**QUEEN'S
UNIVERSITY
BELFAST**

Optimal and privacy-aware resource management in Artificial Intelligence of Things using osmotic computing

Sharma, V., Tan, T. G., Singh, S., & Sharma, P. K. (2022). Optimal and privacy-aware resource management in Artificial Intelligence of Things using osmotic computing. *IEEE Transactions on Industrial Informatics*, 18(5), 3377-3386. <https://doi.org/10.1109/TII.2021.3102471>

Published in:

IEEE Transactions on Industrial Informatics

Document Version:

Peer reviewed version

Queen's University Belfast - Research Portal:

[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights

Copyright 2021, IEEE.

This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Optimal and privacy-aware resource management in AIoT using osmotic computing

Vishal Sharma, Tan Teik Guan, Saurabh Singh, Pradip Kumar Sharma

Abstract—Critical infrastructure comprising on-demand devices, including secondary servers, comes into play when a situation like an overload is involved. The on-demand servers and devices require smart management solutions that form an integral part of Artificial Intelligence of Things (AIoT). This work considers AIoT as a combination of Mobile-Internet of Things (M-IoT) and AI requiring immediate response, secondary support system, and computational resources. Privacy in AIoT is always a concern when sharing information as intruders can eavesdrop on the settings of the system. This paper uses an osmotic computing paradigm, which enables the derivation of strategies to decide on the methods of sharing services via optimal and privacy-aware resource management in AIoT. A safety competition is built on top of configuration rewards that help to attain privacy-by-design. The contributions of this work are expressed using theoretical analysis and numerical simulations.

Index Terms—Osmotic computing, privacy-aware, resource management, Internet of Things, Mobility.

I. INTRODUCTION

Resource management ensures that the network is available to handle unscheduled or emergency tasks required for supporting excessive service requests, setup of rescue operations and even deploying on-demand nodes facilitated by edge computing [1], [2]. With the pressure of connecting billions of intelligent devices and having mobile terminals as a part of the Internet of Things (IoT), the intensity of migrating the services and assuring the safety of operations are critical [3]. Most recent works focus on service migration when an additional load is encountered in the network or the case of a change in the configuration of the network [4], [5]. These works utilize cost functions and build a priority system to handle incoming requests. This facilitation is covered as a part of resource management. The management of resources demands high-quality operations, better control, security and privacy which withholds all the information that may help an intruder to pick its target based on the service-requests [6], [7].

In the general model of resource management, the handling entities are fixed in terms of count and location and do not exhibit mobility. Such situations are comparatively easier to tackle than those with mobile terminals. Further with additional layers, the complexity of resource management

increases. Self-configuration and self-learning like solutions are a choice for these scenarios which bring the utilization of Machine Learning and Artificial Intelligence into the service migration and resource provisioning [8]. Artificial Intelligence Internet of Things (AIoT) offers a wide range of applications in robotics, surveillance, navigation, search and rescue, adversary detection in cyber-physical systems, and autonomous systems [9]. AI can facilitate security enhancements in IoT. Specifically, AI can accelerate security management by supporting quick training and learning of models [10]. The authors in [10] have highlighted several aspects of using AI for fast authentication in large-scale IoT. The authors explored progressive authorization and the use of classifiers for lightweight authentication.

Privacy is no longer limited to IoT devices, but a necessary consideration for resource management in large-scale IoT setups. Leaking information during allocation causes an immense loss to both the service providers as well as the infrastructure. Privacy becomes even a bigger issue when sensitive information such as medical records are involved as a part of IoT [11]. If the conditions of information are critical [12], [13], the sensitivity of the setup increases, which needs to ensure not only information but the parameters involved in the decision of allocations are protected. An attacker may take hold of the configuration and decide to attack the servers based on their handling capacity. It is of utmost importance that privacy is given due consideration along with resource management under mission-critical and AI-empowered setup of M-IoT devices. The network architecture must be such that it supports the privacy of entities as well as the information and can be dynamically adjusted. Thus, leading to the provision of privacy-by-design.

To overcome this challenge, this article utilizes an osmotic computing paradigm [14] that considers the infrastructure as well as information and builds strategies of resource management following privacy-by-design. The model uses *gap* rewards to decide on the privacy-aware characteristics and ensures that the configurations of the network are unique and prevents an attacker from deriving the details of the servers involved in the handling of the services. This prohibits the prediction of allocation of resources even if the configurations of the available servers are the same. Such a mechanism is attached by forming a safety competition driven by the service graph. The analytical evaluations and numerical study help to understand the effectiveness and efficiency of the proposed model.

V. Sharma is with the School of Electronics, Electrical Engineering and Computer Science (EEECs) at the Queen's University Belfast (QUB), Northern Ireland, United Kingdom.

T.T. Guan is with the ISTD Pillar, Singapore University of Technology and Design, Singapore.

S. Singh is with the Department of Industrial Systems Engineering, Dongguk University, South Korea.

P.K. Sharma is with the Cybersecurity department, University of Aberdeen, UK.

Fig. 1. An illustration of the network model with the utilization of reverse-osmotic computing [15].

A. Motivation, Problem Statement and Contributions

AIoT, comprising AI-empowered M-IoT, involves several mobile entities incorporating a decentralized setup offering intelligent, efficient, and reliable end-to-end transmissions in mission-critical scenarios [16], [17]. These networks often require a lot of reconfigurations and several load-balancing solutions, which must ensure that the resources are appropriately allocated. Here, resources are provided by the handling servers and the load refers to the requests generated from the end-users. However, along with decentralization, resource management requires better privacy as an advertisement of the handling servers can expose the network. An intruder can prevent the network from handling the load and deny services with a lot of unauthorized requests and spoofed messages. While we expect that AI can partially alleviate such attacks by detecting and denying false messages by managing information flow, it also exposes the underlying configurations and logic to other possible threats [18]. Such a situation can be corrected through privacy-aware resource management, which is the problem-statement of this work.

