Data Center-Enabled High Altitude Platforms: A Green Computing Alternative

Wiem Abderrahim, Osama Amin and Basem Shihada

Abstract—Information technology organizations and companies are seeking greener alternatives to traditional terrestrial data centers to mitigate global warming and reduce carbon emissions. Currently, terrestrial data centers consume a significant amount of energy, estimated at about 1.5% of worldwide electricity use. Furthermore, the increasing demand for data-intensive applications is expected to raise energy consumption, making it crucial to consider sustainable computing paradigms. In this study, we propose a data center-enabled High Altitude Platform (HAP) system, where a flying data center supports the operation of terrestrial data centers. We conduct a detailed analytical study to assess the energy benefits and communication requirements of this approach. Our findings demonstrate that a data center-enabled HAP is more energy-efficient than a traditional terrestrial data center, owing to the naturally low temperature in the stratosphere and the ability to harvest solar energy. Adopting a data center-HAP can save up to 14% of energy requirements while overcoming the offloading outage problem and the associated delay resulting from server distribution. Our study highlights the potential of a data center-enabled HAP system as a sustainable computing solution to meet the growing energy demands and reduce carbon footprint.

Index Terms—Data centers, Data center-Enabled High Altitude Platforms, Energy-efficiency (EE), High altitude platforms (HAP), Workload offloading.

1 Introduction

Data centers are one of the top enabling technologies for the information technology industry, with a global market spending of 216,095 million US dollars in 2022 [1], [2]. They are considered the mission-critical infrastructure of this computing era because they play a pivotal role in processing and storing our continuously-growing data [3]. This role is unprecedentedly crucial given the data-intensive applications used in nowadays' advanced fields such as artificial intelligence and internet-of-everything applications [1], [4]–[6], which require extensive growth of this infrastructure's sizes and functionalities [1], [7]. However, data centers face major energy-efficiency issues [3], [8] as a large-scale computing infrastructure. Indeed, they consume not only substantial amounts of energy, around 1.5% of the worldwide electricity use but also the annual growth of the consumed energy by data centers is predicted to rise exponentially in the upcoming years [1], [9]–[12]. For example, the electricity demand of Google's data centers increased twenty fold over the last ten years [10]. Moreover, China's data centers will be devouring more than 400 billion kWh by 2030, accounting for 3.7 percent of the country's total electricity consumption [13]. In addition, the annual energy cost of data centers is expected to surpass their construction cost and equipment price within the upcoming few years [1], [3], [10]. These statistics are particularly alarming because the non-renewable energy is still the predominant source to generate electricity nowadays [9]. Therefore, serious research efforts should be conducted towards finding practical solutions that improve the current data center energy efficiency, which can help reduce global carbon emissions. According to Gartner, Inc., it is expected that by 2027, around 75% of organizations will have implemented a data center infrastructure sustainability program due to cost optimization and stakeholder pressures. [14].

The authors are with the Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Divison, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Makkah Prov., Saudi Arabia (e-mail: wiem.abderrahim, osama.amin, basem.shihada @kaust.edu.sa)

1.1 Literature Review

Data center energy consumption is primarily distributed between the cooling infrastructure (30%-40%) and the computing infrastructure (26%-56%) [3], [9], [15]. As a result, various energy-aware approaches have been investigated in the literature to improve data center energy efficiency from both computation and cooling perspectives [1], [9], [10], [16]–[20]. On the one hand, cooling energy can be reduced through strategies such as raised floors, racks' arrangement following hot/cold aisles, chillers configuration, and fan optimization [9], [16]. Additionally, load balancing across geographically distributed data centers can decrease cooling power. For instance, workload distribution and scheduling should consider the energy/cooling efficiency of servers, as well as electricity costs and experienced delays [1], [10], [17], [18]. On the other hand, data center computational energy can be reduced through dynamic management of capacity by controlling idle servers [3], [12], [17], [21]. Furthermore, optimizing server speed by adjusting central processing unit (CPU) frequency can help reduce associated computational power [19]. Another way to decrease energy consumption in data centers is to leverage green options. For example, renewable energy sources like wind and solar power can supply electricity to data centers, as demonstrated by companies such as Apple, Google, Microsoft, and Facebook [18], [22]. Consequently, it is vital to explore how data centers can utilize renewable energy and study the factors that favor

renewable energy over traditional sources, especially for geographically distributed data centers [18], [20], [23]. Most existing research efforts mainly focus on reducing computational energy and overlook cooling energy, a significant factor in terrestrial data centers [8] Moreover, renewable energy sources, such as solar and wind power, can be unreliable for mission-critical and large-scale data centers due to their dependence on variable weather conditions like cloud cover and wind patterns [11], [22].

1.2 Contributions

To overcome the continual growth of data centers' energy consumption, non-traditional energy-efficient computation paradigms are needed. We advocate data center-enabled High Altitude Platform (HAP) as a practical and green alternative to terrestrial data centers [24]. HAPs can be a core futuristic airborne network component that will revolutionize the networking frontier in the stratospheric range at an altitude between 17 km and 20 km [24]–[26]. HAPs offer several unique advantages, particularly from energy and ubiquity perspectives. Firstly, being located in the stratosphere saves cooling energy thanks to the naturally low atmospheric temperature, which ranges between 50 C and 15 C. Therefore, a HAP-enabled data center can offload some workloads from terrestrial data centers, saving the associated cooling energy. Additionally, HAPs can host large solar panels on their large surfaces that supply electricity to the data center servers and partially cover the required computational power [24]–[26]. Secondly, HAPs have a pervasiveness advantage over terrestrial data centers thanks to the large footprint offered by the line of sight (LoS) links to the terrestrial infrastructure, wireless communication abilities and flexible relocation abilities [24]–[26]. HAPs can host multiple-input multiple-output (MIMO) antennas on their large surfaces and provide higher data rates to their users [24], [26], [27]. However, it is necessary to define the limits and conditions under which HAP-enabled data centers can be beneficial, considering the practical limitations of HAPs, the quality of communication links, the range of proper offloaded workloads, and the impact of transmission delay on queuing time. Nonetheless, HAP-enabled data centers offer a promising and sustainable solution to the increasing energy consumption of traditional terrestrial data centers.

In this paper, we analyze the operation of a data center on a HAP considering practical and realistic operational conditions. Firstly, we compare the energy models of HAP-enabled and terrestrial data centers, demonstrating significant energy savings with the former. We then analyze the harvested energy requirements to maintain the HAP's flying condition while performing the required computation for the offloaded terestrial workload. Next, we examine the reliability of the transmission link between terrestrial data centers and HAP-based systems by studying the outage probability. To address transmission outages, we propose a re-transmit dropped workloads solution using a portion of the saved energy. Finally, we explore the delay experienced in transmitting or re-transmitting workloads to a HAP-enabled data center without affecting the queuing time and verify the conditions that enable reliable operation. Our study presents an extensive and pioneering analysis of energy, communication, and delay performance aspects of HAP-enabled data centers. The system model is presented in Section 2, followed by our system analysis in Section 3. In Section 4, we discuss and analyze our results. We conclude the paper in Section 5.

TABLE 1: Description of Main Notations

Notation	Description	Unit
I	Number of servers in the terrestrial data center	
J	Number of cooling units in the terrestrial data center	
I′	Number of servers in the HAP	
$_i$	Workload arrival rate to server s_i	task/s
u_i	Utilization ratio of server s_i	%
$_i$	Service rate of server s_i	MIPS
	Computational task length	Bits
P_i^{idle}	Average power of idle server s_i	Watt
P_i^{peak}	Average power of fully utilized server s_i	Watt
P^{fan}	Fan power	Watt
Q_j	Heat amount removed by cooling unit ac_j	Watt
COP	Performance coefficient of cooling unit ac_j	%
P_i^{comp}	Computational power of server s_i	Watt
E_i^{comp}	Computational energy of server s_i	Joule
$E_{\text{TDC}}^{\text{comp}}$	Computational energy of the terrestrial data center	Joule
P_j^{cool}	Cooling power of cooling unit ac_j	Watt
E_j^{comp}	Cooling energy of cooling unit ac_j	Joule
$E_{\text{TDC}}^{\text{cool}}$	Cooling energy of the terrestrial data center	Joule
l	HAP latitude	Degree
d	Considered day of the year	
$_{\text{pv}}$	Efficiency of the photo-voltaic system of the HAP	%
$\mathcal{A}_{\text{pv}}$	Area of the photo-voltaic system of the HAP	m^2
G	Total extra-terrestrial solar radiance per m^2	$W\{\text{m}^2$
	Air density	$kg\{\text{m}^3$
$_{\text{prop}}$	Propeller efficiency	%
v_{wind}	Wind velocity	m/s
v_{HAP}	HAP velocity	m/s
C_{D}	Drag coefficient	%
$_0$	Channel power gain at the reference distance	dB
L_{HAP}	Distance between the terrestrial data center and HAP	m
	Rician factor	
f_{carrier}	Carrier frequency	Hz
$P_{\text{HAP}}^{\text{harv}}$	Harvested power of the HAP	Watt
$E_{\text{HAP}}^{\text{harv}}$	Harvested energy of the HAP	Joule
$P_{\text{HAP}}^{\text{prop}}$	Propulsion power of the HAP	Watt
$E_{\text{HAP}}^{\text{prop}}$	Propulsion energy of the HAP	Joule
$P_{\text{HAP}}^{\text{payload}}$	Payload power of the HAP	Watt
$E_{\text{HAP}}^{\text{payload}}$	Payload energy of the HAP	Joule
$E_{\text{TDC-HAP}}^{\text{trans}}$	Transmission energy of the data center enabled-HAP	Joule

Notations: Lower case boldface letters denote vectors while upper case boldface letters denote matrices. The conjugate-transpose of matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^H$ and the transpose of matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^T$. $k\,k_F^2$ denotes the square of Frobenius norm. $k\,k$ denotes Euclidean norm. I_M denotes the M M identity matrix. $\mathbb{C}^{M\ N}$ denotes the complex space of M N. Prp q denotes the probability. Moreover, Table 1 summarizes the main notations used in this paper and their respective descriptions.

