Predictive Control for Energy Efficient Buildings with Thermal Storage

Yudong Ma, Anthony Kelman, Allan Daly, Francesco Borrelli

The building sector is the largest energy consumer in the world. It is therefore economically, socially, and environmentally significant to reduce the energy consumption of buildings. Achieving substantial energy reduction in buildings may require rethinking the whole processes of design, construction, and operation of a building. This article focuses on the specific issues of model-based predictive control system design for buildings.

Building controls design becomes challenging when predictions of weather, occupancy, renewable energy availability, and energy price are used for closed-loop control. The challenge will be even greater when conventional systems are replaced by innovative heating and cooling systems that use active storage of thermal energy [1]–[5]. Given its success in the process industry, model predictive control (MPC) is the first choice for systematically addressing building system control design. In principle, model predictive control can use all the aforementioned predictions to improve utility efficiency, decrease peak demand, and reduce energy costs by using load shifting. Load shifting makes use of thermal storage available through the building itself or external storage devices so that energy is saved during optimal conditions for later use.

The main idea of model predictive control is to use a model of the process to predict its future evolution [6]–[8]. At each sampling time, starting with the current observed or measured state, an open-loop optimal control problem is solved over a finite horizon. The resulting optimal command signal is applied to the process only during the following sampling interval. At the next time step, a new optimal
control problem based on new state measurements is solved over a shifted horizon. The application of predictive optimal controllers for active and passive building thermal storage is extensively studied in the past years [9]–[15].

This article tries to shed some light on the following questions. Why should MPC be used in the first place, which type of predictive models should be used, what are the benefits of active thermal storage, and is MPC still a good choice when predictions are uncertain? These questions are problem dependent and difficult to answer. The goal of this paper is to present the details of a specific building control architecture and the abstraction levels associated with it, and to show the potential advantages and disadvantages of model predictive control over conventional building control sequences. This article focuses on the tight coupling between active storage and predictive control.

**Model Predictive Control and Thermal Storage: a Simple Example**

In order to make people feel comfortable, energy has to be converted, stored, and delivered to buildings. The simplest measure of comfort is the air temperature of a given space. The objective of this section is to show the potential benefits of model predictive control when applied to building air temperature regulation. A simplified model of the system is used to show the basic principles of active thermal storage, and to demonstrate the merits of model predictive control.

The temperature dynamics of a given space can be modeled by using an RC circuit analogy

\[ C \dot{T} = u + P_d + (T_{oa} - T)/R, \]  

where \( T \) is the temperature of the thermal zone, \( P_d \) is the external disturbance load generated by occupants, direct sunlight, and electrical devices, \( T_{oa} \) is the temperature of outside air, and \( u \) is the heating and cooling power input to the zone. The zone is cooled when \( u \leq 0 \), and heated when \( u \geq 0 \). The lumped parameter \( R \) describes the thermal resistance of walls and windows isolating the zone from the outside environment,
and the parameter $C$ captures the thermal capacitance of the room components including walls, floors, and furniture.

By using Euler discretization with a sampling rate of $\Delta t$, the representation of system (1) in discrete time is

$$T(k + 1) = AT(k) + Bu(k) + d(k), \quad (2)$$

where $A = 1 - \Delta t/RC$, $B = \Delta t/C$, $d = P_d\Delta t/C + T_{oa}\Delta t/RC$.

A simple model predictive control problem is formulated with objectives of minimizing total heating and cooling energy consumption, minimizing the peak power consumption, and maintaining rooms within a desired temperature range despite predicted load changes. The predictive controller solves at each time step the following problem

$$\min_{U_t, \varepsilon} \sum_{k=0}^{N-1} |u_{t+k}| + \kappa \max\{|u_t|, \ldots, |u_{t+N-1}|\} + \rho \sum_{k=1}^{N} (|\varepsilon_{t+k}| + |\varepsilon_{t+k}|) \quad (3)$$

$$T_{t+k+1} = AT_{t+k} + Bu_{t+k} + d_{t+k}, \quad (4)$$

$$\bar{T} - \varepsilon_{t+k} \leq T_{t+k} \leq \bar{T} + \varepsilon_{t+k}, \quad (5)$$

$$\varepsilon_{t+k}, \varepsilon_{t+k} \geq 0, \quad (6)$$

where the symbol $v_{t+k}$ is read as “the variable $v$ at time $t + k$ predicted at time $t$”. For instance, $T_{3|1}$ represents the predicted temperature at time 3 when the prediction is done at time $t = 1$ starting from the current temperature $T(1)$. It is in general different from $T_{3|2}$, which is the predicted temperature at time 3 when the prediction is done at time $t = 2$ starting from the current temperature $T(2)$. With this notation in place, $U_t = [u_t, u_{t+1}, \ldots, u_{t+N-1}]$ is the vector of energy control inputs, $\varepsilon = [\varepsilon_{t+1}, \ldots, \varepsilon_{t+N}]$ is the temperature violations from the lower bounds, $\varepsilon$ the temperature violation from the upper bounds, $T_{t+k}$ is the thermal zone temperature, $d_{t+k}$ is the load prediction, and $\bar{T}$ and $\bar{T}$ are the lower and upper bounds on the zone temperature, respectively. $\rho$ is the penalty on the comfort constraint violations, and
$\kappa$ is the penalty on peak power consumption.

Let $U_t^* = \{u^*_{t|t}, \ldots, u^*_{t+N-1|t}\}$ be the optimal solution of Problem (3)–(6) at time $t$. Then, the first element of $U_t^*$ is applied to system (2)

$$u(t) = u^*_{t|t}.$$  

(7)

The optimization problem (3)–(6) is repeated at the next time step $t + 1$ based on the new measured temperature $T_{t+1|t+1} = T(t + 1)$, yielding a moving or receding horizon control strategy.

The following parameters are used in the simulations presented next. Thermal capacitance $C = 9.2 \times 10^3$ kJ/°C, thermal resistance $R = 50$ °C/kW, sampling rate $\Delta t = 1$ hour, prediction horizon $N = 24$ hours, and thermal comfort interval $[T, \bar{T}] = [21, 26]$ °C. It is assumed that weather and load are periodic with a period of one day, and that their predictions are perfect without mismatch between predictions and actual measurements. The outside air temperature profile $T_{oa}(t)$ used in (1) is depicted in Figure 1(a), and the disturbance load profile $P_d(t)$ used in (1) is depicted in Figure 1(b).

Two controllers are considered. Controller C1 is a proportional controller designed to reject the load without predictive information. Controller C1 input is zero when the space temperature is within the comfort range, otherwise the proportional control law

$$u(t) = \begin{cases} 
K(\bar{T} - T(t)) & T(t) \geq \bar{T}, \\
0 & \bar{T} \leq T(t) \leq T, \\
K(T - T(t)) & T(t) \leq \bar{T}
\end{cases}.$$  

(8)

is applied to system (2). Controller C2 implements the MPC Problem (3)–(6)–(7).

Simulations of system (2) in closed-loop with C1 and C2 are performed until the system settles to static periodic behavior. The performance of the controllers is measured by the closed loop total energy
consumption

\[ J^u = \sum_{k=0}^{N-1} |u^*(k)| \Delta t, \]  

(9)

the peak power consumption

\[ J^p = \max\{|u^*(0)|, \ldots, |u^*(N-1)|\}, \]  

(10)

and the total comfort violation

\[ J^\varepsilon = \sum_{k=0}^{N} (|\varepsilon^*(k)| + |\varepsilon^*(k)|) \Delta t. \]  

(11)

Closed-loop simulations are performed with various gains \( K \) of the controller C1 and various tuning \( \rho \) of the MPC controller with no penalty on peak power consumption, i.e. \( \kappa = 0 \). Figure 2 shows the tradeoff between comfort violation and total energy. It is observed that C1 and C2 use the same energy for the same amount of constraint violation. As expected, increased comfort violation corresponds to a lower energy usage for both controllers.

