IEICE TRANS. COMMUN., VOL.E104–B, NO.7 JULY 2021

INVITED PAPER *Special Section on Future Directions of Research and Development on Communication Quality*

Traffic Reduction Technologies and Data Aggregation Control to Minimize Latency in IoT Systems

Hideaki YOSHINO[†a)], ***Fellow***, **Kenko OTA**[†], ***Member***, ***and*** **Takefumi HIRAGURI**[†], ***Senior Member***

SUMMARY The spread of the Internet of Things (IoT) has led to the generation of large amounts of data, requiring massive communication, computing, and storage resources. Cloud computing plays an important role in realizing most IoT applications classified as massive machine type communication and cyber-physical control applications in vertical domains. To handle the increasing amount of IoT data, it is important to reduce the traffic concentrated in the cloud by distributing the computing and storage resources to the network edge side and to suppress the latency of the IoT applications. In this paper, we first present a recent literature review on fog/edge computing and data aggregation as representative traffic reduction technologies for efficiently utilizing communication, computing, and storage resources in IoT systems, and then focus on data aggregation control minimizing the latency in an IoT gateway. We then present a unified modeling for statistical and nonstatistical data aggregation and analyze its latency. We analytically derive the Laplace–Stieltjes transform and average of the stationary distribution of the latency and approximate the average latency; we subsequently apply it to an adaptive aggregation number control for the time-variant data arrival. The transient traffic characteristics, that is, the absorption of traffic fluctuations realizing a stable optimal latency, were clarified through a simulation with a time-variant Poisson input and non-Poisson inputs, such as a Beta input, which is a typical IoT traffic model.

key words: *IoT, fog, edge, data aggregation, QoS, latency, control, communication quality, communication traffic*

1. Introduction

With the global spread of the Internet of Things (IoT) applications and services, the numbers of IoT devices such as sensors and actuators connected to the Internet are increasing at an unprecedented rate [1]. The spread of IoT has led to the generation of large amounts of data, which requires massive system resources, that is, communication, computing, and storage resources. To adequately deploy IoT systems, a communication traffic and quality design that balances the efficient use of these resources and quality of service (QoS) is an important issue.

Various IoT use cases have been investigated and a few practical services have already been deployed across wide ranging fields such as factories, agriculture, healthcare, and smart cities [2]. IoT systems realizing such IoT use cases need to support different QoS required, such as reliability, latency, and bandwidth, for each use case and each usage scenario. QoS issues in IoT systems have been the subjects of numerous studies. To categorize QoS approaches in IoT systems, White et al. [3] conducted a systematic mapping of 162 selected papers using a number of automated searches from the most relevant academic databases. They identified that such approaches most often take into account quantitative quality factors such as the reliability, performance efficiency, and functional stability among eight quality factors based on the ISO/IEC quality model for software products [4].

Among the QoS factors, latency is one of the most important metrics regarding the performance efficiency to realize critical IoT applications such as factory automation and a smart grid, which require a latency of 0.25–10 ms and 3–20 ms, respectively [5]. To satisfy such rigorous QoS requirements and realize latency-critical IoT applications, various communication technologies that suppress the end-to-end latency and realize an efficient use of system resources have been proposed.

The 3rd Generation Partnership Project (3GPP) published its release 16 of 5th Generation (5G) New Radio (NR) including enhanced ultra-reliable low-latency communication (URLLC) [6] for mission critical applications such as factory automation, transport industry, and electrical power distribution. The end-to-end latency of factory automation and power distribution as representative use cases for the Release 16 NR URLLC evaluation was set to 2 and 5 ms, respectively [6]. Some of the above-mentioned latency-critical IoT use cases will be covered by 5G URLLC technologies. The highly reliable URLLC services, however, require redundant transmission over the 5G network, which leads to a further increase in data volume. Furthermore, to support the URLLC services, it is necessary to use high-functionality user equipment (UE) as IoT devices. For most IoT applications classified in massive machine type communication (mMTC) or cyber-physical control applications in vertical domains [7], it is necessary to take into account restrictions on the device functions and power consumption. Therefore, without relying on UE functionality, traffic reduction technologies suppressing the massive volume of data transmitted and the end–to–end latency play an important role in realizing most IoT applications.

One of the effective solutions to a traffic reduction is fog [1], [8], [9] or edge computing [10]. Fog computing focuses more on the server and network infrastructure side, whereas edge computing focuses more on the IoT device side [10]. Both technologies have several advantages over traditional cloud computing in certain aspects such as latency, battery consumption, and network bandwidth.

Manuscript received August 28, 2020.
Manuscript revised November 21, 2020.
Manuscript publicized February 4, 2021.

[†]The authors are with Faculty of Engineering, Nippon Institute of Technology, Saitama-ken, 345-8501 Japan.
a) E-mail: yoshino@nit.ac.jp

DOI: 10.1587/transcom.2020CQI0002

Another solution to reducing traffic and latency is data aggregation, which summarizes spatially distributed data and transmits the aggregated-data to the fog/edge node or cloud through the Internet. IoT gateways [11] or M2M gateways [12] generally include data aggregation functions to act as a data bridge between sensors and the fog/edge node or cloud. Data aggregation can be broadly classified into statistical and nonstatistical approaches. Statistical aggregation summarizes multiple data into a statistical value using statistical functions, such as Avg, Sum, Min, and Max. This type of aggregation is effective with regard to failures or anomaly detection in wide-area factories or farms using numerous sensor data. Meanwhile, nonstatistical aggregation bundles multiple data and combines them into a chunk without compression. Packet or frame aggregation in wireless LAN [13]–[15] and VoIP [16] are typical examples of nonstatistical data aggregation.

In our previous study [17], we analyzed two fundamental statistical data aggregation schemes: the constant interval aggregation (CIA) and constant number aggregation (CNA) schemes, and derived the Laplace–Stieltjes transform (LST) of the latency distribution. Furthermore, we clarified the existence of the optimal aggregation interval and number, which minimize the latency, and derived simple and accurate estimation formulae for these parameters. In addition, we analyzed the nonstatistical CNA (nCNA) scheme in [18] and derived the LST of the latency distribution, an approximation of the average latency, and an estimation formula that enables us to determine the optimal aggregation number and an optimal latency that minimizes the average latency. By applying the above estimation formulae for the statistical and nCNA models, we proposed an adaptive aggregation number control when the arrival rate fluctuates over time [19], [20].

In our previous analysis of the statistical aggregation scheme [17], the transmission time of aggregated data was assumed to have an exponential distribution. However, in the analysis of the nonstatistical aggregation scheme [18], it is assumed that the transmission time is a unit distribution comprising a fixed-length header and the number of aggregations deployed by the fixed unit data length. That is, the modeling for both aggregation schemes was not uniform, and it was difficult to compare these schemes with each other under common conditions.