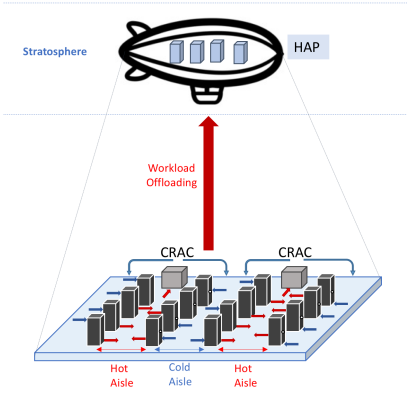

Fig. 1: Data Center-enabled HAP Architecture

2 System Model

Our system model comprises a terrestrial data center and airborne data centers to offload some of the workloads. The terrestrial data center includes a set $I = \{s_1, \ldots, s_I\}$ of I servers arranged in a hot aisle/cold aisle configuration. The cold aisles face each other, while the hot aisles face each other. Additionally, the terrestrial data center has a cooling system consisting of a set $J = \{ac_1, \ldots, ac_J\}$ of J computer room air conditioning (CRAC) units, as shown in Fig. 1. The purpose of the CRAC units is to maintain a temperature range of 18 C to 26 C, following the recommendation of the American Society of Heating, Refrigerating, and Air-Conditioning Engineers [28]. On the other hand, the data center-enabled HAP comprises a set $I^1 = \{s_{I+1}, \ldots, s_{I+I^1}\}$ of I^1 servers. Since the stratosphere's average temperature falls between 50 C and 15 C, we assume that the data center-enabled HAP does not require any CRAC units [25], [26].

2.1 Workload Model

The controller station (dispatcher) schedules the total workload that needs to be processed by our system among the servers $s_{i=1,\ldots,I+I^1}$ in the terrestrial data center and data center-enabled HAP, according to a Poisson process [10]. Let λ denote the total workload arrival rate, such that $K = \int_{t_1}^{t_2} \lambda(t)\,dt$, where K is the total number of tasks arriving in the system. The number of tasks K_i processed by server s_i can be expressed as $K_i = \int_{t_1}^{t_2} \lambda_i(t)\,dt$, where λ_i is the workload arrival rate to server s_i. We assume that the arrival rate follows a Poisson process when distributed by the controller station. Therefore, $\sum_{i=1}^{I+I^1} K_i = \sum_{i=1}^{I+I^1} \lambda_i (t_2 - t_1) = K$.

When the arriving workload is processed by server s_i, it transitions from the idle state to the active state and generates a CPU utilization ratio of $0 < u_i(\lambda_i) \leq \bar{u}_i \leq 100\%$, where $\bar{u}_i$ is the desirable utilization ratio of server s_i. We can express the utilization ratio u_i as a function of the arrival rate $u_i(\lambda_i) = \lambda_i \bar{\gamma}$, where $\bar{\gamma}$ is the expected value of the task length assumed to be constant for all tasks. The service rate of server s_i, denoted by μ_i and measured in millions of instructions per second (MIPS), refers to the capacity of its CPU to process tasks.

2.2 Power Model of the Terrestrial Data Center

A breakdown of the energy consumed by a data center shows that the computational infrastructure and the cooling infrastructure are the two primary components that absorb the data center energy [3]. Therefore, we focus on studying the energy consumed by these components in the following subsections.

2.2.1 Computational Model

The I servers of the data center have different characteristics in terms of CPU capacity. We assume that the server power consumption and the CPU utilization have a linear relationship [3]. Therefore, the computational power P_i^{comp} consumed by s_i can be expressed as [3], [9], [29]:

$$P_i^{\text{comp}}(\lambda_i) = P_i^{\text{idle}} + (P_i^{\text{peak}} - P_i^{\text{idle}})\, u_i(\lambda_i), \quad (1)$$

where P_i^{idle} is the average power value when server s_i is idle (i.e. $u_i(\lambda_i(t)) = 0\%$) and P_i^{peak} is the average power value when the server is fully utilized, i.e., $u_i(\lambda_i(t)) = \bar{u}_i$.

2.2.2 Cooling Model

The compressor and the fans of the CRAC units are the main consumers of the cooling energy in a data center [3], [10]. In a data center, the cold air supplied by the CRAC units enters through the cold aisle to the server inlets in the front of the server racks and is exhausted into the hot aisle through the server outlets in the back of the server racks. As a result of this air circulation, each node in the data center, whether a server $s_{i=1,\ldots,I}$ or a CRAC unit $ac_{j=1,\ldots,J}$, has an inlet temperature T^{in} and an outlet temperature T^{out}. The inlet temperature T^{in} represents the amount of heat received from other nodes, while the outlet temperature T^{out} represents the amount of heat contained in a given node. Therefore, we can express the cooling power $P^{\text{cool}}(j)$ of CRAC unit $ac_{j=1,\ldots,J}$ as the power necessary to cool the servers under its coverage. This cooling power can be calculated as follows [10], [30],

$$P_j^{\text{cool}}(t) = P_j^{\text{fan}} + \frac{Q_j(t)}{\text{COP}(T_j^{\text{out}})}, \quad (2)$$

where P^{fan} is the fan power, Q_j is the heat amount removed by CRAC unit ac_j and COP is the ac_j performance coefficient given by $\text{COP}(T_j^{\text{out}}) = 0.0068\,(T_j^{\text{out}})^2 + 0.008\,T_j^{\text{out}} + 0.458$.

2.3 Power Model of the Data Center-enabled HAP

Managing energy in HAPs is crucial because they are typically designed for long-duration missions (at least one year). Throughout this subsection, we study the energy harvested and consumed by a data center-enabled HAP.

2.3.1 Energy Harvesting Model

The main energy source for the daytime operation of HAPs is the solar energy [25], [26]. HAPs also incorporate energy storage components; which are typically Lithium-Sulphur batteries or hydrogen fuel cells [26], [31]. These batteries support the nighttime operation of the HAP and are fed by the solar energy harvested during the daytime [26], [31]. By

assuming that the energy harvested can be stored for 24 h, the mean solar power harvested by the HAP at latitude l on day d can be expressed as [31],

$$P^{harv}_{HAP}(l,d) = \eta_{pv} A_{pv} G(l,d), \tag{3}$$

where η_{pv} is the efficiency of the photo-voltaic system, A_{pv} is the area of the photo-voltaic system and G is the total extra-terrestrial solar radiance per m^2 and is given by [31],

$$G(l,d) = \frac{\tau(l,d) G_{max}(l,d)}{\pi\,\alpha_{max}(l,d)}\left(1 - \cos(\alpha_{max}(l,d))\right), \tag{4}$$

where $\tau(l,d)$ is the daylight duration at latitude l on day d, $G_{max}(l,d)$ is corresponding maximum radiation intensity and $\alpha_{max}(l,d)$ is the corresponding maximum altitude of the sun (in rd). $\tau(l,d)$ is expressed in hours from [31],

$$\tau(l,d) = 1 - \frac{1}{\pi}\cos^{-1}\left(\frac{\tan(l)\sin(\epsilon)\sin(\phi(d))}{\sqrt{1-\sin^2(\epsilon)\sin^2(\phi(d))}}\right), \tag{5}$$

where $\epsilon = 0.4093$ is the angle of obliquity of the Earth and $\phi(d)$ is the azimuthal angle of the sun and is given by [31],

$$\phi(d) = 1.3411 + M(d) + 0.0334\sin(M(d)) + 0.0003\sin(2M(d)) \tag{6}$$

such that $M(d) = -0.041 + 0.017202d$ is the angular distance of the sun, which is also known as mean anomaly [31]. $G_{max}(l,d)$ can be computed from [31],

$$G_{max}(l,d) = G_{solar}\left(1 + 0.033\cos\frac{360d}{365}\right)\cos(l - \delta(d)), \tag{7}$$

where $G_{solar} = 1366.1\ W/m^2$ is the standard solar constant at zero air mass and δ is the solar declination angle expressed as (in rd) [31]:

$$\delta(d) = 0.4093\sin\left(\frac{2\pi(d - 79.75)}{365}\right). \tag{8}$$

Whereas α_{max} is given by $\alpha_{max}(l_r,d) = \frac{\pi}{2} - l_r + \delta(d)$, where l_r is the latitude in rd [31].

