicated hardware system (which is also dual-homed) is deployed. Similarly, one interface is attached to a network that can reach out to the internet minimally to the port listening on the VPN server, and a second interface is connected to a private LAN segment managed by a switch. By downloading a client configuration via the web frontend of PNNL's OpenVPN access server using a pre-shared key, UTRC can use OpenVPN to connect to PNNL. The OpenVPN client system then connects to the OpenVPN access server and establishes a TLS tunnel. The client configuration file that is downloaded specifies two scripts that get triggered when the OpenVPN service starts and stops. These scripts configure the OpenVPN client system to use the same bridging strategy as the OpenVPN access server in PNNL's OpenStack cloud. The TAP interface of the OpenVPN connection and the private side interface are housed on a Linux bridge.

Upon launching the OpenVPN service, a script to turn on the bridge is executed. Upon stopping an OpenVPN service, a script to turn off the bridge is executed.

Once this site-to-site bridge is set up, systems in the virtual private LAN in PNNL's OpenStack testbed and systems in the private LAN segment at the client site can communicate over Ethernet/Layer 2 in the OSI model. Figure C.9 shows a simple network diagram of what the connection looks like and below that is the step-by-step instructions for setting up the OpenVPN client box using Ubuntu 16.04 as the operating system.

Software and Hardware Components at UTRC: On the UTRC side the architecture consists of the following:

1. A VOLTTRON agent running on a RedHat Linux machine

2. A Python wrapper that provides an application programing interface (API) to the ZMQ layer needed to exchange communication with the VOLTTRON agent

3. A set of MATLAB functions that encapsulates and hides details of the Python layer above allows to send and receive information with the VOLTTRON agent at a level of abstraction suitable for the driving MATLAB application.

4. A MATLAB application that performs required elaboration on data exchanged with VOLTTRON agent.

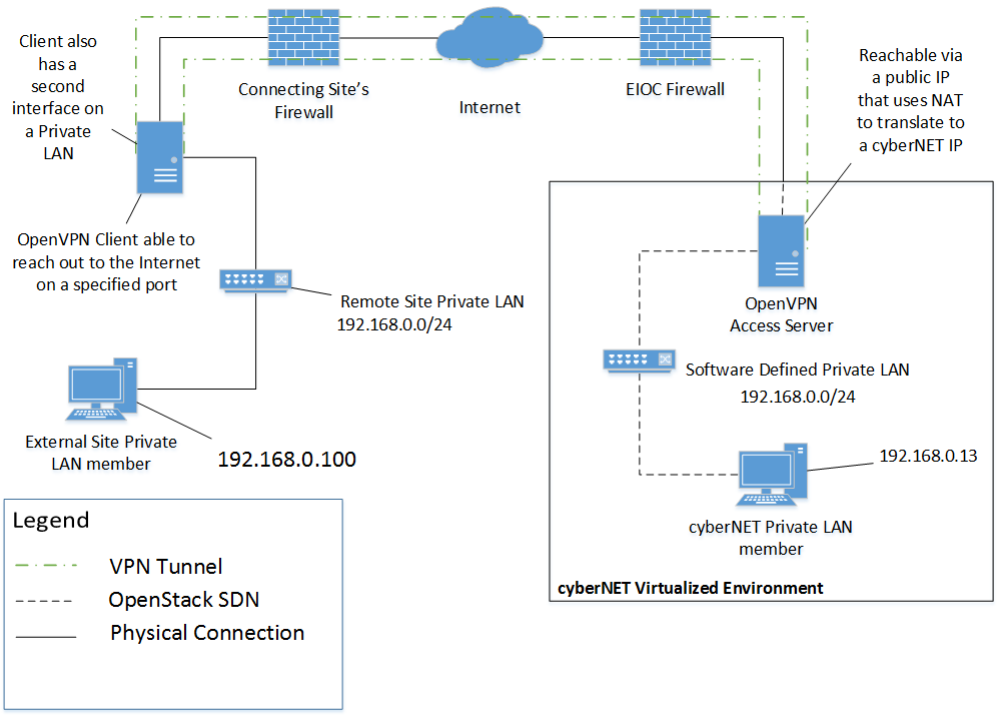


Figure C.9. Illustrative Network Diagram of Layer-2 Federation Connection

The communication between the Python/MATLAB code and the UTRC VOLTTRON agent is handled via a ZMQ publish/subscribe mechanism. The information exchange is completely asynchronous; therefore, a simple application level protocol is defined to associate each message with its corresponding response. As the exchanges are typically very infrequent (with period measured in minutes) the possible overhead required to guarantee a non-lossy communication is negligible.

The VOLTTRON agents in PNNL and UTRC also communicate via a ZMQ based publish/-subscribe mechanism. Since the two VOLTTRON instances are running on separate machines, each instance provides a forwarder agent that uses.

ZMQ to send messages to the other instance – PNNL forwards to UTRC, and UTRC forwards to PNNL. Within each VOLTTRON instance, agents subscribe to messages of interest.

