GreenContinuum: A Formal Model of a Smart Grid-Aware Edge-Cloud Continuum for Carbon and Energy Management

Rohail Gulbaz
Department of Computing
Umeå University,
Umeå, Sweden
rohail.gulbaz@umu.se

Paul Townend
Department of Computing
Umeå University,
Umeå, Sweden
paul.townend@umu.se

Per-Olov Östberg
Department of Computing
Umeå University,
Umeå, Sweden
per-olov.ostberg@umu.se

***Abstract*—The Edge-Cloud Continuum is a large-scale, loosely coupled system consisting of multiple stakeholders, regions, dynamic infrastructures, and conflicting objectives. With surging growth and demand, the Continuum's energy and carbon footprint have massively increased, resulting in great operational expense, environmental impact, and strain on power grids. Methods to mitigate this face significant challenges: Quality of Service (QoS) guarantees must be balanced against not only carbon emissions, but the loadings, capacities, and QoS of the (smart) grids that power the underlying infrastructure. Integrated models to enable reasoning across both a Continuum and its associated Smart Grids are therefore required.**

This work presents a *formal model to reason across the integration of Smart Grids and the Edge-Cloud Continuum*. Firstly, we identify the components, interactions, and properties crucial to mitigating cross-Continuum energy and carbon footprint while maintaining user, provider, and power grid QoS. We then present associated mathematical models to enable a model-based simulation to be developed based on our work. We present this simulation (*all code is available for download*) and use a simple scheduling algorithm to demonstrate the feasibility of utilizing knowledge from both the Smart Grid and Edge-Cloud Continuum for carbon and energy management, showing that significant savings are possible.

Keywords—Modelling, Simulation, Edge-Cloud Continuum, Smart Grid, Energy Consumption, Carbon Footprint

I. INTRODUCTION

Cloud Computing has evolved from a centralized service model to a federated model encompassing complex technologies (e.g., 5G/6G cellular networks), diverse services, federation strategies, hybrid infrastructures, and wide-area application orchestration. This federated model has resulted in the combination of Fog, Edge, and Mist Computing, under the concept of the Edge-Cloud Continuum [1-3], regarded by the European Commission as a driving force for regional digitalization [4]. In comparison to traditional Edge architectures, the Continuum emphasizes seamless orchestration of geographically distributed heterogeneous resources, infrastructure providers, and conflicting constraints [5] at the expense of added management complexity.

Data centers (**DCs**) are projected to consume 10% of global energy by 2030 [6], whilst demand and complexity are further expected to increase to support over 50 billion Internet-of-Things (IoT) devices [7]. Added to this, a reliance on machine learning-based training and inference is causing DCs to draw massive electrical power, with approximately 40-50% of a DC's operational costs attributed to electricity bills [8]. This has profound consequences financially and in terms of environmental impact. Many DCs have onsite power generation facilities, but still typically buy large amounts of grid power. Moreover, green energy sources are intermittent; consequently, approximately 80% of the world's energy demand, including by DCs, is met by environmentally unfriendly brown (non-renewable) power sources [9]. A Continuum may have several service providers, i.e., Edge-Cloud Providers (**ECPs**), where each provider can have a network of locally federated DCs and computational edge devices. Furthermore, a Smart Grid, providing energy to ECPs, can have saturated loads in some regions, different Carbon Intensity (**CI**) across regions [9], etc.

Given the cost and impact of this energy consumption, there is a clear need for ECPs to reduce energy consumption and carbon footprint. From a hardware perspective, this is typically approached by the development of more efficient servers, cooling infrastructure, etc. From a software perspective, this can include methods to reduce data transmission costs, manage containerized and virtualized resources for maximum resource utilization, process data closer to users to reduce latency-induced overheads, distribute workloads for better load balancing, and distribute compute load based on CI and green energy availability in a particular region to maximize green energy utilization [2, 10].

This latter goal shows high potential – but effective solutions require knowledge of power grid infrastructures, utilizations, costs, and carbon intensities; *this is typically unavailable* to compute resource management systems operating on traditional power grids. However, *Smart Grid systems offer new opportunities* – providing decentralized power generation and real-time monitoring of utilization and carbon intensity. Continuum systems connected with a Smart Grid can make decisions, such as purchasing power when prices are lower, running workloads in regions with less CI, suspending workloads when CI is high and resuming them when renewable energy becomes available, splitting workloads to reduce the load on a region with high CI, and offloading or migrating load among other federated DCs [11].

Many such temporal and spatial decisions can be considered by ECPs alongside regular QoS constraints such as latency, makespan, resource utilization, load balancing, response time, etc. These conflicting objectives can be treated as a multi-objective optimization problem to meet QoS and sustainability goals defined in Service Level Objectives (**SLOs**) of stakeholders of ECPs and Energy Providers (**EPs**).

Although much work (including formal models) in traditional Cloud resource management addresses energy and carbon efficiency, *there is a notable gap in the literature* for models that integrate energy infrastructure with the Edge-Cloud Continuum [2]. This is essential for reasoning, simulation, and management of combined Edge-Cloud resources and power grids. *This work aims to create a foundation for solutions to integrate Smart Grids for carbon and energy management, with three key contributions*:

1. **Theoretical Model:** We present a theoretical model to capture the components, properties, and interactions of a Continuum and its supporting energy infrastructures.
2. **Mathematical Models:** We present mathematical models that show the way to incorporate interactions and regional concepts highlighted in the theoretical model.
3. **Demonstration:** We demonstrate the use of our formal model by developing a model-based simulation on a realistic multi-stakeholder FaaS application scenario. We use a demonstrator (CPriority) to highlight the significance of our model. *All code is available, with a link given in Section V.*

This work is structured as follows: Section II elaborates on research gaps in state-of-the-art models. Section III presents a theoretical model. Section IV provides a mathematical view of the model. Section V demonstrates the use of our model. Section VI validates the significance of our model. Section VII outlines a summary and future directions.

II. RELATED WORK & RESEARCH GAP

A. Formal Model Definition

A formal model aiming to reduce carbon emissions defines the components, properties, and interactions of the Smart Grid and Continuum [12]. A formal model-based solution has more potential, as it is validated through rationales and simulations, e.g., [12, 13]. Smart Grid and Continuum models can be integrated or non-integrated, encompassing formal models and solution-oriented approaches such as schedulers.

