From Conception to Retirement: a Lifetime Story of a 3-Year-Old Wireless Beacon System in the Wild

Yi Ding
Alibaba Group, University of Minnesota

Ling Liu
Shanghai Jiao Tong University

Yu Yang
Rutgers University

Yunhuai Liu
Peking University

Desheng Zhang
Rutgers University

Tian He
Alibaba Group, University of Minnesota

Abstract

We report a 3-year city-wide study of an operational indoor sensing system based on Bluetooth Low Energy (BLE) called `aBeacon` (short for alibaba Beacon). `aBeacon` is pilot-studied, A/B tested, deployed, and operated in Shanghai, China to infer the indoor status of Alibaba couriers, e.g., arrival and departure at the merchants participating in the Alibaba Local Services platform. In its full operation stage (2018/01-2020/04), `aBeacon` consists of customized BLE devices at 12,109 merchants, interacting with 109,378 couriers to infer their status to assist the scheduling of 64 million delivery orders for 7.3 million customers with a total amount of $600 million USD order values. Although in an academic setting, using BLE devices to detect arrival and departure looks straightforward, it is non-trivial to design, build, deploy, and operate `aBeacon` from its conception to its retirement at city scale in a metric-based approach by considering the tradeoffs between various practical factors (e.g., cost and performance) during a long-term system evolution. We report our study in two phases, i.e., an 8-month iterative pilot study and a 28-month deployment and operation in the wild. We focus on an in-depth reporting on the five lessons learned and provide their implications in other systems with long-term operation and large geospatial coverage, e.g., Edge Computing.

1 Introduction

Instant delivery is an emerging business where online orders (e.g., groceries or foods) are delivered within a short time (e.g., 30 mins) from merchants (e.g. grocery stores and restaurants) to customers. This business grows rapidly in recent years with the emergence of several online platforms, e.g., Prime Now [6], Uber Eats [50], Instacart [26], DoorDash [16], Deliveroo [14], and Alibaba Local Services [17]. In an instant delivery service, a customer uses an APP on a platform to place an order at a merchant; the platform assigns a courier to pick up this order at the merchant and then deliver it to the customer. It is essential for the platform to know its couriers' real-time arrival status at merchants, which is used to assign new orders to the most suitable couriers based on their locations to avoid an order delivery overdue given short delivery window [57]. While the outdoor status of couriers can be obtained by smartphone GPS, inferring the indoor status is always challenging due to a lack of infrastructure at scale.

In this paper, we report a 3-year study for a system named `aBeacon` developed by Alibaba Inc. [5] in Shanghai to infer its couriers' indoor status, i.e., arriving and departing at merchants. The indoor status inference is of great significance for Alibaba Local Services (a subsidiary of Alibaba Inc. for instant delivery), since couriers spend almost one third of total working time indoor based on our data. The goal of `aBeacon` is to provide a city-wide indoor sensing solution with practical cost/performance tradeoffs when deploying in the wild. *We share one-month data of `aBeacon` for future research*[1] [1].

Admittedly, indoor arrival and departure status detection is not technically challenging and has been widely investigated in controlled environments, e.g., labs, museums, and airports. However, it is still an open question for city-wide detection in the wild. In industry, current solutions mainly rely on either courier's smartphone GPS (which is inaccurate in indoor environments) [29] or manual reporting (which suffers from intentional or unintentional human errors). In academia, the solutions are based on Wi-Fi [11, 13, 23, 31, 39], LED fixtures [32, 49, 52, 54], and RFID [2, 51]. However, each of them has limitations for a city-wide deployment with more than 12,000 merchants and 109,000 couriers with only commodity smartphones. Wi-Fi based solutions are limited because continuous scanning is required to keep the Wi-Fi list updated, which brings much extra power consumption for courier's smartphones, and more importantly, for merchants without Wi-Fi Access Point devices, it is costly to deploy new ones [9, 30, 33, 41]. LED solutions do not scale up due to hardware modification costs [49]. RFID solutions require additional equipment on both receivers and transmitters.

In this work, we argue the Bluetooth Low Energy (BLE) device [15, 19, 24, 56] is a promising solution to achieve our goal. BLE is not a new technology, and BLE-based iBeacon was introduced by Apple [25] in 2013. However, the new features provided in BLE 5.0 [45] in 2016 (e.g., longer range and faster speeds) offer us the opportunity to build `aBeacon` starting from 2017/05. We deploy 12,109 customized `aBeacon` devices to 12,109 merchants on Alibaba platform in Shanghai. An `aBeacon` device is a low-cost ($8 USD) broadcast-only BLE device, and does not have GPS or cellular/ Wi-Fi connections, so it cannot receive any update, and it also cannot directly communicate with back-end servers. An `aBeacon` device deployed in a merchant constantly broadcasts its ID tuple (UUID, major, minor) following the BLE protocol, which will be received by couriers' smartphone APP if in proximity

[1] https://tianchi.aliyun.com/dataset/dataDetail?dataId=75174

and then uploaded to a server by APP using smartphones' Internet connection. Based on the uploaded ID tuple, the server is aware of the couriers' arrival to this merchant given previously-mapped device-merchant pairs in the deployment.

BLE devices have several advantages. Continuous scanning in BLE only introduces less than 2% extra power consumption on couriers' smartphones based on our experiments, which is much less than Wi-Fi [7]; compared to RFID-based solutions, no hardware modification is needed on the courier end, since `aBeacon` only requires a courier to have a smartphone; compared to LED, battery-powered BLE devices can be installed in many places due to their small size and portability. We note that a key limitation of `aBeacon` is we need to deploy an `aBeacon` device at every merchant, which introduces both hardware and deployment costs. However, our deployment has a low cost since we utilize an Alibaba in-house team and its members visit merchants periodically for business development; the hardware cost of `aBeacon` can be managed if we only remain core functions, e.g., no GPS, no cellular/Wi-Fi, and no Over-The-Air (OTA) updates.

In a control environment, using BLE devices to detect arrival and departure is straightforward. However, it is nontrivial to build, deploy, and operate `aBeacon` from the ground up, considering the tradeoffs between various practical factors, e.g., cost and performance, in a **metric-based approach**. BLE devices are already applied in real-world applications, e.g., interaction in museums [34] and indoor localization in airports [47]. However, we argue that these indoor environments are normally under the control of BLE system operators. Still, the indoor environments for instant delivery (e.g., shopping mall) are not under the company's control, i.e., in the wild. To our knowledge, there are few studies, if any, on a practical city-wide BLE device deployment in the wild. We introduce `aBeacon` based on Alibaba Local Services for courier indoor status monitoring (i.e., arrival and departure) in a 36 month two-phase study from 2017/5 to 2020/4.

- **Phase I: 8-Month Iterative Pilot Study (2017/5-12)**. We deployed three types of commodity BLE devices in 18 merchants and built an APP to test the feasibility of BLE. Based on the promising results, we customized `aBeacon` devices for lower cost and new functions. We deployed one customized `aBeacon` device and one commodity device in 200 merchants to A/B test their performance.
- **Phase II: 28-Month Deployment and Operation in the Wild (2018/1-2020/4)**. We deploy and operate 12,109 `aBeacon` devices in Shanghai with one device in each merchant. In this phase, `aBeacon` interacts with 109,378 couriers to provide their status to assist the scheduling of 64 million delivery orders for unique 7.3 million customers with a total amount of $600 million USD order values.

As of 2020/4, `aBeacon` is being retired and replacing by a new system `aBeacon`+ (introduced in the Discussion section). In this Operational Systems track submission, we focus on 5 lessons we learned in our 3 year study of `aBeacon` from its conception to retirement to provide new insights for the existing design assumptions based on our successes and failures.

Lesson learned 1: Explicitly Quantifying the System Gain. During our interaction with the Alibaba executives team who makes the decision to fund `aBeacon`, we utilize a metric-based approach to quantify `aBeacon`'s monetary gain (i.e., benefit minus cost) to justify `aBeacon`. In particular, we explore the fundamental tradeoff between cost and benefit (proportional to its performance) to optimize the gain of `aBeacon` by (i) reducing the cost by customizing new devices and utilizing our in-house team without technical expertise for configuration-free deployment, and (ii) increasing the benefit by improving lifetime, reliability, and utility. We study the system gain in an evolving cumulative fashion at the fine-grained device level. `aBeacon` achieves the break-even point where its benefit is equal to its cost after 12 months of the deployment, and then generate 14 months of benefits. In retrospect, a batch deployment, instead of an "one-shot" deployment, could make `aBeacon` break even earlier.

