Design Tradeoffs for Cloud-Based Ambient Assisted Living Systems

Yi Dong*, Yonggang Wen†, Han Hu†, Chunyan Miao*† and Cyril Leung*
*Interdisciplinary Graduate School (IGS) and †School of Computer Science and Engineering (SCSE)
Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
Email: {ydong004,ygwen,hhu,ascymiao,cleung}@ntu.edu.sg

***Abstract*—Ambient assisted living (AAL) has received considerable attention due to its ability to provide services to the elderly by sensors and actuators. However, building such a system is challenging on two fronts. First, the tradeoff between accuracy and monetary cost should be understood. Accuracy of each sensor can be empirically estimated from its sample rate. Typically, higher rate indicates higher accuracy. As a result, higher rate requires more computation resources to process the sampled data, incurring more monetary cost. Second, user needs change frequently. Thus, we need a resource allocation scheme that is (a) able to strike a good balance between accuracy and monetary cost and (b) adaptive enough to meet the frequently changing needs. Unfortunately, several seemingly natural solutions fail on one or more fronts (e.g., simple one shot optimizations). As a result, the potential benefits promised by these prior efforts remain unrealized. To fill the gap, we address these challenges and present the design and analysis of a low-complexity online algorithm to minimize the long-term accuracy-monetary cost on a queue length based control. The design is driven by insights that queue-lengths can be viewed as Lagrangian dual variables and the queue-length evolutions play the role of subgradient updates. Therefore, the control decisions depend only on the instantaneous information and can adapt to the changing needs. Simulations demonstrate that the proposed algorithm can strike a good balance between accuracy and monetary costs. Moreover, the asymptotic optimality of the proposed algorithm has been shown by rigorous analysis and numerical results.**

I. Introduction

Caring for elderly people and the disabled is an important but challenging undertaking. According to United Nations, the number of people aged 60 years or over, is increasing rapidly and will reach nearly 2.1 billion in 2050 [1]. Although the elderly have wisdom and wealth gathered from their life experience, they often require long-term assistance. Ambient assisted living (AAL) systems open up a new way to address the needs of the aged and the disabled by utilizing the capabilities of information and communications technologies [2]. Figure 1 shows the typical reference architecture of cloud based AAL systems. To fulfill the various needs of the elderly, multiple sensors and actuators coexist in ambient assisted living (AAL) systems. The sensed data are processed in the cloud because many applications are computation intensive. The applications infer knowledge from the data, guiding the actuators to meet the needs of the elderly.

For AAL systems, critical factors include the accuracy, monetary cost and delay of the services. Each sensor samples and transmits data at its own rate with a certain accuracy. The accuracy of the sensed data can be estimated [3]. Experiments show that data sensed at high rate usually means high accuracy. However, it demands more computational resources to process the sensed data, leading to higher monetary cost. Moreover, users' requirements change throughout a day and are highly time-varying. Therefore, it is important to design a dynamic scheme to allocate the resources effectively so that the operational costs are reduced and the user experience is satisfactory.

There are several existing efforts on sensor accuracy estimation, AAL systems and mobile cloud. The accuracy estimation for sensor systems has been done mainly through inference [4] and learning [5, 6] techniques. Synthesizing previous works, [3] proposed a framework to estimate the accuracy of certain sensor measurements invoking KL divergence of the distributions. For the AAL systems, most of the works focus on energy efficiency. For example, Wu et al. [7] propose asynchronous flow scheduling taking into account energy efficiency and implementation simplicity. There have been extensive studies on efficient operations of mobile cloud [8–12]. However, none of these publications consider accuracy as an important performance metric and jointly solve this problem.

We investigate the tradeoff between accuracy and monetary costs in a typical AAL scenario with multiple sensors and applications, where the user pattern is not predictable. We propose a low-complexity online algorithm by invoking the Lyapunov optimization and mathematically prove that it minimizes accuracy-monetary cost for given delay constraints. Simulations based on real measurements demonstrate the effectiveness of the proposed algorithm. Our investigation may provide insights into the real world deployment of AAL systems.

Contributions and Roadmap:

- Joint investigation of the accuracy, monetary cost and queueing delay for cloud based AAL systems (§II).
- Proof of theoretical bounds and design of online algorithm for optimal accuracy-monetary cost minimization (§III).
- Evaluations that demonstrate the effectiveness of the proposed algorithm (§IV).

