Optimal Energy and IT Service Emergency Schedule for Internet Data Center

Zhongqi Zhao[1,2], *Student Member, IEEE,* Lei Fan[1,2], *Senior Member, IEEE,* and Zhu Han[1] *Fellow, IEEE*
Dept. of Electrical and Computer Engineering[1] *and Dept. of Engineering Technology*[2]
University of Houston, TX, USA

***Abstract*—In the era of Artificial intelligence (AI), internet data centers (IDCs) play a crucial role in supporting the global information infrastructure. However, large language model (LLM) training consumes a lot of energy, which poses unique challenges to IDC, especially in power emergencies. In order to prevent power outages from causing huge economic losses to IDC, a good energy management system is essential. In this study, we propose a new mixed-integer linear programming (MILP) model of an IDC during an electricity emergency to maximize the profit while minimize operation cost. On power side, we considered renewable energy, traditional thermal power generators, uninterruptible power supply (UPS), heating, ventilation, and air conditioning (HVAC) system. On IT job scheduling side, we allow jobs to be done, held, discarded and transferred to other IDC. Finally, we analyze and visualize the computational results and verify the correctness of the proposed model.**

***Index Terms*—Data Center, Mixed Integer Linear Programming, Optimization, Energy Management, Emergency Response**

I. Introduction

With the rise of AI-generated content (AIGX) and large language models (LLMs), internet data centers (IDCs) have become a popular topic due to their ability to meet the enormous demand for computational resources. However, costs and energy management are key factors in the operation of IDCs. It has been reported that IDC outages are not uncommon globally. According to [1], between January 2016 and June 2018, the Uptime Institute reported five "extremely severe" outages involving power or cooling systems, resulting in considerable income loss and brand damage, potentially posing an existential threat to IDC operators or their clients. To prevent similar incidents, we need to study the emergency response protocols of IDCs to avoid or reduce financial losses.

On the power side, IDCs are rapidly emerging as significant electricity consumers. It is predicted that by 2030, IDCs will account for 9.1% of annual U.S. electricity consumption, up from an estimated 4% today [2]. A recent study found that individual IDCs, such as Digital Realty's Lakeside Technology Center in Chicago, Illinois, frequently demand around 85 megawatts of power [3]. According to the statistics in [1], there is a 15% chance that the loss from an outage event will exceed $1 million, underscoring the need for an effective emergency energy management system.

Moreover, IT job scheduling has also attracted attention in the network development of IDCs. Numerous papers have explored this area. The authors in [4] leveraged uninterruptible power supply (UPS) batteries to temporarily augment the utility supply during emergencies. Additionally, [5] considered renewable energy generators to reduce carbon footprints and increase stability. Furthermore, [6] proposed a post-emergency event restoration method for IDCs, treating them as critical loads. Recently, [7] proposed a mixed-integer linear programming (MILP) approach for IDC energy management, which considers most of the previously mentioned devices and energy sources. Although these papers provide in-depth research on daily energy management, emergency, and post-emergency event handling, they do not address the need for detailed scheduling of all major equipment during critical moments. In particular, research on the scheduling of new energy sources, power sources, and various power-consuming equipment during critical moments is lacking. This will be the first challenge we encounter.

IDCs require an operating system that can fairly allocate physical resources to user-submitted tasks. Since every computer has a limited amount of resources, schedulers are of paramount importance for maintaining system performance, stability, and reliability, which will eventually lead to better monetary outcomes. This field has also been well explored. The authors in [8] used enhanced particle swarm optimization to improve the job scheduling approach. Meanwhile, an incentive-based job scheduling approach is proposed in [9], addressing the flaws of conventional IDC job scheduling approaches that do not compensate resource users for jobs that miss deadlines. Moreover, [10] proposed a virtual network functions scheduling approach for the 5G network, solved by quantum computers. Although there has been significant research on IDC job scheduling, there is limited research on job scheduling under emergency conditions. Solving this problem can help IDCs reduce losses and gain considerable benefits when energy is limited.

Therefore, inspired by [7], [10], this article focuses on addressing the aforementioned challenges and proposes a general model. For the job scheduling problem, we designed jobs with different states and considered the computing performance of various computing nodes to simulate real job scheduling in IDCs as accurately as possible. For the energy management problem, we incorporated real-time energy demand, the cooling demand of each job, and the new challenge of generator scheduling for the entire IDC during crises, based on [11]. Finally, we proposed the entire problem as a mixed-integer linear programming problem.

The contributions of this paper are summarized as follows:

- We propose a new MILP-based model for the IDC scheduling problem during emergencies. We considered factors such as energy consumed by the job scheduling itself, renewable energy, batteries, indoor temperature, HVAC, and traditional heat generators.
- We verify the correctness of the proposed model formulation by setting up a test case. The visualization of the job scheduling and all other factors proves the soundness of our problem formulation.

The rest of this paper is organized as follows. Section II introduces the problem definition and formulations. Section III validates our algorithm using experiments. Finally, Section IV concludes the paper.

II. Problem Definition and Formulations

In this section, we propose a new MILP-based mathematical programming formulation for IDC energy system operations, considering IT services based on the previous works mentioned in [7], [10], [11].