B. Non-integrated Continuum Models

In [14], the benefit of Continuum is emphasized as the application execution time is less than in Cloud-only scenarios. [15] presents Continuum as a solution for the limitation of resources and also describes mathematical models to compute energy consumption at different layers of the Continuum. However, the mathematical models lack the interactions with the smart grid. An approach in [16] presents the idea of network zones where network topology connects Fog DCs in distributed Geographical Regions (**GRs**). From [17-18], it appears that the Continuum is a very complex environment with different settings, providers, and Dynamic Infrastructures. Apart from the Smart Grid characteristics, all models [14-18] have omitted the decentralized control and parallel workloads in the Continuum.

C. Non-integrated Smart Grid Models

According to [19], the unpredictability of renewable energy production causes frequency and voltage imbalances, which require models to balance power supply and demand. The framework [20] reasons that proper modelling, having consideration like locating the Power Distribution Network (**PDN**) near the source of generation, adds flexibility in the system despite the challenge of integrating new power sources. [21] exploits game theory for the demand-side energy management and also suggests a cost-effective distributed energy storage plan. These models [19-21] have primarily focused on supply and demand balance, but not all models have collectively considered multiple EPs, import/export of power, intermittent energy production, battery storage, and the dynamic addition of power sources, etc.

D. Integrated Smart Grid & Continuum Models

1) Subset of Continuum Layers: The authors in [22] explain that grid energy CI varies across regions with fewer intra-regional hourly fluctuations. The carbon-aware approach in [23] dynamically adjusts server allocation to batch jobs based on carbon cost, exploiting the resource elasticity of workloads, while CASPER [24] optimizes latency and CI through load balancing for applications hosted in multiple regions with varying CI. In [25], a Smart Grid-aware scheduler for DCs makes job placement decisions. The first strategy considers smart meters for energy, and the second approach predicts future energy prices based on the weather forecast of wind and solar. In these models [22-25], either renewable energy or CI is taken to represent the Smart Grid information for the subset of Continuum layers; however, multiple EPs, multiple ECPs, parallel workloads, and decentralized control are not properly modeled.

2) Formal Models: In [12], a high-level formal model to outline the need for resource management in the Continuum with an interface with energy systems is proposed. The model is not demonstrated and lacks multiple ECPs. A formal model of integrated Smart Grid and Edge Federations [13] shows that integration can significantly improve green energy consumption, but does not cover multiple ECPs, multiple EPs, GRs, parallel workloads, and decentralized control.

3) Solution-oriented Approaches: The authors in [26] present a controller for container management to reduce carbon footprint and interference from collocated workloads, whereby maximum utilization of renewable energy sources results in reduced carbon emissions. However, multiple ECPs, multiple EPs, decentralized control, and infrastructural dynamicity have been omitted. In [27], a multi-cloud model to improve resource utilization balances the distribution of residual energy by offloading tasks to energy-rich nodes. It ignores multiple EPs, GRs, parallel workloads, decentralized control, and infrastructural dynamicity.

Although some integrated formal models and solutions *(excluding papers focusing on a few layers of the Continuum [22-25])* exist, their architectures significantly omit interdependent necessary characteristics, as shown in Table I.

TABLE I. RESEARCH GAP ANALYSIS

Integrated Models	Key Characteristics						Demonstration
	Multiple Edge-Cloud Providers	*Multiple Energy Providers*	*Geographical Regions*	*Parallel Workloads*	*Decentralized Control*	*Infrastructural Dynamicity*	
[12]	✕	✓	✓	✓	✓	✓	✕
[13]	✕	✕	✕	✕	✕	✓	✓
[26]	✕	✕	✓	✓	✕	✕	✓
[27]	✓	✕	✕	✕	✕	✕	✓
Proposed Formal Model	✓	✓	✓	✓	✓	✓	✓

III. THEORETICAL MODELLING

A. Integrated Continuum Modelling

The Continuum involves horizontally and vertically federated DCs and device layers. Fig. 1 presents a scenario of six DCs and devices (*devices owned by end-users are not depicted in Fig. 1*) located across four types of regions (ECP, Workload, GR, and EP). These overlapping regions can be seen as *Sets* in a set theory terminology, and optimization insights based on the objectives of different stakeholders can be derived from these sets.

There can be multiple ECPs (e.g., AWS, Google, Alibaba). Furthermore, service consumers may have their own on-premises DC (usually less powerful). Processing of users' tasks can begin anywhere from on-device to DCs such as on-premise, far edge, near edge, and cloud, or a combination thereof. Service providers, through various SLOs with other providers and users, formulate loosely coupled federations, enabling massive scalability and flexible decision-making. Fig. 1 shows the ECP regions E = {ECP_1, ECP_2}, where ECP_1 = {DC_1, DC_2, DC_3, DC_4, d_1, d_2, d_3, d_4, d_5}, ECP_2 = {DC_5, DC_6, d_6, d_7} and DC_5 belonging to ECP_2 is shared with ECP_1. If one ECP's DCs are not able to meet some QoS, like latency, energy demand, etc., then the federated DC can be utilized in two different ways to ensure decentralized control:

- Local Controller (LC): In this type, by looking at the accessible logs about resource availability, the LC of an ECP can request another LC to handle some of their jobs.
- Global Controller (GC): In this type, an LC can itself take job offloading decisions, acting as a GC, rather than handing the job to another LC.

Fig. 1 also shows multiple workload or application regions W = {W_1, W_2, W_3} of collaborating ECPs, facilitated by multiple DCs, represented as W_1 = {DC_1}, W_2 = {DC_1, DC_2, DC_3, DC_4, DC_5, d_1, d_2, d_3, d_4, d_5, d_6}, and W_3 = {DC_5, DC_6, d_6, d_7}, where {W_1, W_2} ∈ ECP_1 and W_3 ∈ ECP_2. DC_1 and DC_5 have two parallel workloads, and DC_5 also facilitates the workload of another ECP. For a realistic Continuum representation, it is necessary to consider workload regions having conflicting QoS, like a response time requiring processing in the nearest DC/device, but resource limitations not allowing additional load in that DC/device. A Continuum system has infrastructural dynamicity because of resource limitations, mobility (like an autonomous vehicle), whilst users of an application also vary over time.

