Just Enough, Just in Time, Just for “Me”: Fundamental Principles for Engineering IoT-native Software Systems

Zheng Li
University of Concepción
Concepción, Chile
imlizheng@gmail.com

Rajiv Ranjan
Newcastle University
Newcastle upon Tyne, UK
raj.ranjan@ncl.ac.uk

ABSTRACT

By seamlessly integrating everyday objects and by changing the way we interact with our surroundings, Internet of Things (IoT) is drastically improving the quality of households and enhancing the productivity of businesses. Given the unique IoT characteristics, IoT applications have emerged distinctively from the mainstream application types. Inspired by the outlook of a programmable world, we further foresee an IoT-native trend in designing, developing, deploying, and maintaining software systems. However, although the challenges of IoT software projects are frequently discussed, addressing those challenges are still in the “crossing the chasm” period. By participating in a few various IoT projects, we gradually distilled three fundamental principles for engineering IoT-native software systems, such as just enough, just in time, and just for “me”. These principles target the challenges that are associated with the most typical features of IoT environments, ranging from resource limits to technology heterogeneity of IoT devices. We expect this research to trigger dedicated efforts, techniques and theories for the topic IoT-native software engineering.

1 INTRODUCTION

Driven by the revolutionary development of modern computer, information and communication technologies (e.g., from large mainframes to small chips and from coaxial cables to 5G), all the sectors of our lives are experiencing a rapid evolution in the use of Internet of Things (IoT) [8, 13, 27]. An estimation shows that about 50 billion devices will be connected and in use worldwide by 2030 for creating smart environments [24], while the recent statistics reveal that “an increase of 10 percentage points in the growth of IoT connections per inhabitant is associated with a 0.23 percentage points increase in the total factor productivity (TFP) growth” [4]. Correspondingly, large amounts of companies have started investing on IoT applications to grow their businesses and to strengthen their competitive advantages.

Despite the de facto cloud-native solutions that satisfy IoT application requirements mainly in the data centers (e.g., AWS IoT and Azure IoT), there is a clear trend in deploying and executing custom application logics directly on IoT devices [26]. In fact, by foreseeing richer embedded execution environments, researchers have predicted that our everyday things will all become not only connected but also programmable dynamically [21]. Inspired by the prospect of a programmable (Internet) edge world, we advocate IoT-native solutions for achieving IoT-intensive business values, by composing distributed software components across IoT and edge nodes. In specific, we define IoT-native software to be executable artefacts designed, developed and deployed directly in the IoT environments, with little or limited cloud-side orchestration.

However, there are inevitable challenges of engineering IoT-native software systems, and the challenges are closely associated with the IoT characteristics. In addition to the massive device distribution [14] that leads to the generic challenges of distributed software engineering [20], this research particularly highlights two fundamental features of IoT. First, individual IoT devices generally have limited compute, storage and network capabilities. For example, the embedded RAM is only 4KB for current taximounted GPS devices. Although the optimistic opinion about future IoT hardware is that any type of device will enable virtualized software stacks (ranging from fully fledged OSs to dynamic language runtimes) [21], those everyday things shall always have considerably less resources than the dedicated computing machines of the same generation.

Second, an IoT system may involve non-scalable integration of heterogeneous technologies produced by different manufacturers. On one hand, the IoT industry has not employed any globally standard protocols and universal interfaces yet, which is similar to the early computer industry [5]. On the other hand, to secure the niche market shares against competitors, different device

CCS CONCEPTS

• Software and its engineering → Development frameworks and environments: Distributed systems organizing principles; • Human-centered computing → Ubiquitous and mobile computing systems and tools.