A. Data Center Energy System Modeling Overview

The IDC electricity emergency response management problem aims to maximize a IDC's net income given time periods $t \in \mathcal{T}$, where $\mathcal{T} = \{1, 2, \ldots, T\}$. The length of the time interval is denoted by Δt, and $T\Delta t$ represents the total time of interest. On the one hand, the IDC must complete as many jobs as possible while bearing the financial cost of discarded jobs. On the other hand, it is critical for IDCs to reduce the operational cost of their energy consumption. Furthermore, $t = 0$ is a special case that is not in $\mathcal{T}$ and it is simply a notation for the device's or ambient initial status and does not factor into the cost objective calculation. For the symbol notation, all lowercase letters such as $\mathbf{c}$ and $\mathbf{y}$ represent vectors of variables, and most uppercase letters represent known parameters, with a few exceptions. The details of parameters and variables will be clarified when explaining the constraints. Among those variables, $\{\mathbf{x}, \mathbf{u}, \mathbf{v}\} \in \{\mathbf{0}, \mathbf{1}\}$, which are binary variables.

$$\textbf{(P)} \max\ c^{\text{profit}} - c^{\text{loss}} - c^{\text{transfer}} - \sum_{t=1}^{T} c_t^{\text{Op}}, \tag{1}$$

$$\text{s.t. } c^{\text{profit}} = \sum\nolimits_{j=1}^{J} C_j^{\text{profit}} \left(x_j^{\text{Done}} + x_j^{\text{TF}}\right), \tag{2}$$

$$c^{\text{loss}} = \sum\nolimits_{j=1}^{J} C_j^{\text{loss}} x_j^{\text{Abort}}, \tag{3}$$

$$c^{\text{transfer}} = \sum\nolimits_{j=1}^{J} C_j^{\text{transfer}} x_j^{\text{TF}}. \tag{4}$$

The objective function (1) aims to maximize the net income of the IDC within the time period $\mathcal{T}$. In constraint (2), c^{profit} represents the total profit generated from jobs that are successfully completed or transferred. In constraint (3), c^{loss} accounts for any losses incurred, such as penalties from aborted or failed jobs. In constraint (4), c^{transfer} includes costs associated with transferring jobs to other IDCs. Returning to Equation (1), $\sum_{t=1}^{T} c_t^{\text{Op}}$ sums up the operational costs over all time periods, including costs related to power generation and other operational expenses.

B. Job States and Scheduling

In this section, constraints related to job status and scheduling are introduced in sequence. First, we introduce the constraints on job states, followed by the general scheduling constraints, and finally the specific constraints for each type of job.

1) Job States:

$$x_j^{\text{Done}} + x_j^{\text{Abort}} + x_j^{\text{TF}} + x_j^{\text{Hold}} = 1,\ \forall j \in \mathcal{J}, \tag{5}$$

$$\begin{cases} x_j^{\text{TF}} = 0, & \text{if } j \in \mathcal{J}^{\text{w}}, \\ x_j^{\text{Hold}} = 0, & \text{if } j \in \mathcal{J}^{\text{w}} \cup \mathcal{J}^{\text{TS}}, \\ x_j^{\text{Abort}} = 0, & \text{if } j \in \mathcal{J}^{\text{NTS}}. \end{cases} \tag{6}$$

Equation (5) ensures that for each job j in the set of all jobs $\mathcal{J}$, exactly one of the four states (Done, Abort, Transferred (TF), Hold) is true. The set of constraints (6) defines specific conditions under which certain states cannot be true for different subsets of jobs. If a job is ongoing ($\mathcal{J}^{\text{w}}$), it cannot be transferred or held. If a job is time-sensitive ($\mathcal{J}^{\text{TS}}$), it cannot be held. If a job is a normal job ($\mathcal{J}^{\text{NTS}}$), it cannot be aborted. This set of conditions is flexible and can be redefined by users.

2) General Scheduling Constraints:

$$\sum\nolimits_{n=1}^{N} x_{j,n}^{\text{job}} = x_j^{\text{Done}},\ \forall j \in \mathcal{J}, \tag{7}$$

$$u_{j,n,t}^{\text{job}} \le x_{j,n}^{\text{job}},\ \forall j \in \mathcal{J}, t \in \mathcal{T}, n \in \mathcal{N}, \tag{8}$$

$$u_{j,n,t-1}^{\text{job}} - u_{j,n,t}^{\text{job}} - v_{j,n,t}^{\text{sd}} + v_{j,n,t}^{\text{su}} = 0,\ j \in \mathcal{J}, \forall t \in \mathcal{T}, n \in \mathcal{N}, \tag{9}$$

$$v_{j,n,t}^{\text{sd}} + v_{j,n,t}^{\text{su}} \le 1,\ \forall j \in \mathcal{J}, t \in \mathcal{T}, n \in \mathcal{N}. \tag{10}$$

Equation (7) ensures that the sum of the job states over all N nodes is equal to the binary variable state x_j^{Done} for each working job $j \in \mathcal{J}$, meaning there is at most one node processing job j. The variable $x_{j,n}^{\text{job}}$ indicates whether job j is assigned to node n. Equation (8) guarantees that a job will not be assigned to any node at any time before $x_{j,n}^{\text{job}}$ is confirmed. Constraint (9) shows the logical relationship between the state of processing (u^{job}), start (v^{su}), and shutdown (v^{sd}). Here, $(\cdot)_{j,n,t} = 1$ implies that job j is in the corresponding state on node n at time t. Constraint (10) ensures that a node n cannot both start and finish a job j at the same time.