From Fig. 1, we can also see that the Continuum spans multiple geographical boundaries, representing cities, states, countries, Cloud/Fog/Edge/device layers, data regulations, security-defined regions, etc. There are two GRs G = {G_1, G_2}, where G_1 = {DC_1, DC_2, DC_3, DC_5, d_1, d_2, d_3, d_6} and G_2 = {DC_4, DC_6, d_4, d_5, d_7}. If the CI of $G_2 > G_1$ at a certain time, then a job of W_3 belonging to [G_1, G_2] can be placed in G_1.

Fig. 1 also shows that there can be multiple Power Sources, Power Distribution Networks (PDN), and EPs. A Power Source Region (PSR) can supply green, brown or power-mix depending on agreements (if any) with ECPs and availability. Furthermore, there can be more than one EP having its own PDN (often PDN and EP are used interchangeably). A PDN can have many sources of energy, and a provider can have multiple PDNs. Overall, energy-related decision-making is complex due to the versatility of providers, sources, intermittent production behaviors,

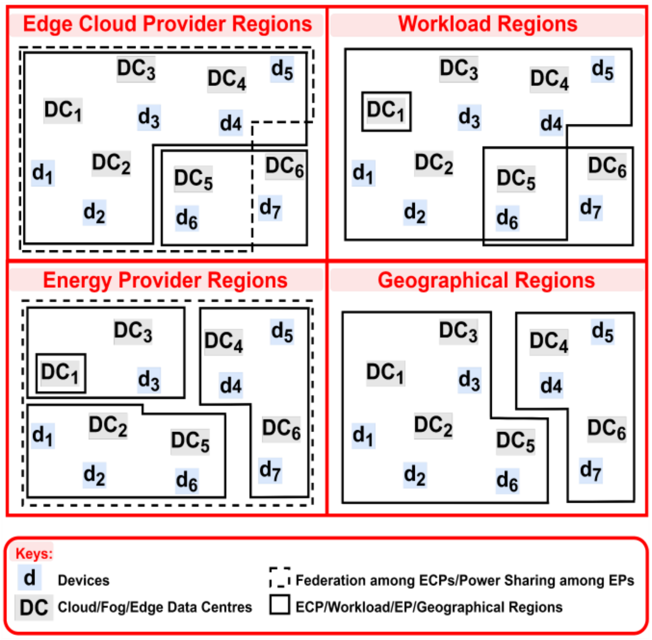

Fig. 1. Sample Regional Representation

agreements, load distribution, balancing, costs, CI, and energy factors conflicting with other QoS. In the figure there are four EP regions EP = {EP_1, EP_2, EP_3, EP_4}, where EP_1 = {DC_1} i.e. onsite energy, EP_2 = {DC_1, DC_3, d_3}, EP_3 = {DC_2, DC_5, d_1, d_2, d_6}, and EP_4 = {DC_4, DC_6, d_4, d_5, d_7}. Furthermore, these EPs also import/export powers with other EPs.

B. Integrated Smart Grid Modelling

In Fig. 2, an integrated Smart Grid model is shown with three layers: power, logic, and distribution. The power layer includes various sources such as green, renewable (not all renewable sources are entirely green), brown, and nuclear power (low carbon). The production from green energy sources varies significantly depending on region, weather, season, and time. The users of the power can also be a source; therefore, new sources can be added or removed dynamically at any time. Smart Grid users with solar panels can either feed onsite surplus into the Smart Grid or feed all the harvested power directly into the Smart Grid, with adjustments made in their bills. PDNs also have dedicated green and brown sources. Although energy is produced as needed to balance the frequency, batteries are used to store the surplus.

The logic layer represents the decision-making phase of the Smart Grid. There can be multiple PDNs, e.g., a national grid having power infrastructure spread across multiple cities. Furthermore, there can be multiple power providers in the form of microgrids, the national grid, and agreements of import with other EPs. Communication among multiple PDNs can be addressed by the Meta-PDN. Green-only sources can come from green onsite surplus or dedicated green sources. Brown can only come from dedicated brown sources. Power-mix comes from dedicated brown, surplus battery storage and imported power from other regions. Surplus storage can also be used to export power to other regions.

The third layer is the distribution layer, representing a Continuum consisting of several DCs powered by multiple sources, e.g., onsite and grid. According to the requirement, source availability, and power distribution agreements (if any), power sources are utilized, where PDN is responsible for dynamically deciding the source to power a particular region.

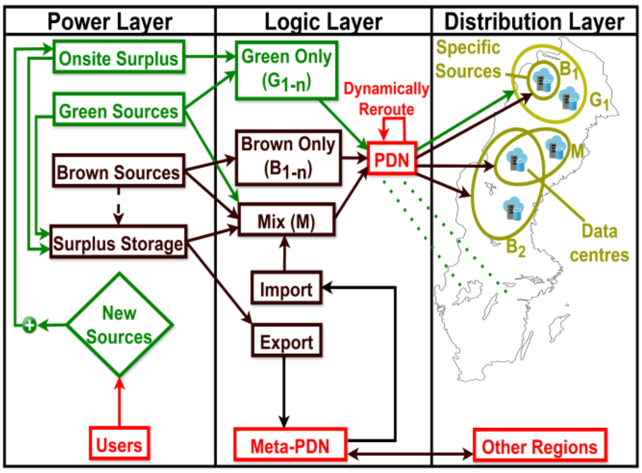

Fig. 2. Smart Grid Representation

C. Carbon vs Energy

Accessing maximum information about the Smart Grid for making various energy or carbon-aware decisions is ideal. However, relying solely on the percentage of Green or Renewable energy when both CI and Green Energy data are available can be misleading, as contradictory patterns can sometimes be observed. From Fig. 3, we can see that during the first week of January 2023, the Green Energy percentage of South Sweden is mostly better than South Central Sweden, as reported in [28]. South Sweden may appear 'greener', yet interestingly, South Central Sweden has lower CI, as shown in Fig. 4. CI depends on the sources, transmission loss, which requires more production from brown sources, demand, and the size of the region; due to these factors, a region with more green energy may not necessarily have lower CI.

