



Original software publication

# LiveDewStream: A stream processing platform for running in-lab distributed deep learning inferences on smartphone clusters at the edge



Cristian Mateos\*, Matías Hirsch, Juan Manuel Toloza, Alejandro Zunino

ISISTAN-UNICEN-CONICET, Tandil, Argentina

## ARTICLE INFO

## Article history:

Received 23 June 2022

Received in revised form 7 November 2022

Accepted 15 November 2022

## Keywords:

Mobile devices, Stream processing, Deep learning, Dew computing, Android

## ABSTRACT

Dew computing, an evolution of Fog computing, aims at fulfilling computing needs, such as deep learning applied to object classification, close to where data is originated and using computing resources that include consumer electronic devices such as smartphones. Simulation tools like DewSim aid the study of resource allocation mechanisms for exploiting clusters of smartphones, however, there is a gap w.r.t software tools that allow to perform similar studies over real Dew computing testbeds. We have developed LiveDewStream, an open source project to model executable tasks derived from data streams to be run on real smartphone clusters. The project offers a key functionality missing in other tools: reproducibility of battery-driven Dew experiments. Our major contribution is to provide the community a common in vivo platform to study best-performing allocation mechanisms under different stream processing scenarios and/or deep learning inference models.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## Code metadata

Current code version  
 Permanent link to code/repository used for this code version  
 Permanent link to reproducible capsule  
 Legal code license  
 Code versioning system used  
 Software code languages, tools and services used  
 Compilation requirements, operating environments and dependencies

V1.0  
<https://github.com/ElsevierSoftwareX/SOFTX-D-22-00180>  
<https://github.com/matieber/livedewstream>  
 GNU GPL  
 git  
 Python, Android, shell scripting  
 The *emanager\_server* and *scrunner* modules run on Linux-based machines using Python 3.7+; Normapp runs on Android 6+ (it is built using the provided Android Studio project). See dependencies and installation instructions at the GitHub repository  
<https://github.com/matieber/livedewstream/doc>  
[matias.hirsch@isistan.unicen.edu.ar](mailto:matias.hirsch@isistan.unicen.edu.ar)

If available, link to developer documentation/manual  
 Support email for question

## 1. Motivation and significance

Fog computing [1] was introduced in 2012 to provide highly-scalable network and computing infrastructures for latency and location-aware (mobile) IoT applications, while augmenting resource-constrained devices with processing/storage resources in their proximity. Several alternatives to realize this idea, including cloudlets, mobile edge computing, micro datacenters, nano

\* Corresponding author.

E-mail addresses: [cristian.mateos@isistan.unicen.edu.ar](mailto:cristian.mateos@isistan.unicen.edu.ar) (C. Mateos), [matias.hirsch@isistan.unicen.edu.ar](mailto:matias.hirsch@isistan.unicen.edu.ar) (M. Hirsch), [juanmanuel.toloza@isistan.unicen.edu.ar](mailto:juanmanuel.toloza@isistan.unicen.edu.ar) (J.M. Toloza), [alejandrozunino@isistan.unicen.edu.ar](mailto:alejandrozunino@isistan.unicen.edu.ar) (A. Zunino).

datacenters and femto Clouds, have been introduced [2], which aim at processing data/computations using computing resources located at the edge of the network and optionally using remote resources in the distant Cloud when necessary.

Since then, there has been a tremendous growth in the amount of resource-rich devices and hence computational resources available at the closest edge. According to Statista.com, nearly 84% of the current world's population owns a smartphone. Modern smartphones contain, on average, more than a dozen sensors, up to eight cores, and powerful GPUs. Likewise, thousands of purpose-specific sensing devices such as surveillance cameras, smoke detectors, noise detectors, and so on are being deployed across buildings and cities around the globe. This has led to another paradigm shift by which relying on not-so-close Fog servers to support IoT applications consuming and processing the streams of data from such devices might not suffice. This is particularly true considering that today's IoT applications are becoming more commonplace and intelligent, and therefore the timely and efficient execution of increasingly complex tasks and the context-aware processing of larger amounts of data in urban scenarios is needed, which is difficult for centralized Clouds and challenging even for Fog infrastructures.