B. Our Contributions

The major contributions of this work are:

- An optimal and privacy-aware resource management approach is presented using the osmotic computing paradigm.
- The proposed model is built using the optimal control theory [19], which ensures the formation of the resource management providing privacy-by-design using the design principles in [20].
- The algorithms presented in this work help to identify legitimate requests and support privacy-aware service mapping using safety competition.
- Theoretical and numerical studies are used to evaluate the proposed concepts.

TABLE I
NOTATIONS AND MEANING

Notation	Meaning
$\mathcal{S}$	set of servers
$\mathcal{A}$	set of edge-based access points
$\mathcal{W}$	set of inter-zone handling switches
$\mathcal{U}$	set of users
$\mathcal{N}$	superset representing entire network
$\mathcal{H}$	set of mobile hubs
τ	timestamp
Ψ	optimal control function
$\mathbf{M}$	memory utilization rewards
$\mathbf{C}$	capacity rewards
$\mathbf{R}$	offered rate rewards
$\mathbf{E}$	energy utilization rewards
η	reward decision constants
β_i	available memory in i th timestamp
m_x	memory utilization of a single process
$\mathcal{X}$	number of the processes
C_i	request handling capacity of the network
ω	requests per second
$\mathcal{U}_l$	left-out users
P_R	number of pending requests
H_R	number of successfully handled requests
$\mathcal{R}_i$	rate of the network
λ	arrival rate of requests for a given number of processes
μ	size of requests for a given number of processes
$\mathcal{E}$	expected average available energy
ϵ	error rate for handling the given processes
$\mathcal{N}_a$	set of available entities
α_h	associated Priori for the handling network entities
α_p	associated Priori for pending to respond network entities
ϑ	reward balancing constant
$\mathbb{S}_C$	safety competition
δ	safety competition strategies
$\Gamma(t)$	gap (configuration rewards)

II. SYSTEM MODEL

The entire solution uses an osmotic computing paradigm, which involves the process of osmosis that helps balance the concentration. The concentration is the ratio of the solute to the solvent, which is mapped specifically to a computing

problem to fit in the requirements of the system and settle the given constraints. Table I lists different symbols used in the system model. In this work, the filtering of services is managed through reverse-osmosis that helps with the privacy-preservation. Reverse osmosis uses selection criteria built on the top of several governing equations to make sure that the privacy of the system is always preserved when the service requests are allocated to the servers. Different components used in the proposed model (see Figure 1) and as a part of the network infrastructure related to reverse-osmotic computing are expressed below [8], [21]:

- **Fixed-infrastructure as a solute:** The components of the network which cannot be shifted although modifications to their settings can be performed as termed as a solute of the network under osmotic computing. Devices that are managed with power, memory and fixed bands are considered this part. The solute can be used to have different optimization criteria. For example, having resources as solute and then identifying optimizing services (solvent) around the resources.
- **Services as a solvent:** Services including the user requests, processes, authorization and access policies, and deployment metrics are considered to be solvent. These control the optimization criteria and privacy aspect of a model.
- **Rules as membrane:** Criteria used to control the services are done with the help of certain rules, which are related to the membrane in osmosis. All the control algorithms are included in this part of the osmotic computing.
- **Load as concentration:** It defines the ratio involving load to the capacity, or load to the resources, etc. Concentration defines the configuration of the model and helps to ensure that the model/network is always in a balanced state and all the optimization criteria are fully met.

The network comprises a core regularly serving the users through a set of servers denoted by $\mathcal{S}$. These servers are the backbone of the network and handle the global requests coming from the end devices. These servers integrate public and private cloud setup for the M-IoT, which can be used with the majority of upcoming networks, such as 5G and beyond communications, where several functions can be deployed on these servers. These servers are connected to several edge-based access points denoted by a set $\mathcal{A}$. The intermediate layer between the edge and the core includes another set of inter-zone handling switches denoted by $\mathcal{W}$. The users in the network are the actual devices serving as M-IoT and are represented by a set $\mathcal{U}$. All the decisions to routing and resource management are either taken at the edge- or intermediate-layer. The entire network excluding the end-user is denoted by a superset $\mathcal{N}$. The mobility of the network is affected by another component, termed as a mobile hub, inspired by the network architecture used in [22] [15], which is denoted by a set $\mathcal{H}$, as shown in Figure 1. These mobile hubs connect the access points and the switches and ensure the connectivity between the switches and edge is always maintained to facilitate the “Network-as-a-Service (NaaS)” part of the model. The entire model is driven by the optimal

control theory [23], [24] which projects the NaaS and allows effective and privacy-aware resource management to ensure the regulated connectivity in the network even during the high number of failures or unavailability of the components. To understand the system, let τ be the timestamp when NaaS needs to be invoked and services need to be shifted across the edge or at the hub-switch architecture considering the availability and making sure that the privacy of the network remains intact. Let $\tau' (\leq \tau)$ be the timestamp until which the system is safe and aware of its state and can decide on a load of each node based on some set criteria to maximize the reward of an optimal control function for a node in the network ($\mathcal{N}$), defined as:

$$\Psi_{\tau'}(x_i) = \sup_{0 \leq \tau' \leq \tau} \mathbb{E}_{x_i}(\mathbf{M}, \mathbf{C}, \mathbf{R}, \mathbf{E}), \quad (1)$$

where x_i defines the associated non-failure rewards w.r.t. $\mathcal{N}$, $\mathbf{M}$, $\mathbf{C}$, $\mathbf{R}$ and $\mathbf{E}$ are the rewards associated with memory utilization, capacity, offered rate, and energy utilization, respectively. These metrics define the service-exchange as well as the access points and servers that are involved in the resource management. Let η_1 , η_2 , η_3 , and η_4 be the reward decision constants for memory utilization, capacity, offered rate, and energy utilization, respectively, such that $\eta_1 + \eta_2 + \eta_3 + \eta_4 \leq 1$, where lesser value is in case the system considers the residuals as a part of the rewards. These constants prohibit exchange of information related to the resources as an eavesdropper cannot account for actual settings from these decision constants. Following it, the rewards for memory are calculated as:

