Stochastic Modeling and Approaches for Managing Energy Footprints in Cloud Computing Service

Siqian Shen
Assistant Professor

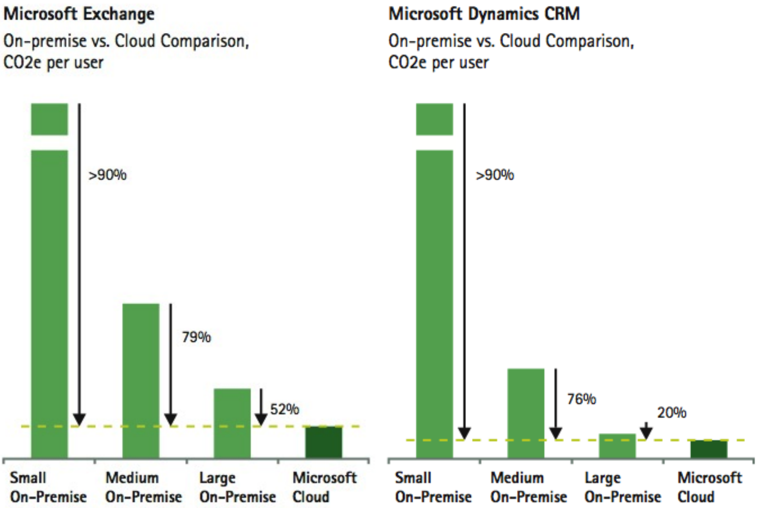
Industrial and Operations Engineering
University of Michigan

October 8, 2013

Emerging Trends of Cloud Computing (CC)

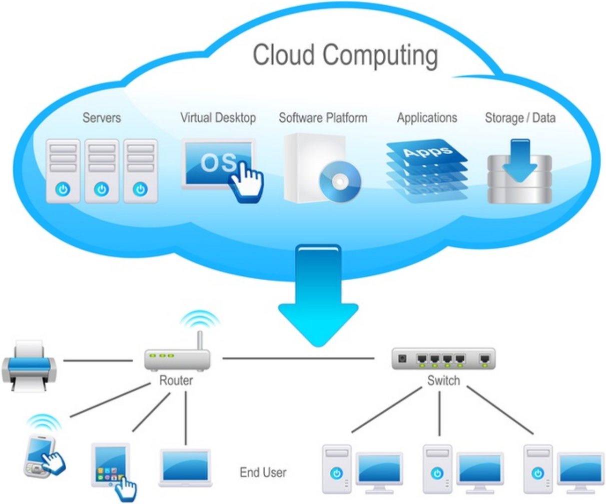
Source: www.cloudtweaks.com by David Fletcher

CC Advantages: Reducing Carbon Emission



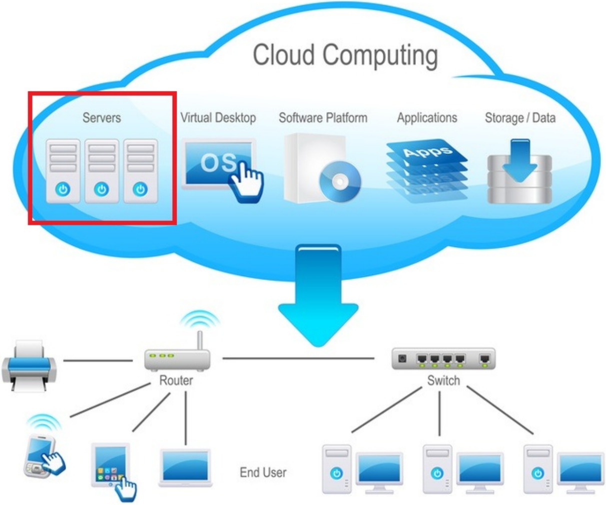
Source: Accenture (2010) “Cloud Computing and Sustainability: Environmental Benefits of Moving to the Cloud”

How CC Works...

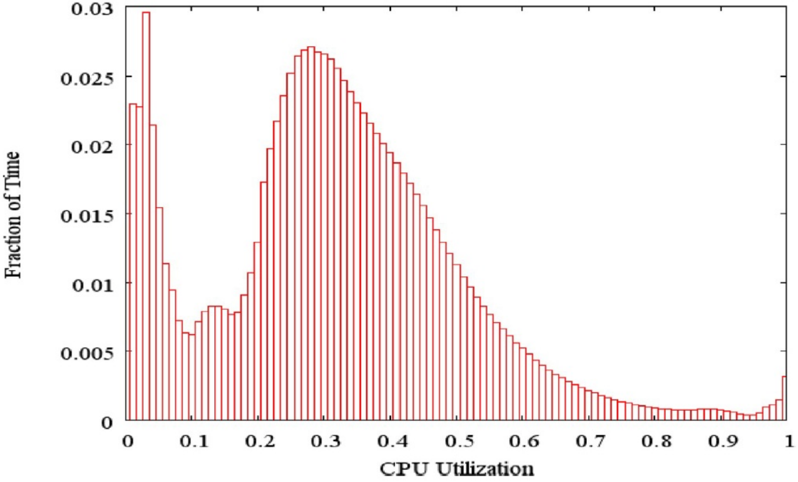


Source: www.veterangeek.com

How CC Works...



Source: www.veterangeek.com



- Moreover, an idle server consumes 60%+ energy at full mode.

Virtual Machine Consolidation

Large-scale servers with low utilization

Consolidate the work to fewer Cloud servers

Source: Google's official blog - Energy efficiency in the cloud.

Our data centers use 50% less energy than typical data centers through server (Virtual Machine) consolidation. — Google.