Figure C.10 describes one application using the communication architecture described above. Per collected data from UTRC campus operation and simulated data from PNNL via the communication described above, a MATLAB application optimizes building operation.

C.1.1 VOLTTRON Agent development

The UTRC VOLTTRON agent is structured as a dual thread that reacts simultaneously to input received from the PNNL agent and the Python/MATLAB client. The response to input from the remote agent is triggered by the VOLTTRON subscription mechanism (a wrapper for a ZMQ publish/subscribe mechanism.) The response to input from the client is also using the ZMQ publish/subscribe mechanism but relies on an explicit time-based polling as the client is not part of the VOLTTRON infrastructure. The polling rate is currently 10/sec, significantly faster than the expected rate of stimuli which will be in the minutes range.

Application specific scripts have been added to the standard VOLTTRON deployment on the

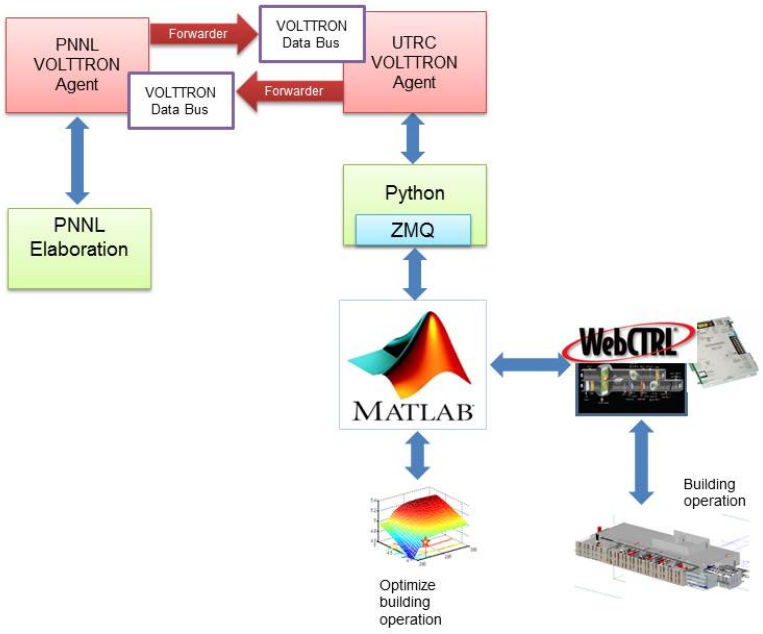


Figure C.10. Fed-in-a-box application connected to building control application

UTRC side to simplify operations such as the opening and closing of sessions, and enabling and disabling of agents and features, among other things.

C.2 Spirae Testbed

The solution employed to federate with Spirae involves a Layer-3 point-to-site VPN tunnel using an OpenVPN access server and TLS encryption. PNNL has set up an OpenVPN access server in the CyberNET testbed environment that is reachable by the internet through PNAT on a public IP.

At PNNL's end, the OpenVPN access server is deployed within an OpenStack cloud environment. Similar to the Layer-2 connection, it is dual-homed, and the first interface is attached to a private software-defined subnet that is reachable externally via NAT. A second interface is attached to a private software-defined subnet that houses the NODES VM. By configuring static routes in the OpenVPN access server settings, clients can reach PNNL's private network in the testbed on Layer 3 of the OSI model.

On Spirae's end, a dedicated VM is deployed. One interface is attached to a network that can reach out to the internet minimally to the port the PNNL VPN server is listening on. By downloading a client configuration via the web frontend of the OpenVPN access server using a pre-shared key, Spirae can use OpenVPN to connect to PNNL. The OpenVPN client system then connects to the OpenVPN access server and establishes a TLS tunnel. The client configuration file that is downloaded specifies that the client can route to a private LAN behind the OpenVPN access server.

Once this point-to-site tunnel is set up (see Figure C.11), systems in the virtual private LAN in PNNL's OpenStack testbed and the connected client system on the Spirae network can com-