3) Current Working Jobs:

$$x_{j,N_j^w}^{\text{job}} = x_j^{\text{Done}},\ \forall j \in \mathcal{J}^{\text{w}}, \tag{11}$$

$$u_{j,N_j^w,t}^{\text{job}} = x_j^{\text{Done}},\ \forall j \in \mathcal{J}^{\text{w}}, t \in [1, T_j^w]. \tag{12}$$

Equation (11) states that for each working job $j \in \mathcal{J}^{\text{w}}$, on its specified node $N_j^w \in \mathcal{N}$, the job state is equal to the Done state of the job, linking the specific working instance directly to job completion. Moreover, Equation (12) ensures that for each working job j, and for all time steps t within the working time period $\left[1, T_j^w\right]$, the processing status (u^{job}) of job j on node N_j^w aligns with the Done state of the job. This

enforces that the job's processing status is consistent with its completion state throughout its entire working period.

4) Non-time-sensitive Jobs:

$$u^{\text{job}}_{j,n} = \sum\nolimits_{t=1}^{T} v^{\text{sd}}_{j,n,t},\ \forall j \in \mathcal{J}^{\text{NTS}}, n \in \mathcal{N}, \tag{13}$$

$$\sum\nolimits_{t=1}^{T} u^{\text{job}}_{j,n,t} = T^{\text{job}}_{j,n} \cdot x^{\text{job}}_{j,n},\ \forall j \in \mathcal{J}^{\text{NTS}}, n \in \mathcal{N}, \tag{14}$$

$$\sum_{\tau=1}^{T^{\text{job}}_{j,n}} v^{\text{sd}}_{j,n,(t-\tau+1)} \le u^{\text{job}}_{j,n,t},\ \forall j \in \mathcal{J}^{\text{NTS}}, n \in \mathcal{N}, t \in \mathcal{T}, \tag{15}$$

$$\sum\nolimits_{n\in\mathcal{N}} \sum\nolimits_{t\in\mathcal{T}} v^{\text{sd}}_{j,n,t} = x^{\text{Done}}_{j},\ \forall j \in \mathcal{J}^{\text{NTS}}, \tag{16}$$

$$\sum\nolimits_{n\in\mathcal{N}} \sum\nolimits_{t\in\mathcal{T}} v^{\text{su}}_{j,n,t} = x^{\text{Done}}_{j},\ \forall j \in \mathcal{J}^{\text{NTS}}. \tag{17}$$

Non-time-sensitive jobs ($\mathcal{J}^{\text{NTS}}$) do not have stringent deadlines or immediate urgency for completion and can be regarded as normal jobs. Equation (13) states that the total utilization u^{job} of job j on node n is equal to the sum of job shutdown variables v^{sd} over the time period $\mathcal{T}$ for normal jobs ($j \in \mathcal{J}^{\text{NTS}}$). This links the cumulative shutdown events to the processing status. Constraint (14) ensures that the total processing state over the time period $\mathcal{T}$ for job j on every node equals the product of the job duration $T^{\text{job}}_{j,n}$, which is a parameter, and the job assignment variable $x^{\text{job}}_{j,n}$. This implies that the job's processing state is consistent with its duration and assignment. Constraint (15) guarantees that once the node starts processing job j, the node must process it for the required time $T^{\text{job}}_{j,n}$. Constraints (16) and (17) ensure that every job in $\mathcal{J}^{\text{NTS}}$ is started or shut down at most once.

5) Time-sensitive Jobs:

$$u^{\text{job}}_{j,n} = \sum\nolimits_{t=1}^{T^{\text{TS}}_j} v^{\text{sd}}_{j,n,t},\ \forall j \in \mathcal{J}^{\text{TS}}, n \in \mathcal{N}, \tag{18}$$

$$\sum\nolimits_{t=1}^{T^{\text{TS}}_j} u^{\text{job}}_{j,n,t} = T^{\text{job}}_{j,n} \cdot x^{\text{job}}_{j,n},\ \forall j \in \mathcal{J}^{\text{TS}}, n \in \mathcal{N}, \tag{19}$$

$$\sum\nolimits_{\tau=1}^{T^{\text{job}}_{j,n}} v^{\text{sd}}_{j,n,(t-\tau+1)} \le u^{\text{job}}_{j,n,t},\ \forall j \in \mathcal{J}^{\text{TS}}, n \in \mathcal{N}, t \in [1, T^{\text{TS}}_j], \tag{20}$$

$$\sum\nolimits_{n\in\mathcal{N}} \sum\nolimits_{t\in[1,T^{\text{TS}}_j]} v^{\text{sd}}_{j,n,t} = x^{\text{Done}}_{j},\ \forall j \in \mathcal{J}^{\text{TS}}, \tag{21}$$

$$\sum\nolimits_{n\in\mathcal{N}} \sum\nolimits_{t\in[1,T^{\text{TS}}_j]} v^{\text{su}}_{j,n,t} = x^{\text{Done}}_{j},\ \forall j \in \mathcal{J}^{\text{TS}}. \tag{22}$$

Time-sensitive jobs must be completed within a specific time frame to meet the deadline. Accordingly, the constraints for this part are similar to those for non-time-sensitive jobs, with the only difference being the due time. Therefore, (18)-(22) achieve something similar to (13)-(17). However, the time frame is replaced by a new due time T^{TS}_j, which is also a known set of parameters.