Other benefits:

- more robust operations schedules
- more idle servers reacting to demand surges

Our Work

- Stochastic mixed-integer programming models to optimize energy footprints while ensure various Quality of Service (QoS) guarantees for managing servers in Cloud Computing service.

Our Work

- Stochastic mixed-integer programming models to optimize energy footprints while ensure various Quality of Service (QoS) guarantees for managing servers in Cloud Computing service.
- Estimate demand based on distributions of historical data, and dynamically consolidate or distribute jobs on servers through operational scheduling.

Our Work

- Stochastic mixed-integer programming models to optimize energy footprints while ensure various Quality of Service (QoS) guarantees for managing servers in Cloud Computing service.
- Estimate demand based on distributions of historical data, and dynamically consolidate or distribute jobs on servers through operational scheduling.
- Vary QoS levels by using joint/multiple chance constraints, to bound chances of job delay and incompleteness.

Our Work

- Stochastic mixed-integer programming models to optimize energy footprints while ensure various Quality of Service (QoS) guarantees for managing servers in Cloud Computing service.
- Estimate demand based on distributions of historical data, and dynamically consolidate or distribute jobs on servers through operational scheduling.
- Vary QoS levels by using joint/multiple chance constraints, to bound chances of job delay and incompleteness.

Outline of Our Research

- Formulations: Stochastic & Chance-Constrained Programs
- Algorithms: the Benders Decomposition and Heuristics
- Computational Design
- Result Analyses
- Conclusions and Future Research

$\mathcal{N}_m$ set of servers in a data center

$\mathcal{N}_m$	set of servers in a data center
Ω	set of finite scenarios for realizing uncertain demand

$\mathcal{N}_m$	set of servers in a data center
Ω	set of finite scenarios for realizing uncertain demand
T	total number of time periods considered
ℓ^t	length of period t (in hours) for all $t = 1, \ldots, T$

$\mathcal{N}_m$	set of servers in a data center
Ω	set of finite scenarios for realizing uncertain demand
T	total number of time periods considered
ℓ^t	length of period t (in hours) for all $t = 1, \ldots, T$
$\widetilde{d}^t$	random job requests (demand) received at period t

$\mathcal{N}_m$	set of servers in a data center
Ω	set of finite scenarios for realizing uncertain demand
T	total number of time periods considered
ℓ^t	length of period t (in hours) for all $t = 1, \ldots, T$
$\widetilde{d}^t$	random job requests (demand) received at period t

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \widetilde{d}^t \quad \forall 1 \leq t \leq T && \text{(1b)} \\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1c)} \\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)} \\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\, 2 \leq t \leq T && \text{(1e)} \\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1f)} \\
& y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

The basic model consolidates demand on severs to minimize the total energy consumed by all servers over $t = 1, \ldots, T$.

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \widetilde{d}^t \quad \forall 1 \leq t \leq T && \text{(1b)} \\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1c)} \\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)} \\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\ 2 \leq t \leq T && \text{(1e)} \\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1f)} \\
& y_i^t, z_i^t \in \{0,1\} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

$g_i y_i^t$: energy used for booting machine i at period t.

$y_i^t \in \{0,1\}$: $= 1$ if server i is switched to "on" at period t.

$$\text{min:} \quad \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) \tag{1a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \widetilde{d}^t \quad \forall 1 \leq t \leq T \tag{1b}$$

$$\ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{1c}$$

$$y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m \tag{1d}$$

$$y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\ 2 \leq t \leq T \tag{1e}$$

$$0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{1f}$$

$$y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{1g}$$

$v_i x_i^t$: energy for job processing in machine i at period t.
$x_i^t \geq 0$: percentage of server i's capacity used at period t.

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \widetilde{d}^t \quad \forall 1 \leq t \leq T && \text{(1b)} \\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1c)} \\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)} \\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\, 2 \leq t \leq T && \text{(1e)} \\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1f)} \\
& y_i^t, z_i^t \in \{0,1\} \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

$f_i z_i^t$: energy used at "idle" of machine i at period t.
$z_i^t \in \{0,1\}$: $= 1$ if server i is "idle" at period t.

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \widetilde{d}^t \quad \forall 1 \leq t \leq T && \text{(1b)} \\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1c)} \\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)} \\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\ 2 \leq t \leq T && \text{(1e)} \\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1f)} \\
& y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

Computational time allocated to each period t is no less than the random demand $\widetilde{d}^t$.

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)}\\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \widetilde{d}^t \quad \forall 1 \leq t \leq T && \text{(1b)}\\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1c)}\\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)}\\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\, 2 \leq t \leq T && \text{(1e)}\\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1f)}\\
& y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

If $\widetilde{d}^t$ is discretely distributed,

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \max_{\omega \in \Omega} d^{t\omega} \quad \forall 1 \leq t \leq T && \text{(1b)} \\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1c)} \\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)} \\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\ 2 \leq t \leq T && \text{(1e)} \\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1f)} \\
& y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

If $\widetilde{d}^t$ is discretely distributed, and let $d^{t\omega}$ represent a realization of $\widetilde{d}^t$ in scenario $\omega \in \Omega$,

- reformulate (1b) as a set of deterministic constraints

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \max_{\omega \in \Omega} d^{t\omega} \quad \forall 1 \leq t \leq T && \text{(1b)} \\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1c)} \\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)} \\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\, 2 \leq t \leq T && \text{(1e)} \\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1f)} \\
& y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m,\, 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

The total “on” time of server i at period t is no less than computational time plus the time of booting the server (if there is any).

$$\text{min:} \quad \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) \tag{1a}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \max_{\omega \in \Omega} d^{t\omega} \quad \forall 1 \leq t \leq T \tag{1b}$$

$$\ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{1c}$$

$$y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m \tag{1d}$$

$$y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\ 2 \leq t \leq T \tag{1e}$$

$$0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{1f}$$

$$y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{1g}$$

Server i is "on" at period 1 if we switch it to "on."