Dew computing proposes to establish clusters of mobile devices at the very edge [3], as a form of "ubiquitous and opportunistic computing" within data sensing contexts, specially in public places such as transport, classrooms and coffee shops, where many nearby devices are present. Hence, cost-effective platforms for intelligent IoT applications emerge, provided computational resources are managed wisely. Therefore, platforms and tools to study the individual/collective capabilities and limitations of smartphones for intelligent data stream processing are needed.

We present a software platform to support experimentation with a specific but broad family of such applications, i.e. those using deep learning over data streams. The software allows users, i.e. in-lab Dew researchers, to specify automatic, repeatable batch benchmark plans, while indicating specific smartphone-ready deep learning models and task scheduling algorithms for the cluster. Even when we have already proposed a Dew simulation software [3], our framework represents the first step towards "in vivo" benchmarking of mobile devices for such applications, pretty much like commercial platforms such as BrowserStack and LambdaTest allow users to create in-Cloud device farms with the goal of test automation of web and mobile applications. Given that there is a need for platforms for fairly and comparatively evaluating Edge deep learning performance [4–6], this is the first platform of this kind, particularly aiming at providing a common testbed for studying scheduling of deep learning inference tasks under horizontal execution over mobile device clusters.

The scientific value of this experimentation platform is threefold, namely to allow researchers to (a) realistically characterize and compare smartphone hardware capabilities for executing deep learning codes over arbitrary data streams using multi-core microprocessors and GPUs, (b) to experiment with different cluster settings and task scheduling criteria, and (c) to gather smartphone profile data that might be in turn employed to feed back our existing Dew simulators, thus creating a virtuous circle in deriving task schedulers [7]. As we support arbitrary streams and Tensorflow models, in practice, derived knowledge using our platform might impact many disciplines where Dew computing is the killer computing paradigm.

## 2. Software description

### 2.1. Software architecture

From an architectural standpoint, the platform is a client-server software system, complemented with a hardware device

called Motrol (see Fig. 1(a)). Clients are, on one hand, the mobile devices being exercised, which run a native Android 6+ application (Normapp). Another client is the *scnranner* module, which parses Dew computing scenario parameters, builds corresponding data streams, derives deep learning-based jobs, decides which node executes which job, and submits jobs to the server and hence indirectly exploiting attached mobile devices.

The server works by assuming an energy managing device called Motrol for which Python-based drivers are provided. Via its support for dynamic energy supply switching, Motrol allows researchers to automatically repeat/reproduce job set executions involving several smartphones configured with a specific battery level. Currently, we support an USB-interfaced Arduino device called Motrol 1.0 [8], and a WiFi-enabled ESP8266-based microcontroller device called Motrol 1.5 (see Fig. 1(b)). A more complex prototype based on the Raspberry Pi 4 Model B, called Motrol 2.0 [9], is under development but it is still not considered in this paper. For example, this model will support USB-charging in addition to AC-charging, a feature missing. In terms of software design, *emanager\_server* exposes several Rest APIs. Please refer to the project's documentation for detailed API specifications in the popular Swagger format (<http://swagger.io>).

Motrol 1.0 and Motrol 1.5 use electromechanical relays and provide low-level operations to control energy supply for attached mobile devices. We also provide a mock energy manager to operate the whole platform without Motrol, which prompts the user to manually plug/unplug a specific smartphone<sup>1</sup> from the power grid during test execution as needed. Naturally, this plug/unplug behavior, when using Motrol, remains automatic.