II. System Model and Problem Formulation

In this section, we first introduce the typical cloud based AAL system which is based on real world deployments. Then, we present various models for the cloud based AAL system, including the sensor accuracy model, the queueing model and

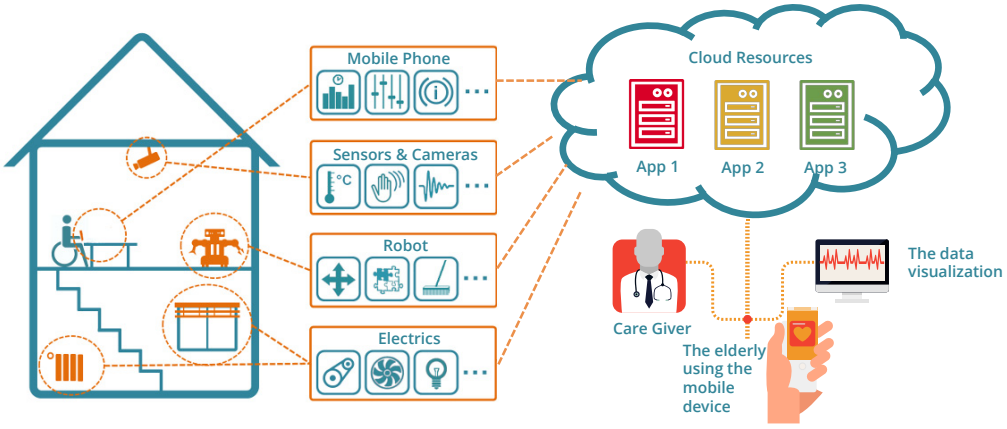


Figure 1. **The reference architecure of a cloud based Ambient Assisted Living system.** An AAL system consists of three parts: data sensing; data mining and processing; control and feedback.

the cloud monetary model. Finally, we formulate the rate selection as a constrained optimization problem. For future reference, we summarize the models and key notations in Figure 2.

A. System Architecture

We aim to design an AAL system which can fulfill the diverse needs of the elderly. For this purpose, various sensors and actuators are deployed in the AAL environment, as shown in Figure 1. Based on this infrastructure, different applications can provide personalized services to the elderly, meeting their specific needs comprehensively.

The typical AAL system consists of three parts: data sensing and transmission, data processing and mining, control and feedback. There exists multiple data sources in the AAL. Each sensor samples the elderly or the environment periodically with a certain rate. The loss of accuracy can be measured by the Kullback-Leiber (KL) distance between the distance of reported data and real value. Generally speaking, the accuracy of the sampled data are proportional to the sample rate. We employ cloud to process the sensed data because of the limited computation capacity of the on-chip processor. Moreover, the cloud works as a central entity to fuse all the sensor data. Due to monetary constraints, we allocate limited computing capacity for processing the incoming data. Therefore, some data are stored in a queue if the data arrival rate exceeds the departure rate. Each sensor maintains one queue. After the data are mined in the cloud, the results are sent back to mobile devices or actuators.

In this system, the main factors that impact user experience are accuracy, monetary cost and delay of the services. Generally speaking, accuracy will improve when the sample and transmission rate increases. However, the increased rate will incur more monetary cost. How to strike a good balance is our main concern. Moreover, we investigate the relationship between the queueing delay and monetary costs. We jointly analyze these factors using a Lyapunov framework. Two knobs are provided for users to tune this AAL system to achieve the desired balance between accuracy, monetary cost and queueing delay.

B. System Model

In the AAL system, there are N sensors which monitor the environment. Sensor i samples and transmits at a rate $r_i(t)$ at time t. We consider a time-sloted system with time slots of equal length indexed by $t = 0, 1, \ldots$. The actual duration of each time slot is application dependent. We now describe sub models as follows.

1) Sensor Data and Loss of Accuracy: The accuracy of sensors can be easily estimated using the observed data. In this paper, we adopt the framework proposed in [3]. The Kullback-Leibler (KL) distance is used to measure the similarity of the distribution of real value and the distribution of reported data. Measurements on the datasets [14] show that there exists a logarithmic relationship between the loss of accuracy and the sample rate in terms of KL distance, formally given by

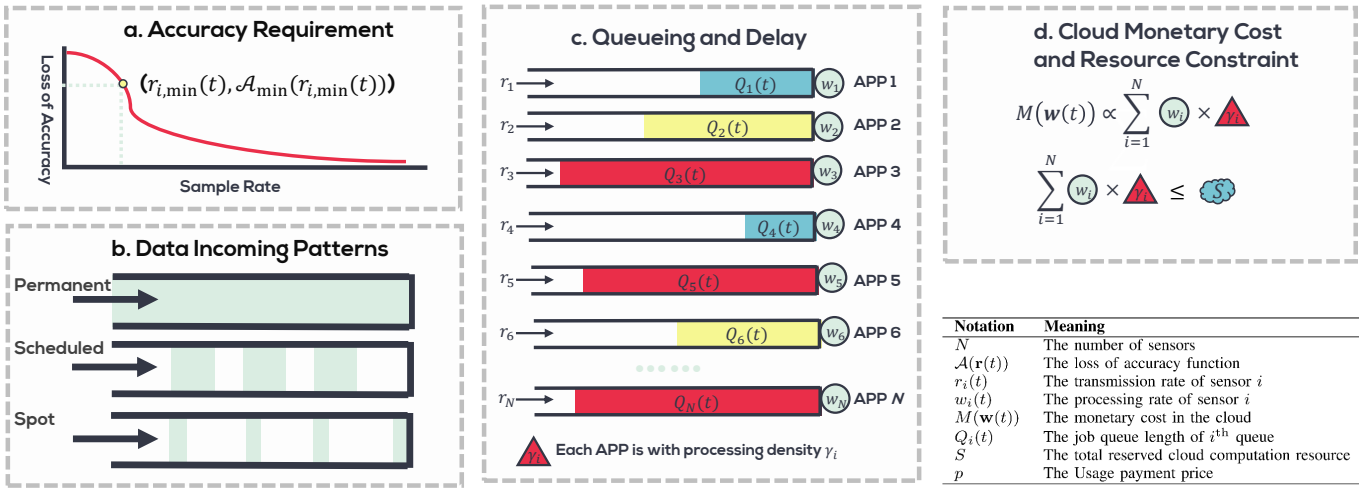
$$A(r_i(t)) = -\log(a_1 r) + a_0\,, \tag{1}$$