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) && \text{(1a)} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}_m} e_i \ell^t x_i^t \geq \max_{\omega \in \Omega} d^{t\omega} \quad \forall 1 \leq t \leq T && \text{(1b)} \\
& \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m, \, 1 \leq t \leq T && \text{(1c)} \\
& y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m && \text{(1d)} \\
& y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m, \, 2 \leq t \leq T && \text{(1e)} \\
& 0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m, \, 1 \leq t \leq T && \text{(1f)} \\
& y_i^t, z_i^t \in \{0, 1\} \quad \forall i \in \mathcal{N}_m, \, 1 \leq t \leq T && \text{(1g)}
\end{aligned}
$$

If server i is "off" at $t-1$ but "on" at t, then it means that

- server i is switched to "on" at period t

GOAL:

- Minimize energy consumption of all servers over $1, \ldots, T$ + the expected penalty cost of backlogging.

Allow **backlogging** such that

- Job (j, t) can be partitioned and processed on multiple servers, at any time that is no more than L periods after period t ("time of submission").

Define Sets:

- $B_1(t)$: backlogging periods such that if $t = 1, \dots, T - L$, then $B_1(t) = t, \dots, t + L$; if $t = T - L + 1, \dots, T$, then $B_1(t) = t, \dots, T$.
- $B_2(t)$: possible periods for submitting jobs due at t, such that if $t \le L$, then $B_2(t) = 1, \dots, t$; if $t = L + 1, \dots, T$, then $B_2(t) = t - L, \dots, t$.

Additional Parameter:

- $\mathcal{N}_c$: Set of user groups who submit computational demand.
- $\widetilde{d}_j^t$: random job (j, t) submitted by user j at period t.
- p_j^{tk}: unit penalty of unfinished job (j, t) at period k, $\forall k \in B_1(t)$.

New Variables:

- u_{ji}^{tk}: percentage of ℓ^t for processing job (j, t) on server i in period k, $\forall i \in \mathcal{N}_m$, $j \in \mathcal{N}_c$, $t = 1, \dots, T$, and $k \in B_1(t)$.
- $b_j^{tk\omega}$: unfinished job (j, t) at period k in scenario ω, $\forall k \in B_1(t)$ and $\omega \in \Omega$

$$\min: \quad \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) + \sum_{\omega \in \Omega} \rho^\omega \left(\sum_{t=1}^{T} \sum_{j \in \mathcal{N}_c} \sum_{k \in B_1(t)} p_j^{tk} b_j^{tk\omega} \right)$$

s.t. (1c)–(1g) $\Rightarrow$ Constraints from Model (1)

$$\sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \tilde{d}_j^t \quad \forall j \in \mathcal{N}_c,\ 1 \leq t \leq T \tag{2a}$$

$$x_i^t \geq \sum_{k \in B_2(t)} \sum_{j \in \mathcal{N}_c} u_{ji}^{kt} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{2b}$$

$$b_j^{tk\omega} = \max\left\{0,\ d_j^{t\omega} - \sum_{l=t}^{k} \sum_{i \in \mathcal{N}_m} e_i \ell^l u_{ji}^{tl}\right\} \quad \forall j \in \mathcal{N}_c,\ 1 \leq t \leq T,\ k \in B_1(t),\ \omega \in \Omega \tag{2c}$$

$$0 \leq u_{ji}^{tk} \leq 1,\ b_j^{tk\omega} \geq 0. \tag{2d}$$

ρ^ω: the probability of scenario $\omega \in \Omega \Rightarrow$ penalize unfinished job requests in the objective, and minimize the expected penalty.

$$\text{min:} \quad \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) + \sum_{\omega \in \Omega} \rho^\omega \left(\sum_{t=1}^{T} \sum_{j \in \mathcal{N}_c} \sum_{k \in B_1(t)} p_j^{tk} b_j^{tk\omega} \right)$$

s.t. (1c)–(1g) $\Rightarrow$ Constraints from Model (1)

$$\sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \max_{\omega \in \Omega} d_j^{t\omega} \quad \forall 1 \leq t \leq T \tag{2a}$$

$$x_i^t \geq \sum_{k \in B_2(t)} \sum_{j \in \mathcal{N}_c} u_{ji}^{kt} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T \tag{2b}$$

$$b_j^{tk\omega} = \max\left\{0,\ d_j^{t\omega} - \sum_{l=t}^{k} \sum_{i \in \mathcal{N}_m} e_i \ell^l u_{ji}^{tl}\right\} \quad \forall j \in \mathcal{N}_c,\ 1 \leq t \leq T,\ k \in B_1(t),\ \omega \in \Omega \tag{2c}$$

$$0 \leq u_{ji}^{tk} \leq 1,\ b_j^{tk\omega} \geq 0. \tag{2d}$$

$d_j^{t\omega}$: the realization of $\widetilde{d}_j^t$ in scenario $\omega \in \Omega \Rightarrow$ replace stochastic constraints (2a) by equivalent deterministic constraints.

Model 3: Backlogging with a Joint Chance Constraint

Relax Model (2) by allowing job incompleteness after L backlogging periods, however, **bounded by a certain risk tolerance**.