6) Transferred Jobs:

$$D_j \cdot x^{\text{TF}}_j = \sum\nolimits_{t=1}^{T_{\text{range}}} y^{\text{bw}}_{j,t},\ \forall j \in \mathcal{J}^{\text{NTS}}, t \in \mathcal{T}, \tag{23}$$

$$D_j \cdot x^{\text{TF}}_j = \sum\nolimits_{t=1}^{T_{\text{range}}} y^{\text{bw}}_{j,t},\ \forall j \in \mathcal{J}^{\text{TS}}, t \in [1, T^{\text{TS}}_j], \tag{24}$$

$$\sum\nolimits_{j\in\mathcal{J}} y^{\text{bw}}_{j,t} \le \overline{\text{DCBW}},\ \forall t \in \mathcal{T}, \tag{25}$$

$$y^{\text{bw}}_{j,t} \le \overline{\text{BW}}_j \cdot x^{\text{TF}}_j,\ \forall j \in \mathcal{J}, t \in \mathcal{T}. \tag{26}$$

Node Type	Power (KW)	Performance (petaFLOPS)
10 DGX1	350	17.3
10 DGX2	100	20
4 DGXA100	26	20
10 DGXA100s	15	13
1 DGXA200	14.3	72
1 DGXH100	10.2	32

TABLE I: Power and Performance Specifications of Nodes

Equations (23) and (24) ensure that for each non-time-sensitive and time-sensitive job, once it needs to be transferred out of the IDC, the job's pre-trained dataset D_j, a known parameter set, will be transferred out of the IDC within the corresponding time range. The amount of data transferred for each task j at time t is denoted by the continuous variable $y^{\text{bw}}_{j,t}$. Constraint (25) ensures that the total bandwidth utilization for all jobs j at any time t does not exceed the maximum IDC bandwidth ($\overline{\text{DCBW}}$). This prevents overloading the IDC's capacity. Constraint (26) specifies that at any time t, if job j needs to be transferred, the bandwidth utilization of the job must not exceed the maximum data transfer rate of the corresponding job ($\overline{\text{BW}}_j$). This ensures that when tasks are transferred, the bandwidth utilization of any job remains within its specified limit.

7) Energy Consumption:

$$e^{\text{O,Node}}_{n,t} = e^{\text{O,N,idle}}_{n,t} + e^{\text{O,N,w}}_{n,t},\ \forall n \in \mathcal{N}, t \in \mathcal{T}, \tag{27}$$

$$e^{\text{O,N,idle}}_{n,t} = E^{\text{O,N,idle}}_{n} u^{\text{power}}_{n,t},\ \forall n \in \mathcal{N}, t \in \mathcal{T}, \tag{28}$$

$$e^{\text{O,N,w}}_{n,t} = \sum_{j\in\mathcal{J}} E^{\text{O,N,w}}_{j,n} \cdot \beta^{\text{TDP}}_{j} \cdot u^{\text{job}}_{j,n,t},\ \forall n \in \mathcal{N}, t \in \mathcal{T}, \tag{29}$$

$$u^{\text{power}}_{n,0} = 1,\ \forall n \in \mathcal{N}^*, \tag{30}$$

$$u^{\text{job}}_{j,n,t} \le u^{\text{power}}_{n,t},\ \forall j \in \mathcal{J}, \forall t \in \mathcal{T}, n \in \mathcal{N}, \tag{31}$$

$$u^{\text{power}}_{n,t-1} - u^{\text{power}}_{n,t} - v^{\text{power,sd}}_{n,t} + v^{\text{power,su}}_{n,t} = 0,\ \forall t \in \mathcal{T}, n \in \mathcal{N}. \tag{32}$$

Constraint (27) calculates the total operational energy ($e^{\text{O,Node}}$) for node n at time t as the sum of the idle energy ($e^{\text{O,N,idle}}$) and the working energy ($e^{\text{O,N,w}}$) for that node. Equation (28) defines the idle energy ($e^{\text{O,N,idle}}$) for node n at time t as the product of the parameter idle energy rate ($E^{\text{O,N,idle}}$) and the power state ($u^{\text{power}}_{n,t}$) of the node. Equation (29) defines the working energy ($e^{\text{O,N,w}}$) for node n at time t as the sum of the products of the working power consumption rate ($E^{\text{O,N,w}}_{j,n}$), the thermal design power factor (β^{TDP}_j), and the job processing state ($u^{\text{job}}_{j,n,t}$) over all jobs j. Equation (30) sets the initial power state of node n. $\mathcal{N}^*$ is a user-defined node set. Equation (31) ensures that a node is always turned on when there is a job running on it. Similar to Constraint (9), Constraint (32) shows the logical relationship between the state of being powered on (u^{power}), turned on ($v^{\text{power,su}}$), and shut down ($v^{\text{power,sd}}$) of a node.