$$\mathbf{M} = \sum_{i=0}^{\tau'} \left(\eta_{1,i} \left(\beta_i - \sum_{j=0}^i (\mathcal{X} \times m_x) \right) \right), \quad (2)$$

where β_i is the available memory in the i th timestamp considered equivalent to a round, m_x is the memory utilization of a single process, $\mathcal{X}$ is the number of processes, which can be presented using a continuous function $f(\mathcal{X})$ considering that all involved entities are executing without a delay. For simplicity, $\int_{|\mathcal{N}|} f(\mathcal{X}) = 1$, which means each of the involved entity is performing at least a single computation in every round. However, with practical constraints, it is believed that $f(\mathcal{X})$ will be the arrival function operating with a Poisson Distribution ($f(\mathcal{X}, z) = \frac{\mathcal{X}^z \cdot e^{-\mathcal{X}}}{z!}$), where z is the servers per service in the network. Using capacity modelling,

$$\mathbf{C} = \sum_{i=0}^{\tau'} \left(\eta_{2,i} \left(\mathcal{C}_i - \frac{\omega_i}{|\mathcal{U}_{i,l}| \times |\mathcal{S}_i|} \left(\log \left(1 + \frac{H_{R,i}}{P_{R,i}} \right) \right) \right) \right), \quad (3)$$

where $\mathcal{C}_i$ is the request handling capacity of the network subject to the availability of servers, ω is the requests per second of the setup with $\mathcal{U}_l$ being the left-out users leading to a certain number of pending requests, P_R , and H_R is the number of successfully handled requests. The rate at which the requests are handled impacts the rewards associated with the optimal control. It also regulates the capacity of the entire system. In the given model, the rate awards are modeled as:

$$\mathbf{R} = \sum_{i=0}^{\tau'} \left(\eta_{3,i} \left(\mathcal{R}_i - \left(\frac{\lambda_i \mathcal{X}}{\mu_i} \right) \right) \right), \quad (4)$$

where $\mathcal{R}_i$ is the rate of the network based on the arrival rate of requests λ with a size, μ , for a given number of processes. In the model, if memory is fixed, μ severely impacts β . Considering that memory drives the model, and the servers can accommodate requests based on their capacity, the energy awards are marked by understanding the difference between the expected average available energy, $\mathcal{E}$, and the actual consumption in a round, such that, prior to failures, the awards are:

$$\mathbf{E} = \sum_{i=0}^{\tau'} \left(\eta_{4,i} \left(\mathcal{E}_i - \sup_{0 \leq j \leq i} \mathbb{E}_{\mathcal{X}} \left(\frac{\mathcal{E}_{g,j}}{\beta_{g,j}} + \frac{\mathcal{E}_{\mathcal{N},j}}{\beta_{\mathcal{N},j}} \right) \right) \right), \quad (5)$$

where $\frac{\mathcal{E}_g}{\beta_g}$ is the energy per memory ratio for the generated operations, j , and $\frac{\mathcal{E}_{\mathcal{N}}}{\beta_{\mathcal{N}}}$ is the energy memory ratio for the network at specified time. Based on the above formulation, the objective of the resource management is given as

$$\max(\Psi_{\tau'}(x)), \quad (6)$$

such that

$$\sup_{0 \leq \tau' \leq \tau} \epsilon_{\mathcal{X}_{\tau'-1}}, \quad (7)$$

where ϵ is the error rate for handling the given processes. The error rate can be accounted for a previous state that a system has passed. It is to be noted that in practical situations with critical failures, the action time of the network should be minimum with zero tolerance. The optimal control theory helps to operate through constraints to allow efficient shifting of services with zero tolerance. This prevents an eavesdropper from interfering with the network and even if it controls a certain number of servers, the checkpoint will help ensure the safety of the system. However, the entire model needs to identify the safe points under the optimal control at which the network can regain its state before any failure or attack. Another constraint to follow is that the time to reach the optimal point where the system is said to be safe, and it should be critically lower than τ' .

The shifting of services to facilitate resource management under osmotic computing are based on the prior probabilities [25], say for instance, κ , only for the network entities excluding the end users, such that

$$\alpha_h(\tau') = \frac{P(\kappa || \mathcal{N}_a) \cdot \frac{|\mathcal{N}_a|}{|\mathcal{N}|}}{P(\kappa || \mathcal{N}_a) \cdot \frac{|\mathcal{N}_a|}{|\mathcal{N}|} + P(\kappa || \mathcal{N}_a) \cdot \frac{|\mathcal{N}| - |\mathcal{N}_a|}{|\mathcal{N}|}}, \quad (8)$$

and

$$\alpha_p(\tau') = \frac{P(\kappa || \mathcal{N} - |\mathcal{N}_a|) \cdot \frac{|\mathcal{N}| - |\mathcal{N}_a|}{|\mathcal{N}|}}{P(\kappa || \mathcal{N} - |\mathcal{N}_a|) \cdot \frac{|\mathcal{N}_a|}{|\mathcal{N}|} + P(\kappa || \mathcal{N}_a) \cdot \frac{|\mathcal{N}| - |\mathcal{N}_a|}{|\mathcal{N}|}}, \quad (9)$$

where α_h and α_p are the associated Priori for the handling and pending to respond network entities, such that $\mathcal{N}_a \subseteq \mathcal{N}$ is the set of available entities. The shift of services is needed when the number of failures in the existing network increases, i.e., shift if $\alpha_p(\tau') > \alpha_h(\tau')$. This, however, do not cover the load and exceeding requests, which are only modeled around (6). To cover these, the optimal control point [19] for shifting the

services for any metric in $\Psi_{\tau'}$, is given based on the rewards, as:

$$\begin{aligned} & \sum_{i=1}^{|M|} \vartheta_i \int_0^{t_2} \left(m_i(t) - \overline{m_i(t)} \right)^2 f(m_i(t)) dm - \sum_{i=1}^{|M|} \frac{\vartheta_i \times \overline{m_i(t)}}{|M|} \\ & < \sum_{i=1}^{|M|} \vartheta_i \int_0^{t_1} \left(m_i(t) - \overline{m_i(t)} \right)^2 f(m_i(t)) dm - \sum_{i=1}^{|M|} \frac{\vartheta_i \times \overline{m_i(t)}}{|M|} \end{aligned} \quad (10)$$

where the set, M , $\forall m \in M$, includes the metrics used for obtaining $\Psi_{\tau'}$ and $t_1 < t_2$, and $t_2 \leq \tau'$, $f(m_i)$ is the probability density function, and $\overline{m_i}$ is equivalent to $\int m_i f(m_i) dm$. Here, ϑ is the reward balancing constant that helps to control the scenarios where one metric may dominate the other or the system is unavailable due to adjustments.