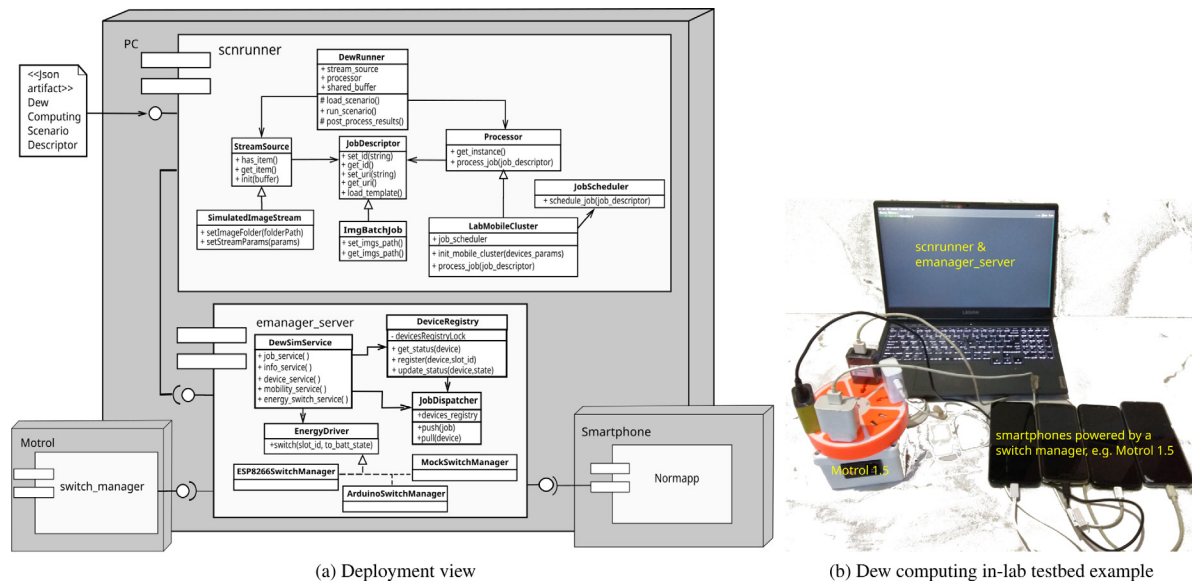
In respect to *scnranner*, it is a subproject designed to encapsulate logic that facilitates stream-derived jobs modeling and its online execution using real Dew computing testbeds. Modeling stream-derived jobs consists in generating partitions from a continuous data flow, for example, frames captured within a time window, so its processing is treated by the computing infrastructure (e.g., mobile cluster) as an indivisible workload (atomic job) to be solved. By online execution we mean that once an atomic job is available, a scheduling mechanism is activated to delegate its execution to a smartphone within the Dew cluster. For now, jobs perform object recognition and classification operations with deep learning over images. By extending the *stream.py* module, i.e., providing custom implementations to the *has\_items* and *get\_item* methods of the *StreamSource* class, new stream types can be supported, e.g., audio streams and text streams.

In conjunction with *emanager\_server*, *scnranner* eases the reproduction of experimental settings representing Dew computing scenarios to study relevant metrics (e.g. throughput, battery utilization, latency) as a result of using different scheduling criteria under real Dew computing environments. An experimental setting is mainly characterized by a workload generation and node states reset. *scnranner* provides a `run.sh` script that is the entry point to execute a given set of Dew computing scenarios. The script receives a path to a directory containing Dew computing scenarios descriptor files as the only required argument.

### 2.2. Software functionalities

Once installed in a PC, the server must be started, by optionally indicating the total number of mobile devices to be employed in the benchmark session. By default, this parameter is the maximum number of smartphones (connection slots) supported by the configured energy device (please see `src/emanager_server/serverConfig.json`).

<sup>1</sup> Under this operation mode, smartphones must be connected directly to the power grid



**Fig. 1.** Architectural view of LiveDewStream. The diagram depicts the UML deployment view, and within each node, the UML components and classes modeling the various entities.