With \( \kappa \neq 0 \), a different behavior can be observed. Simulation results for C1 with \( K = 400 \) and C2 with \( \rho = 1000 \) and \( \kappa = 2 \) are plotted in Figure 3. The peak power consumption \( J^p \) of the proportional controller C1 is reduced by 89\% when the MPC control C2 is used. For both controllers the comfort violation indexes \( J^\varepsilon \) are zero. This nice behavior is obtained by taking advantage of the predictive information of the disturbance and using the space thermal storage. In fact the MPC precools the space temperature to 22.8\(^\circ\)C before the peak occupancy load. This “precooling” behavior helps shift the peak load and flatten the control profile. In the previous simulation, the total energy consumption \( J^p \) of the MPC controller increases by 6.3\% compared to the proportional controller. The increase in energy consumption is due to energy losses though the resistance \( R \) while doing precooling. This tradeoff between the total energy losses \( \frac{J^u_{C2} - J^u_{C1}}{J^u_{C1}} \) and the peak power reduction \( \frac{J^p_{C1} - J^p_{C2}}{J^p_{C1}} \) is further explored in Figure 4. The tradeoff lines are generated for different tunings of \( \kappa \) taken from the interval \([0, 5]\) and three different thermal resistances with \( R_0 = 50 \text{ }^\circ\text{C}/kJ. \). The MPC controller can achieve lower peak
power consumption at the cost of higher energy consumption. MPC energy losses with respect to the proportional controller decreases as the thermal resistance $R$ in (1) increases.

The above example shows the benefit from the use of MPC. However, two important elements are not captured by the example. The first one is the complexity of the real problem. Buildings are more complicated than simple RC systems. For details, see “The problem complexity”. The main question is about system abstraction. What is a simple yet descriptive system abstraction that can be used in the context of MPC and bring substantial savings to the real world? The second question is related to the effect of the uncertainty in the predictions. Is it still reasonable to claim benefits of predictive control when model, load, and weather predictions are uncertain?

In the next sections we try to address the first point by providing a brief description of how building cooling and heating systems work, describing abstraction models at different hierarchical levels, and presenting MPC algorithms for each hierarchical level. In the last part of this article we go back to the simple example of this section, and briefly discuss the second point on uncertainty and address future research directions.

This work focuses on a popular hierarchical architecture used in building HVAC systems where cooling and heating rely on centralized chilled water and hot water generation systems, respectively. For the sake of brevity, we concentrate on the cooling side. Despite of different central equipment for heating generation, the heating architecture is similar to cooling. As a result, the abstraction levels and control methodology introduced in this work can be extended to heating systems with minimal effort.

**Cooling Control Systems**

This section describes a building HVAC system where cooling relies on a centralized chilled water generation system. It is also assumed that an energy storage device is available. This article focuses on thermal storage, a stratified water tank in particular. At a lower level of the system hierarchy, subsystems
known as air handling units (AHUs) transfer energy from the distributed chilled water into localized air flows. Those air flows are transported to buildings spaces, delivering cooling or heating energy where required. We first describe the high level chilled water generation and storage, and then the lower level AHU systems.

**High-Level: Components Outside a Building**

Figure 5 depicts the main components of a cooling system based on chilled water generation, storage and distribution. The system can serve either a single building or multiple buildings. The chillers and the cooling towers are responsible for generating the chilled water. The chillers remove heat from the chilled water loop by means of a vapor-compression or absorption refrigeration cycle. The cooling towers have electrical fans so that the condenser water temperature can be lowered by heat conduction and evaporation to the outside air.

The chilled water storage element in Figure 5 can shift the peak cooling demand so that the chillers and cooling towers can run only when it is most efficient to do so. The logic behind the load shifting can depend on a variety of factors, which include time varying utility price, availability of renewable energy, lower ambient temperature, and load shedding signals coming from the utility-grid. The chilled water is distributed through pipelines by using hydraulic pumps. The valve in Figure 5 controls the chilled water flow to either fill the storage element or serve the buildings with the required mass flow rate of chilled water.

Several thermal storage elements are available in the building industry. This article focuses on a stratified tank that can store the chilled water, such as the one that has been installed on the campus of University of California at Merced [16], [17]. Other thermal storage elements use ice balls or concrete slabs to store cooling capacity for later use.
Low-Level: Components Inside a Building

The main equipment used to produce and distribute cool air in a building are air handling units (AHUs), supply fans, and variable air volume (VAV) boxes as depicted in Figure 6. The AHU recirculates return air from building spaces, and mixes it with fresh outside air. The proportion of return air to outside air is controlled by corresponding damper positions in the AHU (Figure 6). The mixed air is cooled by the cooling coils that extract the cooling energy from the chilled water produced by chillers.

The air temperature after the cooling coils depends on the cooling coil valve position that determines the mass flow rate of the chilled water, the temperature of the chilled water, the temperature of mixed air entering the cooling coils, the mass flow rate of the mixed air, and the physical characteristics as well as thermal effectiveness of the cooling coil. Cool air is delivered to the building spaces by electrical fans. Before reaching a given space, the air goes through variable air volume (VAV) boxes. At each VAV box the mass flow rate of the air supplied to the space is adjusted by a damper position. In addition, air temperature can be increased by using reheat coils installed in the VAV box when needed. Space served by one VAV box is denoted by a thermal zone. The delivered air enters a room through diffusers that are designed to fully mix the incoming air with the air in thermal zone.

Abstraction Levels and System Modeling

We follow the structure of the previous sections and present two abstraction levels, each one with associated dynamics, components that use energy, and demands. At the high level the relevant dynamics are those of the storage devices. The building’s cooling or heating demand is modeled as a lumped load, and the main components using energy are chillers and cooling towers. At the low level the relevant dynamics are those of the thermal zones. The loads include the ones generated by occupants, sun, and electrical devices. The components using energy are fans, cooling, and heating coils.
A model of a water tank used for actively storing chilled water is presented and validated with data collected from the campus of the University of California, Merced, USA. The water in the tank is subject to minor mixing, thus the tank can be modeled as a stratified system with layers of warmer water at the top and cooler water at the bottom. Figure 7 depicts the temperature of water measured inside the tank at different heights at 8:30 on November 29th, 2007. A thin layer of water can be observed, known as a thermocline, that has a steep temperature gradient over the height of the tank. For this reason, warmer water above the thermocline and cooler water below the thermocline are lumped up to obtain a four states system describing the heights and temperatures of the warmer and cooler water, respectively.

The tank is assumed to be part of a closed hydraulic loop, that is, the mass flow rate entering the tank is equal to the mass flow rate exiting the tank. Subsequently, the total height of the tank $z_{tank}$ is the sum of the height of warm water $z_a$ and the height of cool water $z_b$ in the tank. The tank dynamics are governed by mass and internal energy conservation laws

$$
\dot{z}_b = (\dot{m}_{CHWS} - \dot{m}_{cmp})/\rho/(\pi r^2), \quad \dot{z}_a + \dot{z}_b = 0, \quad (12)
$$

$$
\dot{U}_a = \dot{H}_a + \dot{Q}_{oa>a} + \dot{Q}_{oa>a}, \quad \dot{U}_b = \dot{H}_b + \dot{Q}_{a>b} + \dot{Q}_{oa>b}, \quad (13)
$$

where $\rho$ is the density of the water, $r$ is the inner radius of the tank, $\dot{m}_{CHWS}$ is the mass flow rate of water supplied to the buildings, and $\dot{m}_{cmp}$ is the one returning from the buildings. $U_* = \rho \pi r^2 z_* c_p T_*$ is the internal energy of the water in the tank, where $* = a$ denotes variables for warmer water, and $* = b$ denotes variables for cooler water. $\dot{Q}_{oa>a}$ is the heat transferred from ambient to the warmer water in the tank, $\dot{Q}_{oa>b}$ is the heat transferred from ambient to the cooler water in the tank

$$
\dot{Q}_{oa>a} = (T_{oa} - T_a)(2\pi r_{tank} z_a)k_1, \quad \dot{Q}_{oa>b} = (T_{oa} - T_b)(2\pi r_{tank} z_b)k_1, \quad (14)
$$
\( \dot{Q}_{a>b} \) is the heat conducted from warmer water to cooler water in the tank and \( \dot{Q}_{a>b} \) is the heat conducted cooler water to warmer water in the tank

\[
\dot{Q}_{a>b} = (T_a - T_b)(\pi r_{\text{tank}}^2)k_2, \quad \dot{Q}_{b>a} = -\dot{Q}_{a>b},
\]

where \( k_1 \) and \( k_2 \) are heat transfer coefficients, \( \dot{H}_a \) is the enthalpy rate for the warm in the tank contributed by water flow and similarly \( \dot{H}_b \) for the cool water.

