computers

MDPI

Review

IoT Serverless Computing at the Edge: A Systematic Mapping Review

Vojdan Kjorveziroski *, Sonja Filiposka and Vladimir Trajkovik

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia; sonja.filiposka@finki.ukim.mk (S.F.); trvlado@finki.ukim.mk (V.T.)
* Correspondence: vojdan.kjorveziroski@finki.ukim.mk

Abstract: Serverless computing is a new concept allowing developers to focus on the core functionality of their code, while abstracting away the underlying infrastructure. Even though there are existing commercial serverless cloud providers and open-source solutions, dealing with the explosive growth of new Internet of Things (IoT) devices requires more efficient bandwidth utilization, reduced latency, and data preprocessing closer to the source, thus reducing the overall data volume and meeting privacy regulations. Moving serverless computing to the edge of the network is a topic that is actively being researched with the aim of solving these issues. This study presents a systematic mapping review of current progress made to this effect, analyzing work published between 1 January 2015 and 1 September 2021. Using a document selection methodology which emphasizes the quality of the papers obtained through querying several popular databases with relevant search terms, we have included 64 entries, which we then further categorized into eight main categories. Results show that there is an increasing interest in this area with rapid progress being made to solve the remaining open issues, which have also been summarized in this paper. Special attention is paid to open-source efforts, as well as open-access contributions.

Keywords: serverless computing; edge computing; function as a service; Internet of Things; systematic review

Citation: Kjorveziroski, V.; Filiposka, S.; Trajkovik, V. IoT Serverless Computing at the Edge: A Systematic Mapping Review. *Computers* **2021**, *10*, 130. https://doi.org/10.3390/computers10100130

Academic Editors: Paolo Bellavista, Kiran Kumar Pattanaik and Sourabh Bharti

Received: 11 September 2021
Accepted: 6 October 2021
Published: 15 October 2021

1. Introduction

The ever increasing progress in hardware development and computer networking paved the way for the introduction of cloud computing, which in turn has led to a new revolution, allowing computing capacity to be perceived as just another utility, used on-demand, with virtually limitless capacity [1]. Both academia and industry have invested in the creation of different cloud computing infrastructure, depending on their needs, currently available resources, and cost, resulting in the deployment of various private, public, community, and hybrid clouds [2]. However, to allow regular users to benefit from such vast computing capacity, additional abstractions are introduced, in the form of infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) offerings. IaaS provides the lowest level of abstraction, allowing users to rent computing, networking, and storage capacity on-demand, usually in the form of virtual machines (VMs), and utilize them as they see fit, building their own infrastructure on top. PaaS goes a step further, and is primarily aimed at software developers, abstracting away the necessary VM management, and instead providing the building blocks and interfaces for directly hosting developed applications, along with any other prerequisites, such as databases and message queues. Finally, SaaS, aimed at end-users, provides the highest level of abstraction, where the service being offered is a finished software product, ready to be used, without any additional requirements in terms of maintenance, updates, or management.

These three offerings are by no means the only products available as a service today. The idea of abstracting complicated tasks away from the users is natural and proved

very popular, resulting in * (anything) as a service [3], which serves as a general term for products that free the end-user from performing a demanding task, and instead offloading it to a professional service provider, thus freeing up customers' time and reducing the time to market.

Even though most service providers support granular billing policies for all of the above service offerings, and customers can be billed on a minute-by-minute or even per-second intervals, costs are incurred for simply leaving the infrastructure running, no matter the amount of visitors that it serves. Serverless computing is a recent paradigm shift that aims to overcome these issues, while making the application development process even simpler for developers. The major question faced by developers is no longer "where to deploy", but instead "how to create" the application, focusing foremost on its features. Serverless is comprised of Function as a Service (FaaS) and Backend as a Service (BaaS), and despite its name, it still relies on underlying servers for hosting the workload and data processing. However, compared to the other *as a service* approaches, it provides an even greater abstraction layer to the developers, who no longer have to think in terms of the infrastructure, resource requirements, or even scaling, and can instead focus on writing granular functions with a well defined role, and integrating their functionality to achieve more complex systems or applications. Thus, the main point of function as a service, and serverless computing in general, is to allow the developer to write functional code with a well defined task, using the desired programming language, which can then be uploaded and hosted as an atomic unit on a provider's infrastructure. This FaaS approach combined with common backend functionality such as databases, or message queues which are offered as a service as part of BaaS offerings, accessible through provider-defined APIs, unburdens users from any server management. Furthermore, by billing per function invocation, and allowing function instances to be scaled down to zero replicas when not being utilized, customers are billed only for the time that the function is active, while having access to seamless scalability, monitoring, and security features. The serverless approach is beneficial to service providers as well, since it advances the ever-present ideal of executing more workload on the same amount of resources. By transferring the responsibility for resource dimensioning away from the customers, service providers are better able to manage their computing capacity, utilized resources, as well as power usage.

The first public FaaS offering dates to 2015, when Amazon AWS introduced its Lambda computing service [4], aimed primarily at web developers. Others quickly followed [5–7] by introducing competing services inspired by the initial success and the potential benefits that the serverless approach might unlock. Open-source FaaS solutions are widely popular as well [8–10], and there are even cases where commercial service providers have either based their FaaS offerings on one of the existing open-source solutions, as is the case with IBM and OpenWhisk [11], or have open-sourced either in part, or completely the underlying components of their FaaS architecture [12], contributing to the open-source community, and thus directly investing in the serverless ecosystem.

Web development is not the only area where serverless computing unlocks interesting new opportunities. Another research area which has seen an enormous growth in recent years is the Internet of Things (IoT). Serverless for IoT would be particularly beneficial as a result of the inherently real-time and event-based workload of these systems [13,14]. However, IoT faces a different set of challenges in comparison to the typical client-server applications that were the primary targets for the initial serverless push. Even though the cloud has been utilized to a great extent in IoT scenarios, offering endless computing capacity, and data storage, means of actually transferring the data in an acceptable time-frame, without much delay, have always proved a challenge. While the cloud is an excellent choice for applications that run at human perception speed and response times of hundreds of milliseconds or even seconds are acceptable, optimizations have to be made to meet requirements for real-time IoT applications, running at machine perception speed [15]. This latency and network capacity problem will become even more pronounced with the advent of billions of new IoT devices that will find their way in our lives. Moving

the computing capacity towards the devices that actually generate the data is one of the solutions attracting great research interest. Edge computing reduces the network latency by allowing time-sensitive computations to be executed on compute infrastructure close to the data sources and can be seen as the missing piece to bring the simplicity of serverless computing to the event driven IoT world. Utilizing serverless edge computing transforms the previously utilized ship-data-to-code paradigm, which incurred high network latency and transmission costs, to a ship-code-to-data paradigm [16]. Furthermore, by initially preprocessing the data at the edge, not only can network bandwidth be saved and faster response time obtained, but compliance with data protection laws can be ensured as well. In this manner, customer data can be anonymized closer to the data source, in the same jurisdiction before being shipped to the cloud for long term storage and aggregation.

Many infrastructure providers have adapted their service offerings to include serverless products aimed at the network edge, such as AWS Greengrass [17], and Azure IoT Hub [18], bringing the associated benefits such as fast development, easy deployment, and seamless scalability to this part of the network. A number of open-source initiatives are also present, either adapting the existing open-source serverless platforms for the network edge, or starting from a clean slate, without any pre-existing technical debt, and developing entirely new solutions. While there is a perpetual discussion of centralized versus decentralized architectures, and the cycle seems to reverse itself during the years, serverless at the edge is still a novel research area with many outstanding issues left to be resolved.

The aim of this review paper is to describe and examine the current state of serverless research in relation to IoT and outline the open issues. Throughout the paper we use the widely accepted definition for serverless computing as introduced before, combining the function as a service, and backend as a service offerings. When moving to the edge, we use the term serverless edge computing to refer to such serverless workloads that can be executed either on the data generating devices themselves, or on infrastructure deployed in their vicinity.

The rest of the paper is structured as follows: in Section 2 we present related research papers to this topic and how they have tackled the associated problems. In Section 3 we present the research method that we have used, outlining the searching procedure, inclusion criteria for papers, as well as the analysis and classification processes. In Section 4 we describe the results, showcasing the developed categorization framework and outlining the state-of-the-art research for applying serverless computing to the edge of the network in an IoT context. We then proceed with Section 5 where we analyze the current trends and offer a discussion regarding open issues and threats to validity. We conclude with Section 6, where we summarize our research findings.

2. Related Work

Serverless computing is an active research topic which has attracted a noticeable research interest in recent years with a large number of both primary and secondary literature. The majority of this work is focused on serverless computing in the cloud, categorizing it as an emerging technology with potentially great impact to various fields and use-cases in the future.

Varghese et al. [19] argue that with further advancements to the serverless paradigm, it can become a viable alternative for many more applications, including IoT ones which are primarily event driven. The authors of [13] share this vision for serverless computing, classifying it as the driving force behind sensor networks at the edge in the future, together with the help of blockchain and artificial intelligence (AI). The large applicability of this new paradigm is evident even now, with vastly different use-cases available today, such as the ability to run JavaScript serverless functions on provider edge infrastructure, offering faster response time to web users across the globe [20]. Other areas that might benefit from serverless are further discussed by Shafiei [21] et al. and Hassan et al. [22], including

real-time collaboration and analytics, video processing, scientific computing, serving of machine learning models, file processing, smart grid, information retrieval, and chatbots.