A. Conditions of reverse-osmosis

The conditions of reverse-osmosis to decide on the resources and requests are as follows:

- *Preventing Isolation:* The entire approach operates with the aim that none of the nodes in the network is isolated. An upper limit is set on the degree of each node to ensure that the flow of information is not to the sufficiently large groups that may impact the privacy of the system.
- *Shortest-pass Authentication:* In the network, privacy is governed by ensuring that only intended nodes receive the particular information. This also accompanies the selection of the shortest path possible for the connected graph. The rule of the shortest path is for the nodes already having established a session between them. According to this rule, the number of intermediates needs to be lower, and the services must be allocated with criteria that a lesser number of nodes are used to settle (6).
- *Obeying Security-policies:* The proposed model needs to ensure that the rules of osmosis are accompanied by the security policies which are never violated. Supporting services on the servers which are having high availability but lacks the security of the sessions and have a high degree of connections should be avoided.
- *Traceable Allocations:* Concurrent requests end up at different servers and layers. Concurrent requests make it difficult for the allocating entity to decide on the control of information across the network. These allocations need to be traceable to their source such that the de-registration process is concluded under the privacy settings.
- *Decentralization:* A decentralized architecture that can trace honest and legitimate nodes is helpful to ensure that the privacy policies are not controlled by a single entity, and this prevents a single point of failure as well as a single-node authorization on the resource management.

B. Assumptions of the operational model

The reverse osmotic computing-based privacy-aware resource management happens with the difference of the concentration strategy used to filter out the requests. If the concentration strategy is not the same, then the attacker cannot decide how the resources are allocated. However, here it is assumed

TABLE II
DESIGN PRINCIPLES OF PRIVACY-AWARE RESOURCE MANAGEMENT FOR OSMOTIC COMPUTING USING SIX DESIGN CONCEPTS IN [20].

Privacy-awareness design concepts	Relevance to osmotic computing
Notice (Lemma 4.1)	The osmotic membrane which acts upon the predefined set of rules become the advertised code of practice for resource management. The notice of openness is decided in the formulation phase of the osmotic computing, where the resources are decided for handling the services.
Choice and consent (Lemma 4.1)	Users can attach additional filters to the membrane or specify their reward expectations in the safety competition (12) as an explicit indication of their consent to use the platform. This allows users to fixate their choice based on the current values of balancing constants. It also facilitates developing a system with additional settings and support to the users. Safety by configuration is a key in the choice and consent of the users.
Anonymity and pseudonymity (Lemma 4.1, Lemma 4.2)	The unpredictable nature of resource management through the transitioning function (11) allows for additional memory and bandwidth to be allocated to achieve anonymity required at the expense of reward (12). Users can also choose to split the processing across multiple devices to remove the concentration risk of data being found at any one location.
Proximity and locality (Lemma 4.1)	Locations that are deemed not-feasible for data use can be defined as unrecognizable components where the the reverse-osmosis process to prevent the solvents (or services) from reaching these locations.
Security	The osmotic computing paradigm does not prevent the use of encryption to protect the data in transit or in storage. However, at this instance, such details are beyond the modelling of the proposed system.
Access and recourse (Lemma 4.1)	Additional filters can be attached to the membrane to define a view, modify, and delete operations. Since the underlying configurations are actually hidden from all users (including intruders), the filters will by default be dormant and only trigger if conditions are met.

that the servers are secure and cannot be compromised by the intruder. Such methods in the real world may require additional protocols to assure the safety of the system. Moreover, to prevent an attacker from using patterns of service shifts to trace the resources through secure migration, two strategies can be adapted.

Case-1: All the servers store the information of another available server, however, this may not be following the principle of M-IoT and cannot support mobility. To resolve mobility constraints, separate exchange protocols are expected to be employed for sharing configurations between the servers.

Case-2: Once the network is initialized, a server is expected to manage devices that show the configuration during the initial session and then the server maintains the calculation on its own to decide on the resources. The acknowledgement message (ACK) in these case help to identify how the allocated services are handled. The ACK can come as a flag that can be pre-decided during the session establishment and can be changed for each session. A situation may arise where an attacker sends the reports to the servers to understand the threshold used for filtering. As there is no exchange aspect in the resource management, the decision values cannot be manipulated. With these assumptions and support of protocols, even if an attacker acquires some of the resources, it cannot affect the management of resources. However, the fake advertisement of resources can be one of the issues in such an approach which is only possible if the attacker can replace the physical credential provided by the manufacturer. Such a threat model is beyond the scope of this solution. In building privacy-awareness into resource management, six design concepts of privacy-awareness [20] and anonymity through constant randomness [26] are adapted and applied to osmotic computing. The description and application are found in Table II. Of the six concepts, *Notice* and *Proximity and locality* map directly by design into the reward constants and operations of the osmotic membrane. *Choice and consent* and *Access and resource* are achieved through additional filters that specify reward expectations in the safety competition, while

Anonymity and pseudonymity are achieved by the distributed nature of the setup and over-allocating of resources to accommodate indistinguishability of the services run. *Security* by way of using encryption to ensure confidentiality is not affected by osmotic computing. In access management, the osmotic servers classify the services based on the available resources and the decision module uses distinct constants. It prevents the identification of an individual as every component will be marked with a different index and different constants will be issued for the incoming data.