This work is an extension of our previous studies [18], [20] as we propose a unified model for statistical and nonstatistical data aggregation and analyze its latency. We analytically derive the LST and average of the stationary distribution of the latency, and approximate the average latency. Applying the average latency approximation, we propose an adaptive control based on a boundary value estimation formula that minimizes the average latency with respect to the time-varying arrival rate. The transient and average characteristics of the proposed control were compared through a simulation.

The main contributions of this paper are: (1) A unified modeling and latency analysis that can handle general constant number data aggregation in IoT gateways, including both statistical and nonstatistical schemes, (2) a newly derived estimation formula for optimal aggregation number control to minimize latency, and (3) extensive simulation studies to support the proposed modeling, analysis, and control, which can realize acceptable performance for latency-critical IoT applications.

The remainder of this paper is organized as follows: Section 2 provides a literature review on representative traffic reduction technologies: Fog/edge computing and data aggregation in IoT systems. Section 3 focuses on data aggregation in IoT gateways and presents the unified modeling for statistical and nonstatistical data aggregation schemes. Section 4 analyzes the latency of the unified model and derives the exact and approximation formulae. Section 5 applies the analytical results to the adaptive control of the aggregation number and evaluates the transient and average characteristics of the proposed control. Finally, the concluding remarks are provided in Sect. 6.

2. Traffic Reduction Technologies and Related Studies

2.1 Fog/Edge Computing

Fog/edge computing [1], [8], [9] has been introduced to provide services by bringing the available computing and storage resources closer to end-users at the edge of the network from the cloud. With fog/edge computing, the massive data generated by IoT devices can be processed at the network edge instead of utilizing communication resources as well as computing and storage resources at the centralized cloud. Fog/edge computing can provide services with lower latency and greater QoS for IoT applications in comparison with cloud computing.

Lin et al. [9] conducted a comprehensive overview of IoT with respect to the architectures, enabling technologies, security, and privacy issues, and presented the foundation of fog/edge computing-based IoT and applications. They also clarified the relation between cyber-physical systems (CPSs) and IoT. Mouradian et al. [21] presented a comprehensive survey on fog computing not only for IoT applications but also for fields such as content delivery. Piao et al. [22] provided a comprehensive survey of the recent advances of edge caching techniques and their corresponding effects on the network performances in radio access network for 5G.

ETSI Multi-access edge computing (MEC), formerly known as mobile edge computing, is an effective architecture for providing low-latency services by deploying computing resources near base stations [23], [24]. Computation offloading in MEC or fog environments, which assigns computing tasks to either edge/fog nodes or cloud servers, has recently gained attention from researchers [25]–[29]. As examples of conducting a quantitative evaluation of the fog/edge system, Yi et al. [30] built a proof-of-concept fog computing platform and compared the latency and bandwidth provided in the fog and cloud. The round trip time

(RTT) as a metric of latency was reduced from 18.0 to 1.4 ms, and the up-link throughput was improved from 1.8 to 83.7 Mbps by moving the computing resources from the cloud to the fog. Furthermore, Gia et al. [31] proposed an IoT-based health monitoring architecture with fog computing and showed a reduction in latency of 73% and data size of up to 93% when using fog computing. Fog/edge computing does indeed enable reduced traffic, and hence latency, in IoT systems with respect to cloud computing.

2.2 Data Aggregation

Studies on data aggregation in IoT systems originated in the field of RFID data aggregation. Chen et al. [32] and the references therein treat a hierarchical aggregation model for distributed RFID data streams. They propose a QoS-aware framework and dynamic aggregation algorithms.

As we reviewed in the previous study [17], statistical data aggregation has been intensively studied, particularly in the field of wireless sensor networks (WSNs) [33]–[35]. With respect to the 93 data aggregation techniques listed in Table 4, in [33], most of the approaches focus on data aggregation protocols based on network architectures, such as cluster-, tree-, and grid-based networks. The performances of all of these protocols were evaluated through a simulation for the network models, and the latency characteristics at each WSN node have not been fully analyzed. In addition, the latency characteristics of the data aggregation for MTC defined in 3GPP have been analyzed in [36]–[39].

By contrast, as we also reviewed in the previous studies [17], [18], nonstatistical aggregation has been modeled and analyzed in detail using queueing theory [40]–[44]. Hong et al. [40] analyzed a model with a limit on the maximum number of packets that can be aggregated in a single frame. However, the transmission times of the successive frames are dependent, rendering the analysis difficult. Although the LSTs of the queueing time and the total system time were formulated, the closed-form solutions were not provided even for an exponential packet length. Razi et al. [41] analyzed the CIA scheme, and Chen and Zhou [42] and Shrader and Ephremides [43] analyzed the CNA scheme. Even in these analyses, closed-form solutions were not given, and the waiting time approximation of a GI/G/1 queue was applied because the transmission time depends on the amount of aggregated data. Kim et al. [44] precisely analyzed the queueing time for aggregating on–off traffic sources. However, their model did not include the queueing time for the transmission of the aggregated packets.

In the aforementioned related studies, however, analyses were limited to steady state conditions, and the optimal aggregation parameters that minimize latency were not derived. In addition to the above survey, readers can find related studies on latency analysis of data aggregation, including the relationship with the age of information (AoI) in [17] and the batch service queue in [18].

In the following sections, we present a unified model for statistical and nonstatistical data aggregation and analyze its latency.

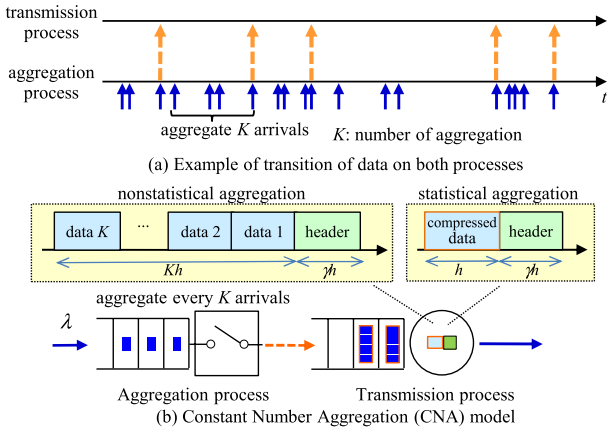

Fig. 1 Constant number aggregation schemes and queueing models.

3. Data Aggregation Schemes and Queueing Models

Figure 1(a) illustrates an example of a transition of data during the aggregation and transmission processes for the nonstatistical and statistical CNA schemes, where data that arrive in the aggregation process are aggregated until the number of arrivals reaches a certain number K. We assume that the aggregated data are immediately sent to the transmission process after the K-th data arrival.

Data aggregation and transmission processes can be represented by a tandem queueing model depicted in Fig. 1(b). The first node is a gate with a buffer, which represents the aggregation process and opens immediately after the number of data arrived reaches K. The second node is a single server first-in-first-out (FIFO) queue, which represents the transmission process. Latency is the total system time defined as the duration between the data reaching the first node and the end of the aggregated data transmission.