where $A(r_i(t))$ denotes the loss of accuracy of sensor i at rate $r_i(t)$ in terms of KL distance. It follows the sensor dependent distribution, which can be measured when the system has been set up.

We aim to set the rate of sensors to meet users requirement of accuracy. In general, there is no closed form expressions about the relationship of the rate and loss of accuracy although it is expected that the result tends to be more accurate if the data is transmitted at a higher rate (with better quality) [15]. In this paper, we assume that the loss of final output accuracy $\mathcal{A}(r_i(t))$ to be a decreasing function of $r_i(t)$, which is similar to (1).

2) Service and Data Incoming Types: Services in an AAL environment can be divided into three types of services based on its user pattern. As shown in Figure 2, we categorize all the services into three types:

- *Permanent Service.* These type of services operate continuously. Therefore we call them permanent services.



Notation	Meaning
N	The number of sensors
$\mathcal{A}(\mathbf{r}(t))$	The loss of accuracy function
$r_i(t)$	The transmission rate of sensor i
$w_i(t)$	The processing rate of sensor i
$M(\mathbf{w}(t))$	The monetary cost in the cloud
$Q_i(t)$	The job queue length of i^{th} queue
S	The total reserved cloud computation resource
p	The Usage payment price

Figure 2. **Models in the cloud based AAL system. a,** Illustration of the relationship between the loss of accuracy and sample rate. We assume a convex of concave function in this paper. **b**, Three typical types of services exist in AAL systems. (i) Permanent service. The type of services should operate throughout the day. (ii) Scheduled service. Some services have strong daily pattern and highly related to user's biological clock. (iii) Spot service. Some services are driven by users' behaiviours, which is unpredictable. The duration of these services tend to be small compared to other services. **c,** The data transmitted by sensor are stored in queues for the cloud to process and analyze. Each sensor maintains one queue. We can use $\mathbf{r}(t) = r_1(t), r_2(t), \ldots, r_N(t)$ to quantify the input/arrival rate to the job queue and use the processing rate $\mathbf{w}(t) = w_1(t), w_2(t), \ldots, w_i(t)$ as the *departure rate* of the queue. Therefore, we introduce the γ_i to denote the processing density of sensor i. Let $Q_i(t)$ be the job queue length, which indicates the queueing delay. It evolves over time. **d,** We assume the reserved pricing model [13] since it fits the need of ambient assisted living environment. Users reserve a certain computation resource S and pay as they go.

For example, the surveillance camera should be running continuously in most cases.

- *Scheduled Service.* Some services have strong daily pattern and highly related to user's biological clock. For example, sleep related monitoring tend to be active only in the night time. These activities can be scheduled at a daily pattern.
- *Spot Service.* Some services are driven by users' behaviours, which is unpredictable. We call it spot services. For example, monitoring of skeletal movement when the user is doing rehabilitation exercises. The durations of these services tend to be small compared to other services.

Since the spot services are unpredictable, we need to develop an algorithm to dynamically select the rates and cloud resource provision to meet the user's requirements.

3) Queue Dynamics: The data transmitted by a sensor are stored in its individual queue for the cloud to process and analyze. Each sensor maintains one queue. We can use $\mathbf{r}(t) = r_1(t), r_2(t), \ldots, r_N(t)$ to quantify the input/arrival rate to the job queue and use the processing rate $\mathbf{w}(t) = w_1(t), w_2(t), \ldots, w_N(t)$ as the *departure rate* of the queue. Let $Q_i(t)$ be the job queue length. It evolves over time following the dynamics specified by

$$Q_i(t+1) = \max[Q_i(t) - w_i(t), 0] + r_i(t)\,, \tag{2}$$

assuming that the initial queue is empty, i.e., $Q_i(0) = 0$.