2.3.2 Consumption Model

The power consumed by a HAP is determined mainly by both the payload and the propulsion subsystems [26], [31]. The required power by the payload is the computational power of the servers carried by the HAP. The required power by the HAP propeller can be expressed as [32],

$$P^{prop}_{HAP}(l,d) = \frac{\rho}{2\eta_{prop}} v^3_{wind}(l,d) v^{2/3}_{HAP} C_D, \tag{9}$$

where ρ is the air density, η_{prop} is the propeller efficiency, v_{wind} is the wind velocity, v_{HAP} is the HAP velocity and C_D is the drag coefficient, which is expressed as $C_D = N_C C_{envelope}$, such that N_C is a constant equal to 1.8 in the scope of this paper and $C_{envelope}$ is given by [32],

$$C_{envelope} = \frac{0.172 f_r^{\frac{1}{3}} + 0.252 f_r^{-1.2} + 1.032 f_r^{-2.7}}{R_e^{\frac{1}{6}}}, \tag{10}$$

where $f_r = \frac{L_{HAP}}{D_{HAP}}$ is the fineness ratio of the HAP body length L_{HAP} to its maximum width/diameter D_{HAP}, and R_e is the Reynolds number given by [32],

$$R_e = \frac{\rho v_{wind}(l,d) D_{HAP}}{\mu}, \tag{11}$$

where μ is the air dynamic viscosity [32].

2.4 Channel Model

We assume that both the terrestrial data center and the data center-enabled HAP utilize MIMO technology to overcome path loss and fading [27]. Specifically, the terrestrial data center is equipped with N antennas, and the HAP is equipped with M antennas. The channel between the terrestrial data center and the HAP can be modeled using Rician fading, which accounts for both a line-of-sight component and a scattered component. This channel model can be expressed as follows [27], [33]–[35]:

$$H_{HAP}(d) = \sqrt{\frac{\beta_0}{L^2_{HAP}}}\left(\sqrt{\frac{\kappa}{\kappa+1}}\bar{H}_{HAP}(d) + \sqrt{\frac{1}{\kappa+1}}\hat{H}_{HAP}\right), \tag{12}$$

where β_0 is the channel power gain at the reference distance; L_{HAP} is the distance between the terrestrial data center and the HAP; κ is the Rician factor; $\bar{H}_{HAP} \in \mathbb{C}^{M\times N}$ is the LoS channel component and $\hat{H}_{HAP} \in \mathbb{C}^{M\times N}$ is the Rayleigh fading component. The LoS channel component is given by $\bar{H}_{HAP}(d) = \exp\left(-j\frac{2\pi f_{carrier}}{C} r_{HAP}(d)\right)$ where $f_{carrier}$ is the carrier frequency; C is the speed of light and r_{HAP} is the length of the direct path between the transmit and receive antennas. The Rayleigh fading component $\hat{H}_{HAP}$ follows the distribution $\mathcal{CN}(0, I)$. Accordingly, the achievable data rate between the terrestrial data center and the HAP can be expressed as [27],

$$R_{HAP}(d) = B\log\det\left(I_M + H_{HAP}(d)q_{HAP}q^H_{HAP}H^H_{HAP}(d)n^{-1}\right), \tag{13}$$

where B is the bandwidth available for the transmission between the terrestrial data center and the HAP; $q_{HAP} \in \mathbb{C}^{N\times 1}$ is the precoding matrix of the terrestrial data center; $n \in \mathbb{C}^{M\times 1}$ is the additive white Gaussian noise (AWGN) with distribution $\mathcal{CN}(0, \sigma^2 I_M)$.

3 System Analysis

In this section, we analyze the performance of a data center-enabled HAP and compare it with a terrestrial data center. We begin by examining the conditions under which the energy harvested by the HAP's solar panels can sustain the HAP's flight and power the hosted servers. Next, we establish a comprehensive energy model for both systems and demonstrate that a data center-enabled HAP can achieve significant energy savings compared to a terrestrial data center. We then investigate the outage probability of the transmission link between the terrestrial data center and the HAP and explore whether the saved energy can be utilized to manage dropped workloads. Finally, we analyze the delay experienced in a data center-enabled HAP, which consists of the transmission delay and the mean waiting time that represents the queuing time of tasks in the servers before processing. To facilitate our analysis, we define the

workload arrival rate vector for the terrestrial data center as $_{TDC}$ $r_i s \in R^I$ and the workload arrival rate vector for the data center-enabled HAP as $_{HAP}$ $r_i s \in R^{I-1}$. By using these vectors, we can compare and evaluate the workload performance of the two systems.

3.1 HAP Flying Condition

Throughout the following analysis, we aim to determine whether the energy harvested by the solar panels on the HAP (E^{harv}_{HAP}) is sufficient to meet the energy requirements of both the payload ($E^{payload}_{HAP}$) and propulsion (E^{prop}_{HAP}). In our study, we consider the HAP payload to be a modular data center comprising a set of $I-1$ servers denoted by $I-1$ $s_{I-1}; ::s_{I-1}$. The power consumption of the payload is determined by the computational power of the servers, as given in equation (1). Therefore, the total power consumption of the payload can be expressed as,

$$P^{payload}_{HAP}(_{HAP}) \sum_{i=1}^{I-1} P^{idle}_i + P^{peak}_i - P^{idle}_i u_i(_i). \quad (14)$$

Hence, $E^{payload}_{HAP}$ can be determined by integrating equation (14), resulting in the following expression:

$$E^{payload}_{HAP}(_{HAP}) \sum_{i=I}^{I-1} P^{idle}_i + \frac{i}{i} P^{peak}_i - P^{idle}_i (t_2 - t_1). \quad (15)$$

Similarly, E^{prop}_{HAP} is derived by integrating equation (9), which yields

$$E^{prop}_{HAP}(d) = v^{\frac{17}{6}}_{wind}(d) C^{-1}_D (t_2 - t_1). \quad (16)$$

where

$$C^{-1}_D \frac{\frac{5}{6} v^{2\{3}_{HAP}}{2 _{prop}} \frac{\frac{1}{6} N_C \ 0{:}172 f_r^{\frac{1}{3}} - 0{:}252 f_r^{1:2} + 1{:}032 f_r^{2:7}}{D^{\frac{1}{6}}_{HAP}}.$$

Therefore, we define the flying condition of the HAP as follows:

$$E^{prop}_{HAP}(d) \le E^{payload}_{HAP}(_{HAP}) + E^{harv}_{HAP}(d). \quad (17)$$

The flying condition of the HAP is affected by three parameters: I, d and $_{HAP}$. To maintain the desired flying condition, we control the workload arrival rate that can be processed by the servers on the HAP, i.e., $_{HAP} \in R^{I-1}$. Specifically, for a fixed latitude I and day of operation d, the workload arrival rate $_{HAP}$ must satisfy the constraint $_{HAP} \le ^{max}_{HAP}$, where $^{max}_{HAP}$ is the maximum allowable workload arrival rate. Hence, the corresponding energy consumption is $E^{payload}_{HAP}(^{max}_{HAP}) - E^{harv}_{HAP}(I; d) - E^{prop}_{HAP}(I; d)$.

Assuming symmetrical characteristics for the servers, including the same service rate ($_i$), desirable utilization ratio ($\bar{u}_i$ $\bar{u}$), and power consumption (P^{idle}_i P^{idle} and P^{peak}_i P^{peak}), the workload can be evenly distributed across multiple servers with $_i$ ^{max} ^{max}. To determine the maximum allowable workload arrival rate ^{max}, we can use the following expression:

$$^{max}(I; d) = \frac{_{pv} A_{pv} (I; d) G_{max}(I; d)(1 - \cos(_{max}(I; d)))}{24 _{max}(I; d)} - v^{\frac{17}{6}}_{wind}(I; d) C^{-1}_D \ I^{-1} P^{idle} \frac{}{I^{-1}(P^{peak} - P^{idle})}. \quad (18)$$

3.2 Energy Saving

In this section, we investigate and compare the energy consumed by a terrestrial data center E^{cons}_{TDC} to the energy consumed by a data center-enabled HAP $E^{cons}_{TDC\text{-}HAP}$. Hence, we assume that all the workload directed to the HAP is successfully transmitted and no link outage is experienced. To establish a fair comparison, we assume that both systems have I $I-1$ servers in total. Therefore, I $I-1$ servers are placed in the terrestrial data center. In the data center-enabled HAP, I servers are placed in the terrestrial data center and $I-1$ servers are placed in the HAP.