The thermal storage tank can operate in two modes. When the chilled water flow \( \dot{m}_{\text{CHWS}} \) is greater than the water flow demanded by buildings \( \dot{m}_{\text{cmp}} \), the excessive flow fills the tank. Hence the water flow enthalpy rates are calculated as

\[
\dot{H}_a = -(\dot{m}_{\text{CHWS}} - \dot{m}_{\text{cmp}})c_p T_a, \quad \dot{H}_b = (\dot{m}_{\text{CHWS}} - \dot{m}_{\text{cmp}})c_p T_{\text{CHWS}}.
\]

When the chilled water flow \( \dot{m}_{\text{CHWS}} \) is less than the water flow demanded by buildings, the water in the tank compensates the chilled water supply. Hence the water flow enthalpy rates are calculated as

\[
\dot{H}_a = -(\dot{m}_{\text{CHWS}} - \dot{m}_{\text{cmp}})c_p T_{\text{cmp,r}}, \quad \dot{H}_b = (\dot{m}_{\text{CHWS}} - \dot{m}_{\text{cmp}})c_p T_b,
\]

where \( T_{\text{cmp,r}} \) is the temperature of chilled water returned from buildings.

The simplified model (12)–(17) is validated by using measured data collected from May 22nd–29th, 2007. The measured inputs are applied to the tank model, and the output of the model \([z_a, z_b, T_a, T_b]\) is compared with the measurements in Figure 8. Figure 8(a) shows the tank water temperature validation results. The solid lines are the temperature measurements of top layer water \( T_a \) and bottom layer water \( T_b \) in the tank, and the dotted lines show the predicted cooler and warmer water temperature. The tank model successfully matches the temperature dynamics of the top and bottom layer of the tank water. The tank model also manages to capture the dynamics of the height of the cool water in the tank (Figure 8(b)). More details can be found in [18].
Load Modeling

At the high level, buildings can be modeled as “load demand” elements. A lumped load model predicts the total energy requested by a building based on date, time, occupancy, and weather. In the approach studied in [15], the building load model has two subcomponents, namely, the “Solar and Internal Load Predictor” and the “Building Thermal Load Predictor” in Figure 9.

The Solar and Internal Load Predictor uses time, date, and cloud coverage as its inputs, and calculates inside and outside solar loads as well as internal loads. The outside solar load reflects the solar energy on the outer surface of the building, while the inside solar load is the solar radiation into the building through windows. The internal load includes the heat from people, lights, and equipment.

The Building Thermal Load Predictor predicts the cooling load of buildings, whose main components includes walls and windows. These components are conventionally modeled by RC circuit analogy [19]. The building thermal load model is sketched in Figure 10. $R_1$ represents the thermal resistance of windows. The walls are separated into two layers, where $C_{in}$ and $C_{out}$ capture the thermal capacitance of the wall when influenced by outside and inside solar load, respectively. The thermal resistance between $C_{in}$ and $C_{out}$ is modeled by $R_3$ while $R_2$ and $R_4$ capture the thermal resistance caused by heat convection. The interconnection of the thermal components is also shown in Figure 10. The model inputs are outside air temperature $T_{oa}$, outside solar load $\dot{Q}_{Solar,\text{out}}$, inside solar load $\dot{Q}_{Solar,\text{in}}$, internal load $Q_{\text{internal}}$, and the indoor temperature set-point $T_{sp}$. The internal states of the model are the temperatures of the thermal masses $[T_{in}, T_{out}]$, and the model output is the cooling load demand $\dot{Q}_{Load}$. 
The detailed equations describing the model in Figure 10 are

\[
\begin{align*}
\dot{T}_{\text{in}} &= \dot{Q}_{\text{Solar, in}} + \frac{T_{\text{out}} - T_{\text{in}}}{R_3} + \frac{T_{sp} - T_{\text{in}}}{R_4}, \\
\dot{T}_{\text{out}} &= \dot{Q}_{\text{Solar, out}} + \frac{T_{oa} - T_{\text{out}}}{R_2} + \frac{T_{\text{in}} - T_{\text{out}}}{R_3}, \\
\dot{Q}_{\text{Load}} &= \max(0, \dot{Q}_{\text{internal}} + \frac{T_{\text{in}} - T_{sp}}{R_4} + \frac{T_{oa} - T_{sp}}{R_1}).
\end{align*}
\]

The load model (18)–(20) has been used to model all the buildings on the UC Merced campus where the parameters \(R_1, R_2, R_3, R_4, C_{in}, C_{out}\) have been estimated by using historical data. The load model with estimated parameters is validated by using load measurements from June 1st to June 5th, 2009 at UC Merced. Figure 11 presents the validation results. The measured building load is depicted as the dotted line, and the solid line shows the building load prediction by the building load model (18)–(20). The load dynamics are successfully captured by the simplified model. In general, it might be required to re-identify the parameters on a new set of measured data when the prediction mismatch exceeds desired tolerances [20]. For other types of building load models see [20], [21].

**Main Components Modeling**

At the high level, the main components consuming energy are chillers, cooling towers, and pumps. It is assumed that the setpoints sent to these components are tracked instantly and without errors.

**Chillers:** A simple regression-based centrifugal chiller model can be formulated as in [16]. The chiller model uses polynomial curve fits to predict the chiller capacity and coefficient of performance. The power used by the chillers can be modeled as a static function of the chilled water supply temperature \(T_{CHWS}\), condenser water supply temperature \(T_{CWS}\), mass flow rate of the chilled water supply \(m_{CHWS}\), and chilled water return temperature \(T_{CHWR}\)

\[
P = \text{Chiller}(T_{CHWS}, T_{CWS}, m_{CHWS}, T_{wb}, T_{CHWR}),
\]

12
where $T_{wb}$ is the wet bulb temperature. Historical data can be used to populate this map or improve maps provided by chiller vendors. The model (21) was used to model the chillers at UC Merced. Figure 12 presents the validation results for the chillers from June 1st to June 3rd, 2009. The predicted chillers’ power in solid line successfully captures the measurements in dotted line.

**Cooling Towers:** The cooling tower uses variable speed fans to track a setpoint for the condenser water supply temperature by rejecting the condenser water heat to the ambient environment through evaporation and conduction. A simple cooling tower model [17] employs a regression-based implicit model which links the condenser water supply temperature $T_{CWS}$ and flow rate $\dot{m}_{CWS}$ with the wet bulb temperature $T_{wb}$, fan speed $\omega_{fan}$, and condenser water return temperature $T_{CWR}$

$$f_{CT}(T_{CWS}, T_{CWR}, T_{wb}, \dot{m}_{CWS}, \omega_{fan}) = 0. \quad (22)$$

Given a desired condenser water supply temperature $T_{CWS}$, the fan speed $\omega_{fan}$ can be computed by solving the implicit equation (22) with the information of $T_{wb}$, $T_{CWR}$, and $\dot{m}_{CWS}$. The cooling tower power consumption is approximated as a cubic function of the fan speed

$$P_{ct} = c(\omega_{fan})^3. \quad (23)$$

The model parameters in (22)–(23) can be determined by fitting historical measurements. This model has been applied to the cooling tower at UC Merced, and the validation results are reported in Figure 13.

**Pumps:** When modeling the energy consumption of the pumps, it is assumed that the enthalpy change and pressure difference of water through the pumps are constant. It is also assumed that the pump volumetric water flow $q$, the pump speed $\omega_{pump}$, the pressure difference $\Delta p$ generated by the pump, and their corresponding nominal values denoted with the superscript $0$ satisfy the equations

$$\frac{q}{q^0} = \frac{\omega_{pump}}{\omega_{pump}^0}, \quad \frac{\Delta p}{\Delta p^0} = \left(\frac{\omega_{pump}}{\omega_{pump}^0}\right)^2. \quad (24)$$
The pressure difference \( \Delta p^0(q^0) \) and efficiency \( \eta^0(q^0) \) under nominal operation conditions are approximated as polynomial functions of the nominal volumetric water flow \( q^0 \), and the corresponding coefficients can be obtained by fitting the historical pump performance data. Given the pressure difference \( \Delta p \) and the volumetric water flow rate \( q \), the value of \( \omega^{\text{pump}}_0 \) can be solved in the implicit equation

\[
\Delta p \left( \frac{\omega^{\text{pump}}_0}{\omega^{\text{pump}}} \right)^2 = \Delta p^0 \left( q \frac{\omega^{\text{pump}}_0}{\omega^{\text{pump}}} \right).
\]  
(25)

The power of the pump is calculated as

\[
P^{\text{pump}} = \Delta pq / \eta^0(q \omega^{\text{pump}}_0 / \omega^{\text{pump}}).
\]  
(26)

The pump model (25)–(26) is validated by measured data from the chilled water supply pump at UC Merced. The validation results in Figure 14 suggest that the model manages to correctly predict the power consumption of the pumps.