4) Cloud Computing and Monetary Cost: The cloud process the data in each queue with the rate $w_i(t)$. The CPU cycles for different data and applications are different. For example, facial recognition would be more computation-intensive compared with temperature monitoring. Therefore, we introduce γ_i to denote the processing density of sensor i. Higher processing density means more CPU cycles required for every bit of sensed data. Table I summarizes the processing densities of various tasks which are used in several studies [16–19]. At the t^{th} slot, the required CPU cycles for sensor i per unit time can be calculated as $s_i(t) = w_i(t) \cdot \gamma_i$.

We assume the reserved pricing model [13] since it fits the need of the AAL environment. Users reserve a certain computation resource S and pay as they go. Different sensors and applications require s_i CPU cycles. Therefore, the total CPU cycles needed are limited by the reserved resources, formally given by $\sum_{i=1}^{N} s_i \leq S$. The total cost consists of one-time payment, p_0 and usage payment p for each CPU cycle per unit time. We only consider the usage monetary cost $M(\mathbf{w}(t))$, which can be expressed as

$$M(\mathbf{w}(t)) = \sum_{i=1}^{N} s_i(t) \cdot p = \sum_{i=1}^{N} w_i(t) \cdot \gamma_i \cdot p\,. \tag{3}$$

C. Problem Formulation

Our objective for a AAL system in Figure 1 is to develop a control algorithm. We make the control decision $\mathbf{\Theta}(t) = (\mathbf{r}(t), \mathbf{w}(t))$ at the beginning of each time slot t. Specifically, the sensor i will sample and transmit at $r_i(t)$ according to user's accuracy demand. Then, the cloud will process the transmitted data. Let the cloud processing rate for queue i be $w_i(t)$. Therefore the control decision of the whole system is $\mathbf{\Theta}(t) = (\mathbf{r}(t), \mathbf{w}(t)) = (r_1(t), r_2(t), \ldots, r_N(t), w_1(t), w_2(t), \ldots, w_N(t))$.

The accuracy-monetary cost can be expressed as

$$g(t) = \sum_{i=1}^{N} \mathcal{A}(r_i(t)) + \beta M(\mathbf{w}(t))\,, \tag{4}$$

where $\beta \geq 0$ is referred to as the accuracy-cost parameter indicating the relative importance of the monetary cost of the cloud. The long-term average cost can therefore be written as

$$\bar{g} \triangleq \lim_{t\to\inf} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}[g(\tau)]\,, \tag{5}$$

where the expectation is taken with respect to the random environment. Similarly, we define $\bar{r}_i \triangleq \lim_{t\to\inf} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}[r_i(\tau)]$, $\bar{w} \triangleq \lim_{t\to\inf} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}[w(\tau)]$. To minimize the long-term accuracy-monetary cost, we aim at developing a control policy that solves the following problem

$$\min_{\mathbf{r}(t),\mathbf{w}(t),t=0,1,2,\ldots} \quad \bar{g} \triangleq \lim_{t\to\inf} \frac{1}{t}\sum_{\tau}^{t-1} \mathbb{E}[g(\tau)] \tag{6}$$

$$s.t. \quad \bar{r}_i \leq \bar{w}_i, \tag{7}$$

$$\sum_{i=1}^{N} s_i \leq S\,, \tag{8}$$

where constraint (7) ensures the mean-rate queue stability in the long term, and constraint (8) limited the whole computational resources is under the reserved value.

There exists a stationary and randomized offline control policy solving (6)-(8) based on Caratheodory's theorem. However, it requires full knowledge of the queue arrival statistics and is rather challenging, if not possible, to obtain the optimal control policy. Hence, we develop a practical and online algorithm which optimize the objective slot by slot to achieve the asymptotic optimality.

III. Dynamic Rate Selection Algorithm

We adopt Lyapunov drift plus penalty technique [20–22] to design the algorithm for sensor rate and processing rate selection. Unlike the stationary offline policy, this algorithm only needs the current queue lengths and throughput. Since user patterns (workload arrivals) are not predictable, this fits the use case well.

A. Make Slot-by-slot Objective

We aim to minimize the loss of accuracy and monetary cost as in (6) while stabilizing total task queues in (7). First, to measure the "size" of the vector $\mathbf{Q}(t)$, we define a quadratic Lyapunov function $L(\mathbf{Q}(t))$ as follows.

$$L(\mathbf{Q}(t)) \triangleq \frac{1}{2}\sum_{n=1}^{N} Q_n(t)^2\,. \tag{9}$$

This function can fairly stabilize all the queues, i.e., the queue lengths should be analogously maintained. Then we define the one-slot conditional Lyapunov drift $\Delta(\mathbf{Q}(t))$ as follows.

$$\Delta(\mathbf{Q}(t)) \triangleq \mathbb{E}\{L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t))|\mathbf{Q}(t)\}\,. \tag{10}$$

We aim to minimize the Lyapunov drift $\Delta(\mathbf{Q}(t))$. However, it requires knowledge of prior statistics. Instead, we derive and minimize its upper bound.