3.2.1 Energy Consumption in Terrestrial Data Center

To understand the energy consumption of a terrestrial data center, i.e., E^{cons}_{TDC}, we break it down into two components: the computational energy consumed by the servers ($E^{comp}_{TDC}(_{TDC})$) and the cooling energy consumed by CRAC units ($E^{cool}_{TDC}(_{TDC})$), which is written as,

$$E^{cons}_{TDC}(_{TDC}) = E^{comp}_{TDC}(_{TDC}) + E^{cool}_{TDC}(_{TDC}).$$

E^{comp}_{TDC} captures the computational energy consumed by all the servers expressed as,

$$E^{comp}_{TDC}(_{TDC}) \sum_{i=1}^{I-1} E^{comp}_i.$$

To calculate the computational energy consumed by each server, we integrate equation (1) and obtain the value of E^{comp}_i as,

$$E^{comp}_i(_i) = P^{idle} \frac{}{} + \frac{i}{} P^{peak} - P^{idle} (t_2 - t_1). \quad (19)$$

E^{cool}_{TDC} captures the energy consumed by the J CRAC units, i.e.,

$$E^{cool}_{TDC}(_{TDC}) \sum_{j=1}^{J} E^{cool}_j;$$

where the cooling energy consumed by CRAC unit j, denoted as E^{cool}_j, is obtained by integrating (2) as,

$$E^{cool}_j \int_{t_1}^{t_2} P^{fan}_j \frac{Q_j(t)}{COP(T^{out}_j)} dt; \quad (20)$$

where P^{fan}_j is computed from

$$P^{fan}_j \frac{\dot{V}^{fan}_j (p^{fan}_j)}{_{fan,motor}};$$

with $\dot{V}^{fan}_j$ denoting the air flow rate of the CRAC unit, p^{fan}_j is the pressure loss due to air flow resistances, $_{fan}$ and

$_j^{fan,motor}$ are the efficiencies of the fan and motor fan, respectively [36].

The heat removed by CRAC unit ac_j from I_1 I servers in the data center, i.e., $Q_j(t)$ is found from [9],

$$Q_j(t) = \sum_{i=1}^{I_1} C_{air} f_{air} (T_i^{out}(t) - T_i^{in}(t)); \quad (21)$$

where C_{air} is the air heat capacity, f_{air} is the air flow rate through the server CPU; T_i^{in} and T_i^{out} are respectively the inlet temperature and the outlet temperature of server s_i. T_i^{in} is expressed as [10],

$$T_i^{in}(t) = T_j^{out} + (T_i^{in}(0) - T_j^{out})e^{-t}; \quad (22)$$

where $T_i^{in}(0)$ is the initial inlet temperature of server s_i; is the temperature raise imposed by the recirculated exhausted air in the data center and reflects the temperature influence of the closest CRAC. On the other handm, T_i^{out} is found from [9],

$$T_i^{out}(t) = (1 - \frac{1}{C_{air} f_{air} R}) T_i^{in}(t) + \frac{1}{C_{air} f_{air} R} T_i^{CPU}(t); \quad (23)$$

where R is the thermal resistance of the CPU and T_i^{CPU} is the CPU temperature such that [10], [37]:

$$T_i^{CPU}(t) = T_i^{in}(t) + RP_i^{comp} + (T_i^{CPU}(0) - T_i^{in}(t) - RP_i^{comp})e^{-\frac{t}{RC_i}}; \quad (24)$$

where $T_i^{CPU}(0)$ is the CPU temperature at the initial time and C_i is the server heat capacity. Substituting (22), (23) and (24) in (21) yields:

$$Q_j(t) = \frac{1}{R}\sum_{i=1}^{I_1} R(1 - e^{-\frac{t}{RC_i}})P_i^{comp} + (T_i^{CPU}(0) - T_j^{out})e^{-\frac{t}{RC_i}} + (T_i^{in}(0) - T_j^{out})(e^{-\frac{1}{RC_i}t} - e^{-\frac{1}{RC_i}t}). \quad (25)$$

By substituting (25) in (20), E_j^{cool} consumed by ac_j can be re-written as,

$$E_j^{cool} = \sum_{i=1}^{I_1} \frac{C_i(T_i^{CPU}(0) - T_j^{out})(e^{-\frac{t_1}{RC_i}} - e^{-\frac{t_2}{RC_i}})}{0.0068(T_j^{out})^2 + 0.008T_j^{out} + 0.458} + \sum_{i=1}^{I_1} \frac{C_i(T_i^{in}(0) - T_j^{out})(e^{-\frac{1}{RC_i}t_2} - e^{-\frac{1}{RC_i}t_1})}{(RC_i - 1)(0.0068(T_j^{out})^2 + 0.008T_j^{out} + 0.458)} + \sum_{i=1}^{I_1} \frac{(P^{idle} + \frac{i}{I}(P^{peak} - P^{idle}))(t_2 - t_1)}{0.0068(T_j^{out})^2 + 0.008T_j^{out} + 0.458} P_j^{fan}(t_2 - t_1) + \sum_{i=1}^{I_1} \frac{(P^{idle} + \frac{i}{I}(P^{peak} - P^{idle}))RC_i(e^{-\frac{t_2}{RC_i}} - e^{-\frac{t_1}{C_i}})}{0.0068(T_j^{out})^2 + 0.008T_j^{out} + 0.458}. \quad (26)$$

By summing over (19) and (26), $E_{TDC}^{cons}(_{TDC})$ can be re-written as,

$$E_{TDC}^{cons}(_{TDC}) = \sum_{i=1}^{I_1}(P^{idle} + \frac{i}{I}(P^{peak} - P^{idle}))(t_2 - t_1) + \sum_{j=1}^{J} P_j^{fan}(t_2 - t_1) + \sum_{i=1}^{I_1}\sum_{j=1}^{J} \frac{C_i(T_i^{CPU}(0) - T_j^{out})(e^{-\frac{t_1}{RC_i}} - e^{-\frac{t_2}{RC_i}})}{0.0068(T_j^{out})^2 + 0.008T_j^{out} + 0.458} + \sum_{i=1}^{I_1}\sum_{j=1}^{J} \frac{C_i(T_i^{in}(0) - T_j^{out})(e^{-\frac{1}{RC_i}t_2} - e^{-\frac{1}{RC_i}t_1})}{(RC_i - 1)(0.0068(T_j^{out})^2 + 0.008T_j^{out} + 0.458)} + \sum_{i=1}^{I_1}\sum_{j=1}^{J} \frac{(P^{idle} + \frac{i}{I}(P^{peak} - P^{idle}))(t_2 - t_1)}{0.0068(T_j^{out})^2 + 0.008:T_j^{out} + 0.458} + \sum_{i=1}^{I_1}\sum_{j=1}^{J} \frac{(P^{idle} + \frac{i}{I}(P^{peak} - P^{idle}))RC_i(e^{-\frac{t_2}{RC_i}} - e^{-\frac{t_1}{RC_i}})}{0.0068(T_j^{out})^2 + 0.008T_j^{out} + 0.458}. \quad (27)$$

3.2.2 Energy Consumption in Data Center-enabled HAP

The energy consumed by the data center-enabled HAP is expressed as a function of I, d and the processed workload $r = (_{HAP}, _{TDC})^T$ P^{I-1}:

$$E_{TDC\text{-}HAP}^{cons}(\ ; I; d) = E_{HAP}^{payload}(_{HAP}) + E_{HAP}^{prop}(I; d) + E_{TDC\text{-}HAP}^{trans}(I; _{HAP}) + E_{TDC\text{-}HAP}^{comp}(_{TDC}) + E_{TDC\text{-}HAP}^{cool}(_{TDC}); \quad (28)$$

Specifically, $E_{TDC\text{-}HAP}^{cons}$ comprises the computational energy (19) and the cooling energy (26) besides the payload energy (15), the propulsion energy of the HAP (16) and the transmission energy. The transmission energy required to send the workload from the terrestrial data center to the HAP is expressed as [27]:

$$E_{TDC\text{-}HAP}^{trans}(I; _{HAP}) = \frac{b(_{HAP})}{R_{HAP}(I)} k(_{HAP})k_F^2(t_2 - t_1); \quad (29)$$

where is the ratio of the transmitted data size to the original task data size due to the transmission overhead; b is the size of the input data (in bits) and $k(_{HAP})k_F^2$ is the transmit power from the terrestrial data center to the HAP. The size of the input data b can be expanded as $b(_{HAP}) = \sum_{i=I_{-1}}^{I-1}(_i - b)$ such that b is the size of the instruction in bits.