That is, replace Constraint (2a)

$$\sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \widetilde{d}_j^t \quad \forall j \in \mathcal{N}_c,\, 1 \leq t \leq T$$

with

$$\mathbb{P}\left(\sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \widetilde{d}_j^t, \forall j \in \mathcal{N}_c,\, 1 \leq t \leq T \right) \geq \alpha$$

Model 3: Backlogging with a Joint Chance Constraint

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) + \sum_{\omega \in \Omega} \rho^\omega \left(\sum_{t=1}^{T} \sum_{j \in \mathcal{N}_c} \sum_{k \in B_1(t)} p_j^{tk} b_j^{tk\omega} \right) \\
\text{s.t.} \quad & \text{(1c)–(1g)} \;\Rightarrow\; \text{Constraints from Model (1)} \\
& \text{(2b)–(2d)} \;\Rightarrow\; \text{Constraints from Model (2)} \\
& \mathbb{P} \left(\sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \widetilde{d}_j^t, \forall j \in \mathcal{N}_c,\, 1 \leq t \leq T \right) \geq \alpha
\end{aligned}
$$

Model 3: Backlogging with a Joint Chance Constraint

$$\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) + \sum_{\omega \in \Omega} \rho^\omega \left(\sum_{t=1}^{T} \sum_{j \in \mathcal{N}_c} \sum_{k \in B_1(t)} p_j^{tk} b_j^{tk\omega} \right) \\
\text{s.t.} \quad & \text{(1c)–(1g)} \ \Rightarrow \ \text{Constraints from Model (1)} \\
& \text{(2b)–(2d)} \ \Rightarrow \ \text{Constraints from Model (2)} \\
& \sum_{\omega \in \Omega} \rho^\omega \zeta^\omega \leq 1 - \alpha
\end{aligned}$$

where, for each $\omega \in \Omega$, binary variables $\zeta^\omega = 1$ if $\forall j \in \mathcal{N}_c, \ 1 \leq t \leq T$, there exists at least one

$$\sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} < d_j^{t\omega},$$

and 0 otherwise.

Model 3: Backlogging with a Joint Chance Constraint

$$
\begin{aligned}
\text{min:} \quad & \sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t) + \sum_{\omega \in \Omega} \rho^\omega \left(\sum_{t=1}^{T} \sum_{j \in \mathcal{N}_c} \sum_{k \in B_1(t)} p_j^{tk} b_j^{tk\omega} \right) \\
\text{s.t.} \quad & \text{(1c)–(1g)} \ \Rightarrow \ \text{Constraints from Model (1)} \\
& \text{(2b)–(2d)} \ \Rightarrow \ \text{Constraints from Model (2)} \\
& \sum_{\omega \in \Omega} \rho^\omega \zeta^\omega \leq 1 - \alpha \\
& \sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} + M_j^t \zeta^\omega \geq d_j^{t\omega} \\
& \qquad\qquad \forall \omega \in \Omega,\ j \in \mathcal{N}_c,\ 1 \leq t \leq T \\
& \zeta^\omega \in \{0, 1\} \quad \forall \omega \in \Omega.
\end{aligned}
$$

where M_j^t is set as the maximal standard time for processing job (j, t), e.g., $M_j^t = \max_{\omega \in \Omega} d_j^{t\omega},\ \forall j \in \mathcal{N}_c,\ 1 \leq t \leq T$.

Model 4: Backlogging with Multiple Chance Constraints

Instead of a joint chance constraint

$$\mathbb{P}\left(\sum_{k\in B_1(t)}\sum_{i\in\mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \tilde{d}_j^t,\ \forall j\in\mathcal{N}_c,\ 1\leq t\leq T\right) \geq \alpha,$$

we formulate a series of job-based constraints, each of which is associated with a risk tolerance α_j^t, for job (j,t), $\forall j\in\mathcal{N}_c$ and $1\leq t\leq T$.

$$\mathbb{P}\left(\sum_{k\in B_1(t)}\sum_{i\in\mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \tilde{d}_j^t\right) \geq \alpha_j^t \quad \forall j\in\mathcal{N}_c,\ 1\leq t\leq T.$$

Solution Algorithms

Computational challenges from:

Large-Scale Time Intervals $(1, \ldots, T)$

Large Number of Users and Servers ($|\mathcal{N}_c|$ and $|\mathcal{N}_m|$)

Large Number of Scenarios ($|\Omega|$) for Describing the Uncertainty ($\tilde{d}$)

Binary Server Operational Decisions (y and z)

$$\text{min:} \quad \boxed{\sum_{t=1}^{T} \sum_{i \in \mathcal{N}_m} (g_i y_i^t + v_i x_i^t + f_i z_i^t)} + \boxed{\sum_{\omega \in \Omega} \text{Prob}^{\omega} \left(\sum_{t=1}^{T} \sum_{j \in \mathcal{N}_c} \sum_{k \in B_1(t)} p_j^{tk} b_j^{tk\omega} \right)}$$

$$\text{s.t.} \quad \ell^t x_i^t + s_i y_i^t \leq \ell^t z_i^t \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T$$

$$y_i^1 \geq z_i^1 \quad \forall i \in \mathcal{N}_m$$

$$y_i^t \geq z_i^t - z_i^{t-1} \quad \forall i \in \mathcal{N}_m,\ 2 \leq t \leq T$$

$$0 \leq x_i^t \leq 1 \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T$$

$$y_i^t, z_i^t \in \{0,1\} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T$$

Major Operations Decisions (Relaxed Master Problem)

$$\sum_{k \in B_1(t)} \sum_{i \in \mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \tilde{d}_j^t \quad \forall j \in \mathcal{N}_c,\ 1 \leq t \leq T$$

$$\hat{x}_i^t \geq \sum_{k \in B_2(t)} \sum_{j \in \mathcal{N}_c} u_{ji}^{kt} \quad \forall i \in \mathcal{N}_m,\ 1 \leq t \leq T$$

$$b_j^{tk\omega} = \max\left\{0,\ d_j^{t\omega} - \sum_{l=t}^{k} \sum_{i \in \mathcal{N}_m} e_i \ell^l u_{ji}^{tl}\right\}$$

