Paving the Way for an Energy Efficient and Sustainable Future Internet of Things

Frank Loh, Simon Raffeck, Stefan Geißler, Tobias Hoßfeld
University of Würzburg, Institute of Computer Science, Würzburg, Germany
{firstname.lastname}@informatik.uni-wuerzburg.de

I. Introduction and Problem Formulation

The continuous development of new applications and monitoring solutions fosters the success of the Internet of Things (IoT). According to Statista, there have been more than 11 billion connected devices worldwide in the IoT in 2021, and forecasts predict nearly 30 billion by 2030 [1]. In this context, many wireless access network technologies arise from general cellular access over 5G and future 6G to different Low Power Wide Area Network (LPWAN) solutions. Furthermore, several proprietary technologies for very specific use cases or sensor networks have been developed.

Nevertheless, for many IoT application areas, the general behavior is rather similar: end devices collect data and that needs to be transmitted to a gateway. The data is processed, aggregated based on its specific purpose, or stored in a centralized data center. Because of an increasing number of end devices and applications, the resulting traffic load increases and as a result, more resources are consumed to realized IoT use cases. In contrast, to provide a more sustainable network, a general need for resource sparing and energy efficient operation is required.

This involves two major tasks: selecting the correct technology for the correct use case and using the technology in the most effective and efficient manner. To this end, it is the task of the network operator to provide the end user with the appropriate technology to guarantee service level agreements (SLAs) and provide the end user a high Quality of Service (QoS) for the applications. On the other hand, the operator's goal is to work economically and, with the increasing interest in green and sustainable operation, use as little resources as possible. In this context, literature already exist to improve general service quality for specific access network technologies by investigating different optimization potential (e.g. [2]), but sustainability is often not yet taken into consideration.

For data collection and transmissions from monitoring devices to gateways, much research has been conducted, studying the general performance of different wireless access network solutions in recent years. Examples are, among others, measurement and classification of 5G traffic [3], very specific IoT traffic like signaling [4], or the general behavior of different other access network technologies in IoT [5], [6], [7]. Furthermore, energy models are provided for low power IoT in general [8], for very specific access network technologies like Long Range Wide Area Network (LoRaWAN) [9], or general energy efficiency in narrowband IoT [10]. However, a holistic view of the network in general and the impact of improvement during data transmission on the complete access network, and also on other network parts is still missing.

For example, the deployment of more gateways in a LoRaWAN can improve the transmission quality in the access network [11], [12] and also the energy consumption and efficiency of the individual nodes [13]. In return, more gateways must be deployed, maintained, and supplied with power [11]. Another example is complex, real time resource provisioning that can improve the performance of the access network by in time resource, frequency, and time slot allocation in an optimized manner. In return, it has a negative impact on the energy footprint in the cloud or in data centers due to the required comprehensive analysis of the available resources and application of, for example, computationally expensive machine learning solutions. Besides this, the energy requirements and the carbon footprint for the used devices itself needs to be considered during manufacturing since large differences are visible between a simple sensor, a complex and suitable device in a 5G or future 6G network, or a high power server rack. Often, applications only require a small subset of the features complex sensors provide and high resource saving is possible in this context. To the best of our knowledge, the consideration of the entire network under a green and sustainable point of view has not been done yet in literature and is rather challenging because of the multitude of impact factors and inter-dependencies. This also includes a comprehensive technology selection for different use cases, starting from the access network, service and performance management and monitoring systems, and data storage and evaluation in different locations and by different hardware.

For that reason, we describe the network from data collection to storage as a system with different components in this work. We outline the different components and highlight dependencies with regard to energy consumption and energy efficiency. With this view, we can identify different challenges and questions a future system architect must answer to provide a more sustainable, green, resource friendly, and energy efficient application or system. Therefore, all system components must be considered individually but also in conjunction for future IoT solutions. This requires, in particular, novel sustainability metrics in addition to current Quality of Service (QoS) and Quality of Experience (QoE) metrics,