IV. Mathematical Modelling

Existing computational models capture various aspects of energy in Clouds, but the nature of the Continuum requires novel mathematical models to drive the simulation and show its interaction with the Smart Grid. In the proposed mathematical models, container-level granularity, idle power, busy power, workload idle proportion (as workloads from different providers share the server's idle power), and CI are collectively incorporated more realistically as contributors to energy and other metrics. Million Instructions (MI) and Million Instructions per Second (MIPS) are used as an abstraction in the simulation for CPU-level system metrics.

A. Proposed Energy Computation Models

MS_i indicates the makespan of container 'i', which is the maximum or overall time taken by a set of jobs k.

$$MS_i = Max\ (T_k), \forall\ Jobs\ k\ \in Container\ i \tag{1}$$

The makespan of server 'j', denoted as MS_j, is the maximum makespan of containers that have run on it for the time under consideration.

$$MS_j = Max\ (MS_i), \forall\ Containers\ i \in Server\ j \tag{2}$$

The share ratio of container 'i', denoted as SR_i, defines the limit of processing power that a container can have. In the simulation, the limit ratio can be calculated by dividing the MIPS of a container by the MIPS of the server.

$$SR_i = \frac{MIPS_i}{MIPS_j}, Container\ i \in Server\ j \tag{3}$$

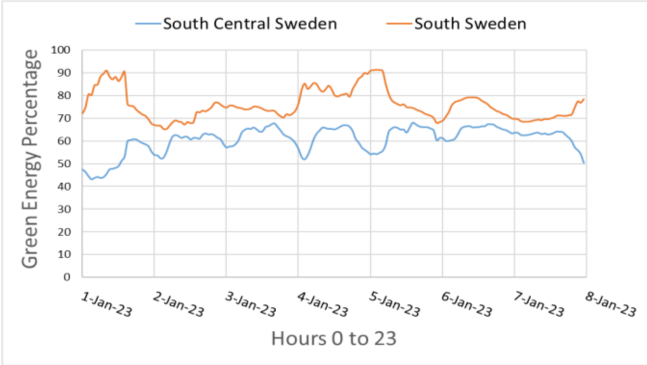

Fig. 3. Green Energy Comparison

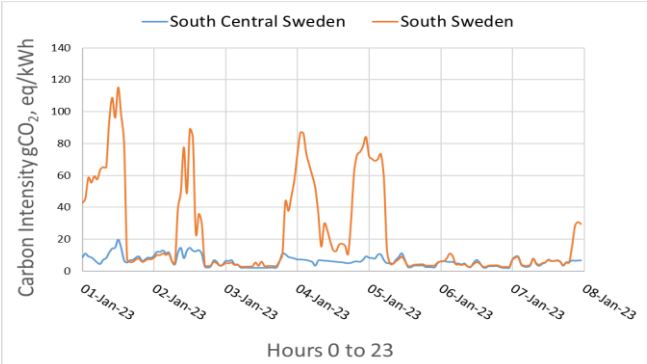

Fig. 4. Carbon Intensity Comparison

The energy consumption by container 'i', represented as EC_i, is the subset of busy power of server 'j' ($P_{busy_j} \times SR_i$) dependent on the share ratio and consumed for MS_i time.

$$EC_i = MS_i \times P_{busy_j} \times SR_i, \quad Container\ i \in Server\ j \tag{4}$$

The energy consumption by a server, denoted as EC_j, is based on both idle and busy power of the server. The P_{idle_j} power is always consumed, and busy is in addition to it. The summation of total idle power and energy consumed by all containers is the energy consumption by the server.

$$EC_j = MS_j \times P_{idle_j} + \sum_{i=1}^{n}(EC_i), \quad \forall\ Containers\ i \in Server\ j \tag{5}$$

The energy consumption by the DC is based on all servers 'j' that are part of that DC.

$$EC_{DC} = \sum_{j=1}^{n} EC_j, \forall\ Server\ j \in DC \tag{6}$$

The energy consumption by edge, fog, cloud, or geographical region depends on the summation of energy consumption by all DCs that belong to the region.

$$EC_{E/F/C/G} = \sum_{DC=1}^{n} EC_{DC}, \forall\ DC \in E/F/C/G \tag{7}$$

The Server 'j' Energy Consumption Full Busy portion, denoted as EC_{SFB_j}, is the maximum possible power drawn from the server regardless of the workload.

$$EC_{SFB_j} = MS_j \times P_{busy_j} \quad (8)$$

The energy consumption of the Server for which containers of the workload are responsible, denoted as EC_{busy_w}, depends on the energy consumed by all containers belonging to the workload running on the server 'j'. Server 'j' running the containers of the workload 'w' can be multiple.

$$EC_{busy_w} = \sum_{i=1}^{n} (EC_i), \quad (9)$$
$$\forall\, Containers\ i \in Workload\ w,$$
$$Server\ j \in S_w, where\ w\ is\ running\ on\ j$$

Share Ratio of Workload 'w', i.e. the percentage of energy consumed by containers of the workload out of the Full Possible Energy drawn from the Server, is given by SR_w.

$$SR_w = \frac{\sum_{i=1}^{n}(EC_i)}{\sum_{j=1}^{m}\left(EC_{SFB\,j}\right)}, \quad (10)$$
$$\forall\, Containers\ i \in Workload\ w$$

A workload 'w' is not responsible for all the idle power consumed by the set of Servers 'j' running the workload. Therefore, to calculate the Energy of the workload EC_W, we propose to assign a subset of idle consumption of servers in addition to busy power consumption by containers. We attribute an idle portion of the same ratio as of busy share, achieved through $MS_j \times P_{idle_j} \times SR_w$ for each server.

$$EC_W = \sum_{j=1}^{n}\left(MS_j \times P_{idle_j} \times SR_w\right) + EC_{busy_w}, \quad (11)$$
$$\forall\, Servers\ j\ where\ w\ is\ running$$

Energy consumption by an ECP does not depend on DCs because a DC may be running workloads owned by another ECP. Therefore, energy consumption is determined by the workloads belonging to the ECP.

$$EC_{ECP} = \sum_{w=1}^{n} EC_w, \forall\, Workload\ w \in ECP \quad (12)$$

B. Proposed Carbon Computation Models

Carbon Emission based on the busy effort of container 'i', denoted by CE_{busy_i}, is the sum of the products of kWh EC_i in time interval 'k' and gCO$_2$-eq/kWh CI in every interval 'k' of regions set 'r' where the containers are deployed.