In this model, we assume a system with an infinite buffer for both processes and a Poisson arrival at constant rate λ for aggregation. The aggregated data can be treated as a single packet or frame with a header. After queueing during the transmission process, the data are transmitted to an edge device or directly to a cloud server. The transmission time, h[†], per data unit, is assumed to be constant, and the ratio of header transmission time to h, hereinafter called the overhead ratio, is denoted by γ. Statistical data aggregation compresses multiple data into a statistical value of transmission time h using statistical functions. Meanwhile, nonstatistical data aggregation bundles multiple data and combines them into a chunk of transmission time Kh without compression.

Because the service time in the second queue for both statistical and nonstatistical schemes corresponds to the transmission time of the aggregated data along with the

[†]We used the normalized time unit as is common practice in a queueing analysis.

header, which is a constant value, we refer to this model as the dsCNA model, which represents the CNA scheme for aggregation and the deterministic service time for the transmission.

4. Latency Analysis of the dsCNA Model

In this section, we derive the performance measures for the queueing model, namely, the system time (sojourn time) per process and the latency, that is, the sum of the system times of the aggregation and transmission. We define the random variables and LSTs for these measures as follows:

$W_1, F_1^*(s)$: System time during aggregation and its LST;
$W_q, F_q^*(s)$: Waiting time during transmission and its LST;
$W_2, F_2^*(s)$: System time during transmission and its LST;
$W = W_1 + W_2, F^*(s)$: Latency and its LST.

Throughout this paper, we assume that the systems are stationary and ergodic and that the random variables are nonnegative.

4.1 System Times during Aggregation and Transmission Processes

The aggregation process of the dsCNA model corresponds to that of the CNA model derived in our previous study [17]. The LST and average of the system time distribution in the aggregation process of the dsCNA model are respectively given by the following:

$$F_1^*(s) = \frac{\lambda + s}{Ks}\left\{1 - \left(\frac{\lambda}{s+\lambda}\right)^K\right\}, \ \mathrm{Re}(s) > 0, \tag{1}$$

$$\mathrm{E}[W_1] = \frac{K-1}{2\lambda}. \tag{2}$$

The inter-arrival time distribution of the aggregated-data to the transmission process for the dsCNA model is a K-stage Erlang distribution with rate λ/K. Because the service time is constant, the system time W_2 of the dsCNA model can be analyzed using an E_K/D/1 queueing model as given for the nonstatistical CNA model derived in our previous study [18]. That is, the waiting time distribution of the E_K/D/1 queue is equivalent to that of the M/D/K queue under same traffic intensity for both queueing models [45]. Hence, W_2 in the dsCNA model can be exactly determined using the M/D/K queueing model.

The relations among traffic parameters of the E_K/D/1 and M/D/K queues are summarized in Table 1, where the cases of $l = 1$ and $l = K$ correspond to the statistical and nonstatistical dsCNA models, respectively. Therefore, the analysis in this section can handle both models in a unified manner. Applying the above relations, as we derived in [18], for $\mathrm{Re}(s) > 0$, the LST $F_2^*(s)$ and average $\mathrm{E}[W_2]$ of the system time distribution during transmission are given by the following:

$$F_2^*(s) = \frac{\{K-(l+\gamma)\lambda h\}\lambda^{K-1}s}{\lambda^K - (\lambda - s)^K e^{(h+\gamma)hs}} \prod_{k=1}^{K-1}\left\{1 - \frac{s}{\lambda(1-z_k)}\right\}, \tag{3}$$

Table 1 Relations of traffic parameters between E_K/D/1 and M/D/K queues.

	E_K/D/1 model	M/D/K model
Arrival rate	λ/K	λ
Service time	$(l+\gamma)h$	$(l+\gamma)h$
Offered load	$(l+\gamma)\lambda h/K$	$(l+\gamma)\lambda h$
Traffic intensity	$(l+\gamma)\lambda h/K$	$(l+\gamma)\lambda h/K$

$$\mathrm{E}[W_2] = \frac{1}{\lambda}\left\{\frac{\{(l+\gamma)\lambda h\}^2 - K(K-1)}{2\{K-(l+\gamma)\lambda h\}} + \sum_{k=1}^{K-1}\frac{1}{1-z_k}\right\} + (l+\gamma)\lambda h, \tag{4}$$

where $z_k (k = 1, 2, \ldots, K-1)$ are complex roots of the transcendental equation, $1 - z^K e^{(h+\gamma)\lambda h(1-z)} = 0$, except for $z = 1$ [45].

4.2 LST and Average of Latency Distribution

From Eqs. (1) and (3), the LST of the latency distribution for the dsCNA model is derived for $\mathrm{Re}(s) > 0$ as follows:

$$F^*(s) = F_1^*(s) \cdot F_2^*(s) = \frac{\lambda + s}{Ks}\left\{1 - \left(\frac{\lambda}{\lambda+s}\right)^K\right\} \cdot \frac{\{K-(l+\gamma)\lambda h\}\lambda^{K-1}s}{\lambda^K - (\lambda - s)^K e^{(h+\gamma)hs}} \prod_{k=1}^{K-1}\left\{1 - \frac{s}{\lambda(1-z_k)}\right\}. \tag{5}$$

In addition, from Eqs. (2) and (4), the average latency for the dsCNA model is given by the following:

$$\mathrm{E}[W] = \frac{\{K-(l+\gamma)\lambda h+1\}(l+\gamma)h}{2\{K-(l+\gamma)\lambda h\}} + \frac{1}{\lambda}\sum_{k=1}^{K-1}\frac{1}{1-z_k}. \tag{6}$$

4.3 Latency Approximation of dsCNA Model

The exact analysis above not only exhibits a problem during a complex root calculation; it also presents a no-closed form with respect to the parameters characterizing the system performance. Therefore, as in our previous study on the latency of the nCNA model in [18], we apply an approximation for the average waiting time for the M/D/s queue proposed in [46],

$$\mathrm{E}[\tilde{W}_{q(\mathrm{M/D/}K)}] \approx \frac{1}{2}\{1 + f(K,\rho)g(K,\rho)\}\mathrm{E}[W_{q(\mathrm{M/M/}K)}], \tag{7}$$

where

$$f(K,\rho) = \frac{(1-\rho)(K-1)(\sqrt{4+5K}-2)}{16K\rho}, \tag{8}$$

$$g(K,\rho) = 1 - \exp\left\{-\frac{K-1}{(K+1)f(K,\rho)}\right\}, \tag{9}$$

and $\mathrm{E}[W_{q(\mathrm{M/M/}K)}]$ is the exact average waiting time for the M/M/K queue. This approximation formula has been verified and confirmed to be highly accurate [18]. That is, the relative error is less than 1% for $\rho \geq 0.8$, the absolute value of the average waiting time is small, and the effect of the

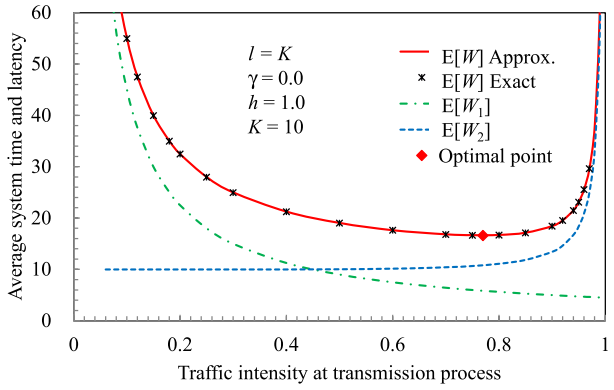

Fig. 2 Average system time and latency for nonstatistical dsCNA model [20].