The server will then initialize each device involved in the session, by asking the user to plug, one by one, each device via USB to the PC. This allows the server to gain root access to the device to (a) pushing essential configuration such as server IP address and port, and mostly (b) copy and run shell scripts directly on the device that are needed for benchmarks to correctly operate, and (c) remotely install (or update) and run our Android application on the device. It is worth noting that tasks (a) and (b) are performed via the ADB (Android Debug Bridge) tool of the Android SDK, which allows a PC to remotely send commands to a daemon which runs on a mobile device. On the other hand, task (c) is done by using ADB in conjunction with Monkey, a program that emulates streams of mobile user events such as clicks, touches, or gestures, as well as a number of system-level events. Via Monkey, we launch and (if configured to do so in `serverConfig.json`) start the application automatically. Once Normapp is running in the smartphones, the user unplugs each mobile device from the PC and plugs it using its original charger to the configured energy supply hardware (i.e. Motrol 1.0 or Motrol 1.5) or to wall sockets (i.e. Mock). Normapp will periodically poll the server via the DeviceService API for jobs to execute, and submit back the results. Job creation and result summarization is done by `scrunner`.

### 2.3. Sample code snippets

Algorithm 1 shows `src/scrunner/dew_runner.py`, which has the main method of `scrunner`. In lines 1–9 there is logic to parse the supplied Dew computing scenario JSON file and to create the corresponding stream and processor objects. At line 8, a shared queue is created (and passed as argument to the `build_stream` factory method) to enable stream and processor entities to communicate using a producer–consumer pattern. The Stream entity, which produces items, runs in a separate thread started upon invocation of `stream.yield_items()` at line 11. Processor entity runs in the main thread consuming items from the shared queue (line 15).

### 3. Illustrative example

In the GitHub repository, we provide a sample scenario configuration file (`/doc/cs402_scn002.json`). The `src/scrunner/run.sh` script accepts as an argument a directory path with at least one of these JSON formatted files. The `scn_id` field identifies the scenario. In the file, this field used to name the results directory and an output log file, all created by the platform and where it stores scenario execution related information. The JSON file also contains configuration parameters of the stream entity mentioned in Section 2.3. For instance, `img_folder` indicates the path where input images are located within the filesystem, `per_job_frames`, `per_burst_jobs`, `millis_bt看_jobs` and `millis_bt看_bursts` are parameters used to shape the stream speed and load introduced to the system. In the provided example, jobs are composed of 30 consecutive frames and these jobs are generated every one second (`millis_bt看_bursts`). Job generation finishes when consuming all images of `img_folder`. Images are served to Normapp through a simple HTTP server which can be accessed through the specified port. The configuration also indicates the Python class loaded by the platform to create the next job item every time the `get_item()` method is invoked.

The processor key is where parameters of the processor entity are configured. The `hardsupp.mobile_cluster.LabMobileCluster` class is loaded by the platform to perform nodes initialization and jobs scheduling. The class uses the `emanager_server` Rest API. Nodes in this case are four smartphones whose model names and initial battery level are indicated under `devs_batt_init`. Battery level is a two-decimal number in the range [0–1]. This is, in the configuration, battery level is a number with at most two decimals, being 0 equals to 0% of battery (in the mobile device). The same applies to battery level equals to 1 (it represents 100%). Other examples are 0.04, 0.55 and 0.85 (meaning 4%, 55% and 85%). Besides, “-1” is interpreted as “any battery level”. Desired scheduling logic is configured by providing the path using dot notation to the class implementing a processor logic, in this case, `job.job_scheduling.RoundRobin`. Paths are relative to where `dew_runner.py` is within the project.

Fig. 2 shows, on the top left, a sample directory tree of files generated by the platform after running a test. The `scn_id`

**Algorithm 1:** Parsing/creating the stream and processor entities declared in scenario descriptor file.

```

1  if __name__ == '__main__':
2
3      args = parser.parse_args()
4      with open(args.scenarioDescriptor, "r") as scnfile:
5          scn_data = json.load(scnfile)
6
7      processor = pb.build_processor(scn_data["processor"])
8      jobs_queue = queue.Queue(0)
9      stream = sb.build_stream(scn_data["scn_id"], scn_data["stream_source"], jobs_queue)
10
11     stream.yield_items()
12     test_end = False
13     while not test_end:
14         try:
15             processor.process_job(jobs_queue.get())
16         except queue.Empty:
17             if stream.is_closed() and processor.all_jobs_completed():
18                 test_end = True