Lesson learned 2: System Scale Evolution in the Wild. Even though a device has an expected lifetime of 24 months, `aBeacon`'s scale (i.e., number of live devices) has been constantly shrinking, immediately after fully deployed in the wild, for the entire 26 months of the operation. In particular, the decrease is steady in the first 20 months due to various factors (e.g., deployment, hardware, and merchants closed) yet with a stable citywide spatial coverage; whereas the decrease is dramatic in the last 6 months due to clustered device battery run-outs. This observation has the potential to provide some guidance on the re-deployment strategies (e.g., timing and priority) to keep the system scale and a positive gain (as suggested in the Lesson Learned 1), e.g., large-scale redeployment much earlier than expected battery lifetime. In retrospect, `aBeacon`'s scale shrinking is much worse than our expectation, making us rethink the initial rationale of deploying physical devices in the wild. It motivated us to virtualize the next generation of `aBeacon`, i.e., `aBeacon`+.

Lesson learned 3: Lifetime in the Wild. During the `aBeacon` operation, the lifetime of 42% devices is longer than deployment environment (e.g., a device is live but the merchant it was deployed is closed). However, once deployed in the wild, large-scale recycling of low-cost ($8 USD) devices from these short-lifetime environments is not practical due to significant labor. In retrospect, `aBeacon` devices could be designed with different energy modules for different environment lifetime (e.g., predicted based on our order data) to minimize the hardware cost.

Lesson learned 4: Reliability in the Wild. Many existing sensing systems (e.g., proximity [36], gesture [58], and breath [55]) are mainly tested in control environments with high reliability [22]. However, we found that even for simple arrival detection in `aBeacon` the reliability is heavily affected by many real-world factors including smartphone diversity (e.g., 52 phone brands and 672 phone models in our platform), device placement (e.g., non-expert installation), and courier mobility behaviors. In retrospect, we could add an OTA update function to some devices (but not all devices) deployed in uncertain environments, and utilize couriers' phones to update them, e.g., increasing their transmission powers.

Lesson learned 5: Utility in the Wild. Unlike other infrastructures, e.g., Wi-Fi, we found that in our `aBeacon` operation, the locations with more interactions between couriers and devices may not have higher device deployment utility (quantified by the order delivery rate improvement based on courier detection). In contrast, the locations with higher uncertainty of courier mobility behaviors (e.g., higher floors) lead to a higher utility. In retrospect, we could change our deployment strategies to prioritize more uncertain environment.

Based on the above lessons learned, we discuss our limitations and potential applications of `aBeacon` and then discuss their implications to other systems with long-term broad geospatial coverage (e.g., Edge Computing), and finally share the direction of our ongoing work `aBeacon`+.

2 aBeacon Design Goal

In `aBeacon`, a generic workflow is as follows: (1) devices deployed in indoor merchants to continually broadcast their ID tuples; (2) an embedded BLE scanning module in the Alibaba couriers' smartphone APP (mandatory for all couriers) to receive these ID tuples from devices when in proximity and to upload them to a back-end server using the smartphone Internet connectivity; (3) The server updates couriers' arrival and uses them for various functions, e.g., new order scheduling. Based on this workflow, we introduce our metrics as follows.

2.1 Cost and Performance Metrics

Cost C_{Dev}: The costs of a device in `aBeacon` mainly consist of the hardware cost and the deployment cost (i.e., the shipping and labor cost to deploy a device at a merchant).

Lifetime P^i_{Life}: In our design, we envisioned a lifetime of a device for two years, and then redeploy new devices after two years if (i) `aBeacon` was successful (Yes); (ii) the deployment cost was still low (Yes); and (3) `aBeacon` was still the best solution (No since we have `aBeacon`+). The lifetime of a device i is affected by the design (e.g., battery) and the environment (e.g., the deployed merchant is closed).

Reliability P^i_{Reli}: We quantify a device i's reliability by the percentage of couriers we detected among all arrived couriers. The ground truth of the courier arrival is obtained by the delivery order accounting data. P^i_{Reli} is affected by device deployment, smartphone diversity, and courier mobility.

Utility P^i_{Util}: We quantify the utility of a device i by *overdue delivery rates reduction* for the merchant after i was deployed in it. After a merchant was deployed with a device, the platform can better detect and predict the status of couriers around this merchant, which are used to schedule new orders for this merchant by finding nearby couriers (e.g., a courier just left), thus reducing the overdue rate for this merchant. P^i_{Util} is affected by a merchant's features where i was deployed (e.g., merchant locations, floors).

Table 1: Metric Summary

C_{Dev}:	cost of a device, i.e., hardware & deployment
C_{Over}:	cost of overdue penalty per order, e.g., \$1.
P^i_{Life}:	lifetime of a device i
P^i_{Reli}:	reliability of i
P^i_{Util}:	utility of i
t^i_0:	day of i was deployed
T:	# of days since `aBeacon` deployed
N_t:	# of deployed devices until the tth day
O^i_t:	# of orders at tth day in the merchant with i

2.2 Metric-based Approach for Trade-offs

We utilize a metric-based approach to explore the trade-off between costs and performance by Eq. (1). Assuming it has been T days since `aBeacon` was deployed, the cumulative `aBeacon` gain G_T is given by the difference of (i) the cost C_T of deploying `aBeacon` until the Tth day; and (ii) the benefit (i.e., monetary saving) brought by performance improvement, i.e., overdue reduction due to better detection by `aBeacon`.

$$G_T = \sum_{t=1}^{T}\sum_{i=1}^{N_t} B^i_t - \underline{C_T} \tag{1}$$

where N_t is the number of devices deployed until the tth day ($t \le T$) including live and dead devices. $\underline{C_T} = N_T \cdot C_{\text{Dev}}$ is the cost of all devices until Tth day where C_{Dev} is a device cost. B^i_t is the **Benefit** of a device i in the tth day as

$$B^i_t \;=\; \underline{F_1(\,P^i_{\text{Life}},\, t,\, t^i_0\,)} \;\cdot\; \underline{F_2(\,O^i_t,\, P^i_{\text{Reli}},\, P^i_{\text{Util}},\, C_{\text{Over}}\,)}. \tag{2}$$

$\underline{F_1(P^i_{\text{Life}}, t, t^i_0)}$ indicates whether or not a device i reached its lifetime limit by the tth day. It was calculated by remaining lifetime $P^i_{\text{Life}} - (t - t^i_0)$, where P^i_{Life} is the lifetime of i; t^i_0 is number of days that device i has been deployed. $F_1(P^i_{\text{Life}}, t, t^i_0) = 1$ if $P^i_{\text{Life}} - (t - t^i_0) \ge 0$; $F_1(\cdot) = 0$ otherwise, i.e., no remaining lifetime, so we do not have to consider F_2. $\underline{F_2(O^i_t, P^i_{\text{Reli}}, P^i_{\text{Util}}, C_{\text{Over}})}$ indicates the monetary saving by reduced overdue penalty of the orders detected by i. O^i_t is the number of orders at the tth day in a merchant with i, e.g., 100; P^i_{Reli} is the percentage of the orders whose couriers can be detected by i, e.g., 80%; P^i_{Util} is the reduced overdue rate (compared to the overdue rate before the device was deployed) for all orders whose couriers are detected by i, e.g., 20%; C_{Over} is the overdue penalty per order, e.g., \$1. An example of F_2 is the product of all these terms, i.e., $O^i_t \cdot P^i_{\text{Reli}} \cdot P^i_{\text{Util}} \cdot C_{\text{Over}}$ (e.g., saving is $100 \cdot 80\% \cdot 20\% \cdot \$1 = \$16$).

3 aBeacon Life Cycle Overview

Unlike the wireless systems (e.g., Smart Home IoT) that can be updated by OTA, an `aBeacon` device was not designed to be updated after customization to save the hardware cost. Thus, separated by the time we finished the customization

Table 2: Overview of Two Phases

Phase / Stage & Scale / Goal		Phase I: 8-month Pilot Study (2017/5 – 2017/12)		Phase II: 28-month Operation in the Wild (2018/1 – 2020/4)	
		Conception Stage (2017/5-8) 18 merchants,54 devices	**Customization Stage** (2017/8-12) 200 merchants, 400 devices	**Deployment Stage** (2018/1-3) 12,109 merchants & devices	**Operation Stage** (2018/3-2020/4) Evolving
Cost	Hardware	Commodity (Fig.1)	Commodity ($11, Fig.1, T4) Customized ($8, Fig.2)	Customized	—
	Deployment	Our Team (Fig.1)	Our Team	302 Business Managers	—
Performance	Lifetime	Fig. 1	Commodity (2-3 yrs. advertised) Customized (2 yrs. designed)	*In retrospect, we should have selected the merchants with longer lifetime*	Fig. 4 & 5
	Reliability	98%	Both are Close to 98%	Installation Handbook Provided	Fig. 6-8 & Table 5
	Utility	—	—	Highly Profitable Merchants Selected	Fig. 9-12

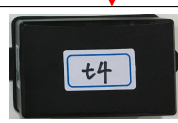
 Size Comparison	 Device 1 (T15)	 Device 2 (T4)	 Device 3 (T11)
Tx Power	-59 dB	-65 dB	-65 dB
Advertised Lifetime	≤ 3 yr	2 ∼ 3 yr	≤ 3 yr
Cost	$11 each	$10 each	$10 each
Encapsulation	—	Water, Dust, Shock Proof	Dust Proof Only

Fig 1: Deployment for Conception Stage

(i.e., 2018/01), we divide the entire 3-year study of `aBeacon` into two phases, i.e., Phase I: 8-Month Pilot Study; and Phase II: 28-month Deployment and Operation.