$$CE_{busy_i} = \sum_{k=1}^{n} EC_{i_k} \times CI_{k_r} \quad (13)$$

Average Carbon Emission based on the busy effort of container 'i', denoted by $Avg\_CE_{busy_i}$, takes the average of CI in region 'r' during the intervals set 'k'. The CI average is helpful when CI readings do not exactly match with the container's life cycle 'k'.

$$Avg\_CE_{busy_i} = EC_i \times Avg\_CI_{k_r} \quad (14)$$

Carbon Emission Idle, denoted as CE_{idle}, is the product of the sum of the idle power of servers 'j' belonging to regions 'r', each interval time 'k', and CI during interval 'k' for which separate CI readings are available. Here, 'k' depends on the available CI data intervals, not on the container's lifetime.

$$CE_{idle} = \sum_{r=1}^{l}\sum_{k=1}^{m}\left(\sum_{j=1}^{n} P_{idle_j}\right) \times k \times CI_{k_r} \quad (15)$$

The total Carbon Emission (CE) is the sum of carbon emitted by containers 'i' and total idle carbon emissions.

$$CE = \left(\sum_{i=1}^{n} CE_{busy_i}\right) + CE_{idle} \quad (16)$$

C. Proposed Resource Utilization Models

The Host Resource Utilization Ratio is the energy consumed by the host out of the total possible consumption.

$$HRUR_j = \frac{\sum_{i=1}^{n}(EC_i)}{EC_{SFB\,j}}, \quad (17)$$
$$\forall\, Containers\ i \in Server\ j$$

Our proposed Average Resource Utilization Ratio is based on an average of HRURs, rather than time estimates.

$$ARUR = \frac{\sum_{j=1}^{n} HRUR_j}{n} \quad (18)$$

V. Simulation-based Validation

Now that we have a theoretical model and mathematical models showing interactions of Smart Grid and Continuum, we extend CloudSim [29] to use these. We validate the simulation with a real Wildfire Detection use case [5], using real smart grid data [28], system specifications [30], and Function as a Service (FaaS) traces [31]. In this use case, IoT fire sensors and cameras data are initially filtered at the edge layer, and then multi-provider ECP schedulers map FaaS invocations (jobs) to Continuum resources while considering the federation and Smart Grid. All code is available at: ***https://github.com/rohailgulbaz/SmartGridAware-ContinuumSim***

TABLE II. CPriority Settings in Regions

Settings	DC_{ID} - {$Host_{ID}$}	ECP_{ID}	EP_{ID}	GR_{ID}
CPriorityN	1 - {1, 2, 3}	1	1	1
	2 - {4, 5, 6}	2	1	1
	3 - {7, 8, 9}	1	3	2
	4 - {10, 11, 12}	2	3	2
	5 - {13, 14, 15}	1	4	2
	6 - {16, 17, 18}	2	4	2
CPriorityLessDCN	DC_3 and DC_5 are excluded.			
CPriorityLessHostN	Hosts 6 and 12 are excluded.			
CPriorityAustraliaN	First 3 DCs use EP_3 and remaining EP_4 in GR_2.			
CPriorityUSAN	First 3 DCs use EP_1 and remaining EP_2 in GR_1.			
CPriorityAustraliaM	Morning hours CI is used instead of Night.			

A. System Model

In our Continuum setup, DCs of different ECPs, powered by various EPs in their respective geographical regions, have heterogeneous resources. The infrastructure is dynamic, as resource availability, workload demand, and carbon intensity fluctuate over time. We simulate six (an arbitrary choice) different Continuum DCs, and three servers per DC to emphasize different settings (a Continuum may comprise thousands of servers, but scalability is not the focus here). We assume typical 2-CPU and 4-CPU servers in each DC with processing power [14000-17000] and [17000-20000] MIPS, respectively, based on the conservative estimate [30] from the z14 mainframe and M-series Azure VM. We assume full power in the range [300-750] watts, with 10-20% idle [32].

B. Smart Grid Integration

In our experiments, energy providers in the USA and Australia represent different (smart) power grids having diverse climates and energy sources. The USA includes the "Arizona Public Service Company" operating in Arizona, and "Western Area Power Administration – Desert Southwest Region" operating in California and surrounding areas. Australia has Victoria and New South Wales (NSW) state-level providers' data. Assume DCs are in California, Arizona, Victoria, and NSW, with their associated EPs as EP_1, EP_2, EP_3, and EP_4, respectively. Table II shows the DCs in each CPriority setting, including a list of hosts, ECPs owning DCs, EPs providing electricity, and the geographical regions of DCs. The workloads W_1 and W_2 belong to ECP_1 and ECP_2, respectively, but these can be placed in any federated DC.

The actual hourly CI values for 2024 (latest available) of EP regions (EP_1 to EP_4) are sourced from [28]. Many aspects of the Smart Grid, such as import/export, are inherently reflected in CI. Before placing a real-time job, a prediction is used to identify the lowest CI region. We observe an over 99% accuracy with the last hour CI value-based prediction mechanism used in CPriority for the given data.

C. Job Model

The Azure Functions real traces [31] from 31-Jan-2021 over two weeks are used in the simulation. We have considered two workloads, W_1 and W_2, originally identified by the hashes 73427… and 85479… (truncated), respectively. The job arrival time spans 0-16740s, covering 1000 function invocations. The job allocation decisions use dedicated channels of communication with a latency of 0.1s. Job sizes in MI are derived from duration x 19000 MIPS, based on M-series Azure VM [30]. To involve container lifetime considerations, application-level concurrency is set to 2, with jobs sharing CPU time. Each (function) container's processing limits (201-204 MIPS) are defined, and images are available in the registry. The power draw is computed proportionally using (3), and the cold start delay is 1s [33].