**Low-Level: Modeling a Building from the Inside**

**Thermal Zones Temperature Dynamics**

In this work, an undirected graph structure is adopted to represent the thermal zones and their dynamic couplings in the following way. The \( i \)-th zone is associated with the \( i \)-th node of a graph, and when an edge \((i, j)\) connecting the \( i \)-th and \( j \)-th node is present, the thermal zones \( i \) and \( j \) are subject to direct heat transfer. The graph \( G \) is defined as

\[
G = (\mathcal{V}, \mathcal{A}),
\]  
(27)

where \( \mathcal{V} = \{1, \ldots, N_v\} \) is the set of nodes or vertices, and \( \mathcal{A} \subseteq \mathcal{V} \times \mathcal{V} \) is the set of edges \((i, j)\) with \( i \in \mathcal{V}, j \in \mathcal{V} \). \( \mathcal{N}^i \) is the set of neighboring nodes of \( i \). Figure 15 shows an example of the graph structure defined in (27), where \( \mathcal{V} = \{1, 2, 3, 4, 5\} \), \( \mathcal{A} = \{(1, 2), (1, 3), (2, 3), (3, 4), (2, 5), (4, 5)\} \), and
Consider a single thermal zone $j \in \mathcal{V}$. The air enters the thermal zone $j$ with a mass flow rate $\dot{m}_s^j$ and temperature $T_s^j$. The temperature of air supplied to thermal zone $j$ ($T_s^j$) is controlled by the cooling $\Delta T_c$ generated at the AHU, the heating $\Delta T_h^j$ from the reheat coils in the VAV box, and the damper position $\delta$ in the AHU system

$$T_s^j = \delta T_r + (1 - \delta)T_{oa} - \Delta T_c + \Delta T_h^j,$$

(28)

where $T_{oa}$ is the outside air temperature, $T_r$ is the return air temperature calculated as weighted average temperature of return air from each thermal zone ($\sum_{i \in \mathcal{V}} \dot{m}_s^i T_i^1 / \sum_{i \in \mathcal{V}} \dot{m}_s^i$), and $\delta$ is the damper position in the AHU. The return air is not recirculated if $\delta = 0$, and no outside fresh air is used if $\delta = 1$. $\delta$ has to be strictly less than one to guarantee minimal ventilation as required by building codes.

A thermal zone can be modeled as a two-mass system. $C_1^j$ is the thermal capacitance of the fast-dynamic mass including air around VAV diffusers, and $C_2^j$ represents the thermal capacitance of the slow-dynamic mass including the solid parts such as floors, walls, and furniture. The appearance of fast and slow temperature dynamics has been observed in [22] for residential buildings. The temperature dynamics in a thermal zone can be modeled as

$$C_1^j \dot{T}_1^j = \dot{m}_s^j c_p (T_s^j - T_1^i) + \frac{T_2^j - T_1^i}{R} + \sum_{i \in \mathcal{N}^j} \frac{T_1^i - T_1^j}{R_{ij}} + \frac{T_{oa} - T_1^j}{R_{oa}} + P_d^j,$$

(29)

$$C_2^j \dot{T}_2^j = \frac{T_1^j - T_2^j}{R},$$

(30)

where $T_s^j$ is defined in (28) while $T_1^j$ and $T_2^j$ are the system states representing the temperature of the lumped masses of $C_1^j$ and $C_2^j$, respectively. The perceived temperature of thermal zone $j$ is assumed to be the temperature of the fast dynamic mass $T^j = T_1^j$. $\mathcal{N}^j$ is the set of neighboring thermal zones of zone $j$, $R_{oa}$ is the thermal resistance between zone $j$ and ambient temperature, and $c_p$ is the specific heat capacity of zone air. $R^j$ models the thermal resistance between $C_1^j$ and $C_2^j$, $R_{ij} = R_{ji}$ models thermal
resistances between zone $i$ and the adjacent zone $j$, and $P^j_d$ is an unmeasured load which is induced by occupancy and solar radiation. Figure 16 depicts the RC network corresponding to the graph $G$ in Figure 15.

The model has been used to model the temperature dynamics of thermal zones in the Bancroft Library located on the campus of University of California at Berkeley, USA. Next we show the results by focusing on one thermal zone, a conference room without windows ($j = 1$). The parameters $p = [C^1_1, C^1_2, R^1, R^1_{12}, R^1_{oa}]$ are identified by using a nonlinear regression algorithm using measured data during weekends when the conference room has no occupants, i.e. $P^1_d = 0$. The result plotted in Figure 17 shows that the model successfully captures the thermal dynamics of the conference room without occupants or solar load. In Figure 17 the solid line depicts the measured room temperature trend, and the dashed line is the room temperature predicted by model (29)–(30) when driven by the measured inputs.

Load

The prediction of thermal zone load $P^j_d(t)$ is important for designing predictive feedback controllers and assessing potential energy savings. A couple of approaches are available in the literature to estimate occupancy load. For instance, the authors of [23] develop an agent-based model to simulate the occupants’ behavior in a building, and the work in [24], [25] focuses on occupancy behavior and occupancy mobility patterns by using wireless camera sensor network.

Time-varying bounds on disturbance load $P_d(t)$ can be learned by computing the mismatch between a nominal model and historical data, and correlating them with shared calendars, weather predictions as well as predicted cloud coverage. This concept can be illustrated by the conference room discussed in the previous section. By using the conference room calendar, we noticed two regularly scheduled group meetings at 10:00 and 14:00 every Wednesday. The meetings can also be identified by inspecting the model mismatch between nominal model predictions and historical data. Figure 18 depicts the envelope-bounded
disturbance load during all Wednesdays in July, 2010. The envelope has been computed as point-wise min and max difference between the measured data and the nominal model described by (29)–(30) with $P_d^1 = 0$. The two peaks in the disturbance load envelope in Figure 18 correspond to two regularly scheduled group meetings.

**Main Components**

The components at the lower level of the architecture that use energy include dampers, supply fans, heating coils, and cooling coils in Figure 6. The supply fan needs electrical power to drive the system while the heating and cooling coils consume the energy of the chilled and hot water. It is assumed that the power to drive the dampers is negligible. A simple energy consumption model for each component is presented next.

**Fan Power:** The fan power can be approximated as a second order polynomial function

$$P_f = c_0 + c_1 \dot{m}_s + c_2 \dot{m}_s^2,$$  \hspace{1cm} (31)

where $c_0$, $c_1$, $c_2$ are parameters to be identified by fitting recorded data, and $\dot{m}_s$ is the mass flow rate of the supply air driven by the supply fan. The simplified fan model (31) is tested by the recorded data from the UC Berkeley Bancroft Library from October 1st to October 10th 2010. The identification results plotted in Figure 19 suggest that the polynomial function successfully predicts the electricity consumption of the fan.

**Cooling and Heating Coils:** Cooling coils and heating coils are air-water heat exchangers. There has been extensive studies to develop simplified yet descriptive models of coil units [26]–[28]. The authors in [26] developed simple empirical equations with four parameters by using a finite difference method to capture the transient response of counterflow heat exchangers. In [27], the authors presented an improved simulation model based on ASHRAE Secondary HVAC Toolkit, and in [28] a simplified control oriented cooling coil unit model is presented based on energy and mass conservation laws.
A even simpler coils model can be derived by assuming that the efficiency of the cooling coil is a constant $\eta_c$ and the efficiency of the heating coil is a constant $\eta_h$. By efficiency of the coil, we mean the efficiency of the coil unit to convert the cooling or heating energy from water-side to air-side of the coil. With this simplification the model is a static energy function of the load on the air-side

$$
P_c = \frac{\dot{m}_{wc}c_p\Delta T_c}{\eta_c COP_c}, \quad P_h = \frac{\dot{m}_{wh}c_p\Delta T_h}{\eta_h COP_h},
$$

(32)

where $P_c$ is the electrical energy related to the chilled water consumed by the cooling coils and $P_h$ is the one related to the hot water consumed by the heating coils. $\dot{m}_{wc}$ is the airflow through the cooling coils, $\dot{m}_{wh}$ the airflow through the heating coils, $c_p$ is the specific heat capacity of room air, $\Delta T_c$ is the temperature difference through the cooling coils, $\Delta T_h$ the temperature difference through the heating coils, $COP_c$ is the chilling coefficient of performance, and $COP_h$ is the heating coefficient of performance. The coefficient of performance (COP) defined as

$$
COP = \frac{E_{\text{thermal}}}{E_{\text{input}}},
$$

(33)

captures the efficiency of the exchange system, i.e., the amount of thermal energy $E_{\text{thermal}}$ (J) generated by the central plant with one Joule of energy consumed $E_{\text{input}}$ by the plant. The input energy $E_{\text{input}}$ can be from different resources such as electricity, fuel, and gas for different systems. Model (32) is oversimplified as compared to the aforementioned literatures. However, the model is adequate to capture the energy consumption if the load on the air side of the coils is calculated, and the coils are operating in a narrow performance range.