B. Derive an upper bound

It can be shown following the Lyapunov optimization technique [20] that

Theorem 1 (Bound of Lyapunov Drift): Consider the quadratic Lyapunov function (9), and assume $\mathbb{E}\{L(\mathbf{Q}(0))\} < \infty$. Then:

$$\Delta(\mathbf{Q}(t)) \leq B + Q(t)\mathbb{E}[\sum_{i=1}^{N} r_i(t) - w(t)|\mathbf{Q}(t)]\,, \tag{11}$$

where B is a finite constant satisfying

$$\begin{aligned} B \geq & \frac{1}{2}\mathbb{E}\Big[\Big(\sum_{i=1}^{N} r_i(t)^2 + w^2(t)|\mathbf{Q}(t)\Big)\Big] \\ & - \mathbb{E}\Big[\Big(\sum_{i=1}^{N} r_i(t)\Big)\cdot \min[Q(t), w(t)]|\mathbf{Q}(t)\Big]\,, \end{aligned} \tag{12}$$

for any $t = 0, 1, \ldots$.

Proof: We first use (9) to derive a bound of the slot-to-slot change in the Lyapunov function:

$$\begin{aligned} L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t)) &= \frac{1}{2}\sum_{i=1}^{2}[Q_i(t+1)^2 - Q_i(t)^2] \\ &= \frac{1}{2}\sum_{i=1}^{N}[(\max[Q_i(t) - w_i(t), 0] + r_i(t))^2 - Q_i(t)^2] \\ &\leq \sum_{i=1}^{N}\frac{[r_i(t)^2 + w_i(t)^2]}{2} + \sum_{i=1}^{2} Q_i(t)[r_i(t) - w_i(t)]\,, \end{aligned} \tag{13}$$

where in the final inequality we have used the fact that for any $Q \geq 0, w \geq 0, r \geq 0$, we have:

$$(\max[Q - w, 0] + r)^2 \leq Q^2 + r^2 + w^2 + 2Q(r - w)\,.$$

Now, plug in (10) and (13), we obtain:

$$\begin{aligned} \Delta(\mathbf{Q}(t)) \leq & \mathbb{E}[\sum_{i=1}^{N}\frac{r_i(t)^2 + w_i(t)^2}{2}|\mathbf{Q}(t)] \\ & + Q(t)\mathbb{E}[\sum_{i=1}^{N} r_i(t) - w(t)|\mathbf{Q}(t)]\,. \end{aligned} \tag{14}$$

Then, define B as a finite constant that bounds the first term on the right-hand-side of the above drift inequality. We can see that

$$\begin{aligned} B \geq & \frac{1}{2}\mathbb{E}\Big[\Big(\sum_{i=1}^{N} r_i(t)^2 + w^2(t)|\mathbf{Q}(t)\Big)\Big] \\ & - \mathbb{E}\Big[\Big(\sum_{i=1}^{N} r_i(t)\Big)\cdot \min[Q(t), w(t)]|\mathbf{Q}(t)\Big]\,. \end{aligned}$$

■

By adding $V\mathbb{E}[g(t)|\mathbf{Q}(t)]$ on both side of (11), we obtain the drift plus monetary cost upper bound as follows.

$$\begin{aligned} \Delta(\mathbf{Q}(t)) + V\mathbb{E}[g(t)|\mathbf{Q}(t)] \leq & B + V\mathbb{E}[g(t)|\mathbf{Q}(t)] \\ & + Q(t)\mathbb{E}[\sum_{i=1}^{N} r_i(t) - w(t)|\mathbf{Q}(t)]\,, \end{aligned} \tag{15}$$

where $V > 0$ is a parameter indicates how close the long term average achieved is to the optimal offline algorithm.

Remark 1: V is a cost-delay tradeoff parameter. Intuitively, we cannot obtain both the minimal queueing length and the minimal cost. If more data are processed in the cloud, the queue length will be shortened, which can meet the critical deadline of delay sensitive applications. But the cloud has to provide more computation resource which incurs more monetary cost. Parameter V is introduced to indicate the importance of the cost.

C. Design the Algorithm

We design an online algorithm based on the upper bound (15) of Lyapunov drift plus monetary cost. The algorithm is formally presented in Algorithm 1. Our algorithm only require the knowledge of current queueing states and workload arrivals. If does not require any statistics about future states or workload arrivals.

Algorithm 1 Online Algorithm for Accuracy Monetary Trade-off Operation

1: At the beginning of every time slot t, observe the queue lengths $\mathbf{Q}(t)$ and the actual incoming packets.

2: The sensor chooses transmission rate $r_i(t)$ for the following optimization problem.

$$\min_{r_i(t),t=0,1,2,\ldots} \quad V \cdot \mathcal{A}(r_i(t)) + Q_i(t) \cdot r_i(t) \tag{16}$$
$$s.t. \quad r_i(t) \in [0, r_{i,\max}]\,.$$

3: The cloud chooses processing rate $w_i(t)$ for the following optimization problem.

$$\min_{\mathbf{w}(t),t=0,1,2,\ldots} \quad V\beta \sum_{i=1}^{N} M(w_i(t)) - \sum_{i=1}^{N} Q_i(t) \cdot w_i(t) \tag{17}$$
$$s.t. \quad \sum_{i=1}^{N} \bar{w}_i \le S$$
$$w_i \ge 0\,.$$

4: Update the job queue $\mathbf{Q}(t)$ as (2).