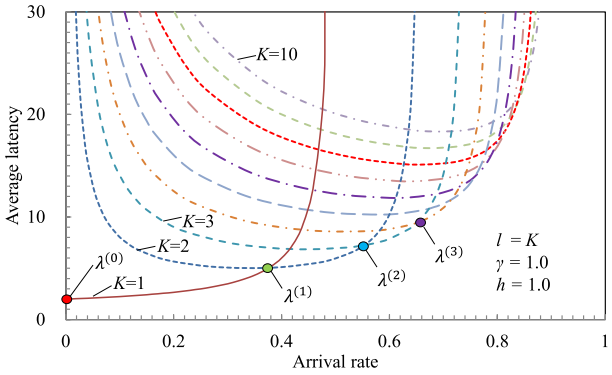

Fig. 3 Average latency versus arrival rate for nonstatistical dsCNA model [20].

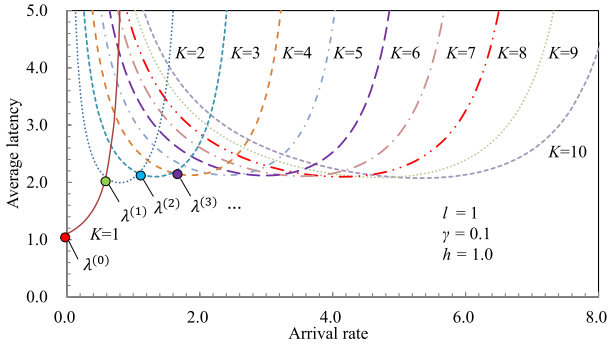

Fig. 4 Average latency versus arrival rate for statistical dsCNA model.

error on the average delay characteristic is relatively low for $\rho < 0.8$. Applying the approximation, we obtained the following approximation formula for the latency of the dsCNA model:

$$\mathrm{E}[W] \approx \frac{K-1}{2\lambda} + \mathrm{E}[\tilde{W}_{q((\mathrm{M/D}/K)}] + (l+\gamma)h. \tag{10}$$

Note that Eqs. (3)–(6) and (10) are extensions of those in Ref. [18] for nonstatistical data aggregation to the unified model, including statistical data aggregation.

Figure 2 shows the average system times $\mathrm{E}[W_1]$ and $\mathrm{E}[W_2]$ and latency $\mathrm{E}[W]$ according to traffic intensity ρ for the nonstatistical ($l = K$) dsCNA model with $\gamma = 0$, $h = 1$, and $K = 10$. Here, $\mathrm{E}[W_2] \approx \mathrm{E}[\tilde{W}_q] + (l+\gamma)h$ and $\mathrm{E}[W]$ were calculated using the approximations in Eqs. (7) and (10), respectively. This figure shows that $\mathrm{E}[W_1]$ is inversely proportional to ρ, $\mathrm{E}[W_2]$ diverges as $\rho \to 1$ as in the average waiting time of general queueing systems, and the latency, that is, the sum of these system times, is consequently a convex function of traffic intensity ρ. The convexity leads to the optimal traffic intensity for minimizing the latency, which is the focus of the following discussion in this study. Figure 2 also shows the exact values for the average latency $\mathrm{E}[W]$ (indicated by symbol $\times$). The exact values were calculated using Eq. (6). From the exact results, we can confirm that the approximation errors in Eqs. (7) and (10) are negligible.

Figures 3 and 4 illustrate the average latencies according to the arrival rate with $h = 1.0$ and $K = 1, \ldots, 10$ for nonstatistical and statistical dsCNA models, respectively, where we set $\gamma = 1.0$ for the nonstatistical model and $\gamma = 0.1$ for the statistical model. This is a condition in which the ratio of the overhead length to the transmitted data length corresponds to 1/10 in both the statistical and nonstatistical models for $K = 10$.

Here, the average delay characteristics for no aggregation (i.e., $K = 1$), which can be obtained using the M/D/1 queueing model, are also plotted in both figures. The adaptive control that takes into account a no-aggregation case, which was not dealt with in our previous studies [17]–[19], is described in the following section.

5. Adaptive Aggregation Number Control and Numerical Results

5.1 Adaptive Control Based on Average Latency Approximation

The results described in the previous section were derived under stationary arrival conditions. In this section, we propose the adaptive aggregation number control for the dsCNA model under the nonstationary arrival cases. A basic idea for the adaptive control proposed here is that even if the arrival rate fluctuates, by adaptively changing the aggregation number according to the average latency characteristics shown in Figs. 3 and 4, the latency can be controlled at the minimum value.

Here, we set a constant measurement interval T, and measure the amount of data that arrive during the aggregation process, that is, λ_i in the i-th ($i = 1, 2, \ldots$) interval. In our previous study [17], we proposed an adaptive aggregate number control that applies the optimal aggregate number estimation formula of the statistical CNA model. Applying the estimation formula for the optimal aggregate number, the aggregate number k_{i+1} in the next interval $i+1$ based on the measured arrival rate in the interval i is calculated through the following formula:

$$\hat{k}_{i+1} = \lceil 1.935\lambda_i/\mu \rceil, \quad i = 0, 1, 2, \ldots, \tag{11}$$

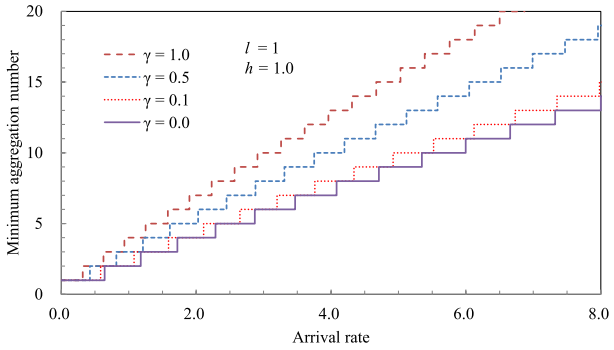

Fig. 5 Minimum aggregation number as a function of arrival rate.

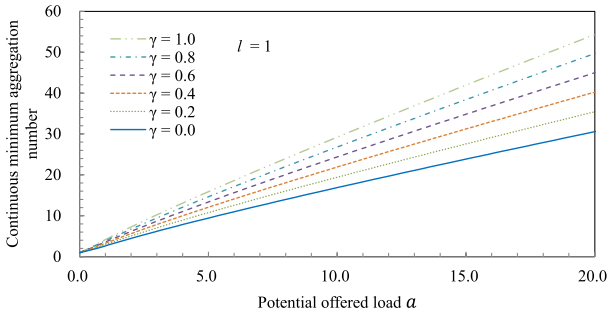

Fig. 6 Continuous minimum aggregation number as a function of potential offered load.