By leveraging the effortless scalability that it offers, serverless computing can also be used for on-demand data processing and execution of resource intensive tasks which can be sped up by parallelly executing the same function on various compute nodes, where each instance would work on a smaller partition of the original data. Buyya et al. [23] drive this concept even further, describing serverless pipelines comprised of multiple functions chained together with the aim of modeling complex data analysis workflows. Real world examples are already available in this case as well [24,25]. The data processing does not need to take place exclusively on serverless platform in the cloud, and instead can be migrated to the edge as well, optimizing bandwidth usage should the computing resources meet the required performance [26].

All these different workloads that have unpredictable load levels and need to cope efficiently with large increases in the number of requests emphasize the need for advanced resource allocation and scheduling algorithms that can better meet the FaaS quality of service (QoS) expectations during peaks [23]. A review of existing scheduling optimizations is offered in [27]. Even though it primarily focusses on the cloud, it is also relevant in network edge environments.

When it comes to the network edge, the authors of [28] argue that there are significant benefits to moving serverless computing to this part of the network, and that it should not be limited to the cloud environment only. The establishment of an edge–cloud continuum which would allow dynamic workload migration and be transparent to the end users would bring the best of both worlds, data preprocessing at the edge when reduced latency is needed, and the vast compute capacity of the cloud for further analysis and long term storage. Unfortunately, before establishing a true edge–cloud continuum, further research is needed into efficiency optimizations in terms of runtime environments, their performance at the edge, and the feasibility of on-the-fly data migration. Hellerstein et al. [16] outline all of the efficiency problems affecting first generation serverless implementations, such as the limited execution time of functions imposed by serverless platforms, slow first invocation of the functions, low performance of input/output (I/O) operations, and limited support for specialized hardware, such as graphics cards. Discussion about potential solutions to the initial start up delay is offered by Kratzke et al. in [29], while reviewing cloud application architectures. Apart from comparing the advantages and disadvantages of serverless, the utilization of unikernels is proposed as a more lightweight runtime environment for serverless function execution. However, in order to effectively test any performance improvements, adequate and standardized benchmarks are needed which would be capable of cross platform execution. The authors of [30] provide a review of existing efforts made to benchmark FaaS platforms.

Real-world serverless platforms that are ready to be used also play an important role in the serverless adoption across its different realms of usage, and they are responsible for implementing all the other advancements in terms of security, scheduling, and efficiency in a comprehensive, ready to use package. Bocci et al. provide [31] a systematic review of serverless computing platforms, focusing on supported languages, models, and methodologies to define FaaS orchestrations. Special attention is also given to security issues, but single node serverless platforms are purposefully excluded. In our opinion, even though not natively scalable, single node platforms are still a valuable resource and can act as a guidance in relevant platform development trends. In a future work they can be expanded to encompass multiple nodes or can serve as an inspiration to other platforms by repurposing individual components. Additional analysis, but in a wider context, reviewing general features of existing popular serverless edge platforms is also available in [20], which can aid the decision making process when choosing a new serverless solution for the network edge.

Even though there are research papers that deal with serverless security and evaluate isolation levels of the various platforms available today [31], the analysis of Stack

Overflow [32] questions related to FaaS products suggests that developers rarely concern themselves with such topics, focusing more on the implementation and functional aspects of their applications instead. Still, many serverless platforms mandate strong runtime isolation between different serverless functions, in part mitigating such security concerns, albeit leading to reduced performance, additional function non-portability, and vendor lock-in [22].

In conclusion, multiple reviews have identified serverless computing as an emerging technology with prospects of being utilized in a variety of different contexts, including IoT. However, to the best of our knowledge, no comprehensive review exists focusing primarily on serverless edge computing from an IoT perspective. In our opinion, IoT is not just another use-case for this new paradigm, instead it is the killer application with a great potential, should the identified open issues be solved.

3. Research Method

In this section we first define the main aim of our systematic review and then proceed to explain in detail the undertaken steps for searching, classifying, and analyzing the relevant papers. The applied research method closely follows the guidelines for systematic mapping studies by Petersen et al. presented in [33].

3.1. Research Aim

The aim of this review paper is to determine the current state-of-the-art research for applying serverless computing to IoT workloads. To do so, we first examine the range, direction, and nature of current research in this subject area related to applying function as a service or backend as a service in an IoT environment. We then proceed to create a classification framework for serverless computing at the network edge, derived by analyzing relevant papers to this topic, and through this framework determine open issues and research gaps, with a focus on scenarios in which serverless computing is applied to resource constrained environments. Apart from identifying future research opportunities, this categorization can also aid new researchers who look for an introduction to the subject area by presenting the recent research on a given subtopic. A detailed explanation of each step performed to derive the classification framework is available in the subsections below.

3.2. Search

We have used 6 different databases for the initial search of relevant articles. The databases that we have selected are: IEEEXplore (https://ieeexplore.ieee.org/Xplore/home.jsp (accessed on 3 September 2021)), ACM Digital Library (https://dl.acm.org/ (accessed on 3 September 2021)), Arxiv (https://arxiv.org/ (accessed on 3 September 2021)), Google Scholar (https://scholar.google.com/ (accessed on 3 September 2021)), Springer Link (https://link.springer.com/ (accessed on 3 September 2021)), and Science Direct (https://www.sciencedirect.com/ (accessed on 3 September 2021)). The database selection decision was based on past experiences by other authors, and public recommendations [34,35]. We have considered all returned articles for inclusion, and have stopped searching once all results have been exhausted [36]. The following searching criteria were applied to article titles, abstracts, and author-provided keywords:

- Studies containing the keywords: “serverless” or “faas” or “function-as-a-service” or “function as a service” or “baas” or “backend-as-a-service” or “backend as a service” AND
- Studies containing the keywords: “IoT” or “internet of things” or “internet-of-things”

Multiple variants of the same search term were provided to account for difference in spelling and the use of abbreviations. In cases where case sensitivity was enabled by default, it was disabled manually, to mitigate any capitalization variations in the spelling of the abbreviations. Table 1 provides more details in terms of the utilized search query, as well as the number of returned results by each source. Grey literature is purposefully omitted from this review since it is not peer reviewed and wider industry trends can

be captured from the included articles themselves, taking into account the high level of interest and number of published items on this topic.

We have avoided more specific keywords relating to the analyzed subject area such as: *edge computing*, *sensors*, *service architecture*, *service oriented architecture*, or *sensor networks* because we have determined that mandating their presence together with the different variations of *IoT* and *serverless* significantly decreased the number of returned results. We have instead opted for a more laborious, albeit more precise analysis process where a general search query was used, and then resorted to individual systematic analysis of the content, as described in the subsections that follow. This approach allowed us to manually determine the relevancy of each entry to serverless edge computing in an IoT context, minimizing the threats to validity imposed by the accuracy of the original keyword categorization. This has led to an inclusion of additional relevant entries which would have been excluded, should more specific terms had been used. Finally, by not requiring the explicit presence of *edge* as a key classifier, we have avoided the problem where authors frequently describe serverless issues in a wider context, not necessarily mentioning or categorizing them as edge-related problems, which in this case would have led to their omission. Nonetheless, in our opinion, a large number of these issues are indeed applicable to the network edge as well, and we deem that they provide a valuable contribution to this review, so they have been included after careful consideration for their relevancy, respectively.

Table 1. Database results and search query.

Database	Results	Accepted	Query
IEEEXplore	77	29	(“serverless” or “faas” or “function as a service” or “function-as-a-service” or “baas” or “backend-as-a-service” or “backend as a service”) and (“iot” or “internet of things” or “internet-of-things”)
ACM	27	14	
Arxiv	10	2	
Google Scholar	45	6	
Springer	56	3	
Science Direct	2	0	

3.3. Study Selection and Quality Assessment

The keyword search using the query shown in Table 1 across all of the selected databases yielded 217 results. However, attention must be paid to the discussion above regarding the choice of keywords and their prospective effect on the final results. The introduction of the *AND (“edge” OR “edge computing”)* condition would have reduced the number of results from 217 to 162, whereas the introduction of *AND “sensors” OR ”sensor networks“* would have produced 92 entries. Mandating the presence of *AND (”service” OR “service architecture”)* in the existing search query would have resulted in the fewest number of entries, only 18. These results would have had great impact on the number of accepted papers after the selection process as well. The presence of *edge computing* would have reduced the number of papers by 12, *sensors* by 42, and *services* by 51.

The criteria for considering a given paper for further analysis were:

- English language conference papers, journal papers or scientific magazine articles;
- Publish date between 1 January 2015 and 1 September 2021;
- Full-text accessible to the authors of this paper;
- Clear relation to serverless computing in an IoT context at the network edge.

The initial 217 results were narrowed down to 206 after excluding duplicates and non-English titles. These were then further analyzed by their titles, abstracts, and in cases of ambiguity, their conclusions as well. This analysis led to a discovery of one paper that had a title in English, but the full-text itself was written in a foreign language, thus bypassing the initial language filter. Another 30 results represented incompatible media, such as: bachelor theses, master theses, doctoral theses, book chapters, or books. Additional 4 papers had to be excluded because even though their content was relevant, they have not underwent a formal review process, and were published only as preprints, thus being

grey literature. An interesting phenomenon was the amount of results that did match the initial search terms, but were not relevant to the topic of interest. This was because the various abbreviations such as "FaaS" or "BaaS" have been used in a different context, such as "blockchain as a service". Their removal reduced the number of results by 21. A number of results did not focus on serverless computing in either an IoT, sensor network, service architecture, or edge context, only briefly mentioning some of the search terms, leading to the exclusion of 79 entries. Finally, even though 6 papers did pass the title and abstract screening, we were not able to obtain their full-text due to our lack of appropriate publisher access, and them not being open-access, which led to their exclusion. After these selection activities, 141 entries were excluded in total, while leaving 65, which further underwent a full-text analysis. The full-text analysis resulted in the exclusion of 11 papers: 6 as a result of not being relevant to the researched topic, which was only concluded after they have been completely read; 3 as a result of low quality and ambiguity; 2 since they were duplicates with other already included papers, bearing different titles, while keeping large portion of the content unchanged. During this full-text analysis, 10 additional papers were identified by applying the snowballing technique [37] scouring the references of the read papers for additional relevant content. All of these activities resulted in the final acceptance of 64 records, which were further classified according to the methodology described in Section 3.5.