C. Heating, Ventilation, and Air Conditioning System

We enhance the model proposed by [7] to fit our problem. Due to space limitations, only the modified constraints will be presented here.

$$e^{\text{G}}_{t} = e^{\text{HVAC}}_{t} + e^{\text{DC}}_{t} + \Delta e^{\text{B}}_{t} - e^{\text{S}}_{t} + e^{\text{misc}}_{t},\ \forall t \in \mathcal{T}. \tag{33}$$

In Constraint (33), we reformulate the power balance constraint and define a new variable e^{G} to denote the power received from the thermal generators.

$$T_{i,t}^{\mathrm{Zone}} = T_{i,t-1}^{\mathrm{Zone}} + \sum_{i' \in \mathcal{N}(i)} \left(\frac{T_{i',t-1}^{\mathrm{Zone}} - T_{i,t-1}^{\mathrm{Zone}}}{C_i^{\mathrm{heat}} R_{i'i}^{\mathrm{Zone}}} \right) + \frac{\theta_{i,t}}{C_i^{\mathrm{heat}}} + \frac{\dot{m}_{i,t}^{\mathrm{Zone}} c^{\mathrm{a,s}} \left(T_{i,t}^{\mathrm{AC}} - T_{i,t-1}^{\mathrm{Zone}} \right)}{C_i^{\mathrm{heat}}}, \ \forall i \in \mathcal{I}, \forall t \in \mathcal{T}, \tag{34}$$

$$\text{where } C_i^{\mathrm{heat}} = c^{\mathrm{a,s}} \cdot \rho^{\mathrm{air}} \cdot S_i^{\mathrm{Zone}} \cdot h_i,$$

$$\dot{m}_{i,t}^{\mathrm{Zone}} = k_i^{\mathrm{AC}} \cdot v_t^{\mathrm{AC}},$$

$$\theta_{i,t} = \sum\nolimits_{n \in I_i^{\mathrm{Node}}} e_{n,t}^{\mathrm{O,Node}}.$$

We assume that the power used by the nodes is dissipated as heat within the IDC zone. Therefore, we have Equation (34). In Constraint (34), we let the variable internal heat generation (θ) be equal to the power used by the nodes within the corresponding zone.

$$x_i^{\mathrm{Zone}} \geq u_{n,t}^{\mathrm{power}}, \ \forall i \in \mathcal{I}, \forall t \in \mathcal{T}, \tag{35}$$

$$T_{i,t}^{\mathrm{Zone}} \leq T_{i,t}^{\mathrm{Zone,+}} + (1 - x_i^{\mathrm{Zone}}) T_{i,t}^{\mathrm{HI,+}}, \ \forall i \in \mathcal{I}, \forall t \in \mathcal{T}. \tag{36}$$

The role of these two new constraints (35) and (36) is that if all nodes in a zone are in the shutdown state, the maximum allowable zone temperature can be appropriately increased by $T_{i,t}^{\mathrm{HI,+}}$, which is a parameter. For completeness and a comprehensive understanding of the underlying constraints and their derivations, readers are referred to [7], where the rest of the constraints are detailed.

D. Generators

As a backup energy resource, thermal generators must be taken into account. Inspired by [11], we model thermal generators in the IDC using a simplified unit commitment (UC) problem as follows:

$$e_t^{\mathrm{G}} = \sum\nolimits_{i \in \mathcal{G}} p_{i,t}, \ \forall t \in \mathcal{T}, \tag{37}$$

$$c_t^{\mathrm{Op}} = \sum_{i \in \mathcal{G}} \left(\mathrm{SU}_{i,t} \cdot v_{i,t} + \beta_i \cdot p_{i,t} + \gamma_i \cdot u_{i,t} \right), \ \forall t \in \mathcal{T}, \tag{38}$$

$$u_{i,t} - u_{i,t-1} \leq v_{i,t}, \ \forall i \in \mathcal{G}, t \in \mathcal{T}, \tag{39}$$

$$\sum_{k=t-UT_i+1}^{t} v_{i,k} \leq u_{i,t}, \ \forall i \in \mathcal{G}, t \in [UT_i, T], \tag{40}$$

$$\sum_{k=t-DT_i+1}^{t} v_{i,k} \leq 1 - u_{i,t-DT_i}, \ \forall i \in \mathcal{G}, t \in [DT_i, T], \tag{41}$$

$$\underline{P}_i \cdot u_{i,t} \leq p_{i,t} \leq \overline{P}_i \cdot u_{i,t}, \ \forall i \in \mathcal{G}, \forall t \in \mathcal{T}, \tag{42}$$

$$p_{i,t} - p_{i,t-1} \leq RU_i \cdot u_{i,t-1} + \overline{RU}_i \cdot (1 - u_{i,t-1}), \quad \forall i \in \mathcal{G}, t \in \mathcal{T}, \tag{43}$$

$$p_{i,t-1} - p_{i,t} \leq RD_i \cdot u_{i,t} + \overline{RD}_i \cdot (1 - u_{i,t}), \quad \forall i \in \mathcal{G}, t \in \mathcal{T}. \tag{44}$$

Constraint (37) aggregates the individual power contributions of each generator to determine the overall energy production at each time step. Constraint (38) represents the expected future dispatch operation cost, including UC costs for flexible generation resources. We approximate quadratic cost functions for regular and flexible generation resources using linear expressions and appropriate factors like β and γ, respectively. Lower bounds for start-up variables are modeled in (39). Constraints (40) and (41) specify the minimum online and offline time requirements. Constraints (42) set limits for output power. Constraints (43) and (44) represent ramp-up and ramp-down constraints for regular generation resources. For completeness and a comprehensive understanding of the underlying constraints, readers are referred to [11], where the rest of the constraints are detailed.