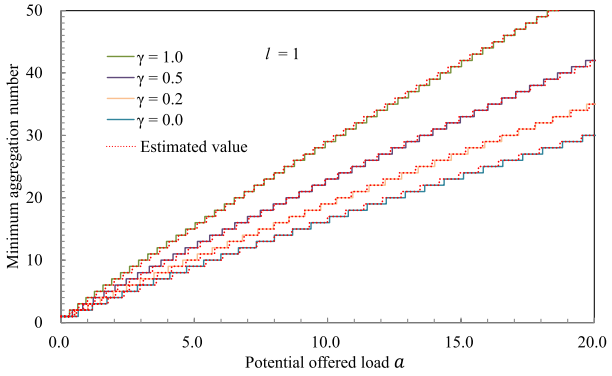

Fig. 7 Estimation accuracy for minimum aggregation number.

where we assumed an exponential service time with a constant rate μ during the transmission process including the header.

By contrast, the proposed control in this paper determines the aggregation number k_{i+1} in the next $i+1$-th interval that minimizes the average latency in Fig. 3 for $l = K$, or for $l = 1$ in Fig. 4 from the arrival rate $\lambda_i (i = 1, 2, \dots)$ in the interval i. That is, the boundary value $\lambda^{(j)} (j = 1, 2, \dots)$ of the arrival rate that gives the intersection of the same average latency for $K = j$ and $K = j+1$ is calculated in advance using Eq. (10), and the aggregation number k_{i+1} of the next $i+1$-th interval that minimizes the average latency is determined by the following logic:

$$if \quad \lambda^{(j-1)} \le \lambda_i < \lambda^{(j)} \quad then \quad k_{i+1} = j, \tag{12}$$

where $\lambda^{(0)} = 0$. Hereafter, we refer to the aggregation number that minimizes the average latency during each interval $[\lambda^{(j-1)}, \lambda^{(j)})$ as a minimum aggregation number. In the following, we deal with the statistical dsCNA model. For a nonstatistical case, please refer to the estimation formula 3 in our previous study [20].

Figure 5 shows the relationship between the minimum aggregation numbers and the boundary values $\lambda^{(j)} (j = 0, 1, 2, \dots)$ of the arrival rates for $h = 1.0$ and $\gamma = 0.1$, 0.5, 1.0, 2.0, 4.0. These graphs are step functions with right continuous and unit step widths. The arrival rates giving the left limits provide the boundary values $\lambda^{(j)} (j = 0, 1, 2, \dots)$.

5.2 Estimation Formula for Optimal Aggregation Number

To obtain an estimation formula for the minimum aggregation number, let each graph in Fig. 5 be regarded as a continuous function according to the potential offered load $a = \lambda h$, and let the vertical axis realize, Fig. 5 can be transformed to Fig. 6. Here, the potential offered load is the offered traffic to the transmission process assuming no aggregation. Applying the relationship shown in Fig. 6 and using the least squares, we derive the estimation formula for the minimum aggregation number as a quadratic function of a as follows:

$$\tilde{k} = \alpha(\gamma) a^2 + \beta(\gamma) a + 1. \tag{13}$$

For the range of parameters of $0 \le a \le 20, 0 \le \gamma \le 1.0$ as shown in Fig. 6, by approximating these factors $\alpha(\gamma)$ and $\beta(\gamma)$ as linear functions of γ, and discretizing the real-valued expression of Eq. (13), we finally obtain the estimation formula for the minimum aggregate number as follows:

$$\tilde{K} = \lfloor \alpha(\gamma) a^2 + \beta(\gamma) a + 1 \rfloor, \tag{14}$$

where

$$\alpha(\gamma) = -0.0076\gamma - 0.0121 \tag{15}$$

$$\beta(\gamma) = 1.319\gamma + 1.722. \tag{16}$$

Figure 7 shows the estimated values (dotted line) by Eq. (14) and the calculated values (solid line) by Eq. (10). We can confirm that the estimation errors are negligible.

Because the estimation formula is based on simple logic, it can be easily implemented using only counter and simple arithmetic operations. This is also an advantage of the proposed control for IoT gateways, where advanced functionality cannot be expected, and a low power consumption is required.

5.3 Transient and Average Characteristics of Proposed Control

Because one of the main features of the proposed control is adaptability to drastic changes in the offered traffic, we simulated the transient characteristics of the control for two types of time-variant input models, that is, linear- and step-types; the arrival rate at the normal state linearly and rapidly

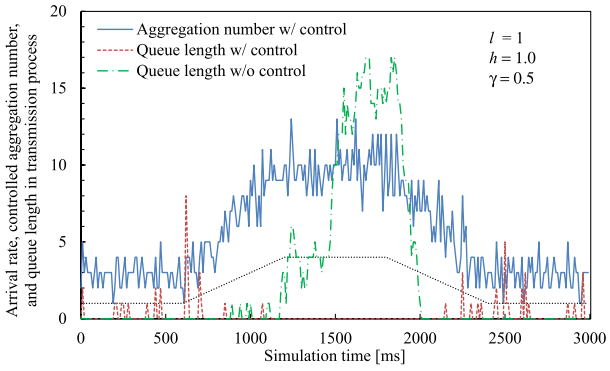

Fig. 8 Transient characteristics of proposed control for linear–type.

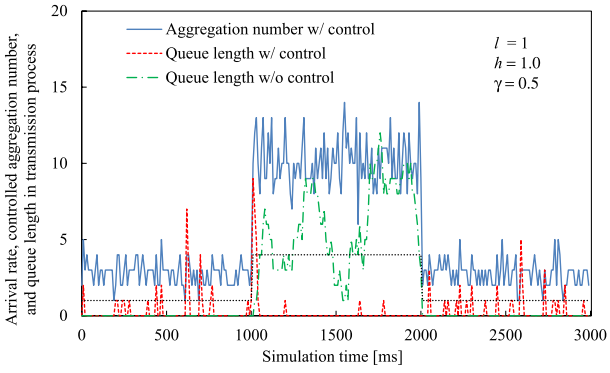

Fig. 9 Transient characteristics of proposed control for step–type.

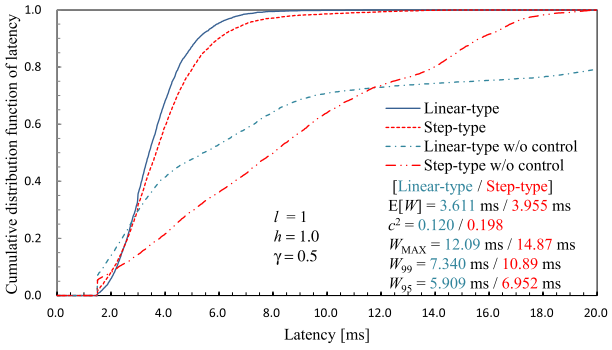

Fig. 10 CDF of latency for linear– and step–types.