```

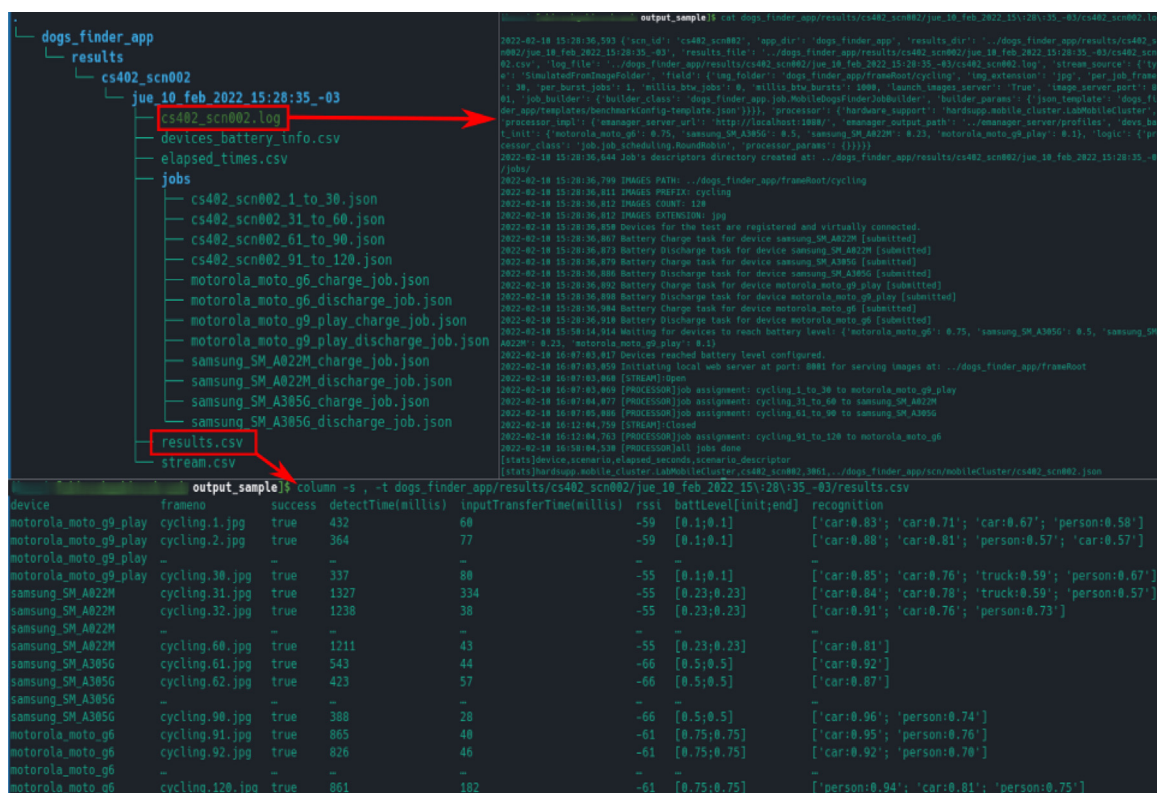


Fig. 2. Output sample: Directory tree generated, events log and results file.

specified in the scenario file and the system date are used to identify and create a directory (`dogs_finder_app/results/cs402_scn002`) where to locate all output files generated by a test. Such directory includes a plain text log containing chronologically-ordered events (see Fig. 2 top right). Depending on which entity is logging information, entries of the event log are tagged with [STREAM] or [PROCESSOR]. There is also scenario descriptor file information to identify which log corresponds to which input parameters. Fig. 2 bottom part shows a sample `results.csv` file, which is built by merging individual results sent by mobile devices to the server. It tells which images were sent to which device, how much time the device spent on downloading data and performing the inference over the image. It also records RSSI indicator, battery level at the time of downloading images and finishing inferences, and the output of the recognition algorithm itself.