As in Table 2, we carefully designed the stage, scale, cost, and performance to serve each phase's purposes.

3.1 Phase I: 8-Month Pilot Study (2017/5-12)

In this phase, we performed two studies in a conception stage to investigate three commodity devices, and a customization stage to design and evaluate new devices with A/B testing.

(i) Conception Stage (2017/5-8): As shown in Table 2, we aim to understand whether a BLE device system can detect the couriers' indoor arrival and departure with reliability higher than 95%. We bought 54 commodity devices in three brands and deployed them in 18 merchants of a shopping mall in Shanghai. Each merchant was equipped with three commodity devices of different brands, as shown in Fig.1 with technical specifications. We set some key configurations of couriers' mobile APP when interacting with commodity devices as parameters for further developing, e.g., scanning duration and intervals, data upload cycle, and working hours. Note that the couriers' APP and the back-end server developing were also the major works in this stage, but we omit them in this paper since they are standard. After this study, we had average reliability of 98%, so we concluded that a beacon-based solution could achieve high reliability.

Table 3: BLE Chip Comparison

BLE Chip	Link Budget	Tx Power Consump. (curr. at 0dB)	Sleep Power Consump. (curr.)	Price $/unit
CC2540 [27]	97 dB	21 mA	0.9 ua	∼1.1
DA14580 [43]	93 dB	12.4 mA	0.5 ua	∼1.1
CSR1010 [40]	93 dB	18 mA	5 ua	∼1.1
nRF51822 [44]	**96 dB**	**8.06 mA**	**2.6 ua**	**∼1.1**

(ii) Customization Stage (2017/8-12): Instead of using commodity devices, we customized our `aBeacon` device for low cost ($8 per device) and longer lifetime. We performed a middle-scale A/B testing between the best one among three commodity devices and our customized device. As in Table 2, after the reliability had been proved in the previous stage, our customization was focused on the hardware cost and lifetime since the large-scale city-wide deployment cost in Phase II is marginal when we utilize our in-house business team. In our customization, three components, i.e., BLE chip, battery, and casing, were carefully customized to achieve overall lower cost and longer lifetime. (1) For the BLE chip, we compared the mainstream BLE chips as in Table 3. Since our BLE devices in `aBeacon` were expected to broadcast for at least two years without external power continuously, we chose the nRF51822 from Nordic Semiconductor as the BLE chip since it has both the minimum Tx power and acceptable other configurations. (2) For the battery, we considered both the lithium battery and the alkaline battery since we expected an `aBeacon` device could operate for at least two years without maintenance. The lithium battery usually has a smaller size, but the alkaline battery has a much better unit capacity (mAh/$), so we used two alkaline AA batteries in cascade . (3) For the casing, we considered dust, water, and shockproof for transportation and operation in various indoor (or potential future outdoor) operations. Finally, we built 200 aBeacon devices, as shown in Fig.2. We A/B tested our 200 customized devices with 200 commodity devices (i.e., Device 2 (T4) [10] in Fig.1). We selected 200 merchants in two malls and placed one customized device and one commodity device side by side to compare their performance. After 2-month testing, we concluded our customized devices are ready for deployment and operation because they have similar reliability with the commodity devices, but have a lower hardware cost and potentially longer lifetime.

Fig 2: Customized Hardware for aBeacon

3.2 Phase II: 28-month Operation (18/1–20/4)

We introduce a 3-month deployment and a 25-month operation stage of 12,109 devices in Shanghai (visualized in Fig.3.)

(i) Deployment Stage (2018/1-3). After we received all aBeacon devices from a manufacturer, we aim to deploy them in 12,109 chosen merchants among 57,223 merchants in Shanghai after consulting with our accounting department to understand these merchants' profitability, potentially decides our aBeacon's utility. We decided to deploy around 12 thousand devices for aBeacon because of the approved \$100,000 budget, i.e., the system cost. Assuming no benefit at all, based on Eq. (1), our system gain G_T is $-\$100{,}000$ (i.e., the trivial lower bound in Fig.3 (iii)). In Phase I, our team deployed 200 devices by ourselves, but 12,109 devices were out of our team's capability. As a result, we utilize our in-house regional business development managers who periodically visit all merchants for regular business meetings to install our aBeacon device. We mailed our aBeacon devices and guided them for aBeacon device deployment and mapping between aBeacon devices and merchants by a detailed handbook, which shows "Where to attach the device?" (e.g., main entrance), "How to attach the device?" (e.g., double-sided tape) and "How to map the device?". The mapping was achieved by scanning a QR code on an aBeacon device and then choosing its merchant from a given merchant list in a business manager APP. 302 managers participated in our deployment process, and it took us around two months to deploy all the devices after one month of shipping and logistics. Since our business managers deploy our devices for free, the main deployment cost is the shipping cost, which is around \$1 per device.

(ii) Operation Stage (2018/3-2020/4). After the deployment, aBeacon is fully operational, and we have been monitoring its status and utilizing it to detect couriers remotely based on the data we collected from couriers' APPs. We embed the device monitoring function in the official smartphone APP of 109,378 Alibaba couriers in Shanghai. When we first receive an aBeacon device ID tuple from a courier's phone, we need to make sure this live device is correctly deployed and works properly. For all devices, their initial status on our server end is "Not Deployed"; once a manager completes the mapping operation on her APP, a "Not Deployed"device becomes "Online". For all "Online", we use order accounting data to validate if the deployment is correct or not indirectly: (1) if a device is heard by more than three couriers whose current orders or GPS would not let them pass the merchant mapped to this device, this "Online" device would be changed to "Wrongly Deployed"; (2) if no ID tuples were received from a device for 24 hours, and if the mapped merchant still has orders during these 24 hours (e.g., more than ten orders), then this "Deployed" device would be considered as "Offline" or "Retired" based on its expected lifetime is reached or not since deployment; (3) if ID tuples were received from a device, but the merchant was closed (based on our accounting data), it would be considered "Closed".

Table 4: Operation Data Collected

(a) aBeacon monitoring data

Attribute	Example	Attribute	Example
Courier ID	C_000001	RSSI	-80dB
Timestamp	2019/08/15 12:30:23	Phone ID	D_000001
Device ID Tuple	(UUID, Major, Minor)	Phone Brand/OS	Apple/iOS
Merchant ID	M_000001	Phone Model	iPhone X

(b) Courier GPS data

Attribute	Example
Courier ID	C_000001
Timestamp	2019/08/15 12:30:23
Latitude	31.243715
Longitude	121.245847
Speed	3.7 km/h
Altitude	40.2 meters

(c) Courier order report data

Attribute	Example
Courier ID	C_000001
Timestamp	2019/08/15 12:30:23
Merchant ID	M_000001
Order ID	O_000001
Report Type	Acceptance/Arrival/ Departure/Delivery

(iii) Operation Data Collected. During our operation, we collected three kinds of data sets to monitor and validate aBeacon. **(a) aBeacon Monitoring Data**. As in Table 4(a), every time an aBeacon broadcast was received by a courier's phone, we recorded the information of the aBeacon device, phone, and the Received Signal Strength Index (RSSI) of the broadcast. **(b) Courier GPS data**. As in Table 4(b), GPS data were collected under courier consent since customers also like to know where his/her order is, and the platform needs to know couriers' locations for order assignment. **(c) Courier Order Report Data**. As in Table 4(c), for each delivery order, the courier needs to report when he/she arrives at or leaves from the merchants manually for real-time order status updates. These report data are used as the ground truth for aBeacon detection. However, in our previous study, we found couriers often forgot to report their status and exaggerate their status (e.g., early reporting) to game the scheduling system for better order assignment. That is why these report data can only be used as post-hoc ground truth, i.e., we know that a courier arrived at a merchant after an order was delivered since a courier often forgets or falsely reports their arrival. Please