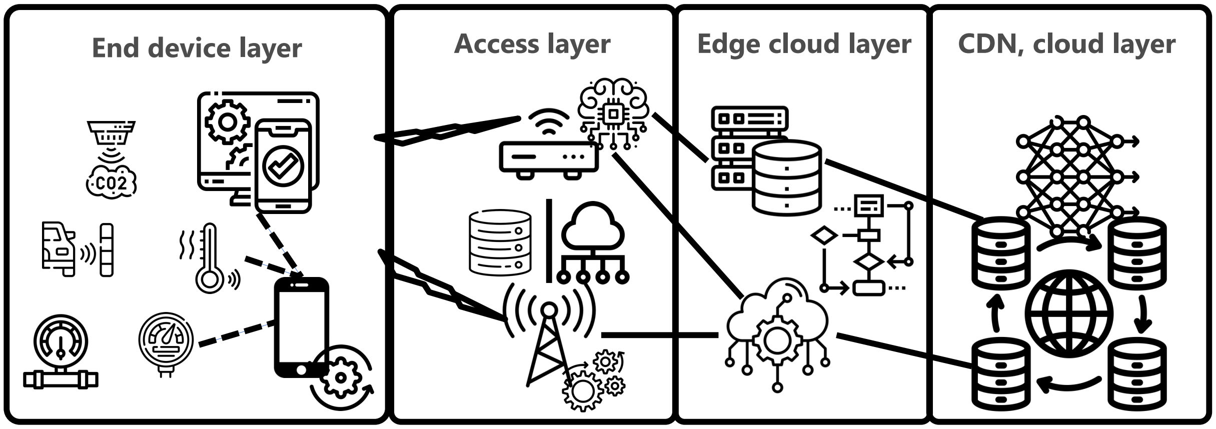

Figure 1: Overview of relevant network layers.

to provide high performing and high quality yet sustainable network operation. In this context, the following Section II introduces the general IoT as a system of different layers, outlines components and their behavior and role within each layer, possible interoperability and influences of improving single components on other ones. Furthermore, the aim for future sustainability metrics is highlighted and key positions in the network where sustainability can be quantified are introduced. Finally, Section III presents and discusses examples for possible IoT system improvements and summarizes future challenges and research topics.

II. Concept and Methodology

According to literature, the network can be split in perception layer, network layer, middleware layer, cloud layer, and application layer [14]. In this layered approach, network parts are aggregated based on their responsibility. Thus, for example, device-to-device communication between sensors is added to the network layer. In contrast, an in-depth energy consumption and sustainability analysis in IoT requires a slightly different separation between network parts. For that reason, we set up layers based on operation location of applications or specific processing locations, as shown in Figure 1. Since IoT applications can operate in different parts of the network and its data can be processed in different locations, improvement by means of energy consumption, energy efficiency, carbon footprint, and sustainability in general is possible dependent on the operation location. In the following, the different layers are introduced and discussed briefly.

End Device Layer: First, the end device layer, shown to the left of the figure, is mainly responsible for raw data measurement. The applications can include simple monitoring tasks like air quality or temperature monitoring where major communication is in uplink direction and downlink is only used for updates or synchronization. Furthermore, more complex tasks like autonomous driving or Industry 4.0 applications can be served in this layer. This can include direct communication between devices without the requirement of another layer for communication, high traffic rates to access points in the network, and also a significant amount of downlink traffic. Furthermore, complex end devices, highlighted by the smartphones and the screen in the example figure, can be deployed here. These devices can pre-process specific data, providing besides monitoring also computation resources.

Thus, dependent on the general application, the task, and the end devices, also different wireless technologies are required, considered, and analyzed based on sustainability. A multitude of different simple monitoring tasks without the requirement of reliable communication can use, for example, a simple LPWAN technology like LoRaWAN. Such technologies have benefits like easy and cheap usage and hardware, consume little resources, can transmit across long distances, and work in a very energy efficient way. The downside is possible message loss [5] because of the often unreliable transmission. Other alternatives are, among others, ZigBee, Sigfox, NB-IoT, or different weightless standards. When it comes to reliable, fast, and timely communication, and the requirement of high data rates, current cellular communication, WiFi, 5G, or future 6G might be more appropriate. The drawback are much smaller cells, more required gateways that consuming energy and other resources during production. In addition, the cost and resource requirements for data transmission and for the end devices is higher. To this end, a wise selection of the correct technology is required for each application. Furthermore, to achieve the highest possible Quality of Information in the most energy efficient and sustainable way, among others, end device locations, number of end devices per task, online and sleep times of the devices, and data monitoring and transmission patterns must be analyzed in detail during application deployment.

Access Layer: The access points to the Internet for the end devices from the end device layer are in the access layer. Gateways can be simple and small home routers deployed for smart home or smart city applications. For typical cellular communication, gateways are often large base station. In recent years, also mobile base stations with, for example, drones increased in popularity to provide additional resources during specific high traffic times. All data that is not directly processed or used within the end device layer and is not lost during the, mainly wireless, transmission to the gateways arrive here. The access layer provides connection to other parts of the Internet, but often also little computational resources as part of the general term Edge or Fog Cloud. Furthermore, it is the first possibility to exchange data between different access network technologies as gateways are usually connected to the Internet via Ethernet or other wired technologies.