To illustrate some of the software output that allow researchers to study the feasibility of Dew computing scenarios, next we include running examples of several Dew computing scenario configuration files (`dogs_finder_app/scn/`). These aim at studying the performance of different versions of the same Tensorflow Lite model on two mobile devices using different tasks load balancing mechanisms. The resulting plots were obtained by processing the `results.csv` file of each Dew computing scenario. Figs. 3(a) and 3(b) show the inference times of different versions of the same Tensorflow Lite model (YOLOv4) when running on a Xiaomi Redmi Note 7 and a Samsung A30 smartphone, respectively. We can see that the lowest inference time is achieved using quantization and multithreading. However, multithreading produces more dispersed times than single threading model execution. Fig. 3(c) shows the performance of all model-thread combinations for a Dew computing scenario



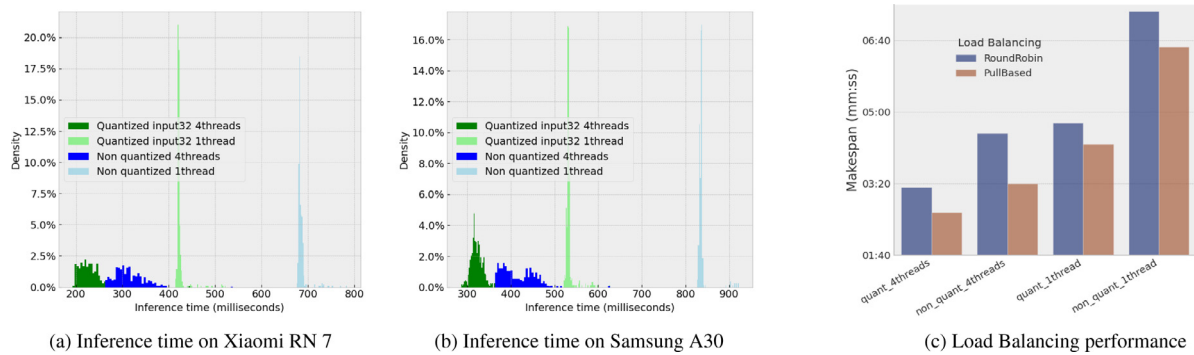


Fig. 3. YOLOv4 model performance.

where two smartphones cooperate in making inferences over a stream of 1000 images given at a rate of 30 FPS. Consistently with previous results, the quantized version of YOLOv4 using multithreading achieves the lowest makespan. Moreover, a pull-based load balancing mechanism beats Round Robin irrespective of the model version.

#### 4. Impact and closing remarks

Due to the difficulty and time-consuming nature of experimentation in the Distributed Computing research community, simulation is an accepted practice. This is demonstrated by the heavy adoption of simulators for Cloud and Fog computing in the literature. Prominent examples are CloudSim (2011) [10] and iFogSim (2017) [11], whose seminal papers have reached around 5,500 and 1,200 citations respectively according to Google Scholar. The software projects behind these simulators have also made an impact in the abovementioned research community. For instance, CloudSim has generated dozens of complementary projects (e.g. WorkflowCloudSim) and its GitHub repo ([github.com/Cloudslab/cloudsim](https://github.com/Cloudslab/cloudsim)) has over 400 forks. In Dew computing, simulation is also a practiced evaluation methodology for studying mobile resource scavenging heuristics [7]. Then, we have proposed DewSim (2020) [3], a trace-based simulation software that uses battery traces to map mobile devices resource utilization levels (e.g. microprocessor) with battery behavior. Despite having many miles ahead to travel yet, DewSim plus its supporting platform [8] has already shown impact in the scientific literature, as it is the support for many Dew schedulers [7,12], which have been published in reputed journals.

Generally speaking, simulations assume that all the involved elemental processes are known and correct, only then built simulators are fully trustworthy. In practice, however, the relentless developing nature of Cloud, Fog and Dew gradually introduce new fundamental knowledge that challenge the assumptions made by built simulators. This means (a) new versions of specific software modules have to be built, or (b) integral software modifications must be carried out in order to comply to certain validation experiments. For example, a trace-based file transfer model has been proposed for CloudSim [13], and the sub-optimal accuracy of the simulator as a whole when modeling resource heterogeneity has been evidenced [14].