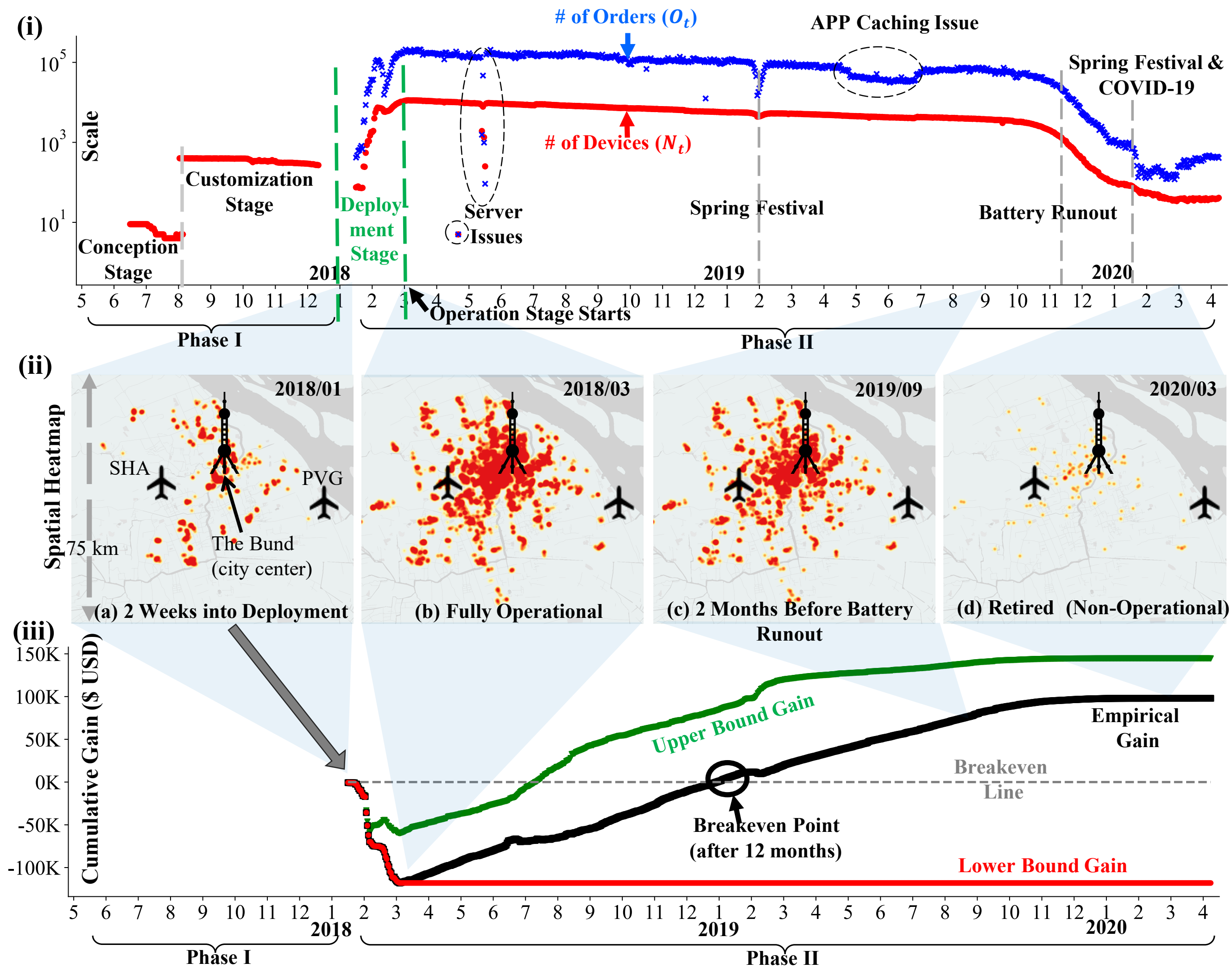


Fig 3: (i) `aBeacon` timeline including Phase I (i.e., conception and customization stage) and Phase II (i.e., deployment and operation stage); (ii) `aBeacon` device heatmaps in Shanghai for four key stages; the background shapes are expended to the corresponding time in (i) and the system gains (iii); (iii) Three `aBeacon` gains with three kinds of utility.

see our Discussion section for details on using `aBeacon` for Anomaly Detection.

4 aBeacon Operation Results

4.1 Result Overview

In Fig.3, we show a panorama of `aBeacon` life cycle with our two phases from 2017/5 to 2020/4.

Quantitative System Evolution Overview: In Fig.3 (i), given a day t, we show both the number of `aBeacon` devices N_t with "Deployed" status in t and the number of delivery orders O_t whose couriers are detected by `aBeacon` in t. We omit the number of couriers detected since it is highly correlated with the number of delivery orders. The detailed analysis on N_t and O_t in Sec.4.3, but we would like to highlight two technical incidents affecting both N_t and O_t as indicated by three circles in Fig.3 (i). On May 16th and 21st, 2018, a configuration exception occurred on the APP server and led to data loss, and our team diagnosed and fixed it quickly. In May 2019, we found an unusual decrease in detected orders, which took our team around two months to diagnose the root cause, i.e., a caching problem in the courier APP of some phone brands. In particular, since the courier APP is not always connected to the server, it would cache some received `aBeacon` device data if the network connection is unavailable. But when the local cache was full, received data got lost without exceptions raised in some smartphones brands. By the end of June, the problem was fixed, and the detected orders increased.

Qualitative Spatial Coverage Evolution: In Fig.3 (ii), we visualize the `aBeacon` spatial evolution in Shanghai at four critical periods. (a) 2018/01: 2 weeks into the deployment stage where `aBeacon` has not been uniformly deployed; (b) 2018/03: `aBeacon` is fully operational, reaches its spatial scale peak, and covers all the central business districts in Shanghai; (c) 2019/09: `aBeacon` has been operating for 20 months, and the spatial cover remains relatively similar, and it is two months away from the starting of clustered battery run-out; (d) 2020/03: aBeacon drops below a critical level and is being retired and replaced by aBeacon+ (see the Discussion section). We found even the scale of aBeacon has been shrinking right after the full deployment due to various real-world issues, the spatial coverage has been relatively stable for 20 months. Based on our field study, we found the most of the dead devices are related to merchants closed, deployment imperfection, hardware malfunction, and vandalism. It provides some practical guidelines for our current project of the next generation of aBeacon, i.e., aBeacon+.

4.2 System Gain Evolution

System Gain Overview: In Fig.3 (iii), we utilize Eq. (1) to show the system gain, i.e., the monetary saving minus the system cost. All metrics in Eq.(1) can be directly measured by our aBeacon data except the system utility P^i_{Util}. We show three cumulative gains (defined in Sec.2.2) based on the empirical value of system utility P^i_{Util} (overdue rates reduction after device i was deployed, discussed in Sec.4.6), along with its lower bound $\underline{P}^i_{\text{Util}}$ (no overdue reduction at all) and its upper bound $\overline{P}^i_{\text{Util}}$ (complete overdue reduction), respectively. We found aBeacon achieved a break-even point after 12 months , which provides empirical guidance for our aBeacon+. Some additional applications of aBeacon for the Alibaba group are shared in the Discussion section.

Lesson Learned 1: **Explicitly Quantifying the System Gain.** Even though the cost of a real-world system can be often explicitly quantified, the benefit of a system is often hard to be, which makes the justification of deploying a system challenging when convincing the decision-makers. Based on our interactions with the Alibaba executive team, who made decisions to initiate and fund aBeacon, we utilized a metric-based approach to quantify the cumulative system gain to justify aBeacon development. In particular, we explore the cumulative system gain by (i) reducing the cost by customizing new devices (e.g., 20% less than commodity devices yet with more functionality) and utilizing our Alibaba in-house business development team without technical expertise for large-scale deployment due to our configuration-free setting, and (ii) increasing the performance by extending device lifetime, improving reliability, and enhancing utility. As shown in Fig.3 (iii), aBeacon achieves a break-even point after 12 months. In retrospect, a few approaches could be used to make sure aBeacon achieves break-even earlier. The most promising one is a batch deployment instead of a "one-shot" deployment in a short time, which have been used in our other physical device deployment projects.

In-depth System Gain Investigation Overview: To provide an in-depth investigation on the cumulative system gain, we analyze seven metrics in Eq.(1) and (2) individually: (i) C_{Dev} and C_{Over} are the individual device cost and the order overdue penalty, which are almost fixed in our setting; (ii) N_t and O_t are related to the system scale and we study them in Sec. 4.3; (iii) P^i_{Life}, P^i_{Reli}, P^i_{Util} are related to the system performance in terms of lifetime, reliability, and utility, which are studied in Sec. 4.4, 4.5, and 4.6, respectively. The correlation between different metrics is introduced in Sec. 4.7.

4.3 Scale Metric: Number of Device & Order

Scale Metric 1: Number of Devices N_t. In Fig.3 (i), starting from our deployment stage in Phase II, the number of aBeacon devices increased significantly until the end of our deployment stage in 2019/3. However, after aBeacon scale peaked in 2019/3, two decreasing trends are observed. (1) The first one is a slow decrease throughout the major part of Phase II from 2018/3 to 2019/10, where we lost some devices every day. In addition to vandalism, deployment, and hardware issues, the primary reason is that some merchants terminate their business with Alibaba every day. The merchant turnover rate in China online platforms is high, and almost 70% of new merchants were closed within one year of the opening in 2017 [20]. We report our empirical merchant lifetime data in Fig.4 and analyze it in detail later. (2) The second one is the sharp decrease from 2019/11 to 2020/2, due to the clustered battery running out after 20 months of operations since 2018/3. Such an observation provided some insights about our potential re-deployment strategies, which we will discuss in the Lesson Learned 2.