III. PROPOSED SOLUTION

In the proposed model, osmotic computing is leveraged to integrate the network which is used for allocating the resources specified in the critical scenarios which involve M-IoT. Given the particulars of mobility and having mobile terminals operating between the end-users' edge layer and switches, the resource management is carried over the rewards associated with the consumption of resources, which include memory, capacity, rate and energy. A competition is formulated to identify the controlling rewards in the system and to mark which reward will control the resources.

A. Resource management

In this model, the network phenomenon relies on reverse osmosis, which is an alternative to osmosis for the flow of the solvent. The reverse osmotic part helps to retain unrecognizable components on the servers. The unrecognizable components include unauthorized service requests and control conditions to select a particular server based on optimization criteria, checkpoints, and privacy policies.

The model operates by deriving an osmotic transition function, f_Ψ , which is defined as

$$f_\Psi : (\mathcal{N}_a) \times (\mathcal{X}) \rightarrow (\mathcal{N}_a), \mathcal{N}_a \subseteq \mathcal{N}, \quad (11)$$

that takes into account the available infrastructure. (11) helps to build a service allocation graph, $G' = (\mathcal{N}_a, \mathcal{X}, \vartheta)$, which

is used to decide on the entities that will handle the given services and the metrics that are accounted for taking the decision on the movement of the services. The function, f_Ψ can take ϵ and evaluate an output ϑ , such that $f_\Psi \times \epsilon \rightarrow \vartheta$. The balancing constant impacts the conditions of transferring services as well as the servers which will be available for handling the requests. To have an effective strategy, the proposed model leads to $\min(f(\epsilon, \tau'))$ and $\max(f(G'))$, i.e., $\mathcal{N}_a$ becomes equivalent to $\mathcal{N}$ after input of more resources. These functions can be defined as per the conditions of the network. However, at a certain timestamp, these functions are expected to follow a particular distribution. These distributions help to apply approximation to predict a particular failure in the network and assure that all services are accurately allocated to the servers, and privacy-aware safety competition has all available input points.

Algorithm 1 Strategies for resource management

```

1: Input: state of the system
2: Output: Strategic shifting of services
3: Init(),  $t_1, t_2$ 
4: while (Incoming( $\mathcal{X}$ ) == True) do
5:   Initialize rewards
6:   set dominance  $\{\eta_1, \eta_2, \eta_3, \eta_4\}$ 
7:   if ( $\epsilon(t_2) \leq \epsilon(t_1)$ ) then
8:     Service_Shuffling(False)
9:   else
10:    Check for available servers:  $\mathcal{S}$ 
11:    Calculate (1) using (2)-(5)
12:    set max optimization criteria for (6)
13:    if ((10) == True) then
14:      Allocate_Resources(True)
15:      if ( $\alpha_p < \alpha_h$ ) then
16:        map: (11)
17:        Resources_Allocated(True)
18:      else
19:        reload and reset the balancing constant
20:      end if
21:    else
22:      No shifting required
23:    end if
24:  end if
25: end while
26: Keep log and continue, increment timestamp

```

Algorithm 1 helps to realize the formation of the strategy once the resource management procedures are initiated on the arrival of additional services. The model operates on the initial system and validates the conditions expressed by the optimal control theory in (1). The effective strategy ensures that the error rate is minimized, and the balancing constant is appropriately derived to maintain the connectivity between the devices. The privacy-aware service mapping is expressed in the Algorithm 2. The procedures control the privacy policies and control the non-legitimate services whenever the network is introduced with new handling requests. The model used timestamps to match and verify the requests and maintain the freshness of data and build trust on top of privacy policies to have an appropriate selection of the servers. Algorithm 2 maintains the continuity with a condition that balancing constants and optimization criteria change for different configurations. The procedures are assuring a formation of the state

Algorithm 2 Privacy-aware service mapping

```

1: Input: Requests, Entities, Algorithm 1 observations
2: Output: Safety-competition and rewards
3: Init(),  $t_1, t_2$ 
4: while (Incoming( $\mathcal{X}$ ) == True) do
5:   Generate  $G' = (\mathcal{N}_a, \mathcal{X}, \vartheta)$ 
6:   if (Is_Distinct( $\delta$ ) == True) then
7:     Obtain  $\Gamma$  using (13)
8:     if ( $\Gamma(t_2) \geq \Gamma(t_1)$ ) then
9:       check for balancing constants
10:      Initiate resource migration and balancing
11:    else
12:      Discard or improve  $\Gamma$ 
13:    end if
14:    Generate incidence and adjacency matrices over  $G'$ 
15:    Match and remove redundant entities
16:    set optimal points for min and max for the mapping functions
17:    if ( $\min(f(\epsilon, \tau')) \&\& \max(f(G'))$ ) then
18:      fetch allocated servers
19:      Trace vacant slots over servers
20:      Refresh  $G'$  and re-instate min and max conditions
21:      Invoke safety-competition for sharing  $\Gamma$ 
22:    else
23:      Migration_Infeasible(True)
24:    end if
25:  else
26:    Update configurations and re-initiate network
27:  end if
28: end while
29: Keep log and continue, increment timestamp

```

diagram which makes sure that no information is leaked to an intruder. The privacy formulation atop of these algorithms is derived as a service competition, which lets to an unbiased decision on the metric to be used for resource management. As an eavesdropper wants to identify these configurations, the protection of the constants will prevent an attacker to identify the servers which will be used for handling the requests. This process is repeated for all timestamps, and it is helpful to control the system till it matches τ' . Algorithm 1 operates over the incoming processes and solving (1) using (2)-(5), which leads to a complexity of $O(\mathcal{X} \times \tau)$. If the number of processes remains the same for every timestamp, then, for an input size of 1,000,000, the output can be obtained in 5.63 seconds (approx.) on an Intel Core i7 Extreme Edition 3960X system. The complexity of Algorithm 2 is dominated by the approach used for generating the mapping graph and checking the value of gap , which can be attained in $O(\mathcal{N}_a + \mathcal{X})$.