KEYWORDS

choorephy, cloud-centric coordination, Internet of Things (IoT), IoT-native software engineering, software engineering principles
work plan. According to the privacy law, organizations must limit the collection of personal data to the minimum extent for their business needs. The spirit of just enough, just in time, and just for “me”. In brief, (1) just enough aims to address the limitation of compute and storage capabilities of IoT devices, by minimizing data collection, storage duration, and processing workloads. (2) Just in time aims to address IoT devices’ limited network capability and/or the environmental bandwidth constraints, by optimizing the data communication among IoT application components. (3) Just for “me” aims to address the heterogeneity and complexity in organizing the distributed IoT application components, by enabling the components’ awareness of environments and by making them collaborate with each other without using orchestration flows. More importantly, in this paper we not only justify the three principles theoretically, but also demonstrate suitable techniques that can satisfy these principles, based on our practices in IoT projects [10–12]. As such, this research makes a twofold contribution:

- In theory, our forward-looking vision about IoT-native software can inspire researchers to keep developing a dedicated suite of software engineering principles for IoT. In fact, it has been identified that IoT applications differ from the mainstream application types [21], and there is still a lack of a common set of approaches, models, and methodologies to facilitate engineering IoT-intensive software systems [28].
- In practice, our initially developed principles together with the demonstrated techniques can help practitioners better implement IoT-native software projects. We argue that the three fundamental principles will stay valid and effective along the evolution of IoT, e.g., even after the fully-fledged software stacks are available on all the IoT devices.

The remainder of this paper is organized as follows. The three principles are introduced and discussed in Section 2, 3, and 4 respectively. Section 5 draws conclusions and specifies our future work plan.

2 PRINCIPLE I: JUST ENOUGH

The current big data business may incentivize collecting more data and holding them for longer periods of time [22]. However, it is impractical and often unnecessary to centralize every bit of data from every single IoT device [7]. Following this idea, we argue just enough collection, storage and processing of data when developing software systems in the IoT environments.

2.1 Minimization of Data Collection

The spirit of just enough in this case naturally matches data minimization that is a fundamental principle of the privacy law [23]. According to the privacy law, organizations must limit the collection of personal data to the minimum extent for their business needs. In addition to the concern about privacy, the limited resources and battery life of IoT devices are the other critical reasons for software systems to reduce the data collection (e.g., lowering the sensing frequency).

Besides the design-time data reduction and runtime data filtering [7], we argue that on-device processing can further help minimize data collection, as it will avoid the centralization of raw data. Furthermore, the real-time data erasing should be recommended and preferred whenever it is possible during the data processing, because erasing the data at the earliest chance will also satisfy the minimization of data storage duration (cf. Section 2.2). For example, the following techniques are particularly suitable for this case.

- **Running Statistics** can calculate the statistical indicators of the incoming data in a single pass. The recursive updates of statistical indicators and internal parameters are based on one data point at a time. Every data point can be erased immediately after being used.
- **Moving Window Statistics** only summarize the data within a predefined window, and keep updating the statistical summary along moving the window over the incoming data. Since the width of the window is fixed, each update of the statistical summary not only includes a new data point but also excludes the oldest one. A data point can be erased once it is excluded from the window.
- **Streaming Algorithms** aim to cope with extremely large (or even infinite) sequential data in one pass (or a small number of passes) by using a constant size of space [16]. By definition it is clear that the raw data streams are not supposed to be saved, and thus the streaming algorithms can be viewed as a superset of running statistics.

2.2 Minimization of Data Storage Duration

Unless the original data are critical for future verification or for cumulative data mining, it is generally unnecessary to keep them forever. This idea intuitively aligns with the decay theory, i.e. the stored information in human mind would gradually become unavailable for later retrieval as memory fades over time [19]. The benefit of forgetting the previous information is that "it lets us act in time, cognizant of, but not shackled by, past events. Through perfect memory we may lose a fundamental human capacity to live and act firmly in the present" [2]. By analogy, to utilize storage resources more efficiently, we advocate to remove the data once they have satisfied the purposes for which they are collected.

Considering the location and awareness of data consumption, we only discuss the data removal from the data sink’s perspective. Note that we recognize two roles in a data flow, i.e. data source and data sink. A data sink can in turn act as a data source when it delivers intermediate processing results to its downstream sinks.