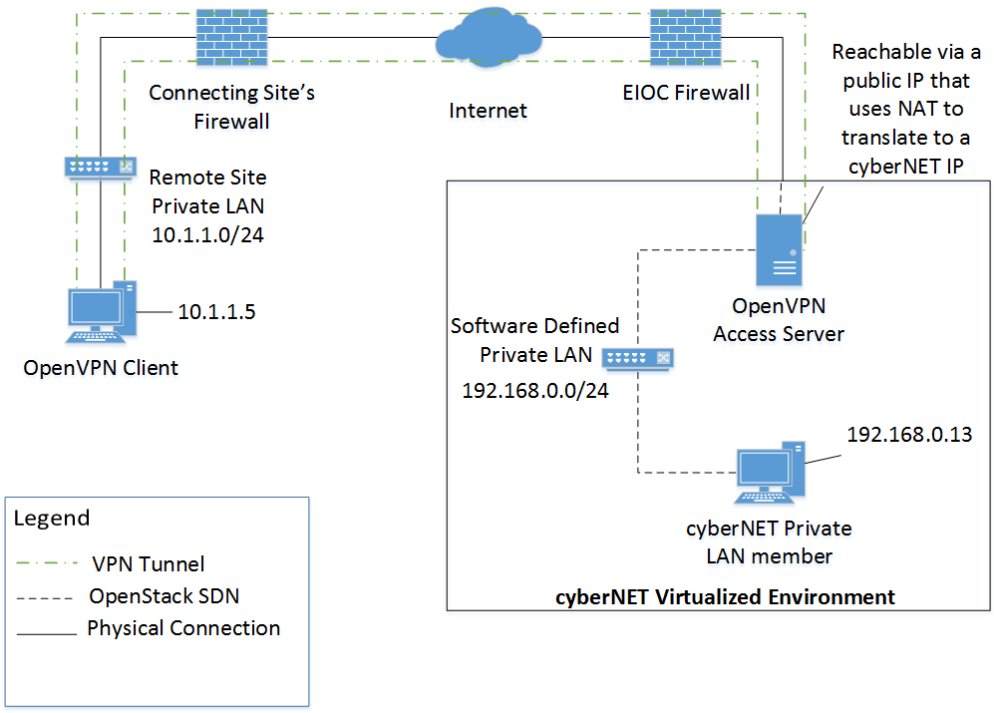


Figure C.11. Illustrative Network Diagram of the Layer-3 Federation Connection

municate over TCP/IP, or Layer 3 in the OSI model. Figure C.12 is a simple network diagram of what the connection looks like and below that are the step-by-step instructions for setting up the OpenVPN client box using Ubuntu 16.04 as the OS.

Software and Hardware Components at Spirae: Under the NODES project, Spirae is providing remote access to around 60 physical power system resources located at two sites powered by the same distribution substation. Resources include curtailable solar inverters, battery energy storage systems, small generation, and a variety of interruptible single- and three-phase loads, including a curtailable electric vehicle charging station. Access is by means of APIs exposed by Spirae's Wave® microgrid control software, which delivers asset- and group-based monitoring and control functionality. Thus, the asset-specific interfaces are abstracted to a common secure format. This way, the larger simulation can interact with those assets, while limiting exposure of Modbus interfaces to the Wave components. For initial testing purposes, Spirae has emulated a version of the physical assets—communicating to the microgrid software via the same Modbus points—thus limiting the need to expose control of physical devices to high-value testing times. This software and hardware architecture is summarized in Figure C.12. The final experiment may involve interaction and control using real hardware systems instead of emulated software systems.

C.3 SCE Testbed Federation

In order to test, evaluate, and demonstrate emerging distribution control system technology, SCE designed and implemented a Controls Testbed. The testbed is both flexible enough to test a wide variety of use cases and robust enough to simulate the operation of hundreds of DERs.

The core of the Controls Testbed is a real time power system simulator. The simulator performs a three-phase unbalanced dynamic RMS simulation synchronized to the system time. The time