To operate economically and sustainably, in-depth gateway and resource planning is required for the future IoT. It is essential that additional small computational resources are close to end devices within the access layer if in return, large overhead in transmission load can be saved compared to transmitting all data to a large data center. Furthermore, data pre-selection and pre-processing can be done to reduce load towards data centers. At the end, it is essential to guarantee SLAs, which is often a trade-off between moving computational resources closer to the end user reducing latency but in return, the requirement for additional servers and energy. In contrast, when data is transmitted to a larger data center less server locations are required and already available resources that can be managed more efficiently. This, however, can lead to additional end-to-end delay. In future IoT, besides typical QoS metrics like latency or bandwidth, data transmission in the most energy efficient and sustainable way is required as additional Key Performance Indicator (KPI). This means that future gateways could provide interfaces for different wireless access network technologies, and thus, coverage for the appropriate technology. This can reduce the number of additionally deployed gateways, required space for new gateways, and reduce total energy consumption by pooling resources together.

Edge Cloud Layer: Next, the edge cloud layer is responsible for data processing and network management if end-to-end delay is a significant SLA. There, even complex processing and Machine Learning tasks can be performed and data can be stored, if required, with small delay guarantees. The drawback is that additional distributed edge clouds can not profit from pooling gain and can, in the best case, only be turned to idle or sleep mode if the resources are not required. Furthermore, turning them back to operation mode could even result in higher delay [15], violating SLAs.

Thus, from an energy efficiency and sustainability point of view, Edge Clouds should be used efficiently, if needed, and deployed in locations where applications require little delay. In contrast, for applications that can deal with higher delay, an in-depth investigation is required whether processing in the Edge is beneficial or data can be transmitted in a more energy efficient way to the data center. Besides delay, also the amount of data and the network structure should be included in this investigation. Therefore, a general trade-off between the benefit of improving the access network and the additional resource consumption in the edge cloud for data processing, network model and management decisions, and Machine Learning model training should be analyzed from a sustainability point of view. Furthermore, units should be turned offline if not needed, to save energy.

CDN and Cloud Layer: The last layer is the Content Delivery Network (CDN) and cloud layer. It contains large data centers able to store and operate on a massive amount of data. Due to the scale of available resources, various complex data analysis and Machine Learning tasks can be performed here. However, especially in these hyperscaler environments, the efficient allocation and orchestration of resources is critical, both from a performance and a sustainability point of view. This includes, for example, the placement of services within data centers and across multiple regions, impacting the need for communication between, potentially far apart, services. In addition, the allocation of the right amount and the right type of resources is critical to operate efficiently. In this area, platforms like AWS and Google Cloud provide highly specialized services for things like machine learning [16], big data pipelines and long term storage, that operate much more efficiently than self-hosted solutions (e.g. in the form of EC2 instances).

Considering large data center operation from a sustainability point of view, besides optimization of the infrastructure itself, a large amount of compute resources is naturally required. This leads to massive energy consumption and carbon footprint. However, since this large amount of energy is required at one geographically small place, the usage of renewable energy must be ensured to ensure sustainable operation. Finally, pre-selection and service tuning by the service operator of the use cases in the end device layer is required to decide which data must be sent to the cloud, to which degree data must be analyzed, learned, and stored. Thus, from a future IoT point of view, efficient selection of applications and the generated data together with resource allocation and consumption is essential to avoid unnecessary transmissions of large amounts of data over large distances.

III. Summary and Discussion

The sustainable operation of end devices as well as network and cloud infrastructure is, without a doubt, one of the most critical and challenging tasks currently being discussed by the research community. To this end, the presented layered approach highlights several challenges for the future Internet of Things.

While each individual layer has been the subject of studies for several years and different optimization possibilities are available in literature, future research should be more application driven, optimizing the requirements, behavior, but also resource consumption. Furthermore, details about the interaction of different layers is often omitted. In particular research is required tackling Quality of Information in future networks. Often, capturing more data does not improve insights in the application behavior and just increases the monitoring overhead, load in the network, and required resources in the cloud. Therefore, a pre-selection of monitoring positions, sensor capability, and the target output is essential. Furthermore, an in-depth investigation of research towards the best data processing location and the required data to be processed is important. The recent heavy usage of Machine Learning based solutions should be re-evaluated based on required resources, energy consumption, carbon footprint, and sustainability in general. One possible approach is to define a resource or energy budget per application or layer. Then, different services, applications, or layers can be improved towards this energy budget. This means, if one layer requires more resources, another one must save resources.

For that reason, it is essential to determine and test quality metrics for sustainability, define new monitoring interfaces, and study the influence of improving sustainability on the overall network quality. One example is pre-selecting or aggregating data at the end device to save resources and improve transmission quality, as shown for LoRaWAN [17]. Another approach is the re-usage of existing Machine Learning models [18] or a more in-depth undertsanding of differnt application behavior to occasionally avoid the requirement of intense data generation and Machine Learning approaches [19].