Furthermore, we distinguish between two patterns of data erasure according to the sink’s controllability over the data’s lifecycle.

- **Controlled Data Erasure** is for the data that are consumed by the sink node only, and thus the erasure will purely depend on the data sink’s decision. To minimize the data storage duration, the sink node may intentionally conduct data
erasure right after its processing tasks are done. In particular, the received data can be discarded on the fly if they are processed in the streaming way (cf. Section 2.1).

- **Data Self-Erasure** is for the data that are stored on the sink node while consumed by external stakeholders (e.g., the end users’ review). In this case, the data erasure would have to wait for the uncertain external interventions. Driven by the concern about privacy, an alternative option is to let data be erased automatically based on some predefined rules. In fact, data self-erasure (a.k.a. self-destruction) has become a successful innovation especially for the social media and communication applications, e.g., Snapchat\(^1\) and WhatsApp\(^2\).

### 2.3 Minimization of Data Processing Workloads

To relieve the limitation of compute capability in the IoT environments, we can try to minimize the processing workloads from both the runtime and the design-time points of view. At runtime, a software system’s data processing workloads would be proportional to the amount of user requests. Moreover, one user may issue the same request multiple times (e.g., due to the network interruption), and different users may request the same data simultaneously or at different times (e.g., in the multicast scenario). Therefore, in suitable cases, it is worth taking into account caching mechanisms to reduce the repeated data processing. A typical example that can demonstrate this strategy in the IoT domain is the mechanism content store [1]. Note that caching processing results does not violate the aforementioned minimization of the storage duration of raw data.

At design time, the functional features of a software system could have essential impacts on the processing workloads. Therefore, we argue that practitioners should avoid over-designing IoT software systems and only implement just enough functional features to deal with real-world problems. This argument is confirmed by Elon Musk’s recent talk about the five-step engineering process [18]. In particular, (software) engineers are strongly suggested to clarify the requirements and to avoid the bias towards adding features when looking for a solution. A typical example that can demonstrate this argument in the IoT domain is: The researchers improved the performance of online skewness monitoring by skipping the explicit variance calculation [9].

### 3 PRINCIPLE II: JUST IN TIME

By guaranteeing at least the just-in-time arrival of key data, we argue to tolerate certain processing delays to improve the communication efficiency or to relieve the network congestion within an IoT system. The “just-in-time” moments on different data sink nodes can be decided by referring to the plain system design and implementation. We use three different scenarios to explain this argument as follows.

The first and the most straightforward scenario is the one-off data transfer between a pair of source and sink nodes. Assume transferring a data block \(D\) takes time \(T_D\) in the plain version of an IoT application. Alternatively, we can compress the data on the source node, transfer the compressed data, and then decompress the data on the sink node. If the alternative approach takes as much time as \(T_D\) or less, the extra overhead of data compression and decompression will pay off, as it will at least save the bandwidth resource. In this case, we claim that the alternative approach satisfies the principle of just in time, because the data arrival is guaranteed no later than its plain counterpart, as regulated in Eq. (1).

\[
T_c + T_D + T_{dec} \leq T_D
\]

where \(T_c\) represents the latency of transferring the compressed data block \(D'\), and \(T_D\) and \(T_{dec}\) indicate the time consumption of data compression and decompression respectively.

The second scenario is the sequential transfer of a finite set of data points between a pair of source and sink nodes. Following the plain first-in-first-out (FIFO) fashion, we assume that transferring the first \(k\) data points takes time \(T_{fifo}(d_k)\), i.e., the wall-clock latency of transferring the \(k\)th data point \(d_k\) is \(T_{fifo}(d_k)\). Inspired by the packet buffering feature of today’s Internet protocols, we have developed application-level buffering mechanisms to improve the efficiency of data communication, as long as the arrival of the buffered key data satisfies Eq. (2). The key data can either be a single data point that has high priority, or a group of data points that enjoys majority benefits. In both situations, we only need to select one data point as the reference to adjust the buffer size, and the reference’s arrival is guaranteed no later than its FIFO counterpart. In practice, the dynamic buffer adjustment will require the prior knowledge about the IoT environment (via performance evaluation).