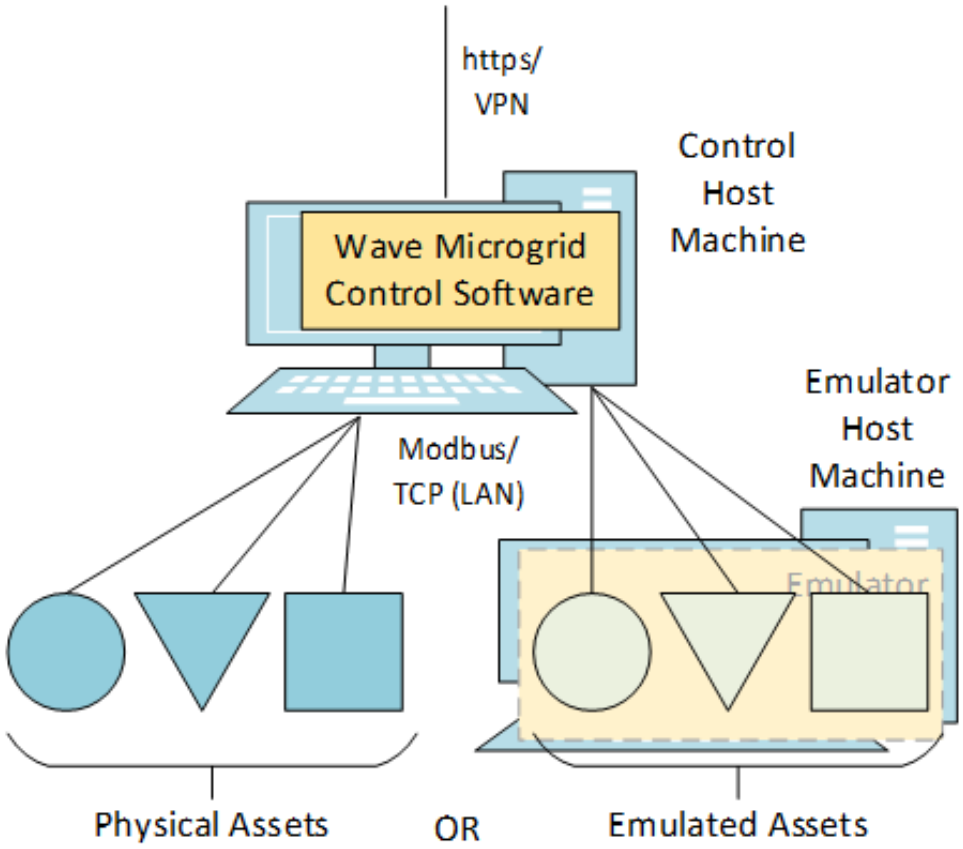


Figure C.12. Depiction of the physical and emulated assets hosted by Spirae

step of the simulation can be adjusted in order to simulate larger networks. This choice of simulation allows for simulating a very detailed distribution feeder model including DERs, localized controls on capacitor banks and voltages regulators, and captures system dynamics (voltage and frequency oscillations for example).

The simulator interfaces with the outside world through a Supervisory Control and Data Acquisition (SCADA) gateway. The SCADA gateway can establish multiple independent Distributed Network Protocols (DNP3), Modbus serial communications, Inter-Control Center Communications Protocol (ICCP), or 61850 MMS clients or servers. The SCADA gateway processes data once per second and exchanges data with the real time power system simulator via OPC. The VOLTTRON platform is connected to the controls testbed through a Modbus server.

An example of a typical connection between the power system simulator and SCADA gateway is shown in Figure C.13.

There were several options for integrating VOLTTRON into the real time simulation component of the Controls Testbed, the best option was to interface the VOLTTRON Modbus Master Driver with the Modbus Servers in the Controls Testbed. This option allowed SCE to leverage the ZeroMQ messaging system and existing agents in the VOLTTRON platform. A diagram of the integration in shown in Figure C.14.