D. Theoretical Analysis

Here we prove that we can develop a processing rate selection algorithm to achieve near optimal performance in terms of average accuracy-monetary cost by appropriately tuning the parameter V. The following theorems highlight our key finds. Specifically, Theorem 2 shows the effectiveness of minimizing the objective in (6). Theorem 3 further ensures the queue stability in (7).

Theorem 2 (Bound of Long-term Average Accuracy-monetary Cost): The long-term average accuracy-monetary cost is bounded as follows.

$$\bar{g} = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[g(\tau)] \le \bar{g}^{\mathrm{opt}} + \frac{B}{V} \tag{18}$$

Table I
PROCESSING DENSITIES OF VARIOUS APPLICATIONS FOR THE CLOUD-BASED AAL SYSTEMS.

Application	Processing density (cycles/bit)
Face recognition [16], [17]	31680 [16], 2339 [17]
Anomaly Detection [19]	240
Video transcoding [18]	200 - 1200

Table II
DEFAULT PARAMETER VALUES

Parameter	value	Parameter	value
N	10	V	0.01
β	1	γ_i	i^2
T	200	S	1000000
p	1	a_1	100
a_0	15	$r_{\max}$	100

Proof: Please see Appendix A. ■

Remark 2: This theorem shows that we can achieve near optimal of the long term average accuracy-monetary cost by setting V to a large value.

Theorem 3 (Queue Stability): All queues $Q_i(t)$ are mean rate stable.

Proof: Please see Appendix B. ■

Theorem 4 (The Bound of Average Job Queue):

$$\lim_{t\to\infty} \sup \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[Q(\tau)] \le \frac{B + V \cdot D}{\epsilon}\,, \tag{19}$$

where $D = g(\epsilon) - \bar{g}^{\mathrm{opt}}$ is an average accuracy-monetary cost such that $\sum_{i=1}^{N} \bar{r}_i \le \bar{w} - \epsilon$ for some $\epsilon > 0$ (i.e., Slater's condition).

Proof: The proof follows the main results in the proofs of Theorem 2 and Theorem 3. For brevity, we omit the details here. We encourage the interested reader refer to [20]. ■

The design of Algorithm 1 follows our derived analytical results. All the previous theorems can hold for this algorithm. Numerical results further showed the effectiveness of our proposed algorithm.

IV. PERFORMANCE EVALUATION

A. Simulation Setup

Real accuracy measurement. To obtain realistic functions of the accuracy and rate in Section II, we measure the accuracy parameters for several types of sensors [14].

Datasets (workloads and processing density). We create the workload/traffic arrivals by generating Bernoulli arrivals. Moreover, the processing densities of various tasks are summarized in Table I, which are used in several studies [16–19].

Parameters. The default parameters are summarized in Table II.

B. Simulation Results

In this section, we will show the effectiveness of the proposed algorithms and demonstrate the impact of various system parameters. Please note the "average" number at time t in

the following simulations is calculated by summing up all the past values (include time t), then divided by $t+1$.

*1) **Cost-delay Tradeoff Parameter** V:* As shown in Figure 3, parameter V can effectively strike the trade-off between delay and accuracy plus monetary cost functions. During this set of simulation, we set $\beta \cdot V = 0.000001$. The figures show that as V increases, both the loss of accuracy and monetary cost decrease, while the queue length increases, which means average queueing delay increases.

*2) **Accuracy-Monetary Cost Tradeoff Parameter** β:* The results in Figure 4 show that β can provide a flexible trade-off between the loss of accuracy and monetary cost. When β increases, the monetary cost drops while the loss of accuracy will increase. The observations confirm our objective function with β to indicate the relative importance of monetary cost.

*3) **Asymptotic optimality of the proposed algorithm**:* Simulation results show that our algorithm can achieve near optimal average accuracy-monetary cost at the cost of larger queue length.

V. Conclusion

We studied the three most cared metrics in AAL systems and developed an algorithm to help users to achieve the balance among accuracy, monetary cost and delay. Specifically, user can tune knob parameters V and β to operate the AAL system in a desirable way. Moreover, since the proposed algorithm only considers the current queue state and workload arrivals, it can coordinate the system well in time varying environment, which is common in practical AAL systems. The effectiveness of the proposed algorithm is proved by rigorous analysis and simulation results.