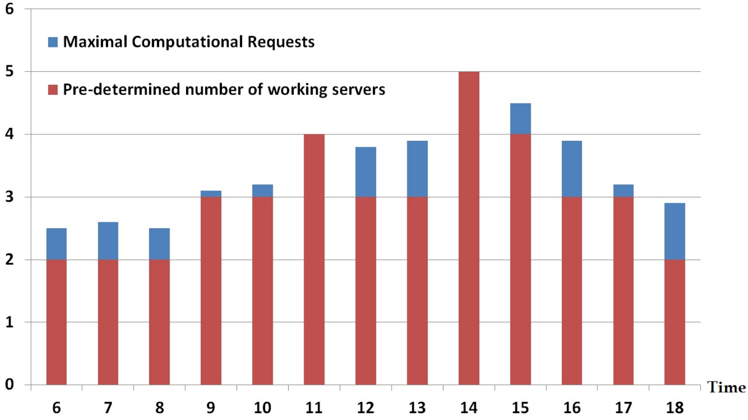
$$\forall j \in \mathcal{N}_c,\ 1 \leq t \leq T,\ k \in B_1(t),\ \omega \in \Omega$$

$0 \leq u_{ji}^{tk} \leq 1,\ b_j^{tk\omega} \geq 0.$ **Continuous!**

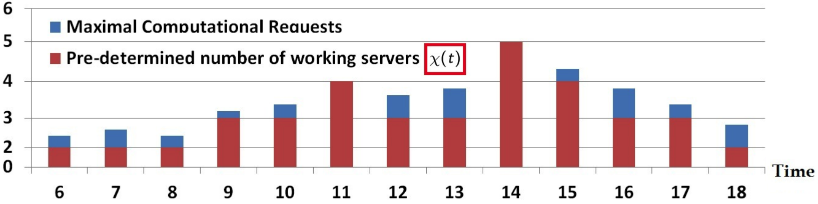
Scenario-based Decisions (Second Stage Subproblem)

A Heuristic Approach

Idea: fix schedules of a subset of servers. Then optimize schedules for the rest of servers using math modeling.



A Heuristic Approach



We pre-determine a subset of servers' schedule by setting

$$x_i^1 = 1 - s_i/\ell^t \quad \forall i = 1, \ldots, \chi(1),$$

$$x_i^t = 1 \quad \forall 2 \le t \le T,\ i = 1, \ldots, \chi^-(t),$$

$$x_i^t = 1 - s_i/\ell^t \quad \forall 2 \le t \le T,\ i = \chi^-(t)+1, \ldots, \chi^-(t)+\chi^+(t) \quad \text{if } \chi^+(t) > 0,$$

where for $t = 1, \ldots, T$,

$$\chi(t) = \left\lfloor \sum_{j \in \mathcal{N}_c} \max_{\omega \in \Omega} d_j^{t\omega}/\ell^t \right\rfloor,$$

$\chi^-(t) = \min\{\chi(t-1), \chi(t)\}$, and $\chi^+(t) = \max\{\chi(t) - \chi(t-1), 0\}$.

- $|\mathcal{N}_c| = 2$ (two types of users) and $|\mathcal{N}_m| =$5, 10, and 20.
- Set $T = 24$ hours.
- Average energy consumption of Off, Idle, Processing, and Booting for a 3.0 Ghz server to be, respectively, 0W, 150W, 250W, and 250W (i.e., $v_i = 100W$, $f_i = 150W$).

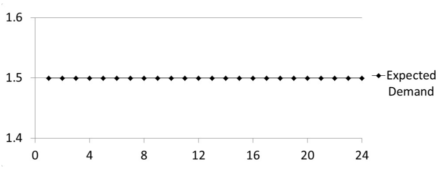
$\mathcal{N}_m$	$\mathcal{I} = 10\%$		$\mathcal{I} = 30\%$		$\mathcal{I} = 50\%$	
	$E[\sum_{t=1}^{T} \tilde{d}^t]$ (hours)	Bk (kWh)	$E[\sum_{t=1}^{T} \tilde{d}^t]$ (hours)	Bk (kWh)	$E[\sum_{t=1}^{T} \tilde{d}^t]$ (hours)	Bk (kWh)
5	12	19.2	36	21.6	60	24
10	24	38.4	72	43.2	120	48
20	48	76.8	144	86.4	240	96

- "$\mathcal{I}$": computational intensity
- $E[\sum_{t=1}^{T} \tilde{d}^t] = \mathcal{I} * |\mathcal{N}_m| * 24$ (hours)
- **Bk**: gives benchmark energy consumption (objective) by having servers first "on" then "idle."

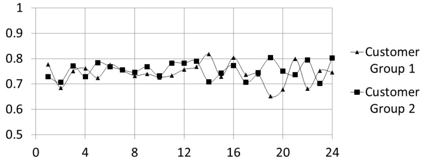
$\mathcal{N}_m$	$\mathcal{I} = 10\%$ $E[\sum_{t=1}^{T} \widetilde{d}^t]$ (hours)	Bk (kWh)	$\mathcal{I} = 30\%$ $E[\sum_{t=1}^{T} \widetilde{d}^t]$ (hours)	Bk (kWh)	$\mathcal{I} = 50\%$ $E[\sum_{t=1}^{T} \widetilde{d}^t]$ (hours)	Bk (kWh)
5	12	19.2	36	21.6	60	24
10	24	38.4	72	43.2	120	48
20	48	76.8	144	86.4	240	96

- "$\mathcal{I}$": computational intensity
- $E[\sum_{t=1}^{T} \widetilde{d}^t] = \mathcal{I} * |\mathcal{N}_m| * 24$ (hours)
- **Bk**: gives benchmark energy consumption (objective) by having servers first "on" then "idle."