By substituting (15), (16), (29) in (28) and summing over (19) and (26), the energy consumed by the data center-

enabled HAP is re-written as:

$$E^{cons}_{TDC\text{-}HAP}(\ ;l;d_q) \quad v_{wind}^{\frac{17}{6}}(l;d_q)C_D^{1}(t_2 \quad t_1) \quad \frac{b}{R_{HAP}(l_q)}k_q{}_{HAP}k_F^2(t_2 \quad t_1)\sum_{j\ 1}^{J}P_j^{fan}(t_2 \quad t_1) \quad \sum_{i\ 1}^{I\ 1}P^{idle} \quad \frac{i}{} \quad P^{peak} \quad P^{idle} \quad (t_2 \quad t_1) \quad \sum_{i\ 1}^{I}\sum_{j\ 1}^{J}\frac{C_i(T_i^{CPU}(0q) \quad T_j^{out} \quad q)(e^{\frac{t_1}{RC_i}} \quad e^{\frac{t_2}{RC_i}})}{0{:}0068\ T_j^{out\ 2} \quad 0{:}008T_j^{out} \quad 0{:}458} \quad \sum_{i\ 1}^{I}\sum_{j\ 1}^{J}\frac{C_i(T_i^{in}(p0q) \quad T_j^{out} \quad q)(e^{p\ \frac{1}{RC_i}(t_2} \quad e^{p\ \frac{1}{RC_i}(t_1)q})}{p(RC_i \quad 1q\ 0{:}0068\ T_j^{out\ 2} \quad 0{:}008T_j^{out} \quad 0{:}458} \quad \sum_{i\ 1}^{I}\sum_{j\ 1}^{J}\frac{P^{idle} \quad \frac{i}{} \quad P^{peak} \quad P^{idle} \quad (t_2 \quad t_1)}{0{:}0068\ T_j^{out\ 2} \quad 0{:}008T_j^{out} \quad 0{:}458} \quad \sum_{i\ 1}^{I}\sum_{j\ 1}^{J}\frac{P^{idle} \quad \frac{i}{} \quad P^{peak} \quad P^{idle} \quad RC_i\ e^{\frac{t_2}{RC_i}} \quad e^{\frac{t_1}{RC_i}}}{0{:}0068\ T_j^{out\ 2} \quad 0{:}008T_j^{out} \quad 0{:}458}: \tag{30}$$

We propose that using a data center-enabled high-altitude platform (HAP) is a more energy-ef cient option compared to a terrestrial data center. This is because the HAP does not require cooling units, as the average temperature in the stratosphere is signi cantly lower than the recommended temperature for a data center. Additionally, the HAP utilizes solar energy harvested during the daytime and stored in Lithium-Sulphur batteries during the nighttime to power its servers, whereas a terrestrial data center relies on electric energy supplied through the electrical grid constantly. As a result, the energy saved by using the HAP can be expressed as the difference between the energy consumed by the HAP data center, denoted as $E^{cons}_{TDC\text{-}HAP}(\ ;l;d_q)$, and that consumed by the terrestrial data center, denoted as $E^{cons}_{TDC}(\ TDC_q)$, which can be written as $E^{sav}(\ ;l;d_q) \quad E_{TDC\text{-}HAP}{}^{cons}(\ ;l;d_q) \quad E^{cons}_{TDC}(\ _{TDC_q})$.

3.3 Outage Probability of Of oading

In this section, we investigate the outage probability of of- oading workload to the data center-enabled high-altitude platform (HAP). Speci cally, we analyze the scenario where the established transmission link between the terrestrial data center and the HAP cannot support the of oaded workload. As we assume that this link uses MIMO technology, we refer to the outage probability analysis of MIMO Rician fading channels presented in [38]. In [38], the authors derive a lower bound and an upper bound of the outage probability's distribution by considering the trace of a non-central Wishart matrix derived from the channel matrix. Therefore, the upper bound Pr_{tr_U} of the complementary cumulative distribution function (CCDF) of the data rate for the transmission link to the HAP is given by [38]:

$$Pr(R_{HAP} \quad i \quad q \leq Pr_{tr\_U}(\ q) \quad Q_{MN}\left(\sqrt{2\ 1};\ 2a\sqrt{N\ 2\bar{\gamma}\ 1}\right); \tag{31}$$

where $_{HAP}$ is the total of oaded workload, $\bar{H}_{HAP}\ ^{2}_{F}$ is the non-centrality parameter, $a\ \frac{1}{1}$ with ^{1} de ned as the average signal to noise ration (SNR) at each receive antenna and $Q_{MN}(a;y_q)$ is the generalized Marcum Q-function of order $M \quad N$ given by:

$$Q_{MN}(a;y_q) \quad \int_{y}^{\infty} x\left(\frac{x}{a}\right)^{MN\ 1} e^{\frac{x^2\ a^2}{2}} I_{MN\ 1}(axq)dx; \tag{32}$$

with $I_{MN\ 1}$ is denoting the modi ed Bessel function of the rst kind of order $MN\ 1$.

Moreover, the lower bound Pr_{tr_L} of the CCDF of the HAP data rate is given by [38]:

$$Pr(R_{HAP} \quad i \quad q \geq Pr_{tr\_L}(\ q) \quad Q_{MN}\left(\sqrt{2\ 1};\ 2a\sqrt{2 \quad 1q}\right): \tag{33}$$

Given the upper bound and the lower bound of the CCDF of the HAP data rate, we investigate the dropping rate of the data center-enabled HAP; which is given by:

$$Pr_{drop}(\ q) \quad 1 \quad Pr(R_{HAP} \quad i \quad q; \tag{34}$$

To mitigate this dropping rate, the dropped workload can be re-transmitted thanks to the energy saved by the data center-enabled HAP (as studied in the previous subsection). Accordingly, the number of re-transmissions N_r can be expressed as a function of the dropped workload ^{drop} as follows:

$$N_r \quad \begin{cases} \left\lfloor \frac{E^{sav}(\ q)}{E^{trans}_{TDC\text{-}HAP}(\ ^{drop}q)} \right\rfloor, & \text{if } \geq {}^{drop*}; \\ 0, & \text{otherwise}: \end{cases} \tag{35}$$

with ^{drop*} denotes the maximum workload arrival with no dropped workload such that $Pr(R_{HAP} \quad i \quad q \quad 1; @ \ ^{drop*}$.

3.4 Delay in Data Center-enabled HAP

In this section, we analyze the delay D that occurs in a data center-enabled high-altitude platform (HAP) system. This delay is primarily composed of two components: the waiting time, also known as the queuing time, denoted as W_i, for each task to be executed on server s_i, and the round trip time (RTT) required to send the task to the HAP and receive the execution result, such that $D \quad \bar{W}_i \quad RTT$. The RTT is determined by the transmission delay and can be expressed as follows:

$$RTT(l;\ _{HAP}q \quad 2t_{trans}(l;\ _{HAP}q \quad \frac{2b(\ _{HAP}q}{R_{HAP}(l_q)}. \tag{36}$$

In order to calculate the mean waiting time, we model the server as an M/G/1 queue with vacations, since the server enters an idle mode when there are no tasks to be executed. Our approach closely follows the standard derivations for M/G/1 queues, as described in [39]–[41]. Speci cally, we employ the following:

- the arriving workload follows a Poisson process with rate $_i$.
- the service time distribution is general because the control commands are assumed to be random.

We are interested in obtaining the mean waiting time $\bar{W}_i$ at server s_i and its second moment $\bar{W}_i^2$. The waiting time of task k scheduled after $K\ 1$ tasks is given by

$W_i = \sum_{k=1}^{k-1} X_i^{k-1} + R_i$; where X_i^{k-1} is the service time of task $k-1$ that arrived before task k at server s_i and R_i is the residual service time of server s_i. R_i can be either residual service time R_i^s or residual vacation time R_i^v depending on utilization. The first moment of this residual time can then be written as:

$$\overline{R_i} = \overline{R_i^s} = \overline{R_i^v} = \frac{1}{2} u_i \frac{\overline{X_i^2}}{\overline{X_i}} + 1 - u_i + \frac{\overline{V_i^2}}{\overline{V_i}} \quad (37)$$

$$\frac{1}{2} \sum_i \overline{X_i^2} + 1 - u_i + \frac{\overline{V_i^2}}{\overline{V_i}} ;$$

where u_i is the utilization ratio of server s_i, $\overline{V_i}$ is the mean vacation time duration at server i and $\overline{V_i^2}$ is the second moment of vacation time duration at server i. By applying Little's formula, we obtain the average waiting time as follows:

$$\overline{W_i} = \frac{\overline{R_i}}{1 - u_i} = \frac{\sum_i \overline{X_i^2}}{2(1 - u_i)} + \frac{\overline{V_i^2}}{2\overline{V_i}}. \quad (38)$$

3.5 Complexity Analysis

In this subsection, we analyze the complexity of the proposed data center enabled HAP system and we discuss its feasibility for implementation in a realistic setup. The complexity analysis considers various factors that include the computational complexity, the communication complexity and the deployment complexity.