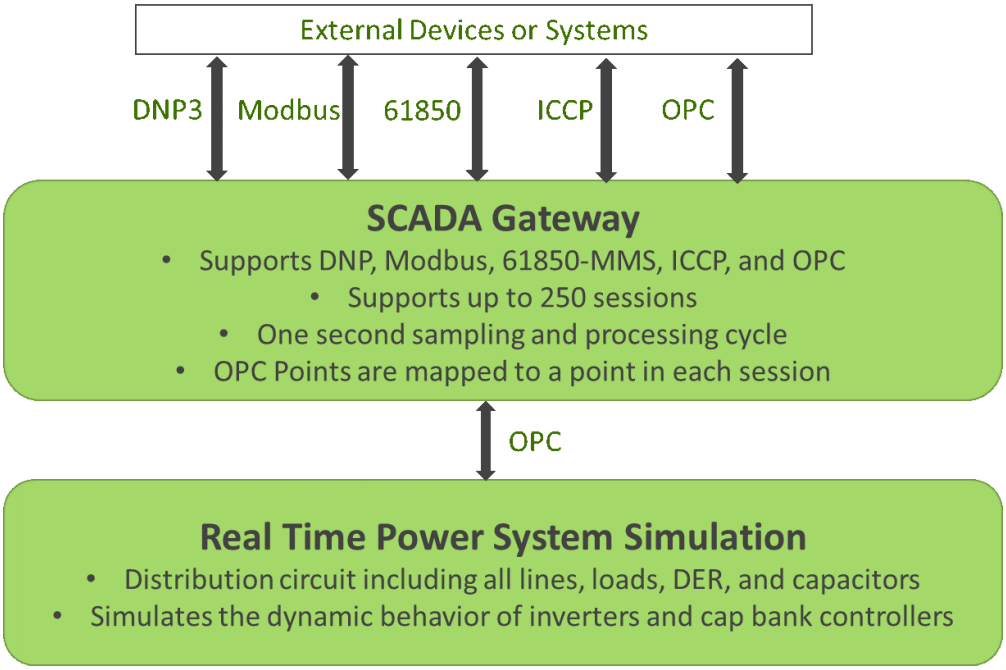


Figure C.13. Real time simulation to SCADA Gateway interface

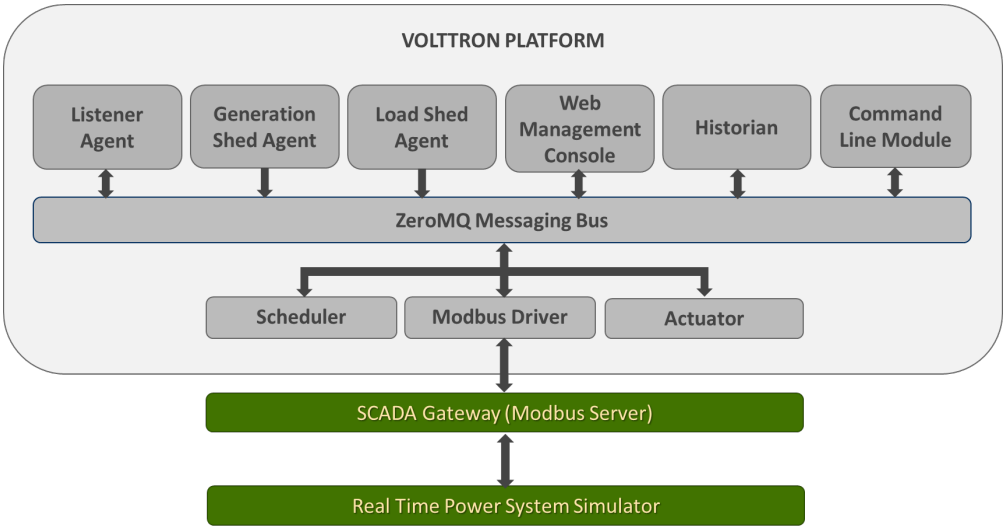


Figure C.14. VOLTTRON integration into Controls Testbed

Appendix D – Data Transfer Specifications between PNNL and Collaborators

D.1 Software Process involved in sending data from UTRC VOLTTRON to PNNL VOLTTRON

1. Install VOLTTRON 5.0 (delete any older instances)

   - `git clone https://github.com/VOLTTRON/volttron -b releases/5.0rc`
   - `cd volttron/`
   - `python bootstrap.py`

2. Activate and Configure

   - `. env/bin/activate`
   - `vcfg`
     - Is this the volttron you are attempting to setup? [Y]: y
     - What is the external instance ipv4 address? [tcp://127.0.0.1]: tcp://<Your external facing IP address>
     - What is the instance port for the vip address? [22916]:
     - Is this instance a volttron central? [N]: n
     - Will this instance be controlled by volttron central? [Y]: n
     - Would you like to install a platform historian? [N]: n
     - Would you like to install a master driver? [N]: y
     - *Configuring /home/volttron/volttron/services/core/MasterDriverAgent*
     - Install a fake device on the master driver? [N]: y
     - Should agent autostart? [N]: n
     - Would you like to install a listener agent? [N]: y
     - *Configuring examples/ListenerAgent*
     - Should agent autostart? [N]: n
     - *Finished configuration*
   - Start volttron and see the log for any errors
     - `./start-volttron // Start volttron. Do this from /volttron location`
     - `tail -f volttron.log //FYI, you don't need to do this`
   - Forwarder configurations and run it
     - `mkdir config //do this from /volttron directory`
     - `cp services/core/ForwardHistorian/config config/forwarderconfig`
     - `vim config/forwarderconfig`
       - `destination-vip: "tcp://<IP>:<port. Ex.22916> //IP of machine you want to send the data to`
       - `destination-serverkey: <get serverkey from Sri at pnnl dot gov and paste it here>`
     - `vim config/updateforwarder`
       - `python scripts/install-agent.py -s services/core/ForwardHistorian -c config/forwarderconfig -start -force -i forwarder.historian`
       - `[/*for help, start with this:*/ cd.. /*now, you should be in /volttron directory. Do this:*/python scripts/install-agent.py -help]`

- cd .. //now you should be in /volttron
- chmod +x config/updateforwarder
- ./config/updateforwarder

- Generate public key and run the forward historian, master_Driver
  - vctl auth publickey //give the public key with "IDENTITY: forwarder.historian" to sri at pnnl dot gov - he then authorizes UTRC's public key so PNNL's volttron can receive the data
  - vctl status //you should see forwarder.historian and master_driver running. In below ex., master_driver is not running:
  - vctl status

Table D.2. VCTL Status Screen

AGENT	IDENTITY	TAG	STATUS
9 forwarderagent-4.0	forwareder.historian		running [1393]
7 listeneragent-3.2	listeneragent-3.2_1	listener	
0 master_driveragent-3.1.1	platform.driver	master_driver	

  - /*Do this to start the master driver:*/vctl start 0 //PNNL's VOLTTRON should now receive the data. See below: 2.1, 2.3 to know how PNNL would see the data

D.2 To receive data from PNNL VOLTTRON to UTRC VOLTTRON

1. Stop forwarder.historian, master_driver and start listeneragent
   - vctl status // shows the agents
   - vctl stop 9 //see ex. In 1.5 about "9". Stops forwarder.historian
   - vctl start 7 // see ex. In 1.5 about "7". Starts listeneragent
2. Authorize PNNL VOLTTRON agent
   - Vctl auth serverkey //give that serverkey to sri at pnnl dot gov. Now he adds the server key to his equivalent of /home/volttron/volttron/config/forwarderconfig
   - vctl auth add --credentials <forwarder.historian publickey> //get the forwarder.historia public key from sri at pnnl dot gov and add authorize it. When ready, ask him to start sending the data
3. Watch the data being received by UTRC VOLTTRON agent
   - tail -f volttron.log //now you should see both the listener heartbeat and the data that is sent from PNNL VOLTTRON agent
4. Stop everything
   - deactivate