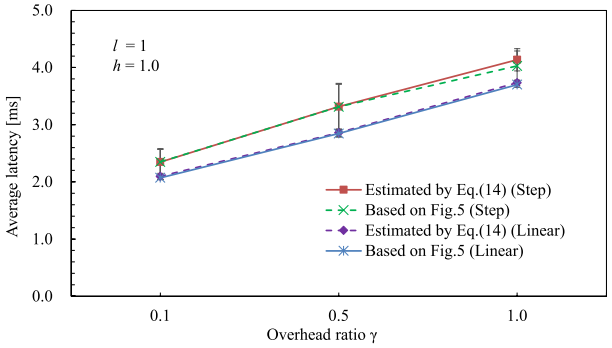

Fig. 11 Average latency as a function of overhead ratio.

becomes overloaded and decreases to the normal state according to a linear step function, as the dotted black lines show in Figs. 8 and 9, respectively. The arrival rates for normal and overloaded states were set to 1.0 and 4.0.

The simulation experiments were run on a workstation (four core 4 GHz CPU with 8 GB of RAM) using an s4 simulation system. We used the following parameters in the simulations.

- measurement interval: $T = 10$ ms,
- simulation time: $T_{sim} = 3$ s,
- $h = 1$ ms, $\gamma = 0.5$, and
- $K = 6$ for cases without control.

Figures 8 and 9 show the simulation results for the transient characteristics of the aggregation number and queue length during the transmission process. These figures indicate that cases without control lead to an increase in the queue length and latency in the transmission process during an overload. By contrast, applying the proposed control, the aggregate number is adaptively controlled according to the increase or decrease in the arrival rate, and the queue length during an overload is kept low. However, the queue length with the proposed control shown in these figures has spikes, and we investigated the overall distribution of the achieved latency in the experiments. Figure 10 shows the cumulative distribution functions (CDF) of the achieved latency for the linear– and step–type experiments in Figs. 8 and 9, respectively. In this figure, c^2, W_{MAX}, W_{99}, and W_{95} represent the squared coefficient of variation, maximum, 99th percentile, and 95th percentile of latency, respectively. We can confirm that the spikes of queue length have little effect on the overall latency distribution, staying within 15 ms at the maximum, and demonstrating a low coefficient of variation. That is, the proposed control achieves stable latency characteristics and can realize acceptable latency requirements of IoT applications, for example, < 20 ms [7].

Figure 11 shows the average latency and 95% confidence interval calculated from 10 simulation runs for a simulation time of 3 s according to the overhead ratio γ as a parameter. The figure also shows the average latency characteristics with control based on Fig. 5 to verify the estimation accuracy of Eq. (14). We can observe that the proposed control achieves stable characteristics close to nearly the theoretically optimal latency, which can be calculated by Eq. (10), as the weighted average latency for the normal state and for the overloaded state. It can also be confirmed that there is almost no estimation error in Eq. (14). Therefore, the adaptive control based on the estimation formula proposed in this paper can be judged to be highly accurate and effective within the range of the verified parameters.

Figure 12 shows the average latency and 95% confidence interval calculated from 10 simulation runs for the simulation time of 3 s according to the mean transmission time $1/\mu$ or $(1 + \gamma)h$ as a parameter. The figure also shows the average latency characteristics with control based on the optimal aggregation number in Eq. (11), as proposed in our previous study [19], as a comparison with the pro-

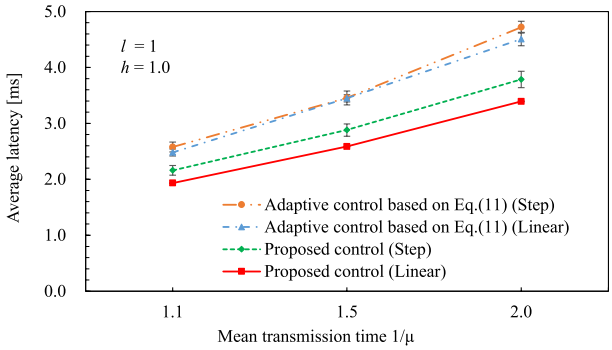

Fig. 12 Comparisons of average latency with our previous study.

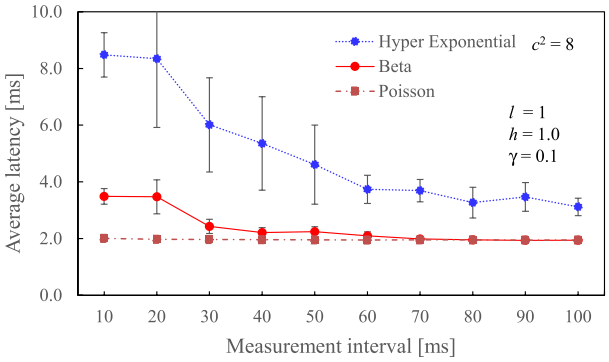

Fig. 13 Average latency for non-Poisson inputs.

posed control. Here, to make the conditions of transmission time the same for comparison, we assume that $h = 1$, and the average transmission times in the previous research, $1/\mu = 1.1, 1.5, 2.0$, correspond to $\gamma = 0.1, 0.5, 1.0$, respectively. We can observe that the average latency can be suppressed to a low level by the proposed control compared to the previous study. The reason why the average latency of the step-type input becomes larger than that of the linear-type is due to the control delay, that is, the arrival rate changes rapidly in the former case and the waiting time for transmission temporarily increases.

Finally, we investigate the influence of IoT data arrival traffic models [47]–[50]. As non-Poisson arrival models of the aggregation process, we used hyper-exponential and Beta distribution models. We assume a two-stage hyper-exponential distribution with balanced mean and its squared coefficient of variation $c^2 = 8$, and a beta distribution model as a periodical sensor data arrival model proposed by 3GPP [47]:

$$p(t) = \frac{t^{\alpha-1}(T_B - t)^{\beta-1}}{T_B^{\alpha+\beta-1} Beta(\alpha, \beta)}, \tag{17}$$

where $p(t)$ represents the arrival intensity in interval $[0, T_B]$, and $Beta(\alpha, \beta)$ represents the Beta function. We assume $T_B = 10$ s, $\alpha = 3$, and $\beta = 4$ as given in [47]. Under these conditions, the cumulative arrival number $A(t)$ is given by [48]

$$A(t) = \frac{20nh(T_B - 1)^3}{T^6}(t + 1)^3, \tag{18}$$

where n represents the number of sensors and h represents the sensor data size. We assume $n = 1000$ and $h = 1$, as given in [48].

Figure 13 shows the average latency and 95% confidence interval calculated from 10 simulation runs for the simulation time of 3 s according to the measurement interval T as a parameter. We can observe that the average latency of the Beta distribution is more stably suppressed than that of the hyper-exponential distribution. As a result, we can confirm that a stable and low-latency control can be realized by appropriately setting the measurement interval even for the periodic IoT traffic inputs.