3.5.1 Computational Complexity

To assess the computational complexity of the proposed framework, we consider the time and resources required to process the offloaded workload form the terrestrial data center. The energy saving complexity of our proposed framework can be evaluated as $O(\frac{N}{\mu})$ where λ is the workload arrival rate N is the total number of servers in the system and μ is the server capability. The time complexity of our proposed framework can be evaluated as $O(\frac{N}{2\mu})$, where we consider the waiting time and the offloading time of the tasks to the HAP. We notice the trade-off between the energy saving offered by the HAP-enabled system and the increased computational requirements due to the limited resources available on the HAP. Specifically, the more servers are deployed in the HAP with increasing workload offloading, the more energy is saved in the terrestrial data center. However, this strategy impacts resource utilization since energy consumption is strongly coupled to servers' capabilities. Accordingly, over-utilization and under-utilization of the flying servers must be avoided during resource' provisioning and allocation [24].

3.5.2 Communication Complexity

The communication complexity of the proposed framework involves the analysis of the communication links between the terrestrial data centers and the HAP. This includes the evaluation of the bandwidth requirements, signal propagation delays and the reliability of these links through the outage probability as studied in section 3.3. We can model the communication complexity by using a combination of metrics such as the bit error rate (BER), signal-to-noise ratio (SNR) and throughput.

3.5.3 Deployment Complexity

The deployment complexity of the proposed framework involves the practical challenges associated with deploying and maintaining HAPs in the stratosphere. This complexity encompasses the HAP positioning and buoyancy control; which are tightly related to weather conditions in the stratosphere determined mainly by the wind speed. Moreover, this complexity includes the use of sophisticated and resilient electronic devices in the flying servers to guarantee their operation in the low temperature of the stratosphere and afford a reliable computing service. The deployment complexity involves also the regulatory rules; which impose strict conformity with the standards of design, transport and operation to guarantee airships' safety in the stratosphere.

In summary, the complexity analysis of the proposed framework demonstrates its feasibility for implementation in a realistic setup while considering the important trade-offs. The energy savings and environmental benefits offered by the data center enabled HAP outweigh the computational and communication complexity. Moreover, the deployment challenges can be addressed through technological advancements and regulatory efforts.

4 Results and Discussion

In this section, we consider a stratospheric airship-based high-altitude platform (HAP) that has a significant surface area and can harvest notable amounts of solar energy. We assume that the HAP can support a maximum payload of 450kg, similar to the Stratobus airship HAP [26], [42]. Additionally, we assume that the hosted data center has a rack weight of 363kg [43], with each server in the rack characterized by a service rate of 580MIPS and a weight of approximately 9kg [43]. The remaining parameters used in our simulations are detailed in Table 2.

Throughout our numerical results, we employ the derived expressions in section 3 to quantify the performance gain of the data center-enabled HAP compared to a terrestrial data center under various operational conditions. To analyze the scalability of our solution, we consider various workload scenarios including the homogeneous workload and heterogeneous workload. The considered workload model in each scenario has a wide range of arrival rates throughout all the conducted simulations. The homogeneous workload scenario involves workloads with the same characteristics in terms of task's size. The heterogeneous workloads scenario involves workloads characterized with small task length and workloads characterized with large task length. For each workload scenario, we investigate the data center enabled HAP's ability to handle the increased demand by augmenting the workload arrival rate without significantly impacting the performance metric.

We consider four simulation scenarios to investigate the impact of HAP location over the course of a year, workload arrival rate, and the number of servers present in the HAP on the relative performance gain. Firstly, we verify the HAP's flying condition in the homogeneous workload scenario by monitoring the energy balance based on the number of servers in the HAP and the maximum workload

Fig. 2: Maximum Workload Arrival Rate under the HAP Flying Condition

arrival rate. Secondly, we evaluate the energy ef ciency of the data center-enabled HAP by assessing the saved energy for different latitudes throughout the year. To further validate the scalability of our proposed framework, we consider the deployment of a constellation of HAPs along with the terrestrial data center. Thirdly, we explore the impact of retransmitting dropped tasks in case of link outage on system performance for heterogeneous workload scenario. Finally, we analyze the delay experienced in the data center-enabled HAP by comparing the queuing delay, i.e., the waiting time, to the transmission delay in the heterogeneous workload scenario.

TABLE 2: Simulation Settings

Type	Parameter	Numerical Value
Cooling/ Thermal Inputs [10]	Supply Temperature	299:15 K
	Server Initial Temperature	310 K
	CPU Initial Temperature	318 K
	Thermal Resistance	0:34 K/W
	Server Heat Capacity	340 J/K
HAP Inputs [31], [32]	Area of the PV	8000 m^2
	Ef ciency of the PV	0:4
	Propeller ef ciency	0:8
	Air Density	0:08891 $kg\{m^3$
	Dynamic Air Viscosity	1:422:10 ^{5} $N:s\{m^2$
Trans- mission Inputs [27]	Antennas in TDC	2
	Antennas in HAP	16
	Carrier Frequency	31 GHz
	Channel Bandwidth	100 MHz

4.1 HAP Flying Condition

The rst simulation example investigates the workload arrival rate, i.e., ^{max}, that maintains the HAP ying condition versus the number of HAP servers, i.e., I^1, for different locations and time along the year, as depicted in Fig. 2. It is necessary to consider the physical capacity of our system beforehand. Speci cally, the maximum payload supported by the HAP imposes a limit on the maximum number of servers hosted in the HAP, as shown in Fig. 2. Moreover, the server's service rate characteristic combined with the desired utilization ratio $\bar{u}$ dictate a limit on the supported workload. Since we want to leverage the full performance of the HAP, we assume that the HAP is fully occupied and consider the Maximum HAP Payload limit in our simulations. Moreover, we suppose that the servers hosted in the HAP are fully-utilized by imposing that $\bar{u}$ 100% This maximum utilization ratio yields a High-load Threshold for the workload arrival rate, as shown in Fig. 2; beyond which the servers are dysfunctional.

We observe rst from Fig. 2 that the harvested energy is under-utilized when the number of airborne servers is low because the desirable utilization ratio $\bar{u}$ restricts the accepted workload. For instance, the supported maximum arrival rate in the Northern Hemisphere is around 2600task/s for ten servers on December Solstice. However, the actual server capacity of the server is below 1000 task/s even when it is highly loaded. Therefore, the number of servers in the HAP should be increased to utilize the harvested energy fully. However, there is a limitation on the maximum number of airborne servers according to the supported HAP payload. Indeed, we notice that the harvested energy and the servers' capacity are ideally utilized when the number of servers reaches the maximum HAP payload. Moreover, we notice a lower maximum arrival rate allowed per server when more servers are present in the HAP. This observation is because the same number of tasks can be distributed and processed by more servers to maintain the ying condition. Otherwise, the harvested energy would not cover the payload energy, which is equivalent to the computational energy of the servers. We also notice an opposite behavior for the maximum arrival rate when comparing the Northern hemisphere to the Southern hemisphere. For instance, in the Northern hemisphere, more tasks are accepted during June than December solstice. However, more tasks are carried out in the Southern hemisphere during December than June solstice. This observation is because the solar radiation and the daylight duration are more critical during June in the Northern hemisphere; because the Northern hemisphere is closer to the Sun during June. However, the solar radiation and the daylight duration are more critical during December in the Southern hemisphere because the Southern hemisphere is closer to the Sun during December. Therefore, more solar energy can be harvested during June in the Northern hemisphere and December in the Southern hemisphere. Hence, more computational energy can be covered, and accordingly, more tasks can be processed in the HAP during June in the Northern hemisphere and December in the Southern hemisphere.

Since HAPs are designed to operate for long-duration missions, we assume that the full payload is used when the airship is launched, and we study the maximum workload arrival rate variation according to days and latitudes. We note that it is crucial to consider the High-load Threshold imposed by the desirable utilization ratio of the servers. Indeed, as depicted in Fig. 3, the servers tend to be over-utilized around June solstice in the Northern hemisphere and under-utilized in the Southern hemisphere if the utiliza-

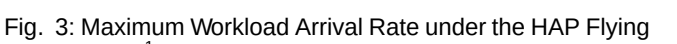

Fig. 3: Maximum Workload Arrival Rate under the HAP Flying Condition. (I ^{1} 40)

tion ratio is overlooked. More generally, the servers' computational capacity is better utilized with higher latitudes in both hemispheres. Moreover, we notice that the maximum arrival rate reaches on June solstice its maximum point in the Northern hemisphere and its minimum point in the Southern hemisphere. The explanation of this observation was deeply detailed in the previous paragraph. We notice also that the maximum arrival rate increases with latitudes in the Southern hemisphere. In the Northern hemisphere, the maximum arrival rate decreases with latitudes when d P r1; 100s Y r250; 366s and increases with latitudes when d P r100; 250s. It is worthy to mention that the zigzag course of the curves is due to the expression of the solar radiation intensity given in (7). We notice also that the maximum arrival rate has less variations around the equator compared to the other latitudes along the year. Indeed, for high latitudes, the maximum arrival rate is less dynamic and is around the full capacity of the servers during different periods of the year. This observation is important because it shed light into the workload management along the year given the long-duration of HAP's missions.