\[
T_{bf(D)} + T_D \leq T_{fifo}(d_k), \quad d_k \in D
\]

where \(T_{bf(D)}\) stands for the time consumption of buffering a group of data points (including \(d_k\)) into a block \(D\). In particular, \(d_k\) acts as the reference that can bring majority benefits to \(D\) without delaying its own arrival time.

The third scenario is the data shuffling among multiple IoT devices. Data shuffling is pervasively required by the modern programming models and distributed computing frameworks (e.g., MapReduce and Dryad) [3, 6], and it may have a dominant effect on the completion time of a computing task. For example, the communication overhead for data shuffling can account for up to 70% of the overall execution time of self-join applications [29]. If a distributed workflow involves large amounts of shuffling events, the previous strategy of frequent data compression and decompression will be impractical and unacceptable. To address this challenge, we have introduced compact data structures (CDS) techniques [15] to the development of data shuffling-intensive IoT applications. Compared with data compression, the CDS techniques can not only maintain data with less space, but they also enable the data to be queried, navigated, and manipulated directly in the compact form, i.e., without being decompressed. By supplementing a CDS layer to the original data processing logic, we can still finish suitable tasks at least on time, with significantly less consumption of storage and network resources.

Overall, this principle essentially promotes the value of preparation before making a job done, as stated by the idiom “a beard well lathered is half shaved”. It should be noted that the increased

\(^1\)https://support.snapchat.com/en-US/a/when-are-snaps-chats-deleted

\(^2\)https://faq.whatsapp.com/general/chats/about-disappearing-messages/
processing workload here (for non-functional "preparation") does not violate the aforementioned sub-principle about minimizing processing workloads (for functional needs).

4 PRINCIPLE III: JUST FOR “ME”
Considering the distributed nature and limited resources of IoT devices, on-device processing would particularly be suitable for single-purpose and loosely coupled tasks, and the corresponding functional components would also follow the UNIX philosophy "do one thing, and do it well". Therefore, we argue that the architectural principle and style of microservices architecture (MSA) are still useful and valuable for building IoT-native software systems. For the purpose of conciseness, here we reuse the concept "microservice" to represent those IoT-native functional components, although the practical implementations may vary and not always fit in the de facto definition. For example, the popular microservice techniques and theoretical standards (e.g., containerization and self-maintained database) could be too heavy to implement on resource-constrained sensors and actuators, at least at this current stage.

Recall that the UNIX philosophy emphasizes high cohesion and loose coupling. Inspired by the idea about living with and exploiting the heterogeneity [14], we further emphasize the awareness of environments. In particular, each microservice only needs to be aware of its own environment, which eventually leads to the principle just for "me". As such, the IoT software development will be able to minimize the orchestration flows and mainly rely on the collaboration among microservices, via notifying and detecting their environments:

- **Notification** indicates the activity that a microservice updates the environment with its own status.
- **Detection** indicates the activity that a microservice requests the information of its environment.

We employ the scenario of conference room illumination [28] to demonstrate how to satisfy just for "me" in practice, as shown in Figure 1. For the convenience of description, we directly use the device names to represent the corresponding on-device microservices, and we assume their business logic partnership in this demonstration as: The window shade needs outdoor brightness sensor (for deciding open or close), the ceiling light needs window shade and indoor brightness sensor (for deciding on or off), while the projector needs ceiling light and window shade (for adjusting the room illumination). To clarify the environmental awareness scope, each microservice maintains a table of its predefined business logic partners, except for the standalone ones. Then:

1. When the projector turns on for the first time, it broadcasts a semantic notification to all the devices within the same room. By matching the device names/interfaces, the ceiling lights and window shades will send their addresses as responses to the projector, and the projector will save these addresses into its partner table. After that, the projector will only broadcast future notifications to the saved addresses until it needs to update the partner table again.
2. After sending responses to the projector and then interpreting the notification, the window shades will close (or remain closed) and the ceiling lights will turn off (or remain off).
3. Before the projector turns off, it will broadcast another notification to the relevant devices and verify their responses.