Overall, future research needs to evaluate and re-evaluate both optimizations within each of the proposed layers and the interaction between layers. Where in the past, we optimized mainly for performance and resilience, future research must include sustainability as one of the key metrics for networks of the future.

References

[1] Statista, "Number of Internet of Things Connected Devices Worldwide from 2019 to 2021, with Forecasts from 2022 to 2030," 2022, accessed: May 09, 2023. [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

[2] F. Loh, S. Geißler, and T. Hoßfeld, "LoRaWAN Network Planning in Smart Environments: Towards Reliability, Scalability, and Cost Reduction," 2022.

[3] M. Dighriri, G. M. Lee, and T. Baker, "Measurement and Classification of Smart Systems Data Traffic over 5G Mobile Networks," *Technology for smart futures*, pp. 195–217, 2018.

[4] V. Vomhoff, S. Geissler, F. Loh, W. Bauer, and T. Hossfeld, "Characterizing Mobile Signaling Anomalies in the Internet-of-Things," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2022, pp. 1–6.

[5] F. Loh, N. Mehling, and T. Hoßfeld, "Towards LoRaWAN Without Data Loss: Studying the Performance of Different Channel Access Approaches," *Sensors*, vol. 22, no. 2, p. 691, 2022.

[6] F. Loh, N. Mehling, S. Geißler, and T. Hoßfeld, "Simulative Performance Study of Slotted Aloha for LoRaWAN Channel Access," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2022, pp. 1–9.

[7] H. A. Alobaidy, M. J. Singh, R. Nordin, N. F. Abdullah, C. G. Wei, and M. L. S. Soon, "Real-World Evaluation of Power Consumption and Performance of NB-IoT in Malaysia," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 614–11 632, 2021.

[8] T. Hoßfeld, S. Raffeck, F. Loh, and S. Geißler, "Analytical Model for the Energy Efficiency in Low Power IoT Deployments," in *2022 IEEE 8th International Conference on Network Softwarization (NetSoft)*. IEEE, 2022, pp. 19–24.

[9] F. Loh, S. Raffeck, S. Geißler, and T. Hoßfeld, "Generic Model to Quantify Energy Consumption for Different LoRaWAN Channel Access Methods," in *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2022, pp. 290–295.

[10] S. Popli, R. K. Jha, and S. Jain, "A Survey on Energy Efficient Narrowband Ineternet of Things (NBIoT): Architecture, Application and Challenges," *IEEE Access*, vol. 7, pp. 16 739–16 776, 2018.

[11] F. Loh, N. Mehling, S. Geißler, and T. Hoßfeld, "Efficient Graph-Based Gateway Placement for Large-Scale LoRaWAN Deployments," *Computer Communications*, vol. 204, pp. 11–23, 2023.

[12] ——, "Graph-based Gateway Placement for Better Performance in LoRaWAN Deployments," in *2022 20th Mediterranean Communication and Computer Networking Conference (MedComNet)*. IEEE, 2022, pp. 190–199.

[13] F. Loh, C. Baur, S. Geißler, H. ElBakoury, and T. Hoßfeld, "Collision and Energy Efficiency Assessment of LoRaWANs with Cluster-based Gateway Placement," in *2nd Workshop on Green and Sustainable Networking (GreenNet2023)*. IEEE, 2023.

[14] W. Zhang and B. Qu, "Security architecture of the Internet of Things oriented to perceptual layer," *International Journal on Computer, Consumer and Control (IJ3C)*, vol. 2, no. 2, pp. 37–45, 2013.

[15] K. Nguyen, F. Loh, T. Nguyen, D. Doan, H. T. Nguyen, and T. Hoßfeld, "Serverless Computing Lifecycle Model for Edge Cloud Deployments," in *2nd Workshop on Green and Sustainable Networking (GreenNet2023)*. IEEE, 2023.

[16] A. W. Services, "AWS - Services for Artificial Intelligence," 2023, accessed: May 10, 2023. [Online]. Available: https://aws.amazon.com/de/machine-learning/ai-services/

[17] F. Loh, S. Raffeck, F. Metzger, and T. Hoßfeld, "Improving LoRaWANs's Successful Information Transmission Rate with Redundancy," in *2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2021, pp. 313–318.

[18] F. Loh, K. Hildebrand, F. Wamser, S. Geißler, and T. Hoßfeld, "Machine Learning Based Study of QoE Metrics in Twitch.tv Live Streaming," in *2023 IEEE/IFIP Network Operations and Management Symposium*.

[19] F. Loh, A. Pimpinella, S. Geißler, and T. Hoßfeld, "Uplink-based Live Session Model for Stalling Prediction in Video Streaming," in *2023 IEEE/IFIP Network Operations and Management Symposium*.