4.2 Energy Saving

After checking the HAP ying condition and determining the maximum arrival rate for different days and latitudes, we explore the energy-saving capabilities of the data center-enabled HAP. This simulation evaluates the consumed energy for the maximum workload arrival rate relative to different latitudes and days. First, we study the consumed energy variation for both data center-enabled HAP and terrestrial data centers versus the latitude for day number 150 in the year, as depicted in Fig. 4. We observe rst that the consumed energy decreases with latitude because more workload can be accepted and hence covered by the HAP's harvested energy. We also notice that the data center-

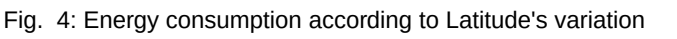

Fig. 4: Energy consumption according to Latitude's variation

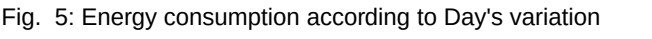

Fig. 5: Energy consumption according to Day's variation

enabled HAP helps to reduce the consumed energy for a large range of latitude r 60 ; 60 s. We assess the saved energy rate E ^{sav}p%q 14:61% achieved for the maximum workload arrival rate ^{max} when the latitude is around 60 for one HAP. Higher energy saving rates can be reached with a lower arrival rate ^{max}. We also note that if a larger payload can be supported by the HAP (increased by ten servers), higher energy saving can be recorded (increased by 3%). Indeed, a more important workload amount can be processed in the HAP in this case, which reduces the cooling energy consumption and yields higher energy savings.

Then, we study the consumed energy variation according to different days in the year, as shown in Fig. 5. The least

Fig. 6: Outage Probability of the Arriving Workload

Fig. 7: Saved Energy in the case of Re-transmission to HAP

consumed energy record is at the middle of the year for the data center-enabled HAP. This observation is because the beginning and the end of the year are characterized by cold weather (the Winter season), where the ambient temperature reaches its lowest levels. Hence, less solar energy can be harvested by the HAP. However, the days around 200^{th} day belong to Summer; where more solar energy can be collected within the data center-enabled HAP. We also notice that the data center-enabled HAP helps to reduce the consumed energy with an energy-saving rate $E^{sav}(\%) = 14.38\%$ achieved for the maximum workload arrival rate ^{max} around the 200^{th} of the year. These saving rates can be substantially enhanced with the deployment of more HAPs. For instance, the saving rate exceeds 17% starting from 2 HAPs deployment along with the terrestrial data center.

4.3 Outage Probability of Of oading

This simulation studies the outage probability of the communication links between terrestrial and HAP-based data centers. To this end, we plot the upper bound (31) and lower bound (33) outage probabilities respectively, versus the workload arrival rates, as illustrated in Fig. 6. We note that the partial dropping of some tasks might start at a workload arrival rate around $^{LB}_{partial} = 360$ task/s by considering the lower bound and at a workload arrival rate around $^{UB}_{partial} = 650$ task/s by considering the upper bound. Hence, $^{UB}_{drop^*} = 360$ task/s (c.f. (35)). The total dropping of the arriving tasks starts at a workload arrival rate around $^{LB}_{total} = 400$ task/s by considering the lower bound and at a workload arrival rate around $^{UB}_{total} = 730$ task/s by considering the upper bound. We also study the impact of the mean task length on the outage probability. As depicted in Fig. 6, we notice that the outage probability is more important for a higher task length because of more data (b in bits (29)) is carried on the transmission link for the same workload arrival rate.

Given the outage probability and the saved energy findings, we can investigate the tasks' retransmission impacts on the saved energy to the HAP. Therefore, we study the saved energy variation according to the workload arrival rate in two cases. In the first case, the dropped workload is processed in the terrestrial data center, while in the second case, we consider that the dropped workload is re-transmitted to the HAP. As shown in Fig. 7, if the dropped workload is not re-transmitted to the HAP, the saved energy percentage decreases notably with high arrival rates. Interestingly, the saved energy through a data center-enabled HAP is around 9% even when the workload is partially processed in the flying data center. However, under the re-transmission assumption, the saved energy becomes around 11.76%, which achieves almost the same performance in the case of zero outage. This observation highlights that the saved energy is not impacted even for high data rates because the transmission energy is significantly lower than the cooling energy.

4.4 Delay in Data Center-enabled HAP

In this simulation, we study the mean waiting time (38) in our system. Therefore, we approximate the general distribution of the considered queuing model to an exponential service time distribution with a mean service rate $_s$ and an exponential vacation time distribution with a mean vacation rate $_v$. Accordingly, the first moment of the residual service time can be simplified, based on (37), as follows:

$$\overline{R_i} \quad \frac{_i}{2_s} \quad \frac{(1 \quad u_i)}{_v}. \tag{39}$$

By using this approximation, we compare the mean waiting time obtained through the analysis in section 3.4 to the simulation results of an M/M/1 queue model with vacations where the service rate is $_s$ and the vacation rate is $_v$. Therefore, we study the mean waiting time according to the variation of the workload arrival rate in both cases. As depicted in Fig. 8, the analytical results coincide with the simulation results for different workload arrival rates. Indeed, the difference between both curves $|\overline{W_i}^{analysis} \quad \overline{W_i}^{simulation}| < 10^{-6}$ as zoomed out for the values of workload arrival rates around 700 task/s. We notice also that higher delays are attained when the workload arrival

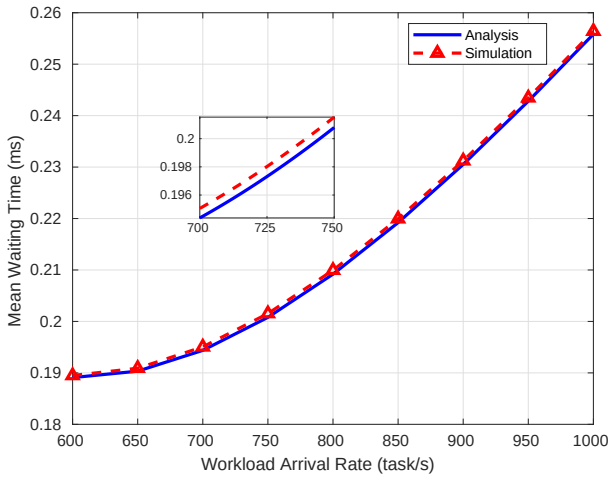

Fig. 8: Mean Waiting Time Variation according to the Arriving Workload

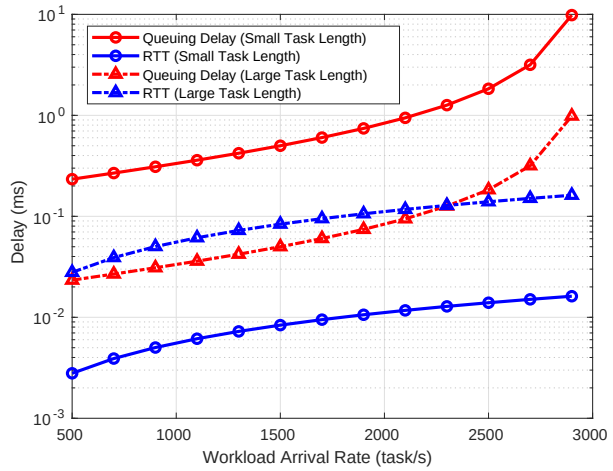

Fig. 9: Comparison between the Queuing Delay and the Transmission Delay

rate increases because more tasks should be waiting before their processing.

Therefore, we study the experienced delay in a data center-enabled HAP according to the variation of the workload arrival rate. As depicted in Fig. 9, we notice that this comparison depends on the mean task length of the arriving workload. For instance, the queuing delay is notably more significant than RTT for different arrival rates in the case of small task length. These results indicate that the tasks' transmission and the re-transmission to the HAP can occur without affecting the experienced delay for a workload characterized by a small task length. However, the queuing delay is sometimes lower than RTT for a given range of arrival rates in the case of large task length. These results indicate that the tasks' transmission and re-transmission to the HAP should be carefully studied for a workload characterized with a large task length because it depends on the arrival rate. For example, the queuing delay is close to RTT for an arrival rate 500 task/s ¤ ¤ 2250 task/s as depicted in Fig. 9. Therefore, the tasks' transmission to the HAP will add an additional and substantial delay to the queuing delay in this case. But, the queuing delay becomes higher than RTT for an arrival rate 2250task/s ¤ ¤ 3000task/s as depicted in Fig. 9. We note that the difference is not important, though, and would limit the number of re-transmitted tasks in the case of an outage.