Scale Metric 2: Number of Orders O_t. As shown in the Cumulative System Gain Eq.(1), the number of orders O_t whose couriers were detected by aBeacon is the central part of deciding the gain of aBeacon. In Fig.3 (i), we found in the full operation stage of Phase II (from 2018/3 to 2019/11), the number of orders detected is around ten times the number of aBeacon devices, which implies each device serves ten orders on average every day. This ratio remains similar throughout Phase II except for the mid-February, during which the overall number of orders decreases sharply. Mid-February is typically the Chinese Spring Festival, i.e., the biggest holiday where the number of total orders reduced since many merchants closed during this time. We observed sharp decreases and recoveries during February of 2018, 2019, and 2020 in Fig.3 (i), and the corresponding impact on the system gain in Fig.3 (iii). In 2020, the impact of COVID-19 lasts after February, so we do not see an apparent recovery at the end of February.

Lesson Learned 2: **System Scale Evolution in the Wild.** The scale of aBeacon (quantified by the number of devices N_t and the number of associated orders O_t) is essential to ensure the cumulative system gain. As in Fig.3 (i), after fully deployed in the wild (2018/3), aBeacon scale has been continuously shrinking for 26 months until 2020/4, even though devices have an expected lifetime of 24 months. In particular, the decrease is steady in the first 20 months (from 2018/3 to 2019/10) due to various factors (e.g., vandalism, hardware malfunction) yet with a stable city-wide spatial coverage in Shanghai (Fig.3 (ii)). In contrast, the decrease is quite sharp in the last six months (from 2019/10 to 2020/4) due to clustered battery run-out. It suggests that if we want to keep the system scale, we should start a full-scale re-deployment much earlier than expected, or perform batch-based small re-deployment continuously if we want to keep `aBeacon` at scale. However, we did neither of them in practice since we move on to a new system `aBeacon`+ without deployed devices as introduced in future work. Further, by using the number of orders as a bridging factor, our results also provided some insights on how to link the traditional system scale (i.e., number of devices) to the business revenue (i.e., reduced order overdue penalty) to justify their potential correlation. The insights help us communicate with the Alibaba executive team when reporting the impact of `aBeacon` on the overall Alibaba ecosystem.

4.4 Performance Metric 1: Lifetime P^i_{Life}

Lifetime Overview: The lifetime of an `aBeacon` device is decided by two primary factors: (i) the battery size, which was considered in the customization stage of Phase I when we design our hardware; (ii) the merchant lifetime, which unfortunately was not considered in the deployment stage of Phase II as shown in Table 2 since we mainly consider the profitability of merchants. In our defense, there should be a strong correlation between the profitability and lifetime of a merchant, but we found that the profitability of many merchants has been rapidly changing, especially on the online platform. Further, given the high real estate rental fees in Shanghai, many merchants move their physical stores frequently. If a merchant deployed with an `aBeacon` device is closed or moved, there is a high chance that the deployed device would be thrown away. The CDF of devices' and merchants' lifetime are given in Fig.4. Around 23% and 40% of merchants closed within 1 or 2 years, respectively; whereas more than 50% of devices died within one year, much less than the expected lifetime based on battery alone.

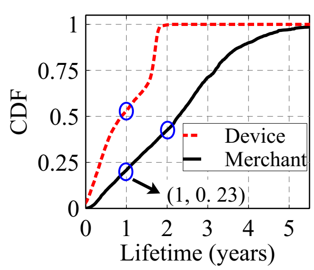


Fig 4: CDF of Device Lifetime and Merchant Lifetime

Lifetime Correlation: An in-depth visualization of the correlation between these two factors is scatter-plotted in Fig.5 where we record the last day a device i was heard as X-axis; the last day the corresponding merchant had orders on our platform as Y-axis. We have three observations: (1) points (15%) around the diagonal ($|x-y| \leq 14$, i.e., 14 days) suggest devices died within two weeks of the closure of the corresponding merchants on our platform; (2) another cluster of points (17%) is around $x = 640$, which means a device is dead after 21 months of operation, as observed in Fig.3 (i). (3) for the points above the diagonal (26%, $x < y$), it indicates the merchant has active orders from our platform, but the `aBeacon` device is dead; for the points below the diagonal (42%, $x > y$), it indicates the merchant closes on our platform (i.e., no orders) but the device can still be heard by couriers, i.e., the device may be in the original locations or nearby, and can be heard when our couriers in proximity. We note that a merchant has no orders on our platform does not necessarily mean the merchant is closed, but it can be used to approximate the merchant's lifetime on our platform. For a closed merchant, an intuitive idea is to recycle the device for re-deployment, but in practice, we did not do it due to two reasons: (1) the platform is not generally informed in advance when the merchant is closing so we cannot prepare in advance to recycle the device; (2) the device recycling introduces significant labor and shipping costs, and the recycled devices may be damaged or with low battery, which makes purchasing a new device a better choice overall. As a result, we did not perform large-scale device recycling in practice.

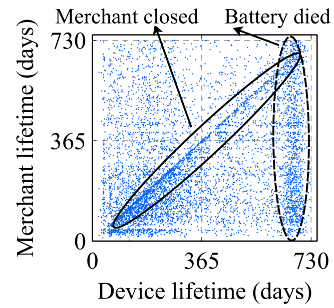


Fig 5: Last day a Device was Heard and Last day the Corresponding Merchant has Orders

Lesson Learned 3: **Lifetime in the Wild.** The lifetime of 42% devices is longer than the lifetime of their deployed environment (e.g., merchants). It provides new insights on our design assumption on mobile device energy since a longer battery life may not increase the device lifetime due to uncertainty of the deployed environment but introduce higher costs. This lesson is especially true when the large-scale device recycling and re-deployment are not practical due to higher labor cost. It motivated us to design devices with different battery capacity and then deploy devices in batches to accommodate the environment's lifetime, which can be predicted by our platform data. We apply this lesson in our `aBeacon`+ where we use merchant phones as our *virtual devices* (instead of deploying physical devices) to broadcast their ID so that the couriers can receive them in proximity. In `aBeacon`+, embedded in merchants' smartphone APPs, the virtual device broadcasting module has different versions, whose parameters were set differently for different merchants.

4.5 Performance Metric 2: Reliability P^i_{Reli}

We quantify the reliability P^i_{Reli} of an `aBeacon` device i with a percentage indicating among all the orders from a merchant deployed with the device i, how many orders we detect. There are three major factors impacting P^i_{Reli}: Stay Duration, Device Deployment, and Smartphone Hardware.

Impact of Stay Duration on P^i_{Reli}. The stay duration is the time between a courier arrives at and departs from a merchant. The stay duration varies due to multiple factors such as the layout of a merchant, the courier's walking speed, and whether an order is ready when the courier arrived, i.e., waiting for the order or not. In Fig.6, we found that the longer that a courier stays, the higher the P^i_{Reli}. The stay duration is computed as the differences of departure and arrival time from the couriers' order report data in Table 4(c) Even though there were inaccurate report data due to human errors, our results are based on 76 million orders for two years, ensuring our results are statistically significant. Two observations can be made from Fig.6: (1) The reliability increases with the staying duration, but does not change much after 7 mins; (2) iOS has a much better performance than Android.

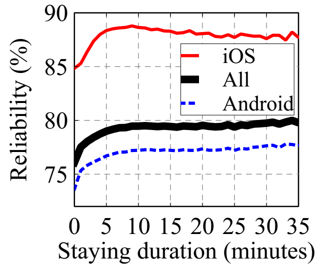


Fig 6: Impact of Stay Duration

Impact of Deployment Position on P^i_{Reli}. The deployment position is an essential factor for reliability, as we found some merchants with an exceptionally low detection ratio. Although our deployment handbook suggested that "Beacons should be attached around the order pickup area", some business managers put devices somewhere else due to various reasons. For example, some merchants do not have a fixed "meal/groceries pickup area"; some merchants prefer the device to be placed somewhere else, e.g., under the counter. We performed some field study, and our findings can be clearly explained with Fig.7 that depicts the layout of a real-world restaurant. In this merchant, there are two entrances with a horizontal wall in the center. Two couriers may pick up orders from both entrances, which leads to the different indoor pickup traces. Unfortunately, because the wall obstructs the device broadcast, only the Courier 1's arrival was detected, which results in a reliability P^i_{Reli} of 46% in our observed period. If the `aBeacon` device were placed in the pickup area, we could have better reliability since both courier traces can be detected. In short, the impact of deployment position is difficult to estimate due to the uncontrollable deployment quality. The reason is we utilize our in-house business team with no deployment expertise (or incentive), and a deployed device can be moved as well, both of which typically leads to low reliability at some merchants.