D. Validation with a CPriority Algorithm

We validate the significance of our formal model using a simple Carbon Priority algorithm (CPriority), which covers all six characteristics in Table I. We assume same CPriority logic across all job-receiving endpoints, enabling decentralized mapping via multiple ECPs accessing global resource availability information. It has been tested in six different settings (named with the prefix CPriority) over five scenarios:

Algorithm: CPriority
Input: $\mathcal{R}\sim$(set of regions), $\mathbf{DC_r}\sim$(DCs in region $\mathbf{r}$), $\mathbf{H}\sim$(set of hosts), $\mathbf{j}\sim$(incoming job)
Output: Job $\mathbf{j}$ assigned to FaaS container
Step 1. Wait for a job j to arrive
Step 2. $r^* \leftarrow \min_{r\in\mathcal{R}} CI(r)$
Step 3. $RA(d) = \frac{totalAvailableMIPS(d)\times 100}{totalMIPS(d)}, \quad \forall d \in DC_{r^*}$
Step 4. $d^* \leftarrow \max_{d\in DC_{r^*}} RA(d)$
Step 5. $h^* \leftarrow NextHost_{RR}(d^*)$
Step 6. If $\exists c \in h^*$, such that $S_j = Service(c)$ and $Load(c) < 2$
Step 7. Then $c \leftarrow j$
Step 8. Else $c' \leftarrow NewContainer(S_j)$ for S_j and $c' \leftarrow j$
Step 9. Return to Step 1

- **Different number of EPs**
(CPriorityN vs CPriorityAustraliaN)
- **Different EPs but same number of EPs**
(CPriorityUSAN vs CPriorityAustraliaN)
- **Different time zones**
(CPriorityAustraliaN vs CPriorityAustraliaM)
- **Different number of host machines**
(CPriorityN vs CPriorityLessHostN)
- **Different number of DCs**
(CPriorityN vs CPriorityLessDCN)

The EPs considered in each setting are highlighted in Table II. In CPriorityN, the CI is of night 'N' time, 1-Jan-2024 00:00-04:39 hours. CPriorityLessDCN has 2 fewer DCs, and CPriorityLessHostN has 2 fewer hosts than CPriorityN. CPriorityAustraliaM considers Australian EPs (EP_3 and EP_4) with morning 'M time CI values, 1-Jan-2024 06:00-10:39 hours. CPriorityUSAN considers only U.S EPs (EP_1 and EP_2).

CPriority selects the region 'r*' with the minimum previous hourly CI to map a function invocation (job). Next, it selects a DC with maximum resource availability (RA) based on available processing power. Later, it selects a host in a round-robin manner. If a suitable container 'c' does not exist in the host, a new container is instantiated from an image.

VI. RESULTS & DISCUSSION

Integrated models are complex but have many benefits, as evident by industrial initiatives aiming to make an integrated multi-provider Continuum. Our formal model is helpful in carbon reduction through a clearer representation of the Continuum, giving guidelines like using CI over relying on solely green energy availability, etc. Experiments involving five scenarios highlight the significance of our integrated formal model, demonstrating that even minor changes in Smart Grid or Continuum settings can substantially affect carbon emissions. This emphasizes the need to adopt our model, which effectively captures such dynamic behaviors.

A. Different Number of EPs

This experiment shows a contrast when EPs differ in count, as CPriorityN has one additional EP than CPriorityAustraliaN. From Fig. 5, although DCs have the same specifications, they are in different EP regions; therefore, container creation or function placement decisions do not map to the same EP and associated DCs. This leads to differences in Makespan, Average Completion Time, Container Count, Average Containers Makespan, Average Resource Utilization Ratio (ARUR), Energy Consumption, and Container Carbon Emissions, though Idle Carbon Emissions are very close, as are overall Carbon Emissions. Fig. 6 shows CPriorityN using two GRs. In CPriorityAustraliaN EPR_4 gets significantly more jobs than EPR_3 due to its lower CI, with ~4% of jobs shifting from NFR_1 to FR_1 for W_1, and ~2% from NFR_2 to FR_2 for W_2.

B. Different EPs but Same Number of EPs

Comparing CPriorityAustraliaN and CPriorityUSAN, where EPs differ, we find that makespan matches due to identical DC count and specifications. However, all other metrics vary, as shown in Fig. 5. Fig. 6 depicts that in CPriorityUSAN, EP region EPR_2 gets no jobs, indicating consistently high CI values. However, NFR_1 and FR_2 receive more jobs due to ECP_1 having DCs in the low CI EP region.