Figure 1 shows a graphical representation of the quality assessment process, providing detailed information about each undertaken step, and the resulting changes in terms of the number of accepted records.

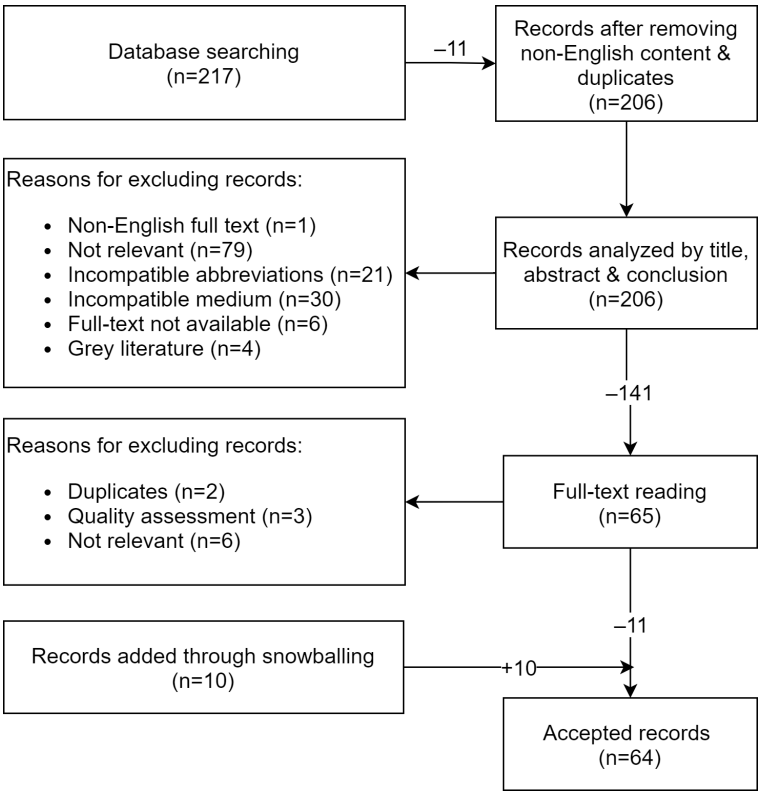

Figure 1. Number of included and excluded records during study selection.

3.4. Data Extraction

To aid the classification effort and to allow us to develop summary statistics for the accepted entries, we have manually maintained a list of parameters, containing the information described in Table 2.

Table 2. Extracted parameters through database searching and full-text reading.

ID	Name	Description
1	Paper ID	Sequential number of the entry
2	Source	The database containing the entry
3	Type	Article type (Conference, Journal, Magazine, Thesis, Other)
4	Venue	Name of the publication where the entry is published
5	Publication year	Publish year of the entry
6	Name	Full name of the entry
7	Open-Access	Whether the publication supports open-access (True/False)
8	DOI	DOI of the entry, if applicable
9	A-Keywords	Keywords as specified by the authors
10	Keywords	Classification keywords, derived after full-text reading
11	Short Note	Structured short description for the acceptance/rejection of the entry
12	Description	Short free-text description of the article content
13	Full-text	Whether full-text is available (True/False)

Based on the gathered data in terms of publication media for the selected papers presented in Figure 2, it can be clearly seen that more records are published as conference papers, with journals following. The number of records published in magazines is the lowest, which is understandable, taking into account the wider target audience that they have.

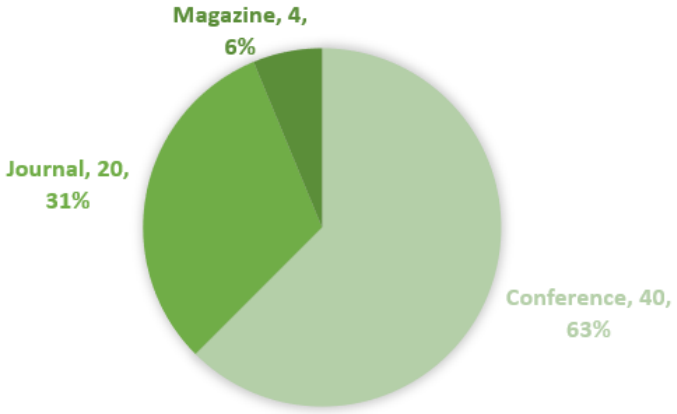

Figure 2. Publication medium for the selected papers.

3.5. Analysis and Classification

The 54 papers that were identified as relevant through the initial database search, as well as the 10 additional ones that were snowballed were fully read. We used the keywording technique as described in [38] to assign relevant and descriptive keywords to each paper, not taking into account the initial keywords specified by the authors. During the reading process, for each paper the following information was independently extracted:

- Classification notes—applicable keywords, as well as relevancy to other selected papers;
- Summary—paper summary, limited to 3 sentences, outlining the main topics;
- General notes—general information about the paper, used technologies, tackled problems;
- Technical notes—technical information regarding the research, detailed description and implementation details for the proposed solution;
- Citations—potentially relevant articles that have been cited by the analyzed paper, subject to further analysis.

Based on the acquired data, and the applied keywords, the relevant columns were filled in the table whose structure is given in Table 2. This aided the process of identifying the main topic of each paper, as well as discovering further related subtopics related to the main one, allowing us to perform more granular classification. Additional details about the derived classification framework are available in the results section below.

An extra classification criteria that was used, albeit not directly relevant to the paper content, was the support for open-access, and whether the analyzed dataset or implementation were publicly available, thus directly contributing to the cause of open-science. As providing open-access were classified all publications that had either a full-text or a preprint available on the official venue web page or at some other relevant location, such as the researchers' home institution, their personal profile pages, or on some pre-print database.

4. Results

4.1. Range and Direction of Existing Research

Figure 3 shows the yearly distribution of published articles, based on the data gathered though database searching and snowballing. It is evident that the idea of applying serverless computing in an IoT context is new and attracts an increasing interest with every passing year. While it is true that there is a slight decrease in the number of published papers in 2020, to an extent this can be attributed to external factors as well, such as the COVID-19 pandemic. The number of papers published so far in 2021 looks promising and by the end of the year it might prove to be the most popular for IoT serverless edge research yet.

Of the 64 analyzed papers, 50 in total, or 78 per cent offer some kind of an open-access, which is very encouraging. However, of these, only 16 were officially made available as open-access by their publisher or conference organizer, while the remainder were found in various online libraries and repositories of the researchers' host institutions, or personal portfolio web pages of the authors. Figure 4 presents the article distribution in terms of the location where they are available for download, while Figure 3 shows the number of open-access articles published per year.

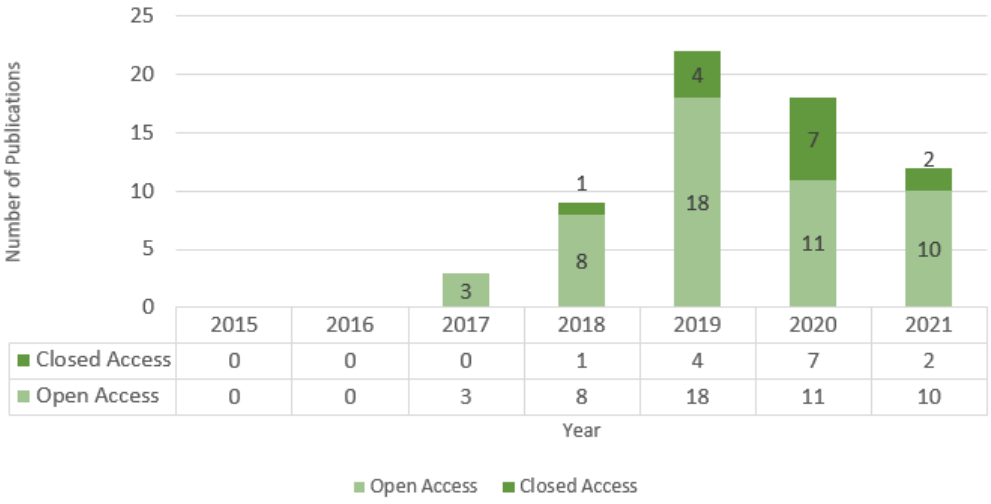

Figure 3. Number of publications per year.

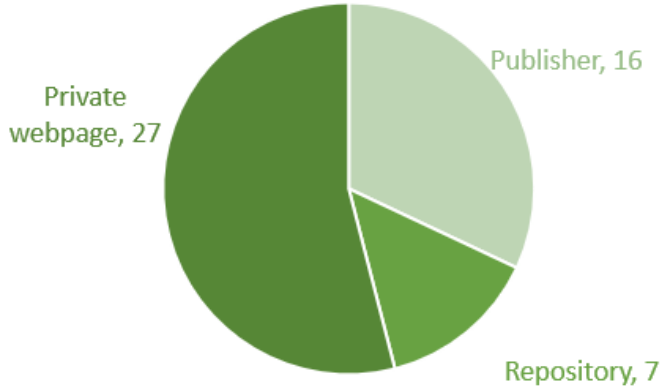

Figure 4. Availability of open-access papers.