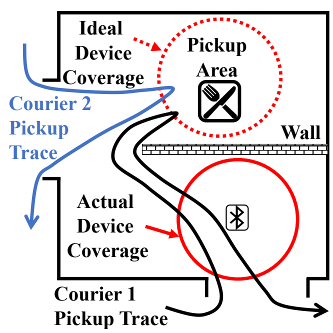


Fig 7: A Field Study of Deployment Position Impact

Impact of Phones Brands and OS on P^i_{Reli}. Our goal is to have most courier smartphones (if not all) to be compatible with `aBeacon` at both the hardware (i.e., phone brands and models) and software level (i.e., OS types). Given more than 109,000 couriers in Shanghai, it is challenging to either force the couriers to use specific smartphone brands or know if a courier uses an un-supported smartphone. To analyze the impact of smartphone OS, we divide all the orders in `aBeacon` merchants into two dimensions: whether its courier was detected by `aBeacon` or not; whether its courier was using an Android or iOS phone. As in Table 5, 63.4% of the orders

Table 5: Detected Ratio of Device OSs over All the Orders

Devices	Detected	Undetected
iOS	13.4%	2.4%
Android	63.4%	20.8%

were detected with the Android couriers (including 52 brands and 672 models), and their average P^i_{Reli} is $\frac{63.4\%}{63.4\%+20.8\%} = 75.2\%$; 13.4% of the orders were detected with the iOS couriers, and their average P^i_{Reli} is $\frac{13.4\%}{13.4\%+2.4\%} = 84.8\%$. We found iOS performs significantly better than Android. For different phone brands (different hardware), the average P^i_{Reli} varies. We show the average P^i_{Reli} of nine well-known brands in China in Fig.8, in which Nexus has the highest P^i_{Reli} of 92%, and iPhones has a P^i_{Reli} of 84%.

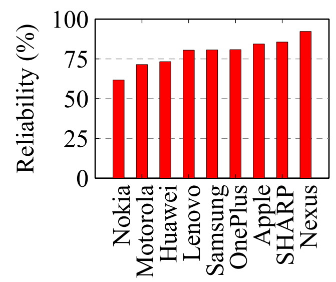


Fig 8: Impact of Smartphone Brand on Reliability

Lesson Learned 4: **Reliability in the Wild.** Many existing wireless sensing systems (e.g., proximity [36], gesture [58], breath [55], human-object interaction [22], and indoor pathway mapping [46]) are mainly tested in the environments with little uncertainty, so they have high reliability. However, we found that even the reliability of a simple presence detection (i.e., courier arrival) is far from guaranteed in a wild, and it is affected by many real-world factors including smartphone software& hardware combination (e.g., 52 phone brands and 672 phone models in Table 5 and Fig.8), and installation position (e.g., low-cost yet unprofessional installation and obstacles in Fig.7), and stay duration (e.g., no waiting time for couriers in Fig.6). In retrospect, we could add an OTA function to some of our devices (but not all devices) deployed in uncertain environments and update them with couriers' phones, e.g., increasing transmission power.

4.6 Performance Metric 3 Utility: P^i_{Util}

The overdue rate reduction is the metric we use to measure the utility P^i_{Util} of deploying an `aBeacon` device i at a particular merchant. For an overdue delivery order (e.g., longer than 30 mins for food), there is an overdue penalty C_{Over} with which the platform will compensate the customer. A typical C_{Over} is \$1, but an overdue penalty could be as high as 200% of the average profit per order if a customer brought delivery insurance. Specifically, the overdue rate is the percentage of the overdue orders among the total orders. So the overdue rate reduction is the difference between the overdue rates before and after an `aBeacon` device deployment. We note that other factors impact the overdue rates of a merchant, e.g., holidays and weathers, but they are out the scope of our paper. We use six months of data before `aBeacon` deployment and 24 months of data after `aBeacon` deployment in the evaluation. There are many features of a merchant that decide the utility of deploying an `aBeacon` device. We study two of them, i.e., Building Floor and City District, due to the space limitation.

Impact of Different Building Floors On P^i_{Util}. To evaluate the impact of deploying an `aBeacon` device on different floors on utility, we aggregate the overdue rate reduction on different floors and compare them with the average overdue rate of all merchants in Shanghai city before and after our `aBeacon` is deployed. The device scale distribution on different floors is given in Fig.9. As shown in Fig.10, the utility is higher

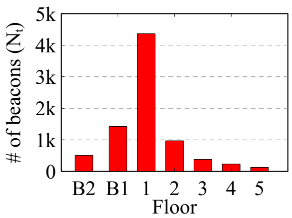


Fig 9: Device Distribution in Different Floors

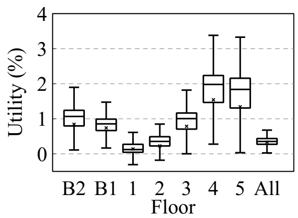


Fig 10: P_{Util} after Deployment in Different Floors

on higher floors or lower basements than the ground floor. This is because the stability of the courier indoor mobility is disproportional to the distance after they enter a building. The higher the floor, the longer the distance, the less stable of courier mobility behaviors (e.g., arrival), the higher benefit for `aBeacon` to detect these behaviors for later order scheduling.

Impact of Different City Districts on P^i_{Util}. To evaluate the impact of districts, we choose five typical districts in Shanghai and compare their average utility, i.e., the overdue rate reduction after `aBeacon` was deployed. As shown in Fig.11, Huangpu is a central business district with a population density of $32,004/km^2$, about three times of New York City ($10,194/km^2$). Songjiang is a suburban area with a population density of $2,892/km^2$, comparable to Los Angeles ($2,910/km^2$). As shown in Fig.12, P^i_{Util} for all merchants with `aBeacon` devices in Songjiang is lower than Shanghai city average; whereas P^i_{Util} in Huangpu is much higher than the average. This is because (1) there are more orders in a more populous area such as Huangpu where each device can serve more orders (we omit the results due to space limitation); (2) the overdue rate is more severe in the city center, and the `aBeacon` can detect couriers more effectively, which leads to better scheduling and thus higher overdue rate improvement.

$\boxed{\textbf{Lesson Learned 5:}}$ **Utility in the Wild.** Unlike other wireless infrastructures, e.g., Wi-Fi, we found the deployment locations with more courier interactions (i.e., demand) during

Fig 11: Five Districts

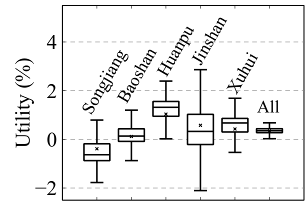


Fig 12: Districts' Utility P_{Util}

our system operation may not have higher utility (quantified by delivery overdue improvement). Instead, we found that the utility of an `aBeacon` device is proportional to the uncertainty of the courier behaviors it can detect (e.g., couriers in higher floors and basements or the downtown area as shown in Fig.9 and Fig.11) because detecting couriers in these uncertain environments can improve the later order scheduling, thus higher utility. This is different from Wi-Fi or cellular device deployment, which are mostly focused on the high user density area. In short, our above findings can provide **practical design guidance** for battery capacity, transmission frequency and power, OTA interface, better installation, and deployment strategies for future wireless systems in the wild.

4.7 Correlation between Different Metrics

Due to the space limitation, we briefly report the results of the correlation between different performance metrics. Our main finding is that for the same `aBeacon` device, when its reliability is low, usually its utility and lifetime are below average; whereas when its reliability is high, the utility is more impacted by the merchant's floors and districts. When reliability $P^i_{\text{Reli}} < 0.5$, this correlation might be caused by improper deployment, which (i) weakens the device utility due to limited data gathered for order scheduling, and (ii) reduces the device lifetime due to potential damage from improper deployment. It also implies that we need to consider other factors if we want to have a longer lifetime and push the utility to the limit when P^i_{Reli} is already high. In our analyses, lifetime and utility are not strongly correlated.

5 Discussions

5.1 Limitations of `aBeacon`

Manual Deployment: For a real-world deployment in the wild, hiring professional teams ensure reliable deployment results but introduces a higher deployment cost per device. In our project, we use our in-house business team to deploy and install `aBeacon` devices at more than 12,000 merchants. Our `aBeacon` system works well in general under this deployment strategy. We admit that our approach may not apply to other settings where such an in-house team is not available. However, we believe our approach can be implemented by Crowdsourcing [42] to deploy wireless devices with a lower labor cost, given little configuration is needed.

No Precise Locations: Another critical limitation of `aBeacon` is that it can only detect couriers' arrival at merchants but cannot perform localization. Given our design goal, a fine-grained localization would be a nice feature to have, but it may not provide higher system gain. Localization increases both hardware cost and deployment cost significantly (it may need onsite configuration or fingerprinting [19]), and may not reduce the overdue rates substantially since the order scheduling only needs coarse-grained locations of couriers.