B. Privacy-aware Safety Competition

The privacy-preservation solution is expressed based on the safety of the balancing constants and the decisions on award constants. The safety of the configuration and constraining them into the periphery of the secure servers allows ensures privacy during the management of the resources. The safety competition, $\mathbb{S}_C$, which is derived from the unification of strategies defined as, $\delta_1, \delta_2, \delta_3, \dots, \delta_k$ for k available servers, $S_1, S_2, S_3, \dots, S_k$, where $S_k \in \mathcal{S}$. The safety competition is given as

$$\mathbb{S}_{C,(k)} = \delta_k(\Psi_{\tau'}, S_k, \epsilon_k, \vartheta_k), \quad (12)$$

such that $\delta_1 \neq \delta_2 \neq \delta_3 \neq \dots \neq \delta_k$, even if, configurations of servers are equal. Let Γ be the *gap* between the strategies defined as the *configuration rewards* (inspired by Markov Modelling [27] [28]), such that for ρ number of alternative-strategies at each instance, the *gap* should be maximum, however, ensuring that the constraints ((7)-(10)) are always settled. This is a trade-off as a higher *gap* ensures better safety, but it lowers the options if k is too large. A smaller value of k supports privacy at lower probabilities, however, with a possibility of brute force. Thus, the choice of strategy will impact the privacy of the system. The *gap* is quantified as:

$$\Gamma(t) = \frac{\sum_{y=1}^k \mathbb{S}_{\mathcal{C},(y)}^{unique}(t)}{\sum_{i=1}^k \mathbb{S}_{\mathcal{C},(i)}^{all}(t) + \sum_{z=1}^k \mathbb{S}_{\mathcal{C},(z)}^{same}(t)}, y \neq z, z + y = i, \quad (13)$$

where *same* refers to equivalent δ whereas *unique* means different δ , which is the desired formation in the proposed model. The privacy is ensured at $\max(\Gamma(t)), \forall k \in \mathcal{S}$. With this, δ_k will be different in each round resulting to the maximum of *gap* $\forall t \in [0, \tau']$. The solution to this safety competition lies in the selection of non-repetitive servers for similar types of service requests and balancing the load in accordance with ϑ . Optimal δ helps to control the privacy of the model and provide effective resource management even when high mobility of devices is involved. The state-diagram, Figure 2, helps to summarize this process of optimizing constraints to have better privacy. The state diagram comprises four processes which include servers and incoming service requests, safety competition, reward-maximization, and system decisions. The network load is taken into account to decide on the strategies of safety competition. Following this, the model aims to maximize the rewards (*gap*) and balance the constants where optimal values help to decide for the system.

Fig. 2. An overview of the state diagram for handling the safety competition in resource management.

IV. THEORETICAL ANALYSIS

This section helps to understand the theoretical reachability of the proposed solution.

Lemma 4.1 (Graph-based privacy preservation): The formation of service-graph ensures the privacy of operations via adjacency defined through the decisive-membrane of the osmotic computing.

Proof. The proposed approach uses the balancing and reward constants where *Notice* to users is achieved as part

of the osmosis. It controls the allocation of the network facilities to request for the handling of the services to the decision module. The strength of privacy is modelled around the non-disclosure of the configurations. It prohibits the intruder from identifying the state of the system, and use any reverse mechanisms, such as brute force, to predict the server handling a particular service. On the other hand, valid users' requirements by way of *Choice and consent* and *Access and recourse* are achieved with the filters associated with the membrane. The entire phenomenon is facilitated by $G'(\tau') = (\mathcal{N}_a, \mathcal{X}, \vartheta)$ which controls the incidence and adjacency properties of the network, where *Proximity and locality* are enforced. The freshness of the strategy, δ , prohibits an intruder from coping up with the network's ability to trace management strategy, thereby, achieving *Anonymity and pseudonymity*. Thus, preserving the privacy of resource management. $\square$

Lemma 4.2 (Connection-secrecy): The proposed modeling ensures secret allocation of resources via non-disclosed metrics of the privacy-aware game.

Proof. In the proposed model, the resource allocation is attained through reward-calculations of the function operating over memory, capacity, rate, and energy. Such a solution can be trivial if the values of the utilization are known. The advantage of the model lies in its strategy that derives safety competition between the servers. This competition aims at a maximum *gap* that is assumed to be shared between the servers. No other information is passed, which prohibits an intruder from guessing even if it breaks into the session. The *gap* prevents the repetition of the allocated servers despite having similar configurations, which may be the ones used in the earlier round. This can be followed in (12), which helps derive the *gap*, $\Gamma(t)$, without exposing the metrics, at any instance, t . $\square$

Lemma 4.3 (Optimal point for balancing constants): The proposed approach helps to privatize the resource management through τ -state Prio by calculating the definite interval for the migration of the services.

Proof. In the proposed work, the privacy-based resource management is monitored according to (10). Considering for a single entity and only a constraint from (1), the condition can be written, deducing discrete event $\sum_{i=1}^{|M|} \frac{\vartheta_i \times \overline{m}_i(t)}{|M|}$ to constant occurrence $\int_0^\tau m_i \frac{e^{-\frac{m_i^2}{2}}}{\sqrt{2\pi}} dm$, as

$$\vartheta \left(\int_0^\tau \frac{\left(x - \frac{1-e^{-\frac{x^2}{2}}}{\sqrt{2\sqrt{\pi}}} \right)^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} - x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \right), \quad (14)$$

where τ is the final instance, and x is the functional capacity from the start of the system for a single input metric. Solving (14) by applying linearity and plugging in Gauss error

Fig. 3. Memory rewards at different values of controlling constants vs. operational time for unlimited availability of energy.