(4) After responding and interpreting the second notification, the window shades will open, unless it is dark night according to the outdoor brightness sensors. The ceiling lights will firstly detect the status of window shades. If the status is "opening", the ceiling lights will wait for a short period of time. Then, they further detect the natural daylight via indoor brightness sensors to decide whether they should turn on.

It is clear that compared to the study [28], our solution does not require any central lighting controller. Particularly, the semantic notifications are not specific control instructions. For example, the projector may notify its environment with "I am turning on and expecting low illumination", which does not physically guide what its business logic partners should do. Instead, the ceiling lights and window shades will be able to make different and appropriate actions, by interpreting the same notification into respective workflows.

Overall, this principle essentially argues to further break the monolithic orchestration mechanism (e.g., cloud-centric coordination [25]), and to employ the choreography mechanism to compose individual microservices into an IoT application.

5 CONCLUSIONS AND FUTURE WORK
Since the term "Internet of Things" coined in 1999 [17], the movement towards IoT has been happening and accelerating unprecedentedly in every sector of the world. In addition to the evolution of sensor technologies and IoT infrastructure, the emerging IoT applications have also become distinctive from the current mainstream application types. Given the trend in making everyday things programmable, we foresee the replacement of cloud-centric IoT solutions with IoT-native solutions that will implement software projects directly in the IoT environments.

Correspondingly, we argue the needs of dedicated efforts, techniques and theories for engineering IoT-native software systems. Unfortunately, the software engineering research in IoT is still at the early stage, not to mention the potential topic IoT-native software engineering. Benefiting from attending a few various IoT projects, we particularly extracted experience and lessons from those software sections. This research summarizes our experience and lessons into three principles, and we claim them to be fundamental because (1) they were born from the unique IoT characteristics and (2) they match the IoT-native evolution direction. Based on the current effort, we plan to enlarge the collaborate scope and enrich the engineering principles to cover broader phases (ranging from system design to maintenance) of IoT-native software projects.
REFERENCES

[1] Marica Amadeo, Claudia Campolo, Antonella Molinaro, and Giuseppe Ruggieri. 2018. IoT Data Processing at the Edge with Named Data Networking. In Proceedings of the 24th European Wireless Conference (EW2018). IEEE Press, Catania, Italy, 38–43.

[2] Oshrat Ayalon and Eran Toch. 2013. Retrospective privacy: managing longitudinal privacy in online social networks. In Proceedings of the 9th Symposium on Usable Privacy and Security (SOUPS 2013). ACM Press, NewCastle, United Kingdom. Article no. 4.

[3] Mohsharaf Chowdhury, Matei Zaharia, Justin Ma, Michael I. Jordan, and Ion Stoica. 2011. Managing data transfers in computer clusters with orchestra. In Proceedings of the ACM SIGCOMM 2011 Conference on Applications, Technologies, Architectures, and Protocols for Communication ACM Press, Toronto, ON, Canada, 98–109.

[4] Harald Edquist, Peter Goodridge, and Jonathan Haskel. 2021. The Internet of Things and economic growth in a panel of countries. Economics of Innovation and New Technology 30, 3 (2021), 262–283.

[5] B. L. Gorman, D.R. Resseguie, and C. Tomkins-Tinch. 2009. Sensorpedia: Information sharing across incompatible sensor systems. In Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems (CTS 2009). IEEE Press, Baltimore, Maryland, USA, 448–454.

[6] Hai Jin, Lin Jia, and Zhi Zhou. 2021. Boosting Edge Intelligence With Collaborative Cross-Edge Analytics. IEEE Internet of Things Journal 8, 4 (February 2021), 2444–2458.