**Predictive Control Design with Active Thermal Storage**

It is obvious to control practitioners that as buildings get “smarter” the large number of decoupled local controllers together with coordination strategies based on rules-of-thumb become difficult to design, tune, maintain, and upgrade. For more details on current approaches for building control design, see
“Current Building Operation and Control Logics”. Model predictive control (MPC) naturally enters in the picture as a systematic control methodology that can handle large-scale multi-input multi-output (MIMO) dynamically coupled systems, with performance guarantees, and with the unique capability to explicitly handle the system constraints. The models presented in the previous section can be used to design MPC algorithms at both the high-level and low-level of building control architectures.

A possible MPC architecture for the systems considered in this article, is depicted in Figure 20. A high level MPC (HMPC) is deployed to optimize the operation and schedule the cooling and heating systems with active thermal storage. A low level MPC (LMPC) controls the VAV boxes and the air handling units in Figure 6 by considering thermal comfort constraints of the occupants. At both levels a variety of predictions can be included in the models and in the cost function to run the system in a most efficient and effective way. These predictions include building loads, load shed signals from the power grid, utility prices, weather, occupancy, and solar loads. In addition, HMPC and LMPC can exchange information to achieve better performance. For example, the occupancy load prediction from LMPC can help HMPC achieve better accuracy of building loads prediction. The chilled and hot water temperature predictions from HMPC impose constraints on achievable supply air temperature after the cooling and heating coils, respectively. The efficiency model of the coil units (32) in LMPC can benefit from the prediction of the chilling and heating coefficient of performance $COP_c$ and $COP_h$ from the HMPC.
The following optimization problem is used to describe both the HMPC and LMPC

\[
J^*(x(t), t) = \min_{u_{t[t]}, \ldots, u_{t[N-1]|t}} \sum_{k=0}^{N-1} J(x_{t+k|t}, u_{t+k-1|t}, k) + J_N(x_{t+N|t})
\]

(34)

s.t. \( x_{t+k+1|t} = f(x_{t+k|t}, u_{t+k|t}, d_{t+k|t}, k) \), for all \( k = 0, 1, \ldots, N - 1 \),

(35)

\( y_{t+k|t} = g(x_{t+k|t}, u_{t+k-1|t}, d_{t+k|t}, k) \), for all \( k = 0, 2, \ldots, N \),

(36)

\( y_{t+k|t} \in \mathbb{Y} \), for all \( k = 1, 2, \ldots, N \),

(37)

\( u_{t+k|t} \in \mathbb{U} \), for all \( k = 0, 1, \ldots, N - 1 \),

(38)

\( d_{t+k|t} \in \mathbb{D}(t+k) \), for all \( k = 1, 2, \ldots, N \),

(39)

where \( \mathbb{Y} \) is set of feasible system outputs \( y \), \( \mathbb{U} \) is the feasible set of control inputs \( u \), \( J_N(x) \) is the terminal cost function, \( g(x, u, d, k) \) is the time-varying state update equation, \( d \) is the disturbance, and \( \mathbb{D} \) is the set of possible realizations of the disturbances. Disturbances \( d_k \) can be predicted by a dynamic model such as building load model (18)–(20). An alternative approach is to obtain the future admissible set of disturbances \( \mathbb{D}(t+k) \) by external modules such as the occupant model in Figure 18.

Let \( U^*_{t-t+N-1|t} = \{u^*_{t|t}, \ldots, u^*_{t+N-1|t}\} \) be the optimal solution of problem (34)–(39) at time \( t \). Then, the first element of \( U^*_{t-t+N-1|t} \) is implemented to the system \( u(t) = u^*_{t|t} \). The optimization problem (34)–(39) is repeated at \( t + \Delta t \), with the updated state \( x_{t+\Delta t|t+\Delta t} = x(t + \Delta t) \), yielding a moving or receding horizon control strategy.

The cost function, model dynamics, constraints, and disturbances depend on the abstraction level and the specific problem of interest. Two detailed implementations of HMPC and LMPC scheme are presented in the following sections. We remark that when a nominal model of the disturbances \( d_k \) is replaced by a set valued model, i.e., \( d_k \in \mathbb{D}(k) \) with some probability distribution function, then robust or stochastic MPC formulations need to be used in place of (34)–(39).
Model Predictive Control for High-Level Control

The objective of HMPC is to minimize the electrical energy consumption while generating enough chilled and hot water. A typical cost of the HMPC in (34) penalizes total electricity cost and the deviation from the building demand satisfaction. The cost can be further extended to include the peak load requests, time-varying utility prices, time-varying availability of renewable energy. The control variables include the chilled water supply temperature $T_{CHWS}$, condenser water supply temperature $T_{CWS}$, chilled water supply flow rate $\dot{m}_{CHWS}$, chilling system start time $t_s$, and chilling system end time $t_{end}$. The dynamic system $g(x, u, d, k)$ includes the storage dynamics in (12)–(13), and the disturbance $d$ includes weather and building load demand.

University of California at Merced Experimental Testing

A version of the HMPC controller has been implemented at UC Merced. The detailed experimental setup can be found in [15] and briefly described below. The MPC controller computes the setpoints for cooling towers, chillers, and the thermal storage tank at the central plant. The MPC algorithm is implemented in Matlab and running in real-time on a Pentium 4 Intel processor. The MPC algorithm receives and sends data to the campus through the building automation system “Automated Logics Web Control” (ALC) system.

Two experiments have been studied to evaluate the benefits of MPC compared to conventional controllers. Experiment 1 (E1) is the baseline performance. The plant is operated manually by using the policy defined by the plant managers, and no optimal control algorithm is involved. The control policy is based on the operators’ experience. The data for experiment E1 are collected from May 27th to May 31st, 2009. Experiment 2 (E2) implements the HMPC controller described in the previous section. The data for experiment E2 are collected from Oct. 6th to Oct. 10th, 2009. The quantity of chilled water stored in the tank at the end of the experiment is forced to be equal to those available at the beginning of
the experiment. Despite the difference in time, the weather conditions during experiments E1 and E2 are similar as shown in Figure 21(a). For this reason we can fairly compare the HMPC performance to the baseline control logic. The performance of the cooling system is measured by coefficient of performance (COP) defined by (33).

The results of real time experiment indicate that the performance of the central plant controlled by HMPC, in terms of COP, is improved by 11.9%, which corresponds to a total of $1280 weekly savings compared to their original baseline control E1. The improvement can be explained by the control input profiles plotted in Figure 21. We notice the following differences. First, HMPC in E2 applies higher mass flow rate of chilled water supply (Figure 21(b)) than E1, and it decides to run the chilling system in a different schedule so that the charging window has the lowest ambient temperature. As a result, the combined chillers and cooling tower efficiency is enhanced. Second, HMPC in E2 applies slightly higher chilled water supply temperature, improving the COP of the chilling system according to the performance curve of the chilling system (21). Third, HMPC in E2 applies higher condenser water supply temperature in Figure 21(d). This adjustment enables the chilling system to better balance the loads on chillers and cooling towers. Forth, HMPC makes use of predictive knowledge of the buildings load so that right amount of energy is stored in the thermal tank without overcharging.

Model Predictive Control for Low-Level Control

The objective of LMPC is to minimize the energy consumption in the form of cold and hot water as well as electricity while controlling the thermal zones in the building within the comfort range. The model for this level is described by equations (28)-(30). The control inputs optimized by LMPC are AHU fan speeds, recirculation damper position, cooling and heating coil water valve positions, and zone VAV damper positions. The cost function is the following combination of terms from (31)-(32)

\[
J(x, u, k) = r_e(k)P_f(u) + r_e(k)P_c(u) + r_h(k)P_h(u),
\]
where $r_e(k)$ and $r_h(k)$ refer to the utility rate in dollars per unit energy for electricity and heating fuel, respectively.

Simulation results

To demonstrate MPC results we construct a simple 5-zone building model with input thermal loads as shown in Figure 22. The first four zones have equal and negative load that requires heating except briefly in the afternoon. Zone 5 has high positive load that requires cooling during occupied hours, with a small negative load in unoccupied hours.