4.2. Classification Framework

As a result of the full-text reading, the article summarization, and attached keywords to each included paper, we have derived a classification framework containing eight main topics, and 30 unique subtopics in total. This classification framework is presented in Figure 5, showcasing the relationship between the various categories and subcategories.

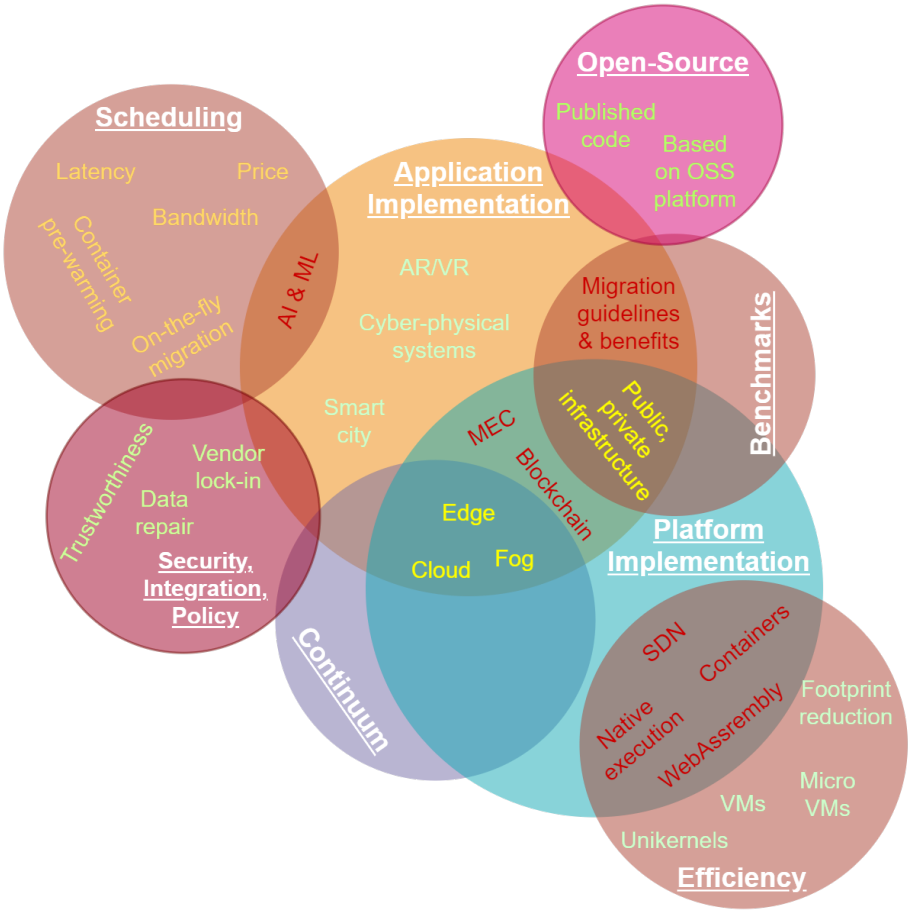

Figure 5. Derived IoT serverless categories and related subcategories.

Some of the subcategories are mapped to more than one category, such as: *edge*, *fog*, *cloud*, where they are present in the Platform Implementation, Application Implementation, and Continuum categories. In the case of platform and application implementation the distinction is clear, a given paper might discuss a new platform architecture capable of hosting various applications at either of these various locations in the network, or just a

single novel application hostable at either the edge, fog or cloud. On the other hand, the Continuum category is reserved for articles that offer a strategy for dynamic workload migration from one part of the network to another, exploiting their respective advantages, such as network latency in the case of the network edge or processing capacity for the cloud case. In this way, it is clear from a given paper's classification whether different execution locations are supported, albeit at the discretion of the administrator with no dynamic migration between them, or the workload execution location can be dynamically selected based on some rules and conditions.

Similarly, *containers*, *native execution*, and *WebAssembly* are shared between both the Efficiency and Platform Implementation categories. When discussing Efficiency, these subcategories relate to a novel approach or optimization made to these runtimes, potentially applicable to existing serverless solutions as well. On the other hand, in the Platform Implementation case, the subcategories are used simply for explaining the choice for an existing runtime architecture. Additionally, *AI & ML* is also present in both the Scheduling and Application Implementation categories. In the first case, AI is used in the process of workload scheduling, optimizing metrics such as latency, price or bandwidth and has no direct relation to the functionality of the instantiated applications whatsoever. However, in the second case, AI is simply part of the introduced application, aiding its use-case, and is not related to the runtime efficiency of the platform itself. A similar discussion can be made about the presence of *blockchain* and *MEC* in both Application Implementation and Platform Implementation.

Finally, even though there is a complete overlap between the subcategories of Benchmarking and Application Implementation, their meaning is once again very distinctive. Both *private infrastructure* and *public infrastructure* can be related to the execution location where an application can be run, describing whether a commercial service is required as in the case of public infrastructure, or the use of private, self-hosted infrastructure is also supported. Contrary to this, the same two subcategories present in the Benchmarking section relate to specific performance tests developed to evaluate the capabilities of the given infrastructure where they are performed, and thus have no direct use-case for end-users. *Migration guidelines & benefits* from the application perspective relates to tips and recommendations for how the serverless paradigm can benefit various applications, discussing the associated benefits. From the benchmarking perspective, however, the presence of this subcategory designates that the presented solution within the research article, such as an application, new platform, or an advanced scheduling algorithm is evaluated in terms of the performance that it offers in comparison to existing products.

4.3. Classification of Existing Literature

Using the previously described framework, all 64 papers were categorized according to the topics that they discuss. Each entry is rated on a scale from 0 to 3 in terms of how relevant it is to the given category. One primary category is assigned to each entry, denoted by underlining the three stars representing the rating. The other ratings are derived by counting the number of subtopics that the entry tackles within the given parent category, as per the relationships in Figure 5. The primary category selection process does not follow this rule—it is instead manually assigned based on the entry's content and discussion among the authors of this survey, and always has a star rating of three. Due to space constraints, papers that tackle only a single category are included in Table 3, while the remaining ones are part of Table 4, grouped by their primary category. Note that some categories are omitted from Table 3, since no paper representatives tackling solely that topic were present. We next provide an overview in terms of the content and tackled subcategories of the papers presented in Tables 3 and 4.

Table 3. Classification of papers discussing a single serverless issue.

Category Name	Discussed By
Application Implementation	[13,14,16,19,21–23,32,39–41]
Efficiency	[29]
Benchmarks	[42]

Table 4. Classification of Papers Discussing Multiple Serverless Issues.

Paper	A. Impl.	Eff.	Sched.	Bench.	P. Impl.	Cont.	SIP	OSS
[43]	★★★	★★☆	☆☆☆	★☆☆	★★☆	☆☆☆	☆☆☆	☆☆☆
[44]	★★★	★☆☆	☆☆☆	★☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆
[45]	★★★	☆☆☆	☆☆☆	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[46]	★★★	☆☆☆	☆☆☆	★☆☆	☆☆☆	☆☆☆	☆☆☆	★★☆
[47]	★★★	☆☆☆	☆☆☆	★☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆
[48]	★★★	☆☆☆	☆☆☆	★☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆
[49]	★★★	☆☆☆	☆☆☆	★☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆
[50]	★★★	☆☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆	★☆☆	☆☆☆
[51]	★☆☆	★★★	★☆☆	★☆☆	★★☆	☆☆☆	☆☆☆	★☆☆
[52]	★☆☆	★★★	☆☆☆	★☆☆	★★☆	☆☆☆	☆☆☆	☆☆☆
[15]	☆☆☆	★★★	★☆☆	★☆☆	★★☆	☆☆☆	☆☆☆	★☆☆
[53]	★★☆	☆☆☆	★★★	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[54]	★☆☆	☆☆☆	★★★	★☆☆	☆☆☆	★★☆	☆☆☆	☆☆☆
[27]	★☆☆	☆☆☆	★★★	☆☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆
[55]	☆☆☆	★★☆	★★★	★☆☆	☆☆☆	☆☆☆	☆☆☆	★★☆
[56]	☆☆☆	★☆☆	★★★	★☆☆	★★★	★★☆	☆☆☆	☆☆☆
[57]	☆☆☆	★☆☆	★★★	★☆☆	☆☆☆	★★☆	★☆☆	☆☆☆
[58]	☆☆☆	★☆☆	★★★	★☆☆	☆☆☆	★★☆	☆☆☆	☆☆☆
[59]	☆☆☆	☆☆☆	★★★	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[60]	☆☆☆	☆☆☆	★★★	★☆☆	★★☆	☆☆☆	☆☆☆	☆☆☆
[61]	☆☆☆	☆☆☆	★★★	★☆☆	★☆☆	★★☆	☆☆☆	☆☆☆
[62]	☆☆☆	☆☆☆	★★★	★☆☆	☆☆☆	☆☆☆	☆☆☆	★☆☆
[63]	☆☆☆	☆☆☆	★★★	★☆☆	☆☆☆	☆☆☆	☆☆☆	★☆☆
[64]	★★☆	☆☆☆	☆☆☆	★★★	☆☆☆	☆☆☆	☆☆☆	★☆☆
[65]	☆☆☆	☆☆☆	☆☆☆	★★★	☆☆☆	☆☆☆	☆☆☆	★☆☆
[66]	☆☆☆	☆☆☆	☆☆☆	★★★	☆☆☆	☆☆☆	☆☆☆	★☆☆
[67]	★★★	☆☆☆	★☆☆	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[68]	★★★	☆☆☆	★☆☆	★☆☆	★★★	★★★	☆☆☆	★★☆
[69]	★★☆	☆☆☆	★★☆	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[70]	★★☆	☆☆☆	★☆☆	☆☆☆	★★★	★★☆	☆☆☆	★☆☆
[26]	★★☆	☆☆☆	☆☆☆	★☆☆	★★★	★★★	☆☆☆	★★☆
[71]	★☆☆	☆☆☆	★★☆	★☆☆	★★★	★★★	☆☆☆	★☆☆
[72]	★☆☆	☆☆☆	★☆☆	★☆☆	★★★	★★☆	☆☆☆	★☆☆
[73]	★☆☆	☆☆☆	☆☆☆	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[74]	☆☆☆	☆☆☆	★★☆	★☆☆	★★★	★★★	★☆☆	☆☆☆