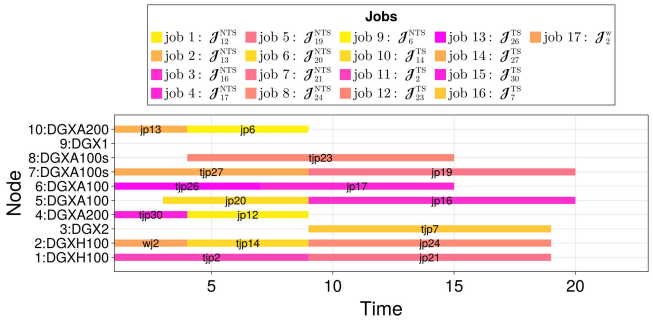

Fig. 1: Job Schedule of the IDC

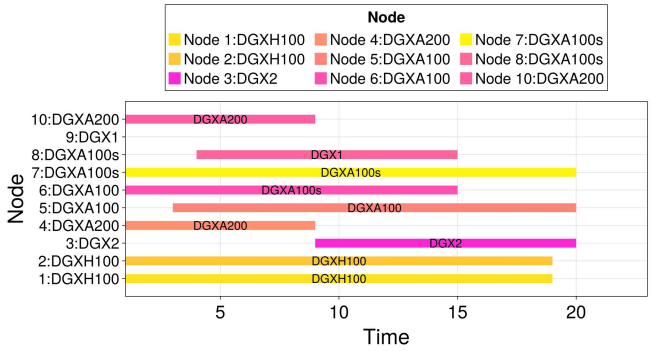

Fig. 2: Node Power Schedule of the IDC

III. Numerical Validation

We validate our proposed model on a hardware platform using the commercial solver Gurobi.

A. Experiments Setup

In our experimental setup, we consider an IDC comprising 10 computational nodes and a total of 33 jobs (3 jobs in $\mathcal{J}^{\mathrm{w}}$, 9 jobs in $\mathcal{J}^{\mathrm{TS}}$, and 21 additional jobs in $\mathcal{J}^{\mathrm{NTS}}$). The experiment spans a time range of 20 discrete time intervals, with each interval representing one hour. The nodes in the system are characterized by their specific power and computational capabilities. The power specifications for the nodes are defined in Table I and are provided in [12]. The nodes are categorized into different types, each corresponding to a

Done			Hold		TF	Abort
$\mathcal{J}_{12}^{\text{NTS}}$	$\mathcal{J}_{13}^{\text{NTS}}$	$\mathcal{J}_{2}^{\text{w}}$	$\mathcal{J}_{1}^{\text{NTS}}$	$\mathcal{J}_{10}^{\text{NTS}}$	$\mathcal{J}_{11}^{\text{NTS}}$	$\mathcal{J}_{29}^{\text{TS}}$
$\mathcal{J}_{16}^{\text{NTS}}$	$\mathcal{J}_{17}^{\text{NTS}}$	$\mathcal{J}_{30}^{\text{TS}}$	$\mathcal{J}_{15}^{\text{NTS}}$	$\mathcal{J}_{18}^{\text{NTS}}$	$\mathcal{J}_{5}^{\text{NTS}}$	$\mathcal{J}_{1}^{\text{w}}$
$\mathcal{J}_{19}^{\text{NTS}}$	$\mathcal{J}_{20}^{\text{NTS}}$	$\mathcal{J}_{7}^{\text{TS}}$	$\mathcal{J}_{18}^{\text{NTS}}$	$\mathcal{J}_{25}^{\text{NTS}}$	$\mathcal{J}_{22}^{\text{TS}}$	$\mathcal{J}_{3}^{\text{w}}$
$\mathcal{J}_{21}^{\text{NTS}}$	$\mathcal{J}_{24}^{\text{NTS}}$	$\mathcal{J}_{26}^{\text{TS}}$	$\mathcal{J}_{28}^{\text{NTS}}$	$\mathcal{J}_{3}^{\text{NTS}}$		
$\mathcal{J}_{6}^{\text{NTS}}$	$\mathcal{J}_{14}^{\text{TS}}$	$\mathcal{J}_{27}^{\text{TS}}$	$\mathcal{J}_{4}^{\text{NTS}}$	$\mathcal{J}_{8}^{\text{NTS}}$		
$\mathcal{J}_{2}^{\text{TS}}$	$\mathcal{J}_{23}^{\text{TS}}$		$\mathcal{J}_{9}^{\text{NTS}}$			

TABLE II: Final Job Statuses

Zone	Node
1	DGXA100s, DGXH100
2	DGXA1, DGXH100, DGXH100s
3	DGXH100, DGXA200, DGX2, DGXA200
4	DGXA100

TABLE III: Mapping of Nodes to Zones

key in the power and compute specifications dictionaries. An idle rate of 20% is assumed based on the findings in [13], indicating that each node consumes roughly 20% of its rated power when idle. This setup allows us to rigorously evaluate the performance and efficiency of our proposed job scheduling and resource allocation strategies within a controlled and detailed experimental environment. The rest of the setup parameters are stored in [14].