Fig. 4. Capacity rewards at different values of controlling constants vs. operational time for unlimited availability of energy.

function, i.e. $\text{erf}(x) = \frac{2}{\pi} \int_0^x e^{-x^2} dx$, at $\tau > 0$, yields

$$\vartheta \left(\frac{e^{-\frac{\tau^2}{2} - 2\tau} \left((2\pi\tau^2 - 2\tau - 4\pi + 3) e^{2\tau} - 4e^\tau + 1 \right) e^{\frac{\tau^2}{2}} + 4\pi e^{2\tau}}{2^{\frac{5}{2}} \pi^{\frac{3}{2}}} \right), \quad (15)$$

which is the desired result expressing the upper limit for initiating the resource management. Using this upper limit the model can predict when a state of unavailability and constants can be balanced before entering such a state. Thus, making the proposed model undecidable for an intruder. $\square$

V. NUMERICAL RESULTS

In continuity with the theoretical evaluations, which help to understand privacy-aware properties of the proposed model, numerical simulations were conducted to understand the performance. In the evaluations, $|\mathcal{N}|=50$, $|\mathcal{S}|=20$, $\mathcal{X}=5$, z range between 0.01 and 0.5 (represents scarcity in the number of available servers), $\beta=20$ GB for each server at $\mathcal{R}=256$ Mbps, $\frac{\lambda}{\mu}=125$ Mbps, $m=512$ MB, $\omega=5$, $|\mathcal{U}|=5$, $\frac{H_R}{P_R}$ between 0.1 and 5, $\mathcal{E}$ is set at twice the amount required with $\frac{\mathcal{E}_g}{\beta_g}=0.1$ and $\frac{\mathcal{E}_N}{\beta_N}=0.1$, κ ranging between 0.1 and 0.9, and ϑ operating at three different values of 0.1, 0.5 and 0.9. In the proposed model, the memory plays a key role in deciding on the servers' abilities to provide additional resources for handling the services and depends on the layout of the network. In the derived model, the memory rewards will decrease with

an increasing operational time considering that the services arrive at a consistent rate following Poisson Distribution. The rewards will help to decide how much consideration needs to be given to the particular metric for deciding on migration policies. With different values of η_1 , shown in Figure 3, memory rewards increases and if the network can match these awards, it can continue in the given mode, otherwise, the services need to be switched to make sure that the *gap* between the derived rewards and the actual rewards is less. A similar strategy works for other metrics, such as capacity rewards, and can be seen in Figure 4. If the reward constant has a higher value, it means much preference is given to the metric, and it becomes the dominating metric. The *gap* needs to be accommodated only for the dominating metric, and the network can continue to operate if other metrics do not result in a maximum for the optimal control function as seen in (12) and (13).

Fig. 5. Optimal resource management points based on τ -Prior knowledge only to the server deciding on the migration and allocation of services vs. difference in τ indicating knowledge up to previous instances.

Fig. 6. Additive rewards of the optimal control function for deciding the movement of the service-requests using the server as a membrane.

The reward-based solution helps to understand the requirement of additional resources. This depends on the prior information until the current instance of the system. Better availability of information means better control and the system can wait for a longer duration before switching to the utilization of resources and sharing the services. This can be observed in Figure 5, which explains the optimal resource management points based on which the server can wait for a certain duration. In general, this is the upper limit, until

which waiting will not impact the performance. A better value of the balancing constant will increase the upper limit of the optimal resource management points and offer better resource management.

TABLE III
QUALITATIVE ANALYSIS OF THE PROPOSED AND EXISTING SOLUTIONS.

Approach	Osmotic paradigm	Resource policing	Privacy-by-design	Optimal policing	On-demand balancing
[15]	Yes	Yes	No	Yes	Yes
[29]	No	Yes	No	Yes	-
[30]	Yes	Yes	No	Yes	Yes
[31]	No	-	Yes	-	Yes
Proposed	Yes	Yes	Yes	Yes	Yes

The entire evaluations come into play when the rewards of the optimal control function are utilized to finally decide on the movement of services, as shown in Figure 6. In numerical evaluation, the additive strategy is used to compute these rewards for memory, capacity, rate, and energy. The reward constants help to decide on the requirements of the system. In Figure 6, it can be observed that if the energy is the dominating factor, the reward range of optimal function decreases indicating the network needs to have more energy resources to balance the resource management that can be followed for other metrics as well. To contrast the contributions of this work, a qualitative comparison is presented in Table III. The table presents a key difference with approaches that offer privacy-by design but does not consider optimal policing and others that cover optimal policing with resource management but do not ensure privacy awareness. In the proposed work, the server can decide on the choice of metrics to use for services that remain unknown to any other party trying to decide on the configurations of the model to affect privacy. However, including an end-to-end security protocol and authentication and access control solutions can enhance the utilization of the proposed model, which are beyond the scope of this article.

VI. CONCLUSION

This work supports optimal and privacy-aware resource management using osmotic computing for AIoT. The proposed solution provides an algorithmic solution to utilize osmosis for building optimal control function and associated rewards, which are operated with service graphs. A safety competition is formed to find the optimal reward *gap* for resource management that maintains a check on the information leakage by capturing τ -state Priori information of the network. The optimal configuration reward helps to attain privacy-by-design using osmotic computing.

REFERENCES

[1] H. Yuan, M. Zhou, Q. Liu, and A. Abusorrah, "Fine-grained resource provisioning and task scheduling for heterogeneous applications in distributed green clouds," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 5, pp. 1380–1393, 2020.

[2] G. Tang, D. Guo, K. Wu, F. Liu, and Y. Qin, "Qos guaranteed edge cloud resource provisioning for vehicle fleets," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 5889–5900, 2020.

[3] S. Deng, Z. Xiang, P. Zhao, J. Taheri, H. Gao, J. Yin, and A. Y. Zomaya, "Dynamical resource allocation in edge for trustable internet-of-things systems: A reinforcement learning method," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 6103–6113, 2020.

[4] H. Yu, J. Yang, and C. Fung, "Fine-grained cloud resource provisioning for virtual network function," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1363–1376, 2020.

[5] Y. Sharma, J. Taheri, W. Si, D. Sun, and B. Javadi, "Dynamic resource provisioning for sustainable cloud computing systems in the presence of correlated failures," *IEEE Transactions on Sustainable Computing*, 10.1109/TSUSC.2020.3025180, 2020.