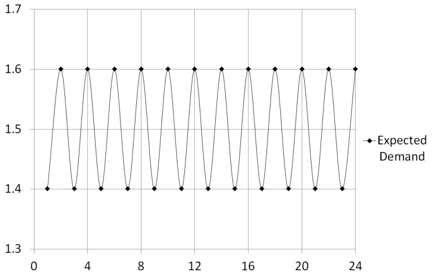
Demand Patterns



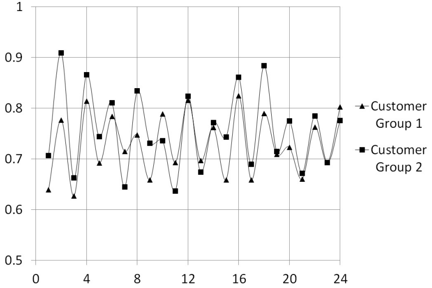
(a) Type 0 Demand Curve



(b) Type 0 Job Demand Sample

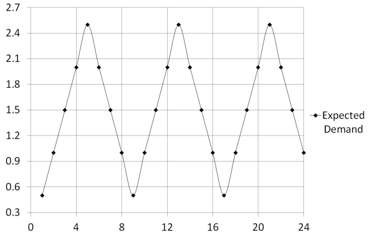


(c) Type 1 Demand Curve

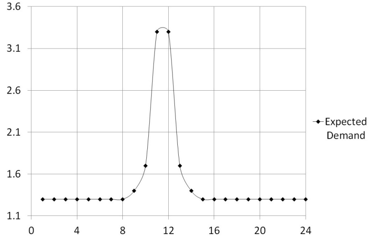


(d) Type 1 Job Demand Sample

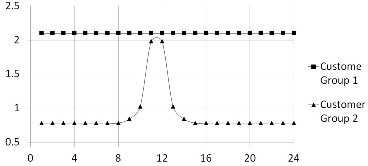
Demand Patterns



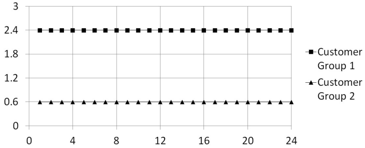
(e) Type 2 Demand Curve



(f) Type 3 Demand Curve



(g) Type 4 Demand Curve



(h) Type 5 Demand Curve

Types $0 \sim 3$: Homogeneous. Types 4 & 5: Heterogeneous.

Computational Design

- CPLEX 12.4 via ILOG Concert Technology with C++
- HP Workstation Z210 with CPU 3.20 GHz and 8GB memory
- CPU time limits =1800 seconds for each instance
- Test five instances for each parameter combination

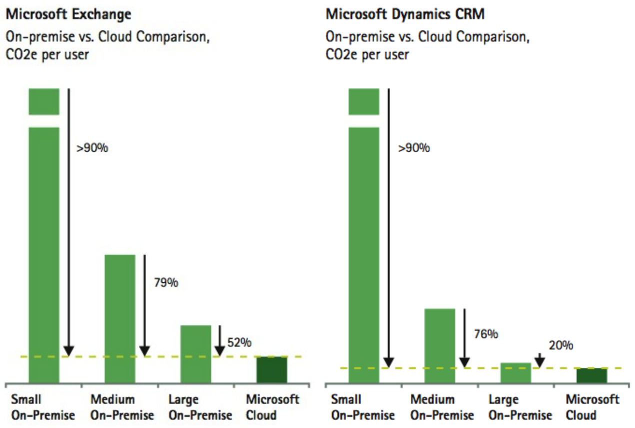
Results of Model 1

		$\mathcal{I} = 10\%$			$\mathcal{I} = 30\%$			$\mathcal{I} = 50\%$		
$\mathcal{N}_m$	Type	Bk	Oper	Save	Bk	Oper	Save	Bk	Oper	Save
5	T0	19.2	4.9	75%	21.6	11.0	49%	24	17.1	29%
	T1	19.2	4.9	75%	21.6	11.0	49%	24	17.1	29%
	T2	19.2	4.9	74%	21.6	12.0	44%	24	17.3	28%
	T3	19.2	5.2	73%	21.6	11.7	46%	24	-	-
10	T0	38.4	9.7	75%	43.2	22.0	49%	48	34.2	29%
	T1	38.4	8.2	79%	43.2	20.4	53%	48	32.7	32%
	T2	38.4	8.5	78%	43.2	22.2	49%	48	34.4	28%
	T3	38.4	7.3	81%	43.2	20.7	52%	48	-	-
20	T0	76.8	15.9	79%	86.4	40.3	53%	96	64.9	32%
	T1	76.8	14.3	81%	86.4	38.8	55%	96	65.1	32%
	T2	76.8	14.8	81%	86.4	41.3	52%	96	67.4	30%
	T3	76.8	13.9	82%	86.4	40.9	53%	96	-	-