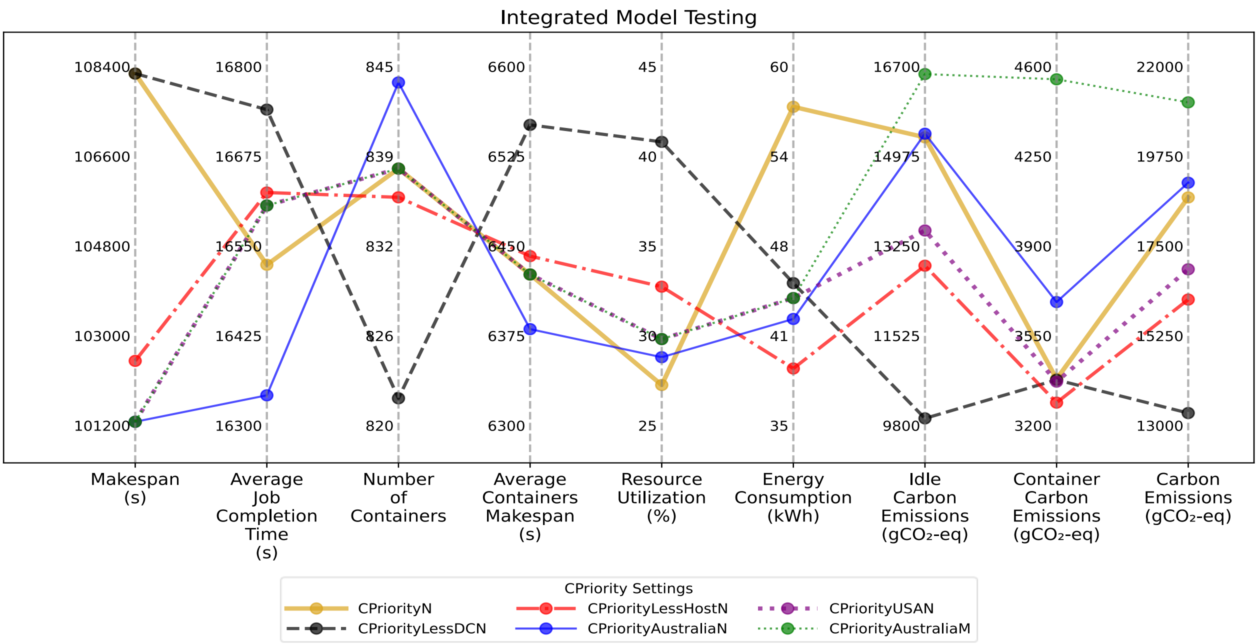

Fig. 5. Integrated Model Testing

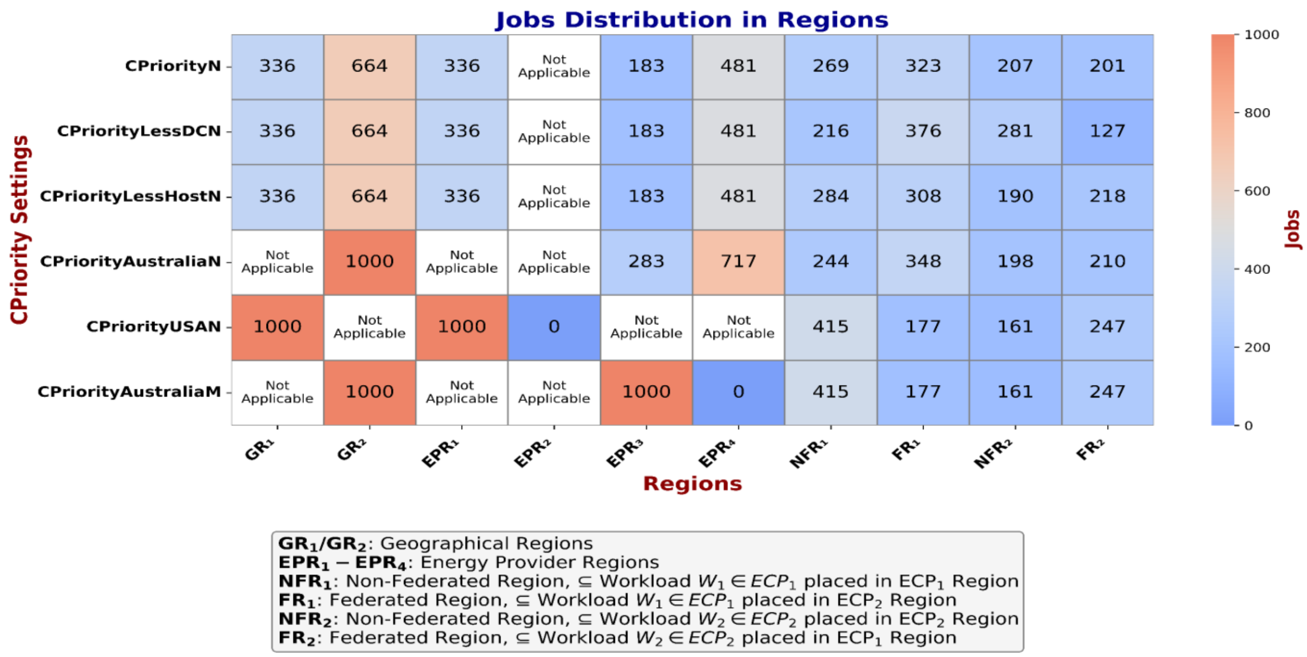

Fig. 6. Jobs Distribution in Regions

C. Different Time Zones

CPriorityAustraliaN and CPriorityAustraliaM differ only in time zone for function invocations, while other settings are the same. Fig. 5 shows that both have the same makespan, but varying CI leads to differences in performance metrics. Fig. 6 reveals that CPriorityAustraliaM assigns exactly no jobs to EP_4 due to its high morning CI, shifting the load to an ECP with more DCs in the low CI region.

D. Different Number of Host Machines

From Fig. 5, we observe that a minor change in the Continuum setting, running CPriorityN with fewer hosts, impacts all metrics. CPriorityLessHostN increases container life (handling multiple invocations), as containers were found warm in fewer hosts, leading to a higher average job completion time due to waiting. This improves resource utilization (RU) by ~5% and reduces carbon emissions by ~13%. This setup change also affects ECP region load, with ~2% of jobs shifting from FR_1 to NFR_1 for W_1 and ~4% from NFR_2 to FR_2 for W_2, as shown in Fig. 6.

E. Different Number of DCs

CPriorityLessDCN has two fewer DCs than CPriorityN for the same Smart Grid settings. From Fig. 5, we can see that fewer DCs lead to fewer container creations, similar to CPriorityLessHostN, resulting in ~13% better RU and a notable ~28% reduction in carbon emissions. In both settings, CI values are the same, but the resource availability pattern differs. Like CPriorityLessHostN, ECP regions' load is also affected, with ~19% of jobs shifting from NFR_1 to FR_1 for W_1 and ~18% from FR_2 to NFR_2 for W_2, as shown in Fig. 6.

VII. CONCLUSION & FUTURE WORK

The emergence of Continuum-based applications and their resulting high carbon emissions encourages stakeholders to take constructive measures to reduce carbon footprint; however, to reason about and mitigate carbon impact properly, it is essential to consider an integrated view of continuums and (smart) power grids.

We have designed a formal model that, for the first time, identifies the components and interactions of an Edge-Cloud Continuum running on Smart Grids, focusing on key characteristics not addressed collectively in the literature. We present a graphical view of the integrated Continuum and provide novel mathematical models showing the interactions required to transform the model into a CloudSim-based simulation, with all code made available.

To demonstrate the validity of the models and simulation, we utilize a real Wildfire Detection use case, and perform all experiments using real data from Smart Grids, system specifications, and FaaS traces. The observed fluctuations in carbon emissions across different experimental settings, such as varying EPs, highlight the need to adopt an integrated modelling approach that explicitly captures these dynamics.