5.2 System Security

In addition to the hardware cost reduction, another significant improvement in our customization stage is that we enable our devices' security functions. In traditional iBeacon protocol [25], a device ID tuple is fixed for each hardware and broadcast in clear words. It reduces the system complexity while making a device sniffer possible. For example, (1) malicious attackers or unauthorized users (free-riders) can easily restore the device map through war-driving around the devices; (2) if they replicate some device IDs somewhere else, wrong detection information will be collected by `aBeacon`. To address this problem, we designed and implemented a Time-based One-Time Password (TOTP) [53] algorithm to encrypt the device ID broadcast by changing the *major* and *minor* in the ID tuple periodically. A shorter period makes the mapping harder to be restored, whereas a longer period reduces the complexity and the server workload. We set a daily periodical change after exploring the trade-off. The mapping of the device IDs and the merchant locations was stored on the server so that only authorized users can access it. A detailed study of system security is out the scope of this work and will be reported in future work.

5.3 Courier Survey

Feedback is collected from couriers every month regarding multiple aspects, e.g., APP performance, order scheduling, employee care, and penalty appeal. Among the 433 negative feedback on "APP performance" in a recent month, we found the following feedback potentially related to `aBeacon` (# of reports): inaccurate localization (23), slow localization (14), cannot report arrival at the merchants (11), too much battery consumption (8), too much data consumption (2), mandatory Bluetooth on (2). The top three criticisms are all about localization. The underlying reason is that the couriers must report "arrival" at the merchants and customers, and the report must be conducted near (e.g. within 500m) the merchants or customers based on the courier's GPS and the latitude/longitude of the merchant or customer. GPS drifting due to the indoor environment is the main reason for failed reports. The feedback results indicate that alternatives are need besides GPS. `aBeacon` can help in some cases, but we still need to fix the cases when GPS and `aBeacon` fail at the same time. We should also take care of the battery and data consumption.

5.4 Ethics and Privacy

All the data sets are collected under the consent of the couriers. In all our analyses, we have been working on aggregate data. As a result, our results cannot be used to trace back to individuals. The courier ID is an anonymous key to join different data sets, and any other ID information cannot be tracked or identified in practice. We did not utilize personal information from the couriers, e.g., age, gender, income, to protect the couriers' privacy.

5.5 Additional Applications of `aBeacon`

In addition to the direct system gain we measure in this paper, Alibaba has been using `aBeacon` data for a few additional applications based on courier arrival detection.

Order Delivery Time Estimation: The Estimated Time of Arrival (ETA) problem is one of the critical issues in the delivery industry, especially hard for the indoor environment. Based on `aBeacon` data, we obtain travel time between different indoor merchants and build a data-driven model for delivery time prediction, which has been used by other Alibaba teams to predict the overdue rate for the order scheduling.

Merchant Location Correction: Accurate merchant locations are essential in the delivery service. Currently, these locations are provided by merchants themselves and consist of unintentional or intentional errors. Based on the `aBeacon` data, we can measure the travel time between different merchants, cross-validate the accuracy of these locations, and potentially correct them based on massive traveling data.

Anomaly Detection: Unlike GPS data that can be faked on the smartphone [37], `aBeacon` data provide a physical presence confirmation. `aBeacon` data have been used to detect cheating in the delivery process, e.g., frauds conducted by merchants and couriers for the platform subsidy. A detailed courier behavior study measured by manually reported data and automatically collected `aBeacon` data is out of this paper's scope and merits additional investigation.

5.6 `aBeacon+`: Next Generation of `aBeacon`

Since it is expected the maximum lifetime of `aBeacon` is two years, we have been working on a new system called `aBeacon+` built upon `aBeacon` to retain its strengths and address its limitations. In `aBeacon+`, under the merchants' consent, we use merchants' smartphones as `aBeacon` devices instead of deploying `aBeacon` devices, to avoid the hardware and deployment cost. `aBeacon+` does not suffer from vandalism, hardware malfunction, and merchant closures. We embed a broadcasting module in the official merchant APP based on the opportunity that almost every merchant owner needs to install a merchant APP to manage orders. The deployment and operation insights we obtained from `aBeacon` have guided our development of `aBeacon+`, e.g., batch-based deployment and merchant targeting (see our Lessons Learned for details).

We acknowledge the incentives and privacy issues need our attention to make `aBeacon+` practical and salable. However, we argue that the APP users may be willing to provide their locations voluntarily with appropriate incentives in some settings. In our case, a merchant provides this virtual `aBeacon` service on its smartphone. The virtual `aBeacon` can help the platform decide whether an overdue order is because of the merchant's long order preparation time or the courier's late pickup. Similar applications have been launched in Singapore and potentially in the US for public health purposes during the recent COVID-19 pandemic. TraceTogether [35], a BLE based APP developed in Singapore operates similarly to an `aBeacon`+ scheme that users nearby can detect each other for contact tracing purpose in response to COVID-19, which is another example of smartphone users' voluntary participation under some practical incentive.

5.7 Implications on Others Systems

Our study offers some interesting implications for current and future networked systems' design, verification, and operation.

Offline Ground Truth Collection for the Verification of Wi-Fi based Solutions: Along 48% of our merchants have stable Wi-Fi access, `aBeacon` can be used to collect offline ground truth for various applications based on Wi-Fi in the wild to verify existing assumptions or models on Wi-Fi systems and contribute to the community.

Deployment Strategies for 5G and Edge Computing: It has been widely accepted that the extreme densities of base stations and devices are needed to support 5G applications due to its high carrier frequencies with massive bandwidths [3, 8]. Edge computing networks also have a similar setting. Although these systems may need professional teams for the deployment since their devices typically require configuration, our five lessons learned on quantifying system gain, scale evolution, and performance metrics (e.g., lifetime, reliability, and utility) may reduce their indoor operation efforts.

6 Related Works

Table 6: Operational BLE Device Systems

Nation	Deployment Site	Application	Scale
Iceland	Eldheimar museum [34]	Localization	54 devices
U.S.	Beale Street [48]	Presence detection	100 devices
U.K.	Gatwick airport [21]	Localization	2,000 devices
India	Railway station [18]	Presence detection	2,000 devices
Brazil	Tom Jobim airport [4]	Localization	3,000 devices

Operational BLE Device System: To our knowledge, as we are proposing one of the largest BLE device systems in the world, it is worthwhile to give a summary of existing operational BLE device systems. As shown in Table 6, most BLE systems are operated in public sites such as airports or museums for presence detection or indoor localization. The largest BLE system we found is deployed in Tom Jobim airport in Brazil with 3,000 devices, which is fewer than the 12,109 devices in `aBeacon` we deployed in Shanghai, China. More importantly, these existing systems are operated in a controlled environment (e.g., airports, museums, train stations), but our operating environment is in the wild and out of control. It enables our system to provide some new insights from our lessons learned from large-scale system lifetime, reliability, utility, and cost.

BLE Device Studies: Existing BLE system studies can be categorized according to their applications: localization or presence detection. Indoor localization with BLE systems is similar to works done with Wi-Fi. Fingerprinting is studied in [19] to achieve the accuracy of < 4.8 m at the density of one device per 100 m^2, compared with < 8.5 m for Wi-Fi. Map matching is used in [56] to estimate a user's route based on devices with known locations. 1,600 BLE devices are deployed in all the classrooms and corridors of an institute for evaluation. Dynamic RSSI propagation modeling is proposed in [12] to achieve fine-grained (< 2 m) localization and tracking. There are also studies exploring the proximity information provided. Dining hall usage and student check-in are studied in [38] and [24] with BLE device proximity information. Hardware modifications such as energy harvester are also studied in [28] for better performance.

Real-world Sensing Systems: Another related topic is the large-scale real-world sensing system. These studies lay more emphasis on the system implementation and operation for practical problems. LiveTag is proposed in [22] to sense human-object interaction passively. [51] attempts to answer why RFID sensing systems remain research prototypes and have not been widely deployed in practice with theoretical analysis and real-world experiments.

7 Conclusion

This paper introduces `aBeacon`, a wireless indoor BLE device system in Alibaba, from its conception to its retirement by a unique operation study in Shanghai. We quantify `aBeacon`'s performance by scale, lifetime, reliability, and utility, for all of which we provide some new insights obtained in our 3-year system operation in the wild. In particular, we built `aBeacon` from the ground up in a metric-based approach consisting of two phases, i.e., an 8-month pilot study and a 28-month deployment and operation in the wild, including devices in 12,109 merchants and interactions with 109,378 couriers. From the long-term city-wide study, we identify five key observations and lessons regarding system gain quantification, system scale evolution, lifetime, reliability, and utility in the wild. We believe these in-depth lessons learned have implications for other systems requiring long-term operations and broad geospatial coverage such as 5G and Edge Computing.

8 Acknowledgments

We sincerely thank reviewers for their insightful feedback. This work is supported in part by Alibaba Group, Local Service Company, National Key R&D Program of China 2018YFB2100300, 2018YFB0803400, and National Natural Science Foundation of China (NSFC) 61925202, 61772046. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the funding agencies or others.