Results of Model 1

		$\mathcal{I}=10\%$			$\mathcal{I}=30\%$			$\mathcal{I}=50\%$		
$\mathcal{N}_m$	Type	Bk	Oper	Save	Bk	Oper	Save	Bk	Oper	Save
5	T0	19.2	4.9	75%	21.6	11.0	49%	24	17.1	29%
	T1	19.2	4.9	75%	21.6	11.0	49%	24	17.1	29%
	T2	19.2	4.9	74%	21.6	12.0	44%	24	17.3	28%
	T3	19.2	5.2	73%	21.6	11.7	46%	24	-	-
10	T0	38.4	9.7	75%	43.2	22.0	49%	48	34.2	29%
	T1	38.4	8.2	79%	43.2	20.4	53%	48	32.7	32%
	T2	38.4	8.5	78%	43.2	22.2	49%	48	34.4	28%
	T3	38.4	7.3	81%	43.2	20.7	52%	48	-	-
20	T0	76.8	15.9	79%	86.4	40.3	53%	96	64.9	32%
	T1	76.8	14.3	81%	86.4	38.8	55%	96	65.1	32%
	T2	76.8	14.8	81%	86.4	41.3	52%	96	67.4	30%
	T3	76.8	13.9	82%	86.4	40.9	53%	96	-	-
	Avg.			80%			50%			30%

Recall...


Microsoft Exchange
On-premise vs. Cloud Comparison, CO2e per user
>90%
79%
52%
Small On-Premise
Medium On-Premise
Large On-Premise
Microsoft Cloud
Microsoft Dynamics CRM
On-premise vs. Cloud Comparison, CO2e per user
>90%
76%
20%
Small On-Premise
Medium On-Premise
Large On-Premise
Microsoft Cloud

Results of Model 1

		$\mathcal{I} = 10\%$			$\mathcal{I} = 30\%$			$\mathcal{I} = 50\%$		
$\mathcal{N}_m$	Type	Bk	Oper	Save	Bk	Oper	Save	Bk	Oper	Save
5	T0	19.2	4.9	75%	21.6	11.0	49%	24	17.1	29%
	T1	19.2	4.9	75%	21.6	11.0	49%	24	17.1	29%
	T2	19.2	4.9	74%	21.6	12.0	44%	24	17.3	28%
	T3	19.2	5.2	73%	21.6	11.7	46%	24	-	-
10	T0	38.4	9.7	75%	43.2	22.0	49%	48	34.2	29%
	T1	38.4	8.2	79%	43.2	20.4	53%	48	32.7	32%
	T2	38.4	8.5	78%	43.2	22.2	49%	48	34.4	28%
	T3	38.4	7.3	81%	43.2	20.7	52%	48	-	-
20	T0	76.8	15.9	79%	86.4	40.3	53%	96	64.9	32%
	T1	76.8	14.3	81%	86.4	38.8	55%	96	65.1	32%
	T2	76.8	14.8	81%	86.4	41.3	52%	96	67.4	30%
	T3	76.8	13.9	82%	86.4	40.9	53%	96	-	-

Results of Model 1

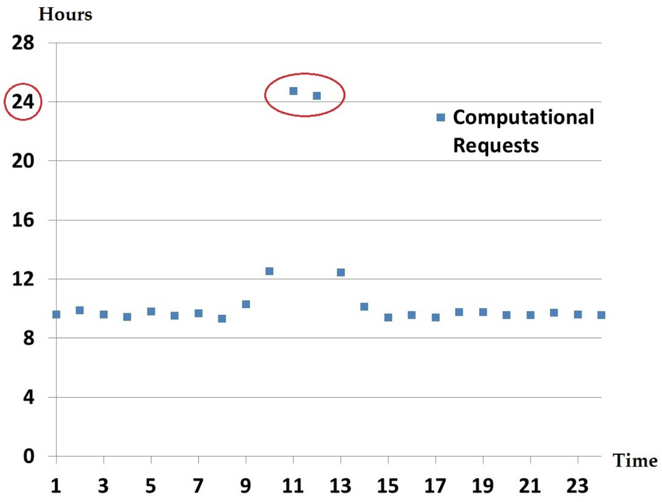


Figure: $\mathcal{N}_m = 20$, $\mathcal{I} = 50\%$, Type 3 Demand

A Revisit of Models

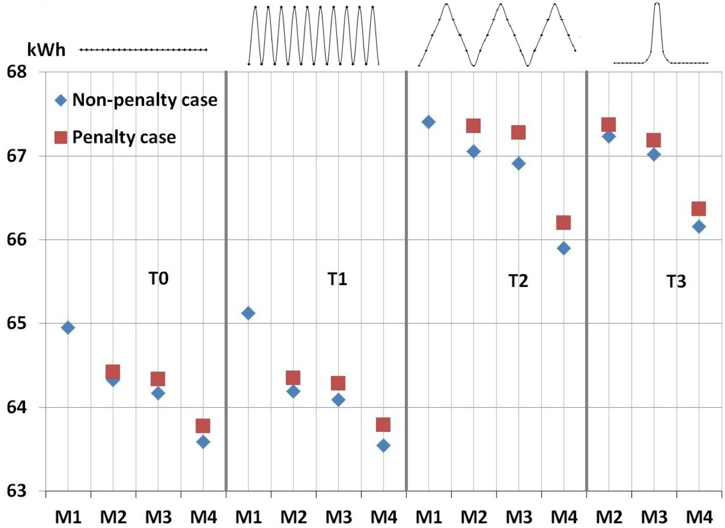
Model (1): $$\sum_{i\in\mathcal{N}_m} e_i \ell^t x_i^t \geq \max_{\omega\in\Omega} d^{t\omega} \quad \forall 1 \leq t \leq T.$$

Model (2): $$\sum_{k\in B_1(t)} \sum_{i\in\mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \max_{\omega\in\Omega} d_j^{t\omega} \quad \forall 1 \leq t \leq T.$$

Model (3): $$\mathbb{P}\left(\sum_{k\in B_1(t)} \sum_{i\in\mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \widetilde{d}_j^t, \forall j \in \mathcal{N}_c,\ 1 \leq t \leq T \right) \geq \alpha.$$

Model 4: $$\mathbb{P}\left(\sum_{k\in B_1(t)} \sum_{i\in\mathcal{N}_m} e_i \ell^k u_{ji}^{tk} \geq \widetilde{d}_j^t \right) \geq \alpha_j^t \quad \forall j \in \mathcal{N}_c,\ 1 \leq t \leq T.$$

Energy Use in Models 1-4 ($\mathcal{N}_m = 20$, $\mathcal{I} = 50\%$)



Unit penalty $p_j^{tk} = 100$ for penalty case, $\forall\, j \in \mathcal{N}_c$, $1 \le t \le T$, and $k \in B_1(t)$.