[6] L. Zhang and J. Li, "Enabling robust and privacy-preserving resource allocation in fog computing," *IEEE Access*, vol. 6, pp. 50384–50393, 2018.

[7] Z. Wang, J. Hu, R. Lv, J. Wei, Q. Wang, D. Yang, and H. Qi, "Personalized privacy-preserving task allocation for mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1330–1341, 2018.

[8] T. L. Duc, R. G. Leiva, P. Casari, and P.-O. Östberg, "Machine learning methods for reliable resource provisioning in edge-cloud computing: A survey," *ACM Computing Surveys*, vol. 52, no. 5, pp. 1–39, 2019.

[9] Ghosh, Ashish and Chakraborty, Debasrita and Law, Anwesha, "Artificial intelligence in Internet of things," *CAAI Transactions on Intelligence Technology*, vol. 3, no. 4, pp. 208–218, 2018.

[10] Fang, He and Qi, Angie and Wang, Xianbin, "Fast authentication and progressive authorization in large-scale IoT: How to leverage AI for security enhancement," *IEEE Network*, vol. 34, no. 3, pp. 24–29, 2020.

[11] Gope, Prosanta and Sikdar, Biplab and Millwood, Owen, "A Scalable Protocol Level Approach to Prevent Machine Learning Attacks on PUF-based Authentication Mechanisms for Internet-of-Medical-Things," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 10.1109/TII.2021.3096048, 2021.

[12] H. Ren, C. Pan, Y. Deng, M. El-kashlan, and A. Nallanathan, "Resource allocation for secure urlc in mission-critical iot scenarios," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5793–5807, 2020.

[13] W. Ning, Y. Wang, Y. Chen, and M. Liu, "Resource allocation in relay-assisted mission-critical industrial internet of things," in *Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–6, IEEE, 2020.

[14] M. Villari, M. Fazio, S. Dustdar, O. Rana, and R. Ranjan, "Osmotic computing: A new paradigm for edge/cloud integration," *IEEE Cloud Computing*, vol. 3, no. 6, pp. 76–83, 2016.

[15] Sharma, Vishal and Jayakody, Dushantha Nalin K and Qaraqe, Marwa, "Osmotic computing-based service migration and resource scheduling in Mobile Augmented Reality Networks (MARN)," *Future Generation Computer Systems*, vol. 102, pp. 723–737, 2020.

[16] J. Xiong, M. Zhao, M. Z. A. Bhuiyan, L. Chen, and Y. Tian, "An ai-enabled three-party game framework for guaranteed data privacy in mobile edge crowdsensing of iot," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 922–933, 2019.

[17] M. Lin and Y. Zhao, "Artificial intelligence-empowered resource management for future wireless communications: A survey," *China Communications*, vol. 17, no. 3, pp. 58–77, 2020.

[18] Sharma, Vishal and You, Ilsun and Andersson, Karl and Palmieri, Francesco and Rehmani, Mubashir Husain and Lim, Jaedeok, "Security, privacy and trust for smart mobile-Internet of Things (M-IoT): A survey," *IEEE Access*, vol. 8, pp. 167123–167163, 2020.

[19] G. Peskir and A. Shiryaev, *Optimal stopping and free-boundary problems*. Lectures in Mathematics, Springer, ETH Zürich, 2006.

[20] Langheinrich, Marc, "Privacy by design—principles of privacy-aware ubiquitous systems," in *International conference on Ubiquitous Computing*, pp. 273–291, Springer, 2001.

[21] L. Carnevale, A. Celesti, A. Galletta, S. Dustdar, and M. Villari, "From the cloud to edge and iot: a smart orchestration architecture for enabling osmotic computing," in *32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pp. 419–424, IEEE, 2018.

[22] V. Sharma, I. You, F.-Y. Leu, and M. Atiquzzaman, "Secure and efficient protocol for fast handover in 5g mobile xhaul networks," *Journal of Network and Computer Applications*, vol. 102, pp. 38–57, 2018.

[23] R. Sargent, "Optimal control," *Journal of Computational and Applied Mathematics*, vol. 124, no. 1–2, pp. 361–371, 2000.

[24] M. Bell and R. Sargent, "Optimal control of inequality constrained dae systems," *Computers & Chemical Engineering*, vol. 24, no. 11, pp. 2385–2404, 2000.

- [25] V. Sharma, G. Choudhary, Y. Ko, and I. You, "Behavior and vulnerability assessment of drones-enabled industrial internet of things (iiot)," *IEEE Access*, vol. 6, pp. 43368–43383, 2018.
- [26] Sweeney, Latanya, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.
- [27] Rashedi, Navid and Tajeddini, Mohammad Amin and Kebriaei, Hamed, "Markov game approach for multi-agent competitive bidding strategies in electricity market," *IET Generation, Transmission & Distribution*, vol. 10, no. 15, pp. 3756–3763, 2016.
- [28] Liao, Peng and Qi, Zhengling and Murphy, Susan, "Batch policy learning in average reward Markov decision processes," *arXiv preprint arXiv:2007.11771*, 2020.
- [29] Gu, Bo and Zhang, Xu and Lin, Ziqi and Alazab, Mamoun, "Deep multiagent reinforcement-learning-based resource allocation for Internet of controllable things," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3066–3074, 2020.
- [30] Kaur, Kuljeet and Garg, Sahil and Kaddoum, Georges and Ahmed, Syed Hassan and Jayakody, Dushantha Nalin K, "En-osco: energy-aware osmotic computing framework using hyper-heuristics," in *Proceedings of the ACM MobiHoc Workshop on Pervasive Systems in the IoT Era*, pp. 19–24, 2019.
- [31] Singh, Saurabh and Ra, In-Ho and Meng, Weizhi and Kaur, Maninder and Cho, Gi Hwan, "SH-BlockCC: A secure and efficient Internet of things smart home architecture based on cloud computing and blockchain technology," *International Journal of Distributed Sensor Networks*, vol. 15, no. 4, pp. 1–18, 2019.