Table 4. *Cont.*

Paper	A. Impl.	Eff.	Sched.	Bench.	P. Impl.	Cont.	SIP	OSS
[75]	☆☆☆	☆☆☆	★☆☆	★☆☆	★★★	★★★	☆☆☆	★☆☆
[76]	☆☆☆	☆☆☆	★☆☆	★☆☆	★★★	★★☆	☆☆☆	★☆☆
[77]	☆☆☆	☆☆☆	★☆☆	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[78]	☆☆☆	☆☆☆	☆☆☆	★☆☆	★★★	★★☆	☆☆☆	★☆☆
[79]	☆☆☆	☆☆☆	☆☆☆	★☆☆	★★★	★★☆	☆☆☆	☆☆☆
[80]	☆☆☆	☆☆☆	☆☆☆	★☆☆	★★★	☆☆☆	☆☆☆	★★☆
[81]	☆☆☆	☆☆☆	☆☆☆	★☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[82]	☆☆☆	☆☆☆	☆☆☆	☆☆☆	★★★	★★☆	☆☆☆	☆☆☆
[20]	☆☆☆	☆☆☆	☆☆☆	☆☆☆	★★★	☆☆☆	★☆☆	★☆☆
[83]	☆☆☆	☆☆☆	☆☆☆	☆☆☆	★★★	☆☆☆	☆☆☆	★☆☆
[84]	★☆☆	☆☆☆	★★☆	★☆☆	☆☆☆	★★★	☆☆☆	☆☆☆
[28]	★☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆	★★★	☆☆☆	☆☆☆
[85]	☆☆☆	☆☆☆	☆☆☆	☆☆☆	★★★	★★★	☆☆☆	☆☆☆
[31]	★☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆	☆☆☆	★★★	★☆☆
[86]	☆☆☆	☆☆☆	☆☆☆	★☆☆	★★☆	★★★	★★★	★★☆
[87]	☆☆☆	☆☆☆	☆☆☆	★☆☆	★★☆	☆☆☆	★★★	★☆☆

Abbreviations: **A. Impl**—Application Implementation; **Eff.**—Efficiency; **Sched.**—Scheduling; **Bench.**—Benchmarks; **P. Impl**—Platform Implementation; **Cont.**—Continuum; **SIP**—Security, Integrity, Policy; **OSS.**—Open-Source Software.

Application Implementation is the topic with most published papers in the reviewed period, with 19 entries in total or 30 per cent of all analyzed papers, as per Tables 3 and 4. Even though serverless computing was initially targeted primarily at web developers to simplify the development process, recently novel use-cases have emerged demanding lower latency and the deployment of edge infrastructure. Serverless computing is especially suitable for event-driven scenarios [14] involving IoT devices. One such area is cyber-physical systems, where a successful implementation of a power grid monitoring solution capable of dynamically responding to unpredicted events and balancing supply according to current demand has been described [41]. Smart city applications like monitoring garbage disposal [39], energy usage optimization [47,48], or improving public transportation systems [49] have also been discussed. However, serverless computing at the edge can also be utilized without a dedicated infrastructure, by harvesting the computing power of nearby devices instead. Using portable JavaScript runtimes, the authors of [43] have created a system which can offload processing to devices in the close vicinity for an AR/VR application [45]. Reports on converting existing serverful applications to a serverless architecture have also been published [46], with the intention of driving a higher adoption and outlining the benefits. Nonetheless, a recent survey on Stack Overflow, analyzing questions related to the topic of serverless computing [32], shows that the majority of encountered problems by developers are related particularly to application implementation. To solve this and to drive a higher level of adoption, formal guidelines should be published educating developers about the limitations of the network edge.

Efficiency improvements have been made to serverless edge platforms, trying to overcome the fact that existing serverless platforms developed initially for environments with plentiful resources are not a good fit for the resource constrained edge. The focus of this research area is finding alternative runtime environments that do not rely on containerization, thus avoiding the slow start-up incurred during the first invocation of a given function. A promising option is WebAssembly [52] with its portability and fast function start-up time [51], albeit further work is needed on improving the execution speed of the

deployed functions. Alternatives include the introduction of unikernels, a surprisingly under researched topic today, and the development of micro virtual machines [29], with some implementations already being open-sourced [88].

Scheduling algorithms optimally determining where and when a given function needs to be executed [53] are another way in which the cold-start problem [63] typical for container based serverless systems can be overcome, apart from introducing new runtime environments. Further optimizations in terms of reduced latency [59], bandwidth [69], and cost [58] have also been described, depending on the use-case and priorities of the administrators. Recently, efforts have been made to develop alternative scheduling systems to popular serverless platforms, utilizing machine learning algorithms [55,62] with the aim of analyzing historical function metric data and adapting the scheduling decisions accordingly. However, scheduling decisions are not limited only to the initial placement of the functions, but can also be extended to live function migration, alleviating unexpected memory pressure, or dynamically pausing and then resuming function execution on the same node while waiting for a synchronous operation to complete [60].

Benchmarks can be used to measure and compare the performance of different efficiency optimizations, scheduling algorithms, and complete serverless platforms in terms of other alternatives. Multiple benchmarking suites have been proposed [64,66] to this effect, utilizing a number of different tests, ranging from purpose built microbenchmarks targeted at measuring raw compute, network, or I/O performance, to all encompassing serverless applications. Unfortunately, lacking a unified abstraction layer that would be supported across all serverless platforms, these benchmarking suites are limited in the number of environments that they support. The addition of a new supported platform is often a tedious process as a result of the different provider application programming interfaces (APIs) available or runtime restrictions. Researchers have attempted to solve this issue by open-sourcing their code and relying on the community to introduce support for popular solutions. This leads to problems where the majority of authors do publish performance results about their implementation, but they are hard to verify, replicate, and compare to other platforms that have not been included in their analysis.

Platform Implementations have decided to adopt the API interfaces of popular cloud-based serverless products [78] with the aim of solving the issue of vendor lock-in and cross-platform incompatibility, thus making all existing functions automatically compatible with the newly presented solution. The development of new serverless edge platforms using existing commercial solutions is not uncommon, and is mostly focused on features that are lacking by default. The authors of [79] extend the AWS Greengrass software to be able to automatically fetch AWS Lambda functions for local execution when there is such demand. This behavior is possible since both AWS Lambda and Greengrass support the same function languages and constructs. Others have instead focused on improving existing open-source serverless platforms and optimizing them for the network edge [67,77]. AI, as one popular use-case of serverless functions, has also incentivized the development of specialized platforms satisfying its requirements [70,72]. However, by offering easy-to-use interfaces, and integration with the cloud, it is possible to leverage the proximity of the edge not only for reduced latency, but also for increased privacy, to preprocess data that would ultimately be analyzed and aggregated in the cloud. This is especially useful for research studies that gather various sensor data containing personally identifiable information, which needs to be anonymized first [26]. A persistent issue faced by all serverless edge platforms is how to connect with the end-users and end-devices who would invoke the available functions. With the continuous improvement in mobile network infrastructure and introduction of new generations of connectivity, the idea of collocating compute infrastructure with providers' base stations becomes a reality. The concept of mobile edge computing (MEC) [73], coupled with serverless can play an important role both for service providers and end-users alike [69]. By deploying serverless platforms capable of offering FaaS to prospective customers [53], operators can rent their in-place edge infrastructure,

while enabling additional IoT use-cases without the need for standalone deployment of new compute or networking equipment.

Continuum describes a hierarchical execution environment comprised of edge, fog, and cloud resources, working in tandem with dynamic workload migration between them. Many serverless edge platforms are not limited to only running at the edge, instead their aim is to develop versatile products that can be run anywhere, at either the edge, fog, or cloud, offering the same function syntax across the whole network [68,75]. When coupled with intelligent scheduling algorithms that can automatically determine the optimal execution location, as opposed to relying on the administrator to make the right decision [85], a true edge-fog-cloud continuum [28] can be established. Attempts have been made to offer such continuums even for commercial products with both cloud and edge counterparts, but not providing a native integration between them [84].

Security, Integrity, Policy is one of the least researched serverless edge topics, even though it is of paramount importance, especially in multi-tenant environments where multiple customers share the same infrastructure for function execution, as depicted by Table 4. Careful attention is warranted to the level of isolation that the chosen runtime offers, as well as the behavior for serving new requests. Aiming to reduce the cold-start latency, many platforms forgo per-invocation isolation, instead reusing the same environment without clearing it and spawning a new one, leaving leftover files or processes [87]. Another problem with serverless execution in scenarios where multiple functions are chained together in a pipeline is the prospect of intermediate data corruption which would require the repeated execution of the whole pipeline to alleviate the problem. Lin et al. [86] describe an append-only system storing function inputs and results, allowing granular reexecution of downstream functions, without affecting the upstream ones in the pipeline, thus minimizing the effects of any data corruption as well as reducing the time needed for repair, with low performance overhead.