B. Simulation Result

We will analyze the results by examining several key aspects of the model. The analysis will include sections on job scheduling, node power management, zone temperature, battery reserves, and generators, culminating in a final analysis that integrates these components to provide comprehensive insights into our model's performance.

1) Job Scheduling and Node Power Management: The final result is presented in Table II, which categorizes the final statuses of various jobs into four distinct columns: "Done," "Hold," "Transferred," and "Abort." The data center ultimately achieves a net income of $2,677,697.15 during this emergency without violating any constraints.

For the jobs marked as done, a detailed graph is provided in Figure 1. The Gantt chart in Figure 1 illustrates the scheduling and execution of various jobs across different computational nodes of the IDC over a specified time range. Each job, identified by a unique color and labeled with an ID and name, is allocated to specific nodes, demonstrating efficient resource utilization. The x-axis represents time, while the y-axis lists the nodes. In addition, *jp* stands for $\mathcal{J}^{\text{NTS}}$, *tjp* stands for $\mathcal{J}^{\text{TS}}$, and *wj* stands for $\mathcal{J}^{\text{w}}$. The concurrent execution of jobs on different nodes maximizes resource usage and prevents conflicts during the emergency, ensuring no overlapping tasks on the same node. Overall, the Gantt chart provides a clear and concise visualization of the job scheduling strategy, proving that our model can achieve balanced resource allocation and effective performance in handling varied computational tasks during the emergency period.

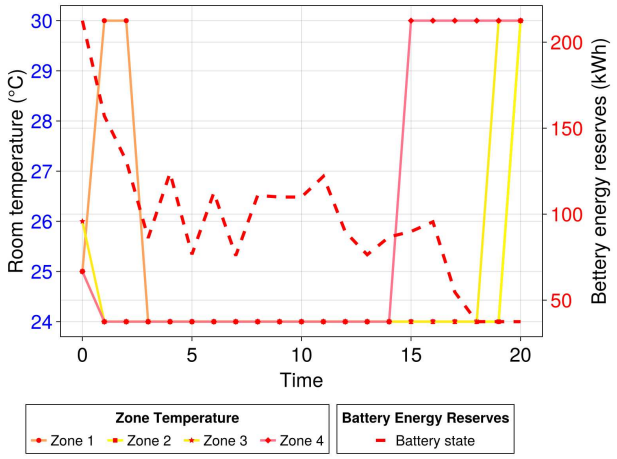

Fig. 3: The Trend of Zone Temperature and Battery Reserves

The Gantt chart in Figure 2 displays the power status of different nodes over time, highlighting the periods during which each node is active. The x-axis represents time, while the y-axis lists the nodes, identified by their specific types. The chart demonstrates the effective resource management of this model, where nodes are actively used without significant idle periods during their scheduled times.

2) Temperature and Battery Reserves: Figure 3 illustrates the variations in room temperature across four different zones and the battery energy reserves over the time period. The x-axis represents time, while the left y-axis (in blue) shows the room temperature in degrees Celsius, and the right y-axis (in red) indicates the battery energy reserves in kilowatt-hours (kWh). The battery energy reserves are represented by the red dashed line, while the temperature of each zone is shown by the solid lines. Overall, the management is exceptional. For example, since nodes 5 and 8 are both turned off during $t \in [1, 2]$, the model raises the upper bound of zone 1's temperature to save energy. Therefore, the model demonstrates a robust energy and thermal management strategy, crucial for maintaining optimal conditions across different zones while efficiently utilizing energy resources.

3) Generators: Figure 4 illustrates the ON/OFF status of three generators (G1, G2, and G3) over $\mathcal{T}$. Each sub-graph corresponds to one generator, with the y-axis indicating the ON/OFF state (with ON at the top and OFF at the bottom) and the x-axis representing time. Next, Figure 5 shows the power output of the same three generators over the same period. Each sub-graph corresponds to one generator, with the y-axis indicating the power output in kW. Since each time interval is one hour, the energy production is in kWh with the same value. The x-axis represents time. Together, these graphs provide a clear overview of the operational patterns and contributions of the three generators. G1 and G2 are the main contributors to the power supply, with G1 operating for most of the time and G2 running continuously. G3 remains inactive, likely reserved for contingencies or specific conditions not encountered during this period. This balanced approach to generator utilization ensures reliability and efficiency in power