The software proposed in this paper provides a testbed for deriving such fundamental knowledge, so DewSim can be correctly evolved, and proper validation experiments can be performed. Not to mention that, as the platform targets deep learning codes over data streams on mobile devices clusters, this specific yet general-enough application scenario, where Dew computing seems to be the killer paradigm, might increment our current user base.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Data availability

No data was used for the research described in the article.

#### Acknowledgments

We thank Eng. Pablo Sanabria (<https://www.linkedin.com/in/tailmon/?originalSubdomain=cl>) from Chile for his assistance regarding running Tensorflow Lite models on Android. We also acknowledge the financial support by CONICET (grant no. 11220170100490CO) and ANPCyT (grant no. PICT-2018-03323).

#### References

- [1] Bonomi Flavio, Milito Rodolfo, Zhu Jiang, Addepalli Sateesh. Fog computing and its role in the internet of things. In: Proceedings of the first edition of the MCC workshop on mobile cloud computing. 2012, p. 13–6.
- [2] Aazam Mohammad, Zeadally Sherali, Harras Khaled A. Offloading in fog computing for iot: Review, enabling technologies, and research opportunities. *Future Gener Comput Syst* 2018;87:278–89.
- [3] Hirsch Matías, Mateos Cristian, Rodríguez Juan Manuel, Zunino Alejandro. Dewsim: A trace-driven toolkit for simulating mobile device clusters in dew computing environments. *Softw - Pract Exp* 2020;50(5): 688–718.
- [4] Wang Xiaofei, Han Yiwen, Leung Victor CM, Niyato Dusit, Yan Xueqiang, Chen Xu. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Commun Surv Tutor* 2020;22(2): 869–904.
- [5] Chen Jiasi, Ran Xukan. Deep learning with edge computing: A review. *Proc IEEE* 2019;107(8):1655–74.
- [6] Murshed MG Sarwar, Murphy Christopher, Hou Daqing, Khan Nazar, Ananthanarayanan Ganesh, Hussain Faraz. Machine learning at the network edge: A survey. *ACM Comput Surv* 2021;54(8):1–37.
- [7] Hirsch Matías, Mateos Cristian, Zunino Alejandro, Majchrzak Tim A, Grønli Tor-Morten, Kaindl Hermann. A task execution scheme for dew computing with state-of-the-art smartphones. *Electronics* 2021;10(16):2006.
- [8] Hirsch Matías, Mateos Cristian, Zunino Alejandro, Toloza Juan. A platform for automating battery-driven batch benchmarking and profiling of android-based mobile devices. *Simul Model Pract Theory* 2021;109:102266.
- [9] Mateos Cristian, Hirsch Matías, Toloza Juan, Zunino Alejandro. Motrol 2.0: A dew-oriented hardware/software platform for batch-benchmarking smartphones. In: 2021 IEEE 45th annual computers, software, and applications conference. IEEE; 2021, p. 1772–7.

- [10] Calheiros Rodrigo N, Ranjan Rajiv, Beloglazov Anton, Rose César AF De, Buyya Rajkumar. Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Softw - Pract Exp* 2011;41(1):23–50.
- [11] Gupta Harshit, Dastjerdi Amir Vahid, Ghosh Soumya K, Buyya Rajkumar. Ifogsim: A toolkit for modeling and simulation of resource management techniques in the internet of things, edge and fog computing environments. *Softw Pract Exp* 2017;47(9):1275–96.
- [12] Hirsch Matías, Rodríguez Juan Manuel, Mateos Cristian, Zunino Alejandro. A two-phase energy-aware scheduling approach for cpu-intensive jobs in mobile grids. *J Grid Comput* 2017;15(1):55–80.
- [13] Chai Anchen, Bazm Mohammad-Mahdi, Camarasu-Pop Sorina, Glatard Tristan, Benoit-Cattin Hugues, Suter Frédéric. Modeling distributed platforms from application traces for realistic file transfer simulation. In: 2017 17th IEEE/ACM international symposium on cluster, cloud and grid computing. IEEE; 2017, p. 54–63.
- [14] Zakarya Muhammad, Gillam Lee. Modelling resource heterogeneities in cloud simulations and quantifying their accuracy. *Simul Model Pract Theory* 2019;94:43–65.