5. Discussion

It is evident that there is a large interest in employing serverless computing at the edge of the network, with various research topics tackled. Figure 6 shows the primary category distribution of the selected papers, with the inclusion of review papers as well. The x-axis represents the percentage of all papers published in the given year. The y-axis represents the percentage of all papers which have a connection to the given category. Please note that the numbers on the y-axis do not add up to 100 per cent because one paper can be relevant to multiple categories, as shown in Table 4. The color coding of the bubbles relates to the open-access policy of the papers, with green denoting that all associated papers within the given category are open-access and yellow representing mixed policy—both open-access and closed-access are present.

The majority of analyzed papers (67 per cent) have been classified as offering some level of benchmarking and comparison with existing solutions, which is understandable taking into account the high representation of both platform implementation (50 per cent) and application implementation (56 per cent), two categories where performance discussion and comparison is commonplace. Open-source is also a highly popular category, accounting for 48 per cent of all entries, with many papers either basing their work on existing open-source code or publishing their implementation in turn. On the other hand, very few papers deal with the security aspects of using FaaS platforms, the integrity of the analyzed and produced data, or with policy in general, such as avoiding vendor lock-in problems.

A topic that is under active research especially is the past few years is the establishment of a true edge-fog-continuum. However, additional advancements are needed in the area of intelligent scheduling algorithms and efficiency optimizations before such erasure of network boundaries can become commonplace. We present a list of open issues which we deem need to be solved in order for serverless computing to achieve an even wider adoption at the edge of the network:

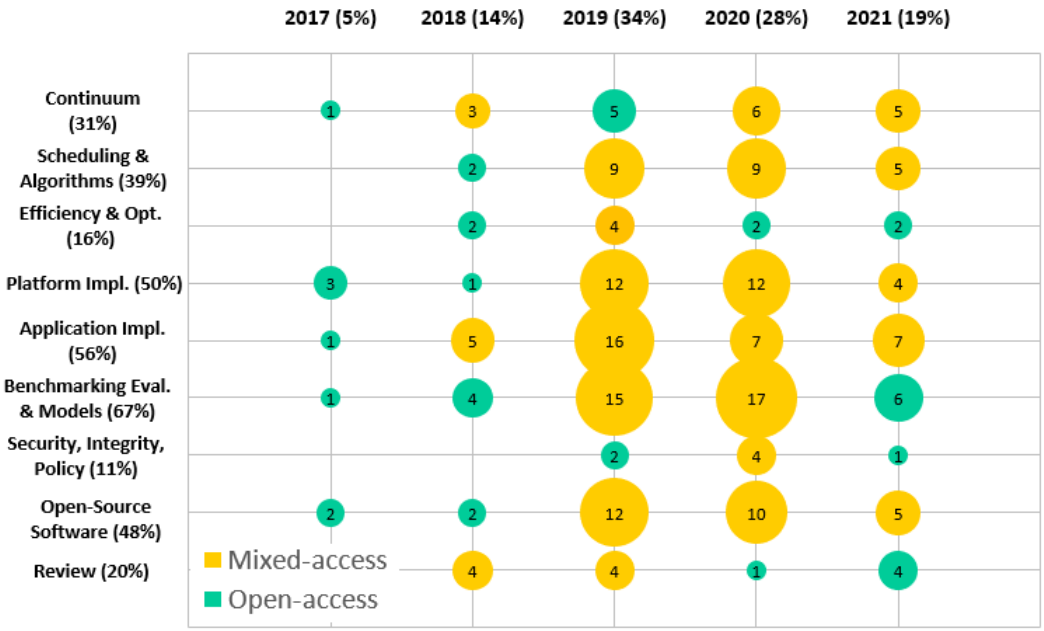

Figure 6. Primary category distribution per year and open-access classification.

- Development of efficient scheduling algorithms that are capable of handling high volumes of function instantiations and deletions in short amounts of times, across different infrastructures, providing an edge–cloud continuum.
- Safe migration of running serverless functions across different environments, allowing for better resiliency and cost effectiveness.
- Performance improvement of existing serverless function runtimes to make them suitable for resource constrained devices located at the edge, and migration away from containerization technologies altogether, by adopting more lightweight alternatives, such as WebAssembly, and unikernels. However, further research is needed in terms of execution speed performance, and development of easy-to-use solutions, which would in turn lead to an increase in popularity.
- Eliminating the cold start problem associated with the dynamic nature of serverless functions and the scale-to-zero feature.
- Eliminating vendor lock-in, as a prerequisite for a wider adoption, as well as constructing more elaborate hierarchical infrastructures, which would include both commercial and private elements. This is also the main issue preventing the establishment of cross-platform function marketplaces where users can freely exchange existing serverless functions.
- Improvements to serverless function security and isolation, especially in multi-tenant environments. Even though security is of great concern for resource constrained IoT devices, innovative ways in which greater function isolation can be established, without resulting in increased execution or start-up time are needed. Exhaustion of resources as a result of ever more present denial of service attacks is also an open issue, especially for serverless functions utilizing a commercial platform, where billing is done depending on the number of invocations and the total runtime. An increase in denial of service attacks aiming to take a given service offline by incurring large monetary cost to its owners is not excluded.
- Improvements to function chaining, and shift to asynchronous execution where possible. One of the main benefits of serverless, the scale-down-to-zero feature, cannot be realized when a chain of subsequent functions is executed in a serial manner, all waiting for an intermediate result before they can be terminated. Not only does this lead to less efficient resource utiliziation, but also to increased cost, as a result of each function being billed independently, even when it is stuck waiting on another one.

- Lack of comprehensive guidelines for development of new serverless IoT applications, or migration of existing ones, taking into account the specifics of this new paradigm.
- Support for hardware-acceleration and utilization of specific hardware, essential for artificial intelligence and video processing workloads.

Finally, limitations to the applied approach must be stated. We have searched six different databases, evaluating the returned results for relevancy to the given topic. This limits the exposure to content not indexed by the selected sources, but we have tried to mitigate this issue by performing a forward snowballing of papers referenced within the obtained results. Furthermore, our exploration focused only on English language papers. In terms of the categorization process, the included papers were classified solely on their actual content, without confirming the feasibility or accuracy of the outlined results, instead relying on the peer review that they have underwent as part of their submission process. For this reason, grey literature has been purposefully omitted, as it is not peer reviewed. We strove to eliminate individual bias during the categorization process by cross checking the decisions made and discussing differences until a consensus was reached. When no consensus was possible, majority voting was performed.

6. Conclusions

Using a systematic mapping approach, we have reviewed the state-of-the-art research in terms of serverless computing in an IoT context, applied at the edge. By searching six popular paper databases, we have identified 64 papers relevant to the topic, from an initial pool of 217 results. After performing a full-text analysis on the accepted entries, we have identified eight areas in which existing serverless edge research is focused, many of them intertwined with one another, blurring the lines between them. These areas are: (i) application implementation; (ii) efficiency; (iii) scheduling; (iv) benchmarks; (v) platform implementation; (vi) continuum; (vii) security, integration, policy; (viii) open-source software.

Using the derived categories in the analysis of the selected papers, we have identified an increasing interest in applying serverless computing at the edge of the network especially in the past three years, with 81 per cent of all included papers published within this time frame. Another interesting trend is the move towards providing support for open-access for recent research, with 50 of the included papers available in this manner either officially through their publishers, as pre-prints, or uploaded to an institutional repository.

Even though IoT has the potential to become the killer use-case for serverless computing at the network edge, nonetheless, a number of discussed issues are unresolved and a suitable solution needs to be found before a wider adoption can be made possible.

Author Contributions: Conceptualization: S.F., V.T. and V.K.; methodology: S.F., V.T. and V.K.; software: V.K. and S.F.; validation: S.F. and V.T.; formal analysis: V.K. and S.F.; investigation: V.K.; writing—original draft preparation: V.K.; writing—review and editing: S.F. and V.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia under the "SCAP" project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Buyya, R.; Yeo, C.S.; Venugopal, S.; Broberg, J.; Brandic, I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Gener. Comput. Syst.* **2009**, *25*, 599–616. [CrossRef]
2. Mell, P.; Grance, T. *The NIST Definition of Cloud Computing*; Technical Report NIST Special Publication (SP) 800-145; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2011. [CrossRef]
3. Duan, Y.; Fu, G.; Zhou, N.; Sun, X.; Narendra, N.C.; Hu, B. Everything as a Service (XaaS) on the Cloud: Origins, Current and Future Trends. In Proceedings of the 2015 IEEE 8th International Conference on Cloud Computing, New York, NY, USA, 27 June–2 July 2015; pp. 621–628. [CrossRef]