In the future, powerful multi-objective optimization solutions for carbon and energy management can be designed based on this integrated modelling work about the integration of the Smart Grid and the Continuum. These solutions, such as intelligent schedulers, can proactively consider factors like smart grid utilization, capacity constraints, dynamic electricity pricing, carbon intensity, and equitable load shifting to reduce stampede burdens on specific regions and communities, among other considerations, thereby balancing energy efficiency with Quality of Service (QoS) requirements. Building on our model, tools can be developed to reason over complex interactions and behaviors, enhancing the interpretability of QoS metrics affecting carbon emissions.

ACKNOWLEDGMENTS

This work was supported by the European Commission through the Horizon Europe project SovereignEdge.Cognit under Grant Agreement 101092711 and by the Wallenberg AI, Autonomous Systems and Software Programme (WASP) through the project Tools for Autonomic Resource Management in the Cognitive Cloud Continuum (ARMC3).

REFERENCES

[1] S. Dustdar et al., "On Distributed Computing Continuum Systems," IEEE Trans. on Know. and Data Eng., vol. 35, no. 4, 2023.

[2] Y. S. Patel, P. Townend, A. Singh, and P-O. Östberg, "Modeling the Green Cloud Continuum: integrating energy considerations into Cloud–Edge models," Cluster Computing, vol. 27, no. 4, 2024.

[3] I. Stoica, "Sky computing: opportunities and challenges," Decision Making and Decision Support in the Information Era: Dedicated to Academician Florin Filip, pp. 15–27, 2024.

[4] F. Montevecchi et al., "Energy-efficient cloud computing technologies and policies for an eco-friendly cloud market," European Commission, final study report, Comms. Networks, Content and Tech., 2020.

[5] P. Townend et al., "Cognit: Challenges and vision for a serverless and multi-provider cognitive cloud-edge continuum" in Proc. of IEEE Int. Conf. Edge Computing and Communications (EDGE), 2023.

[6] E. Masanet et al., "Recalibrating global data center energy-use estimates," Science, vol. 367, no. 6481, 2020.

[7] N. Sami et al. "Future Internet of Things (IOT) from cloud perspective: aspects, applications and challenges," Internet of Things (IoT) Concepts and Applications, pp. 515–532, 2020.

[8] https://encoradvisors.com/data-center-cost/ (accessed on 15-Sep-2025)

[9] W. Li et al., "On enabling sustainable edge computing with renewable energy resources," IEEE Comms. magazine, vol. 56, no. 5, 2018.

[10] D. Maji et al., "The green mirage: Impact of location-and market-based carbon intensity estimation on carbon optimization efficacy," in Proc. 15th ACM Int. Conf. on Future and Sustainable Energy Systems, 2024.

[11] B. Kar et al., "Offloading using traditional optimization and machine learning in federated cloud–edge–fog systems: A survey," IEEE Communications Surveys & Tutorials, vol. 25, no. 2, 2023.

[12] Y. S. Patel et al., "Formal Models for the Energy-Aware Cloud-Edge Computing Continuum: Analysis and Challenges," in Proc. IEEE Int. Conf. on Service-Oriented System Engineering, 2023.

[13] R. Cárdenas, P. Arroba, J. L. Risco-Martín, and J. M. Moya, "Modeling and simulation of smart grid-aware edge computing federations," Cluster Computing, vol. 26, no. 1, 2023.

[14] L. Baresi et al., "A unified model for the mobile-edge-cloud continuum," ACM Trans. on Internet Tech. (TOIT), vol. 19, 2019.

[15] A. Mahapatra et al., "An energy-aware task offloading and load balancing for latency-sensitive IoT applications in the Fog-Cloud continuum," IEEE Access, vol. 12, 2024.

[16] J. Santos et al., "Diktyo: Network-aware scheduling in container-based clouds," IEEE Trans. on Network and Service Management, 2023.

[17] D. Rosendo et. al., "E2clab: Exploring the computing continuum through repeatable, replicable and reproducible edge-to-cloud experiments," in 2020 IEEE Int. Conf. on Cluster Computing, 2020.

[18] A. Mseddi, W. Jaafar, H. Elbiaze, and W. Ajib, "Joint container placement and task provisioning in dynamic fog computing," IEEE Internet of Things Journal, vol. 6, no. 6, 2019.

[19] G. Barone et al., "The role of energy communities in electricity grid balancing: A flexible tool for smart grid power distribution optimization," Renewable and sustainable energy reviews, 2023.

[20] R. Moreno et al., "Planning low-carbon electricity systems under uncertainty considering operational flexibility and smart grid technologies," Phil. Trans. of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 375, no. 2100, 2017.

[21] K. Wang et al., "Green energy scheduling for demand side management in the smart grid," IEEE Trans. on Green Comms. and Networking, 2:2, 2018.

[22] J. Thiede et al., "Carbon containers: A system-level facility for managing application-level carbon emissions," in Proceedings of the 2023 ACM Symposium on Cloud Computing, 2023.

[23] W. A. Hanafy et al., "Carbonscaler: Leveraging cloud workload elasticity for optimizing carbon-efficiency," Proc. of the ACM on Measurement and Analysis of Computing Systems, vol. 7, no. 3, 2023.

[24] A. Souza et al., "Casper: Carbon-aware scheduling and provisioning for distributed web services," in Proceedings of the 14th International Green and Sustainable Computing Conference, 2023.

[25] M. Mäsker et al., "Smart grid-aware scheduling in data centres," Computer Communications, vol. 96, 2016.

[26] K. Kaur et al., "KEIDS: Kubernetes-based energy and interference driven scheduler for industrial IoT in edge-cloud ecosystem," IEEE Internet of Things Journal, vol. 7, no. 5, 2019.

[27] J. Luo et al., "Container-based fog computing architecture and energy-balancing scheduling algorithm for energy IoT," Future Generation Computer Systems, vol. 97, 2019.

[28] https://app.electricitymaps.com/ (accessed 15-Sep-2025)

[29] R. N. Calheiros et al., "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Softw. Pract. Exper., 2011.

[30] https://learn.microsoft.com/en-us/azure/virtual-machines/workloads/mainframe-rehosting/concepts/mainframe-compute-azure (accessed on 15-Sep-2025)

[31] M. Shahrad et al., "Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider," in 2020 USENIX annual technical conference (USENIX ATC 20), 2020.

[32] https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf (accessed on 15-Sep-2025)

[33] https://www.openfaas.com/blog/fine-tuning-the-cold-start/ (accessed on 15-Sep-2025)