- `./stop-volttron // shutdown volttron. Do this from  /volttron location`
- `ps -ef | grep volttron //see what is running`
- `kill <#> //whatever you want to kill`

Appendix E – SCE Agent Scripts for Integration Testing

E.1 Fake Load Shed Agent for Integration Testing

```
import logging
import sys
import datetime

from volttron.platform.vip.agent import Agent, PubSub, Core
from volttron.platform.agent import utils

utils.setup_logging()
_log = logging.getLogger(__name__)
__version__='0.1'

def fakeloadshed(config_path, **kwargs):

    config = utils.load_config(config_path)
    _log.debug(config)
    agent_id = "fls"

    class FakeLoadShedAgent(Agent):

        def __init__(self, **kwargs):
            super(FakeLoadShedAgent, self).__init__(**kwargs)

        @Core.receiver('onsetup')
        def setup(self, sender, **kwargs):
            self._agent_id = "fls"

        @Core.receiver('onstart')
        def startup(self, sender, **kwargs):
            self.use_rpc()


        def use_rpc(self):
            try:
                start = str(datetime.datetime.now())
                end = str(datetime.datetime.now() + datetime.timedelta(minutes=1))

                msg = []

                for der, value in config.items():
                    _log.debug(value['address'])
                    msg.append([value['address'],start,end])

                result = self.vip.rpc.call(
                                        'platform.actuator',
                                        'request_new_schedule',
```

```
                                    agent_id,
                                    "shed load",
                                    'HIGH',
                                    msg).get(timeout=10)
                    _log.debug("schedule result", result)

                except Exception as e:
                    _log.debug ("Could not contact actuator. Is it running?")
                    _log.debug(e)
                    return

                try:
                    if result['result'] == 'SUCCESS':

                        for der, value in config.items():
                            if(value['point'] == 'pcurtail'):
                                result = self.vip.rpc.call(
                                            'platform.actuator',
                                            'set_point',
                                            agent_id,
                                            value['address']+'/'+value['point'],
                                            int(value['size'])).get(timeout=15)
                                _log.debug("Set result", result)

                            elif(value['point'] == 'pref'):
                                result = self.vip.rpc.call(
                                            'platform.actuator',
                                            'set_point',
                                            agent_id,
                                            value['address']+'/'+value['point'],
                                            int(value['size'])).get(timeout=15)
                                _log.debug("Set result", result)
                            elif(value['point'] == 'drsignal'):
                                result = self.vip.rpc.call(
                                            'platform.actuator',
                                            'set_point',
                                            agent_id,
                                            value['address']+'/'+value['point'],
                                            True).get(timeout=15)
                                _log.debug("Set result", result)

                except Exception as e:
                    _log.debug ("Device Failure - Device not found!")
                    _log.debug(e)


        Agent.__name__ = 'fakeloadshed'
        return FakeLoadShedAgent(**kwargs)

    def main(argv=sys.argv):
```

```
    '''Main method called by the eggsecutable.'''
    try:
        utils.vip_main(fakeloadshed, version=__version__)
    except Exception as e:
        _log.debug(e)
        _log.exception('unhandled exception')


if __name__ == '__main__':
    # Entry point for script
    try:
        sys.exit(main())
    except KeyboardInterrupt:
        pass
```

E.1.1 Fake Generation Shed Agent for Integration Testing

```
import logging
import sys
import datetime

from volttron.platform.vip.agent import Agent, PubSub, Core
from volttron.platform.agent import utils

utils.setup_logging()
_log = logging.getLogger(__name__)
__version__='0.1'

def fakegenshed(config_path, **kwargs):

    config = utils.load_config(config_path)
    _log.debug(config)
    agent_id = "fgs"

    class FakeGenShedAgent(Agent):

        def __init__(self, **kwargs):
            super(FakeGenShedAgent, self).__init__(**kwargs)

        @Core.receiver('onsetup')
        def setup(self, sender, **kwargs):
            self._agent_id = "fgs"

        @Core.receiver('onstart')
        def startup(self, sender, **kwargs):
            self.use_rpc()


        def use_rpc(self):
            try:
```

```
            start = str(datetime.datetime.now())
            end = str(datetime.datetime.now() + datetime.timedelta(minutes=1))

            msg = []

            for der, value in config.items():
                _log.debug(value['address'])
                msg.append([value['address'],start,end])

            result = self.vip.rpc.call(
                                    'platform.actuator',
                                    'request_new_schedule',
                                    agent_id,
                                    "shed gen",
                                    'HIGH',
                                    msg).get(timeout=10)
            _log.debug("schedule result", result)

        except Exception as e:
            _log.debug ("Could not contact actuator. Is it running?")
            _log.debug(e)
            return

        try:
            if result['result'] == 'SUCCESS':

                for der, value in config.items():
                    if(value['point'] == 'pcurtail'):
                        result = self.vip.rpc.call(
                                    'platform.actuator',
                                    'set_point',
                                    agent_id,
                                    value['address']+'/'+value['point'],
                                    int(0)).get(timeout=15)
                        _log.debug("Set result", result)

                    elif(value['point'] == 'pref'):
                        result = self.vip.rpc.call(
                                    'platform.actuator',
                                    'set_point',
                                    agent_id,
                                    value['address']+'/'+value['point'],
                                    int(-1*value['size'])).get(timeout=15)
                        _log.debug("Set result", result)
                    elif(value['point'] == 'drsignal'):
                        result = self.vip.rpc.call(
                                    'platform.actuator',
                                    'set_point',
                                    agent_id,
                                    value['address']+'/'+value['point'],
```

```
                                        False).get(timeout=15)
                                _log.debug("Set result", result)


                    except Exception as e:
                        _log.debug ("Device Failure - Device not found!")
                        _log.debug(e)


        Agent.__name__ = 'fakegenshed'
        return FakeGenShedAgent(**kwargs)

def main(argv=sys.argv):
    '''Main method called by the eggsecutable.'''
    try:
        utils.vip_main(fakegenshed, version=__version__)
    except Exception as e:
        _log.debug(e)
        _log.exception('unhandled exception')


if __name__ == '__main__':
    # Entry point for script
    try:
        sys.exit(main())
    except KeyboardInterrupt:
        pass
```