References

[1] Alibaba Group, Local Service BU. abeacon data-set link. https://tianchi.aliyun.com/dataset/dataDetail?dataId=75174, 2020.

[2] Fadel Adib, Zachary Kabelac, and Dina Katabi. Multi-person localization via rf body reflections. In *NSDI'15 Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation*, pages 279–292, 2015.

[3] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 18(3):1617–1655, 2016.

[4] Airport Benchmarking. Airlines and airports are beaconizing. Webpage, 2016.

[5] Alibaba Group. Alibaba group. Webpage, 2020.

[6] Amazon. Amzon prime now. Webpage, 2020.

[7] Ganesh Ananthanarayanan and Ion Stoica. Blue-fi: enhancing wi-fi performance using bluetooth signals. In *ACM MobiSys*, pages 249–262, 2009.

[8] Jeffrey G Andrews, Stefano Buzzi, Wan Choi, Stephen V Hanly, Angel Lozano, Anthony CK Soong, and Jianzhong Charlie Zhang. What will 5g be? *IEEE Journal on selected areas in communications*, 32(6):1065–1082, 2014.

[9] Niranjan Balasubramanian, Aruna Balasubramanian, and Arun Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *ACM SIGCOMM*, pages 280–293, 2009.

[10] Bright Beacon. Indoor lbs, 2020.

[11] Vladimir Brik, Suman Banerjee, Marco Gruteser, and Sangho Oh. Wireless device identification with radiometric signatures. In *ACM MobiCom*, pages 116–127, 2008.

[12] Dongyao Chen, Kang G Shin, Yurong Jiang, and Kyu-Han Kim. Locating and tracking ble beacons with smartphones. In *ACM CoNEXT*, pages 263–275, 2017.

[13] Krishna Chintalapudi, Anand Padmanabha Iyer, and Venkata N Padmanabhan. Indoor localization without the pain. In *ACM MobiCom*, pages 173–184, 2010.

[14] Deliveroo. Deliveroo. Webpage, 2020.

[15] Patrick Dickinson, Gregorz Cielniak, Olivier Szymanezyk, and Mike Mannion. Indoor positioning of shoppers using a network of bluetooth low energy beacons. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8. IEEE, 2016.

[16] DoorDash. Doordash. Webpage, 2020.

[17] Eleme. Eleme. Webpage, 2020.

[18] Fablian Technologies. Eddystone beacon installation at indian railway stations by google. Webpage, 2018.

[19] Ramsey Faragher and Robert Harle. Location fingerprinting with bluetooth low energy beacons. *IEEE journal on Selected Areas in Communications*, 33(11):2418–2428, 2015.

[20] Fast Casual. Everything's bigger in ... china? Webpage, 2017.

[21] Future Travel Experience. Gatwick's beacon installation provides partners with blue dot navigation and augmented reality wayfinding. , 2017.

[22] Chuhan Gao, Yilong Li, and Xinyu Zhang. Livetag: Sensing human-object interaction through passive chipless wifi tags. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 533–546, 2018.

[23] Andreas Haeberlen, Eliot Flannery, Andrew M Ladd, Algis Rudys, Dan S Wallach, and Lydia E Kavraki. Practical robust localization over large-scale 802.11 wireless networks. In *ACM MobiCom*, pages 70–84, 2004.

[24] Yun Huang, Qunfang Wu, and Yaxing Yao. Bluetooth low energy (ble) beacons alone didn't work! In *ACM UbiComp*, pages 62–65, 2018.

[25] Apple Inc. ibeacon. Webpage, 2020.

[26] InstaCart. Instacart. Webpage, 2020.

[27] Texas Instruments. Cc2540 datasheet [eb/ol](2013-6-24).

[28] Kang Eun Jeon, James She, Jason Xue, Sang-Ha Kim, and Soochang Park. luxbeacon-a batteryless beacon for green iot: Design, modeling, and field tests. *IEEE Internet of Things Journal*, 6(3):5001–5012, 2019.

[29] Junghyun Jun, Yu Gu, Long Cheng, Banghui Lu, Jun Sun, Ting Zhu, and Jianwei Niu. Social-loc: Improving indoor localization with social sensing. In *ACM SenSys*, pages 1–14, 2013.

[30] Bryce Kellogg, Vamsi Talla, Shyamnath Gollakota, and Joshua R. Smith. Passive wi-fi: Bringing low power to wi-fi transmissions. In *2016 USENIX Annual Technical Conference (USENIX ATC 16)*, 2016.

[31] Swarun Kumar, Stephanie Gil, Dina Katabi, and Daniela Rus. Accurate indoor localization with zero start-up cost. In *ACM MobiCom*, pages 483–494, 2014.

[32] Ye-Sheng Kuo, Pat Pannuto, Ko-Jen Hsiao, and Prabal Dutta. Luxapose: Indoor positioning with mobile phones and visible light. In *ACM MobiCom*, pages 447–458, 2014.

[33] Jiayang Liu and Lin Zhong. Micro power management of active 802.11 interfaces. In *ACM MobiSys*, pages 146–159, 2008.

[34] Locatify. Eldheimar museum, intuitively designed beacon based museum audio guide, 2020.

[35] Government of Singapore. Help speed up contact tracing with tracetogether, 2020.

[36] Timothy J Pierson, Travis Peters, Ronald Peterson, and David Kotz. Proximity detection with single-antenna iot devices. In *ACM MobiCom*, pages 1–15, 2019.

[37] Google Play. Fake gps location, 2020.

[38] Rachael Purta and Aaron Striegel. Estimating dining hall usage using bluetooth low energy beacons. In *ACM UbiComp*, pages 518–523, 2017.

[39] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. Widar2. 0: Passive human tracking with a single wi-fi link. In *ACM MobiSys*, pages 350–361, 2018.

[40] Qualcomm. Csr1010 datasheet. Webpage, 2015.

[41] Ahmad Rahmati and Lin Zhong. Context-for-wireless: context-sensitive energy-efficient wireless data transfer. In *ACM MobiSys*, pages 165–178, 2007.

[42] Sanae Rosen, Sung-ju Lee, Jeongkeun Lee, Paul Congdon, Z Morley Mao, and Ken Burden. Mcnet: Crowdsourcing wireless performance measurements through the eyes of mobile devices. *IEEE Communications Magazine*, 52(10):86–91, 2014.

[43] Dialog Semiconductor. Da14580 low power bluetooth smart soc. *DataSheet], https://support. dialog-semiconductor. com/downloads/DA14580_DS _v3*, 1, 2015.

[44] Nordic Semiconductor. nrf51822. *Online at http://www. nordicsemi. com/eng/Products/Bluetooth-R-low-energy/nRF51822, visited Nov*, 2013.

[45] Nordic Semiconductor. Bluetooth 5, 2020.

[46] Guobin Shen, Zhuo Chen, Peichao Zhang, Thomas Moscibroda, and Yongguang Zhang. Walkie-markie: Indoor pathway mapping made easy. In *Presented as part of the 10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*, pages 85–98, 2013.

[47] TechCrunch. Gatwick airport now has 2000 beacons for indoor navigation, 2017.

[48] THINKPROXI. Thinkproxi announces famous beale street implemented beacon technology. Webpage, 2017.

[49] Zhao Tian, Yu-Lin Wei, Wei-Nin Chang, Xi Xiong, Changxi Zheng, Hsin-Mu Tsai, Kate Ching-Ju Lin, and Xia Zhou. Augmenting indoor inertial tracking with polarized light. In *ACM MobiSys*, pages 362–375, 2018.

[50] Uber. Uber eat. Webpage, 2020.

[51] Wang, Ju, Liqiong Chang, Omid Abari, and Srinivasan Keshav. Are rfid sensing systems ready for the real world? In *ACM MobiSys*, pages 366–377, 2019.

[52] Yu-Lin Wei, Chang-Jung Huang, Hsin-Mu Tsai, and Kate Ching-Ju Lin. Celli: Indoor positioning using polarized sweeping light beams. In *ACM MobiSys*, pages 136–147, 2017.

[53] Wikipedia. Time-based one-time password. , 2020.

[54] Bo Xie, Guang Tan, and Tian He. Spinlight: A high accuracy and robust light positioning system for indoor applications. In *ACM SenSys*, pages 211–223, 2015.

[55] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *ACM MobiSys*, pages 54–66, 2019.

[56] Daisuke Yamamoto, Ryosuke Tanaka, Shinsuke Kajioka, Hiroshi Matsuo, and Naohisa Takahashi. Global map matching using ble beacons for indoor route and stay estimation. In *ACM SIGSPATIAL*, pages 309–318, 2018.

[57] Yu Yang, Yi Ding, Dengpan Yuan, Guang Wang, Xiaoyang Xie, Yunhuai Liu, Tian He, and Desheng Zhang. Transparent indoor localization with uncertain human participation for instant delivery. In *ACM MobiCom*, pages 116–127, 2020.

[58] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *ACM MobiSys*, pages 313–325, 2019.