The nominal MPC results in Figure 23 show the complicated tradeoff between supply temperature and mass flow rate. Between 6:30 and 10:00, we can see economizer and temperature reset-like operation where cooling of zone 5 is performed using outside air, warmer supply temperatures, and high mass flow rates. This control scheme saves energy because the rest of the zones are in heating mode during this period, and any cooler supply temperature would require reheat to keep those zone temperatures above their lower bounds. Once all of the zones are in cooling mode, controlling the cooling coil to minimal supply temperature and using lower flow rates becomes a more efficient strategy. We see a brief supply temperature reset behavior again near the end of the occupied hours at 18:00. Anticipating less cooling demand for the unoccupied period, the MPC controller starts increasing the supply temperature early.

To compare against the nominal case MPC results in Figure 23(a) and Figure 23(b), we repeat the calculations with modified versions of the cost function. First, in Figure 24(a), we modify the electric utility rate $r_e(k)$ to have a higher value between 12:00 and 16:30. In Figure 24(b) the utility rates are constant throughout, but we add an additional penalty term to the cost function to minimize the peak electric power over the entire day. The modified cost function is

$$J_{\text{mod}}(x, u, k) = r_e(k)P_f(u) + r_e(k)P_c(u) + r_h(k)P_h(u) + \varphi \max_k(P_f(u) + P_c(u)), \quad (41)$$
where $\varphi$ is a penalty weighting factor in dollars per unit power. Both these modified cases in figures 24(a) and 24(b) demonstrate precooling of zone 5 and lengthened cooling of zones 1-4, but with different timing and intent. In Figure 24(a), the peak electric power is not penalized in the cost function but the electric utility rate has a higher value between 12:00 and 16:30. As a consequence, precooling is only performed immediately before noon, with a corresponding spike in cooling power, so that less cooling energy is used between 12:00 and 16:30. In Figure 24(b), the peak electric power is included in the cost function so zone 5 is precooled beginning earlier in the morning. As a result, cooling power is increased at a time when it would normally be low, shifting electric power use away from the times it would normally be at maximum.

**MPC Main Issues and Current Research**

The appealing advantages of MPC shown through simulations and experiments in the previous sections do not come without a price. Several issues have to be considered while designing and implementing MPC for buildings. Among them, the stability and feasibility issues of MPC [6] are well known. In particular, it is well known that stability and feasibility are not ensured by the MPC law without terminal cost and terminal constraints [6]. Usually the MPC problem (34)–(39) is augmented with a terminal cost and a terminal constraint. Typically the terminal set is a robust control invariant set so that the persistent feasibility of the MPC strategy is guaranteed. Persistent feasibility ensures that if the MPC (34)–(39) is feasible for a given initial state $x(0)$, then it is feasible for all $t \geq 0$. Definitions and properties of invariant sets can be found in [6], [29]. In the specific context of the MPC considered in this paper, the terminal set ensures that enough energy is actively stored in thermal storage elements to counteract a bounded unpredicted change in demand. A treatment of sufficient conditions which guarantee persistent feasibility of MPC problems goes beyond the scope of this work and can be found in the survey [6] and in [16] for the specific case of the UC Merced study.
In addition to the aforementioned issues, the computational complexity of the MPC law, the convergence to suboptimal solutions, and the role of prediction uncertainty are also critical for building controls design.

**Computational Complexity of Model Predictive Control.**

As the complexity of the buildings increases, centralized MPC might become computationally intractable due to the limited computational resources available on current building control platforms. This limitation is especially true at the low level where distributed cheap computing platforms are common.

This limitation might be overcome by efficient numerical solvers tailored to the specific hardware or with the use of distributed model predictive control [30], [31]. In distributed MPC, the centralized problem is decomposed on a set of smaller problems which can be associated with different subsystems such as VAV boxes and AHUs. Each subsystem solves local small MPC problems with information from local and neighboring subsystems. The local MPC controllers communicate with each other to achieve globally optimal solutions [31].

**Convergence to Suboptimal Solutions.**

The product between air temperatures and mass flow rates in the thermodynamic energy balance equations (29)–(30) leads to a non-convex MPC problem which might have distinct locally optimal solutions. Fast computational techniques for solving non-convex optimization such as sequential quadratic programming (SQP) can only provide certificates of local optimality. These local optimal solutions might be less efficient than those obtained with simpler control design. We are currently analyzing different types of local optima and their physical interpretation. The analysis can be used to derive branch and bound rules which allow an SQP solver to converge to globally optimal control sequences.
**Prediction Uncertainty.**

The example in the first section of this article showed benefits of MPC under the assumption that MPC has prefect knowledge of predicted disturbances and system dynamics. This section tries to highlight potential issues associated with this assumption. We focus on total energy consumption by using the simple MPC problem (3)–(6) with \( \kappa = 0 \). The control design assumes that weather prediction in Figure 1(a) is perfect and occupancy load prediction in Figure 1(b) is perfect. This time we assume that in reality the occupancy load differs from what predicted. Two scenarios are considered. In scenario S1 the future occupancy load is exactly the same as predicted in Figure 1(b) with probability \( P(S1) \) equal to \( \alpha \). In scenario S2 the occupancy load is zero over the entire day with probability \( P(S2) \) equal to \( 1 - \alpha \). In short, the controller is designed based on S1 but the probability of S1 happening is \( \alpha \).

The expected control input cost \( \mathbb{E}[J_u] \) and constraints violation \( \mathbb{E}[J^e] \) for MPC controller C2 and the proportional controller C1 are computed in closed-loop. The closed-loop simulations use different occupancy load profiles depending on the chosen probability \( \alpha \).

Simulation results for various values of \( \alpha \) and various tunings for MPC controller C2 as well as the proportional controller C1 are summarized in Figure 25. When the prediction is perfect \( \alpha = 1 \), the performance of the MPC controller C2 is the same as the proportional controller C1 in terms of total energy consumption and discomfort. However, the MPC performance deteriorates as \( \alpha \) decreases. In fact, the MPC controller fails to keep the room temperature within the comfort constraints due to the misleading predictions. Also the MPC controller consumes more energy than the proportional controller for \( \alpha = 0.5 \) and \( \alpha = 0 \) because the MPC controller is precooling even when occupants are not expected to enter the space.

Stochastic MPC [32], [33] might be a better approach to address this issue when probability
distribution functions of the loads are available. In this case, we would minimize expected costs and satisfy constraints with a given probability. We are currently investigating this research direction and its real-time computational complexity.
References