Appendix A: Proof of Theorem 1

In Algorithm 1, we minimize the right hand side of the upper bound on the Lyapunov drift plus the accuracy-monetary cost. Thus, for every time slot t, we have

$$\begin{aligned}&\Delta(\mathbf{Q}(t)) + V\mathbb{E}[g(t)|\mathbf{Q}(t)] \\ &\quad\le B + V\mathbb{E}[g^*(t)|\mathbf{Q}(t)] + Q(t)\mathbb{E}[\sum_{i=1}^{N} r_i^*(t) - s^*(t)|\mathbf{Q}(t)]\,,\end{aligned} \tag{A1}$$

where $r_i^*(t)$ and $w^*(t)$ are the arrival rate to the job queue and processing resource provisioning under any alternative control policies, respectively. $g^*(t)$ is the resulting cost. There exist a stationary and randomized offline control policy $\mathcal{Z}$, which can minimize the accuracy-monetary cost. The following holds under policy $\mathcal{Z}$,

$$\mathbb{E}[g^*(t)|\mathbf{Q}(t)] = \mathbb{E}[g^*(t)] = \bar{g}^{opt}, \tag{A2}$$

$$\mathbb{E}[\sum_{i=1}^{N} r_i^*(t) - w^*(t)|\mathbf{Q}(t)] \le 0. \tag{A3}$$

Next, by plugging (A1)-(A3), we can get

$$\Delta(\mathbf{Q}(t)) + V\mathbb{E}[g(t)|\mathbf{Q}(t)] \le B + V\bar{g}^{opt}. \tag{A4}$$

Then, by taking the expectations of (A4), summing up from $0, 1, \ldots, t$ and by dividing the sum by V, we obtain, for any $t > 0$,

$$\begin{aligned}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}[g(t)] &\le \bar{g}^{opt} + \frac{B}{V} + \frac{\mathbb{E}[L(\mathbf{Q}(0))]}{Vt} - \frac{\mathbb{E}[L(\mathbf{Q}(t+1))]}{Vt} \\ &\le \bar{g}^{opt} + \frac{B}{V} + \frac{\mathbb{E}[L(\mathbf{Q}(0))]}{Vt}\,.\end{aligned} \tag{A5}$$

Thus, Theorem 1 can be proved by taking $t \to \infty$.

Appendix B: Proof of Theorem 1

By taking expectations on (A4) and summing up from $0, 1, \ldots, t$, we obtain

$$\mathbb{E}[L(\mathbf{Q}(t+1))] \le \mathbb{E}[L(\mathbf{Q}(0))] + [B + V(\bar{g}^{opt} - g_{\min})]\cdot t\,, \tag{B1}$$

where $g_{\min} \le \mathbb{E}[g(t)]$. Next, by the definition of the Lyapunov function, we have

$$\mathbb{E}[Z_i^2(t)] \le 2\mathbb{E}[L(\mathbf{Q}(0))] + [B + V(\bar{g}^{opt} - g_{\min})]\cdot t\,. \tag{B2}$$

Because the variance of $|Z_i(t)|$ is always non-negative, we have $\mathbb{E}[Z_i^2(t)] - \mathbb{E}[|Z_i(t)|]^2 \ge 0$. Hence, for any $t > 0$. By dividing both sides by t and taking $t \to \infty$, we prove

$$\lim_{t\to\infty} \frac{\mathbb{E}[Z_i(t)]}{t} \le 0\,, \tag{B3}$$

which shows the mean-rate stability.

Acknowledgment

This work is partly supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative, Singapore EIRP02 (Grant NRF2012EWT-EIRP002-013).

References

[1] D. o. E. Population Division and S. Affairs, "World population aging 2015," United Nations, Tech. Rep., 2015.

[2] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 579–590, 2013.

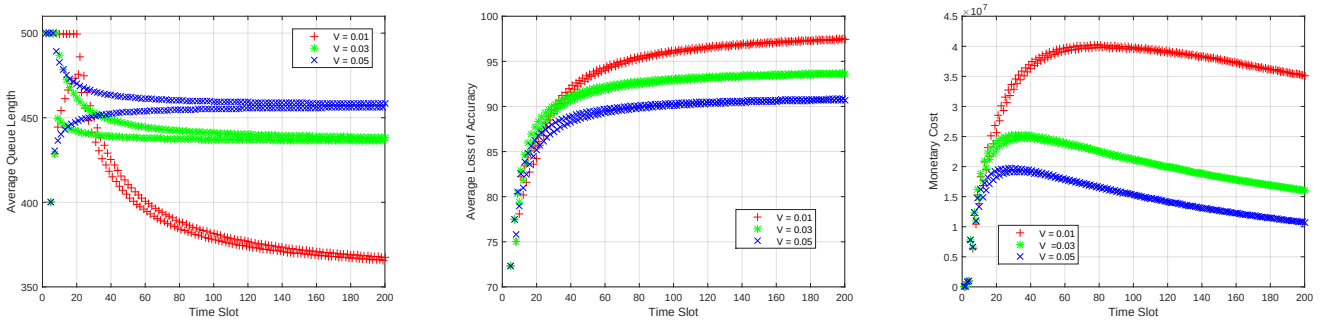
[3] H. Wen, Z. Xiao, A. Markham, and N. Trigoni, "Accuracy estimation for sensor systems," *Mobile Computing, IEEE Transactions on*, vol. 14, no. 7, pp. 1330–1343, 2015.