Solution Approach Comparison

		Model 1					Model 2		
Type	No.	C-Total	B-Time	B-Total	H-Time	H-Total	C-Total	H-Time	H-Total
	1	64.24	17.10	64.24	2.37	64.42	64.56	43.51	64.65
	2	64.21	10.67	64.21	2.36	64.39	64.52	11.84	64.53
T1	3	64.24	949.92	64.24	2.57	64.42	64.57	123.40	64.67
	4	64.41	1827.52	64.41	2.40	64.43	64.62	22.17	64.69
	5	64.21	28.88	64.21	2.59	64.38	64.50	14.35	64.55

Table: $\mathcal{N}_m = 20$, $\mathcal{I} = 50\%$, and Five Instances

"C-", solving Model (2) by directly solving its MIP.

"B-", employing Benders decomposition.

"H-", using the approximation approach.

Solution Approach Comparison

		Model 1					Model 2		
Type	No.	C-Total	B-Time	B-Total	H-Time	H-Total	C-Total	H-Time	H-Total
	1	64.24	17.10	64.24	2.37	64.42	64.56	43.51	64.65
	2	64.21	10.67	64.21	2.36	64.39	64.52	11.84	64.53
T1	3	64.24	949.92	64.24	2.57	64.42	64.57	123.40	64.67
	4	64.41	1827.52	64.41	2.40	64.43	64.62	22.17	64.69
	5	64.21	28.88	64.21	2.59	64.38	64.50	14.35	64.55

Table: $\mathcal{N}_m = 20$, $\mathcal{I} = 50\%$, and Five Instances

The performance of the Benders approach varies among instances and is unstable.

Solution Approach Comparison

		Model 1					Model 2		
Type	No.	C-Total	B-Time	B-Total	H-Time	H-Total	C-Total	H-Time	H-Total
T1	1	64.24	17.10	64.24	2.37	64.42	64.56	43.51	64.65
	2	64.21	10.67	64.21	2.36	64.39	64.52	11.84	64.53
	3	64.24	949.92	64.24	2.57	64.42	64.57	123.40	64.67
	4	64.41	1827.52	64.41	2.40	64.43	64.62	22.17	64.69
	5	64.21	28.88	64.21	2.59	64.38	64.50	14.35	64.55

Table: $\mathcal{N}_m = 20$, $\mathcal{I} = 50\%$, and Five Instances

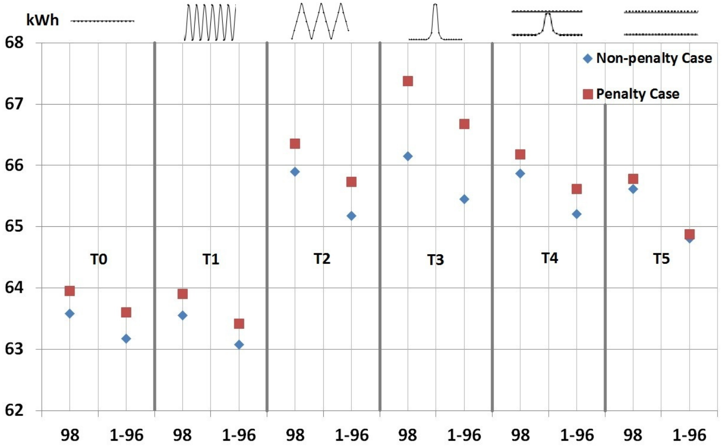
For Model 2, the differences between H-Total and C-Total are within 0.3% gaps for all instances.

Solution Approach Comparison

		Model 1					Model 2		
Type	No.	C-Total	B-Time	B-Total	H-Time	H-Total	C-Total	H-Time	H-Total
T1	1	64.24	17.10	64.24	2.37	64.42	64.56	43.51	64.65
	2	64.21	10.67	64.21	2.36	64.39	64.52	11.84	64.53
	3	64.24	949.92	64.24	2.57	64.42	64.57	123.40	64.67
	4	64.41	1827.52	64.41	2.40	64.43	64.62	22.17	64.69
	5	64.21	28.88	64.21	2.59	64.38	64.50	14.35	64.55

Table: $\mathcal{N}_m = 20$, $\mathcal{I} = 50\%$, and Five Instances

Effects of Use Prioritization (Model 4)



- **98**: $\alpha_j^t = 98\%, \forall j \in \mathcal{N}_c$; **1-96**: $\alpha_0^t = 100\%$, $\alpha_1^t = 96\%, \forall 1 \leq t \leq T$.

- Effectively managing energy footprints and QoS via stochastic optimization models.
- Yield respective 80%, 50%, and 30% of energy savings for 10%, 30%, and 50% demand intensity regardless of demand patterns.
- Backlogging and chance constraints provide additional flexibility in server scheduling and reduce energy use.
- The Benders decomposition and the heuristic approach are faster and yield good results.
- User prioritization via multiple chance constraints can effectively reduce consumed energy.

Thank You!

Questions?