Figure 1. Predicted Ambient Temperature and Load Profiles. The ambient temperature is the measured data from University of California at Merced on July 3rd, 2009. The load profile is artificial, and it is assumed that a group meeting is scheduled between 8:00 and 11:00 everyday with a constant 3 kW occupancy load.
Figure 2. Comparing MPC controller (C2) and proportional controller (C1). Observe that C1 and C2 use the same energy $J_u^{C1} = J_u^{C2}$ for the same amount of constraint violation $J_\varepsilon^{C1} = J_\varepsilon^{C2}$. As expected, increased comfort violation corresponds to a lower energy usage for both controllers.
Figure 3. Closed-Loop Simulations when Using MPC and Proportional Control. The peak power consumption $J^p$ of the proportional controller is reduced by 89% when the MPC control is used. For both controllers the comfort violation indexes $J^c$ are zero. This nice behavior is obtained by taking advantage of the predictive information of the disturbance and using the space thermal storage. In fact the MPC precools the space temperature to $22.8^\circ C$ before the peak occupancy load. This “precooling” behavior helps shifting the peak load and flatten the control profile.
Figure 4. The total energy consumption $J_p$ of the MPC controller (C2) increases when compared to the proportional controller (C1). This increased energy consumption is due to energy losses though the resistance $R$ while doing precooling. This tradeoff between the total energy losses ($\frac{J_p^{C_2} - J_p^{C_1}}{J_p^{C_1}}$) and the peak power reduction ($\frac{J_p^{C_1} - J_p^{C_2}}{J_p^{C_1}}$) is depicted in this figure for different MPC tunings and three different thermal resistances with $R_0 = 50 \degree C/kJ$. The MPC controller can achieve lower peak power consumption at the cost of higher energy consumption. MPC energy losses with respect to the proportional controller decreases as the thermal resistance $R$ in (1) increases.
Figure 5. Schematic of a cooling system based on chilled water generation, storage and distribution. The system can serve either a single building or multiple buildings. The chillers and the cooling towers are responsible for generating the chilled water. The chilled water storage element can shift the peak cooling demand so that the chillers and cooling towers can run only when it is most efficient to do so. The logic behind the load shifting can depend on a variety of factors, which include time varying utility price, availability of renewable energy, lower ambient temperature, and load shedding signals coming from the utility-grid. The chilled water then is distributed through pipelines by using hydraulic pumps. The valve controls the chilled water flow to either fill the storage element or serve the buildings with the required mass flow rate of chilled water.
Figure 6. Schematics of the air distribution system. The main equipment used to produce and distribute cool air in a building are air handling units (AHUs), supply fans, and variable air volume (VAV) boxes. The AHU recirculates return air from building spaces, and mixes it with fresh outside air. The proportion of return air to outside air is controlled by corresponding damper positions in AHUs. The mixed air then is cooled by the cooling coils that extract the cooling energy from the chilled water produced by chillers.
Figure 7. The water in stratified storage tanks is subject to minor mixing. This figure depicts the temperature of water measured inside a tank at different heights. A thin layer of water can be observed, known as a thermocline, that has a steep temperature gradient over the height of the tank.
Figure 8. The tank model (12)–(17) is validated by using measured data collected from the tank at UC Merced. We applied the measured inputs to the tank model, and the output of the model \([z_a, z_b, T_a, T_b]\) is compared with the measurements. The figure shows the tank water temperature validation results. The solid lines are the temperature measurements of top layer water \(T_a\) and bottom layer water \(T_b\) in the tank, and the dotted lines show the predicted cool and warm water temperature. The simplified tank model successfully matches the temperature dynamics of the top and bottom layer of the tank water. The tank model also manages to capture the dynamics of the height of the cool water in the tank.
Figure 9. The building load model has two subcomponents, namely, the “Solar and Internal Load Predictor” and the “Building Thermal Load Predictor”. The Solar and Internal Load Predictor uses time, date, and cloud coverage as its inputs and calculates inside and outside solar loads and internal loads. The Building Thermal Load Predictor predicts the cooling load of buildings whose main components includes walls, and windows.
Figure 10. Building Thermal Load Model. \( R_1 \) represents the thermal resistance of windows. The walls are separated into two layers, where \( C_{in} \) and \( C_{out} \) capture the thermal capacitance of the wall when influenced by outside and inside solar load, respectively. The thermal resistance between \( C_{in} \) and \( C_{out} \) is modeled by \( R_3 \) while \( R_2 \) and \( R_4 \) capture the thermal resistance caused by heat convection. The model inputs are outside air temperature \( T_{oa} \), outside solar load \( \dot{Q}_{Solar,out} \), inside solar load \( \dot{Q}_{Solar,in} \), internal load \( Q_{internal} \), and the indoor temperature set-point \( T_{sp} \). The model output is the cooling load demand \( \dot{Q}_{Load} \).
Figure 11. Campus load validation. The measured building load is depicted as the dotted line, and the solid line shows the building load prediction by the building load model (18)–(20).
Figure 12. Chillers model validation. The dotted line shows the electricity power consumption of the two chillers installed at University of California, Merced, and the solid line is the predicted power consumption of the chillers by the simplified model (21).
Figure 13. Cooling tower model validation. The dotted line depicts the measured electricity power consumption of cooling towers located at University of California, Merced, and the solid line shows the predicted power consumption of cooling towers by model (22)–(23).
Figure 14. Pumps model validation. The dotted line plots the measured electricity power consumption of hydraulic pump 2 in Figure 5, and the solid line shows the predicted power consumption of hydraulic pump 2 by model (25)–(26).
Figure 15. Graph structure of thermal zones. We use an undirected graph structure to represent the thermal zones and their dynamic couplings in the following way. We associate the $i$-th zone to the $i$-th node of a graph, and when an edge $(i, j)$ connecting the $i$-th and $j$-th node is present, then the thermal zones $i$ and $j$ is subject to direct heat transfer. In this example, the vertices are $\mathcal{V} = \{1, 2, 3, 4, 5\}$, the edges are $\mathcal{A} = \{(1, 2), (1, 3), (2, 3), (3, 4), (2, 5), (4, 5)\}$, and the neighboring nodes of the first node are $\mathcal{N}^1 = \{2, 3\}$. 
Figure 16. Network of two-mass thermal zone model to the graph depicted in Figure 15. $C^j_1$ is the thermal capacitance of the fast-dynamic mass including air around VAV diffusers, and $C^j_2$ represents the thermal capacitance of the slow-dynamic mass including the solid parts such as floors, walls, and furniture. $R^j$ models the thermal resistance between $C^j_1$ and $C^j_2$, $R_{ij} = R_{ji}$ models thermal resistances between zone $i$ and the adjacent zone $j$, and $P^j_d$ is a current source modeling an unmeasured load induced by occupancy and solar radiation.
Figure 17. Identification results of the thermal zone model (29)–(30). The solid line is the measured temperature collected from a conference room in the Bancroft Library located on the campus of University of California at Berkeley, USA, on July 4th, 2010. The dashed line is predicted temperature by the simplified thermal zone model (29)–(30) when driven by the measured inputs. The result shows that the model successfully captures the thermal dynamics of the conference room without occupants or solar load.
Figure 18. Envelope-bounded disturbance load during all Wednesdays in July 2010 in the conference room of the Bancroft Library at UC Berkeley. The envelope is computed as point-wise min and max difference between the measured data and the nominal model. The two peaks in the disturbance load envelope correspond to two regularly scheduled group meetings.
Figure 19. A Fan Identification Result. The cross data points are measured fan power consumption collected from Bancroft library, and the solid line is the predicted fan power by model (31).
Figure 20. Hierarchical MPC structure for building control systems. A high level MPC (HMPC) is deployed to optimize the operation and schedule the cooling (heating) systems with active thermal storage. A low level MPC (LMPC) controls the VAV boxes and the Air Handling Units by considering thermal comfort constraints of the occupants. At both levels a variety of predictions can be included in the models and in the cost function to run the system in a most efficient and effective way. These predictions include building loads, load shed signals from the power grid, utility prices, weather, occupancy, and solar loads.
Figure 21. Experimental results of the central cooling plant with thermal storage at UC Merced controlled by the HMPC (E2). The plant Coefficient of Performance is improved by 11.9%, which corresponds to a total of $1280 weekly savings compared to their original baseline control logic (E1). We notice that HMPC applies higher mass flow rate of chilled water supply, and it decides to run the chilling system in a different schedule so that the charging window has the lowest ambient temperature. As a result, the combined chillers and cooling towers efficiency is increased. Also the HMPC applies slightly higher chilled water supply temperature, improving the COP of the chilling system according to the performance curve of the chilling system.
Figure 22. Zone thermal loads $\dot{Q}_i$. The first four zones have equal and negative load that requires heating except briefly in the afternoon. Zone 5 has high positive load that requires cooling during occupied hours, with a small negative load in unoccupied hours.
Figure 23. Simulation results of five thermal zones controlled by MPC. The results show a complicated tradeoff between supply temperature and mass flow rate. Between 6:30 and 10:00, we can see economizer and temperature reset-like operation where cooling of zone 5 is performed using outside air, warmer supply temperatures, and high mass flow rates. This control scheme saves energy because the rest of the zones are in heating mode during this period, and any cooler supply temperature would require reheat to keep those zone temperatures above their lower bounds. Once all of the zones are in cooling mode, controlling the cooling coil to minimal supply temperature and using lower flow rates becomes a more efficient strategy. We see a brief supply temperature reset behavior again near the end of the occupied hours at 18:00. Anticipating less cooling demand for the unoccupied period, the MPC controller starts increasing the supply temperature early.
Figure 24. Simulation results of five thermal zones controlled by MPC. We modify the electric utility rate to have a higher value between 12:00 and 16:30. In the right plot the utility rates are constant throughout, but we add an additional penalty term to the cost function equal to the maximum electric power over the entire day. Both figures demonstrate precooling of zone 5 and lengthened cooling of zones 1-4, but with different timing and intent. In the left plot, peak electric power is not penalized in the cost function but the electric utility rate has a higher value between 12:00 and 16:30. As a consequence, precooling is only performed immediately before noon, with a corresponding spike in cooling power, so that less cooling energy is used between 12:00 and 16:30. In the right plot the peak electric power is included in the cost function so zone 5 is precooled beginning earlier in the morning. As a result, the cooling power is increased at a time when it would normally be low, shifting electric power use away from the times it would normally be at maximum.
Figure 25. The MPC controller is designed based on a certain prediction of occupancy profile. The probability of the occupancy profile being correct is $\alpha$. The figure shows the expected control input cost $\mathbb{E}[J^u]$ and constraints violation $\mathbb{E}[J^\varepsilon]$ for the MPC controller C2 and for the proportional controller C1 in closed-loop. The closed-loop simulations use different occupancy load profiles depending on the chosen probability $\alpha$. When the prediction is perfect ($\alpha = 1$), the performance of the MPC controller is the same as the proportional controller in terms of total energy consumption and discomfort. However, the MPC performance deteriorates as $\alpha$ decreases. In fact, the MPC controller fails to keep the room temperature within the comfort constraints due to the misleading predictions. Also the MPC controller consumes more energy than the proportional controller for $\alpha = 0.5$ and $\alpha = 0$ because the MPC controller is precooling even when occupants are not expected to enter the space.
Sidebar 1: The Problem Complexity

Building control design is not straightforward for a long list of reasons which include the following.