E.1.2 Integration and Controls Testing

To ensure the federated connection between PNNL and UTRC is effective and the data transfer rate is within the requirements to perform NODES experiment, the team conducted data transfer tests between the PNNL and UTRC VOLTTRON agents. For the purposes of the latency and transfer speed test, PNNL VOLTTRON team transmitted five sets of illustrative data. The experiment was conducted such that the data received by UTRC would immediately be sent back to PNNL. This ensured that PNNL could accurately time how long each round-trip communication took. The average time in seconds it took to send a specific message is calculated as below:

$$T_{avg} = \frac{T_{PNNL-UTRC} + T_{UTRC-PNNL}}{2} \tag{E.1}$$

where T_{avg} is the average time in seconds, $T_{PNNL-UTRC}$ is the time to send data from PNNL to UTRC, and $T_{UTRC-PNNL}$ is the time to received data sent from UTRC to PNNL. Table. E.3 shows and overview of 5 tests performed, the first test began with a small array of 100 elements with a total character count of 473. The next four tests were conducted by increasing the array sizes (resulting in increased number of characters). It can be observed that as the array sizes increased, the average transmission time is increased as well. Despite the increase in transmission time, it is still within the fastest requirement of less than 2 seconds.

The five tests with varying messages sizes was performed to understand the delay associated with the connection between PNNL and UTRC. The actual data being exchanged differs in size depending on direction. From PNNL to UTRC the data will be a service request message which

Table E.3. Summary of PNNL to UTRC data exchange test

Array Size	Character count	Average time in sec
100	473	0.126
200	974	0.1135
400	1974	0.817
800	3974	1.066
1600	8574	1.757

is 119 characters long and from UTRC to PNNL it will be the VBM model which is 1424 characters long. Due to the design of the control architecture these messages will only be sent once per allocation period (5-15 minutes) and therefor only the message being sent from PNNL to UTRC is time critical and needs to meet the requirement for this milestone. However, as seen in Fig. E.15 both message sizes are well within the requirements. In Figure 3, Y-axis is the “Average time in sec” and x-axis is the character count.

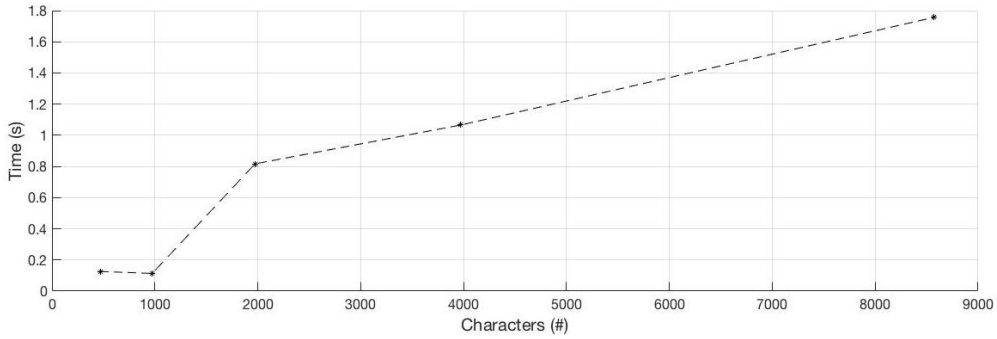


Figure E.15. VOLTTRON – FNCS message bus data exchange

Results from the tests performed demonstrates that the federated communication architecture established between PNNL VOLTTRON instance and UTRC VOLTTRON instance is robust enough to conduct a seamless experiment that involves exchange of VBM, service request, and other flexibility parameters.