[7] Gabor Kesekmeti, Giuliano Casale, Devki Nandan Jha, Justin Lyon, and Rajiv Ranjan. 2017. Modelling and Simulation Challenges in Internet of Things. IEEE Cloud Computing 4, 1 (January-February 2017), 62–69.

[8] Abhishek Khanna and Sannmeet Kaur. 2019. Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture. Computers and Electronics in Agriculture 157, 218–231 (February 2019).

[9] Zheng Li and Jhon Guldames-Retamal. 2021. On IoT-Friendly Skewness Monitoring for Skewness-Aware Online Edge Learning. Applied Science 11, 16 (August 2021). Article no. 7461.

[10] Zheng Li and Esteban J. Pino. 2019. D&D: A Distributed and Disposable Approach to Privacy Preserving Data Analytics in User-Centric Healthcare. In Proceedings of the 12th IEEE International Conference on Service-Oriented Computing and Applications (SOCA 2019). IEEE Computer Society, Kaohsiung, Taiwan, 176–183.

[11] Zheng Li, Mauricio Pradena Miquel, and Pedro Pinacho-Davidson. 2022. Safety-Centric and Smart Outdoor Workplace: A New Research Direction and Its Technical Challenges. In Smart Trends in Computing and Communications. Lecture Notes in Networks and Systems, Vol. 286. Springer, Singapore, 61–74.

[12] Zheng Li, Diego Seco, and José Fuentes-Sepúlveda. 2021. When Edge Computing Meets Compact Data Structures. In Proceedings of the 2021 IEEE Cloud Summit (CloudSummit 2021). IEEE Computer Society, Hemstead, NY, USA, 29–34.

[13] Eva Martínez-Caro, Juan Gabriel Cegarra-Navarro, Alexis Garcia-Pérez, and Monica Fait. 2018. Healthcare Service Evolution Towards the Internet of Things: An End-user Perspective. Technological Forecasting and Social Change 136 (November 2018), 268–276.

[14] Brice Morin, Nicolas Harrand, and Franck Fleurey. 2017. Model-Based Software Engineering to Tame the IoT Jungle. IEEE Software 34, 1 (January-February 2017), 30–36.

[15] Gonzalo Navarro. 2016. Compact Data Structures: A Practical Approach (1st ed.). Cambridge University Press, New York, NY.

[16] Alrazmazha Nujum Navaz, Saad Harous, Mohammed Adel Serhani, and Bhal Taleb. 2019. Real-Time Data Streaming Algorithms and Processing Technologies: A Survey. In Proceedings of the 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE 2019). IEEE Press, Dubai, United Arab Emirates, 246–250.

[17] Postscapes. 2019. Internet of Things (IoT) History. https://www.postscapes.com/iot-history/.

[18] Matt Pressman. 2021. Elon Musk reveals his 5-step engineering protocol. https://evannex.com/blogs/news/elon-musk-reveals-his-5-step-protocol-for-engineering.

[19] Timothy J. Ricker, Evie Vergauwe, and Nelson Cowan. 2016. Decay Theory of Immediate Memory: From Brown (1958) to Today (2014). Quarterly Journal of Experimental Psychology 69, 10 (October 2016), 1969–1995.

[20] Arnon Rotem-Gal-Oz. 2013. Fallacies of massively distributed computing. https://arnon.me/2013/04/fallacies-massively-distributed-computing/.

[21] Antero Taivalsaari and Tommi Mikkonen. 2017. A Roadmap to the Programmable World: Software Challenges in the IoT Era. IEEE Software 34, 1 (January-February 2017), 72–80.

[22] Omer Tene and Jules Polonetsky. 2013. Big Data for All: Privacy and User Control in the Age of Analytics. Northwestern Journal of Technology and Intellectual Property 11, 5 (November 2013). Article no. 1.

[23] The White House. 2012. Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy. CreateSpace Independent Publishing Platform, Washington.