4. AWS Lambda—Serverless Compute—Amazon Web Services. Available online: https://aws.amazon.com/lambda/ (accessed on 27 May 2021).
5. Azure Functions Serverless Compute | Microsoft Azure. Available online: https://azure.microsoft.com/en-us/services/functions/ (accessed on 27 May 2021).
6. IBM Cloud Functions-Overview. Available online: https://www.ibm.com/cloud/functions (accessed on 27 May 2021).
7. Cloud Functions. Available online: https://cloud.google.com/functions (accessed on 27 May 2021).
8. Apache OpenWhisk Is a Serverless, Open Source Cloud Platform. Available online: https://openwhisk.apache.org/ (accessed on 27 May 2021).
9. Available online: https://www.openfaas.com/ (accessed on 27 May 2021).
10. Kubeless. Available online: https://kubeless.io/ (accessed on 27 May 2021).
11. Getting Started with IBM Cloud Functions. Available online: https://cloud.ibm.com/docs/openwhisk?topic=openwhisk-getting-started (accessed on 27 May 2021).
12. Azure/Iotedge. Available online: https://github.com/Azure/iotedge (accessed on 27 May 2021).
13. Gill, S.S.; Tuli, S.; Xu, M.; Singh, I.; Singh, K.V.; Lindsay, D.; Tuli, S.; Smirnova, D.; Singh, M.; Jain, U.; et al. Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges. *Internet Things* **2019**, *8*, 100118. [CrossRef]
14. Aslanpour, M.S.; Toosi, A.N.; Cicconetti, C.; Javadi, B.; Sbarski, P.; Taibi, D.; Assuncao, M.; Gill, S.S.; Gaire, R.; Dustdar, S. Serverless Edge Computing: Vision and Challenges. In *2021 Australasian Computer Science Week Multiconference*; ACM: Dunedin, New Zealand, 2021; pp. 1–10. [CrossRef]
15. Gadepalli, P.K.; Peach, G.; Cherkasova, L.; Aitken, R.; Parmer, G. Challenges and Opportunities for Efficient Serverless Computing at the Edge. In Proceedings of the 2019 38th Symposium on Reliable Distributed Systems (SRDS), Lyon, France, 1–4 October 2019; pp. 261–2615. [CrossRef]
16. Hellerstein, J.M.; Faleiro, J.; Gonzalez, J.E.; Schleier-Smith, J.; Sreekanti, V.; Tumanov, A.; Wu, C. Serverless Computing: One Step Forward, Two Steps Back. *arXiv* **2018**, arXiv:1812.03651.
17. AWS IoT Greengrass—Amazon Web Services. Available online: https://aws.amazon.com/greengrass/ (accessed on 27 May 2021).
18. IoT Hub | Microsoft Azure. Available online: https://azure.microsoft.com/en-us/services/iot-hub/ (accessed on 27 May 2021).
19. Varghese, B.; Buyya, R. Next generation cloud computing: New trends and research directions. *Future Gener. Comput. Syst.* **2018**, *79*, 849–861. [CrossRef]
20. El Ioini, N.; Hästbacka, D.; Pahl, C.; Taibi, D. Platforms for Serverless at the Edge: A Review. In *Advances in Service-Oriented and Cloud Computing*; Zirpins, C., Paraskakis, I., Andrikopoulos, V., Kratzke, N., Pahl, C., El Ioini, N., Andreou, A.S., Feuerlicht, G., Lamersdorf, W., Ortiz, G., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 1360, pp. 29–40.
21. Shafiei, H.; Khonsari, A.; Mousavi, P. Serverless Computing: A Survey of Opportunities, Challenges and Applications. *arXiv* **2019**, arXiv:1911.01296. [CrossRef]
22. Hassan, H.B.; Barakat, S.A.; Sarhan, Q.I. Survey on serverless computing. *J. Cloud Comput.* **2021**, *10*, 39. [CrossRef]
23. Buyya, R.; Srirama, S.N.; Casale, G.; Calheiros, R.; Simmhan, Y.; Varghese, B.; Gelenbe, E.; Javadi, B.; Vaquero, L.M.; Netto, M.A.S.; et al. A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade. *ACM Comput. Surv.* **2019**, *51*, 1–38. [CrossRef]
24. Kubeflow. Available online: https://www.kubeflow.org/ (accessed on 28 September 2021).
25. Argo Workflows—The Workflow Engine for Kubernetes. Available online: https://argoproj.github.io/argo-workflows/ (accessed on 28 September 2021).
26. Risco, S.; Moltó, G.; Naranjo, D.M.; Blanquer, I. Serverless Workflows for Containerised Applications in the Cloud Continuum. *J. Grid Comput.* **2021**, *19*, 30. [CrossRef]
27. Adhikari, M.; Amgoth, T.; Srirama, S.N. A Survey on Scheduling Strategies for Workflows in Cloud Environment and Emerging Trends. *ACM Comput. Surv.* **2019**, *52*, 1–36. [CrossRef]
28. Bittencourt, L.; Immich, R.; Sakellariou, R.; Fonseca, N.; Madeira, E.; Curado, M.; Villas, L.; DaSilva, L.; Lee, C.; Rana, O. The Internet of Things, Fog and Cloud continuum: Integration and challenges. *Internet Things* **2018**, *3–4*, 134–155. [CrossRef]
29. Kratzke, N. A Brief History of Cloud Application Architectures. *Appl. Sci.* **2018**, *8*, 1368. [CrossRef]
30. Scheuner, J.; Leitner, P. Function-as-a-Service performance evaluation: A multivocal literature review. *J. Syst. Softw.* **2020**, *170*, 110708. [CrossRef]
31. Bocci, A.; Forti, S.; Ferrari, G.L.; Brogi, A. Secure FaaS orchestration in the fog: How far are we? *Computing* **2021**, *103*, 1025–1056. [CrossRef]
32. Wen, J.; Chen, Z.; Liu, Y.; Lou, Y.; Ma, Y.; Huang, G.; Jin, X.; Liu, X. An empirical study on challenges of application development in serverless computing. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 416–428. [CrossRef]
33. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.* **2015**, *64*, 1–18. [CrossRef]
34. The Best research DATABASES for Computer Science [Update 2019]. Available online: https://paperpile.com/g/research-databases-computer-science/ (accessed on 27 May 2021).

35. Dyba, T.; Dingsoyr, T.; Hanssen, G. Applying Systematic Reviews to Diverse Study Types: An Experience Report. In Proceedings of the First International Symposium On Empirical Software Engineering And Measurement (ESEM 2007), Madrid, Spain, 20–21 September 2007; pp. 225–234. [CrossRef]
36. Garousi, V.; Felderer, M.; Mäntylä, M.V. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol.* **2019**, *106*, 101–121. [CrossRef]
37. Wohlin, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, London, UK, 13–14 May 2014; pp. 1–10. [CrossRef]
38. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic mapping studies in software engineering. In *Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering*; BCS Learning & Development Ltd.: Swindon, UK, 2008; pp. 68–77.
39. Al-Masri, E.; Diabate, I.; Jain, R.; Lam, M.H.; Reddy Nathala, S. Recycle.io: An IoT-Enabled Framework for Urban Waste Management. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5285–5287. [CrossRef]
40. Pfandzelter, T.; Bermbach, D. IoT Data Processing in the Fog: Functions, Streams, or Batch Processing? In Proceedings of the 2019 IEEE International Conference on Fog Computing (ICFC), Prague, Czech Republic, 24–26 June 2019; pp. 201–206. [CrossRef]
41. Zhang, S.; Luo, X.; Litvinov, E. Serverless computing for cloud-based power grid emergency generation dispatch. *Int. J. Electr. Power Energy Syst.* **2021**, *124*, 106366. [CrossRef]
42. Gorlatova, M.; Inaltekin, H.; Chiang, M. Characterizing task completion latencies in multi-point multi-quality fog computing systems. *Comput. Netw.* **2020**, *181*, 107526. [CrossRef]
43. Salehe, M.; Hu, Z.; Mortazavi, S.H.; Mohomed, I.; Capes, T. VideoPipe: Building Video Stream Processing Pipelines at the Edge. In *Proceedings of the 20th International Middleware Conference Industrial Track*; ACM: Davis, CA, USA, 2019; pp. 43–49. [CrossRef]
44. Christidis, A.; Davies, R.; Moschoyiannis, S. Serving Machine Learning Workloads in Resource Constrained Environments: A Serverless Deployment Example. In Proceedings of the 2019 IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA), Kaohsiung, Taiwan, 18–21 November 2019; pp. 55–63. [CrossRef]
45. Baresi, L.; Filgueira Mendonça, D.; Garriga, M. Empowering Low-Latency Applications Through a Serverless Edge Computing Architecture. In *Service-Oriented and Cloud Computing*; De Paoli, F., Schulte, S., Broch Johnsen, E., Eds.; Springer International Publishing: Cham, Switzerland, 2017; Volume 10465, pp. 196–210.
46. Großmann, M.; Ioannidis, C.; Le, D.T. Applicability of Serverless Computing in Fog Computing Environments for IoT Scenarios. In *Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing Companion*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 29–34. [CrossRef]
47. Albayati, A.; Abdullah, N.F.; Abu-Samah, A.; Mutlag, A.H.; Nordin, R. A Serverless Advanced Metering Infrastructure Based on Fog-Edge Computing for a Smart Grid: A Comparison Study for Energy Sector in Iraq. *Energies* **2020**, *13*, 5460. [CrossRef]
48. Huber, F.; Mock, M. Toci: Computational Intelligence in an Energy Management System. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, 1–4 December 2020; pp. 1287–1296. [CrossRef]
49. Herrera-Quintero, L.F.; Vega-Alfonso, J.C.; Banse, K.B.A.; Carrillo Zambrano, E. Smart ITS Sensor for the Transportation Planning Based on IoT Approaches Using Serverless and Microservices Architecture. *IEEE Intell. Transp. Syst. Mag.* **2018**, *10*, 17–27. [CrossRef]
50. Jonas, E.; Schleier-Smith, J.; Sreekanti, V.; Tsai, C.C.; Khandelwal, A.; Pu, Q.; Shankar, V.; Carreira, J.; Krauth, K.; Yadwadkar, N.; et al. Cloud Programming Simplified: A Berkeley View on Serverless Computing. *arXiv* **2019**, arXiv:1902.03383.
51. Gadepalli, P.K.; McBride, S.; Peach, G.; Cherkasova, L.; Parmer, G. Sledge: A Serverless-first, Light-weight Wasm Runtime for the Edge. In *Proceedings of the 21st International Middleware Conference*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 265–279. [CrossRef]
52. Hall, A.; Ramachandran, U. An execution model for serverless functions at the edge. In *Proceedings of the International Conference on Internet of Things Design and Implementation*; ACM: Montreal, QC, Canada, 2019; pp. 225–236. [CrossRef]
53. Cicconetti, C.; Conti, M.; Passarella, A. Low-latency Distributed Computation Offloading for Pervasive Environments. In Proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications (PerCom), Kyoto, Japan, 11–15 March 2019; pp. 1–10.
54. Patman, J.; Chemodanov, D.; Calyam, P.; Palaniappan, K.; Sterle, C.; Boccia, M. Predictive Cyber Foraging for Visual Cloud Computing in Large-Scale IoT Systems. *IEEE Trans. Netw. Serv. Manag.* **2020**, *17*, 2380–2395. [CrossRef]
55. Wang, B.; Ali-Eldin, A.; Shenoy, P. LaSS: Running Latency Sensitive Serverless Computations at the Edge. In *Proceedings of the 30th International Symposium on High-Performance Parallel and Distributed Computing*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 239–251.
56. Pelle, I.; Paolucci, F.; Sonkoly, B.; Cugini, F. Latency-Sensitive Edge/Cloud Serverless Dynamic Deployment Over Telemetry-Based Packet-Optical Network. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2849–2863. [CrossRef]
57. Pelle, I.; Czentye, J.; Doka, J.; Kern, A.; Gero, B.P.; Sonkoly, B. Operating Latency Sensitive Applications on Public Serverless Edge Cloud Platforms. *IEEE Internet Things J.* **2020**. [CrossRef]
58. Elgamal, T. Costless: Optimizing Cost of Serverless Computing through Function Fusion and Placement. In Proceedings of the 2018 IEEE/ACM Symposium on Edge Computing (SEC), Seattle, WA, USA, 25–27 October 2018; pp. 300–312. [CrossRef]