5 Conclusion

Throughout this study, we have explored the potential of using a data center-enabled HAP system as a green alternative to traditional terrestrial data centers. Our analysis shows that the naturally low temperature and solar power available in the stratosphere make the HAP an ideal environment for flying data centers. We have established the limits of our system in terms of payload capacity and server utilization ratio to determine the optimal workload arrival rate to the HAP. Our study also demonstrates the energy-saving benefits of using a data center-enabled HAP compared to traditional terrestrial data centers under various workload settings, locations, and periods of the year. We have found that energy savings can be further improved by reducing the workload arrival rate to the HAP or increasing the number of servers hosted in the HAP. To increase system reliability and save more energy, reliable communication links are necessary to re-transmit dropped workloads to the HAP. Additionally, we have shown that the distribution of servers between terrestrial data centers and the data center-enabled HAP has no significant impact on the transmission and re-transmission delay for workloads with low task length. However, careful design of workload scheduling is essential to take full advantage of the benefits offered by the HAP system. Overall, our findings suggest that a data center-enabled HAP system presents a promising solution for reducing energy consumption and mitigating the environmental impact of traditional terrestrial data centers. Our study provides valuable insights into the potential benefits of this innovative computing paradigm and highlights the need for continued research in this area.

5.1 Research Challenges

However, it is important to explore the potential research challenges towards a comprehensive study. Therefore, the technical challenges of the data center enabled-HAP should be investigated. For instance, the unfriendly weather conditions in the stratosphere impose the deployment of adapted and more resilient electronic devices in the flying data center. Also, the frequency's maintenance of the airship and the flying servers is crucial to balance the trade-off between the quality of the offered computing service and the HAP's mission duration. Moreover, the economic viability of the data center enabled HAP should be considered. Capital expenditures such as the costs of the HAP platform and aerial servers, as well as operational expenditures like energy costs, must be assessed to evaluate profitability. Accordingly, a dynamic computing pricing model for the HAP is needed. Prices should adjust based on demand and network conditions to provide a satisfactory quality of experience for users while maximizing utilization of HAP resources and ensuring profitability [44]. Specifically, lower prices can be initially offered to encourage offloading to

the HAP during off-peak periods. However, as the arrival rate approaches the HAP's capacity, prices should increase to throttle demand or incentivize offloading some traffic to terrestrial data centers, which can offer lower prices during peak periods.

5.2 Future Directions

To tackle these challenges and improve the performance of the data center enabled-HAP, our future work will be based on different approaches of machine learning. For instance, meta learning can be useful to optimize the workload/network management policies in the flying data center hosted in the HAP. One way to boost the self-organization of the data center-enabled HAP network is to adopt meta-learning. Meta learning can be used in the data center-enabled HAP to learn from the output of the machine learning algorithms commonly used to optimize the network management policies in the terrestrial data center. Then, the generated models can be applied in the flying data center hosted in the HAP. These advantages are particularly valuable in the HAP because the learning process is accelerated and hence more computational energy is saved. Moreover, it is valuable to predict the highly-dynamic workload arriving at the data center by using federated learning. Federated learning can play a crucial role in predicting the workload incoming to the data center in the briefest delays without violating data privacy and less exchanging plain text data to ensure security. Indeed, federated learning can be applied within the different servers of one data center, geographically distributed sites of the same data center, or even within other data centers by only sending the prediction results and models. Accordingly, the predicted workload should be effectively scheduled spatially and temporally to leverage the renewable energy. Moreover, the renewable energy usage can be alternated with the fossil energy usage in a data center enabled-HAP when the appropriate workload amount is offloaded to the HAP; while fulfilling the QoS requirements and respecting the physical capabilities of the data center enabled-HAP.

REFERENCES

[1] N. S. Hogade, S. Pasricha, and H. J. Siegel, "Energy and network aware workload management for geographically distributed data centers," *IEEE Transactions on Sustainable Computing*, 2021.

[2] Gartner. Gartner forecasts worldwide it spending to grow 5.5% in 2023. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2023-04-06-gartner-forecasts-worldwide-it-spending-to-grow-5-percent-in-2023

[3] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2015.

[4] K. Kaur, S. Garg, G. Kaddoum, E. Bou-Harb, and K.-K. R. Choo, "A big data-enabled consolidated framework for energy efficient software defined data centers in iot setups," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2687–2697, 2020.

[5] M. Aslani, O. Amin, F. Nawab, and B. Shihada, "Rethinking blockchain integration with the industrial internet of things," *IEEE Internet Things Mag.*, vol. 3, no. 4, pp. 70–75, Dec. 2020.

[6] W. Abderrahim, O. Amin, M.-S. Alouini, and B. Shihada, "Proactive traffic offloading in dynamic integrated multisatellite terrestrial networks," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4671–4686, July 2022.

[7] V. Dukic, G. Khanna, C. Gkantsidis, T. Karagiannis, F. Parmigiani, A. Singla, M. Filer, J. L. Cox, A. Ptasznik, N. Harland *et al.*, "Beyond the mega-data center: networking multi-data center regions," in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, SIGCOMM'20*, 2020, pp. 765–781.

[8] J. Wan, J. Zhou, and X. Gui, "Sustainability analysis of green data centers with cchp and waste heat reuse systems," *IEEE Transactions on Sustainable Computing*, vol. 6, no. 1, pp. 155–167, 2020.

[9] Y. Berezovskaya, C.-W. Yang, A. Mousavi, V. Vyatkin, and T. B. Minde, "Modular model of a data centre as a tool for improving its energy efficiency," *IEEE Access*, vol. 8, pp. 46 559–46 573, 2020.

[10] X. Li, P. Garraghan, X. Jiang, Z. Wu, and J. Xu, "Holistic virtual machine scheduling in cloud datacenters towards minimizing total energy," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 6, pp. 1317–1331, 2018.

[11] X. Hu, P. Li, K. Wang, Y. Sun, D. Zeng, X. Wang, and S. Guo, "Joint workload scheduling and energy management for green data centers powered by fuel cells," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 2, pp. 397–406, 2019.

[12] S. Albers, "On energy conservation in data centers," *ACM Transactions on Parallel Computing (TOPC)*, vol. 6, no. 3, pp. 1–26, 2019.

[13] Thales. China energy transition status report 2021. [Online]. Available: http://english.www.gov.cn/statecouncil/ministries/202112/09/content_WS61b13edac6d09c94e48a1f81.html

[14] Gartner. Gartner predicts 75% of organizations will have implemented a data center infrastructure sustainability program by 2027. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2023-05-02-gartner-predicts-75-percent-of-organizations-will-have-implemented-a-data-center-infrastructure-sustainability-program-by-2027

[15] L. Ismail and H. Materwala, "Computing server power modeling in a data center: Survey, taxonomy, and performance evaluation," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[16] C. S. Chan, A. S. Akyürek, B. Aksanli, and T. Š. Rosing, "Optimal performance-aware cooling on enterprise servers," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 9, pp. 1689–1702, 2018.

[17] Z. Zhou, F. Liu, R. Zou, J. Liu, H. Xu, and H. Jin, "Carbon-aware online control of geo-distributed cloud services," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 9, pp. 2506–2519, 2015.

[18] N. Hogade, S. Pasricha, H. J. Siegel, A. A. Maciejewski, M. A. Oxley, and E. Jonardi, "Minimizing energy costs for geographically distributed heterogeneous data centers," *IEEE Transactions on Sustainable Computing*, vol. 3, no. 4, pp. 318–331, 2018.

[19] A. Wierman, L. L. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems," in *IEEE INFOCOM 2009*. IEEE, 2009, pp. 2007–2015.

[20] Z. Liu, M. Lin, A. Wierman, S. Low, and L. L. Andrew, "Greening geographical load balancing," *IEEE/ACM Transactions on Networking*, vol. 23, no. 2, pp. 657–671, 2014.

[21] F. Yao, J. Wu, G. Venkataramani, and S. Subramaniam, "Tsbatpro: Improving energy efficiency in data centers by leveraging temporal–spatial batching," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 1, pp. 236–249, 2019.

[22] A. Agarwal, J. Sun, S. Noghabi, S. Iyengar, A. Badam, R. Chandra, S. Seshan, and S. Kalyanaraman, "Redesigning data centers for renewable energy," in *Proceedings of the Twentieth ACM Workshop on Hot Topics in Networks,HotNets'21*, 2021, pp. 45–52.

[23] H. Yuan, H. Liu, J. Bi, and M. Zhou, "Revenue and energy cost-optimized biobjective task scheduling for green cloud data centers," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 817–830, 2020.

[24] W. Abderrahim, O. Amin, and B. Shihada, "How to leverage high altitude platforms in green computing?" *IEEE Communications Magazine*, vol. 61, no. 7, pp. 134–140, 2023.

[25] K. Mershad, H. Dahrouj, H. Sarieddeen, B. Shihada, T. Al-Naffouri, and M.-S. Alouini, "Cloud-enabled high-altitude platform systems: Challenges and opportunities," *Frontiers in Communications and Networks*, vol. 2, p. 30, 2021.

[26] G. K. Kurt, M. G. Khoshkholgh, S. Alfattani, A. Ibrahim, T. S. Darwish, M. S. Alam, H. Yanikomeroglu, and A. Yongacoglu, "A vision and framework for the high altitude platform station (HAPS) networks of the future," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 729–779, 2021.