6. Conclusions

We first presented a recent literature review on fog/edge computing and data aggregation as representative traffic reduction technologies in IoT systems and provided several impressive numerical results. We subsequently proposed a unified modeling for statistical and nonstatistical data aggregation and analyzed its latency. Furthermore, we proposed and evaluated an adaptive control of data aggregation that can be applied in IoT gateways. The simulation results for a time-variant input model including a periodic IoT traffic model indicate that the proposed scheme adaptively changes the aggregation number and absorbs the traffic fluctuations, thereby realizing the minimum stable latency. As the number of connected IoT devices and the accompanying traffic will continue to increase, we believe that traffic reduction technologies have the significant potential to pave the way toward more comfortable and scalable IoT systems in the future. To this end, research on combining fog/edge computing and hierarchical data aggregation can be an interesting future study.

Acknowledgments

This research was supported by the JSPS KAKENHI grant number 17K00133.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," IEEE Commun. Surveys Tuts., vol.17, no.4, pp.2347–2376, Fourthquarter 2015.
[2] ISO/IEC, "ISO/IEC 22417 - Information technology: Internet of Things (IoT) use cases," Technical Report, 2017.
[3] G. White, V. Nallur, and S. Clarke, "Quality of service approaches in IoT: A systematic mapping," J. Syst. Softw., vol.132, pp.186–203, 2017.
[4] ISO/IEC, "ISO/IEC 25010 - Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models," Technical Report, 2010.
[5] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S.A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A.

Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," IEEE Commun. Mag., vol.55, no.2, pp.70–78, 2017.

[6] 3GPP TS 38.824 (2019-03), "Technical Specification Group Radio Access Network; Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC) (Release 16)," March 2019.

[7] 3GPP TS 22.104 (2020-07), "Technical Specification Group Radio Access Network; Service requirements for cyber-physical control applications in vertical domains; Stage 1 (Release 17)," July 2020.

[8] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," Proc. 1st ed. MCC Workshop Mobile Cloud Computing, pp.13–16, Aug. 2012.

[9] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," IEEE Internet Things J., vol.4, no.5, pp.1125–1142, 2017.

[10] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet Things J., vol.3, no.5, pp.637–646, 2016.

[11] A. Papageorgiou, B. Cheng, and E. Kovacs, "Real-time data reduction at the network edge of Internet-of-Things systems," Proc. 11th Int. Conf. Network and Service Management (CNSM), pp.284–291, Nov. 2015.

[12] Y. Nakamura, A. Moriguchi, M. Irie, T. Kinoshita, and T. Yamauchi, "Rule-based sensor data aggregation system for M2M gateways," IEICE Trans. Inf. & Syst., vol.E99-D, no.12, pp.2943–2955, Dec. 2016.

[13] Y. Kim, S. Choi, K. Jang, and H. Hwang, "Throughput enhancement of IEEE 802.11 WLAN via frame aggregation," Proc. IEEE 60th Vehicular Technology Conf., pp.3030–3034, Sept. 2004.

[14] B. Ginzburg and A. Kesselman, "Performance analysis of A-MPDU and A-MSDU aggregation in IEEE 802.11n," Proc. 2007 IEEE Sarnoff Symp., pp.1–5, April 2007.

[15] D. Skordoulis, Q. Ni, H.H. Chen, A.P. Stephens, C. Liu, and A. Jamalipour, "IEEE 802.11n MAC frame aggregation mechanisms for next-generation high-throughput WLANs," IEEE Wireless Commun., vol.15, no.1, pp.40–47, 2008.

[16] K. Kim, S. Ganguly, R. Izmailov, and S. Hong, "On packet aggregation mechanisms for improving VoIP quality in mesh networks," Proc. IEEE 63rd Vehicular Technol. Conf., Melbourne, Vic., pp.891–895, May 2006.

[17] H. Yoshino, K. Ota, and T. Hiraguri, "Queueing delay analysis and optimization of statistical data aggregation and transmission systems," IEICE Trans. Commun., vol.E101-B, no.10, pp.2186–2195, Oct. 2018.

[18] H. Yoshino, K. Ota, and T. Hiraguri, "Optimal parameters of nonstatistical sensor data aggregation minimizing latency in IoT gateway," Proc. IEEE GLOBECOM 2019, Waikoloa, HI, USA, Dec. 2019.

[19] H. Yoshino, K. Ota, and T. Hiraguri, "Adaptive control of statistical data aggregation to minimize latency in IoT gateway," Proc. IEEE Global Info. Infra. and Networking Symp., GIIS 2018, Thessaloniki, Greece, Oct. 2018.

[20] H. Yoshino, K. Ota, and T. Hiraguri, "Adaptive control of nonstatistical sensor data aggregation to minimize latency in IoT gateways," Proc. IEEE 29th Int. Telecom. Networks and Applications Conf., ITNAC 2019, Auckland, New Zealand, Nov. 2019.

[21] C. Mouradian, D. Naboulsi, S. Yangui, R.H. Glitho, M.J. Morrow, and P.A. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," IEEE Commun. Surveys Tuts., vol.20, no.1, pp.416–464, 2018.

[22] Z. Piao, M. Peng, Y. Liu, and M. Daneshmand, "Recent advances of edge cache in radio access networks for Internet of Things: Techniques, performances, and challenges," IEEE Internet Things J., vol.6, no.1, pp.1010–1028, 2019.

[23] ETSI Group Specification MEC 009 V2.2.1 (2020-10), "Multi-access edge computing (MEC); General principles, patterns and common aspects of MEC service APIs," Oct. 2020.

[24] T. Iwai, D. Kominami, M. Murata, R. Kubo, and K. Satoda, "Mobile network architectures and context-aware network control technology in the IoT era," IEICE Trans. Commun., vol.E101-B, no.10, pp.2083–2093, Oct. 2018.

[25] A. Yousefpour, G. Ishigaki, R. Gour, and J.P. Jue, "On reducing IoT service delay via fog offloading," IEEE Internet Things J., vol.5, no.2, pp.998–1010, 2018.

[26] J. Ren, G. Yu, Y. He, and G.Y. Li, "Collaborative cloud and edge computing for latency minimization," IEEE Trans. Veh. Technol., vol.68, no.5, pp.5031–5044, 2019.

[27] M. Sheng, Y. Dai, J. Liu, N. Cheng, X. Shen, and Q. Yang, "Delay-aware computation offloading in NOMA MEC under differentiated uploading delay," IEEE Trans. Wireless Commun., vol.19, no.4, pp.2813–2826, 2020.

[28] A.B. de Souza, P.A.L. Rego, T. Carneiro, J.D.C. Rodrigues, P.P.R. Filho, J.N. de Souza, V. Chamola, V.H.C. de Albuquerque, and B. Sikdar, "Computation offloading for vehicular environments: A survey," IEEE Access, vol.4, pp.198214–198243, DOI: 10.1109/ACCESS.2020.3033828, 2020.