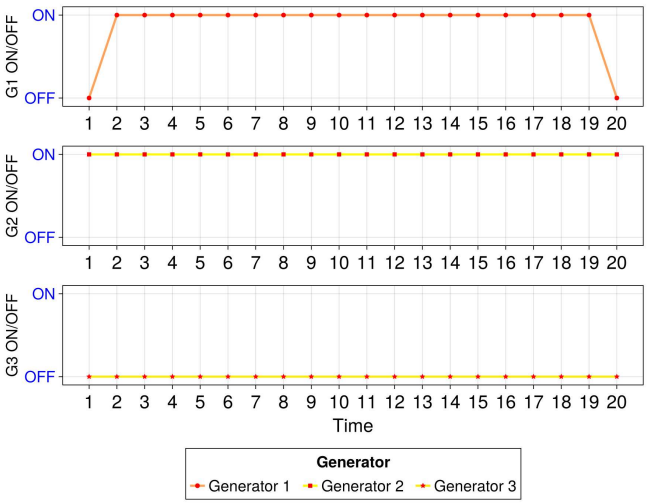

Fig. 4: The ON/OFF State of Generators

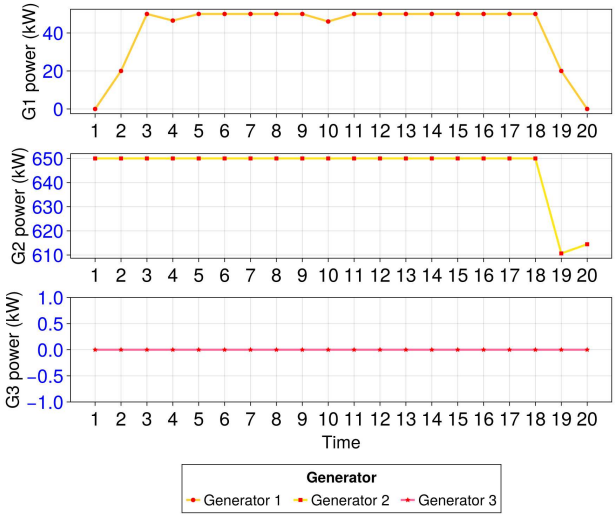

Fig. 5: The Power Generation of Generators

supply management.

4) Analysis: The model demonstrates an effective and balanced approach to resource management after an emergency, combining efficient job scheduling, optimal node utilization, exceptional HVAC management, and strategic power generation. Moreover, it helps IDCs achieve maximum net income while facing multiple constraints. The job scheduling across nodes shows a well-distributed workload, minimizing idle times and maximizing computational efficiency. Overall, the model we proposed successfully integrates power management and job scheduling to achieve high performance and operational stability during the emergency period.

IV. CONCLUSION

In this paper, we first propose a model for IDCs to handle job scheduling during emergency periods. We then use a test case to verify the correctness of our model. From the perspective of solution quality, the model exemplifies an effective and balanced approach to resource management during an emergency, including efficient task scheduling, optimal node utilization, superior HVAC control, and strategic power generation. Furthermore, it enables IDCs to maximize net revenue while addressing a variety of constraints. Additionally, our approach translates the results into a clear and concise visualization of job scheduling, making it easy for users to track job progress, node utilization, and other key functions over time. Therefore, our proposed model efficiently incorporates power management and computational work allocation to provide an optimal energy and IT service schedule for an IDC during an emergency period.

ACKNOWLEDGEMENT

This work is partially supported by NSF CNS-2107216, CNS-2128368, CMMI-2222810, ECCS-2302469, and ECCS-2045978.

REFERENCES

[1] A. Lawrence, "Uptime institute data shows outages are common, costly, and preventable," *Uptime Institute Research, New York, NY*, p. 10, Jun. 2018.

[2] "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption," Electric Power Research Institute (EPRI), Tech. Rep., May 2024.

[3] W. T. D. Centers. Most power consumed (MW). [Online]. Available: http://worldstopdatacenters.com/power/

[4] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar, "Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters," in *Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems*, Nov. 2012, pp. 75–86.

[5] J. Gao, H. Wang, and H. Shen, "Smartly handling renewable energy instability in supporting a cloud datacenter," in *IEEE International parallel and distributed processing symposium (IPDPS)*, May 2020, pp. 769–778.

[6] Y. Liu, S. Lei, and Y. Hou, "Restoration of power distribution systems with multiple data centers as critical loads," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5294–5307, Nov. 2018.

[7] Z. Zhao, L. Fan, and Z. Han, "Optimal data center energy management with hybrid quantum-classical multi-cuts Benders' decomposition method," *IEEE Transactions on Sustainable Energy*, vol. 15, no. 2, pp. 847–858, Aug. 2023.

[8] A. Chatterjee, M. Levan, C. Lanham, and M. Zerrudo, "Job scheduling in cloud datacenters using enhanced particle swarm optimization," in *IEEE International Conference for Convergence in Technology (I2CT)*, Apr. 2017, pp. 895–900.

[9] L. T. Jung and A. A. Haruna, "Green data center by incentive-based job scheduling approach," in *IEEE Conference on Open Systems (ICOS)*, Nov. 2018, pp. 13–18.

[10] W. Xuan, Z. Zhao, L. Fan, and Z. Han, "Minimizing delay in network function visualization with quantum computing," in *IEEE International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, Oct. 2021, pp. 108–116.

[11] S. Wang, C. Zhao, L. Fan, and R. Bo, "Distributionally robust unit commitment with flexible generation resources considering renewable energy uncertainty," *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4179–4190, Feb. 2022.

[12] "Nvidia dgx systems." [Online]. Available: https://docs.nvidia.com/dgx-systems/index.html

[13] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power," *ACM SIGARCH Computer Architecture News*, vol. 37, no. 1, pp. 205–216, Mar. 2009.

[14] Z. Zhao, "Djzts datacenter response." [Online]. Available: https://github.com/djzts/DataCenter_response