59. Cicconetti, C.; Conti, M.; Passarella, A. A Decentralized Framework for Serverless Edge Computing in the Internet of Things. *IEEE Trans. Netw. Serv. Manag.* **2020**. [CrossRef]
60. Karhula, P.; Janak, J.; Schulzrinne, H. Checkpointing and Migration of IoT Edge Functions. In *Proceedings of the 2nd International Workshop on Edge Systems, Analytics and Networking*; ACM Press: Dresden, Germany, 2019; pp. 60–65. [CrossRef]
61. Cho, C.; Shin, S.; Jeon, H.; Yoon, S. QoS-Aware Workload Distribution in Hierarchical Edge Clouds: A Reinforcement Learning Approach. *IEEE Access* **2020**, *8*, 193297–193313. [CrossRef]
62. Agarwal, S.; Rodriguez, M.A.; Buyya, R. A Reinforcement Learning Approach to Reduce Serverless Function Cold Start Frequency. In Proceedings of the 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Melbourne, Australia, 10–13 May 2021; pp. 797–803.
63. Wang, I.; Liri, E.; Ramakrishnan, K.K. Supporting IoT Applications with Serverless Edge Clouds. In Proceedings of the 2020 IEEE 9th International Conference on Cloud Networking (CloudNet), Piscataway, NJ, USA, 9–11 November 2020; pp. 1–4. [CrossRef]
64. Kim, J.; Lee, K. FunctionBench: A Suite of Workloads for Serverless Cloud Function Service. In Proceedings of the 2019 IEEE 12th International Conference on Cloud Computing (CLOUD), Milan, Italy, 8–13 July 2019; pp. 502–504. [CrossRef]
65. Palade, A.; Kazmi, A.; Clarke, S. An Evaluation of Open Source Serverless Computing Frameworks Support at the Edge. In Proceedings of the 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, 8–13 July 2019; pp. 206–211. [CrossRef]
66. Das, A.; Patterson, S.; Wittie, M. EdgeBench: Benchmarking Edge Computing Platforms. In Proceedings of the 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion), Zurich, Switzerland , 17–20 December 2018; pp. 175–180. [CrossRef]
67. Baresi, L.; Filgueira Mendonca, D. Towards a Serverless Platform for Edge Computing. In Proceedings of the 2019 IEEE International Conference on Fog Computing (ICFC), Prague, Czech Republic, 24–26 June 2019; pp. 1–10. [CrossRef]
68. Baresi, L.; Mendonça, D.F.; Garriga, M.; Guinea, S.; Quattrocchi, G. A Unified Model for the Mobile-Edge-Cloud Continuum. *ACM Trans. Internet Technol.* **2019**, *19*, 1–21. [CrossRef]
69. Yang, S.; Xu, K.; Cui, L.; Ming, Z.; Chen, Z.; Ming, Z. EBI-PAI: Towards An Efficient Edge-Based IoT Platform for Artificial Intelligence. *IEEE Internet Things J.* **2020**. [CrossRef]
70. Rausch, T.; Hummer, W.; Muthusamy, V.; Rashed, A.; Dustdar, S. Towards a Serverless Platform for Edge AI. In Proceedings of the 2nd USENIX Workshop On Hot Topics In Edge Computing (HotEdge 19), Renton, WA, USA, 9 July 2019. Available online: https://www.usenix.org/conference/hotedge19/presentation/rausch (accessed on 27 May 2021).
71. Cheng, B.; Fuerst, J.; Solmaz, G.; Sanada, T. Fog Function: Serverless Fog Computing for Data Intensive IoT Services. In Proceedings of the 2019 IEEE International Conference on Services Computing (SCC), Milan, Italy, 8–13 July 2019; pp. 28–35. [CrossRef]
72. Zhang, M.; Krintz, C.; Wolski, R. STOIC: Serverless Teleoperable Hybrid Cloud for Machine Learning Applications on Edge Device. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; pp. 1–6. [CrossRef]
73. Cicconetti, C.; Conti, M.; Passarella, A.; Sabella, D. Toward Distributed Computing Environments with Serverless Solutions in Edge Systems. *IEEE Commun. Mag.* **2020**, *58*, 40–46. [CrossRef]
74. Huang, Z.; Mi, Z.; Hua, Z. HCloud: A trusted JointCloud serverless platform for IoT systems with blockchain. *China Commun.* **2020**, *17*, 1–10. [CrossRef]
75. Pinto, D.; Dias, J.P.; Sereno Ferreira, H. Dynamic Allocation of Serverless Functions in IoT Environments. In Proceedings of the 2018 IEEE 16th International Conference on Embedded and Ubiquitous Computing (EUC), Bucharest, Romania, 29–31 October 2018; pp. 1–8.
76. Avasalcai, C.; Tsigkanos, C.; Dustdar, S. Resource Management for Latency-Sensitive IoT Applications with Satisfiability. *IEEE Trans. Serv. Comput.* **2021**. [CrossRef]
77. Ling, W.; Ma, L.; Tian, C.; Hu, Z. Pigeon: A Dynamic and Efficient Serverless and FaaS Framework for Private Cloud. In Proceedings of the 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 5–7 December 2019; pp. 1416–1421.
78. Wolski, R.; Krintz, C.; Bakir, F.; George, G.; Lin, W.T. CSPOT: Portable, multi-scale functions-as-a-service for IoT. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*; ACM: Arlington, VA, USA, 2019; pp. 236–249. [CrossRef]
79. Quang, T.; Peng, Y. Device-driven On-demand Deployment of Serverless Computing Functions. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; pp. 1–6.
80. Tricomi, G.; Benomar, Z.; Aragona, F.; Merlino, G.; Longo, F.; Puliafito, A. A NodeRED-based dashboard to deploy pipelines on top of IoT infrastructure. In Proceedings of the 2020 IEEE International Conference on Smart Computing (SMARTCOMP), Bologna, Italy, 14–17 September 2020; pp. 122–129. [CrossRef]
81. Pfandzelter, T.; Bermbach, D. tinyFaaS: A Lightweight FaaS Platform for Edge Environments. In Proceedings of the 2020 IEEE International Conference on Fog Computing (ICFC), Sydney, Australia, 21–24 April 2020; pp. 17–24.
82. Nastic, S.; Rausch, T.; Scekic, O.; Dustdar, S.; Gusev, M.; Koteska, B.; Kostoska, M.; Jakimovski, B.; Ristov, S.; Prodan, R. A Serverless Real-Time Data Analytics Platform for Edge Computing. *IEEE Internet Comput.* **2017**, *21*, 64–71. [CrossRef]
83. Persson, P.; Angelsmark, O. Kappa: Serverless IoT deployment. In *Proceedings of the 2nd International Workshop on Serverless Computing*; ACM: Las Vegas, NV, USA, 2017; pp. 16–21. [CrossRef]

84. Zhang, M.; Wang, F.; Zhu, Y.; Liu, J.; Wang, Z. Towards cloud-edge collaborative online video analytics with fine-grained serverless pipelines. In *Proceedings of the 12th ACM Multimedia Systems Conference*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 80–93.
85. Luckow, A.; Rattan, K.; Jha, S. Pilot-Edge: Distributed Resource Management Along the Edge-to-Cloud Continuum. In Proceedings of the 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Portland, OR, USA, 17–21 June 2021; pp. 874–878. [CrossRef]
86. Lin, W.T.; Bakir, F.; Krintz, C.; Wolski, R.; Mock, M. Data Repair for Distributed, Event-based IoT Applications. In *Proceedings of the 13th ACM International Conference on Distributed and Event-Based Systems*; ACM: Darmstadt, Germany, 2019; pp. 139–150. [CrossRef]
87. Datta, P.; Kumar, P.; Morris, T.; Grace, M.; Rahmati, A.; Bates, A. Valve: Securing Function Workflows on Serverless Computing Platforms. In *Proceedings of The Web Conference 2020*; ACM: Taipei, Taiwan, 2020; pp. 939–950. [CrossRef]
88. Firecracker—Secure and Fast microVMs for Serverless Computing. Available online: https://firecracker-microvm.github.io/ (accessed on 27 May 2021).