[4] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones," in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2009, pp. 85–98.

[5] H. Wen, Z. Xiao, N. Trigoni, and P. Blunsom, "On assessing the accuracy of positioning systems in indoor environments," in *European Conference on Wireless Sensor Networks*. Springer, 2013, pp. 1–17.

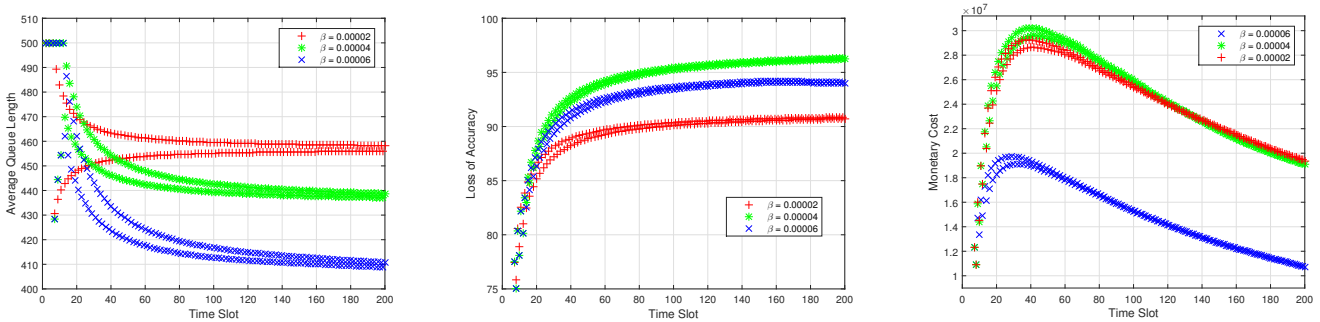
[6] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*. IEEE, 2013, pp. 530–539.

[7] D. Wu, Y. Cai, and M. Guizani, "Asynchronous flow scheduling for green ambient assisted living communications," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 64–70, January 2015.



(a) Average Queue Length (b) Average Loss of Accuracy (c) Average Monetary Cost

Figure 3. **Impact of V.** We set $\beta \cdot V = 0.000001$ in this set of simulations.



(a) Average Queue Length (b) Average Loss of Accuracy (c) Average Monetary Cost

Figure 4. **Impact of β.** We set $\beta \cdot V = 0.000001$ in this set of simulations.

[8] Y. Dong, L. Zhou, J. Chen, B. Zheng, and J. Cui, "Energy efficient virtual machine consolidation in mobile media cloud," in *Picture Coding Symposium (PCS), 2015*, May 2015, pp. 248–252.

[9] L. Zhou, "On data-driven delay estimation for media cloud," *IEEE Transactions on Multimedia*, vol. 18, no. 5, pp. 905–915, May 2016.

[10] L. Zhou and H. Wang, "Toward blind scheduling in mobile media cloud: Fairness, simplicity, and asymptotic optimality," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 735–746, June 2013.

[11] Y. Jin, Y. Wen, H. Hu, and M.-J. Montpetit, "Reducing operational costs in cloud social tv: an opportunity for cloud cloning," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1739–1751, 2014.

[12] Y. Dong, L. Zhou, Y. Jin, and Y. Wen, "Improving energy efficiency for mobile media cloud via virtual machine consolidation," *Mobile Networks and Applications*, vol. 20, no. 3, pp. 370–379, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11036-015-0595-2

[13] J. He, Y. Wen, J. Huang, and D. Wu, "On the cost–qoe tradeoff for cloud-based video streaming under amazon ec2's pricing models," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 4, pp. 669–680, 2014.

[14] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, and R. Thibaux, "Intel lab data," *Online dataset*, 2004.

[15] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review*, vol. 22, no. 3, pp. 177–210.

[16] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "Maui: making smartphones last longer with code offload," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 49–62.

[17] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 945–953.

[18] J. Kwak, O. Choi, S. Chong, and P. Mohapatra, "Processor-network speed scaling for energy-delay tradeoff in smartphone applications," *IEEE/ACM Transactions on Networking*, vol. PP, no. 99, pp. 1–14, 2015.

[19] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 7, pp. 1625–1637, 2013.

[20] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[21] H. Hu, Y. Wen, T.-S. Chua, J. Huang, W. Zhu, and X. Li, "Joint content replication and request routing for social video distribution over cloud cdn: A community clustering method," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2016, in press.

[22] H. Hu, Y. Wen, T.-S. Chua, Z. Wang, J. Huang, W. Zhu, and D. Wu, "Community based effective social video contents placement in cloud centric cdn network," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.