Appendix F – Virtual Battery Formulation: Examples

In this section, we present two examples of calculating VB model parameters for a collection of air-conditioners and electric water-heaters using closed-forms expressions and the individual device parameters [6, 7].

Example 1. *(Air-conditioners) Consider an ensemble of N air-conditioning (AC) loads. The device dynamics is represented by,*

$$\dot{T}(t) = -\frac{(T(t) - T_a(t))}{C\,R} - \frac{\eta\, p(t)}{C}\,, \tag{F.2a}$$

$$p(t^+) = \begin{cases} 0\,, & \text{if } T(t) \le T_{set} - \delta T/2 \\ P\,, & \text{if } T(t) \ge T_{set} + \delta T/2 \\ p(t)\,, & \text{otherwise} \end{cases}\,, \tag{F.2b}$$

where $T(t)$ is the room temperature; $p(t) \in \{0, P\}$ represent the power draw of the AC; $T_a(t)$ denotes the outside air temperature; and C, R, η are the device parameters representing the room thermal resistance, thermal capacitance and the load efficiency, respectively. T_{set} is the temperature set-point and δT represents the width of the temperature hysteresis deadband. The baseline power (p^{base}) and the VBM parameters for a large ($N \gg 1$) ensemble of such devices are given by,

$$p^{base}(t) = \sum_{i=1}^{N} \frac{T_a^i(t) - T_{set}^i(t)}{\eta^i\, R^i}\,, \tag{F.3a}$$

$$\mu^-(t) = -p^{base}(t)\,, \tag{F.3b}$$

$$\mu^+(t) = \sum_{i=1}^{N} P^i - p^{base}(t)\,, \tag{F.3c}$$

$$\kappa^-(t) = -\sum_{i=1}^{N} \frac{C^i\,(\delta T)^i}{2\,\eta^i}\,, \tag{F.3d}$$

$$\kappa^+(t) = \sum_{i=1}^{N} \frac{C^i\,(\delta T)^i}{2\,\eta^i}\,, \tag{F.3e}$$

$$\text{and}\quad \alpha = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{C^i\,R^i}\,. \tag{F.3f}$$

□

Example 2. *(Electric water-heaters) Consider an ensemble of N electric water-heating (EWH) loads. Depending on the requirements, the water temperature dynamics of an EWH can be modeled at varying details [41]. For our purpose, it suffices to use the 'one-mass' thermal model which assumes the temperature inside the water-tank is spatially uniform (valid when the tank is nearly full or nearly empty) [13]:*

$$\dot{T}_w(t) = -a(t)\,T_w(t) + b(s(t), t)\,, \tag{F.4}$$

$$\text{where, } a(t) := \frac{1}{C_w}\left(\dot{m}(t)\,C_p + W\right),$$

$$\text{\& } b(s(t), t) := \frac{1}{C_w}\left(s(t)\,P + \dot{m}(t)\,C_p\,T_{in}(t) + W\,T_a(t)\right).$$

T_w denotes the temperature of the water in the tank, and $s(t)$ denotes a switching variable which determines whether the EWH is drawing power ($s(t) = 1$ or 'on') or not ($s(t) = 0$ or 'off'). Unless otherwise specified, the parameters are assumed to be uniformly distributed in the given range of values, except the hot water-flow rate ($\dot{m}$) which is assumed to follow certain typical water draw profiles [13]. The state of the EWH ('on' or 'off') is determined by the switching condition:

$$s(t^+) = \begin{cases} 0\,, & \text{if } T_w(t) \geq T_{set} + \delta T/2 \\ 1\,, & \text{if } T_w(t) \leq T_{set} - \delta T/2 \\ s(t)\,, & \text{otherwise} \end{cases}\,, \tag{F.5}$$

where T_{set} is the temperature set-point of the EWH with a deadband width of δT. The baseline power (p^{base}) and the VBM parameters for a large ($N \gg 1$) ensemble of such devices are given by,

$$p^{base}(t) = \sum_{i=1}^{N} \left[W^i \left(T^i_{set}(t) - T^i_a(t) \right) + \dot{m}(t)\, C_p \left(T^i_{set}(t) - T^i_{in}(t) \right) \right]\,, \tag{F.6a}$$

$$\mu^-(t) = -p^{base}(t)\,, \tag{F.6b}$$

$$\mu^+(t) = \sum_{i=1}^{N} P^i - p^{base}(t)\,, \tag{F.6c}$$

$$\kappa^-(t) = -\frac{1}{2} \sum_{i=1}^{N} C^i_w \left(\delta T\right)^i\,, \tag{F.6d}$$

$$\kappa^+(t) = \frac{1}{2} \sum_{i=1}^{N} C^i_w \left(\delta T\right)^i\,, \tag{F.6e}$$

$$\text{and} \quad \alpha = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{C^i_w} \left(\dot{m}(t)\, C^i_p + W^i \right)\,. \tag{F.6f}$$

□

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7675)

www.pnnl.gov