**Dynamics.** Building Heating, ventilation and air Conditioning (HVAC) systems convert and transport energy through working fluids, air, and water. The flow dynamics of air and water through distribution networks are described by nonlinear partial differential equations, specifically the Navier-Stokes equations. The computational fluid dynamics technique is extremely computationally intensive and requires complete geometry description at all length scales. This level of detail is rarely available for a real building. A more common approach is to approximate the velocity, temperature, and pressure distributions with reduced order lumped nodal models.

**Component characterization.** HVAC systems are comprised of several components, some of which are described in details later. Each component has distinctive and potentially complicated behavior often described by nonlinear equipment performance curves.

**Building design configuration.** HVAC design practices have resulted in a variety of different component arrangements. A handful of standard configuration types are more common than others, but virtually every building is unique. Therefore, the spatial locations, type of components, and the methods used to implement a control action are highly dependent on the specific HVAC system. For instance, overhead air distribution systems use a different set of actuators from under-floor air distribution systems, and both differ from systems that use water-based radiators for conditioning.

**Comfort.** In this work, we treat occupant thermal comfort as being equivalent to a specific range of spacial air temperatures. A large body of ASHRAE and other literature [34], [35] have introduced more complicated representations of occupant comfort. These more detailed comfort models take into
account metabolism and biological factors, air velocity, humidity, heat transfer through radiation, free convection, and other effects [36].

**Control requirements.** The goal for controlling a building HVAC system is to maintain occupant comfort, preferably at low energy cost. In addition to temperature regulation, HVAC controllers can have additional requirements on humidity regulation, proportions of fresh versus recirculated air for indoor air quality, flow rates for ventilation, and pressurization of spaces.
Sidebar 2: Current Building Operation and Control Logic

Most modern buildings employ some level of automated control. In certain cases the control logic may be complex and optimized, but in the majority of cases, building systems are controlled by basic control logic that err on the side of simplicity over subtlety. These simple control logic are implemented with distinct but interconnected proportional-integral-derivative (PID) control loops and switching logic. These logic respond to setpoints and schedules for building components such as chillers and cooling towers. Controlled parameters include the chilled water supply mass flow rate, chilled water temperature, condenser water mass flow rate, condenser water temperature, operation schedule of the chilling system, and the storage level of the thermal storage element.

Advanced decision systems are available on the market to optimize the high-level system based on components modeling, feedback, and forecasts. A variety of proprietary control sequences for chillers, boilers, and cooling towers are available in the building industry. However, to the best of authors’ knowledge, their implementation is not widespread and often limited to specific configurations and components of the cooling and heating systems.

At the low level, current practice is to use single-input, single-output (SISO) cascaded controllers to achieve a stable behavior and modest disturbance rejection. To this aim, controllers are designed locally, independently, with the goal of decoupling the several components involved.

A classical example for control logic architectures used for AHU and VAV boxes is depicted in Figure S1. In each thermal zone, the control system generates a cooling or a heating request when the zone temperature is higher or lower than the setpoint, respectively. Each individual request is used to adjust setpoints for the corresponding VAV box. Also, the sum of all requests is used to adjust setpoints for the AHU.
The setpoint adjustment usually follows proportional laws as depicted in Figure S2 for a VAV box. When the zone temperature is lower than the comfort range denoted by the dashed lines in Figure S2, the heating coil valve position is proportional to the difference between zone temperature and the lower-bound on comfort level, and the air mass flow rate is set to be at the minimal ventilation level. When the zone temperature is within the comfort set, the VAV box keeps the minimal ventilation level. When the zone temperature is higher than the comfort range, the supply airflow rate increases proportionally with the difference between zone temperature and the upper-bound on comfort level. The total number of cooling requests is used to determine the total cooling energy required to guarantee the thermal comfort. The AHU supply fan is often operated to track the air pressures in VAV boxes to ensure that enough air flow can be provided to each zone.

In addition to basic control logic, a myriad of advanced heuristics are used to reduce the overall energy consumption. For instance but in certain combinations of ambient conditions and zone thermal demands, energy use can be lowered by increasing the proportion of outside air flow. This control logic is known as “economizer” operation.

“Supply temperature reset” is another strategy. Supply temperature reset is implemented to reduce the chilled water consumption. When the zone temperatures are within the comfort constraints, the AHU supply temperature setpoint is increased slowly until one of the zones flags a cooling request. This modification to the supply air temperature setpoints enables the cooling coils to consume less chilled water, thus reducing the energy consumption of the chillers and cooling towers. However, when supply temperature is raised, zones in cooling mode require a higher flow rate to maintain the same zone temperature. Therefore when a zones is in cooling mode, the energy saved is counterbalanced by increased fan energy. In this case energy prediction models need to be used to assess the validity of “supply
temperature reset” strategies.
Figure S1. Current control logic schematics. In each thermal zone, the control system will generate a cooling or heating request if the zone temperature is higher or lower than the set point, respectively. Each individual request will be used to adjust set points for the corresponding VAV box. The total number of cooling and heating requests is used to determine the total cooling and heating energy required to guarantee the thermal comfort. The AHU supply fan is operated to track the air pressures in VAV boxes to ensure that enough air flow can be provided to each zone.
Figure S2. Simple proportional control logic. When the zone temperature is lower than the comfort range denoted by the dashed lines in Figure S2, the heating coil valve position is proportional to the difference between zone temperature and the lower-bound on comfort level, and the air mass flow rate is set to be at the minimal ventilation level. When the zone temperature is within the comfort set, the VAV box keeps the minimal ventilation level. When the zone temperature is higher than the comfort range, the supply airflow rate increases proportionally to the difference between zone temperature and the upper-bound on comfort level.

AUTHOR INFORMATION

Yudong Ma received the B.S in the department of mechanical engineering with a minor in mathematics from Tsinghua University, China, in 2007. He is currently a Ph.D. student in Mechanical Engineering at the University of California, Berkeley, conducting research on applications of model predictive control for energy efficiency.

Address: 2169 Etcheverry hall, 2521 Hearst ave, Berkeley, CA, 94709.

Email: myd07@berkeley.edu.

Anthony Kelman received his B.S. in Mechanical Engineering from the California Institute of Technology in 2007. From 2007 to 2009 he worked on dynamic modeling and controls development of aircraft environmental control systems at Honeywell Aerospace in Torrance, California. He is currently a Ph.D. student in Mechanical Engineering at the University of California, Berkeley, conducting research on applications of model predictive control for energy efficiency.

Allan Daly is a registered mechanical engineer and a Principal at Taylor Engineering, LLC, designing and implementing innovative and sustainable mechanical systems. He specializes in energy efficient and environmentally responsible HVAC system designs that maximize occupant health and comfort, and his current work is focused in institutional projects, commissioning work, energy modeling, and green building projects. They includes the new University of California at Merced, which will be an entire campus of LEED Silver rated buildings. Daly teaches and lectures widely to both professional and academic audiences on energy efficiency, integrated design, building simulation, and UFAD system design. He is a member of the US Green Building Council Energy and Environment Technical Advisory Group, helping to guide the development of LEED 2012.

Francesco Borrelli received the ‘Laurea’ degree in computer science engineering in 1998 from
the University of Naples ‘Federico II’, Italy. In 2002 he received the PhD from the Automatic Control Laboratory at ETH-Zurich, Switzerland. He is currently an Assistant Professor at the Department of Mechanical Engineering of the University of California at Berkeley, USA. He is the author of more than sixty publications and in the field of predictive control. He is author of the book Constrained Optimal Control of Linear and Hybrid Systems published by Springer Verlag, the winner of the ‘Innovation Prize 2004’ from the ElectroSwiss Foundation and the winner of the 2009 NSF CAREER award. In 2008 he was appointed the chair of the IEEE technical committee on automotive control. His research interests include constrained optimal control, model predictive control and its application to advanced automotive control and energy efficient building operation.