[29] A. Shakarami, M. Ghobaei-Arani, and A. Shahidinejad, "A survey on the computation offloading approaches in mobile edge computing: A machine learning-based perspective," Computer Networks, vol.182, no.9, doi.org/10.1016/j.comnet.2020.107496, 2020.

[30] S. Yi, Z. Hao, Z. Qin, and Q. Li, "Fog computing: Platform and applications," Proc. 3rd IEEE Workshop on Hot Topics in Web Syst. and Technol. (HotWeb), Washington, DC, USA, pp.73–78, Nov. 2015.

[31] T.N. Gia, M. Jiang, A. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Fog computing in healthcare Internet of Things: A case study on ECG feature extraction," Proc. IEEE Int. Conf. Comput. Inf. Technol., Liverpool, UK, pp.356–363, Oct. 2015.

[32] W. Chen, Y. Chen, and S. Wu, "Dynamic aggregate: An elastic framework for QoS-aware distributed processing of RFID data on enterprise hierarchy," IEEE Trans. Parallel Distrib. Syst., vol.25, no.7, pp.1724–1734, 2014.

[33] S. Randhawa and S. Jain, "Data aggregation in wireless sensor networks: Previous research, current status and future directions," Wireless Pers. Commun., vol.97, no.3, pp.3355–3425, 2017.

[34] P. Jesus, C. Baquero, and P.S. Almeida, "A survey of distributed data aggregation algorithms," IEEE Commun. Surveys Tuts., vol.17, no.1, pp.381–404, 2015.

[35] R. Rajagopalan and P.K. Varshney, "Data-aggregation techniques in sensor networks: A survey," IEEE Commun. Surveys Tuts., vol.8, no.4, pp.48–63, 2006.

[36] N. Kouzayha, M. Jaber, and Z. Dawy, "M2M data aggregation over cellular networks: Signaling-delay trade-offs," Proc. IEEE GLOBECOM Workshops 2014, Austin, TX, USA, pp.1155–1160, Dec. 2014.

[37] H. Shariatmadari, P. Osti, S. Iraji, and R. Jäntti, "Data aggregation in capillary networks for machine-to-machine communications," Proc. IEEE PIMRC 2015, Hong Kong, pp.2277–2282, Aug. 2015.

[38] M. Vilgelm and W. Kellerer, "Impact of request aggregation on machine type connection establishment in LTE-advanced," Proc. IEEE WCNC 2017, San Francisco, CA, USA, March 2017.

[39] Y.J. Chen, Z.Q. Wang, and L.C. Wang, "Impact of aggregation factor on delay performance in group-based machine type communications," Proc. IEEE PIMRC 2017, Montreal, QC, Canada, Oct. 2017.

[40] J.H. Hong, O. Gusak, K. Sohraby, and N. Oliver, "Performance analysis of packet encapsulation and aggregation," Proc. 14th IEEE Int. Symp. Modeling, Analysis, and Simulation, Monterey, CA, USA, pp.137–146, Sept. 2006.

[41] A. Razi, F. Afghah, and A. Abedi, "Channel-adaptive packetization policy for minimal latency and maximal energy efficiency," IEEE Trans. Wireless Commun., vol.15, no.3, pp.2407–2420, 2016.

[42] X. Chen and J. Zhou, "Adaptive packet length strategy to minimize

end-to-end latency for time-varying channels," Proc. 12th Int. Conf. Mobile Ad-Hoc and Sensor Networks, MSN 2016, Hefei, pp.125–131, Dec. 2016.
[43] B. Shrader and A. Ephremides, "A queueing model for random linear coding," Proc. IEEE MILCOM 2007, Orlando, FL, USA, pp.1–7, Oct. 2007.
[44] U.H. Kim, E. Kong, H.H. Choi, and J.R. Lee, "Analysis of aggregation delay for multisource sensor data with on-off traffic pattern in wireless body area networks," Sensors (Basel), vol.16, no.10, p.1622, Sept. 2016.
[45] H.C. Tijms, A First Course in Stochastic Models, John Wiley & Sons, 2003.
[46] T. Kimura, "Approximations for the delay probability in the M/G/s queue," Math. Comput. Model., vol.22, no.10–12, pp.157–165, 1995.
[47] 3GPP TR 37.868 (2011-09), "Technical Specification Group Radio Access Network; Study on RAN improvements to Machine-type Communications (Release 11)," Sept. 2011.
[48] X. Chen, Z. Li, Y. Chen, and X. Wang, "Performance analysis and uplink scheduling for QoS-aware NB-IoT networks in mobile computing," IEEE Access, vol.7, pp.44404–44415, 2019.
[49] F. Metzger, T. Hoßfeld, A. Bauer, S. Kounev, and P.E. Heegaard, "Modeling of aggregated IoT traffic and its application to an IoT cloud," Proc. IEEE, vol.107, no.4, pp.679–694, 2019.
[50] M. López-Benítez, C. Majumdar, and S.N. Merchant, "Aggregated traffic models for real-world data in the Internet of Things," IEEE Wireless Commun. Lett., vol.9, no.7, pp.1046–1050, 2020.

Hideaki Yoshino received his B.Sc., M.Sc., and D.Sc. degrees in information science from the Tokyo Institute of Technology, Tokyo, Japan in 1983, 1985, and 2010, respectively. He joined NTT Laboratories in 1985, and has been engaged in communication traffic and service quality research for the past 27 years. He was a visiting scholar at the University of Stuttgart, Germany from 1990 to 1991. He is currently a professor at the Department of Electrical and Electronics Engineering in Nippon Institute of Technology. He served as the Chair of the Technical Committee on Communication Quality, IEICE from 2009 to 2011 and as the Chair of the Technical Committee on Communications Quality and Reliability, IEEE ComSoc from 2014 to 2015. Prof. Yoshino was involved in many international conferences related to communication networks and quality, such as the CQRM symposium co-chair of the IEEE ICC and GLOBECOM. He is a member of IEEE, IEICE (Fellow), and the Operations Research Society of Japan.

Kenko Ota received his B.Eng., M.Eng., and Dr.Eng. degrees from Doshisha University, Japan in 2003, 2005, and 2008, respectively. He was an assistant professor in the Tokyo University of Science, Suwa from 2008 to 2012. He is currently an assistant professor at the Nippon Institute of Technology. His research interests include wireless communication systems and signal processing. Dr. Ota is a member of IEEE, IPSJ, IEICE, and ASJ.

Takefumi Hiraguri received the M.E. and Ph.D. degrees from the University of Tsukuba, Ibaraki, Japan, in 1999 and 2008, respectively. In 1999, he joined the NTT Access Network Service Systems Laboratories, Nippon Telegraph and Telephone Corporation, Japan. He has been involved in the research and development of the MAC protocol for high-speed and high-quality communications in wireless systems. He is currently a Professor in the Nippon Institute of Technology. He is a senior member of IEICE and a member of IEEE.